Single Server Queues with State-Dependent Hawkes Arrivals

BO LI* AND GUODONG PANG†

ABSTRACT. We study single-server queues with Hawkes arrivals whose intensity process depends on the queue length or workload through the self-exciting function, and with i.i.d. general service times, under the first-come first-served discipline. We prove the functional law of large numbers (FLLNs) and functional central limit theorems (FCLTs) for the joint processes of the arrivals, queue-length and workload processes, in the heavy traffic regime. The fluid limit is given by a set of nonlinear integral equations such that the fluid queue or workload has a reflection at zero. We analyze the transient and equilibrium behaviors of the fluid limit, in particularly, identifying the equilibrium points for the queue or workload fluid model. We assume that the fluid limit is at an equilibrium point in order to establish the FCLTs for the joint diffusion-scaled arrival, queue-length and workload processes in the critically load regime. When the equilibrium point of the queue is at zero, the limit for the joint processes satisfies a stochastic differential equation such that the queue-length or workload limit has a reflection at zero. In particular, the queue-length or workload process is equivalent in distribution to a generalized (possibly nonlinear drift) Ornstein-Uhlenbeck (OU) diffusion with reflection at zero. When the equilibrium point is positive, the limit for the joint processes is a diffusion process without reflection, in particular, the queue-length or workload process is a generalized OU diffusion. Because of the interacting effects between the Hawkes arrivals and queue-length or workload processes, the standard approaches to prove functional limit theorems for Hawkes processes and for single-server queues cannot be applied directly. We develop a new method to prove the functional limit theorems for the joint Hawkes and queueing dynamics, by using a localization argument and exploiting convergence of martingales and stochastic integrals.

1. INTRODUCTION

Single server queues are fundamental models in applied probability and stochastic models [2, 8, 50], with many applications in service operations, healthcare, manufacturing/production, data centers and computing. The standard single server queues assume Poisson or renewal arrival processes and i.i.d. service times. However, motivated from various applications, the arrival and service processes have been extended in several ways, such as time-dependent, state-dependent and/or random rates. In particular, single server queues with state-dependent arrival and service rates have been extensively studied in [21, 30, 31, 49, 53, 20, 5, 7, 29, 1].

Since the introduction of Hawkes process [22], it has been used in many applications, such as mathematical finance [4, 23], seismology [43], neuron science [39, 41, 11]. See the recent monograph [34]. Depending on its own path, Hawkes process captures the self-exciting, clustering and over-dispersion effects in the dynamics. It has been recently used as an arrival process in queueing models, for example, single server queues with Hawkes arrivals [9, 10], infinite-server queues with Hawkes arrival [19, 15, 32], infinite-server queues with both Hawkes arrival and service processes [45], and multi-server queues [14, 36]. However, in these studies, Hawkes process is only used as an input to the queueing dynamics, while the state of the queues does not affect the arrival process. In

*SCHOOL OF MATHEMATICS AND LPMC, NANKAI UNIVERSITY, TIANJIN, 300071 CHINA

† DEPARTMENT OF COMPUTATIONAL APPLIED MATHEMATICS AND OPERATIONS RESEARCH, GEORGE R. BROWN SCHOOL OF ENGINEERING, RICE UNIVERSITY, HOUSTON, TX 77005

E-mail addresses: libo@nankai.edu.cn, gdpang@rice.edu.

Date: December 27, 2024.

Key words and phrases. Single-server queues, state-dependent Hawkes arrival process, queue-length or workload dependent self-exciting intensity, heavy traffic limits, (reflected) generalized Ornstein-Uhlenbeck diffusion.

this paper, we study single server queues with Hawkes arrivals, whose intensity process depends on the state of the queue or workload process through the self-exciting function, and with i.i.d. service times, under the first-come first-served (FCFS) discipline. Such models have not been studied in queueing theory.

In [51, 42], Hawkes processes are used to model limit order books in finance, where the intensities of the multivariate Hawkes arrivals depend on the stock price, but only empirical studies are conducted. Path-dependent Polya arrival processes are also used as input to the single-server queueing models [17, 18], where heavy-traffic limits are established, but Polya processes differs drastically from Hawkes processes. Shot noise processes with Hawkes arrivals that depend on the state of the process are introduced recently in [38], where functional limit theorems are established in the conventional scaling regime. In this paper, we aim to establish the functional law of large numbers (FLLNs) and the functional central limit theorem (FCLTs) for the joint Hawkes arrivals and queueing dynamics of our single-server models in heavy traffic.

We start with the model where the intensity of Hawkes arrivals depends on the queue length (see equation (2.3)). We first establish the FLLN for the joint fluid-scaled arrival, queueing and workload processes, and find that the joint fluid model of the arrival and queue processes solves a set of nonlinear integral equation with reflection (Theorem 3.1), in particular, the fluid queue length solves a nonlinear differential equation with reflection at zero. We analyze the transient and equilibrium behaviors of the fluid model, by defining properly the sub-critical, critical and super-critical regimes through a state-dependent traffic intensity function, in which the function of the state is determined by the self-exciting function of the Hawkes arrival intensity. We give several examples of this state functional to illustrate the behavior of the fluid queueing model, including monotone (power or exponential) functions and a non-monotone sinusoidal function. For the monotone functions, we identify conditions for zero and nonzero equilibrium points, and for the sinusoidal function, we identify conditions for a countable number of equilibrium points.

We then establish the FCLT for the joint diffusion-scaled arrival, queueing and workload processes that are centered around a critical equilibrium point of the fluid model that is either zero or positive (Theorem 4.1). In the zero equilibrium point case, the diffusion limit for the joint arrival and queueing processes is given by a two-dimensional stochastic differential equation (SDE) such that the limiting queueing process has a reflection at zero. In particular, the diffusion limit for the queueing process can be also written equivalently in distribution as a generalized Ornstein–Uhlenbeck (OU) diffusion with reflection at zero, and a possibly nonlinear drift (see Remark 4.2). We give two examples in which the drift comes as a linear function or a power function (see Remarks 4.3 and 4.4). In the case of a positive equilibrium point, the diffusion limit for the joint arrival and queueing processes is given by a two-dimensional diffusion without reflection because of the nonzero centering for the diffusion-scaled queueing processes. In this case, the diffusion limit for the queue length process can be also written as a generalized OU diffusion with a possibly nonlinear drift. It is also worth noting that although the fluid limit captures both path and state dependences, the diffusion limit only captures the state dependence since the centering for the diffusion-scaled processes is at a critical equilibrium point.

We obtain analogous fluid and diffusion limits for the model where the intensity of Hawkes arrivals depends on the workload (see equation (5.2)). We obtain a similar set of nonlinear integral equations with reflection for the fluid limit of the arrival and workload processes (Theorem 5.1) and similar characterizations of the associated transient and equilibrium behavior (but the state-dependent traffic intensity is now defined through the state of the workload). We then obtain similar two-dimensional SDEs for the diffusion limits of the arrival and workload processes in the two cases of equilibrium points as above (Theorem 5.2).

The proofs for the FLLNs and FCLTs require novel techniques for the following reasons. Scaling limits for standard Hawkes processes are established in [3] under the same scaling regime as our paper. Without the state dependence, there exists a renewal equation representation for the

expected Hawkes counting process, which plays a critical role to prove the FLLN and FCLT. With the state-dependence, such a renewal equation representation can no longer be derived. See further discussions in Section 6.1. Moreover, for standard single server queues, given the scaling limits (FLLN and FCLT) for the standard Hawkes (or any independent) arrival processes, the standard methods of proving FLLN and FCLT for the queue and workload processes in [8, 50] can then be applied; this is a special case of the parallel single-server queues with multivariate Hawkes arrivals in the case without abandonment in [37]. However, with the state-dependence in the Hawkes arrivals, this standard methods for single-server queues can no longer be applied. The state-dependence is also different from those studied in queues and networks with state-dependent arrival/service rates in [49, 52, 40, 35], because of the dependence is through the self-exciting function, which is path-dependent by definition. Also, the recent studies of single-server queues with path-dependent Polya arrival in [17, 18] differ drastically since there is no state-dependence in the arrival intensity process.

The new approach starts with a construction of the LLN-scaled processes that resemble the limiting fluid model, together with some asymptotically negligible residual terms, and uses a localization technique to facilitate the proofs for the convergence. Similarly, for the CLT-scaled processes, we also construct a representation resembling the limiting diffusions together with residual terms. The proofs exploit martingale properties for the localized processes and the associated weak convergence criteria in [26] and convergence of stochastic integrals in [33]. The proofs for the convergence of the residual terms in the constructions to zero in the fluid and diffusion scales are very challenging, and require refined estimates on the increments of the joint arrival and queueing dynamics. Moreover, the analysis for the residual terms at the diffusion scale requires further conditions on the tail behavior of the self-exciting function. Moreover, the proofs for the FCLTs in the two different cases of critical equilibrium points exploit the use of the weak limit theorems for stochastic integrals and SDEs with or without reflections in [33]. The approach in this paper further extends the one that was recently developed in [38] for interactive Hawkes shot noise process. Since shot noise process may be regarded as a generalization of compound processes (still regarded only as an input process in some sense) and may be approximated by a compound process under the conventional scaling regime, the extension is highly nontrivial to single-server queues with both input and output dynamics. We believe that the new approach developed in this paper can be used to study other queueing models with state-dependent Hawkes arrivals.

1.1. **Organization of the paper.** The paper is organized as follows. In Section 2, we give the detailed description of the single-server queueing model with queue-dependent Hawkes arrivals, and also present the scalings for the associated processes and the assumptions. In Section 3, we state the FLLN and discuss the asymptotic behavior of the fluid limit. In Section 4, we state the FCLT and discuss various diffusion limits. In Section 5, we describe the single-server queueing model with workload-dependent Hawkes arrivals, and present both the FLLN and FCLT results. In Section 6, we prove the FLLN and FCLT for the model with queue-dependence. In Section 7, we provide sketch proofs for the FLLN and FCLT in the workload-dependent case. In the Appendix, we collect some additional examples to illustrate the transient and equilibrium behaviors of the fluid limit.

1.2. **Notation.** $\mathbb{N}$ denotes the set of natural numbers. $\mathbb{R}$ and $\mathbb{R}_+ = [0, \infty)$ denote the spaces of real and nonnegative numbers, the Borel sets of $\mathbb{R}_+$ is denoted by $\mathcal{B}(\mathbb{R}_+)$. Let $\mathbb{R}^k$ be the space of k-dimensional real numbers. Let $x^+ := x \vee 0$ and $x^- = -(x \wedge 0)$ for $x \in \mathbb{R}$. Let $\mathbb{D}^k = \mathbb{D}(\mathbb{R}_+, \mathbb{R}^k)$ denote $\mathbb{R}^k$-valued function space of all càdlàg functions on $\mathbb{R}_+$. $(\mathbb{D}^k, J_1)$ denotes space $\mathbb{D}^k$ equipped with the Skorohod J_1 topology (see [6, 50]). We write $\mathbb{D}$ when $k = 1$, and also denote by $\mathbb{C}^k$ and $\mathbb{C}$ the subspace of continuous functions of $\mathbb{D}^k$ and $\mathbb{D}$, respectively. For a measurable function f on $\mathbb{R}$, $||f||_1 := \int_{\mathbb{R}} f(y)\, dy$ denotes the L^1 norm of f whenever integrable. For a real valued random variable ξ, we write $\xi \in L^2(\mathbb{P})$ if $\mathbb{E}[\xi^2] < \infty$. Notations $\to$ and $\Rightarrow$ mean convergence of real numbers and

convergence in distribution, respectively. *u.o.c.* on $\mathbb{R}$ or $\mathbb{R}_+$ means "uniformly on every compact set" on $\mathbb{R}$ or $\mathbb{R}_+$.

2. Hawkes/GI/1 queue with queue-length dependent intensity

2.1. **Model description.** We consider a single-server queue with Hawkes arrivals and i.i.d. service times under the first-come first-served (FCFS) discipline, where the intensity of Hawkes arrival process depends on the state of the queue through the self-exciting function. On a filtered probability space $\{\Omega, \mathscr{F}, \{\mathscr{F}(t)\}_{t\geq 0}, \mathbb{P}\}$, let $A = \{A(t)\}_{t\geq 0}$ be the Hawkes arrival process with arrival times $\{\tau_i, i \geq 1\}$, and $\{\xi_j, j \in \mathbb{Z}\}$ be the sequence of i.i.d. exogenous service times, that is, for some generic variable ξ,

$$\mathbb{P}\big(\xi_j \in [x, x+\mathrm{d}x] \big| \mathscr{F}(\tau_j-), \tau_j\big) = \mathbb{P}\big(\xi \in [x, x+\mathrm{d}x]\big) \quad \forall j \geq 1, \tag{2.1}$$

where $\{\mathscr{F}(t)\}_{t\geq 0}$ is the filtration generated by $\{Q_0, A, \xi\}$, that is,

$$\mathscr{F}(0) = \sigma\{Q_0\} \vee \sigma\{\xi_{-j}, j \geq 1\} \quad \text{and} \quad \mathscr{F}(t) = \mathscr{F}(0) \vee \sigma\{\tau_j, \xi_j, 1 \leq j \leq A(t)\}, \tag{2.2}$$

and the sequence $\{\xi_{-j}, j \geq 1\}$ represents the service times for the jobs initially in the system, and is independent of Q_0. Let $Q = \{Q(t)\}_{t\geq 0}$ be the queue-length process, and $W = \{W(t)\}_{t\geq 0}$ be the workload process. The conditional intensity $\lambda = \{\lambda(t)\}_{t\geq 0}$ of the Hawkes arrival process A is defined as

$$\mathbb{P}\big(A \text{ has a jump in } [t, t+\mathrm{d}t] \big| \mathscr{F}(t-)\big) = \lambda(t)\,\mathrm{d}t\,,$$

where $\lambda(t)$ depends on the state of the queue length and takes the form

$$\lambda(t) = \lambda_0 + \sum_{j\geq 1} h\big(t-\tau_j, Q(\tau_j-)\big) = \lambda_0 + \int_0^t h\big(t-s, Q(s-)\big)\,\mathrm{d}A(s)\,. \tag{2.3}$$

Here, $\lambda_0 > 0$ is a constant represents the baseline intensity, and $h : \mathbb{R}_+^2 \to \mathbb{R}_+$ is a deterministic and measurable function, referred to as the self-exciting function, where we understand that $h(t, y) = 0$ for $t < 0$, and $\mathrm{d}A$ is understand as the Lebesgue-Stieltjes integral induced by A. Let $\Lambda = \{\Lambda(t) : t \geq 0\}$ be the cumulative intensity process, that is,

$$\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s, \quad \forall t \geq 0. \tag{2.4}$$

We remark that if $t \to h(t, y)$ is locally integrable on $\mathbb{R}_+$ for every $y \in \mathbb{R}_+$, the conditional intensity function determines the distributional properties of A uniquely (see, e.g., [12, Proposition 7.3.IV]), and hence, the Hawkes arrival process is well-defined up to the possible explosion time $\tau_\infty = \lim_{n\to\infty} \tau_n \in (0, \infty]$.

To describe the queueing dynamics, we define

$$U(k) = \sum_{j\geq 1} \xi_{-j}\,\mathbf{1}\big(j \leq k \wedge Q_0\big) + \sum_{j\geq 1} \xi_j\,\mathbf{1}\big(j \leq (k-Q_0)^+\big). \tag{2.5}$$

Observe that U is a random partial sum of i.i.d. positive variables. This is slightly different from the usual definition of partial sums in the study of standard GI/GI/1 queues (see, e.g., [8] and [50]). The formulation is necessary because of the dependency between the arrival process and queueing process (hence the service times).

Let $S = \{S(t)\}_{t\geq 0}$ be the renewal process associated with service times $\{\xi_j\}_{j\in\mathbb{Z}}$, that is,

$$S(t) := \max\{k \geq 1 : U(k) \leq t\}, \quad t > 0, \tag{2.6}$$

with $S(0) = 0$ and the convention that $\max\emptyset = 0$. $S(t)$ is interpreted as the number of jobs that can be potentially completed by time t when the server is busy all the time. It is clear that

$S(U(Q_0)) = Q_0$ and $\sigma\{S(t), t \in [0, U(Q_0)]\} \in \mathscr{F}(0)$. The queueing process $Q = \{Q(t)\}_{t\geq 0}$ can be written as

$$Q(t) = Q_0 + A(t) - S\big(B(t)\big), \tag{2.7}$$

where Q_0 is the initial number of jobs in the queue, S is the renewal process defined in (2.6), and $B = \{B(t)\}_{t\geq 0}$ is the cumulative busy time over $[0, t]$, that is,

$$B(t) = \int_0^t \mathbf{1}\big(Q(u) > 0\big)\, \mathrm{d}u\,. \tag{2.8}$$

The workload process can be written as

$$W(t) = U\big(Q_0 + A(t)\big) - B(t), \quad t \geq 0. \tag{2.9}$$

Define the cumulative idleness process $I = \{I(t)\}_{t\geq 0}$ as

$$I(t) = t - B(t) = \int_0^t \mathbf{1}\big(Q(u) = 0\big)\, \mathrm{d}u, \quad t \geq 0. \tag{2.10}$$

With the notations above, one can find that $\big(A, Q, W, \Lambda, B, I, S(B), U(Q_0 + A)\big)$ is $\{\mathscr{F}(t)\}_{t\geq 0}$-adapted.

Remark 2.1. Recalling that for standard Hawkes processes $A(t)$ with an exponential self-exciting process $\lambda(t) = \lambda_0 + \int_0^t \beta e^{-\gamma(t-s)} \mathrm{d}A(s)$ for $\beta, \gamma > 0$, it is well known that $\lambda(t)$ satisfies the SDE:

$$\mathrm{d}\lambda(t) = \gamma\big(\lambda_0 - \lambda(t)\big)\mathrm{d}t + \beta\, \mathrm{d}A(t)$$

and hence, the joint process (A, λ) is Markov (see, e.g., [4]). In our queueing model, if $h(t, y) = \gamma e^{-\gamma t} H(y)$ for some $\gamma > 0$ and measurable function $H : \mathbb{R}_+ \to \mathbb{R}_+$, and service times are i.i.d. with an exponential distribution, one can write

$$\begin{aligned} \mathrm{d}\,\lambda(t) &= \gamma\big(\lambda_0 - \lambda(t)\big)\mathrm{d}t + \gamma H\big(Q(t-)\big)\, \mathrm{d}A(t), \\ \mathrm{d}\,Q(t) &= \mathrm{d}A(t) - \mathrm{d}S(B(t)), \end{aligned}$$

where $B(t)$ is the busy time process given in (2.8), and $S(t)$ becomes a Poisson process. It can be easily shown that (A, λ, Q) is a Markov process. This is an extension of the Markov case of standard Hawkes processes, as well as an extension of the standard M/M/1 queues as a birth-death Markov process.

2.2. **Scalings and Assumptions.** We establish the FLLN and FCLT for the joint processes (A, Q, W), for which we consider a sequence of the queueing models indexed by n, that is, on the filtered probability space $(\Omega, \mathscr{F}^{(n)}, \{\mathscr{F}^{(n)}(t)\}_{t\geq 0}, \mathbb{P})$ with

$$\mathscr{F}^{(n)}(0) = \sigma\{Q_0^{(n)}\} \vee \sigma\{\xi_{-j}^{(n)}, j \geq 1\} \quad \text{and} \quad \mathscr{F}^{(n)}(t) = \mathscr{F}^{(n)}(0) \vee \sigma\{\tau_j^{(n)}, \xi_j^{(n)}, 1 \leq j \leq A^{(n)}(t)\}, \tag{2.11}$$

the conditional intensity process for the arrival process is defined by

$$\lambda^{(n)}(t) = \lambda_0^{(n)} + \int_0^t h^{(n)}\big(t - s, Q^{(n)}(s-)\big)\, \mathrm{d}A^{(n)}(s), \tag{2.12}$$

and

$$\begin{aligned} Q^{(n)}(t) &= Q_0^{(n)} + A^{(n)}(t) - S^{(n)}\big(B^{(n)}(t)\big), \\ W^{(n)}(t) &= U^{(n)}\big(Q_0^{(n)} + A^{(n)}(t)\big) - B^{(n)}(t), \end{aligned}$$

and the sequence of exogenous service times $\{\xi_j^{(n)}\}_{j\in\mathbb{Z}}$ has c.d.f. $F^{(n)}$ on $\mathbb{R}_+$, and where $U^{(n)}$, $S^{(n)}$ and $B^{(n)}$ are the associated cumulative service times, the renewal process and the busy time process, in (2.5), (2.6) and (2.8), respectively. We make the following assumptions on the primitives.

Assumption A1. *(i)* $\lambda_0^{(n)} \to \lambda_0 > 0$ *and* $\mu^{(n)} \to \mu > 0$ *as* $n \to \infty$.

(ii) For some continuous function $H : \mathbb{R}_+ \to [0, 1)$,

$$H^{(n)}(y) := \int_0^\infty h^{(n)}(u, ny)\,\mathrm{d}u \to H(y) \quad \textit{u.o.c. on } \mathbb{R}_+. \tag{2.13}$$

(iii) For every $k > 0$ *and* $\varepsilon > 0$,

$$\limsup_{n\to\infty} \sup_{y\le k} \int_{n\varepsilon}^\infty h^{(n)}(u, ny)\,\mathrm{d}u = 0. \tag{2.14}$$

Remark 2.2. *For a standard Hawkes process, that is,* $h(t, y) = h(t)$ *in (2.3), by the immigration-birth representation, every point gives birth to* $\|h\|_1$ *number of children if* $\|h\|_1 < \infty$, *and generates a family of size* $(1 - \|h\|_1)^{-1}$ *if* $\|h\|_1 < 1$. *The condition* $\|h\|_1 < 1$ *is referred to as the stability condition in the literature, under which a stationary version of Hawkes process exists, and the counting process increases at a constant speed of* $\lambda_0(1-\|h\|_1)^{-1}$ *asymptotically, c.f.* [3, Theorem 1]. *In our model with the state-dependence in the intensity process, every point produces children of size* $\int_0^\infty h(t, y)\,\mathrm{d}t$; *hence, the condition* $H(y) < 1$ *plays the role of the stability condition analogous to the standard case.*

We also make the following assumptions on the service times, under which we obtain the FLLN and FCLT for the processes $U^{(n)}$ and $S^{(n)}$ (see Lemma 6.1).

Assumption A2. *Suppose the service times* $\{\xi_j^{(n)}\}_{j\in\mathbb{Z}}$ *are i.i.d. random variables satisfying*

$$\mathbb{E}\big[\xi^{(n)}\big] = m^{(n)} \to m = \mu^{-1} \quad \textit{and} \quad \mathbb{E}\big[\xi^{(n)}; \xi^{(n)} > n\varepsilon\big] \to 0 \quad \forall \varepsilon > 0, \tag{2.15}$$

$$Var(\xi^{(n)}) = \big(\sigma^{(n)}\big)^2 \to \sigma^2 \quad \textit{and} \quad \mathbb{E}\big[(\xi^{(n)})^2; \xi^{(n)} > \sqrt{n}\varepsilon\big] \to 0 \quad \forall \varepsilon > 0. \tag{2.16}$$

We next introduce the following concept of state-dependent traffic intensity.

Definition 2.1. *Define the state-dependent traffic intensity function*

$$\rho(y) := \frac{\lambda_0}{\mu(1 - H(y))}, \quad y \ge 0. \tag{2.17}$$

Similarly, for the n^{th}*-system, define*

$$\rho^{(n)}(y) := \frac{\lambda_0^{(n)}}{\mu^{(n)}(1 - H^{(n)}(y))}, \quad y \ge 0. \tag{2.18}$$

Assumption A3. *If there exists* $y_0 > 0$ *such that* $\rho(y) > 1$ *for every* $y \ge y_0$, *the following holds:*

$$\int_{y_0}^\infty \frac{1}{\rho(y) - 1} dy = +\infty\,. \tag{2.19}$$

Remark 2.3. *The condition in Assumption A3 will guarantee that the fluid limit does not "explode" (going to infinity in finite time); see further discussions in Remark 3.4. We give an example to illustrate the condition. If* $H(y) = 1 - (2 + y)^{-2} - \frac{1}{2}\sin^2 y^\theta$ *for some* $\theta \ne 0$ *and* $2\lambda_0 > \mu$, *we have*

$$0 < H(y) < 1 \quad \textit{and} \quad \rho(y) = \frac{\lambda_0}{\mu}\frac{2}{2(2 + y)^{-2} + \sin^2 y^\theta} \ge \frac{2\lambda_0}{\mu(1 + 2(2 + y)^{-2})} > 1,$$

for y *large enough, and thus* $\bar{s}_{y_0} = \infty$ *for some* $y_0 > 0$. *One can find that*

$$\frac{1}{\mu(\rho(y) - 1)} = \frac{2(2 + y)^{-2} + \sin^2 y^\theta}{2\lambda_0 - \mu \cdot (\sin^2 y^\theta + 2(2 + y)^{-2})}.$$

For y large enough, we have for some $c_0 > 1$

$$\frac{1}{c_0}\big(2(2+y)^{-2} + \sin^2 y^\theta\big) \le \frac{1}{\mu(\rho(y)-1)} \le c_0\big(2(2+y)^{-2} + \sin^2 y^\theta\big),$$

which imples Assumption A3 holds if and only if $1 + 2\theta \ge 0$.

3. Functional Law of Large Numbers

Define the LLN-scaled processes

$$\big(\bar{A}^{(n)}, \bar{Q}^{(n)}, \bar{W}^{(n)}\big)(t) = n^{-1}\big(A^{(n)}, Q^{(n)}, W^{(n)}\big)(nt), \quad t \ge 0. \tag{3.1}$$

Theorem 3.1. *Under Assumptions A1, A2-(2.15) and A3, assuming that $\bar{Q}_0^{(n)} := n^{-1}Q_0^{(n)} \to \bar{q}_0$ for some constant $\bar{q}_0 \ge 0$, $(\bar{A}^{(n)}, \bar{Q}^{(n)}, \bar{W}^{(n)})$ is a $\mathbb{C}$-tight family in $(\mathbb{D}^3, J_1)$, that is, every limit $(\bar{A}, \bar{Q}, \bar{W})$ takes value in $\mathbb{C}^3$ such that $\bar{W} = \mu^{-1}\bar{Q}$ and $(\bar{A}, \bar{Q})$ satisfies the set of nonlinear integral equations such that $\bar{Q}$ has a reflection at 0:*

$$\begin{aligned} \bar{A}(t) &= \lambda_0 t + \int_0^t H\big(\bar{Q}(s)\big)\,\mathrm{d}\bar{A}(s), \\ \bar{Q}(t) &= \bar{q}_0 + \bar{A}(t) - \mu t + \mu \bar{I}(t), \end{aligned} \tag{3.2}$$

where $\bar{I}$ is the minimal nondecreasing function in $\mathbb{C}$ so that $\bar{Q}(t) \ge 0$ for every $t \ge 0$, and $\bar{I}$ increases only when $\bar{Q}$ is zero, that is,

$$\int_{[0,\infty)} \mathbf{1}\big(\bar{Q}(t) > 0\big)\,\mathrm{d}\bar{I}(t) = 0.$$

One can find from (3.2) that $\bar{Q}$ solves the following nonlinear ODE with reflection at 0:

$$\begin{aligned} \mathrm{d}\bar{Q}(t) &= \Big(\frac{\lambda_0}{1 - H(\bar{Q}(t))} - \mu\Big)\,\mathrm{d}t + \mu\,\mathrm{d}\bar{I}(t) \\ &= \mu\big(\rho(\bar{Q}(t)) - 1\big)\,dt + \mu\,\mathrm{d}\bar{I}(t)\,, \end{aligned} \tag{3.3}$$

with $\bar{Q}(0) = \bar{q}_0$. Given a solution of $\bar{Q}$ above, one can obtain $\bar{A}$ from (3.2) directly.

By Assumption A1-(ii), we know that $\rho(y) \in (0, \infty)$ in (2.17) for every $y \in \mathbb{R}_+$. To have a better understanding about the solution to (3.2), we define the sets:

$$\mathcal{L}^+ := \{y \in \mathbb{R}_+ : \rho(y) > 1\}, \quad \mathcal{L}^- := \{y \in \mathbb{R}_+ : \rho(y) < 1\}, \quad \mathcal{L}^= := \{y \in \mathbb{R}_+ : \rho(y) = 1\}, \tag{3.4}$$

which are referred to as *the sup-critical regime*, *the sub-critical regime* and *the critical regime*, respectively. It is clear that $\mathcal{L}^+$ and $\mathcal{L}^-$ are open subsets of $\mathbb{R}_+$ by the continuity of H.

Remark 3.1. *For the standard Hawkes in Remark 2.2, $H(y) \equiv ||h||_1$ is a constant function for all $y \ge 0$, A is independent of the queue-length process. It is proved in* [3, Theorem 1] *that*

$$\bar{A}^{(n)}(t) \to \bar{A}(t) = \frac{\lambda_0 \cdot t}{1 - ||h||_1} = \frac{\lambda_0}{\mu(1 - ||h||_1)} \cdot \mu t \quad \textit{almost surely and in } L^2(\mathbb{P}).$$

For the single-server queue with Hawkes arrivals as input, it is clear that $\frac{\lambda_0}{\mu(1-||h||_1)}$ is the traffic intensity, its value being $> 1, < 1, = 1$ corresponds to the overloaded, underloaded and critically loaded regimes, respectively. This observation inspires the definitions of $\mathcal{L}^+, \mathcal{L}^-, \mathcal{L}^=$ above.

Proposition 3.1. *Let $\bar{Q}$ be a weak limit for some convergent subsequence of $\{\bar{Q}^{(n)}\}_{n\ge1}$ in Theorem 3.1 with $\bar{Q}(0) = \bar{q}_0 \ge 0$.*

(i) $\bar{Q}$ increases in $\mathcal{L}^+$. If $\bar{q}_0 \in \mathcal{L}^+$, let $\bar{s}_{\bar{q}_0} := \inf\{y > \bar{q}_0, [\bar{q}_0, y] \subset \mathcal{L}^+\} \in (\bar{q}_0, +\infty]$, then

$$\frac{1}{\mu}\int_{\bar{q}_0}^{s} \frac{1}{\rho(y) - 1}\mathrm{d}y\Big|_{s=\bar{Q}(t)} = t, \quad \forall t < \frac{1}{\mu}\int_{\bar{q}_0}^{\bar{s}_{\bar{q}_0}} \frac{1}{\rho(y) - 1}\mathrm{d}y; \tag{3.5}$$

(ii) $\bar{Q}$ *decreases in* $\mathcal{L}^-$. *If* $\bar{q}_0 \in \mathcal{L}^-$, *let* $\underline{s}_{\bar{q}_0} := \sup\{y < \bar{q}_0, [y, \bar{q}_0] \subset \mathcal{L}^-\} \in [0, \bar{q}_0]$, *then*

$$\frac{1}{\mu}\int_s^{\bar{q}_0} \frac{1}{1-\rho(y)}\mathrm{d}y\Big|_{s=\bar{Q}(t)} = t, \quad \forall t < \frac{1}{\mu}\int_{\underline{s}_{\bar{q}_0}}^{\bar{q}_0} \frac{1}{1-\rho(y)}\mathrm{d}y; \tag{3.6}$$

(iii) $\bar{Q}$ *stays constant in the interior of* $\mathcal{L}^=$. *If* $\bar{q}_0 \in (a,b) \subset \mathcal{L}^=$ *for some* $b > a > 0$, *then*

$$\bar{Q}(t) = \bar{q}_0 \quad \textit{and} \quad \bar{A}(t) = \mu\, t, \quad \forall t > 0;$$

(iv) $\bar{I} > 0$ *only if* $0 \in \mathcal{L}^-$. *If* $\bar{q}_0 = 0 \in \mathcal{L}^-$, *then* $\rho(0) < 1$,

$$\bar{Q}(t) = 0, \quad \bar{A}(t) = \mu\rho(0)\, t \quad \textit{and} \quad \bar{I}(t) = \big(1-\rho(0)\big)\, t, \quad \forall t > 0.$$

Proof. The monotonicity of $\bar{Q}$ in $\mathcal{L}^+$ and $\mathcal{L}^-$ of the first two assertions (i) and (ii) follows directly from the ODE expression of $\bar{Q}$ in (3.3) above.

Since $\bar{I} = 0$ before $\bar{Q}$ hitting 0, we have for the case $\bar{q}_0 \notin \mathcal{L}^=$, $\rho(y) \neq 1$ for $y \in (\underline{s}_{\bar{q}_0}, \bar{s}_{\bar{q}_0})$ and

$$\frac{\mathrm{d}\bar{Q}(t)}{\mu(\rho(\bar{Q}(t)) - 1)} = \mathrm{d}t \qquad \forall t < (\tau_{+,\bar{s}_{\bar{q}_0}} \wedge \tau_{-,\underline{s}_{\bar{q}_0}}) \leq \tau_{-,0}$$

where $\tau_{+,\underline{s}_{\bar{q}_0}}$ and $\tau_{-,\underline{s}_{\bar{q}_0}}$ are the passage times for $\bar{Q}$ defined in Remark 3.4. The identities for $\bar{Q}$ in (3.5) and (3.6) are proved by change of variables, respectively.

If, however, $\bar{q}_0 \in (a,b) \subset \mathcal{L}^=$ for some $b > a > 0$, we also have from (3.3) that

$$\bar{Q}(t) - \bar{q}_0 = \mu\int_0^t \big(\rho(\bar{Q}(s)) - 1\big)\,\mathrm{d}s = 0 \quad \text{for } t < \tau_{-,a} \wedge \tau_{+,b}.$$

We thus have $\bar{Q}(t) = \bar{q}_0$ for all $t < \tau_{-,a} \wedge \tau_{+,b}$, and which also implies $\tau_{-,a} \wedge \tau_{+,b} = \infty$, that is, $\bar{Q}$ stays constant in the interior of $\mathcal{L}^=$.

By the fact that $\bar{I}$ increases only on the set $\{t > 0 : \bar{Q}(t) = 0\}$, one can also find that $\bar{I} > 0$ if and only if 0 can be reached and $\rho(0) < 1$. Moreover, in the case $\bar{q}_0 = 0 \in \mathcal{L}^-$, we have $\bar{Q}(t) \equiv 0$ for all $t > 0$, which also gives the identity for $\bar{I}$ in case (iv). □

Remark 3.2. *The set* $\mathcal{L}^=$ *is a closed subset of* $\mathbb{R}_+$, *which can be rewritten as*

$$\mathcal{L}^= = \big\{y \in \mathbb{R}_+ : H(y) = 1 - \frac{\lambda_0}{\mu}\big\}.$$

If H *is a strictly monotone function on* $\mathbb{R}_+$, *there is at most one point in* $\mathcal{L}^=$ *as shown in Case* (3) *in Example 1. There can also be a countable number of points in* $\mathcal{L}^=$, *as shown in Cases* (3), (4) *and* (5) *in Example 2.*

Remark 3.3. *Proposition 3.1 does not discuss the case* $\bar{q}_0$ *being an isolated point in* $\mathcal{L}^=$, *so that only a local version of FLLN is derived. This is related to the uniqueness of a solution to* (3.3). *For example, if* $\mu(\rho(x) - 1) = \frac{3}{2}(x-1)^{1/3}$ *locally, then* $x_0 = 1 \in \mathcal{L}^=$ *and the local differential equation of* (3.3) *for* $\bar{Q}$ *is given by*

$$\mathrm{d}f(t) = \frac{3}{2}(f(t) - 1)^{1/3}\,\mathrm{d}t.$$

However, one can see that $f_1(t) \equiv 1$, $f_2(t) = 1 + t^{3/2}$ *and* $f_3(t) = 1 - t^{3/2}$ *are all local solutions to the equation above. Starting from* $\bar{q}_0 = 1$, $\bar{Q}$ *could be any solutions above. Further analysis is required to identify the correct solution.*

Remark 3.4. *For* $\bar{q}_0 \in \mathcal{L}^+$, *one can find from* (3.5) *that* $\bar{Q}$ *is well defined up to hitting* $\bar{s}_{\bar{q}_0}$, *increases strictly and hits* $z \in (\bar{q}_0, \bar{s}_{\bar{q}_0})$ *at a finite time*

$$\tau_{+,z} := \inf\{t > 0 : \bar{Q}(t) > z\} = \frac{1}{\mu}\int_{\bar{q}_0}^z \frac{\mathrm{d}y}{\rho(y) - 1} < \infty.$$

(1) *If* $\bar{s}_{\bar{q}_0} < \infty$, *then* $\bar{s}_{\bar{q}_0} \in \mathcal{L}^=$, $\bar{Q}$ *hits* $\bar{s}_{\bar{q}_0}$ *at time* $\tau_{+,\bar{s}_{\bar{q}_0}}$, *which is finite if and only if*

$$\tau_{+,\bar{s}_{\bar{q}_0}} = \frac{1}{\mu}\int_{\bar{q}_0}^{\bar{s}_{\bar{q}_0}} \frac{\mathrm{d}y}{\rho(y)-1} < \infty.$$

Observe that at $\bar{s}_{\bar{q}_0} \in \mathcal{L}^=$, *we can further rewrite* ρ *in* (2.17) *by*

$$\rho(y) - 1 = \frac{1-H(\bar{s}_{\bar{q}_0})}{1-H(y)} - 1 \sim \frac{H(y)-H(\bar{s}_{\bar{q}_0})}{1-H(\bar{s}_{\bar{q}_0})} \quad \text{as } y \uparrow \bar{s}_{\bar{q}_0}, \tag{3.7}$$

from which we obtain

$$\tau_{+,\bar{s}_{\bar{q}_0}} < \infty \;\Leftrightarrow\; \int_{\bar{q}_0}^{\bar{s}_{\bar{q}_0}} \frac{dy}{H(y)-H(\bar{s}_{\bar{q}_0})} < \infty.$$

(2) *If* $\bar{s}_{\bar{q}_0} = \infty$, *then* $\rho(y) > 1$ *for every* $y > \bar{q}_0$ *and thus* $\bar{s}_y = \infty$ *for all* $y > \bar{q}_0$. *In this case,* $\bar{Q}$ *is strictly increasing to* ∞, *and one can also check from* (3.5) *that, there exists some* $t_0 > 0$ *so that* $\bar{Q}(t) < \infty$ *for all* $t < t_0$ *and* $\bar{Q}(t) \to \infty$ *as* $t \uparrow t_0$ *if and only if*

$$t_0 = \frac{1}{\mu}\int_{\bar{q}_0}^{\infty} \frac{\mathrm{d}y}{\rho(y)-1} < \infty. \tag{3.8}$$

That is, the fluid queue $\bar{Q}$ *reaches infinity in a finite time, which is called "explosive". This also justifies Assumption A3 being the non-explosive condition.*

The analogous property can be observed for $\bar{q}_0 \in \mathcal{L}^-$. *Specifically, if* $\bar{q}_0 \in \mathcal{L}^-$ *and* $\bar{q}_0 > 0$, *then* $\bar{Q}$ *is well defined up to hitting* $\underline{s}_{\bar{q}_0}$, *decreases strictly and hits* $z \in (\underline{s}_{\bar{q}_0}, \bar{q}_0)$ *at a finite time*

$$\tau_{-,z} := \inf\{t > 0 : \bar{Q}(t) < z\} = \frac{1}{\mu}\int_{z}^{\bar{q}_0} \frac{\mathrm{d}y}{1-\rho(y)} < \infty.$$

(1) *If* $\underline{s}_{\bar{q}_0} \in \mathcal{L}^-$, *then* $\underline{s}_{\bar{q}_0} = 0$, $\bar{Q}$ *hits* 0 *at finite time and stays at* 0 *afterward.*
(2) *If* $\underline{s}_{\bar{q}_0} \in \mathcal{L}^=$, $\bar{Q}$ *hits* $\underline{s}_{\bar{q}_0}$ *at a finite time* $\tau_{-,\underline{s}_{\bar{q}_0}}$ *if and only if*

$$\tau_{-,\underline{s}_{\bar{q}_0}} = \frac{1}{\mu}\int_{\underline{s}_{\bar{q}_0}}^{\bar{q}_0} \frac{\mathrm{d}y}{1-\rho(y)} < \infty \;\Leftrightarrow\; \int_{\underline{s}_{\bar{q}_0}}^{\bar{q}_0} \frac{dy}{H(\bar{s}_{\bar{q}_0})-H(y)} < \infty,$$

where the second equivalence can be obtained from (3.7). *In the case where the integral above is infinite, the fluid queue* $\bar{Q}$ *decreases and converges to* $\underline{s}_{\bar{q}_0}$ *without hitting the boundary at a finite time.*

We next introduce the concept of *equilibrium point* for the fluid model.

Definition 3.1. *We call* $y_0 \geq 0$ *an equilibrium point for the fluid model, provided that if* $\bar{Q}(t_0) = y_0$ *for some* $t_0 \geq 0$, *then* $\bar{Q}(t) = y_0$ *for every* $t \geq t_0$.

By Remark 3.4, we find that (i) $y_0 \in \mathcal{L}^-$ is an equilibrium point if and only if $y_0 = 0$, and (ii) $y_0 > 0$ is an equilibrium point only if $y_0 \in \mathcal{L}^=$. If (3.3) has a unique solution starting from $y_0 \in \mathcal{L}^=$, that is, $f(\cdot) \equiv y_0$, then y_0 is an equilibrium point for the fluid model and the solution stays at y_0.

Remark 3.5. *Considering the case that* $[x_0, y_0) \subset \mathcal{L}^+$ *and* $(y_0, z_0] \subset \mathcal{L}^-$ *for some* $x_0 < y_0 < z_0$, *where we understand that* $[0,0) = \emptyset$ *if* $y_0 = 0$, *then* y_0 *is an equilibrium point.*

Let f *be a solution to* (3.3) *with* $f(0) = y_0 \in \mathcal{L}^=$. *One can find from Proposition 3.1 and Remark 3.4 that* f *increases in* (x_0, y_0) *and decreases in* (y_0, z_0). *Thus,* f *can neither go down into* (x_0, y_0) *nor go up into* (y_0, z_0), *but only stay at* y_0, *thus* $f(\cdot) \equiv y_0$ *is the unique solution to* (3.3).

In particular, if H *is a decreasing function on* $\mathbb{R}_+$ *and* $H(y_0) = 1 - \frac{\lambda_0}{\mu}$ *for some unique* $y_0 \in [0, \infty)$, *then* $[0, y_0) = \mathcal{L}^+$ *and* $(y_0, \infty) = \mathcal{L}^-$. *Thus,* y_0 *is the unique equilibrium point for the system, Assumption A3 is satisfied and the fluid limit stays at* y_0.

3.1. **Examples of H.** We give two examples of H to illustrate the trajectory behaviors and different cases of equilibrium points. More examples are given in the Appendix. Let $\bar{q}_0$ be the starting point of $\bar{Q}$.

Example 1: $H(y) = \beta e^{-\gamma y}$ for $y \in \mathbb{R}_+$ with $\beta \in (0,1)$ and $\gamma > 0$.

Recalling ρ in (2.17), we have

$$\rho(y) = \frac{\lambda_0}{\mu(1-\beta e^{-\gamma y})} \quad \text{and} \quad \mu\left(\rho(y) - 1\right) = \frac{\lambda_0 - \mu(1-\beta e^{-\gamma y})}{1-\beta e^{-\gamma y}}.$$

(1) If $\lambda_0 > \mu$, we have $\mathcal{L}^+ = \mathbb{R}_+$ and $\mathcal{L}^= = \emptyset$. $\bar{Q}$ increases strictly in $\mathbb{R}_+$ and $\bar{I} \equiv 0$, one can find from (3.3) that $\bar{Q}$ solves equation

$$\frac{\lambda_0}{\gamma\mu(\lambda_0-\mu)} \ln \frac{(\lambda_0-\mu)e^{\gamma y} + \mu\beta}{(\lambda_0-\mu)e^{\gamma \bar{q}_0} + \mu\beta} - \frac{y - x_0}{\mu}\Big|_{y=\bar{Q}(t)} = t \quad \forall t > 0, \tag{3.9}$$

and $\bar{Q}(t) \sim (\lambda_0 - \mu) \cdot t$ as $t \to \infty$.

(2) If $\lambda_0 = \mu$, we have $\mathcal{L}^+ = \mathbb{R}_+$ and $\mathcal{L}^= = \emptyset$. $\bar{Q}$ increases strictly in $\mathbb{R}_+$ and $\bar{I} \equiv 0$, one can find from (3.3) that $\bar{Q}$ solves equation

$$\frac{1}{\gamma\mu\beta}\left(e^{\gamma y} - e^{\gamma \bar{q}_0}\right) - \frac{y - \bar{q}_0}{\mu}\Big|_{y=\bar{Q}(t)} = t \quad \forall t > 0,$$

and $\bar{Q}(t) \sim \gamma^{-1} \ln t$ as $t \to \infty$.

(3) If $\mu > \lambda_0 > (1-\beta)\mu$, we have

$$\mathcal{L}^+ = [0, y_0), \quad \mathcal{L}^- = (y_0, \infty), \quad \mathcal{L}^= = \{y_0\} \quad \text{where} \quad y_0 = \frac{1}{\gamma} \ln \frac{\mu\beta}{\mu - \lambda_0}.$$

(a) If $\bar{q}_0 = y_0$, then one see from Remark 3.5 that $\bar{Q}(t) = y_0$, for all $t \geq 0$.

(b) If $\bar{q}_0 \neq y_0$, $\bar{Q}$ solves (3.9) which can be rewritten as

$$\frac{\lambda_0}{\gamma\mu(\lambda_0-\mu)} \ln \frac{e^{\gamma y} - e^{\gamma y_0}}{e^{\gamma x_0} - e^{\gamma y_0}} - \frac{y - x_0}{\mu}\Big|_{y=\bar{Q}(t)} = t \quad \forall t > 0.$$

Moreover, if $\bar{q}_0 > y_0$, $\bar{Q}(t) > y_0$ and decreases strictly to y_0; if $\bar{q}_0 < y_0$, $\bar{Q}(t) > y_0$ and increases strictly to y_0, and

$$-\ln\left|\bar{Q}(t) - y_0\right| \sim \frac{\gamma\mu(\mu-\lambda_0)}{\lambda_0} \cdot t \quad \text{as} \quad t \to \infty.$$

Thus, one can conclude in this example that y_0 is the unique critical and equilibrium point for the fluid limit.

(4) If $\lambda_0 = (1-\beta)\mu$, we have $\mathcal{L}^- = (0, \infty)$ and $\mathcal{L}^= = \{0\}$. $\bar{Q}$ decreases strictly in $(0, \infty)$ and solves equation

$$\frac{1-\beta}{\gamma\mu\beta} \ln \frac{e^{\gamma x_0} - 1}{e^{\gamma y} - 1} - \frac{x_0 - y}{\mu}\Big|_{y=\bar{Q}(t)} = t \quad \forall t > 0,$$

where $\bar{Q}(t) > 0$ for all $t > 0$, and one can check that for $\bar{q}_0 > 0$

$$-\ln\left(\bar{Q}(t)\right) \sim \frac{\gamma\mu\beta}{1-\beta} \cdot t \quad \text{as} \quad t \to \infty.$$

Thus, 0 is the unique critical and equilibrium point.

(5) If $\lambda_0 < (1-\beta)\mu$, we have $\mathcal{L}^- = \mathbb{R}_+$ and $\mathcal{L}^= = \emptyset$. Starting from $\bar{q}_0 > 0$, $\bar{Q}$ decreases strictly, solves (3.9) and hits 0 at finite time

$$\tau_{-,0} := \int_0^{\bar{q}_0} \frac{1-\beta e^{-\gamma y}}{\mu(1-\beta e^{-\gamma y}) - \lambda_0} \, \mathrm{d}y < \infty,$$

and stays at 0 afterward. Thus, 0 is a sub-critical and equilibrium point.

Observe that there are two cases for 0 to be an equilibrium point, as discussed after Definition 3.1, in Case-(4) and Case-(5), one in $\mathcal{L}^=$ and the other in $\mathcal{L}^-$, respectively.

Example 2: $H(y) = \beta \cos^2 y^\theta$ **for** $y \in \mathbb{R}_+$ **with** $\beta \in (0,1)$ **and** $\theta > 0$**.**
Recalling (2.17), we have

$$\rho(y) = \frac{\lambda_0}{\mu(1-\beta\cos^2 y^\theta)} \quad \text{and} \quad \mu\big(\rho(y)-1\big) = \frac{\lambda_0 - \mu(1-\beta\cos^2 y^\theta)}{1-\beta\cos^2 y^\theta}. \tag{3.10}$$

(1) If $\lambda_0 > \mu$, we have $\mathcal{L}^+ = \mathbb{R}_+$. $\bar{Q}$ increases strictly on $\mathbb{R}_+$, and one can check that

$$\int^x \frac{\mathrm{d}y}{\mu(\rho(y)-1)} \sim \frac{x}{\pi} \cdot \int_0^\pi \frac{(1-\beta\cos^2 y)\mathrm{d}y}{\lambda_0 - \mu(1-\beta\cos^2 y)} \quad \text{as} \quad x \to \infty.$$

We have from Proposition 3.1-(3.5) that

$$\bar{Q}(t) \sim \Big(\int_0^\pi \frac{(1-\beta\cos^2 y)\mathrm{d}y}{\lambda_0 - \mu(1-\beta\cos^2 y)}\Big)^{-1} \pi \cdot t \quad \text{as} \quad t \to \infty.$$

In this case, the increment rate of $\bar{Q}$ oscillates in $[\lambda_0 - \mu, \frac{\lambda_0}{1-\beta} - \mu]$.

(2) If $\lambda_0 < \mu(1-\beta)$, we have $\mathcal{L}^- = \mathbb{R}_+$. $\bar{Q}$ decreases on $\mathbb{R}_+$, hits 0 within finite time, and stays at 0 afterward. Thus, 0 is a sub-critical and equilibrium point.

(3) if $\lambda_0 = \mu$, one can find from (3.10) that, with $y_0 := 0 \in \mathcal{L}^+$,

$$\mathcal{L}^= = \{y \geq 0, \cos y^\theta = 0\} = \{0 < y_1 < y_2 < \cdots\} \quad \text{and} \quad \mathcal{L}^+ = \mathbb{R}_+ \setminus \mathcal{L}^=.$$

At every critical point $y_k > 0$, we can rewrite $\mu\big(\rho(y)-1\big)$ in (3.10) as

$$\begin{aligned} \mu\big(\rho(y)-1\big) &= \frac{\mu\beta\cos^2 y^\theta}{1-\beta\cos^2 y^\theta} = \frac{\mu\beta\sin^2(y^\theta - y_k^\theta)}{1-\beta\cos^2 y^\theta} \\ &\sim \mu\beta\cdot(y^\theta - y_k^\theta)^2 \sim \mu\beta\theta^2 y_k^{2(\theta-1)}\cdot(y-y_k)^2 \quad \text{as} \quad y \to y_k. \end{aligned} \tag{3.11}$$

Hence, every $y_k > 0$ is a critical and equilibrium point, and $\bar{Q}$ increases strictly in each interval (y_{k-1}, y_k) and is trapped inside the interval. Moreover, starting from $\bar{q}_0 \in (y_{k-1}, y_k)$, $\bar{Q}$ converges to y_k without hitting y_k and satisfies

$$(y_k - \bar{Q}(t)) \sim \big(\mu\beta\theta^2 y_k^{2(\theta-1)}\big)^{-1} \cdot t^{-1} \quad \text{as} \quad t \to \infty.$$

(4) if $\lambda_0 = \mu(1-\beta)$, one can find from (3.10) that similar to Case-(3) above,

$$\mathcal{L}^= = \{y \geq 0, \sin y^\theta = 0\} = \{0 = y_0 < y_1 < y_2 < \cdots\} \quad \text{and} \quad \mathcal{L}^- = \mathbb{R}_+ \setminus \mathcal{L}^=,$$

which implies $\bar{Q}$ decreases strictly within each interval (y_{k-1}, y_k).

(a) At $y_0 = 0$, we can rewrite $\mu\big(\rho(y)-1\big)$ in (3.10) as

$$\mu\big(\rho(y)-1\big) = \frac{-\mu\beta\,\sin^2 y^\theta}{1-\beta\cos^2 y^\theta} \sim \frac{-\mu\beta}{1-\beta}\cdot y^{2\theta} \quad \text{as} \quad y \to 0+$$

which implies 0 is a critical and equilibrium point. For $\bar{Q}$ starting from $\bar{q}_0 \in (0, y_1)$,

(i) if $2\theta < 1$, $\bar{Q}$ will hit 0 for some finite time and stay at 0 afterward;

(ii) if $2\theta = 1$, $\bar{Q}$ converges to 0 without hitting 0 and

$$-\ln \bar{Q}(t) \sim \frac{\mu\beta}{1-\beta}\cdot t \quad \text{as} \quad t \to \infty;$$

(iii) if $2\theta > 1$, $\bar{Q}$ converges to 0 without hitting 0 and

$$\bar{Q}(t) \sim \Big(\frac{\mu\beta(2\theta-1)}{1-\beta}\Big)^{\frac{1}{1-2\theta}} \cdot t^{\frac{-1}{2\theta-1}} \quad \text{as} \quad t \to \infty.$$

(b) At y_k for $k \geq 1$, we have in (3.10) similarly

$$\mu\left(\rho(y)-1\right) = \frac{-\mu\beta\,\sin^2(y^\theta - y_k^\theta)}{1-\beta\cos^2 y^\theta} \sim \frac{-\mu\beta\theta^2}{1-\beta} y_k^{2(\theta-1)} \cdot \left(y-y_k\right)^2 \quad \text{as} \quad y \to y_k.$$

Hence, y_k is a critical and equilibrium point. Similar to the case (3), starting from $\bar{q}_0 \in (y_k, y_{k+1})$, $\bar{Q}$ decreases strictly to y_k without hitting y_k and

$$\bar{Q}(t) - y_k \sim \frac{1-\beta}{\mu\beta\theta^2} \cdot y_k^{2(1-\theta)} \cdot t^{-1} \quad \text{as} \quad t \to \infty.$$

(5) if $\mu > \lambda_0 > \mu(1-\beta)$, one can find in (3.10) that with $y_0 := 0$,

$$\mathcal{L}^= = \left\{ y \geq 0, \cos^2 y^\theta = \frac{\mu-\lambda_0}{\mu\beta} \right\} = \{0 < y_1 < y_2 < \cdots\},$$
$$\mathcal{L}^+ = \cup_{k\geq 0}(y_{2k}, y_{2k+1}) \quad \text{and} \quad \mathcal{L}^- = \cup_{k\geq 0}(y_{2k+1}, y_{2k+2}),$$

that is, $\mathbb{R}_+$ is consecutively partitioned into sup-critical regimes and sub-critical regimes. At each $y_k > 0$,

$$\begin{aligned}\mu\left(\rho(y)-1\right) &= \mu\beta\frac{\cos^2 y^\theta - \cos^2 y_k^\theta}{1-\beta\cos^2 y^\theta} \sim \frac{-2\mu\beta\cos y_k^\theta \sin y_k^\theta}{1-\beta\cos^2 y_k^\theta} \cdot \left(y^\theta - y_k^\theta\right) \\ &\sim \frac{2\mu(\mu-\lambda_0)}{\lambda_0} \cdot \tan y_k^\theta \cdot \theta \cdot y_k^{\theta-1} \cdot (y_k - y) \quad \text{as} \quad y \to y_k.\end{aligned}$$

We see that each $y_k > 0$ is a critical and equilibrium point. Moreover, it is interesting to find that within each interval, the trajectory of $\bar{Q}$ behaves like a directed magnetic field line flowing from $\{y_{2k}\}_{k\geq 0}$ toward $\{y_{2k+1}\}_{k\geq 0}$. Specifically, starting from $\bar{q}_0 \in (y_{2k}, y_{2k+1})$, $\bar{Q}$ leaves y_{2k} and increases toward y_{2k+1}; starting from $\bar{q}_0 \in (y_{2k-1}, y_{2k})$, $\bar{Q}$ leaves y_{2k} and decreases toward y_{2k-1}, Moreover, $\bar{Q}$ above cannot reach y_{2k+1} within a finite time, and

$$-\ln\left|\bar{Q}(t) - y_{2k+1}\right| \sim \frac{2\mu\theta(\mu-\lambda_0)}{\lambda_0}\sqrt{\frac{\lambda_0-\mu(1-\beta)}{\mu-\lambda_0}}\, y_{2k+1}^{\theta-1} \cdot t \quad \text{as} \quad t \to \infty.$$

4. Functional Central Limit Theorem

In the following, we assume that the FLLN holds at a critical and equilibrium point $\bar{q}_0 \geq 0$ in $\mathcal{L}^=$ such that

$$\bar{A}(t) = \mu\, t, \quad \bar{Q}(t) = \bar{q}_0, \quad \bar{W}(t) = \bar{w}_0 = \mu^{-1}\bar{q}_0, \quad \bar{I}(t) = 0 \quad \text{and} \quad \bar{B}(t) = t. \tag{4.1}$$

We will study the CLT-scaled processes that are centered around the fluid dynamics at a critical and equilibrium point under heavy traffic, that is, the CLT-scaled processes are defined by

$$\left(\hat{A}^{(n)}, \hat{Q}^{(n)}, \hat{W}^{(n)}\right) = \sqrt{n}\left(\bar{A}^{(n)} - \bar{A}, \bar{Q}^{(n)} - \bar{q}_0, \bar{W}^{(n)} - \bar{w}_0\right). \tag{4.2}$$

In the FCLT, we will distinguish the two cases for $\bar{q}_0 \in \mathcal{L}^=$: $\bar{q}_0 = 0$ and $\bar{q}_0 > 0$.

We further make the following assumptions.

Assumption A4. *(i) Given λ_0 and μ in Assumption A1-(i), there exist constants $\hat{\lambda}_0, \hat{\mu} \in \mathbb{R}$ such that as $n \to \infty$,*

$$\left(\hat{\lambda}_0^{(n)}, \hat{\mu}^{(n)}\right) = \sqrt{n}\left(\lambda_0^{(n)} - \lambda_0, \mu^{(n)} - \mu\right) \to \left(\hat{\lambda}_0, \hat{\mu}\right).$$

(ii) Given H in Assumption A1-(ii), there exists $\hat{H}_{\bar{q}_0} \in \mathbb{C}(\mathbb{R})$ such that as $n \to \infty$,

$$\hat{H}_{\bar{q}_0}^{(n)}(y) := \sqrt{n}\left(H^{(n)}(\bar{q}_0 + y/\sqrt{n}) - H(\bar{q}_0)\right) \to \hat{H}_{\bar{q}_0}(y) \quad \textit{u.o.c. on } \mathbb{R}, \tag{4.3}$$

and $\hat{H}_{\bar{q}_0}$ *satisfies the linear growth condition, that is, for some* $c_0 > 0$,

$$\hat{H}_{\bar{q}_0}(y) \le c_0\,(1+|y|) \quad \forall y \in \mathbb{R}. \tag{4.4}$$

(iii) For every $k > 0$ *and* $\varepsilon > 0$,

$$\limsup_{n\to\infty} \sup_{|y|\le k} \sqrt{n} \int_{n\varepsilon}^{\infty} h^{(n)}(u, n\bar{q}_0 + \sqrt{n}y)\,\mathrm{d}u = 0, \tag{4.5}$$

$$\sup_{|y|\le k} \sup_{n\ge 1} \sup_{t>0} \sqrt{t} \int_{t}^{\infty} h^{(n)}(u, n\bar{q}_0 + \sqrt{n}y)\,du < \infty. \tag{4.6}$$

Remark 4.1. (4.3) *in the second condition* (ii) *indicates the convergence of the oscillations from function* $H^{(n)}$ *around* H *in the ball of radius of order* $n^{-1/2}$ *centered at* $\bar{q}_0$. *The continuity condition on* $\hat{H}_{\bar{q}_0}(\cdot)$ *is a technical condition to apply the convergence result of stochastic integrals in* [33, Theorem 5.4] *in the proofs. It also ensures the existence of a strong solution to SDEs, as well as their uniqueness up to a possibly explosive time. However, applying* [33, Theorem 5.4], *the existence of a non-explosive solution to SDEs is needed (see also condition (3) in Propositions 6.2 and 6.3). The linear growth requirement in* (4.4) *is a sufficient condition (commonly imposed) for the stochastic boundedness (see Remark 4.2 for further discussions on solutions to the limiting SDE). Nevertheless,* (4.4) *may not be necessary for some cases as shown in Remark 4.4. The third condition* (iii) *is stronger than the condition* (iii) *in Assumption A1, and is imposed to control the tail behavior of the functions* $h^{(n)}$ *under the square root scale.*

If $h^{(n)}(t, ny) = \phi(t)H(y)$ *for some positive measurable functions* ϕ *and* H *with* $\|\phi\|_1 = 1$, *we have* $H^{(n)} \equiv H$ *and Assumption A1-*(ii) *is simply* $H(y) < 1$ *for all* $y \in \mathbb{R}_+$. *In this case,*

$$\sqrt{n} \int_{n\varepsilon}^{\infty} h^{(n)}(u, n\bar{q}_0 + \sqrt{n}y)\,\mathrm{d}u = H(n\bar{q}_0 + \frac{y}{\sqrt{n}}) \cdot \sqrt{n} \int_{n\varepsilon}^{\infty} \phi(u)\,du.$$

One can see that the condition (4.5) *reduces to*

$$\limsup_{t\to\infty} \sqrt{t} \int_{t}^{\infty} \phi(s)\,ds = 0,$$

which is the usual condition required in the FCLT of standard Hawkes process (see, [38, Proposition 2.2.], [24, Theorem 2.9]*). Since* $t \to \sqrt{t}\int_t^\infty \phi(s)\,ds$ *is a continuous function on* $\mathbb{R}_+$ *vanishing at* 0 *and* ∞, *we must have*

$$\sup_{t>0} \sqrt{t} \int_{t}^{\infty} \phi(s)\,ds < \infty,$$

which ensures (4.6).

Theorem 4.1. *Under Assumptions A1, A2 and A4, assuming that there exists a random variable* $\hat{Q}_0$ *such that* $\hat{Q}_0^{(n)} := \sqrt{n} \cdot (\bar{Q}_0^{(n)} - \bar{q}_0) \Rightarrow \hat{Q}_0$ *as* $n \to \infty$, *we have*

$$\big(\hat{A}^{(n)}, \hat{Q}^{(n)}, \hat{W}^{(n)}\big) \Rightarrow \big(\hat{A}, \hat{Q}, \hat{W}\big) \quad \text{in} \quad (\mathbb{D}^3, J_1) \quad \text{as} \quad n \to \infty.$$

The limit process $\big(\hat{A}, \hat{Q}, \hat{W}\big)$ *is given below in the two cases:*

(i) Case of $\bar{q}_0 = 0$: $(\hat{A}, \hat{Q})$ *is the unique strong solution to the SDE such that* $\hat{Q}$ *has a reflection at* 0*:*

$$\begin{aligned} \mathrm{d}\hat{A}(t) &= \frac{\hat{\lambda}_0 + \mu\,\hat{H}_0(\hat{Q}(t))}{1 - H(0)}\,\mathrm{d}t + \frac{\mathrm{d}\hat{X}(t)}{1 - H(0)}, \\ \mathrm{d}\hat{Q}(t) &= -\,\hat{\mu}\,\mathrm{d}t + \mathrm{d}\hat{A}(t) - \mathrm{d}\hat{S}(t) + \mu\,\mathrm{d}\hat{I}(t), \end{aligned} \tag{4.7}$$

with the initial condition $(\hat{A}(0), \hat{Q}(0)) = (0, \hat{Q}_0)$, $\hat{X}$ *and* $\hat{S}$ *are two independent mean-zero Brownian motions with variance coefficients* μ *and* $\mu^3\sigma^2$, *respectively, and where* $\hat{I}$ *is the*

minimal nondecreasing process in $\mathbb{C}$ *so that* $\hat{Q} \geq 0$*, that is,*

$$\int_{[0,\infty)} \mathbf{1}\big(\hat{Q}(t) > 0\big)\, \mathrm{d}\hat{I}(t) = 0.$$

Moreover,

$$\hat{W} = \mu^{-1} \cdot \hat{Q}.$$

(ii) Case of $\bar{q}_0 > 0$*:* $(\hat{A}, \hat{Q})$ *is the unique strong solution to the SDE :*

$$\begin{aligned} \mathrm{d}\hat{A}(t) &= \frac{\hat{\lambda}_0 + \mu\, \hat{H}_{\bar{q}_0}(\hat{Q}(t))}{1 - H(\bar{q}_0)}\, \mathrm{d}t + \frac{\mathrm{d}\hat{X}(t)}{1 - H(\bar{q}_0)}, \\ \mathrm{d}\hat{Q}(t) &= -\,\hat{\mu}\, \mathrm{d}t + \mathrm{d}\hat{A}(t) - \mathrm{d}\hat{S}(t), \end{aligned} \tag{4.8}$$

with the initial condition $(\hat{A}(0), \hat{Q}(0)) = (0, \hat{Q}_0)$*,* $\hat{X}$ *and* $\hat{S}$ *are Brownian motions as given above in case (i). Moreover, for* $\bar{w}_0 = \mu^{-1}\bar{q}_0$ *in* (4.1)*,*

$$\hat{W}(t) = \mu^{-1}\Big(\hat{Q}(t) + \big(\hat{S}(t) - \hat{S}(t + \bar{w}_0)\big) - \hat{\mu}\bar{w}_0\Big). \tag{4.9}$$

Remark 4.2. *Recall the state-dependent intensity* $\rho^{(n)}(\cdot)$ *in* (2.18)*. Since* $\lambda_0 = \mu(1 - H(\bar{q}_0))$ *at* $\bar{q}_0 \in \mathcal{L}^{=}$*, we have by Assumption A4-*(i) *and* (ii) *that, as* $n \to \infty$*,*

$$\begin{aligned} \hat{\rho}^{(n)}_{\bar{q}_0}(y) &:= \sqrt{n}\big(1 - \rho^{(n)}(\bar{q}_0 + y/\sqrt{n})\big) = \sqrt{n}\Big(1 - \frac{\lambda_0^{(n)}\mu}{\mu^{(n)}\lambda_0}\frac{1 - H(\bar{q}_0)}{1 - H^{(n)}(\bar{q}_0 + y/\sqrt{n})}\Big) \\ &= \sqrt{n}\Big(1 - \frac{\lambda_0^{(n)}}{\lambda_0}\Big) + \frac{\lambda_0^{(n)}}{\lambda_0}\sqrt{n}\Big(1 - \frac{\mu}{\mu^{(n)}}\Big) - \frac{\lambda_0^{(n)}\mu}{\mu^{(n)}\lambda_0}\frac{\sqrt{n}(H^{(n)}(\bar{q}_0 + \frac{y}{\sqrt{n}}) - H(\bar{q}_0))}{1 - H^{(n)}(\bar{q}_0 + y/\sqrt{n})} \\ &\to \frac{\hat{\mu}}{\mu} - \frac{\hat{\lambda}_0}{\lambda_0} - \frac{\hat{H}_{\bar{q}_0}(y)}{1 - H(\bar{q}_0)} =: \hat{\rho}_{\bar{q}_0}(y) \in \mathbb{C}(\mathbb{R})\,. \end{aligned} \tag{4.10}$$

In case (i) with $\bar{q}_0 = 0$*, it follows that* $\hat{Q}$ *in* (4.7) *solves the SDE with reflection at* 0*:*

$$\begin{aligned} \mathrm{d}\hat{Q}(t) &= \left(\frac{\hat{\lambda}_0 + \mu\, \hat{H}_0(\hat{Q}(t))}{1 - H(0)} - \hat{\mu}\right)\mathrm{d}t + \frac{\mathrm{d}\hat{X}(t)}{1 - H(0)} - \mathrm{d}\hat{S}(t) + \mu\, \mathrm{d}\hat{I}(t) \\ &= -\,\mu\hat{\rho}_0(\hat{Q}(t))\, dt + \mu\left(\frac{1}{\lambda_0}\mathrm{d}\hat{X}(t) - \frac{1}{\mu}\mathrm{d}\hat{S}(t)\right) + \mu\, \mathrm{d}\hat{I}(t). \end{aligned} \tag{4.11}$$

By Assumption A4-(ii)*,* $\hat{\rho}_0(\cdot)$ *is a continuous function on* $\mathbb{R}_+$ *and satisfies the linear growth condition, the existence and uniqueness of a strong solution to the reflected SDE above follow from* [54, Theorem 3.1]*. Moreover, the linear growth condition guarantees that the solution does not explode at any finite time, c.f.* [28, Problem V.3.15]*.*

In case (ii) with $\bar{q}_0 > 0$*, it follows that* $\hat{Q}$ *in* (4.8) *solves the SDE:*

$$\begin{aligned} \mathrm{d}\hat{Q}(t) &= \left(\frac{\hat{\lambda}_0 + \mu\, \hat{H}_{\bar{q}_0}(\hat{Q}(t))}{1 - H(\bar{q}_0)} - \hat{\mu}\right)\mathrm{d}t + \frac{\mathrm{d}\hat{X}(t)}{1 - H(\bar{q}_0)} - \mathrm{d}\hat{S}(t) \\ &= -\,\mu\hat{\rho}_{\bar{q}_0}(\hat{Q}(t))\, dt + \mu\left(\frac{1}{\lambda_0}\mathrm{d}\hat{X}(t) - \frac{1}{\mu}\mathrm{d}\hat{S}(t)\right). \end{aligned} \tag{4.12}$$

By Assumption A4-(ii)*,* $\hat{\rho}_{\bar{q}_0}(\cdot)$ *is a continuous function on* $\mathbb{R}$ *and satisfies the linear growth condition, the existence and uniqueness of a strong solution to the SDE above follows from* [47, Theorem 1]*, see also* [54]*. The linear growth condition* (4.4) *is sufficient for non-explosion, c.f.* [28, Problem V.3.15]*.*

In both cases (i) and (ii), the limit $\hat{A}$ *in* (4.7) *and* (4.8) *is equal to*

$$\hat{A}(t) = \hat{Q}(t) + \hat{\mu}\, t + \hat{S}(t) - \mu\hat{I}(t),$$

where $\hat{I} \equiv 0$ in the case $\bar{q}_0 > 0$. In the case of a standard Hawkes process in Remarks 3.1 and 2.2, the system is independent of n, $\lambda_0^{(n)} = \lambda_0$ and $H^{(n)} = H \equiv ||h||_1 < 1$, we have $\hat{\lambda}_0 = 0$ and $\hat{H}_{\bar{q}_0}(y) \equiv 0$ for all $\bar{q}_0, y \geq 0$ in (4.3), and thus by (4.7), $\hat{A}$ is given by

$$\hat{A}(t) = \frac{1}{1 - ||h||_1} \hat{X}(t).$$

This limit is proved as a special case of Theorem 2 in [3] for multivariate Hawkes processes.

Remark 4.3. *In the case $h^{(n)}(t, ny) = h(t, y)$ for some measurable function h on $\mathbb{R}_+^2$, we have $H^{(n)}(y) = \int_0^\infty h(t, y)\, dt = H(y)$ independent of n, similar to the case in the multiplicative model in Remark 4.1. Then we have in (4.3), assuming H is right differentiable at 0,*

$$\hat{H}^{(n)}(y) = \sqrt{n}\big(H(\frac{y}{\sqrt{n}}) - H(0)\big) \to H'_+(0) \cdot y = \hat{H}_0(y) \quad u.o.c.$$

Hence, in this case, $\hat{\rho}_0(y) = \hat{\rho}_0(0) - \frac{H'_+(0)}{1-H(0)}\, y = \hat{\rho}_0 - \frac{H'_+(0)}{1-H(0)}\, y$ by (4.10). By (4.11), we obtain

$$\mathrm{d}\hat{Q}(t) = -\mu\bigg(\hat{\rho}_0 - \frac{H'_+(0)}{1 - H(0)} \cdot \hat{Q}(t)\bigg)\mathrm{d}t + \frac{\mathrm{d}\hat{X}(t)}{1 - H(0)} - \mathrm{d}\hat{S}(t) + \mu\, \mathrm{d}\hat{I}(t) \tag{4.13}$$

which is a reflected OU process taking values in $\mathbb{R}_+$. The existence and uniqueness of a strong solution to the reflected OU process follow from [46] and [48]. Moreover, by [48, Proposition 1], we can obtain the explicit formula for the stationary distribution for this reflected OU process.

In addition, one can rewrite (4.13) as

$$\mathrm{d}\hat{Q}(t) = -\mu\bigg(\hat{\rho}_0 - \frac{H'_+(0)}{1 - H(0)} \cdot \hat{Q}(t)\bigg)\mathrm{d}t + \sigma_Q\, \mathrm{d}\hat{B}(t) + \mu\, \mathrm{d}\hat{I}(t),$$

for some standard Brownian motion $\hat{B}$, where by the fact $\lambda_0 = \mu(1 - H(0))$,

$$\sigma_Q = \sqrt{\frac{\mu}{(1 - H(0))^2} + \mu^3\sigma^2} = \sqrt{\frac{\mu(\mu^2 - \lambda_0^2)}{\lambda_0^2} + \mu(1 + c_s^2)}\,. \tag{4.14}$$

Here $c_s^2 = \mu^2\sigma^2$ is the squared coefficient of variation (SCV) of the service times, representing their variability. When the service times are exponentially distributed, $c_s^2 = 1$.

In the case of a standard Hawkes process, that is, $h(t, y) \equiv h(t)$ and the arrival process is independent of the queueing process, the limit $\hat{Q}(t)$ in (4.11) reduces to a reflected Brownian motion (RBM) as studied in [37], in particular, under the additional condition that $\hat{\rho}_0 = \hat{\rho}_0(0) > 0$ (recalling $\hat{\rho}_0(y)$ in (4.10)),

$$\mathrm{d}\hat{Q}(t) = -\mu\hat{\rho}_0\mathrm{d}t + \frac{\mathrm{d}\hat{X}(t)}{1 - H(0)} - \mathrm{d}\hat{S}(t) + \mu\, \mathrm{d}\hat{I}(t)\,, \tag{4.15}$$

or equivalently (in distribution),

$$\mathrm{d}\hat{Q}(t) = -\mu\hat{\rho}_0\mathrm{d}t + \sigma_Q\, \mathrm{d}\hat{B}(t) + \mu\, \mathrm{d}\hat{I}(t), \tag{4.16}$$

where $\hat{I}$ is the regulator for $\hat{Q} \geq 0$, and σ_Q is defined in (4.14). For the reflected OU diffusion in (4.13), however, we do not need $\hat{\rho}_0$ in (4.7) to be positive as long as $H'_+(0) < 0$ to have a stationary distribution. Hence, in our model with state-dependent Hawkes arrivals, our limit $\hat{Q}$ in (4.13) has a linear functional of $\hat{Q}$ in the drift, which extends the previous studies with Hawkes arrivals without state-dependence.

Also, in the case of Poisson arrival processes, the arrival rate will be simply λ_0 and $h^{(n)} \equiv 0$, and the critical load condition would simply imply that $\lambda_0 = \mu$, so that the term $\frac{\mu(\mu^2 - \lambda_0^2)}{\lambda_0^2}$ in the

expression of σ_Q in (4.14) vanishes. The RMB $\hat{Q}(t)$ in (4.16) reduces to the well-known RBM limit for M/GI/1 queues in heavy traffic:

$$\mathrm{d}\hat{Q}(t) = -\mu\hat{\rho}_0\,\mathrm{d}t \ + \sqrt{2\mu}\,\mathrm{d}\hat{B}(t) + \mu\,\mathrm{d}\hat{I}(t).$$

However, the limit process $\hat{Q}$ is much more general than the reflected OU process, since the drift in the expression in (4.11) has a possibly nonlinear function $\hat{H}_0(y)$. In the next remark, we give an example which leads to a reflected diffusion with a fractional or power drift.

Remark 4.4. *Let $\mu > \lambda_0 > 0$, $\alpha > 0$, $\beta \in (0,1]$ and $\gamma \in (0, \frac{1}{2})$. For $y \in [0,1]$, define*

$$H^{(n)}(y) = 1 - \frac{\lambda_0}{\mu} - \frac{\lambda_0}{\mu}\,\kappa \cdot \begin{cases} y^\alpha & \text{if } n^\gamma \cdot y \geq 1, \\ n^{\frac{\beta-1}{2}} y^\beta & \text{if } n^\gamma \cdot y < 1, \end{cases}$$

where $\kappa > 0$ is a constant so that $H^{(n)} \geq 0$ on $[0,1]$. It is easy to check that as $n \to \infty$,

$$H^{(n)}(y) \to H(y) = 1 - \frac{\lambda_0}{\mu} - \frac{\lambda_0}{\mu}\,\kappa \cdot y^\alpha \quad \text{uniformly on } [0,1]. \tag{4.17}$$

Therefore, $0 \in \mathcal{L}^=$ and $(0,1] \subset \mathcal{L}^-$, and we see from Remarks 3.4 and 3.5 that 0 is a critical and equilibrium point, if $\bar{q}_0 \in (0,1)$ and $\alpha < 1$, $\bar{Q}$ hits 0 at some finite time and stays at 0 afterward. Moreover, one can check that under the condition $2\gamma < 1$,

$$\hat{H}_0^{(n)}(y) \to \hat{H}_0(y) = -\frac{\lambda_0}{\mu}\,\kappa \cdot y^\beta \quad \text{u.o.c. in } \mathbb{R}_+. \tag{4.18}$$

The condition $\beta \leq 1$ guarantees the linear growth condition (4.4). Therefore, the FCLT holds and the limit diffusion $\hat{Q}$ satisfies, by (4.10), (4.11) and for $\hat{\rho}_0 = \frac{\hat{\mu}}{\mu} - \frac{\hat{\lambda}_0}{\lambda_0}$,

$$\mathrm{d}\hat{Q}(t) = -\mu\big(\hat{\rho}_0 + \kappa\,\hat{Q}^\beta(t)\big)\,\mathrm{d}t + \mu\,\mathrm{d}\Big(\frac{\hat{X}}{\lambda_0} - \frac{\hat{S}}{\mu}\Big)(t) + \mu\,\mathrm{d}\hat{I}(t). \tag{4.19}$$

This is a reflected generalized OU process with a fractional functional in the drift.

For the case of positive $y_0 \in \mathcal{L}^=$, define for $y \in [-1,1]$,

$$H^{(n)}(y_0 + y) = 1 - \frac{\lambda_0}{\mu} - \frac{\lambda_0}{\mu}\kappa \cdot \begin{cases} y\,|y|^{\alpha-1} & \text{if } n^\gamma \cdot |y| \geq 1, \\ n^{\frac{\beta-1}{2}}\, y\,|y|^{\beta-1} & \text{if } n^\gamma \cdot |y| < 1, \end{cases}$$

where $\kappa > 0$ is a constant such that $H^{(n)}(y_0 + y) \geq 0$ on $[-1,1]$. One can find that

$$H^{(n)}(y_0 + y) \to H(y_0 + y) = 1 - \frac{\lambda_0}{\mu} - \frac{\lambda_0}{\mu}\kappa \cdot y\,|y|^{\alpha-1} \quad \text{uniformly on } [-1,1]. \tag{4.20}$$

Therefore, $y_0 \in \mathcal{L}^=$, $[y_0 - 1, y_0) \subset \mathcal{L}^+$ and $(y_0, y_0 + 1] \subset \mathcal{L}^-$, we see from Remark 3.5 that y_0 is a critical and equilibrium point. Moreover, one can check directly in (4.3) that for $\bar{q}_0 = y_0 \in \mathcal{L}^=$,

$$\hat{H}_{\bar{q}_0}^{(n)}(y) \to \hat{H}_{\bar{q}_0}(y) = -\frac{\lambda_0}{\mu}\kappa \cdot y\,|y|^{\beta-1} \quad \text{u.o.c. in } \mathbb{R},$$

where the assumption $\beta \leq 1$ guarantees the linear growth condition (4.4). By (4.10) and (4.12) we obtain that $\hat{Q}$ taking values in $\mathbb{R}$, solves the SDE:

$$\mathrm{d}\hat{Q}(t) = \begin{cases} -\mu\Big(\hat{\rho}_0 + \kappa\,\hat{Q}(t)^\beta\Big)\,\mathrm{d}t + \mu\,\mathrm{d}\Big(\frac{\hat{X}}{\lambda_0} - \frac{\hat{S}}{\mu}\Big)(t) & \text{if} \quad \hat{Q}(t) \geq 0, \\ -\mu\Big(\hat{\rho}_0 - \kappa\,|\hat{Q}(t)|^\beta\Big)\,\mathrm{d}t + \mu\,\mathrm{d}\Big(\frac{\hat{X}}{\lambda_0} - \frac{\hat{S}}{\mu}\Big)(t) & \text{if} \quad \hat{Q}(t) < 0. \end{cases} \tag{4.21}$$

This is a generalized OU process with a fractional functional in the drift.

Next, we claim in this example that the linear growth condition (4.4) may not be necessary, that is, the FCLT also holds for $1 < \beta < \frac{1}{1-2\gamma}$ with the same diffusion limit given by (4.19) and (4.21)

respectively, where the condition $\beta < \frac{1}{1-2\gamma}$ *is only used in* (4.17) *and* (4.20) *for* $H^{(n)}$ *at* $y = n^{-\gamma}$. *Take the case* $\bar{q}_0 = 0 \in \mathcal{L}^{=}$ *for example. Noticing that in the proof of the FCLT, Proposition* 6.2 *concerning the convergence of a sequence of solutions to SDEs is applied, where the existence and local uniqueness of a non-explosive solution (in weak sense) to SDE* (4.11) *is necessary, c.f. Remarks* 6.4 *and* 6.5. *The linear growth condition* (4.4) *is only served as a sufficient condition for its non-explosive property, see also Remarks* 4.1 *and* 4.2, *which will be unnecessary for our particular SDE* (4.19) *if* $\beta > 0$.

More precisely, by the continuity of the drift term, one can find from the "step 1" in the proof of Theorem 3.1 in [54] *that* (4.19) *does have a localized unique strong solution, that is, a pathwise uniqueness of solutions for* (4.19) *holds up to leaving an arbitrary interval* $[0, b]$. *Therefore, to show its stochastic bounded property, it is sufficient to show that for some constant* $c_0 > 0$,

$$\mathbb{E}\big[\hat{Q}^2(t)\big] \le c_0(1+t) \quad \forall t > 0. \tag{4.22}$$

For every $b > 0$, *let* $\tau_b := \inf\{t > 0 : \hat{Q}(t) > b\}$. *Applying Itô formula, we obtain*

$$\mathbb{E}\big[\hat{Q}^2(t \wedge \tau_b)\big] = \mathbb{E}\big[\hat{Q}^2(0)\big] + \mathbb{E}\bigg[\int_0^{t\wedge\tau_b} \big(\sigma_Q^2 - 2\mu\,(\hat{\rho}_0\hat{Q}(s) + \kappa\hat{Q}^{\beta+1}(s)\big)\,\mathrm{d}s\bigg],$$

where σ_Q *is the coefficient in* (4.14). *Since* $-\mu\kappa < 0$, *for every* $\beta > 0$, *we have*

$$\sigma_Q^2 - 2\mu\big(\hat{\rho}_0 y + \kappa y^{\beta+1}\big) \le c_0 \quad \forall y \ge 0,$$

for some constant $c_0 > 0$ *independent of* b *and* t, *which gives*

$$\mathbb{E}\big[\hat{Q}^2(t \wedge \tau_b)\big] \le \mathbb{E}\big[\hat{Q}^2(0)\big] + c_0 \cdot t.$$

Letting $b \to \infty$ *gives* (4.22) *and proves the validity of the FCLT for* $\beta \in (1, \frac{1}{1-2\gamma})$ *in this example.*

5. Hawkes/GI/1 queues with workload dependent intensity

We consider another Hawkes/GI/1 queue whose Hawkes arrival process has an intensity depending on the workload process. Specifically, the model description is the same as in the model with queue-length dependent intensity, except that the conditional intensity $\lambda(t)$ in (2.3), is given by the following

$$\lambda(t) = \lambda_0 + \sum_{j\ge1} h(t - \tau_j, W(\tau_j-)\big) = \lambda_0 + \int_0^t h\big(t-u, W(u-)\big)\,\mathrm{d}A(u). \tag{5.1}$$

For the system indexed by n,

$$\lambda^{(n)}(t) = \lambda_0^{(n)} + \int_0^t h^{(n)}\big(t-s, W^{(n)}(s-)\big)\,\mathrm{d}A^{(n)}(s). \tag{5.2}$$

We obtain the following FLLN for the LLN-scaled processes in (3.1).

Theorem 5.1. *Under Assumptions* A1, A2-(2.15) *and* A3, *assuming that* $\bar{Q}_0^{(n)} := n^{-1}Q_0^{(n)} \to \bar{q}_0$ *for some constant* $\bar{q}_0 \ge 0$, $(\bar{A}^{(n)}, \bar{Q}^{(n)}, \bar{W}^{(n)})$ *is a* $\mathbb{C}$*-tight family in* $(\mathbb{D}^3, J_1)$, *that is, every limit* $(\bar{A}, \bar{Q}, \bar{W})$ *takes value in* $\mathbb{C}^3$ *such that* $\bar{Q} = \mu\bar{W}$ *and* $(\bar{A}, \bar{W})$ *satisfies the set of nonlinear integral equations such that* $\bar{W}$ *has a reflection at* 0*:*

$$\begin{aligned} \bar{A}(t) &= \lambda_0 t + \int_0^t H\big(\bar{W}(u)\big)\,\mathrm{d}\bar{A}(u), \\ \bar{W}(t) &= \bar{w}_0 + \mu^{-1}\,\bar{A}(t) - t + \bar{I}(t), \end{aligned} \tag{5.3}$$

where $\bar{w}_0 = \mu^{-1}\bar{q}_0$ *and* $\bar{I}$ *is the minimal nondecreasing process in* $\mathbb{C}$ *so that* $\bar{W}(t) \geq 0$ *for every* $t \geq 0$*, and* $\bar{I}$ *increases only when* $\bar{W}$ *is zero, that is,*

$$\int_{[0,\infty)} \mathbf{1}\big(\bar{W}(t) > 0\big)\, \mathrm{d}\bar{I}(t) = 0.$$

The analysis of the solution $(\bar{A}, \bar{W})$ to (5.3) follows from exactly the same argument as in Proposition 3.1 and remarks afterwards, except that, $\bar{W}$ solves the nonlinear ODE with reflection at 0:

$$\mathrm{d}\bar{W}(t) = \big(\rho(\bar{W}(t)) - 1\big)\, \mathrm{d}t + \mathrm{d}\bar{I}(t) \quad \text{with } \bar{W}(0) = \bar{w}_0,$$

recalling $\rho(\cdot)$ in (2.17) (now as a function of $\bar{W}$ instead of $\bar{Q}$). Comparing with (3.3), *the sup-critical regime*, *the sub-critical regime* and *the critical regime* in (3.4) is now defined for $\bar{w}_0$ instead of $\bar{q}_0$. There can be also multiple equilibrium points in $\mathbb{R}_+$ for the workload dependent model in (5.2).

For the FCLT, we assume that FLLN limit holds at an equilibrium point $\bar{w}_0 \geq 0$ in $\mathcal{L}^=$:

$$\bar{A}(t) = \mu\, t, \quad \bar{W}(t) = \bar{w}_0, \quad \bar{I}(t) = 0 \quad \text{and} \quad \bar{B}(t) = t. \tag{5.4}$$

The CLT-scaled processes are as defined in (4.2) with the above $\bar{A}$ in (5.4).

Theorem 5.2. *Under Assumptions A1, A2 and A4, assuming that there exists a random variable* $\hat{Q}_0$ *such that* $\hat{Q}_0^{(n)} := \sqrt{n} \cdot \big(\bar{Q}_0^{(n)} - \bar{q}_0\big) \Rightarrow \hat{Q}_0$ *as* $n \to \infty$*, we have*

$$\big(\hat{A}^{(n)}, \hat{Q}^{(n)}, \hat{W}^{(n)}\big) \Rightarrow \big(\hat{A}, \hat{Q}, \hat{W}\big) \quad \text{in} \quad (\mathbb{D}^3, J_1) \quad n \to \infty.$$

The limit process $\big(\hat{A}, \hat{Q}, \hat{W}\big)$ *is given as follows in the two cases.*

(i) Case of $\bar{q}_0 = 0$*:* $(\hat{A}, \hat{W})$ *is the unique strong solution to the SDE such that* $\hat{W}$ *has a reflection at* 0*:*

$$\begin{aligned} \mathrm{d}\hat{A}(t) &= \frac{\hat{\lambda}_0 + \mu\, \hat{H}_0(\hat{W}(t))}{1 - H(0)}\, \mathrm{d}t + \frac{\mathrm{d}\hat{X}(t)}{1 - H(0)}, \\ \mathrm{d}\hat{W}(t) &= \mu^{-1}\big(-\hat{\mu}\, \mathrm{d}t + \mathrm{d}\hat{A}(t) - \mathrm{d}\hat{S}(t)\big) + \mathrm{d}\hat{I}(t), \end{aligned} \tag{5.5}$$

with $\hat{A}(0) = 0$ *and* $\hat{W}(0) = \mu^{-1}\hat{Q}_0$*, where* $\hat{I}$ *is the minimal nondecreasing process in* $\mathbb{C}$ *so that* $\hat{W}(t) \geq 0$ *for every* $t \geq 0$*, and* $\hat{I}$ *increases only when* $\hat{W}$ *is zero, that is,*

$$\int_{[0,\infty)} \mathbf{1}\big(\hat{W}(t) > 0\big)\, \mathrm{d}\hat{I}(t) = 0.$$

where $\hat{X}$ *and* $\hat{S}$ *are as given in Theorem 4.1. Moreover,*

$$\hat{Q} = \mu \cdot \hat{W}.$$

(ii) Case of $\bar{q}_0 > 0$*:* $(\hat{A}, \hat{W})$ *is the unique strong solution to the SDE:*

$$\begin{aligned} \mathrm{d}\hat{A}(t) &= \frac{\hat{\lambda}_0 + \mu\, \hat{H}_{\bar{w}_0}(\hat{W}(t))}{1 - H(\bar{w}_0)}\, \mathrm{d}t + \frac{\mathrm{d}\hat{X}(t)}{1 - H(\bar{w}_0)}, \\ \mathrm{d}\hat{W}(t) &= \mu^{-1}\big(-\hat{\mu}\, \mathrm{d}t + \mathrm{d}\hat{A}(t) - \mathrm{d}\hat{S}(t + \bar{w}_0)\big), \end{aligned} \tag{5.6}$$

with $\hat{A}(0) = 0$ *and* $\hat{W}(0) = \mu^{-1}(\hat{Q}_0 - \bar{w}_0\hat{\mu} - \mu^{-1}\hat{S}(\bar{w}_0))$*, where* $\hat{X}$ *and* $\hat{S}$ *are as given in Theorem 4.1. Moreover,*

$$\hat{Q}(t) = \mu\, \hat{W}(t) + \hat{S}(t + \bar{w}_0) - \hat{S}(t) + \bar{w}_0\hat{\mu} = \hat{Q}_0 - \hat{\mu}\, t + \hat{A}(t) - \hat{S}(t).$$

Remark 5.1. *Recalling* $\hat{\rho}_{\bar{w}_0}(\cdot)$ *defined in (4.10), with* $\lambda_0 = \mu\big(1 - H(\bar{w}_0)\big)$ *for* $\bar{w}_0 \in \mathcal{L}^=$ *in (3.4).*

(1) In the case $\bar{w}_0 = 0$, it follows from (5.5) that $\hat{W}$ solves the SDE with reflection at 0:

$$\mathrm{d}\hat{W}(t) = -\hat{\rho}_0\big(\hat{W}(t)\big)\,\mathrm{d}t + \mathrm{d}\Big(\frac{\hat{X}(t)}{\lambda_0} - \frac{\hat{S}(t)}{\mu}\Big) + \mathrm{d}\hat{I}(t)\,, \tag{5.7}$$

which has a unique strong solution as discussed in Remark 4.2.

(2) In the case $\bar{w}_0 > 0$, it follows that $\hat{W}$ in (5.6) solves the SDE:

$$\begin{aligned}\mathrm{d}\hat{W}(t) &= \Big(\frac{\hat{\lambda}_0 + \mu\,\hat{H}_{\bar{w}_0}(\hat{W}(t))}{\mu(1-H(\bar{w}_0))} - \frac{\hat{\mu}}{\mu}\Big)\,\mathrm{d}t + \frac{\mathrm{d}\hat{X}(t)}{\mu(1-H(\bar{w}_0))} - \frac{\mathrm{d}\hat{S}(t+\bar{w}_0)}{\mu} \\ &= -\,\hat{\rho}_{\bar{w}_0}(\hat{W}(t))\,dt + \mathrm{d}\Big(\frac{1}{\lambda_0}\hat{X}(t) - \frac{1}{\mu}\hat{S}(t+\bar{w}_0)\Big),\end{aligned} \tag{5.8}$$

which has a unique strong solution as discussed in Remark 4.2.

6. Proofs in the queue-length dependent case

6.1. **Preliminaries.** We first observe the following local martingale property associated with the compensated Hawkes process. Recall that for the standard Hawkes process A, the compensated process $X(t) = A(t) - \Lambda(t)$ is a martingale with respect to the natural filtration $\{\mathscr{F}(t)\}_{t\geq 0}$, applying [13, Theorem 14.2.IV] (see also [3]). However, in our case with state-dependence, we can only prove that $X(t)$ is a local martingale (as explained below).

Remark 6.1. *In the case of $h(t,y) = h(t)$ in (2.3) in Remark 2.2, A is a standard Hawkes process and independent of the service times $\{\xi_j\}_{j\geq 1}$. In this case, $\mathbb{E}[A(t)] < \infty$ satisfies the renewal equation*

$$\mathbb{E}[A(t)] = \mathbb{E}[\Lambda(t)] = \lambda_0\, t + \int_0^t h(t-s)\mathbb{E}\big[A(s)\big]\,ds,$$

which gives directly

$$\mathbb{E}\big[A(t)\big] = \lambda_0\Big(t + \int_0^t \psi(t-s)s\,ds\Big) = \lambda_0\int_0^t \Big(1 + \int_0^s \psi(u)\,du\Big)\,ds,$$

c.f., [3, Lemma 4]*, where $\psi = h + h*\psi = \sum_{k\geq 1} h^{*k}$ is the renewal kernel of h.*

Moreover, for the study of standard Hawkes processes, $X = N - \Lambda$ is a martingale with respect to the natural filtration $\{\mathscr{F}(t)\}_{t\geq 0}$, and the following martingale representation is crucial for the proof of the scaling limits (c.f. [3, Lemma 4]*),*

$$\Lambda(t) - \mathbb{E}\big[\Lambda(t)\big] = \int_0^t \psi(t-s)X(s)\,ds.$$

However, because of the state-dependence of the Hawkes process upon the queueing process, the expressions above can no longer be obtained. It is in general unable to find an explicit formula for $\mathbb{E}[A(t)]$, and it could be infinite (see the example below).

We nevertheless have a local martingale property, and we shall use a localization arguments in the proofs.

Remark 6.2. *Considering the case $\xi \sim \exp(1)$, $Q(0) = 0$, $\lambda_0 = 1$ and $h(t,y) = 3 + 2\,y$, for every $T > 0$ on the set $\{\xi_1 > T\}$ (that is, no service is completed before T), we have in (2.3) and (2.7) for every $t \leq T$,*

$$Q(t) = A(t) \quad \textit{and} \quad \lambda(t) = 1 + \int_0^t \big(3 + 2Q(s-)\big)\,\mathrm{d}A(s) = (1+Q(t))^2.$$

From this, we obtain that conditioning on $\{\xi_1 > T\}$, Q is a Markov process on $[0,T]$. It can be easily checked that

$$\mathbb{E}\big[Q(t)\big|\xi_1 > T\big] = \infty \quad \textit{for every } t \in (0,T].$$

Proposition 6.1. *Suppose that $t \to h(t,y)$ is locally integrable on $\mathbb{R}_+$ for every $y \geq 0$. Then the process $X = \{X(t) : t \geq 0\}$ defined by $X(t) = A(t) - \Lambda(t)$ is an $\{\mathscr{F}(t)\}_{t\geq 0}$-local martingale with optional quadratic variation A, recalling the filtration $\{\mathscr{F}(t)\}_{t\geq 0}$ in (2.2). The same property holds for the sequence of processes $X^{(n)} = \{X^{(n)}(t) : t \geq 0\}$, defined by $X^{(n)}(t) = A^{(n)}(t) - \Lambda^{(n)}(t)$, that is, $X^{(n)}$ is a local martingale with respect to the filtration $\{\mathscr{F}^{(n)}(t)\}_{t\geq 0}$ in (2.11).*

Proof. Applying Fubini's theorem, one can find that for every $t > s > 0$ and $n \in \mathbb{N}$,

$$\Lambda(t\wedge\tau_n) - \Lambda(s\wedge\tau_n) = \int_s^t \lambda(u)\mathbf{1}(u < \tau_n)\,du = \sum_{k=1}^{n}\int_s^t \lambda_k(u)\mathbf{1}(\tau_{k-1} \leq u < \tau_k)\,du, \tag{6.1}$$

where $\lambda_k(\cdot)$ denotes the local hazard function for the k^{th}-event time and is defined by

$$\lambda_k(t) = \lambda_0 + \sum_{1\leq j\leq k-1} h\big(t-\tau_j, Q(\tau_j-)\big) \in \mathscr{F}(\tau_{k-1}) \quad \forall t \geq \tau_{k-1}.$$

By definition, for every $t > s > 0$,

$$\mathbb{P}\big(\tau_k > t\big|\mathscr{F}(\tau_{k-1})\big) = \exp\Big(-\int_{\tau_{k-1}}^t \lambda_k(u)\,du\Big) \quad \text{on the set } \{\tau_{k-1} \leq t\}, \tag{6.2}$$

$$\mathbb{P}\big(\tau_k > t\big|\mathscr{F}(s)\big) = \exp\Big(-\int_{s}^t \lambda_k(u)\,du\Big) \quad \text{on the set } \{\tau_{k-1} \leq s < \tau_k\}. \tag{6.3}$$

We claim first in (6.1) that for every $t > s \geq 0$ and $k \in \mathbb{N}$,

$$\mathbb{E}\Big[\int_s^t \lambda_k(u)\mathbf{1}(\tau_{k-1} \leq u < \tau_k)\,du\Big|\mathscr{F}(s)\Big] = \mathbb{P}\big(\tau_k \in (s,t]\big|\mathscr{F}(s)\big). \tag{6.4}$$

Plugging (6.4) into (6.1) gives

$$\mathbb{E}\big[\Lambda(t\wedge\tau_n) - \Lambda(s\wedge\tau_n)\big|\mathscr{F}(s)\big] = \mathbb{E}\Big[\sum_{k=1}^{n}\mathbf{1}(\tau_k \in (s,t])\Big|\mathscr{F}(s)\Big] = \mathbb{E}\big[A(t\wedge\tau_n) - A(t\wedge\tau_n)\big|\mathscr{F}(s)\big],$$

which proves the martingale property of $t \to X(t\wedge\tau_n)$.

To prove (6.4), on the set $\{s < \tau_{k-1}\}$ by conditioning on $\mathscr{F}(\tau_{k-1})$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{1}(s < \tau_{k-1})\int_s^t \lambda_k(u)\mathbf{1}(\tau_{k-1} \leq u < \tau_k)\,du\Big|\mathscr{F}(\tau_{k-1})\Big]\\
&= \int_s^t \mathbb{E}\big[\lambda_k(u)\mathbf{1}(s < \tau_{k-1} \leq u)\mathbf{1}(u < \tau_k)\big|\mathscr{F}(\tau_{k-1})\big]\,du\\
&= \mathbf{1}(s < \tau_{k-1} \leq t)\int_{\tau_{k-1}}^t \lambda_k(u)\mathbb{P}\big(\tau_k > u\big|\mathscr{F}(\tau_{k-1})\big]\,du.
\end{aligned}$$

Further applying (6.2), the identity above is equal to

$$\begin{aligned}
&\mathbf{1}(s < \tau_{k-1} \leq t)\int_{\tau_{k-1}}^t \lambda_k(u)\exp\Big(-\int_{\tau_{k-1}}^u \lambda_k(v)\,dv\Big)\,du\\
&= \mathbf{1}(s < \tau_{k-1} \leq t)\Big(1 - \exp\Big(-\int_{\tau_{k-1}}^t \lambda_k(v)\,dv\Big)\Big) = \mathbb{P}\big(s < \tau_{k-1} \leq \tau_k \leq t\big|\mathscr{F}(\tau_{k-1})\big).
\end{aligned}$$

By conditioning on $\mathscr{F}(s)$ on the set $\{s < \tau_{k-1}\}$, we obtain

$$\mathbb{E}\Big[\mathbf{1}(\tau_{k-1} > s)\int_s^t \lambda(u)\mathbf{1}(\tau_{k-1} \leq u < \tau_k)\,du\Big|\mathscr{F}(s)\Big] = \mathbb{P}\big(s < \tau_{k-1} \leq \tau_k \leq t\big|\mathscr{F}(s)\big). \tag{6.5}$$

Similarly, on the set $\{\tau_{k-1} \le s\}$, by the fact $\lambda_k(u)\mathbf{1}(\tau_{k-1} \le s < \tau_k) \in \mathscr{F}(s)$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{1}(\tau_{k-1} \le s) \int_s^t \lambda_k(u)\mathbf{1}(\tau_{k-1} \le u < \tau_k)\, du \Big| \mathscr{F}(s)\Big] \\
&= \int_s^t \mathbb{E}\big[\lambda_k(u)\mathbf{1}(\tau_{k-1} \le s < \tau_k)\mathbf{1}(u < \tau_k)\big|\mathscr{F}(s)\big]\, du \\
&= \int_s^t \lambda_k(u)\mathbf{1}(\tau_{k-1} \le s < \tau_k)\mathbb{P}\big(\tau_k > u\big|\mathscr{F}(s)\big)\, du.
\end{aligned}$$

Applying (6.3), the identity above gives

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{1}(\tau_{k-1} \le s) \int_s^t \lambda_k(u)\mathbf{1}(\tau_{k-1} \le u < \tau_k)\, du \Big| \mathscr{F}(s)\Big] \\
&= \mathbf{1}(\tau_{k-1} \le s < \tau_k) \int_s^t \lambda_k(u) \exp\Big(-\int_s^u \lambda_k(v)\, dv\Big)\, du \\
&= \mathbf{1}(\tau_{k-1} \le s < \tau_k)\Big(1 - \exp\Big(-\int_s^t \lambda_k(v)\, dv\Big)\Big) = \mathbb{P}\big(\tau_{k-1} \le s < \tau_k \le t\big|\mathscr{F}(s)\big).
\end{aligned}$$

This identity, together with (6.5), proves (6.4). This finishes the proof. □

Under Assumption A2, we have the following scaling limits for the processes $U^{(n)}$ and $S^{(n)}$.

Lemma 6.1. *Under Assumption A2-(2.15),*

$$\big(\bar{U}^{(n)}, \bar{S}^{(n)}\big)(t) = n^{-1}\big(U^{(n)}([nt]), S^{(n)}(nt)\big) \to (m, \mu)\cdot t \quad \textit{u.o.c. in probability,}$$

as $n \to \infty$, where $\mu = m^{-1}$. If, in addition, (2.16) holds, then as $n \to \infty$,

$$\big(\hat{U}^{(n)}, \hat{S}^{(n)}\big) := \sqrt{n}\big((\bar{U}^{(n)} - m^{(n)}\mathfrak{e}), (\bar{S}^{(n)} - \mu^{(n)}\mathfrak{e})\big) \Rightarrow (\hat{U}, \hat{S}) \quad \textit{in } (\mathbb{D}^2, J_1),$$

where $\mathfrak{e}$ denotes the identity function on $\mathbb{R}_+$, and $\hat{U}$ is a mean-zero Brownian motion with variance σ^2, and $\hat{S}(t) = -\mu\hat{U}(\mu t)$.

Remark 6.3. *Lemma 6.1 holds under more general conditions on the triangular array $\{\xi_j^{(n)}\}_{n,j}$ by localization. For example, one can replace (2.15) for the FLLN by the following weaker condition:*

$$\tilde{m}^{(n)} := \mathbb{E}\big[\xi^{(n)}; \xi^{(n)} < n\big] \to m \quad \textit{and} \quad n\cdot\mathbb{P}\big(\xi^{(n)} > n\varepsilon\big) \to 0,$$

c.f. [16, Theorem 2.2.11].

We recall the definition of reflection mapping for functions in $\mathbb{D}(\mathbb{R}_+)$.

Definition 6.1 (reflection mapping)**.** *For any $x \in \mathbb{D}(\mathbb{R}_+)$ with $x(0) \ge 0$,*

$$y(t) := \psi(x)(t) = \sup_{s\le t}(-x(s))^+, \quad \forall t \ge 0,$$

is the unique increasing process in $\mathbb{D}(\mathbb{R}_+)$ such that

$$y(0) = 0, \quad z = x + y =: \phi(x) \ge 0 \quad \textit{and} \quad \int_0^\infty \mathbf{1}\big(z(t) > 0\big)\, \mathrm{d}y(t) = 0,$$

where z is called the reflected process of x, and y is called the regulator of x.

It is shown that the mappings ψ and ϕ are Lipschitz continuous on $\mathbb{D}[0,T]$ under the uniform topology for every $T > 0$ (see, e.g., [8, Theorem 6.1] and [50, Chapter 14.2]).

6.2. **Proof of the FLLN.** Define the associated LLN-scaled processes, besides $(\bar{A}^{(n)}, \bar{Q}^{(n)}, \bar{W}^{(n)})$ in (3.1):

$$\big(\bar{X}^{(n)}, \bar{\Lambda}^{(n)}, \bar{B}^{(n)}, \bar{I}^{(n)}\big)(t) = \frac{1}{n}\big(X^{(n)}, \Lambda^{(n)}, B^{(n)}, I^{(n)}\big)(nt),$$

and let $\bar{\mathscr{F}}^{(n)}(t) = \mathscr{F}^{(n)}(nt)$ for $t \geq 0$. It is clear from (2.8) and (2.10) that

$$\bar{I}^{(n)}(t) = \int_0^t \mathbf{1}\big(\bar{Q}^{(n)}(s) = 0\big)\,\mathrm{d}s \quad \text{and} \quad \bar{B}^{(n)}(t) + \bar{I}^{(n)}(t) = t. \tag{6.6}$$

We also rewrite (2.7) as

$$\bar{Q}^{(n)}(t) = \bar{q}_0^{(n)} + \bar{A}^{(n)}(t) - \bar{S}^{(n)}(\bar{B}^{(n)}(t)) = \bar{Z}^{(n)}(t) + \mu^{(n)}\bar{I}^{(n)}(t), \tag{6.7}$$

where

$$\bar{Z}^{(n)}(t) = \bar{q}_0^{(n)} + \bar{A}^{(n)}(t) - \mu^{(n)}t - \big(\bar{S}^{(n)}(\bar{B}^{(n)}(t)) - \mu^{(n)}\bar{B}^{(n)}(t)\big). \tag{6.8}$$

By (6.6), $\bar{I}^{(n)}$ is an increasing process such that $\bar{Q}^{(n)} \geq 0$ and

$$\int_0^\infty \mathbf{1}\big(\bar{Q}^{(n)}(t) > 0\big)\,\mathrm{d}\bar{I}^{(n)}(t) = 0,$$

which means that $\mu^{(n)}\bar{I}^{(n)}$ is the regulator of $\bar{Z}^{(n)}$ in (6.7), that is,

$$\mu^{(n)}\bar{I}^{(n)}(t) = \psi\big(\bar{Z}^{(n)}\big)(t) \quad \text{and} \quad \bar{Q}^{(n)}(t) = \phi\big(\bar{Z}^{(n)}\big)(t). \tag{6.9}$$

By (2.9), we also have

$$\bar{W}^{(n)}(t) = \bar{U}^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t)\big) - \bar{B}^{(n)}(t). \tag{6.10}$$

Next, by change of variables, we obtain from (2.12) that

$$\begin{aligned}
\bar{A}^{(n)}(t) &= \bar{X}^{(n)}(t) + \bar{\Lambda}^{(n)}(t) \\
&= \bar{X}^{(n)}(t) + \lambda_0^{(n)}t + \int_0^t \bigg(\int_0^{n(t-u)} h^{(n)}\big(v, n\bar{Q}^{(n)}(u-)\big)\,\mathrm{d}v\bigg)\mathrm{d}\bar{A}^{(n)}(u) \\
&= \lambda_0^{(n)}t + \int_0^t H^{(n)}\big(\bar{Q}^{(n)}(u-)\big)\,\mathrm{d}\bar{A}^{(n)}(u) + \bar{X}^{(n)}(t) - \bar{\varepsilon}_1^{(n)}(t),
\end{aligned} \tag{6.11}$$

recalling $H^{(n)}$ in Assumption A1-(ii), where

$$\bar{\varepsilon}_1^{(n)}(t) = \int_0^t \bigg(\int_{n(t-u)}^\infty h^{(n)}\big(v, n\bar{Q}^{(n)}(u-)\big)\,\mathrm{d}v\bigg)\mathrm{d}\bar{A}^{(n)}(u). \tag{6.12}$$

We will use a localization technique in the proofs, that is, fixing arbitrary $k_0 > \bar{q}_0$, let

$$\bar{\tau}_{+,k_0}^{(n)} := \inf\{t > 0, \bar{Q}^{(n)}(t) > k_0\}, \tag{6.13}$$

with the convention $\inf\emptyset = \infty$, and denote by

$$\alpha_{k_0} := \sup_{n\geq 1, y\leq k_0} H^{(n)}(y) \ \text{ and } \ \bar{J}_{k_0}^{(n)}(t) := \sup_{y\leq k_0}\int_t^\infty h^{(n)}(u, ny)\,\mathrm{d}u \ \text{ for } t \geq 0. \tag{6.14}$$

For notational brevity we drop the subscript k_0 in $\bar{\tau}_{+,k_0}^{(n)}$, α_{k_0} and $\bar{J}_{k_0}^{(n)}$, and c_0 denotes a constant that may vary from line to line. Under Assumption A1-(ii) and (iii), the variables above are finite. The following facts on $h^{(n)}$ are frequently used

$$\begin{aligned}
&H^{(n)}(y) \leq \alpha \ \text{ for all } n \in \mathbb{N} \text{ and } y \leq k_0, && \bar{J}^{(n)}(0) \leq \alpha < 1, \\
&\limsup_{n\to\infty} \bar{J}^{(n)}(n\varepsilon) = 0 \ \text{ for every } \varepsilon > 0, && \bar{J}^{(n)}(\infty) = 0 \ \text{ for every } n \in \mathbb{N}.
\end{aligned} \tag{6.15}$$

We have the following lemmas, which will be used extensively in the proof.

Lemma 6.2. *Under Assumption A1, $\bar{X}^{(n)}(\cdot \wedge \bar{\tau}_+^{(n)})$ is a martingale with respect to $\{\bar{\mathscr{F}}^{(n)}(t)\}_{t\geq 0}$, recalling that $\bar{\mathscr{F}}^{(n)}(t) = \mathscr{F}^{(n)}(nt)$. There is a constant $c_0 > 0$, such that, for every $t > 0$,*

$$\mathbb{E}\Big[\sup_{s\leq t}(\bar{X}^{(n)})^2(s \wedge \bar{\tau}_+^{(n)})\Big] \leq \frac{c_0}{n} \cdot t. \tag{6.16}$$

Hence, $\bar{X}^{(n)}(\cdot \wedge \bar{\tau}_+^{(n)})$ converges to 0 in $L^2(\mathbb{P})$.

Proof. By the fact $\bar{\varepsilon}_1^{(n)} \geq 0$, we already have from (6.11) that for every $t > 0$

$$\bar{A}^{(n)}(t) \leq \bar{X}^{(n)}(t) + \lambda_0^{(n)} t + \int_0^t H^{(n)}\big(\bar{Q}^{(n)}(s-)\big)\mathrm{d}\bar{A}^{(n)}(s)\,.$$

It follows from the local martingale property of $\bar{X}^{(n)}$ and the fact $1 - H^{(n)} > 0$ that

$$\mathbb{E}\Big[\int_0^{t\wedge\bar{\tau}} \Big(1 - H^{(n)}\big(\bar{Q}^{(n)}(s-)\big)\Big)\mathrm{d}\bar{A}^{(n)}(s)\Big] \leq \lambda_0^{(n)} \cdot t$$

for every stopping time $\bar{\tau}$. In particular, for $\bar{\tau} = \bar{\tau}_+^{(n)}$ in (6.13) we have

$$\mathbb{E}\big[\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)})\big] \leq \frac{\lambda_0^{(n)}}{1-\alpha} \cdot t, \tag{6.17}$$

with the bound in (6.15) applied to every $H^{(n)}(\bar{Q}^{(n)}(s-))$.

Applying the Burkholder-Davis-Gundy (BDG) inequality, c.f. [27, Theorem 20.12], to the local martingales $X^{(n)}$, which has quadratic variation $A^{(n)}$, gives

$$\mathbb{E}\Big[\sup_{s\leq t}\big(\bar{X}^{(n)}\big)^2(s \wedge \bar{\tau}_+^{(n)})\Big] \leq c_2 \cdot \mathbb{E}\big[[\bar{X}^{(n)}](t \wedge \bar{\tau}_+^{(n)})\big] = \frac{c_2}{n} \cdot \mathbb{E}\big[\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)})\big],$$

where c_2 is the constant from BDG's inequality. Further applying (6.17) and [44, Theorem 51] proves the martingale property of $\bar{X}^{(n)}(\cdot \wedge \bar{\tau}_+^{(n)})$, as well as its convergence in $L^2(\mathbb{P})$. □

Recalling the the following modulus of continuity of an arbitrary function x on $[0,T]$ from [6, equation (7.1)]:

$$w_\delta(x,T) := \sup_{0\leq u\leq v\leq T, u-v\leq\delta} \big|x(u) - x(v)\big|, \quad \forall \delta > 0. \tag{6.18}$$

Lemma 6.3. *Under Assumption A1, for every $T > 0$ and $\delta > 0$,*

$$\limsup_{n\to\infty} \mathbb{E}\Big[w_\delta\big(\bar{A}^{(n)}(\cdot \wedge \bar{\tau}_+^{(n)}), T\big)\Big] \leq c_0 \cdot \delta, \tag{6.19}$$

and

$$\limsup_{n\to\infty} \mathbb{E}\Big[\sup_{t\leq T} \bar{\varepsilon}_1^{(n)}(t \wedge \bar{\tau}_+^{(n)})\Big] = 0. \tag{6.20}$$

Lemma 6.3 is proved with the help of Lemma 6.4 below.

Lemma 6.4. *Under Assumption A1, for every $t > 0$,*

$$\limsup_{n\to\infty} \mathbb{E}\Big[\bar{\varepsilon}_1^{(n)}(t); t \leq \bar{\tau}_+^{(n)}\Big] = 0.$$

Proof. We show first that there exists some $c_0 > 0$ such that the following inequality holds:

$$\limsup_{n\to\infty} \mathbb{E}\Big[\bar{A}^{(n)}\big(t \wedge \bar{\tau}_+^{(n)}\big) - \bar{A}^{(n)}\big(s \wedge \bar{\tau}_+^{(n)}\big)\Big] \leq c_0 \cdot (t-s) \quad \forall t > s > 0. \tag{6.21}$$

Making use of (6.11) and the fact $\bar{\varepsilon}_1^{(n)} \geq 0$, we have for every $t > s > 0$,

$$\begin{aligned}
\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}(s \wedge \bar{\tau}_+^{(n)}) &= \big(\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}(s)\big)\mathbf{1}(s \leq \bar{\tau}_+^{(n)}) \\
&\leq \big(\bar{X}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{X}^{(n)}(s \wedge \bar{\tau}_+^{(n)})\big) + \lambda_0^{(n)}(t-s) \\
&\quad + \int_s^t H^{(n)}\big(\bar{Q}^{(n)}(u-)\big)\mathrm{d}\bar{A}^{(n)}(u \wedge \bar{\tau}_+^{(n)}) + \bar{\varepsilon}_1^{(n)}(s)\mathbf{1}(s \leq \bar{\tau}_+^{(n)}) \\
&\leq \big(\bar{X}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{X}^{(n)}(s \wedge \bar{\tau}_+^{(n)})\big) + \lambda_0^{(n)}(t-s) \\
&\quad + \alpha \cdot \big(\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}(s \wedge \bar{\tau}_+^{(n)})\big) + \bar{\varepsilon}_1^{(n)}(s)\mathbf{1}(s \leq \bar{\tau}_+^{(n)}),
\end{aligned}$$

where the bound for $H^{(n)}$ in (6.15) is applied in the last line above, which gives

$$\begin{aligned}
&(1-\alpha) \cdot \big(\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}(s \wedge \bar{\tau}_+^{(n)})\big) \\
&\leq \big(\bar{X}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{X}^{(n)}(s \wedge \bar{\tau}_+^{(n)})\big) + \lambda_0^{(n)}(t-s) + \bar{\varepsilon}_1^{(n)}(s)\mathbf{1}(s \leq \bar{\tau}_+^{(n)}).
\end{aligned} \tag{6.22}$$

For $\bar{\varepsilon}_1^{(n)}$ in (6.12), applying (6.14) and the monotonicity of $\bar{J}^{(n)}$, we have for every $r \in (0, s)$,

$$\begin{aligned}
\bar{\varepsilon}_1^{(n)}(s)\mathbf{1}(s \leq \bar{\tau}_+^{(n)}) &\leq \int_0^s \bar{J}^{(n)}\big(n(s-u)\big)\,\mathrm{d}\bar{A}^{(n)}(u \wedge \bar{\tau}_+^{(n)}) \\
&\leq \int_0^r \bar{J}^{(n)}\big(n(s-r)\big)\,\mathrm{d}\bar{A}^{(n)}(u \wedge \bar{\tau}_+^{(n)}) + \bar{J}^{(n)}(0)\int_r^s \mathrm{d}\bar{A}^{(n)}(u \wedge \bar{\tau}_+^{(n)}).
\end{aligned} \tag{6.23}$$

Taking expectations on both sides of (6.22) and (6.23), and making use of the martingale property of $\bar{X}^{(n)}(\cdot \wedge \bar{\tau}_+^{(n)})$ in Lemma 6.2 gives

$$\begin{aligned}
&(1-\alpha) \cdot \mathbb{E}\big[\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}(s \wedge \bar{\tau}_+^{(n)})\big] \\
&\leq \lambda_0^{(n)}(t-s) + \bar{J}^{(n)}\big(n(s-r)\big) \cdot \mathbb{E}\big[\bar{A}^{(n)}(r \wedge \bar{\tau}_+^{(n)})\big] + \mathbb{E}\big[\bar{A}^{(n)}(s \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}(r \wedge \bar{\tau}_+^{(n)})\big].
\end{aligned} \tag{6.24}$$

Fixing $t > s > 0$, letting $m_0 \in \mathbb{N}$ and $\delta < (t-s) \wedge 2s$, considering a partition on $[s-\delta, s]$ with

$$t_k = s - \frac{\delta}{m_0}k \in [s-\delta, s] \qquad \forall k = 0, \cdots, m_0,$$

and taking $(s, r) = (t_k, t_{k+1})$ in (6.24) and summing over k gives

$$\begin{aligned}
&m_0 \cdot (1-\alpha) \cdot \mathbb{E}\Big[\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}\big(s \wedge \bar{\tau}_+^{(n)}\big)\Big] \\
&\leq (1-\alpha) \cdot \sum_{k=0}^{m_0-1} \mathbb{E}\Big[\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}\big(t_k \wedge \bar{\tau}_+^{(n)}\big)\Big] \\
&\leq m_0 \cdot \left(\lambda_0^{(n)}\big((t-s)+\delta\big) + \bar{J}^{(n)}\big(\frac{n\delta}{m_0}\big) \cdot \mathbb{E}\big[\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)})\big]\right) + \mathbb{E}\big[\bar{A}^{(n)}(s \wedge \bar{\tau}_+^{(n)})\big].
\end{aligned}$$

Letting $n \to \infty$, applying (6.17) and the fact $\bar{J}^{(n)}(n\varepsilon) \to 0$ in (6.15), we have

$$m_0 \cdot \limsup_{n\to\infty} \mathbb{E}\Big[\bar{A}^{(n)}\big(t \wedge \bar{\tau}_+^{(n)}\big) - \bar{A}^{(n)}\big(s \wedge \bar{\tau}_+^{(n)}\big)\Big] \leq c_0 \cdot \big(m_0(t-s) + t\big),$$

where c_0 is a constant independent of m_0. Further letting $m_0 \to \infty$ proves (6.21).

For the last, for arbitrary $0 < \delta < t$, taking $(s, r) = (t, t-\delta)$ in (6.23), we have

$$\bar{\varepsilon}_1^{(n)}(t)\mathbf{1}(t \leq \bar{\tau}_+^{(n)}) \leq \bar{J}^{(n)}(n\delta) \cdot \bar{A}(t \wedge \bar{\tau}_+^{(n)}) + \big(\bar{A}(t \wedge \bar{\tau}_+^{(n)}) - \bar{A}((t-\delta) \wedge \bar{\tau}_+^{(n)})\big).$$

Then (6.17) and (6.21) can be applied to finish the proof. □

Proof of Lemma 6.3. Fix $\delta > 0$ and take

$$t_0 = 0 < t_1 < \cdots < t_v = T \quad \text{with} \quad t_j - t_{j-1} = \delta \quad \text{for } 2 \le j \le v \text{ and } t_1 \le \delta.$$

By the monotonicity of $\bar{\Lambda}^{(n)}$, we obtain that

$$w_\delta\big(\bar{A}^{(n)}(\cdot \wedge \bar{\tau}_+^{(n)}), T\big) \le 2 \max_{1 \le j \le v} \Big(\bar{A}^{(n)}(t_j \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}(t_{j-1} \wedge \bar{\tau}_+^{(n)})\Big),$$

similar to [6, equation (7.10)]. Taking $(t, s) = (t_{j+1}, t_j)$ in (6.22) for $j \ge 1$, we have

$$\begin{aligned} w_\delta\big(\bar{A}^{(n)}(\cdot \wedge \bar{\tau}_+^{(n)}), T\big) &\le 2 \cdot \bar{A}^{(n)}(t_1 \wedge \bar{\tau}_+^{(n)}) \\ &\quad + \frac{2}{1-\alpha} \cdot \Big(\lambda_0^{(n)} \delta + 2 \cdot \sup_{t \le T} \big|\bar{X}^{(n)}(t \wedge \bar{\tau}_+^{(n)})\big| + \sum_{j=1}^{v-1} \bar{\varepsilon}_1^{(n)}(t_j) \mathbf{1}(t_j \le \bar{\tau}_+^{(n)})\Big). \end{aligned}$$

Taking expectation on both sides above, noticing that the last summand is of $[T/\delta]$ many, letting $n \to \infty$, we have from (6.17), Lemmas 6.2 and 6.4 that

$$\limsup_{n \to \infty} \mathbb{E}\Big[w_\delta\big(\bar{A}^{(n)}(\cdot \wedge \bar{\tau}_+^{(n)}), T\big)\Big] \le \frac{2\,\lambda_0^{(n)}}{1-\alpha}\,(t_1 + \delta) \le \frac{4\,\lambda_0^{(n)}}{1-\alpha} \cdot \delta. \tag{6.25}$$

This proves (6.19).

Observe that $\bar{J}^{(n)}$ is a decreasing function on $\mathbb{R}_+$ with $\bar{J}^{(n)}(0) < 1$ and $\bar{J}^{(n)}(\infty) = 0$ as shown in (6.15). So we can write $(-\bar{J}^{(n)})(\mathrm{d}v)$ to denote the associated Stieltjes measure. Applying Fubini's theorem to (6.23), we obtain for $t > 0$,

$$\begin{aligned} \bar{\varepsilon}_1^{(n)}(t \wedge \bar{\tau}_+^{(n)}) &\le \int_0^{t \wedge \bar{\tau}_+^{(n)}} \int_{t \wedge \bar{\tau}_+^{(n)} - u}^{\infty} (-\bar{J}^{(n)})(n\,\mathrm{d}v) \bar{A}^{(n)}(u) \\ &= \int_0^{\infty} \big(\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)} - v)\big)\,(-\bar{J}^{(n)})(n\,\mathrm{d}v). \end{aligned} \tag{6.26}$$

It follows for every $\delta > 0$ and $t \le T$ that

$$\bar{\varepsilon}_1^{(n)}(t \wedge \bar{\tau}_+^{(n)}) \le \bar{J}^{(n)}(n\delta) \cdot \bar{A}^{(n)}(T \wedge \bar{\tau}_+^{(n)}) + w_\delta\big(\bar{A}^{(n)}(\cdot \wedge \bar{\tau}_+^{(n)}), T\big).$$

Hence, applying (6.17) and (6.25) the continuity for $\bar{A}^{(n)}$ proves (6.20). □

Now, we are ready to prove our FLLN by making use of Lemmas 6.2 and 6.3. We show first that the joint prelimit process is a $\mathbb{C}$-tight family, and then that every limit satisfies (3.2).

Proof of Theorem 3.1. For every $t > s \ge 0$, by (6.8) and the fact $\bar{B}^{(n)}(t) \le t$, we have

$$\big|\bar{Z}^{(n)}(t \wedge \bar{\tau}_+^{(n)})\big| \le \bar{q}_0^{(n)} + \bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) + \mu^{(n)} t + \sup_{r \le t} \big|\bar{S}^{(n)}(r) - \mu^{(n)} r\big|,$$

and

$$\begin{aligned} &\big|\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}(s \wedge \bar{\tau}_+^{(n)})\big| + \big|\bar{Z}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{Z}^{(n)}(s \wedge \bar{\tau}_+^{(n)})\big| \\ &\le 2 \cdot \big|\bar{A}^{(n)}(t \wedge \bar{\tau}_+^{(n)}) - \bar{A}^{(n)}(s \wedge \bar{\tau}_+^{(n)})\big| + \mu^{(n)}(t - s) + 2 \cdot \sup_{r \le t} \big|\bar{S}^{(n)}(r) - \mu^{(n)} r\big|. \end{aligned}$$

We obtain from the expressions above that $\{\big(\bar{A}^{(n)}, \bar{Z}^{(n)}\big)(\cdot \wedge \bar{\tau}_+^{(n)})\}$ is a tight family in the product space $(\mathbb{D}^2[0, T], J_1)$ by Lemmas 6.1 and 6.3. Since $\Delta \bar{A}^{(n)} = n^{-1} = \Delta \bar{Z}^{(n)}$ is uniformly bounded, it is clear that every limit point has a continuous sample path, and thus $\{\big(\bar{A}^{(n)}, \bar{Z}^{(n)}\big)(\cdot \wedge \bar{\tau}_+^{(n)})\}$ is a $\mathbb{C}$-tight family.

Now, for every $k_0 > 0$, let $(\bar{q}_0, \bar{A}_{k_0}, \bar{Z}_{k_0})$ be a limit point of $(\bar{q}_0^{(n)}, \bar{A}^{(n)}, \bar{Z}^{(n)})$ over some subsequence $\{n_k\}_{k\geq 1}$, that is, in the product space $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+)) \times (\mathbb{D}^2[0,T], J_1)$,

$$\big(\bar{q}_0^{(n)}, \bar{A}^{(n)}, \bar{Z}^{(n)}\big)(\cdot \wedge \bar{\tau}_{+,k_0}^{(n)})\Big|_{n=n_k} \Rightarrow \big(\bar{q}_0, \bar{A}_{k_0}, \bar{Z}_{k_0}\big). \tag{6.27}$$

Applying the limit for $\bar{S}^{(n)}$ in Lemma 6.1 to (6.8), we obtain that for the same subsequence,

$$\big(\bar{A}^{(n)}, \bar{Z}^{(n)}, \mathfrak{e}\big)(\cdot \wedge \bar{\tau}_{+,k_0}^{(n)})\Big|_{n=n_k} \Rightarrow \big(\bar{A}_{k_0}, \bar{Z}_{k_0}, \mathfrak{e}_{k_0}\big) \quad \text{in} \quad (\mathbb{D}^3[0,T], J_1), \tag{6.28}$$

where $t \to \mathfrak{e}_{k_0}(t)$ is a continuous function on $[0,T]$ and satisfies

$$\bar{Z}_{k_0}(t) = \bar{q}_0 + \bar{A}_{k_0}(t) - \mu \cdot \mathfrak{e}_{k_0}(t).$$

Then using the Lipschitz continuity property of the reflection mapping in (6.9), and applying the continuous mapping theorem, we obtain

$$\big(\bar{A}^{(n)}, \bar{Z}^{(n)}, \mathfrak{e}, \bar{I}^{(n)}, \bar{Q}^{(n)}\big)(\cdot \wedge \bar{\tau}_{+,k_0}^{(n)})\Big|_{n=n_k} \Rightarrow \big(\bar{A}_{k_0}, \bar{Z}_{k_0}, \mathfrak{e}_{k_0}, \bar{I}_{k_0}, \bar{Q}_{k_0}\big) \quad \text{in} \quad (\mathbb{D}^5[0,T], J_1), \tag{6.29}$$

where

$$\mu \cdot \bar{I}_{k_0} = \psi\big(\bar{Z}_{k_0}\big), \quad \bar{Q}_{k_0} = \phi\big(\bar{Z}_{k_0}\big) \quad \text{and} \quad t_{k_0} = t \wedge \bar{\tau}_{+,k_0}, \tag{6.30}$$

with $\bar{\tau}_{+,k_0} := \inf\{s > 0, \bar{Q}_{k_0}(s) \geq k_0\}$.

Notice that, $(\bar{A}_{k_0}, \bar{Z}_{k_0}, \bar{I}_{k_0}, \bar{Q}_{k_0})$ may not be a deterministic function at this stage, and thus the convergence in (6.29) may fail to hold in probability. Neither do we have $t_{k_0} = t \wedge \bar{\tau}_{+,k_0}$ in (6.28) at that moment. The identity $t_{k_0} = t \wedge \bar{\tau}_{+,k_0}$ in (6.30) follows from the continuous mapping theorem.

Next, we claim that the limit $(\bar{q}_0, \bar{A}_{k_0}, \bar{Z}_{k_0})$ in (6.27) is consistent in k_0. For every $k_0' < k_0$, we have by the definition of $\bar{\tau}_{+,k_0}^{(n)}$ for the pre-limit processes, for the same subsequence $\{n_k\}_{k\geq 1}$,

$$\Big(\bar{q}_0^{(n)}, \bar{A}^{(n)}(t \wedge \bar{\tau}_{+,k_0'}^{(n)}), \bar{Z}^{(n)}(t \wedge \bar{\tau}_{+,k_0'}^{(n)})\Big)\Big|_{n=n_k} \Rightarrow \big(\bar{q}_0, \bar{A}_{k_0'}, \bar{Z}_{k_0'}\big),$$

in $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+)) \times (\mathbb{D}^2[0,T], J_1)$. The same argument can be applied and results in the following:

$$\big(\bar{A}^{(n)}, \bar{Z}^{(n)}, \mathfrak{e}, \bar{I}^{(n)}, \bar{Q}^{(n)}\big)(\cdot \wedge \bar{\tau}_{+,k_0'}^{(n)})\Big|_{n=n_k} \Rightarrow \big(\bar{A}_{k_0'}, \bar{Z}_{k_0'}, \mathfrak{e}_{k_0'}, \bar{I}_{k_0'}, \bar{Q}_{k_0'}\big) \quad \text{in } (\mathbb{D}^5[0,T], J_1),$$

comparing with (6.29). Moreover,

$$\big(\bar{A}_{k_0}, \bar{Z}_{k_0}, \bar{I}_{k_0}, \bar{Q}_{k_0}\big)(t) = \big(\bar{A}_{k_0'}, \bar{Z}_{k_0'}, \bar{I}_{k_0'}, \bar{Q}_{k_0'}\big)(t) \quad \forall t < \bar{\tau}_{+,k_0'} := \inf\{t > 0 : \bar{Q}_{k_0}(t) > k_0'\},$$

which is equivalent to the claim that for some common $(\bar{A}, \bar{Z}, \bar{I}, \bar{Q})$,

$$\big(\bar{A}_{k_0}, \bar{Z}_{k_0}, \bar{I}_{k_0}, \bar{Q}_{k_0}\big)(t) = \big(\bar{A}, \bar{Z}, \bar{I}, \bar{Q}\big)(t \wedge \bar{\tau}_{+,k_0}) \quad \text{and} \quad \bar{\tau}_{+,k_0} = \inf\{t > 0 : \bar{Q}(t) > k_0\}. \tag{6.31}$$

Note that the argument above is basically a discussion about the *strong local uniqueness* of solutions, c.f. [33, Page 1058].

Plugging (6.29) and (6.31) into (6.11), we obtain from [33, Theorem 2.2] that

$$\bar{A}(s) = \lambda_0 s + \int_0^s H\big(\bar{Q}(u)\big)\mathrm{d}\bar{A}(u)\Big|_{s=t\wedge\bar{\tau}_{+,k_0}} \quad \text{and} \quad \bar{Q}(s) = \phi(\bar{Z})(s)\Big|_{s=t\wedge\bar{\tau}_{+,k_0}} \quad \forall t \leq T,$$

that is, $(\bar{A}, \bar{Q})$ satisfies (3.2) on $[0, T \wedge \bar{\tau}_{+,k_0}]$, where Lemma 6.2 for $\bar{X}^{(n)}$, Lemma 6.3 for $\bar{\varepsilon}_1^{(n)}$, and the convergence of $H^{(n)}$ to H u.o.c. in Assumption A1-(ii) are applied. Since $\bar{Q}$ is non-explosive under Assumption A3 as discussed in Remark 3.4, we can always take k_0 large enough so that

$$\sup_{t\leq T} \bar{Q}(t) \leq k_0$$

holds for every solution to (3.2). By the continuous mapping theorem, we obtain that

$$\mathbb{P}\big(T \leq \bar{\tau}_{+,k_0}^{(n)}\big) \to \mathbb{P}(T \leq \bar{\tau}_{+,k_0}) = 0$$

which proves the limit for $\big(\bar{A}^{(n)}, \bar{Q}^{(n)}\big)$ without stopping.

For the last, for every fixed k_0, we obtain from (6.7) and (6.10) and Lemma 6.1 that

$$\begin{aligned}\big(\bar{W}^{(n)} - m\cdot\bar{Q}^{(n)}\big)(t\wedge\bar{\tau}_+^{(n)}) &= \big(\bar{U}^{(n)} - m\cdot\mathfrak{e}\big)(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t\wedge\bar{\tau}_+^{(n)}))\\ &\quad + m\cdot\big(\bar{S}^{(n)} - \mu\mathfrak{e}\big)\big(\bar{B}^{(n)}(t\wedge\bar{\tau}_{+,k_0}^{(n)})\big) \to 0 \quad \text{u.o.c. in probability.}\end{aligned}$$

To summarize, we have shown that the joint convergence over $\{n_k\}_{k\geq 1}$ holds for $(\bar{A}^{(n)}, \bar{Z}^{(n)}, \bar{I}^{(n)}, \bar{Q}^{(n)}, \bar{W}^{(n)})$ with

$$\bar{Z}(t) = \bar{q}_0 + \bar{A}(t) - \mu t, \quad \mu\cdot\bar{I}(t) = \psi\big(\bar{Z}\big)(t),$$
$$\bar{Q}(t) = \phi\big(\bar{Z}\big)(t), \quad \bar{W} = m\cdot\bar{Q} \quad \text{and} \quad \bar{A}(s) = \lambda_0 s + \int_0^s H\big(\bar{Q}(u)\big)\mathrm{d}\bar{A}(u),$$

which finishes the proof. □

6.3. **Proof of the FCLT.** Suppose that $\bar{q}_0 \geq 0$ is an equilibrium point in $\mathcal{L}^{=}$, and the fluid limits are as given in (4.1). We define the diffusion-scaled process

$$\hat{X}^{(n)}(t) = \sqrt{n}\,\bar{X}^{(n)}(t) = \sqrt{n}(\bar{A}^{(n)}(t) - \bar{\Lambda}^{(n)}(t))\,, \quad t\geq 0.$$

Recalling $U(k)$ in (2.5), we define

$$U_-^{(n)}(k) = \sum_{j\geq 1}\xi_{-j}^{(n)}, \quad U_+^{(n)}(k) = \sum_{j\geq 1}\xi_j^{(n)},$$

which are the cumulative service times for the jobs initially in the queue and new arrivals associated to the n^{th}-system, respectively. Define the LLN-scaled processes:

$$\bar{U}_-^{(n)}(t) = \frac{1}{n}U_-^{(n)}\big([nt]\big), \quad \hat{U}_-^{(n)}(t) = \sqrt{n}\,\big(\bar{U}_-^{(n)}(t) - m^{(n)}t\big),$$
$$\bar{U}_+^{(n)}(t) = \frac{1}{n}U_+^{(n)}\big([nt]\big), \quad \hat{U}_+^{(n)}(t) = \sqrt{n}\,\big(\bar{U}_+^{(n)}(t) - m^{(n)}t\big).$$

Let $S_-^{(n)}, S_+^{(n)}$ be the associated renewal processes of $U_-^{(n)}(k)$ and $U_+^{(n)}(k)$, respectively. Define the LLN-scaled processes:

$$(\bar{S}_-^{(n)}, \bar{S}_+^{(n)})(t) := n^{-1}\big(S_-^{(n)}, S_+^{(n)}\big)(nt) \quad \text{and} \quad (\hat{S}_-^{(n)}, \hat{S}_+^{(n)})(t) := \sqrt{n}\Big(\big(\bar{S}_-^{(n)}, \bar{S}_+^{(n)}\big)(t) - \mu^{(n)}(t,t)\Big).$$

Recalling $(\hat{U}^{(n)}, \hat{S}^{(n)})$ in Lemma 6.1, we obtain from (2.5) and (2.6) that for $t\geq 0$,

$$\begin{aligned}\hat{S}^{(n)}(t) &= \hat{S}_-^{(n)}\big(t\wedge\bar{U}_-^{(n)}(\bar{q}_0^{(n)})\big) + \hat{S}_+^{(n)}\big((t-\bar{U}_-^{(n)}(\bar{q}_0^{(n)}))^+\big),\\ \hat{U}^{(n)}(\bar{q}_0^{(n)} + t) &= \hat{U}_-^{(n)}(\bar{q}_0^{(n)}) + \hat{U}_+^{(n)}(t) = \hat{U}_+^{(n)}(t) - m^{(n)}\hat{S}_-^{(n)}\big(\bar{U}_-^{(n)}(\bar{q}_0^{(n)})\big).\end{aligned} \tag{6.32}$$

To prove the FCLT, we will apply [33, Theorem 5.4] concerning the convergence of a sequence of solutions to SDEs, where the measurability is used. It is clear that the processes

$$\big\{\bar{A}^{(n)}, \bar{Q}^{(n)}, \bar{W}^{(n)}, \bar{B}^{(n)}, \bar{I}^{(n)}, \bar{\Lambda}^{(n)}, \bar{X}^{(n)}, \bar{S}^{(n)}(\bar{B}^{(n)}), \bar{U}^{(n)}\big(\bar{q}_0 + \bar{A}^{(n)}\big)\big\}$$

are $\{\bar{\mathscr{F}}^{(n)}(t)\}_{t\geq 0}$-adapted process, recalling $\bar{\mathscr{F}}^{(n)}(t) = \mathscr{F}^{(n)}(nt)$ in (2.11).

Lemma 6.5. *Suppose Assumptions A1 and A2 and (4.1) hold for $\bar{q}_0 \in \mathcal{L}^{=}$. We have*

$$\big(\hat{X}^{(n)}, \hat{S}_-^{(n)}, \hat{U}_+^{(n)}(\bar{A}^{(n)})\big) \Rightarrow \big(\hat{X}, \hat{S}_-, -\mu^{-1}\hat{S}_+\big) \quad \textit{in} \quad (\mathbb{D}^3[0,T], J_1),$$
$$\hat{U}_+^{(n)}(\mu t) + \mu^{-1}\hat{S}_+^{(n)}(t) \to 0 \quad \textit{u.o.c. in probability,}$$

where $\hat{X}$, $\hat{S}_-$ and $\hat{S}_+$ are independent mean-zero Brownian motions, where $\hat{X}$ is the one in Theorem 4.1, and $\hat{S}_-$ and $\hat{S}_+$ are two independent copies of $\hat{S}$ in Lemma 6.1.

Proof. By definition and Lemma 6.1, we have $\hat{S}_-^{(n)}(\cdot) \in \bar{\mathscr{F}}^{(n)}(0)$ and $\bar{U}_-^{(n)}(\bar{q}_0^{(n)}) \in \bar{\mathscr{F}}^{(n)}(0)$, and also $\bar{U}_-^{(n)}(\bar{q}_0^{(n)}) \to m\bar{q}_0$ in probability. Moreover, $\hat{X}^{(n)}$ and $\hat{U}_+^{(n)}(\bar{A}^{(n)})$ are $\{\bar{\mathscr{F}}^{(n)}(t)\}_{t\geq 0}$-adapted locally-square-integrable martingales. The convergence for $\hat{S}_-^{(n)}$ follows directly from Lemma 6.1. It thus only remains to check the convergence of $(\hat{X}^{(n)}, \hat{U}_+^{(n)}(\bar{A}^{(n)}))$. Note that the prelimit process $\bar{S}_-^{(n)}$ and $(\hat{X}^{(n)}, \hat{U}_+^{(n)})$ is independent.

We apply [26, Theorem VIII.3.11] to prove the convergence of $\big(\hat{X}^{(n)}(t), \hat{U}_+^{(n)}(\bar{A}^{(n)}(t))\big) \in \bar{\mathscr{F}}^{(n)}(t)$. Since $\hat{X}^{(n)}$ has jumps uniformly bounded by $n^{-1/2}$, and for every $\delta > 0$,

$$\begin{aligned}
&\mathbb{E}\bigg[\sum_{t\leq T}\big(\Delta\hat{U}_+^{(n)}(\bar{A}^{(n)})\big)^2(t\wedge\bar{\tau}_+^{(n)})\mathbf{1}\big(|\Delta\hat{U}_+^{(n)}(\bar{A}^{(n)})|(t\wedge\bar{\tau}_+^{(n)})>\delta\big)\bigg]\\
=&\frac{1}{n}\sum_{j\geq 1}\mathbb{E}\Big[(\xi_j^{(n)}-m^{(n)})^2; |\xi_j^{(n)}-m^{(n)}|>\delta, j\leq \bar{A}^{(n)}(T\wedge\bar{\tau}_+^{(n)})\Big]\\
=&\mathbb{E}\big[(\xi^{(n)}-m^{(n)})^2; \xi^{(n)}>\sqrt{n}\delta+m^{(n)}\big]\cdot\mathbb{E}\big[\bar{A}^{(n)}(T\wedge\bar{\tau}_+^{(n)})\big]\to 0,
\end{aligned}$$

where Fubini's theorem is applied in the second line, (6.16) and the condition (2.16) on the tail behavior of service times $\{\xi_j^{(n)}\}_{j\in\mathbb{Z}}$ in Assumption A2 are used in the third line. Thus, condition (3.23) in [26, Theorem VIII.3.11] holds for $\big(\hat{X}^{(n)}, \hat{U}_+^{(n)}(\bar{A}^{(n)})\big)$.

Next, it is easy to check that for every $\beta, \gamma \in \mathbb{R}$, the quadratic process

$$\begin{aligned}
\big[\beta\hat{X}^{(n)}+\gamma\hat{U}_+^{(n)}(\bar{A}^{(n)})\big](t) &= \frac{1}{n}\sum_{j\geq 1}\big(\beta+\gamma(\xi_j^{(n)}-m^{(n)})\big)^2\mathbf{1}(j\leq A^{(n)}(nt))\\
\to \big(\alpha^2+\gamma^2\sigma^2\big)\bar{A}(t)+\beta^2 t &= \big(\alpha^2+\gamma^2\sigma^2\big)\mu t+\beta^2 t \quad \text{u.o.c. in probability},
\end{aligned}$$

where the FLLN for triangular arrays and (4.1) are used. This proves that

$$\big(\hat{X}^{(n)}, \hat{S}_-^{(n)}, \hat{U}_+^{(n)}(\bar{A}^{(n)})\big) \Rightarrow \big(\hat{X}, \hat{S}_-, \hat{U}_+(\bar{A})\big) \quad \text{in} \quad (\mathbb{D}^3[0,T], J_1),$$

where $\hat{X}$, $\hat{S}_-$ and $\hat{U}_+$ are independent Brownian motions, and $(\hat{S}_-, \hat{U}_+)$ is the limit in Lemma 6.1.

For the last, it is easy to see that for all $t > 0$

$$0 \geq \hat{U}_+^{(n)}\big(\bar{S}_+^{(n)}(t)\big) + m^{(n)}\hat{S}_+^{(n)}(t) > \frac{-1}{\sqrt{n}}\xi^{(n)}_{S_+^{(n)}(nt)+1}.$$

Assumption A2-(2.16) can be used to prove the convergence $\hat{U}_+^{(n)}(\mu t) + \mu^{-1}\hat{S}_+^{(n)}(t) \to 0$ u.o.c. This finishes the proof. □

Applying Lemma 6.5 to the expressions in (6.32), we have the following joint convergence result.

Corollary 6.1. *Suppose Assumptions A1 and A2 and (4.1) hold for $\bar{q}_0 \in \mathcal{L}^=$. We have*

$$\begin{aligned}
&\big(\hat{X}^{(n)}(t), \hat{U}^{(n)}(\bar{q}_0^{(n)}+\bar{A}^{(n)}(t)), \hat{S}^{(n)}(\bar{B}^{(n)}(t))\big)\\
\Rightarrow&\Big(\hat{X}(t), -\mu^{-1}\big(\hat{S}_+(t)+\hat{S}_-(\bar{w}_0)\big), \hat{S}_+\big((t-\bar{w}_0)^+\big)+\hat{S}_-(t\wedge\bar{w}_0)\Big)\\
=&\big(\hat{X}(t), -\mu^{-1}\hat{S}(t+\bar{w}_0), \hat{S}(t)\big) \quad \textit{in } (\mathbb{D}^3[0,T], J_1),
\end{aligned}$$

where $\bar{w}_0 = \mu^{-1}\bar{q}_0 \geq 0$ and

$$\hat{S}(t) := \hat{S}_+\big((t-\bar{w}_0)^+\big) + \hat{S}_-(t\wedge\bar{w}_0), \quad \forall t\geq 0,$$

is a Brownian motion independent of $\hat{X}$ and has the same distribution as that in Lemma 6.1.

6.3.1. **The case $\bar{q}_0 = 0 \in \mathcal{L}^=$.** Recalling $\bar{I}^{(n)}$ in (6.6) and $\bar{Z}^{(n)}$ in (6.8), in addition to $(\hat{A}^{(n)}, \hat{Q}^{(n)}, \hat{W}^{(n)})$ in (4.2), we further define the diffusion-scaled processes

$$\big(\hat{Z}^{(n)}, \hat{I}^{(n)}\big) = \sqrt{n}\,\big(\bar{Z}^{(n)}, \bar{I}^{(n)}\big).$$

It is easy to check from (6.9) that

$$\begin{aligned} \hat{Z}^{(n)}(t) &= \hat{q}_0^{(n)} + \hat{A}^{(n)}(t) - \hat{\mu}^{(n)} t - \hat{S}^{(n)}\big(\bar{B}^{(n)}(t)\big), \\ \mu^{(n)} \hat{I}^{(n)}(t) &= \psi\big(\hat{Z}^{(n)}\big)(t) \quad \text{and} \quad \hat{Q}^{(n)}(t) = \phi\big(\hat{Z}^{(n)}\big)(t). \end{aligned} \tag{6.33}$$

Similar to the proof for FLLN, we use a localization technique in $\sqrt{n}$-scaled sense, for which we define

$$\hat{\tau}_+^{(n)} = \inf\{t > 0, \hat{Q}^{(n)}(t) > k_0\} \ \text{ and } \ \hat{J}^{(n)}(t) := \sup_{y \le k_0} \int_t^\infty h^{(n)}(u, \sqrt{n}y)\, \mathrm{d}u, \tag{6.34}$$

for arbitrary $k_0 > 0$, comparing with $\bar{\tau}_+^{(n)}$ in (6.13) and $\bar{J}^{(n)}$ in (6.14), where for notational brevity we have dropped the subscript k_0. In addition to (6.15), we make use of the following facts:

$$\begin{aligned} &\limsup_{n\to\infty} \sqrt{n}\, \hat{J}^{(n)}(n\varepsilon) \to 0 \ \text{ for every } \varepsilon > 0, \\ &\sup_{n\ge 1} \sup_{t>0} \sqrt{t}\, \hat{J}^{(n)}(t) < \infty \ \text{ and } \ \hat{J}^{(n)}(\infty) = 0 \ \text{ for every } n \in \mathbb{N}, \end{aligned} \tag{6.35}$$

under Assumption A4-(4.5) and (4.6) at $\bar{q}_0 = 0$.

Since $\lambda_0 = \mu(1 - H(0))$ at $\bar{q}_0 = 0 \in \mathcal{L}^=$, we have from (6.11) that

$$\begin{aligned} \hat{A}^{(n)}(t) &= \sqrt{n}\big(\bar{A}^{(n)}(t) - \bar{A}(t)\big) = \sqrt{n}\big(\bar{A}^{(n)}(t) - \lambda_0 t - H(0)\bar{A}(t)\big) \\ &= \hat{\lambda}_0^{(n)} t + \int_0^t \sqrt{n}\,\Big(H^{(n)}\big(\bar{Q}^{(n)}(s-)\big) - H(0)\Big)\, \mathrm{d}\bar{A}^{(n)}(s) \\ &\quad + H(0) \cdot \sqrt{n}\,\big(\bar{A}^{(n)}(t) - \bar{A}(t)\big) + \hat{X}^{(n)}(t) - \hat{\varepsilon}_1^{(n)}(t), \end{aligned}$$

where, recalling $\bar{\varepsilon}_1^{(n)}$ in (6.12), $\hat{\varepsilon}_1^{(n)}(t)$ is defined by

$$\hat{\varepsilon}_1^{(n)}(t) := \sqrt{n}\,\bar{\varepsilon}_1^{(n)}(t) = \sqrt{n} \int_0^t \Big(\int_{n(t-u)}^\infty h^{(n)}\big(v, \sqrt{n}y\big)\, \mathrm{d}v \Big)\Big|_{y = \hat{Q}^{(n)}(u-)} \mathrm{d}\bar{A}^{(n)}(u). \tag{6.36}$$

Recalling $\hat{H}_0^{(n)}$ in Assumption A4-(4.3), we obtain from the expression of $\hat{A}^{(n)}(t)$ above that

$$\hat{A}^{(n)}(t) = \frac{\hat{\lambda}_0^{(n)}\, t + \hat{X}^{(n)}(t)}{1 - H(0)} + \int_0^t \frac{\hat{H}_0(\hat{Q}^{(n)}(s-))}{1 - H(0)}\, \mathrm{d}\bar{A}^{(n)}(s) + \frac{\hat{\varepsilon}_2^{(n)}(t) - \hat{\varepsilon}_1^{(n)}(t)}{1 - H(0)}, \tag{6.37}$$

where

$$\hat{\varepsilon}_2^{(n)}(t) := \int_0^t \Big(\hat{H}_0^{(n)}(y) - \hat{H}_0(y)\Big)\Big|_{y = \hat{Q}^{(n)}(u-)} \mathrm{d}\bar{A}^{(n)}(u). \tag{6.38}$$

We show first that the residual terms above are negligible for large n.

Lemma 6.6. *Under Assumptions A1, A2 and A4,*

$$\hat{\varepsilon}_1^{(n)}(t \wedge \hat{\tau}_+^{(n)}) \to 0 \quad \textit{and} \quad \hat{\varepsilon}_2^{(n)}(t \wedge \hat{\tau}_+^{(n)}) \to 0 \quad \textit{u.o.c. in probability.}$$

Proof. For $\hat{\varepsilon}_2^{(n)}$ in (6.38), we can check directly that

$$\big|\hat{\varepsilon}_2^{(n)}\big|(t \wedge \hat{\tau}_+^{(n)}) \le \sup_{y \le k_0} \big|\hat{H}_0^{(n)}(y) - \hat{H}_0(y)\big| \cdot \bar{A}^{(n)}(t \wedge \hat{\tau}_+^{(n)}) \to 0 \quad \text{u.o.c. in probability.} \tag{6.39}$$

where Assumption A4-(4.3) and the stochastic boundedness of $\bar{A}^{(n)}(t \wedge \hat{\tau}_+^{(n)})$ from (6.17) are applied.

For $\hat{\varepsilon}_1^{(n)}$ in (6.38) on the set $\{t \le \hat{\tau}_+^{(n)}\}$, we have from (6.34),

$$\begin{aligned}\hat{\varepsilon}_1^{(n)}(t) &= \sqrt{n}\int_0^t \mathrm{d}\bar{A}^{(n)}(s)\int_{n(t-s)}^{\infty} h^{(n)}(u,\sqrt{n}y)\,\mathrm{d}u\Big|_{y=\hat{Q}^{(n)}(s-)} \\ &\le \sqrt{n}\int_0^t \hat{J}^{(n)}\big(n(t-u)\big)\,\mathrm{d}\bar{A}^{(n)}(u) =: \check{\varepsilon}_1^{(n)}(t).\end{aligned} \tag{6.40}$$

By the identity $\bar{A}^{(n)} = \bar{X}^{(n)} + \bar{\Lambda}^{(n)}$ and the definition of $\lambda^{(n)}$ from (2.12), we further have

$$\begin{aligned}\check{\varepsilon}_1^{(n)}(t) &= \int_0^t \hat{J}^{(n)}\big(n(t-u)\big)\big(\mathrm{d}\hat{X}^{(n)}(u) + \sqrt{n}\lambda^{(n)}(nu)\,\mathrm{d}u\big) \\ &= \int_0^t \hat{J}^{(n)}\big(n(t-u)\big)\,\mathrm{d}\hat{X}^{(n)}(u) + \lambda_0^{(n)}\cdot\sqrt{n}\int_0^t \hat{J}^{(n)}\big(n(t-u)\big)\,\mathrm{d}u \\ &\quad + \sqrt{n}\int_0^t \mathrm{d}\bar{A}^{(n)}(v)\int_0^{n(t-v)} \hat{J}^{(n)}\big(n(t-v)-u\big)h^{(n)}(u,\sqrt{n}y)\,\mathrm{d}u\Big|_{y=\hat{Q}^{(n)}(v-)} \\ &= \check{\varepsilon}_{1,1}^{(n)}(t) + \lambda_0^{(n)}\cdot\check{\varepsilon}_{1,2}^{(n)}(t) + \check{\varepsilon}_{1,3}^{(n)}(t).\end{aligned} \tag{6.41}$$

Notice by definition that $\hat{J}^{(n)}$ in (6.34) is a decreasing function on $[0,\infty)$ with $\hat{J}^{(n)}(0) < 1$ and $\hat{J}^{(n)}(\infty) = 0$ as shown in (6.35), so we can denote $(-\hat{J}^{(n)})(dv)$ as the associated Lebesgue-Stieltjes measure on $\mathbb{R}_+$. Applying Fubini's theorem, we obtain that for every $t > 0$ and $y \ge 0$,

$$\begin{aligned}&\int_0^t \hat{J}^{(n)}\big(t-u\big)h^{(n)}(u,y)\,\mathrm{d}u \\ &= \int_0^t\int_{t-u}^{\infty}(-\hat{J}^{(n)})(\mathrm{d}v)h^{(n)}(u,y)\,\mathrm{d}u = \int_0^{\infty}(-\hat{J}^{(n)})(\mathrm{d}v)\int_{t-v}^t h^{(n)}(u,y)\,\mathrm{d}u \\ &= \int_0^t(-\hat{J}^{(n)})(\mathrm{d}v)\Big(\int_{t-v}^{\infty}h^{(n)}(u,y)\,\mathrm{d}u - \int_t^{\infty}h^{(n)}(u,y)\,\mathrm{d}u\Big) + \hat{J}^{(n)}(t)\int_0^t h^{(n)}(u,y)\,\mathrm{d}u \\ &= \int_0^t(-\hat{J}^{(n)})(\mathrm{d}v)\int_{t-v}^{\infty}h^{(n)}(u,y)\,\mathrm{d}u - \hat{J}^{(n)}(0)\int_t^{\infty}h^{(n)}(u,y)\,\mathrm{d}u + \hat{J}^{(n)}(t)\int_0^{\infty}h^{(n)}(u,y)\,du.\end{aligned}$$

where we understand that $h^{(n)}(u,y) = 0$ for $u < 0$. Substituting into $\check{\varepsilon}_{1,3}^{(n)}$ in (6.41), and recalling the identities for $\hat{\varepsilon}_1^{(n)}$ and $\check{\varepsilon}_1^{(n)}$ in (6.40), one can find that

$$\begin{aligned}\check{\varepsilon}_{1,3}^{(n)}(t) &= \int_0^t \hat{\varepsilon}_1^{(n)}(t-v)(-\hat{J}^{(n)})(n\,\mathrm{d}v) - \hat{J}^{(n)}(0)\cdot\hat{\varepsilon}_1^{(n)}(t) \\ &\quad + \sqrt{n}\int_0^t \hat{J}^{(n)}\big(n(t-v)\big)H^{(n)}\big(\bar{Q}^{(n)}(v-)\big)\,\mathrm{d}\bar{A}^{(n)}(v) \\ &\le \int_0^t \hat{\varepsilon}_1^{(n)}(t-v)(-\hat{J}^{(n)})(n\,\mathrm{d}v) - \hat{J}^{(n)}(0)\cdot\hat{\varepsilon}_1^{(n)}(t) + \alpha\cdot\check{\varepsilon}_1^{(n)}(t),\end{aligned}$$

where the bound $H^{(n)}(\bar{Q}^{(n)}(v-)) \le \alpha$ from (6.15) is applied in the last term above. Plugging the inequality above into (6.41) and eliminating $\alpha\cdot\check{\varepsilon}_1^{(n)}$ on both sides gives

$$(1-\alpha)\cdot\check{\varepsilon}_1^{(n)}(t) \le \check{\varepsilon}_{1,1}^{(n)}(t) + \lambda_0^{(n)}\cdot\check{\varepsilon}_{1,2}^{(n)}(t) + \int_0^t \hat{\varepsilon}_1^{(n)}(t-v)(-\hat{J}^{(n)})(n\,\mathrm{d}v) - \hat{J}^{(n)}(0)\cdot\hat{\varepsilon}_1^{(n)}(t),$$

and further applying to the fact $\hat{\varepsilon}_1^{(n)} \le \check{\varepsilon}_1^{(n)}$ from (6.40) gives for every $t \le \hat{\tau}_+^{(n)}$,

$$\big(1-\alpha+\hat{J}^{(n)}(0)\big)\cdot\hat{\varepsilon}_1^{(n)}(t) \le \check{\varepsilon}_{1,1}^{(n)}(t) + \lambda_0^{(n)}\cdot\check{\varepsilon}_{1,2}^{(n)}(t) + \int_0^t \hat{\varepsilon}_1^{(n)}(t-v)(-\hat{J}^{(n)})(n\,\mathrm{d}v).$$

In particular, taking the supremum over $t \in [0, T \wedge \hat{\tau}_+^{(n)}]$ above gives

$$(1-\alpha) \cdot \sup_{t \le T} \hat{\varepsilon}_1^{(n)}(t \wedge \hat{\tau}_+^{(n)}) \le \sup_{t \le T} |\check{\varepsilon}_{1,1}^{(n)}|(t \wedge \hat{\tau}_+^{(n)}) + \lambda_0^{(n)} \cdot \sup_{t \le T} \check{\varepsilon}_{1,2}^{(n)}(t \wedge \hat{\tau}_+^{(n)}). \tag{6.42}$$

For the last, by integration by parts, one can find similar to (6.26) that

$$\check{\varepsilon}_{1,1}^{(n)}(t \wedge \hat{\tau}_+^{(n)}) = \int_0^\infty \Big(\hat{X}^{(n)}\big(t \wedge \hat{\tau}_+^{(n)}\big) - \hat{X}^{(n)}\big((t-u) \wedge \hat{\tau}_+^{(n)}\big) \Big) (-\hat{J}^{(n)})(n \, \mathrm{d}u)$$

and thus, for every $\delta > 0$ and $t \le T$,

$$\big|\check{\varepsilon}_{1,1}^{(n)}\big|(t \wedge \hat{\tau}_+^{(n)}) \le w_\delta\big(\hat{X}^{(n)}(\cdot \wedge \hat{\tau}_+^{(n)}), T\big) + 2 \cdot \sup_{t \le T} \big|\hat{X}^{(n)}\big|(t \wedge \hat{\tau}_+^{(n)}) \cdot \hat{J}^{(n)}(n\delta),$$

recalling the modulus of continuity in (6.18). By change of variables, for every $t \le T$,

$$\begin{aligned} \check{\varepsilon}_{1,2}^{(n)}(t \wedge \hat{\tau}_+^{(n)}) &\le \sqrt{n} \int_0^t \hat{J}^{(n)}(nu) \, \mathrm{d}u \le \sqrt{n} \int_0^T \hat{J}^{(n)}(nu) \, \mathrm{d}u \\ &\le \sup_{t>0} \sqrt{t} \, \hat{J}^{(n)}(t) \cdot \int_0^\delta u^{-1/2} \, \mathrm{d}u + \sqrt{n} \, \hat{J}^{(n)}(n\delta) \cdot T. \end{aligned}$$

Letting $n \to \infty$ and then $\delta \to 0+$, we can conclude from the facts in (6.35) and the conclusion for $\hat{X}^{(n)}$ in Lemma 6.5 that

$$\check{\varepsilon}_{1,1}^{(n)}(t \wedge \hat{\tau}_+^{(n)}) \to 0 \quad \text{and} \quad \check{\varepsilon}_{1,2}^{(n)}(t \wedge \hat{\tau}_+^{(n)}) \to 0 \quad \text{u.o.c. in probability.}$$

The claim for $\hat{\varepsilon}_1^{(n)}$ is proved by (6.42) and this finishes the proof. □

Now, we are ready to prove the FCLT, where the SDE for $\hat{Z}^{(n)}$ in (6.33) is analyzed and the following version of [33, Theorem 5.4] is applied.

Proposition 6.2. *Let ϕ and ψ be the operators in Definition 6.1 and f be a continuous function on $\mathbb{R}_+$. Suppose that $(x_n, y_n, z_n, u_n, \varepsilon_n)$ is adapted to $\{\mathscr{F}_n(t)\}_{t \ge 0}$ and satisfies*

$$\begin{aligned} x_n(t) &= u_n(t) + \int_0^t f\big(z_n(s-)\big) \, \mathrm{d}y_n(s) + \varepsilon_n(t), \\ z_n(t) &= \phi(x_n)(t) = x_n(t) + l_n(t) \quad \text{and} \quad l_n(t) = \psi(x_n)(t). \end{aligned} \tag{6.43}$$

(1) $(y_n, u_n) \Rightarrow (y, u)$ in the Skorohod topology and $y_n \in \mathbb{D}$ is an increasing process such that

$$\lim_{b \to \infty} \sup_{n \ge 1} \mathbb{P}\big(y_n(t) > b\big) = 0 \quad \forall t > 0;$$

(2) for every $b > 0$, let $\tau_{n,b} := \inf\{t > 0 : |z_n|(t-) \wedge |z_n|(t) \ge b\}$, then as $n \to \infty$

$$\varepsilon_n\big(t \wedge \tau_{n,b}\big) \to 0 \quad \textit{u.o.c. in probability;}$$

(3) the equation

$$z(t) = u(t) + \int_0^t f\big(z(s-)\big) \, \mathrm{d}y(s) + l(t), \tag{6.44}$$

has a global solution and the weak local uniqueness holds, where l is the minimal increasing process such that $z \ge 0$ and $\displaystyle\int_0^\infty \mathbf{1}\big(z(s) > 0\big) \, \mathrm{d}l(s) = 0$.

Then $(x_n, y_n, z_n, u_n, l_n) \Rightarrow (x, y, z, u, l)$ in the Skorohod J_1 topology, where $l = \psi(x)$ and $z = x + l$.

Remark 6.4. *Recall the following terminologies from the discussion before* [33, Theorem 5.4] *for solution to* (6.44)*. Firstly, a set of processes* (u, y, z, l, τ) *is called a weak local solution of* (6.44)*, if there is a filtered probability space* $\big(\Omega, \mathscr{F}, \{\mathscr{F}(t)\}_{t\geq 0}, \mathbb{P}\big)$ *such that* u, y, z, l *are adapted,* τ *is an* $\{\mathscr{F}(t)\}_{t\geq 0}$*-stopping time,* (u, y) *has the desired distribution in condition* (1) *and*

$$z(t\wedge\tau) = u(t\wedge\tau) + \int_0^{t\wedge\tau} f\big(z(s-)\big)\,\mathrm{d}y(s) + l(t\wedge\tau).$$

Secondly, we say that the weak local uniqueness holds for (6.44)*, if any two weak local solutions* $\big(u_1, y_1, z_1, l_1, \tau_1\big)$ *and* $\big(u_2, y_2, z_2, l_2, \tau_3\big)$ *with* $\tau_1 = h_1(z_1)$ *and* $\tau_2 = h_2(z_2)$ *for measurable functions* h_1, h_2 *on* $\mathbb{D}[0,\infty)$*,* $\big(u_1, y_1, z_1, l_1, h_1\wedge h_2(z_1)\big)$ *and* $\big(u_2, y_2, z_2, l_2, h_1\wedge h_2(z_2)\big)$ *have the same distribution. For the last,* (u, y, z, l) *is called a global solution, if it satisfies* (6.44) *for all* $t > 0$*. Basically, a weak sense of solution is a set of adapted processes so that* (6.44) *holds. The localizing stopping time is usually taken as the first passage time of the process* z*. The weak uniqueness is referred to the coincidence of laws of solutions on the space of functions. A global solution exists if there is a non-explosive solution or equivalently, a locally stochastic bounded solution. For the problem of weak convergence in Skorohod* J_1 *topology, a solution in weak sense would be enough.*

As far as our paper is concerned, (6.44) *takes the form of the SDE with reflection* (4.11) *in Remark 4.2. By the assumption of the continuity of drift term,* [54, Theorem 3.1] *ensures the existence and uniqueness of a strong local solution, c.f.* [25, IV.1]*. The linear growth condition is used so that the solution is stochastically bounded and thus non-explosive, c.f. Remark 4.1.*

Remark 6.5. *Proposition 6.2 is a modification of* [33, Theorem 5.4] *in the following sense.*

(1) Comparing (6.43) *with* [33, eqn (5.2)]*, one can find that*

$$x_n(t) = u_n(t) + \int_0^t f\big(\phi(x_n)(s-)\big)\,\mathrm{d}y_n(s) + \varepsilon_n(t),$$

where for the notations in their paper

$$F_n(x,t) = F(x,t) = f\big(\phi(x)(t)\big) \quad \forall x\in\mathbb{D} \text{ and } t>0.$$

Let $T_1(\mathbb{R}_+)$ *denote the collection of nondecreasing and* 1*-Lipschitz mappings of* $\mathbb{R}_+$ *onto* $\mathbb{R}_+$ *in that paper. It is straightforward to check that for every* $\theta\in T_1(\mathbb{R}_+)$*,*

$$(F(x)\circ\theta)(t) = f\big(\phi(x)(\theta(t))\big) = f\big(\phi(x\circ\theta)(t)\big) = F(x\circ\theta)(t).$$

Therefore, $G(x,\theta) = G(x) = F(x)$ *for the mapping on* $\mathbb{D}(\mathbb{R}_+)\times T_1(\mathbb{R}_+)$ *in* [33]*, which is indeed continuous under the uniform topology on* $\mathbb{D}[0,T]$*, so that condition C5.4 in that paper is satisfied.*

(2) The integrator y_n *is an increasing process so that it is a semi-martingale and Condition C.2.2.(i) in that paper is simplified to its stochastic boundedness in condition* (1)*.*

(3) There is an additional term ε_n *in the integral equation* (6.43) *besides* (u_n, y_n)*, which is assumed to be negligible for large* n *in the stopped version in condition* (2)*. Therefore, the limit*

$$\limsup_{n\to\infty} \mathbb{P}\big(\tau_{n,b} > t\big) = 0 \quad \forall t>0, \tag{6.45}$$

is necessary for the weak convergence of $(x_n, y_n, z_n, u_n, l_n)$ *without stopping.*

(4) The existence of a weak and locally stochastic bounded solution for (6.44)*, as well as its locally uniqueness, is needed for the convergence in Skorohod* J_1 *topology (*[33, Theorem 5.4]*). Here for our SDE, the existence and uniqueness of a strong solution is ensured discussed in Remark 4.2.*

Proof of Proposition 6.2. For every $b > 0$, let $x_{n,b}$ denote the solution of

$$x_{n,b}(t) = u_n(t) + \int_0^t \mathbf{1}(s<\tau_{n,b}) f\big(\phi\big(x_{n,b}\big)(s-)\big)\,\mathrm{d}y_n(s) + \varepsilon_n(t\wedge\tau_{n,b}),$$

that agrees with x_n on $[0,\tau_{n,b})$. The first conclusion of [33, Theorem 5.4] shows that $\{(x_{n,b},y_n,u_n)\}_{n\geq 1}$ is relatively compact under conditions (1) and (2) in Proposition 6.2, and any limit point (x_b,y,u) gives a local solution (x_b,τ) of

$$x(t)=u(t)+\int_0^t f\big(\phi(x)(s)\big)\,\mathrm{d}y(s), \tag{6.46}$$

with $\tau=\tau_c:=\inf\{t>0:|\phi(x)|(t-)\vee|\phi(x)|(t)\geq c\}$ for any $c<b$. Moreover, by the continuities of ψ and ϕ, letting $z_{n,b}=\phi(x_{n,b})$ and $l_{n,b}=\psi(x_{n,b})$, $\{(x_{n,b},y_n,z_{n,b},u_n,l_{n,b})\}_{n\geq 1}$ is also relatively compact on $(\mathbb{D}^5[0,T],J_1)$, and hence, the weakly convergence holds for the same subsequence. The associated limit point, say (x_b,y,z_b,u,l_b) satisfies on $[0,\tau]$,

$$x_b(t)=u(t)+\int_0^t f\big(z_b(s)\big)\,\mathrm{d}y(s)\quad\text{and}\quad(l_b,z_b)=\big(\psi(x_b),\phi(x_b)\big),$$

where l_b is the minimal increasing process such that $z_b\geq 0$, that is,

$$\int_0^\infty \mathbf{1}\big(z_b(s)>0\big)\,\mathrm{d}l_b(s)=0.$$

Given the existence and uniqueness of weak solution to (6.44), say (z,l), one can find that

$$x:=z-l=u(t)+\int_0^t f\big(z(s)\big)\,\mathrm{d}y(s)=u(t)+\int_0^t f\big(\phi(x)(s)\big)\,\mathrm{d}y(s),$$

is well-defined, and the weak uniqueness holds. Therefore, the joint convergence in [33, Theorem 5.4] can be applied to show the convergence of $\{(x_n,y_n,z_n,u_n,l_n)\}_{n\geq 1}$ in the Skorohod topology and this finishes the proof. □

Proof of Theorem 4.1 (the case $\bar{q}_0=0$). Plugging (6.37) into the expression of $\hat{Z}^{(n)}$ in (6.33), we obtain

$$\begin{aligned}\hat{Z}^{(n)}(t)&=\hat{\Xi}^{(n)}(t)+\int_0^t\frac{\hat{H}_0\big(\hat{Q}^{(n)}(s-)\big)}{1-H(0)}\,\mathrm{d}\bar{A}^{(n)}(s)+\hat{\varepsilon}^{(n)}(t),\\ \hat{Q}^{(n)}(t)&=\phi\big(\hat{Z}^{(n)}\big)(t)=\hat{Z}^{(n)}(t)+\mu^{(n)}\hat{I}^{(n)}(t)\quad\text{and}\quad\mu^{(n)}\hat{I}^{(n)}(t)=\psi\big(\hat{Z}^{(n)}\big)(t),\end{aligned} \tag{6.47}$$

comparing with (6.43), where

$$\hat{\Xi}^{(n)}(t):=\hat{q}_0^{(n)}-\hat{\mu}^{(n)}t-\hat{S}^{(n)}\big(\bar{B}^{(n)}(t)\big)+\frac{\hat{\lambda}_0^{(n)}\,t+\hat{X}^{(n)}(t)}{1-H(0)}\quad\text{and}\quad\hat{\varepsilon}^{(n)}(t):=\frac{\hat{\varepsilon}_2^{(n)}(t)-\hat{\varepsilon}_1^{(n)}(t)}{1-H(0)}.$$

For every fixed $k_0>0$, applying the joint convergence result in Corollary 6.1, together with the fact that $(\bar{A}^{(n)},\bar{B}^{(n)})$ converges to a deterministic limit in (4.1), we obtain

$$\big(\bar{A}^{(n)},\hat{X}^{(n)},\hat{S}^{(n)}(\bar{B}^{(n)}),\hat{\Xi}^{(n)}\big)(t)\Rightarrow\big(\bar{A},\hat{X},\hat{S},\hat{\Xi}\big)(t)\quad\text{in}\quad(\mathbb{D}^4[0,T],J_1),$$

where

$$\hat{\Xi}(t)=\hat{q}_0-\hat{\mu}\,t-\hat{S}(t)+\frac{\hat{\lambda}_0\,t+\hat{X}(t)}{1-H(0)}.$$

Moreover, we can check that (6.17) ensures the condition (1) in Proposition 6.2, and Lemma 6.6 ensures the condition (2) in Proposition 6.2. It follows directly from (6.47) that the possible limit of $\big(\bar{A}^{(n)},\hat{X}^{(n)},\hat{S}^{(n)}(\bar{B}^{(n)}),\hat{\Xi}^{(n)},\hat{Z}^{(n)},\hat{Q}^{(n)},\hat{I}^{(n)}\big)$ satisfies

$$\hat{Z}(t)=\int_0^t\frac{\hat{H}_0(\hat{Q}(s))}{1-H(0)}\,\mathrm{d}\bar{A}(s)+\hat{\Xi}(t)\quad\text{and}\quad\big(\mu\hat{I},\hat{Q}\big)=\big(\psi(\hat{Z}),\phi(\hat{Z})\big).$$

It can be checked directly that the SDE for $\hat{Q}$ above is exactly (4.11), where the existence and uniqueness of a strong solution are discussed in Remark 4.2. Therefore, Proposition 6.2 gives

$$\big(\bar{A}^{(n)},\hat{X}^{(n)},\hat{S}^{(n)}(\bar{B}^{(n)}),\hat{Z}^{(n)},\hat{\Xi}^{(n)},\hat{Q}^{(n)},\hat{I}^{(n)}\big)\Rightarrow\big(\bar{A},\hat{X},\hat{S},\hat{Z},\hat{\Xi},\hat{Q},\hat{I}\big)\quad\text{in}\quad(\mathbb{D}^7[0,T],J_1). \tag{6.48}$$

We next obtain from (6.33) that

$$\hat{A}^{(n)}(t) = -\hat{q}_0^{(n)} + \hat{Z}^{(n)}(t) + \hat{\mu}^{(n)}t + \hat{S}^{(n)}\big(\bar{B}^{(n)}(t)\big),$$

and from (4.2) with $\bar{q}_0 = 0$, (6.7) and (6.10) that

$$\begin{aligned}\hat{W}^{(n)}(t) &= m^{(n)}\hat{Q}^{(n)}(t) + \sqrt{n}\big(\bar{W}^{(n)} - m^{(n)}\bar{Q}^{(n)}\big)(t)\\ &= m^{(n)}\hat{Q}^{(n)}(t) + \hat{U}^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t)\big) + m^{(n)}\hat{S}^{(n)}\big(\bar{B}^{(n)}(t)\big).\end{aligned}$$

By the joint convergence results in (6.48) and Corollary 6.1, we obtain the joint convergence for $(\hat{A}^{(n)}, \hat{Q}^{(n)}, \hat{W}^{(n)})$ from the above expressions.

For the last, given the joint convergence above and the fact $\bar{q}_0 = 0$, the SDE for $(\hat{A}, \hat{Q})$ in (4.7) follows directly from (6.37) and (6.33). This finishes the proof. □

6.3.2. **The case $\bar{q}_0 \in \mathcal{L}^=$ and $\bar{q}_0 > 0$.** The proof for the equilibrium point $\bar{q}_0 > 0$ is similar to that when $\bar{q}_0 = 0$, so we mainly highlight the differences below.

Firstly, note that $\hat{Q}_0^{(n)}$ takes value in $\mathbb{R}$ instead of $\mathbb{R}_+$ in this case. Recalling $(\hat{A}^{(n)}, \hat{Q}^{(n)}, \hat{W}^{(n)})$ in (4.2), it is expected that $\bar{Q}^{(n)}$ stays around $\bar{q}_0$ of order $n^{-1/2}$, so that the CLT-scaled regulator of reflection mapping is expected to be negligible for n large enough. More precisely, for $\bar{I}^{(n)}$ in (6.6) and $\bar{Z}^{(n)}$ in (6.8), by defining

$$\big(\hat{Z}^{(n)}(t), \hat{I}^{(n)}(t)\big) = \sqrt{n}\,\big(\bar{Z}^{(n)}(t) - \bar{q}_0, \bar{I}^{(n)}(t)\big),$$

similar to (6.33), we obtain that

$$\hat{Q}^{(n)}(t) = \hat{Z}^{(n)}(t) + \mu^{(n)}\,\hat{I}^{(n)}(t) \quad \text{and} \quad \hat{Z}^{(n)}(t) = \hat{q}_0^{(n)} + \hat{A}^{(n)}(t) - \hat{\mu}^{(n)}t - \hat{S}^{(n)}\big(\bar{B}^{(n)}(t)\big). \tag{6.49}$$

However, $\hat{I}^{(n)}$ is no longer the regulator of reflection for $\hat{Z}^{(n)}$, that is, by (6.9),

$$\mu^{(n)}\,\hat{I}^{(n)} = \mu^{(n)}\sqrt{n}\,\bar{I}^{(n)} = \sqrt{n}\,\psi\big(\bar{Z}^{(n)}\big) = \psi\big(\hat{Z}^{(n)} + \sqrt{n}\,\bar{q}_0\big) \neq \psi(\hat{Z}^{(n)}).$$

With abuse of notations, fixing arbitrary $k_0 > 0$, define in $\sqrt{n}$-scaled sense,

$$\hat{\tau}_+^{(n)} = \inf\big\{t > 0 : |\hat{Q}^{(n)}(t)| > k_0\big\} \quad \text{and} \quad \hat{J}^{(n)}_{(\bar{q}_0,k_0)}(t) := \sup_{|y|\le k_0}\int_t^\infty h^{(n)}\big(u, n\bar{q}_0 + \sqrt{n}y\big)\,\mathrm{d}u, \tag{6.50}$$

comparing with (6.34), where we drop the subscript k_0 and $\bar{q}_0$ for notational brevity. In this case, we make use of the following facts in addition to (6.15),

$$\begin{aligned}&\limsup_{n\to\infty}\sqrt{n}\,\hat{J}^{(n)}(n\varepsilon) \to 0 \ \text{ for every } \varepsilon > 0,\\ &\sup_{n\ge1}\sup_{t>0}\sqrt{t}\,\hat{J}^{(n)}(t) < \infty \ \text{ and } \ \hat{J}^{(n)}(\infty) = 0 \ \text{ for every } n \in \mathbb{N},\end{aligned} \tag{6.51}$$

under Assumption A4-(4.5) and (4.6), which only differs from (6.35) with omitted $\bar{q}_0$.

Since $\lambda_0 = \mu(1 - H(\bar{q}_0))$ at $\bar{q}_0 \in \mathcal{L}^=$, we can rewrite from (6.11),

$$\begin{aligned}\hat{A}^{(n)}(t) &= \sqrt{n}\big(\bar{A}^{(n)}(t) - \bar{A}(t)\big) = \sqrt{n}\big(\bar{A}^{(n)}(t) - \lambda_0 t - H(\bar{q}_0)\bar{A}(t)\big)\\ &= \hat{X}^{(n)}(t) + \hat{\lambda}_0^{(n)}t + \int_0^t \hat{H}^{(n)}_{\bar{q}_0}\big(\bar{Q}^{(n)}(s-)\big)\,\mathrm{d}\bar{A}^{(n)}(u) + H(\bar{q}_0)\cdot\hat{A}^{(n)}(t) - \hat{\varepsilon}_1^{(n)}(t)\\ &= \hat{\lambda}_0^{(n)}t + \int_0^t \hat{H}_{\bar{q}_0}\big(\hat{Q}^{(n)}(s-)\big)\,\mathrm{d}\bar{A}^{(n)}(s) + H(\bar{q}_0)\cdot\hat{A}^{(n)}(t) + \hat{X}^{(n)}(t) - \hat{\varepsilon}_1^{(n)}(t) + \hat{\varepsilon}_2^{(n)}(t)\end{aligned}$$

where by the fact $\bar{Q}^{(n)} = \bar{q}_0 + \hat{Q}^{(n)}/\sqrt{n}$,

$$\begin{aligned}
\hat{\varepsilon}_1^{(n)}(t) &:= \sqrt{n}\,\bar{\varepsilon}_1^{(n)}(t) = \sqrt{n}\int_0^t \Big(\int_{n(t-u)}^{\infty} h^{(n)}\big(v, n\bar{q}_0 + \sqrt{n}y\big)\,\mathrm{d}v\Big)\Big|_{y=\hat{Q}^{(n)}(u-)}\mathrm{d}\bar{A}^{(n)}(u),\\
\hat{\varepsilon}_2^{(n)}(t) &:= \int_0^t \Big(\hat{H}_{\bar{q}_0}^{(n)}(y) - \hat{H}_{\bar{q}_0}(y)\Big)\Big|_{y=\hat{Q}^{(n)}(u-)}\mathrm{d}\bar{A}^{(n)}(u),
\end{aligned} \tag{6.52}$$

comparing with (6.38), which is equivalent to, similar to (6.37),

$$\hat{A}^{(n)}(t) = \frac{\hat{\lambda}_0^{(n)}\,t + \hat{X}^{(n)}(t)}{1 - H(\bar{q}_0)} + \int_0^t \frac{\hat{H}_{\bar{q}_0}(\hat{Q}^{(n)}(s-))}{1 - H(\bar{q}_0)}\,\mathrm{d}\bar{A}^{(n)}(s) + \frac{\hat{\varepsilon}_2^{(n)}(t) - \hat{\varepsilon}_1^{(n)}(t)}{1 - H(\bar{q}_0)}. \tag{6.53}$$

We show first, analogous to Lemma 6.6, that the residual terms above is negligible for large n.

Lemma 6.7. *Under Assumptions A1, A2 and A4, we have*

$$\hat{\varepsilon}_1^{(n)}(t \wedge \hat{\tau}_+^{(n)}) \to 0 \quad \textit{and} \quad \hat{\varepsilon}_2^{(n)}(t \wedge \hat{\tau}_+^{(n)}) \to 0 \quad \textit{u.o.c. in probability.}$$

Proof. For $\hat{\varepsilon}_2^{(n)}$ in (6.52), we directly obtain that

$$\big|\hat{\varepsilon}_2^{(n)}\big|(t \wedge \hat{\tau}_+^{(n)}) \le \sup_{y \le k_0}\big|\hat{H}_{\bar{q}_0}^{(n)}(y) - \hat{H}_{\bar{q}_0}(y)\big| \cdot \bar{A}^{(n)}(t \wedge \hat{\tau}_+^{(n)}).$$

By Assumption A4-(4.3) and the stochastic boundedness of $\bar{A}^{(n)}(t \wedge \hat{\tau}_+^{(n)})$ from (6.17), we can prove the convergence of $\hat{\varepsilon}_2^{(n)}$ to zero u.o.c. in probability.

For $\hat{\varepsilon}_1^{(n)}$ in (6.52) on the set $\{t \le \hat{\tau}_+^{(n)}\}$, we have from (6.50),

$$\begin{aligned}
\hat{\varepsilon}_1^{(n)}(t) &= \sqrt{n}\int_0^t \mathrm{d}\bar{A}^{(n)}(s)\int_{n(t-s)}^{\infty} h^{(n)}(u, n\bar{q}_0 + \sqrt{n}y)\,\mathrm{d}u\Big|_{y=\hat{Q}^{(n)}(s-)}\\
&\le \sqrt{n}\int_0^t \hat{J}^{(n)}\big(n(t-u)\big)\,\mathrm{d}\bar{A}^{(n)}(u) =: \check{\varepsilon}_1^{(n)}(t).
\end{aligned} \tag{6.54}$$

By the fact $\bar{A}^{(n)} = \bar{X}^{(n)} + \bar{\Lambda}^{(n)}$ and the definition of $\bar{\Lambda}^{(n)}$ from (2.12), we further have

$$\begin{aligned}
\check{\varepsilon}_1^{(n)}(t) &= \int_0^t \hat{J}^{(n)}\big(n(t-u)\big)\big(\mathrm{d}\hat{X}^{(n)}(u) + \sqrt{n}\lambda^{(n)}(nu)\,\mathrm{d}u\big)\\
&= \int_0^t \hat{J}^{(n)}\big(n(t-u)\big)\,\mathrm{d}\hat{X}^{(n)}(u) + \lambda_0^{(n)}\cdot\sqrt{n}\int_0^t \hat{J}^{(n)}\big(n(t-u)\big)\,\mathrm{d}u\\
&\quad + \sqrt{n}\int_0^t \mathrm{d}\bar{A}^{(n)}(v)\int_0^{n(t-v)} \hat{J}^{(n)}\big(n(t-v)-u\big)h^{(n)}(u, n\bar{q}_0 + \sqrt{n}y)\,\mathrm{d}u\Big|_{y=\hat{Q}^{(n)}(v-)}\\
&= \check{\varepsilon}_{1,1}^{(n)}(t) + \lambda_0^{(n)}\cdot\check{\varepsilon}_{1,2}^{(n)}(t) + \check{\varepsilon}_{1,3}^{(n)}(t).
\end{aligned} \tag{6.55}$$

Noticing that $\hat{J}^{(n)}$ in (6.50) and (6.51) is also a decreasing function on $[0,\infty)$ with $\hat{J}^{(n)}(0) < 1$ and $\hat{J}^{(n)}(\infty) = 0$, we can also check that

$$\begin{aligned}
&\int_0^t \hat{J}^{(n)}\big(t-u\big)h^{(n)}(u,y)\,\mathrm{d}u\\
&= \int_0^t (-\hat{J}^{(n)})(\mathrm{d}v)\int_{t-v}^{\infty} h^{(n)}(u,y)\,\mathrm{d}u - \hat{J}^{(n)}(0)\int_t^{\infty} h^{(n)}(u,y)\,\mathrm{d}u + \hat{J}^{(n)}(t)\int_0^{\infty} h^{(n)}(u,y)\,du.
\end{aligned}$$

Substituting into $\check{\varepsilon}_{1,3}^{(n)}$ in (6.41), recalling (6.15), $\hat{\varepsilon}_1^{(n)}$ and $\check{\varepsilon}_1^{(n)}$ in (6.40), we also have

$$\check{\varepsilon}_{1,3}^{(n)}(t) \le \int_0^t \hat{\varepsilon}_1^{(n)}(t-v)(-\hat{J}^{(n)})(n\,\mathrm{d}v) - \hat{J}^{(n)}(0)\cdot\hat{\varepsilon}_1^{(n)}(t) + \alpha\cdot\check{\varepsilon}_1^{(n)}(t).$$

Plugging the inequality above into (6.55) and eliminating $\alpha\cdot\check{\varepsilon}_1^{(n)}$ on both sides gives

$$(1-\alpha)\cdot\check{\varepsilon}_1^{(n)}(t)\le\check{\varepsilon}_{1,1}^{(n)}(t)+\lambda_0^{(n)}\cdot\check{\varepsilon}_{1,2}^{(n)}(t)+\int_0^t\hat{\varepsilon}_1^{(n)}(t-v)(-\hat{J}^{(n)})(n\,\mathrm{d}v)-\hat{J}^{(n)}(0)\cdot\hat{\varepsilon}_1^{(n)}(t),$$

and further applying to the fact $\hat{\varepsilon}_1^{(n)}\le\check{\varepsilon}_1^{(n)}$ from (6.54) gives for every $t\le\hat{\tau}_+^{(n)}$,

$$\big(1-\alpha+\hat{J}^{(n)}(0)\big)\cdot\hat{\varepsilon}_1^{(n)}(t)\le\check{\varepsilon}_{1,1}^{(n)}(t)+\lambda_0^{(n)}\cdot\check{\varepsilon}_{1,2}^{(n)}(t)+\int_0^t\hat{\varepsilon}_1^{(n)}(t-v)(-\hat{J}^{(n)})(n\,\mathrm{d}v).$$

In particular, taking the supremum over $t\in[0,T\wedge\hat{\tau}_+^{(n)}]$ above gives

$$(1-\alpha)\cdot\sup_{t\le T}\hat{\varepsilon}_1^{(n)}(t\wedge\hat{\tau}_+^{(n)})\le\sup_{t\le T}\check{\varepsilon}_{1,1}^{(n)}(t\wedge\hat{\tau}_+^{(n)})+\lambda_0^{(n)}\cdot\sup_{t\le T}\check{\varepsilon}_{1,2}^{(n)}(t\wedge\hat{\tau}_+^{(n)}).$$

Following the same procedures as in the proof of Lemma 6.6, we can obtain

$$\check{\varepsilon}_{1,1}^{(n)}(t\wedge\hat{\tau}_+^{(n)})\to0\quad\text{and}\quad\check{\varepsilon}_{1,2}^{(n)}(t\wedge\hat{\tau}_+^{(n)})\to0\quad\text{u.o.c. in probability.}$$

Combining the above results finishes the proof. □

Now, we are ready to prove the FCLT, where the SDE for $\hat{Q}^{(n)}$ is analyzed and the following version of [33, Theorem 5.4] is applied.

Proposition 6.3. *Suppose that $(x_n,y_n,u_n,\varepsilon_n)$ is adapted to $\{\mathscr{F}_n(t)\}_{t\ge0}$ and satisfies*

$$x_n(t)=u_n(t)+\int_0^t f\big(x_n(s-)\big)\,\mathrm{d}y_n(s)+\varepsilon_n(t)\tag{6.56}$$

(1) $(y_n,u_n)\Rightarrow(y,u)$ in the Skorohod J_1 topology and $y_n\in\mathbb{D}$ is an increasing process such that

$$\lim_{c\to\infty}\sup_{n\ge1}\mathbb{P}\big(y_n(t)>c\big)=0\quad\forall t>0;$$

(2) for every $b>0$, let $\tau_{n,b}:=\inf\{t>0:|z_n|(t-)\wedge|z_n|(t)\ge b\}$, then

$$\varepsilon_n\big(t\wedge\tau_{n,b}\big)\to0\quad\textit{u.o.c. in probability;}$$

(3) the equation

$$x(t)=u(t)+\int_0^t f\big(x(s-)\big)\,\mathrm{d}y(s),\tag{6.57}$$

has a weak solution with any initial distribution on $\mathbb{R}$, and the weak uniqueness holds.

Then $(x_n,y_n,z_n,u_n)\Rightarrow(x,y,z,u)$ in the Skorohod J_1 topology.

Proof of Theorem 4.1 (the case $\bar{q}_0>0$). Plugging (6.53) into the expression of $\hat{Z}^{(n)}$ in (6.49), we obtain

$$\hat{Z}^{(n)}(t)=\int_0^t\frac{\hat{H}_{\bar{q}_0}(\hat{Q}^{(n)}(s-))}{1-H(\bar{q}_0)}\,\mathrm{d}\bar{A}^{(n)}(s)+\hat{\Xi}^{(n)}(t)+\hat{\varepsilon}^{(n)}(t),\tag{6.58}$$

similar to (6.47), where

$$\hat{\Xi}^{(n)}(t):=\hat{q}_0^{(n)}-\hat{\mu}^{(n)}t-\hat{S}^{(n)}\big(\bar{B}^{(n)}(t)\big)+\frac{\hat{\lambda}_0^{(n)}t+\hat{X}^{(n)}(t)}{1-H(\bar{q}_0)}\quad\text{and}\quad\hat{\varepsilon}^{(n)}(t):=\frac{\hat{\varepsilon}_2^{(n)}(t)-\hat{\varepsilon}_1^{(n)}(t)}{1-H(\bar{q}_0)}.$$

Noticing that for arbitrary fixed $k_0>0$ and n large enough, we have

$$\hat{I}^{(n)}(t\wedge\hat{\tau}_+^{(n)})=0,\quad\hat{Z}^{(n)}(t\wedge\hat{\tau}_+^{(n)})=\hat{Q}^{(n)}(t\wedge\hat{\tau}_+^{(n)}),$$
$$\text{and}\quad\hat{\varepsilon}^{(n)}(t\wedge\hat{\tau}_+^{(n)})\to0\quad\text{u.o.c. in probability.}$$

Then the same procedure as in the proof of Theorem 4.1 can be applied to (6.58), which results in the convergence

$$\big(\bar{A}^{(n)},\hat{X}^{(n)},\hat{S}^{(n)}(\bar{B}^{(n)}),\hat{Z}^{(n)},\hat{\Xi}^{(n)},\hat{Q}^{(n)},\hat{I}^{(n)}\big)\Rightarrow\big(\bar{A},\hat{X},\hat{S},\hat{Z},\hat{\Xi},\hat{Q},\hat{I}\big)\quad\text{in}\quad(\mathbb{D}^7[0,T],J_1).\tag{6.59}$$

comparing with (6.48), where $\hat{Q} = \hat{Z}$, $\hat{I} = 0$ and $\hat{Q}$ solves the SDE:

$$\mathrm{d}\hat{Q}(t) = \frac{\mu\, \hat{H}_{\bar{q}_0}(\hat{Q}(s-))}{1 - H(\bar{q}_0)}\mathrm{d}t + \mathrm{d}\hat{\Xi}(t) \quad \text{and} \quad \hat{\Xi}(t) = \hat{q}_0 - \hat{\mu}\, t - \hat{S}(t) + \frac{\hat{\lambda}_0\, t + \hat{X}(t)}{1 - H(\bar{q}_0)},$$

which is the same as (4.12). The existence and uniqueness of a solution to the SDE above is as discussed in Remark 4.2.

On the other hand, we have from (6.49),

$$\hat{A}^{(n)}(t) = -\hat{q}_0^{(n)} + \hat{Z}^{(n)}(t) + \hat{\mu}^{(n)}t + \hat{S}^{(n)}\big(\bar{B}^{(n)}(t)\big),$$

and also, from (4.2) with $\bar{q}_0 > 0$, (6.7) and (6.10) that

$$\begin{aligned}\hat{W}^{(n)}(t) =& m^{(n)}\hat{Q}^{(n)}(t) + \sqrt{n}\big(\bar{W}^{(n)} - m^{(n)}\bar{Q}^{(n)}\big)(t) + \hat{m}^{(n)}\bar{q}_0 \\ =& m^{(n)}\hat{Q}^{(n)}(t) + \hat{U}^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t)\big) + m^{(n)}\hat{S}^{(n)}\big(\bar{B}^{(n)}(t)\big) - \frac{\hat{\mu}^{(n)}\bar{w}_0}{\mu^{(n)}}.\end{aligned}$$

By the joint convergence in (6.59) and Corollary 6.1, we obtain the joint convergence for $(\hat{A}^{(n)}, \hat{Q}^{(n)}, \hat{W}^{(n)})$, noticing that $\bar{w}_0 = \mu^{-1}\bar{q}_0 > 0$.

For the last, given the joint convergence results above, the SDE for $(\hat{A}, \hat{Q})$ follows directly from (6.49) and (6.53), and the fact $\hat{I} = 0$. This finishes the proof. □

7. Proofs in the workload dependent case

The proofs for the workload-dependent case follow from the same procedures as those for queue-length dependent case. Here, we mainly highlight the differences in the proofs.

7.1. **Proof of the FLLN.** In this case, the reflection mapping is applied to $\bar{W}^{(n)}$ in (6.10), that is, we rewrite from (6.10),

$$\bar{W}^{(n)}(t) = \bar{Y}^{(n)}(t) + \bar{I}^{(n)}(t) = \bar{Y}^{(n)}(t) + \psi(\bar{Y}^{(n)})(t) = \phi(\bar{Y}^{(n)})(t), \tag{7.1}$$

where

$$\begin{aligned}\bar{Y}^{(n)}(t) &= \bar{U}^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t)\big) - t \\ &= \big(\bar{U}^{(n)}(s) - m^{(n)}s\big)\Big|_{s=\bar{q}_0^{(n)}+\bar{A}^{(n)}(t)} + m^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t)\big) - t.\end{aligned} \tag{7.2}$$

comparing with (6.8) and (6.9). By change of variables, similar to (6.11) but with $\lambda^{(n)}$ defined in term of $\bar{W}^{(n)}$ in (5.2), we can write

$$\begin{aligned}\bar{A}^{(n)}(t) =& \bar{X}^{(n)}(t) + \lambda_0^{(n)}t + \int_0^t \left(\int_0^{n(t-u)} h^{(n)}\big(v, n\bar{W}^{(n)}(u-)\big)\,\mathrm{d}v\right)\mathrm{d}\bar{A}^{(n)}(u) \\ =& \bar{X}^{(n)}(t) + \lambda_0^{(n)}t + \int_0^t H^{(n)}\big(\bar{W}^{(n)}(u-)\big)\,\mathrm{d}\bar{A}^{(n)}(u) - \bar{\varepsilon}_1^{(n)}(t),\end{aligned} \tag{7.3}$$

with abuse of notations, where

$$\bar{\varepsilon}_1^{(n)}(t) = \int_0^t \left(\int_{n(t-u)}^{\infty} h^{(n)}\big(v, n\bar{W}^{(n)}(u-)\big)\,\mathrm{d}v\right)\mathrm{d}\bar{A}^{(n)}(u).$$

The localization technique is also applied in the proofs, again with abuse of notations, we denote by

$$\bar{\tau}_{+,k_0}^{(n)} := \inf\{t > 0 : \bar{W}^{(n)}(t) > k_0\}$$

comparing with (6.13), where the first passage times are defined for $\bar{W}^{(n)}$ instead of $\bar{Q}^{(n)}$. Let α and $\bar{J}^{(n)}$ be defined in (6.14), and recall the facts in (6.15). By analogous arguments, we can prove the following lemmas similar to Lemmas 6.2 and 6.3.

Lemma 7.1. *Under Assumption A1, $\bar{X}^{(n)}(\cdot\wedge\bar{\tau}_+^{(n)})$ is a martingale with respect to $\{\bar{\mathscr{F}}^{(n)}(t)\}_{t\geq 0}$. There is a constant $c_0>0$, such that for every $t>0$,*

$$\mathbb{E}\big[\bar{A}^{(n)}(t\wedge\bar{\tau}_+^{(n)})\big]\leq\frac{\lambda_0^{(n)}}{1-\alpha}\cdot t\quad\text{and}\quad\mathbb{E}\Big[\sup_{s\leq t}(\bar{X}^{(n)})^2(s\wedge\bar{\tau}_+^{(n)})\Big]\leq\frac{c_0}{n}\cdot t.$$

Hence, $\bar{X}^{(n)}(\cdot\wedge\bar{\tau}_+^{(n)})$ converges to 0 in $L^2(\mathbb{P})$.

Lemma 7.2. *Under Assumption A1, for every $T>0$ and $\delta>0$,*

$$\limsup_{n\to\infty}\mathbb{E}\Big[w_\delta\big(\bar{A}^{(n)}(\cdot\wedge\bar{\tau}_+^{(n)}),T\big)\Big]\leq c_0\cdot\delta\quad\text{and}\quad\limsup_{n\to\infty}\mathbb{E}\Big[\sup_{t\leq T}\bar{\varepsilon}_1^{(n)}(t\wedge\bar{\tau}_+^{(n)})\Big]=0.$$

Proof of Theorem 5.1. Fixing arbitrary $T>0$ and $k_0>\bar{w}_0$, for $(\bar{q}_0^{(n)},\bar{A}^{(n)},\bar{Y}^{(n)})$ in (7.3) and (7.2), one can check from Lemmas 6.1, 7.1 and 7.2 that it is tight in $(\mathbb{R}_+,\mathcal{B}(\mathbb{R}_+))\times(\mathbb{D}^2[0,T],J_1)$.

Let $\big(\bar{q}_0,\bar{A}_{k_0},\bar{Y}_{k_0}\big)$ be a limit point over some converging subsequence $\{n_k\}_{k\geq 1}$, that is,

$$\big(\bar{q}_0^{(n)},\bar{A}^{(n)}(t\wedge\bar{\tau}_{+,k_0}^{(n)}),\bar{Y}^{(n)}(t\wedge\bar{\tau}_{+,k_0}^{(n)})\big)\Big|_{n=n_k}\Rightarrow\big(\bar{q}_0,\bar{A}_{k_0}(t),\bar{Y}_{k_0}(t)\big),$$

in $(\mathbb{R}_+,\mathcal{B}(\mathbb{R}_+))\times(\mathbb{D}^2[0,T],J_1)$. Then it is clear that $\bar{A}_{k_0},\bar{Y}_{k_0}\in\mathbb{C}$.

Plugging the above convergence result into (7.2) and making use of Lemma 6.1 gives

$$\big(\bar{A}^{(n)},\bar{Y}^{(n)},\mathfrak{e}\big)(t\wedge\bar{\tau}_{+,k_0}^{(n)})\Rightarrow\big(\bar{A}_{k_0},\bar{Y}_{k_0},\mathfrak{e}_{k_0}\big)(t)\quad\text{in }(\mathbb{D}^3[0,T],J_1),$$

where $\mathfrak{e}_{k_0}$ is a continuous function and

$$\bar{Y}_{k_0}(t)=m\big(\bar{q}_0+\bar{A}_{k_0}(t)\big)-\mathfrak{e}_{k_0}(t).$$

Further plugging into (7.1) proves

$$\big(\bar{A}^{(n)},\bar{Y}^{(n)},\mathfrak{e},\bar{W}^{(n)},\bar{I}^{(n)}\big)(t\wedge\bar{\tau}_{+,k_0}^{(n)})\Rightarrow\big(\bar{A}_{k_0},\bar{Y}_{k_0},\mathfrak{e}_{k_0},\bar{W}_{k_0},\bar{I}_{k_0}\big)(t)\quad\text{in }(\mathbb{D}^5[0,T],J_1),\tag{7.4}$$

where

$$\bar{I}_{k_0}(t)=\psi\big(\bar{Y}_{k_0}\big)(t),\quad\bar{W}_{k_0}(t)=\phi\big(\bar{Y}_{k_0}\big)(t)\quad\text{and}\quad\mathfrak{e}_{k_0}(t)=t\wedge\bar{\tau}_{+,k_0},\tag{7.5}$$

with $\bar{\tau}_{+,k_0}:=\inf\{t>0:\bar{W}_{k_0}(t)>k_0\}$. By the argument for the strong local uniqueness of solutions, we can obtain that

$$\big(\bar{A}_{k_0},\bar{Y}_{k_0},\bar{I}_{k_0},\bar{W}_{k_0}\big)(t)=\big(\bar{A},\bar{Y},\bar{I},\bar{W}\big)(t\wedge\bar{\tau}_{+,k_0}),\tag{7.6}$$

for some common $(\bar{A},\bar{Y},\bar{I},\bar{W})$ and $\bar{\tau}_{+,k_0}:=\inf\{t>0:\bar{W}(t)>k_0\}$.

Next, plugging (7.4), (7.5) and (7.6) into (7.3), and further applying Lemma 7.1 for $\bar{X}^{(n)}$, Lemma 7.2 for $\bar{\varepsilon}_1^{(n)}$ and [33, Theorem 2.2], we obtain

$$\bar{A}(s)=\lambda_0 s+\int_0^s H\big(\bar{W}(u)\big)\mathrm{d}\bar{A}(u)\Big|_{s=t\wedge\bar{\tau}_{+,k_0}}\quad\text{and}\quad\bar{W}(s)=\phi(\bar{Y})(s)\Big|_{s=t\wedge\bar{\tau}_{+,k_0}}\quad\forall t\leq T.$$

Then, the rest part in the proof of Theorem 3.1 can be applied to finish the proof. □

7.2. **Proof of the FCLT.** To prove the FCLT, Lemma 6.5 and Corollary 6.1 as well as notations in the proof for the queue-dependent $\lambda^{(n)}$ in (2.12) are applied.

7.2.1. **The case $\bar{q}_0 = 0 \in \mathcal{L}^{=}$.** In addition to $(\hat{A}^{(n)}, \hat{Q}^{(n)}, \hat{W}^{(n)})$ in (4.2), we further define

$$\big(\hat{Y}^{(n)}, \hat{I}^{(n)}\big) := \sqrt{n}\,\big(\bar{Y}^{(n)}, \bar{I}^{(n)}\big).$$

Recalling (7.1) and (7.2), we have

$$\hat{Y}^{(n)}(t) = \sqrt{n}\cdot\bar{Y}^{(n)}(t) = \hat{U}^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t)\big) + m^{(n)}\big(\hat{q}_0^{(n)} + \hat{A}^{(n)}(t) - \hat{\mu}^{(n)}\,t\big) \tag{7.7}$$

where $m^{(n)}\mu^{(n)} = 1$, and

$$\hat{I}^{(n)}(t) = \psi\big(\hat{Y}^{(n)}\big)(t) \quad \text{and} \quad \hat{W}^{(n)}(t) = \hat{Y}^{(n)}(t) + \hat{I}^{(n)}(t) = \phi\big(\hat{Y}^{(n)}\big)(t). \tag{7.8}$$

Moreover, we also have from (7.3) that

$$\begin{aligned}\hat{A}^{(n)}(t) &= \sqrt{n}\big(\bar{A}^{(n)}(t) - \bar{A}(t)\big) = \sqrt{n}\big(\bar{A}^{(n)}(t) - \big(\lambda_0 t + H(0)\cdot\bar{A}(t)\big)\big)\\ &= \hat{\lambda}_0^{(n)}\,t + \hat{X}^{(n)}(t) + \sqrt{n}\int_0^t \Big(H^{(n)}\big(\bar{W}^{(n)}(u-)\big) - H(0)\Big)\,d\bar{A}^{(n)}(u)\\ &\quad + H(0)\cdot\sqrt{n}\big(\bar{A}^{(n)}(t) - \bar{A}(t)\big) - \hat{\varepsilon}_1^{(n)}(t)\end{aligned}$$

where $\hat{\varepsilon}_1^{(n)} = \sqrt{n}\,\bar{\varepsilon}_1^{(n)}$, and similar to (6.37), we obtain

$$\hat{A}^{(n)}(t) = \frac{\hat{\lambda}_0^{(n)}\,t + \hat{X}^{(n)}(t)}{1 - H(0)} + \int_0^t \frac{\hat{H}_0^{(n)}(\hat{W}^{(n)}(u-))}{1 - H(0)}\,\mathrm{d}\bar{A}^{(n)}(u) - \frac{\hat{\varepsilon}_1^{(n)}(t)}{1 - H(0)}, \tag{7.9}$$

recalling $\hat{H}_0^{(n)}$ in (4.3). The localization technique is used in $\sqrt{n}$-scaled sense, that is, we define

$$\hat{\tau}_+^{(n)} = \inf\{t > 0 : \hat{W}^{(n)}(t) > k_0\} \ \text{ and } \ \hat{J}^{(n)}(t) := \sup_{y\le k_0}\int_t^\infty h^{(n)}(u, \sqrt{n}y)\,\mathrm{d}u,$$

similar to the ones defined in (6.34). Similar to Lemma 6.6, we can prove the following result.

Lemma 7.3. *Under Assumptions A1, A2 and A4, we have*

$$\hat{\varepsilon}_1^{(n)}(t \wedge \hat{\tau}_+^{(n)}) \to 0 \quad \textit{u.o.c. in probability.}$$

Now, we are ready to prove the FCLT in the workload-dependent case. Instead of working on $\hat{Z}^{(n)}$ in (6.33) for the proof of the queue-dependent model, we work on $\hat{Y}^{(n)}$ in (7.7) and (7.8) for the workload-dependent model.

Proof of Theorem 5.2 ($\bar{q}_0 = 0$)**.** Substituting (7.9) into (7.7), we have

$$\begin{aligned}\hat{Y}^{(n)}(t) &= m^{(n)}\int_0^t \frac{\hat{H}_0^{(n)}(\hat{W}^{(n)}(u-))}{1 - H(0)}\,\mathrm{d}\bar{A}^{(n)}(u) + \hat{\Xi}^{(n)}(t) - m^{(n)}\frac{\hat{\varepsilon}_1^{(n)}(t)}{1 - H(0)}\\ &= \int_0^t \frac{\hat{H}_0(\hat{W}^{(n)}(u-))}{\mu(1 - H(0))}\,\mathrm{d}\bar{A}^{(n)}(u) + \hat{\Xi}^{(n)}(t) + m^{(n)}\frac{\hat{\varepsilon}_2^{(n)}(t) - \hat{\varepsilon}_1^{(n)}(t)}{1 - H(0)}\end{aligned} \tag{7.10}$$

where comparing with (6.37) and (6.47),

$$\begin{aligned}\hat{\Xi}^{(n)}(t) &= \hat{U}^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t)\big) + m^{(n)}\Big(\hat{q}_0^{(n)} + \frac{\hat{\lambda}_0^{(n)}\,t + \hat{X}^{(n)}(t)}{1 - H(0)} - \hat{\mu}^{(n)}\,t\Big),\\ \hat{\varepsilon}_2^{(n)}(t) &:= \int_0^t \Big(\hat{H}_0^{(n)}(y) - \mu^{(n)} m\,\hat{H}_0(y)\Big)\Big|_{y=\hat{W}^{(n)}(s-)}\,\mathrm{d}\bar{A}^{(n)}(s).\end{aligned}$$

We observe that

$$|\hat{\varepsilon}_2^{(n)}|(t \wedge \hat{\tau}_+^{(n)}) \le \sup_{y\le k_0}\big|\hat{H}_0^{(n)}(y) - \mu^{(n)} m\,\hat{H}_0(y)\big|\cdot\bar{A}^{(n)}(t \wedge \hat{\tau}_+^{(n)}) \tag{7.11}$$

from which we obtain the convergence of $\hat{\varepsilon}_2^{(n)}$ to 0 u.o.c. in probability by Assumption A4-(ii) and the stochastic boundedness for $\bar{A}^{(n)}(t \wedge \hat{\tau}_+^{(n)})$ from Lemma 7.1.

Applying Corollary 6.1, we have the joint convergence

$$\big(\bar{A}^{(n)}, \hat{X}^{(n)}, \hat{U}^{(n)}(\bar{q}_0^{(n)} + \bar{A}^{(n)}), \hat{\Xi}^{(n)}\big) \Rightarrow \big(\bar{A}, \hat{X}, \hat{U}(\mu\mathfrak{e}), \hat{\Xi}\big) \quad \text{in} \quad (\mathbb{D}^4[0,T], J_1),$$

recalling that $\bar{q}_0 = 0$, where by Corollary 6.1,

$$\hat{U}(\mu\, t) = -\mu^{-1}\hat{S}(t) \quad \text{and} \quad \hat{\Xi}(t) = -\mu^{-1}\hat{S}(t) + \mu^{-1}\Big(\hat{q}_0 + \frac{\hat{\lambda}_0\, t + \hat{X}(t)}{1 - H(0)} - \hat{\mu}\, t\Big).$$

We next observe from Lemmas 7.1, 7.3 and (7.11) that conditions (1) and (2) in Proposition 6.2 hold. Thus, applying the proposition to (7.10) and (7.8), any limit point of

$$\big(\bar{A}^{(n)}, \hat{X}^{(n)}, \hat{U}^{(n)}(\bar{q}_0 + \bar{A}^{(n)}), \hat{\Xi}^{(n)}, \hat{Y}^{(n)}, \hat{W}^{(n)}, \hat{I}^{(n)}\big), \tag{7.12}$$

say $(\bar{A}, \hat{X}, \hat{U}, \hat{\Xi}, \hat{Y}, \hat{W}, \hat{I})$, solves the SDE:

$$\mathrm{d}\hat{Y}(t) = \frac{\hat{H}_0(\hat{W}(t))}{1 - H(0)}\,\mathrm{d}t + \mathrm{d}\,\hat{\Xi}(t), \quad \hat{I} = \psi(\hat{Y}) \quad \text{and} \quad \hat{W} = \phi(\hat{Y}),$$

noticing that $\bar{A}(t) = \mu\, t$. Recalling $\lambda_0 = \mu(1 - H(0))$ and $\hat{\rho}_0$ in (4.10), we get

$$\begin{aligned} \mathrm{d}\hat{W}(t) &= \mathrm{d}\hat{Y}(t) + \mathrm{d}\hat{I}(t) = \Big(\frac{\hat{H}_0(\hat{W}(t))}{1 - H(0)} + \frac{\hat{\lambda}_0}{\lambda_0} - \frac{\hat{\mu}}{\mu}\Big)\,\mathrm{d}t + \mathrm{d}\Big(\frac{\hat{X}(t)}{\lambda_0} - \frac{\hat{S}(t)}{\mu}\Big) + \mathrm{d}\hat{I}(t) \\ &= -\,\hat{\rho}_0\big(\hat{W}(t)\big)\,\mathrm{d}t + \mathrm{d}\Big(\frac{\hat{X}(t)}{\lambda_0} - \frac{\hat{S}(t)}{\mu}\Big) + \mathrm{d}\hat{I}(t), \end{aligned} \tag{7.13}$$

where $\hat{I}$ is the regulator for $\hat{W}$. As discussed in Remark 5.1, the SDE above has a unique strong solution on $\mathbb{R}_+$, so that condition (3) in Proposition 6.2 is satisfied. This proves the convergence for the processes in (7.12) in Skorohod J_1 topology. Also, from the fact that the solution to (7.13) cannot explode, we obtain that for every $T > 0$,

$$\limsup_{k\to\infty} \limsup_{n\to\infty} \mathbb{P}\big(T \le \hat{\tau}_{+,k}^{(n)}\big) = 0. \tag{7.14}$$

Next, given the joint convergence in (7.12), (7.14) and the limit for $\hat{\varepsilon}_1^{(n)}$ in Lemma 7.4, we can apply [33, Theorem 2.2] to (7.9) to show the joint convergence of

$$\big(\bar{A}^{(n)}, \hat{X}^{(n)}, \hat{U}^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}\big), \hat{Y}^{(n)}, \hat{\Xi}^{(n)}, \hat{W}^{(n)}, \hat{I}^{(n)}, \hat{A}^{(n)}\big), \tag{7.15}$$

in the Skorohod J_1 topology. By the facts $\bar{q}_0 = 0$, (4.2), (6.7) and (6.10), we obtain that

$$\begin{aligned} \hat{Q}^{(n)}(t) &= \sqrt{n}\,\bar{Q}^{(n)}(t) = \sqrt{n}\big(\bar{Q}^{(n)}(t) - \mu^{(n)}\bar{W}^{(n)}(t)\big) + \mu^{(n)}\hat{W}^{(n)}(t) \\ &= -\,\hat{S}^{(n)}\big(\bar{B}^{(n)}(t)\big) - \mu^{(n)}\hat{U}^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t)\big) + \mu^{(n)}\hat{W}^{(n)}(t). \end{aligned}$$

Applying the joint convergences of the processes in (7.15) and Corollary 6.1, we obtain the joint convergence for $(\hat{A}^{(n)}, \hat{Q}^{(n)}, \hat{W}^{(n)})$.

For the last, given the results of the joint convergence above, the SDE for $(\hat{A}, \hat{W})$ follows directly from (7.7), (7.8), (7.9) and Corollary 6.1. This finishes the proof. □

7.2.2. **The case $\bar{q}_0 \in \mathcal{L}^=$ and $\bar{q}_0 > 0$.** The proof of this workload-dependent case follows essentially the same procedure as that for queue-dependent case, with $\bar{Q}^{(n)}$ replaced by $\bar{W}^{(n)}$.

Firstly, we have from (7.3) and the fact $\lambda_0 = \mu\,(1 - H(\bar{w}_0))$ that

$$\begin{aligned}\hat{A}^{(n)}(t) &= \sqrt{n}\Big(\bar{A}^{(n)}(t) - \lambda_0\, t - H(\bar{w}_0)\bar{A}(t)\Big)\\ &= \hat{X}^{(n)}(t) + \hat{\lambda}_0^{(n)}\, t + \int_0^t \hat{H}_{\bar{w}_0}^{(n)}\big(\hat{W}^{(n)}(s-)\big)\,\mathrm{d}\bar{A}^{(n)}(s) + H(\bar{w}_0)\cdot\hat{A}^{(n)}(t) - \hat{\varepsilon}_1^{(n)}(t)\end{aligned}$$

where $\hat{\varepsilon}_1^{(n)} := \sqrt{n}\,\bar{\varepsilon}_1^{(n)}$, and we further obtain

$$\hat{A}^{(n)}(t) = \frac{\hat{\lambda}_0^{(n)}\, t + \hat{X}^{(n)}(t)}{1 - H(\bar{w}_0)} + \int_0^t \frac{\hat{H}_{\bar{w}_0}^{(n)}(\hat{W}^{(n)}(s-))}{1 - H(\bar{w}_0)}\,\mathrm{d}\bar{A}^{(n)}(s) - \frac{\hat{\varepsilon}_1^{(n)}(t)}{1 - H(\bar{w}_0)}, \tag{7.16}$$

comparing with (6.53). Denoting by

$$\big(\hat{Y}^{(n)}(t), \hat{I}^{(n)}(t)\big) = \sqrt{n}\,\big(\bar{Y}^{(n)}(t) - \bar{w}_0, \bar{I}^{(n)}(t)\big),$$

and recalling $\hat{W}^{(n)}$ defined in (4.2), we have from (7.1) and (7.2) that

$$\hat{W}^{(n)}(t) = \hat{Y}^{(n)}(t) + \hat{I}^{(n)}(t) \tag{7.17}$$

where

$$\begin{aligned}\hat{Y}^{(n)}(t) &= \hat{U}^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t)\big) + \sqrt{n}\big(m^{(n)}\bar{q}_0^{(n)} - \bar{w}_0\big) + m^{(n)}\big(\bar{A}^{(n)}(t) - \mu^{(n)}\, t\big)\\ &= \hat{w}_0^{(n)} + \hat{U}^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t)\big) + m^{(n)}\big(\hat{A}^{(n)}(t) - \hat{\mu}^{(n)}\, t\big),\end{aligned} \tag{7.18}$$

where

$$\hat{w}_0^{(n)} := \sqrt{n}\big(m^{(n)}\bar{q}_0^{(n)} - \bar{w}_0\big) = m^{(n)}\hat{q}_0^{(n)} - \bar{w}_0 m^{(n)}\hat{\mu}^{(n)}\,.$$

Under Assumption A4, we obtain that as $n \to \infty$,

$$\hat{w}_0^{(n)} \to \hat{q}_0\mu^{-1} - \bar{w}_0\hat{\mu}\mu^{-1} = \hat{w}_0. \tag{7.19}$$

It is expected that the regulator $\hat{I}^{(n)}$ is negligible for large n.

For arbitrary $k_0 > 0$, let the passage time for $\hat{W}^{(n)}$ be defined by

$$\hat{\tau}_+^{(n)} := \inf\{t > 0 : |\hat{W}^{(n)}(t)| > k_0\}.$$

Then, we can show the following result as in Lemma 6.7.

Lemma 7.4. *Under Assumptions A1, A2 and A4, we have*

$$\hat{\varepsilon}_1^{(n)}(t \wedge \hat{\tau}_+^{(n)}) \to 0 \quad \textit{u.o.c. in probability.}$$

Proof of Theorem 5.2 (the case $\bar{q}_0 > 0$). Plugging (7.16) into the expression of $\hat{Y}^{(n)}$ in (7.18), we obtain

$$\begin{aligned}\hat{Y}^{(n)}(t) &= m^{(n)}\int_0^t \frac{\hat{H}_{\bar{w}_0}^{(n)}(\hat{W}^{(n)}(s-))}{1 - H(\bar{w}_0)}\,\mathrm{d}\bar{A}^{(n)}(s) + \hat{\Xi}^{(n)}(t) - \frac{m^{(n)}\hat{\varepsilon}_1^{(n)}(t)}{1 - H(\bar{w}_0)}\\ &= \int_0^t \frac{\hat{H}_{\bar{w}_0}(\hat{W}^{(n)}(s-))}{\mu(1 - H(\bar{w}_0))}\,\mathrm{d}\bar{A}^{(n)}(s) + \hat{\Xi}^{(n)}(t) + m^{(n)}\frac{\hat{\varepsilon}_2^{(n)}(t) - \hat{\varepsilon}_1^{(n)}(t)}{1 - H(\bar{w}_0)},\end{aligned} \tag{7.20}$$

similar to (6.47), where

$$\begin{aligned}\hat{\Xi}^{(n)}(t) &:= \hat{w}_0^{(n)} + \hat{U}^{(n)}\big(\bar{q}_0^{(n)} + \bar{A}^{(n)}(t)\big) + m^{(n)}\Big(\frac{\hat{\lambda}_0^{(n)}\, t + \hat{X}^{(n)}(t)}{1 - H(\bar{w}_0)} - \hat{\mu}^{(n)}\, t\Big),\\ \hat{\varepsilon}_2^{(n)}(t) &:= \int_0^t \Big(\hat{H}_{\bar{w}_0}^{(n)}(y) - \mu^{(n)}\mu^{-1}\hat{H}_{\bar{w}_0}(y)\Big)\Big|_{y=\hat{W}^{(n)}(s-)}\,\mathrm{d}\bar{A}^{(n)}(s).\end{aligned}$$

Noticing that for every n large enough, we have from (7.17),

$$\hat{I}^{(n)}(t\wedge\hat{\tau}_+^{(n)})=0 \quad \text{and} \quad \hat{Y}^{(n)}(t\wedge\hat{\tau}_+^{(n)})=\hat{W}^{(n)}(t\wedge\hat{\tau}_+^{(n)}).$$

We have

$$\big|\hat{\varepsilon}_2^{(n)}\big|(t\wedge\hat{\tau}_+^{(n)})\le \sup_{|y|\le k_0}\big|\hat{H}_{\bar{w}_0}^{(n)}(y)-\mu^{(n)}\mu^{-1}\hat{H}_{\bar{w}_0}(y)\big|\cdot\bar{A}^{(n)}(t\wedge\hat{\tau}_+^{(n)}) \tag{7.21}$$

from which we obtain $\hat{\varepsilon}_2^{(n)}$ converges to 0 u.o.c. in probability, by Assumption A4-(ii) for the function $\hat{H}_{\bar{w}_0}^{(n)}$ and the stochastic boundedness for $\bar{A}^{(n)}$ from Lemma 7.1. Together with the convergence of $\hat{\varepsilon}_1^{(n)}$ from Lemma 7.4, the convergence of $\hat{\varepsilon}_2^{(n)}$ from (7.21), and the joint convergence for $(\hat{X}^{(n)},\hat{U}^{(n)})$ from Corollary 6.1, the same procedure as in the proof of Theorem 4.1 can be applied to (7.20) and results in the following

$$\begin{aligned}&\big(\bar{A}^{(n)},\hat{X}^{(n)},\hat{U}^{(n)}\big(\bar{q}_0^{(n)}+\bar{A}^{(n)}\big),\hat{Y}^{(n)},\hat{\Xi}^{(n)},\hat{W}^{(n)},\hat{I}^{(n)}\big)\\ \Rightarrow&\big(\bar{A},\hat{X},-\mu^{-1}\hat{S}(\mathfrak{e}+\bar{w}_0),\hat{Y},\hat{\Xi},\hat{W},\hat{I}\big)\quad\text{in}\quad(\mathbb{D}^7[0,T],J_1).\end{aligned} \tag{7.22}$$

comparing with (6.48), where $\hat{W}=\hat{Y}$, $\hat{I}=0$ and $\hat{W}$ is given by

$$\hat{W}(t)=\hat{w}_0+\int_0^t\Big(\frac{\hat{H}_{\bar{w}_0}(\hat{W}(s))}{1-H(\bar{w}_0)}+\frac{\hat{\lambda}_0}{\lambda_0}-\frac{\hat{\mu}}{\mu}\Big)\,\mathrm{d}s+\Big(\frac{\hat{X}(t)}{\lambda_0}-\frac{\hat{S}(t+\bar{w}_0)}{\mu}\Big)$$

recalling $\hat{w}_0$ in (7.19), where $\hat{W}(0)=\hat{w}_0-\mu^{-1}\hat{S}(\bar{w}_0)$.

Next, given the joint convergence in (7.22), the convergences of $\hat{\varepsilon}_1^{(n)}$ in Lemma 7.4 and $\hat{\varepsilon}_2^{(n)}$ from (7.21), [33, Theorem 2.2] can be applied to (7.20) and we obtain the joint convergence of

$$\big(\bar{A}^{(n)},\hat{X}^{(n)},\hat{U}^{(n)}\big(\bar{q}_0^{(n)}+\bar{A}^{(n)}\big),\hat{Y}^{(n)},\hat{\Xi}^{(n)},\hat{W}^{(n)},\hat{I}^{(n)},\hat{A}^{(n)}\big).$$

From (4.2) for $\bar{q}_0>0$, (6.7) and (6.10), we obtain

$$\begin{aligned}\hat{Q}^{(n)}(t)&=\sqrt{n}\big(\bar{Q}^{(n)}(t)-\mu^{(n)}\bar{W}^{(n)}(t)\big)+\mu^{(n)}\sqrt{n}\big(\bar{W}^{(n)}(t)-\bar{w}_0\big)+\bar{w}_0\sqrt{n}(\mu^{(n)}-\mu)\\&=-\hat{S}^{(n)}(\bar{B}^{(n)}(t))-\mu^{(n)}\hat{U}^{(n)}\big(\bar{q}_0^{(n)}+\bar{A}^{(n)}(t)\big)+\mu^{(n)}\hat{W}^{(n)}(t)+\bar{w}_0\hat{\mu}^{(n)}.\end{aligned}$$

By the joint convergences in (7.22) and Corollary 6.1, we obtain the joint convergence for $(\hat{A}^{(n)},\hat{Q}^{(n)},\hat{W}^{(n)})$.

For the last, given the joint convergence results above, the SDE (5.6) for $(\hat{A},\hat{W})$ follows directly from (7.16), (7.17), (7.18), the fact $\hat{I}=0$ and Corollary 6.1. This finishes the proof. □

8. Appendix: Additional examples of H

In this section we present three additional examples of the function H, in addition to Examples 1 and 2 in Section 3.1. Example 1 is a decreasing exponential function, while Example 2 is a sinusoidal function. Here the examples include a decreasing power function, an increasing power function and an increasing exponential function.

8.1. **Decreasing power function $H(y)=\beta(1+y)^{-\gamma}$ with $\beta\in(0,1)$ and $\gamma>0$.** Recalling ρ in (2.17), we have

$$\rho(y)=\frac{\lambda_0}{\mu(1-\beta(1+y)^{-\gamma})}\quad\text{and}\quad\mu\big(\rho(y)-1\big)=\frac{\lambda_0-\mu(1-\beta(1+y)^{-\gamma})}{1-\beta(1+y)^{-\gamma}}.$$

(1) If $\lambda_0>\mu$, we have $\mathcal{L}^+=\mathbb{R}_+$. $\bar{Q}$ decreases strictly in $\mathbb{R}_+$ and $\bar{I}\equiv0$,

$$\lim_{y\to\infty}\frac{1}{\mu(\rho(y)-1)}=\frac{1}{\lambda_0-\mu}>0,$$

thus $\bar{Q}(t)\sim(\lambda_0-\mu)\cdot t$ as $t\to\infty$.

(2) If $\lambda_0 = \mu$, we have $\mathcal{L}^+ = \mathbb{R}_+$, $\bar{Q}$ decreases strictly in $\mathbb{R}_+$, $\bar{I} \equiv 0$ and $\bar{Q}$ solves equation

$$\frac{(1+y)^{1+\gamma} - (1+x_0)^{1+\gamma}}{\mu\beta(1+\gamma)} - \frac{y - x_0}{\mu}\Big|_{y=\bar{Q}(t)} = t,$$

thus $\bar{Q}(t) \sim (\mu\beta(1+\gamma))^{\frac{1}{1+\gamma}} \cdot t^{\frac{1}{1+\gamma}}$ as $t \to \infty$.

(3) If $\mu > \lambda_0 > (1-\beta)\mu$, we have similar to Case-(3) of Example 1,

$$\mathcal{L}^+ = [0, y_0), \quad \mathcal{L}^- = (y_0, \infty), \quad \mathcal{L}^= = \{y_0\} \quad \text{and} \quad y_0 = \big(\frac{\mu\beta}{\mu - \lambda_0}\big)^{\frac{1}{\gamma}} - 1.$$

$\bar{Q}$ increases in $[0, y_0)$, decreases in (y_0, ∞) and y_0 is the unique critical equilibrium point. One can write

$$\mu\left(\rho(y) - 1\right) = \mu\beta\frac{(1+y)^{-\gamma} - (1+y_0)^{-\gamma}}{1 - \beta(1+y)^{-\gamma}} \sim \frac{\gamma\mu(\mu - \lambda_0)}{\lambda_0(1+y_0)} \cdot (y_0 - y) \quad \text{as } y \to y_0,$$

thus for $\bar{Q}$ starting from $x_0 \neq y_0$,

$$-\ln\left|\bar{Q}(t) - y_0\right| \sim \frac{\gamma\mu(\mu - \lambda_0)}{\lambda_0(1+y_0)} \cdot t \quad \text{as } t \to \infty.$$

(4) If $\lambda_0 = \mu(1-\beta)$, we have $\mathcal{L}^- = (0, \infty)$ and $\mathcal{L}^= = \{0\}$. $\bar{Q}$ decreases strictly in $(0, \infty)$, 0 is the unique equilibrium point.

(5) If $\lambda_0 < \mu(1-\beta)$, we have $\mathcal{L}^- = \mathbb{R}_+$, thus $\bar{Q}$ decreases strictly, hits 0 at some finite time and stays at 0 afterward.

8.2. **Increasing power function $H(y) = \beta(1 + y^{-\gamma})^{-1}$ with $\beta \in (0, 1]$ and $\gamma > 0$.** Recalling (2.17), we have

$$\rho(y) = \frac{\lambda_0}{\mu}\frac{1 + y^\gamma}{1 + (1-\beta)y^\gamma} \quad \text{and} \quad \mu\left(\rho(y) - 1\right) = \frac{(\lambda_0 - \mu) + (\lambda_0 - \mu(1-\beta))y^\gamma}{1 + (1-\beta)y^\gamma}.$$

(1) If $\lambda_0 \geq \mu$, we have $(0, \infty) \subset \mathcal{L}^+$. $\bar{Q}$ increases on $(0, \infty)$, one can check that

(a) if $\beta < 1$, we have $\mu\left(\rho(y) - 1\right) \to \dfrac{\lambda_0}{1-\beta} - \mu$ as $y \to \infty$, thus

$$\bar{Q}(t) \sim \big(\frac{\lambda_0}{1-\beta} - \mu\big) \cdot t \quad \text{as } t \to \infty;$$

(b) if $\beta = 1$, we have $\mu\left(\rho(y) - 1\right) \sim \lambda_0\, y^\gamma$ as $y \to \infty$, and thus,

(i) if $\gamma > 1$, Assumption A3 fails to hold, and $\bar{Q}$ explodes at some finite time.

(ii) if $\gamma = 1$, $\bar{Q}$ increases to infinity and we have

$$\ln(\bar{Q}(t)) \sim \lambda_0 \cdot t \quad \text{as } t \to \infty;$$

(iii) if $\gamma < 1$, $\bar{Q}$ increases to infinity and we have

$$\bar{Q}(t) \sim \big(\lambda_0(1-\gamma)\big)^{\frac{1}{1-\gamma}} \cdot t^{\frac{1}{1-\gamma}} \quad \text{as } t \to \infty.$$

One can check that $0 \in \mathcal{L}^=$ if and only if $\lambda_0 = \mu$.

(2) If $\mu > \lambda_0 > (1-\beta)\mu$, we have similar to Case-3 of Example 1,

$$\mathcal{L}^- = [0, y_0), \quad \mathcal{L}^+ = (y_0, \infty), \quad \mathcal{L}^= = \{y_0\} \quad \text{and} \quad y_0 = \Big(\frac{\mu - \lambda_0}{\lambda_0 - \mu(1-\beta)}\Big)^{1/\gamma},$$

Applying Proposition 3.1, one can find that

(a) if $x_0 < y_0$, $\bar{Q}$ decreases strictly, hits 0 at some finite time and stays at 0 afterward, thus 0 is a subcritical and equilibrium point;

(b) if $x_0 > y_0$, $\bar{Q}$ increases strictly and its asymptomatic behavior at ∞ is the same as the Case 1 for both $\beta < 1$ and $\beta = 1$;

(c) if $x_0 = y_0$, then $\bar{Q}(t) = x_0$ for all $t > 0$.

(3) If $\lambda_0 \le \mu(1-\beta)$, we have $\rho(y) < 1$ for all $y \ge 0$, which implies starting from $x_0 \ge 0$, $\bar{Q}$ hits 0 at some finite time, and stays at 0 afterward. Thus, 0 is the unique equilibrium point.

8.3. **Increasing exponential function $H(y) = \beta e^{-y^{-\gamma}}$ with $\beta \in (0,1]$ and $\gamma > 0$.** Recalling ρ in (2.17), we have

$$\rho(y) = \frac{\lambda_0}{\mu(1-\beta e^{-y^{-\gamma}})} \quad \text{and} \quad \mu\big(\rho(y)-1\big) = \frac{(\lambda_0-\mu)e^{y^{-\gamma}} + \mu\beta}{e^{y^{-\gamma}} - \beta}.$$

(1) If $\lambda_0 \ge \mu$ and $\beta \in (0,1)$, we have $(0,\infty) \subset \mathcal{L}^+$. $\bar{Q}$ increases in $\mathbb{R}_+$ and $\bar{I} = 0$. One can check that $\mu(\rho(y)-1) \to \dfrac{\lambda_0}{1-\beta} - \mu$ as $y \to \infty$, thus

$$\bar{Q}(t) \sim \big(\frac{\lambda_0}{1-\beta} - \mu\big) \cdot t \quad \text{as } t \to \infty.$$

(2) If $\lambda_0 \ge \mu$ and $\beta = 1$, we have $(0,\infty) \subset \mathcal{L}^+$. $\bar{Q}$ increases in $(0,\infty)$ and $\bar{I} = 0$. One can check that $\mu(\rho(y)-1) \sim \lambda_0\, y^\gamma$ as $y \to \infty$, thus for $\bar{Q}$ starting from $x_0 > 0$,

(a) if $\gamma > 1$, Assumption A3 fails to hold, and $\bar{Q}$ explodes at some finite time.

(b) if $\gamma = 1$, we have

$$\ln(\bar{Q}(t)) \sim \lambda_0 \cdot t \quad \text{as } t \to \infty.$$

(c) if $\gamma < 1$, we have

$$\bar{Q}(t) \sim \big(\lambda_0(1-\gamma)\big)^{\frac{1}{1-\gamma}} \cdot t^{\frac{1}{1-\gamma}} \quad \text{as } t \to \infty.$$

(3) If $\mu > \lambda_0 > (1-\beta)\mu$, we have

$$\mathcal{L}^- = [0, y_0), \quad \mathcal{L}^+ = (y_0, \infty), \quad \mathcal{L}^= = \{y_0\} \quad \text{and} \quad y_0 = \Big(\ln \frac{\mu\beta}{\mu - \lambda_0}\Big)^{-1/\gamma}.$$

One can find that

(a) if $x_0 < y_0$, $\bar{Q}$ decreases strictly, hits 0 at some finite time and stays at 0 afterward, thus 0 is a subcritical and equilibrium point;

(b) if $x_0 > y_0$ and $\beta \in (0,1)$, $\bar{Q}$ increases strictly and one can find that

$$\bar{Q}(t) \sim \big(\frac{\lambda_0}{1-\beta} - \mu\big) \cdot t \quad \text{as } t \to \infty;$$

(c) if $x_0 > y_0$ and $\beta = 1$, $\bar{Q}$ increases strictly and one can find that $\mu(\rho(y)-1) \sim \lambda_0 y^\gamma$ as $y \to \infty$, and thus,

(i) if $\gamma > 1$, $\bar{Q}$ will explode at some finite time;

(ii) if $\gamma = 1$, $\bar{Q}$ increases to infinity and

$$\ln\big(\bar{Q}(t)\big) \sim \lambda_0 t \quad \text{as } t \to \infty;$$

(iii) if $\gamma < 1$, $\bar{Q}$ increases to infinity and

$$\bar{Q}(t) \sim (\lambda_0(1-\gamma))^{\frac{1}{1-\gamma}} \cdot t^{\frac{1}{1-\gamma}} \quad \text{as } t \to \infty.$$

(d) if $x_0 = y_0$, then $\bar{Q}(t) = x_0$ for all $t > 0$.

(4) If $\lambda_0 \le \mu(1-\beta)$, we have $\mathcal{L}^- = \mathbb{R}_+$ which implies starting from $x_0 \ge 0$, $\bar{Q}$ hits 0 at some finite time, and stays at 0 afterward. 0 is the unique equilibrium point.

References

[1] H. Abouee-Mehrizi and O. Baron. State-dependent M/G/1 queueing systems. *Queueing Systems*, 82(1):121–148, 2016.

[2] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, 2003.

[3] E. Bacry, S. Delattre, M. Hoffmann, and J.-F. Muzy. Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 123(7):2475–2499, 2013.

[4] E. Bacry, I. Mastromatteo, and J.-F. Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

[5] R. Bekker, S. C. Borst, O. J. Boxma, and O. Kella. Queues with workload-dependent arrival and service rates. *Queueing Systems*, 46:537–556, 2004.

[6] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 1999.

[7] O. J. Boxma and M. Vlasiou. On queues with service and interarrival times depending on waiting times. *Queueing Systems*, 56(3):121–132, 2007.

[8] H. Chen and D. D. Yao. *Fundamentals of Queueing Networks*. Springer New York, 2001.

[9] X. Chen. Perfect Sampling of Hawkes Processes and Queues with Hawkes Arrivals. *Stochastic Systems*, 11(3):264–283, sep 2021.

[10] X. Chen and G. Hong. Steady-state analysis and online learning for queues with Hawkes arrivals. *arXiv preprint arXiv:2311.02577*, 2023.

[11] J. Chevallier. Fluctuations for mean-field interacting age-dependent Hawkes processes. *Electronic Journal of Probability*, 22:1–49, 2017.

[12] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes Volume I: Elementary Theory and Methods*. Springer, New York, 2nd edition, Aug 2003.

[13] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes Volume II: General Theory and Structure*. Springer New York, 2008.

[14] A. Daw, A. Castellanos, G. B. Yom-Tov, J. Pender, and L. Gruendlinger. The co-production of service: Modeling services in contact centers using Hawkes processes. *Management Science*, to appear, 2024.

[15] A. Daw and J. Pender. Queues driven by Hawkes processes. *Stochastic Systems*, 8(3):192–229, 2018.

[16] R. Durrett. *Probability Theory and Examples*. Cambridge University Press, fifth edition, 2019.

[17] K. Fendick and W. Whitt. Queues with path-dependent arrival processes. *Journal of Applied Probability*, 58(2):484–504, 2021.

[18] K. Fendick and W. Whitt. Heavy traffic limits for queues with non-stationary path-dependent arrival processes. *Queueing Systems*, 101(1):113–135, 2022.

[19] X. Gao and L. Zhu. Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. *Queueing Systems*, 90(1-2):161–206, 2018.

[20] U. Gupta and T. Srinivasa Rao. On the analysis of single server finite queue with state dependent arrival and service processes: M (n)/G (n)/1/K. *Operations-Research-Spektrum*, 20:83–89, 1998.

[21] C. M. Harris. Queues with state-dependent stochastic service rates. *Operations Research*, 15(1):117–130, 1967.

[22] A. G. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443, 1971.

[23] A. G. Hawkes. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198, 2017.

[24] U. Horst and W. Xu. Functional limit theorems for Hawkes processes. *arXiv.2401.11495*, 2024.

[25] N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*. North-Holland/Kodansha, 1989.

[26] J. Jacod and A. Shiryaev. *Limit theorems for stochastic processes*. Springer Science & Business Media, 2003.

[27] O. Kallenberg. *Foundations of Modern Probability*. Springer International Publishing, third edition, 2021.

[28] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer New York, 1998.

[29] Y. Kerner. The conditional distribution of the residual service time in the M_n/G/1 queue. *Stochastic Models*, 24(3):364–375, 2008.

[30] C. Knessl, B. Matkowsky, Z. Schuss, and C. Tier. On the performance of state-dependent single server queues. *SIAM Journal on Applied Mathematics*, 46(4):657–697, 1986.

[31] C. Knessl, C. Tier, B. Matkowsky, and Z. Schuss. A state-dependent GI/G/1 queue. *European Journal of Applied Mathematics*, 5(2):217–241, 1994.

[32] D. T. Koops, M. Saxena, O. J. Boxma, and M. Mandjes. Infinite-server queues with Hawkes input. *Journal of Applied Probability*, 55(3):920–943, 2018.

[33] T. G. Kurtz and P. Protter. Weak limit theorems for stochastic integrals and stochastic differential equations. *Annals of Probability*, 19(3):1035–1070, 1991.

[34] P. J. Laub, Y. Lee, and T. Taimre. *The elements of Hawkes processes*. Springer, 2021.

[35] C. Lee and A. A. Puhalskii. Non-Markovian state-dependent networks in critical loading. *Stochastic Models*, 31(1):43–66, 2015.
[36] B. Legros. Transient analysis of an affine queue-Hawkes process. *Operations Research Letters*, 49(3):393–399, 2021.
[37] B. Li and G. Pang. Heavy-traffic limits for parallel single-server queues with randomly split Hawkes arrival processes. *Journal of Applied Probability*, 61(2):490–514, Aug. 2024.
[38] B. Li and G. Pang. Scaling limits for interactive Hawkes shot noise processes. *Working paper*, 2024.
[39] E. Löcherbach. Spiking neurons: Interacting Hawkes processes, mean field limits and oscillations. *ESAIM: Proceedings and Surveys*, 60:90–103, 2017.
[40] A. Mandelbaum and G. Pats. State-dependent stochastic networks. Part I. Approximations and applications with continuous diffusion limits. *Annals of Applied Probability*, 8(2):569–646, 1998.
[41] H. Mei and J. M. Eisner. The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764, 2017.
[42] M. Morariu-Patrichi and M. S. Pakkanen. State-dependent Hawkes processes and their application to limit order book modelling. *Quantitative Finance*, 22(3):563–583, 2022.
[43] Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27, 1988.
[44] P. E. Protter. *Stochastic Integration and Differential Equations*. Springer Berlin Heidelberg, 2005.
[45] D. Selvamuthu and P. Tardelli. Infinite-server systems with Hawkes arrivals and Hawkes services. *Queueing Systems*, 101(3):329–351, 2022.
[46] H. Tanaka. Stochastic differential equations with reflecting. *Stochastic Processes: Selected Papers of Hiroshi Tanaka*, 9:157, 1979.
[47] A. J. Veretennikov. On strong solutions and explicit formulas for solutions of stochastic integral equations. *Mathematics of the USSR-Sbornik*, 39(3):387–403, 1981.
[48] A. R. Ward and P. W. Glynn. Properties of the reflected Ornstein–Uhlenbeck process. *Queueing Systems*, 44:109–123, 2003.
[49] W. Whitt. Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems*, 6:335–351, 1990.
[50] W. Whitt. *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*. Springer, 2002.
[51] P. Wu, M. Rambaldi, J.-F. Muzy, and E. Bacry. Queue-reactive Hawkes models for the order flow. *arXiv preprint arXiv:1901.08938*, 2019.
[52] K. Yamada. Diffusion approximation for open state-dependent queueing networks in the heavy traffic situation. *Annals of Applied Probability*, 5(4):958–982, 1995.
[53] H. Zhang. On Whitt's conjecture for queues in which service times and interarrival times depend linearly and randomly upon waiting times. *Queueing systems*, 22(3):345–366, 1996.
[54] T.-S. Zhang. On the strong solutions of one-dimensional stochastic differential equations with reflecting boundary. *Stochastic Processes and their Applications*, 50(1):135–147, 1994.