MATHEMATICS OF OPERATIONS RESEARCH
Vol. 41, No. 3, August 2016, pp. 1135–1160
ISSN 0364-765X (print) | ISSN 1526-5471 (online)

http://dx.doi.org/10.1287/moor.2015.0772

A Unified Approach to Diffusion Analysis of Queues with General Patience-Time Distributions

Junfei Huang
Department of Decision Sciences and Managerial Economics, CUHK Business School, The Chinese University of Hong Kong, Hong Kong, junfeih@cuhk.edu.hk

Hanqin Zhang
Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore, bizzhq@nus.edu.sg

Jiheng Zhang
Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology, Hong Kong, China, jiheng@ust.hk

We propose a unified approach to establishing diffusion approximations for queues with impatient customers within a general framework of scaling customer patience time. The approach consists of two steps. The first step is to show that the diffusion-scaled abandonment process is asymptotically close to a function of the diffusion-scaled queue length process under appropriate conditions. The second step is to construct a continuous mapping not only to characterize the system dynamics using the system primitives, but also to help verify the conditions needed in the first step. The diffusion approximations can then be obtained by applying the continuous mapping theorem. The approach has two advantages: (i) it provides a unified procedure to establish the diffusion approximations regardless of the structure of the queueing model or the type of patience-time scaling; and (ii) it makes the diffusion analysis of queues with customer abandonment essentially the same as the diffusion analysis of queues without customer abandonment. We demonstrate the application of this approach via the single-server system with Markov-modulated service speeds in the traditional heavy-traffic regime and the many-server system in the Halfin-Whitt regime and the nondegenerate slowdown regime.

Keywords: customer abandonment; single-server queue; many-server queue; heavy traffic; Halfin-Whitt regime; nondegenerate slowdown regime; diffusion approximation

MSC2000 subject classification: Primary: 60K25; secondary: 68M20, 90B22

OR/MS subject classification: Primary: queues: balking and reneging; secondary: queues: limit theorems, queues: diffusion models

History: Received August 15, 2014; revised April 2, 2015, and September 14, 2015. Published online in *Articles in Advance* March 3, 2016.

1. Introduction. Motivated by its frequent occurrence in many service systems, customer abandonment has been extensively studied in various queueing models. For example, outstanding orders may be canceled in manufacturing industries, data packets may be dropped if the waiting time in the transmission channel is too long, and customers may hang up at a call center after waiting for a while. Abandonment is modeled by assuming each customer (order, data packet, etc.) has a patience time, which is a random variable. A customer abandons the system once his waiting time exceeds his patience time. The study of customer abandonment dates back to Palm [19], who noticed the impatient behavior of telephone switchboard customers. Many studies focus on the diffusion analysis of queueing processes as they often yield tractable and meaningful approximations. This paper aims to provide a unified approach to diffusion analysis with general patience-time distributions.

In the literature, there are two main streams of studies on abandonment that differ by the patience-time scaling. The first one keeps the patience-time distribution fixed in a heavy-traffic regime. This stream can be further classified depending on the assumed heavy-traffic regime. In the conventional heavy-traffic regime, Ward and Glynn [29] identified the diffusion limit as a reflected Ornstein-Uhlenbeck process for the $M/M/1+M$ model. Later, Ward and Glynn [30] extended the result to the general $G/GI/1+GI$ model. In the Halfin-Whitt regime, Garnett et al. [12] obtained the diffusion limit for the $M/M/n+M$ model. Dai et al. [9] extended the diffusion analysis to a more general $G/PH/n+GI$ model by applying a general continuous map to both the fluid and diffusion-scaled processes and the random-time-change theorem. Mandelbaum and Momčilović [18] derived diffusion approximations for the $G/GI/n+GI$ queue building on the work on the $G/GI/n$ queue by Reed [21]. In the nondegenerate slowdown regime (NDS), Atar [1] established the diffusion approximation for the model with Poisson arrivals and exponential service and patience times. Results of all the above studies share the common feature that only the density of the patience-time distribution at the origin plays a role in the diffusion limit.

Based on a statistical study of call center data, however, Zeltyn and Mandelbaum [35] pointed out that the estimate of the hazard-rate function of patience times at a single point often turns out to be unstable. To preserve more information about the patience-time distribution, another stream of the literature scales the patience-time distribution by the hazard rate, rather than by the density at a single point. Reed and Ward [24] obtained the diffusion

TABLE 1. Diffusion approximations for systems with abandonment.

	No scaling	With scaling
Conventional	Ward and Glynn [29] $M/M/1+M$ Ward and Glynn [30] $G/GI/1+GI$ Lee and Weerasinghe [16] $G/GI/1+GI$	Reed and Ward [24] $G/GI/1+GI$ Lee and Weerasinghe [16] $G/GI/1+GI$
NDS	Atar [1] $M/M/n^{\alpha}+M$	
Halfin-Whitt	Garnett et al. [12] $M/M/n+M$ Dai et al. [9] $G/PH/n+GI$ Mandelbaum and Momčilović [18] $G/GI/n+GI$	Reed and Tezcan [23] $G/M/n+GI$ Weerasinghe [31] $G/M/n+GI$ Katsuda [13] $G/PH/n+GI$

approximations for both the offered waiting-time process and the queue length process for the $G/GI/1+GI$ model in the conventional heavy-traffic regime. Their approach was to use a nonlinear generalized regulator mapping to establish weak convergence results. Taking advantage of the memoryless property of exponential distributions, recently, Reed and Tezcan [23] applied the same hazard-rate scaling to study the diffusion limit of the queue length process for the $G/M/n+GI$ model, which was extended by Weerasinghe [31] to allow a state-dependent service rate. Katsuda [13], again by taking advantage of the memoryless property, extended the service time to be phase-type and allowed patience times to be more general. The extensive numerical experiments of Reed and Tezcan [23] showed that the approximations involving the entire hazard-rate function outperformed those that relied only on the density at the origin when the density of the patience-time distribution changes rapidly near the origin. Table 1 summarizes the existing studies on the diffusion analysis of queueing systems by classifying them into three heavy-traffic regimes and two scalings of the patience-time distribution. Readers are referred to Ward [28] for a comprehensive survey on the study of customer abandonment both without scaling and with hazard-rate scaling of patience times.

Based on the intuition developed by Reed and Ward [24], recently Dai and He [8] proposed a neat approximation for the scaled abandonment process when the service time distribution is generalized from exponential to phase-type. The approximation is expressed as an integral whose integrand is just the hazard rate function and the integral limit is given by the diffusion approximation for the number of customers in the system. Numerically, they showed that their approximation is remarkably accurate. *But one would hope to see a rigorous proof of their proposed approximation for phase-type service times. Furthermore, it would be interesting to build the diffusion approximation for $G/GI/n+GI$ with hazard rate scaling of the patience-time distribution.*

From the methodological perspective, the above-mentioned works are about different models and set in different heavy-traffic regimes (see Table 1). The analysis for single-server queues in the conventional heavy-traffic regime and that for many-server queues in the Halfin-Whitt regime and the NDS regime require different methods. For example, Ward and Glynn [30] used the virtual waiting time for the single-server queues while Mandelbaum and Momčilović [18] relied on the analysis of the queue length process for $G/GI/n+G$ queues in the Halfin-Whitt regime; in contrast with these two papers, however, Atar [1] directly constructed a Poisson process to represent the abandonment process by taking advantage of the memoryless property of the exponential patience-time distribution for $M/M/n^{\alpha}+M$ in the NDS regime. Moreover, for patience time with and without scaling, the methods are quite different even in the same regime. In the Halfin-Whitt regime, for instance, when considering $G/GI/n+G$ without scaling, Mandelbaum and Momčilović [18] constructed an auxiliary system with which to analyze the queue length process of the original system; while considering $G/M/n+G$ with scaling, Reed and Tezcan [23], and Weerasinghe [31] directly proved the asymptotic equivalence between the queue length process and the virtual waiting-time process to obtain the diffusion limit of the queue length process. *It would be nice to have a unified approach that applies across different regimes, and that can treat the patience time with or without scaling.*

Motivated by the above problems, our goal in this paper is to provide a uniform approach to the diffusion analysis of single-server queues and many-server queues with and without hazard-rate scaling. The framework for modeling the patience-time distribution described in (4) can cover no scaling, hazard-rate scaling, and several other types of scalings, which can potentially be used to analyze customer abandonment behaviors. We focus on the unified approach in establishing the diffusion limits under the general scaling for the customer patience time (4). The approach has two steps. The first step is to identify an asymptotic relationship between the customer abandonment process and the queue length process in Theorem 1 based on the general scaling (4) for the patience-time distribution. When (4) is specialized to the case without scaling, our result reduces to that of Dai and He [7]. Such an asymptotic relationship is established by using the patience-time distribution to connect the abandonment process to the virtual waiting time process, which can be approximated with the queue length process by proving a generalization of Little's law. The challenge caused by the general scaling (4) is that the queue length processes are required to be tight, whereas only stochastic boundedness is needed for the case without scaling as in Dai and He [7]. Tightness, in particular, the modulus of continuity (7), is usually difficult to verify. To tackle this challenge, we establish the tightness of the abandonment processes based only on the stochastic boundedness of the queue length processes, which forms a part of Theorem 1. Having tightness of the abandonment processes allows us to verify the tightness of the queue length processes via the second step of our approach. The second step is to construct a mapping, which would reveal a functional relationship between the system status (such as the queue length process) and the stochastic primitives (such as the arrival process, service, and patience times). The mapping, with some nice properties, not only helps to verify the tightness of the queue length processes required by Theorem 1, but also provides diffusion analysis by applying the continuous mapping theorem. Within the unified framework described in the above two steps, to develop diffusion analysis for queueing systems with abandonment, it is enough to construct such continuous mappings and verify some mild assumptions. Those assumptions can be verified in the same way as that for systems without abandonment.

We demonstrate how to use our approach to establish diffusion approximations via three examples, which are all new results in the literature. In the first example (§3.1), we study the single-server queue with Markov-modulated service speeds in the traditional heavy-traffic regime. See Mahabhashyam and Gautam [17] and Takine [26] for a wide range of applications of such models in telecommunications and Web servers. The classical single-server queue studied by Ward and Glynn [30] and Reed and Ward [24] can be viewed as a special case where the service speed is constant. In the second example (§3.2), we establish the diffusion approximations for many-server queues in the Halfin-Whitt regime with general service times. The special case of no scaling is the result in Mandelbaum and Momčilović [18] and the special case with exponential service times and scaling is the result in Reed and Tezcan [23] and Weerasinghe [31]. Moreover, the diffusion approximation established here justifies the approximation of the scaled abandonment processes proposed by Dai and He [8]. In the third example (§3.3), we study the many-server queues in the NDS regime by extending the work of Atar [1] to general patience-time distribution. These three examples shows that the advantage of our unified approach is to simplify the diffusion analysis of queues with customer abandonment by making it essentially the same as the diffusion analysis of queues without customer abandonment.

The rest of this paper is organized as follows. We introduce our unified approach in §2, but postpone the proof to §4. Section 3 demonstrates the application of the unified approach. In particular, we consider three systems: the $G/GI/1+GI$ queue with Markov-modulated service speeds in the conventional heavy-traffic regime in §3.1, the $G/GI/n+GI$ queue in the Halfin-Whitt regime in §3.2, and the $G/M/n^{\alpha}+GI$ queue with $\alpha \in (0,1)$ in the NDS regime in §3.3. Several technical proofs are presented in the appendix.

Before we conclude this section, we introduce some notation and definitions that are used throughout the paper. All random variables and processes are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ unless otherwise specified. We denote by $\mathbb{Z}_+$, $\mathbb{R}$, and $\mathbb{R}_+$ the sets of positive integers, real numbers, and nonnegative numbers, respectively. The space of right continuous functions with left-hand limits on $\mathbb{R}_+$ taking values in $\mathbb{R}$ is denoted by $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$, and the subspace of the continuous functions in $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ is denoted by $\mathbf{C}(\mathbb{R}_+, \mathbb{R})$. The space $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ is assumed to be endowed with the Skorohod J_1-topology (see Billingsley [3]). For $g \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$, $g(t-)$ represents its left limit at $t > 0$, and the uniform norm of $g(\cdot)$ on the interval $[a, b]$ is defined as

$$\|g\|_{[a,b]} = \sup_{t\in[a,b]} |g(t)| \quad \text{with } \|g\|_{[0,b]} \text{ abbreviated to } \|g\|_b.$$

For a sequence of random elements $\{X^n, n \in \mathbb{Z}_+\}$ taking values in a metric space, we write $X^n \Rightarrow X$ to denote the convergence of X^n to X in distribution. $X \overset{d}{=} Y$ means that random elements X and Y have the same distribution. For $a \in \mathbb{R}$, $a^+ = \max\{a, 0\}$, $a^- = \max\{-a, 0\}$, and $\lfloor a \rfloor$ is the largest integer not greater than a. We use $\mathbf{1}_A$ to denote the indicator function of set $A \subset \Omega$.

2. Model and asymptotic framework.

Consider a sequence of first-come first-served (FCFS) $G/GI/N_n+GI$ queues indexed by $n \in \mathbb{Z}_+$, where N_n is deterministic and represents the number of servers in the nth system. Denote by $Q^n(t)$ the number of customers in the queue at time t, by $X^n(t)$ the total number of customers in the system at time t, and by $G^n(t)$ the number of customers who have abandoned the queue by time t, in the nth system. In this paper, we assume, for technical convenience, that the patience times of the customers who are initially in the system are infinite, i.e., the initial customers in the queue are infinitely patient (this assumption is not restrictive; see Mandelbaum and Momčilović [18] for the study on the many-server queue). Clearly, $G^n(0) = 0$ and $Q^n(0)$ is the number of customers waiting in the queue at time zero. Define the diffusion-scaled processes $\tilde{Q}^n = \{\tilde{Q}^n(t): t \ge 0\}$, $\tilde{X}^n = \{\tilde{X}^n(t): t \ge 0\}$, and $\tilde{G}^n = \{\tilde{G}^n(t): t \ge 0\}$ as

$$\tilde{Q}^n(t) = \frac{Q^n(t)}{\sqrt{n}}, \qquad \tilde{X}^n(t) = \frac{X^n(t) - N_n}{\sqrt{n}}, \qquad \tilde{G}^n(t) = \frac{G^n(t)}{\sqrt{n}}. \tag{1}$$

Our objective in this section is to prove an asymptotic relationship (Theorem 1) between $\tilde{Q}^n$ and $\tilde{G}^n$ under appropriate assumptions.

Let $E^n(t)$ denote the number of arrivals by time t in the nth system, and define the diffusion-scaled arrival process $\tilde{E}^n = \{\tilde{E}^n(t)\colon t \geq 0\}$ by

$$\tilde{E}^n(t) = \frac{E^n(t) - \lambda^n t}{\sqrt{n}},$$

where λ^n is called customer arrival rate for the nth system and satisfies

$$\lim_{n\to\infty} \frac{\lambda^n}{n} = \mu > 0. \tag{2}$$

We assume that

$$\tilde{E}^n \Rightarrow \tilde{E} \quad \text{as } n \to \infty \tag{3}$$

for some process $\tilde{E} = \{\tilde{E}(t)\colon t \geq 0\} \in \mathbf{C}(\mathbb{R}_+, \mathbb{R})$. Here, μ in (2) is usually related to customer service times. The customer service times (characterized by customer service requirements and system service speed to process the requirements) will be specified when a concrete system is investigated. Let γ_i^n be the patience time of the ith arriving customer in the nth system. A customer waiting in the system will leave without receiving service once his patience time is exhausted. $\{\gamma_i^n, i \in \mathbb{Z}_+\}$ is assumed to be a sequence of i.i.d. random variables, and independent of the arrival process E^n for each n. We denote the patience-time distribution by $F^n(\cdot)$ and assume that for each $x \geq 0$,

$$\sqrt{n} F^n\left(\frac{x}{\sqrt{n}}\right) \to f(x) \quad \text{as } n \to \infty, \tag{4}$$

where $f(\cdot)$ is nondecreasing. We assume that $f(\cdot)$ is locally Lipschitz continuous function, i.e., for any $T \geq 0$, there is a constant Λ_T such that for all $x, y \in [0, T]$,

$$|f(x) - f(y)| \leq \Lambda_T |x - y|. \tag{5}$$

As pointed out in the introduction, not only can this framework cover the two well-known ways of scaling patience-time distributions, namely, no scaling and hazard-rate scaling, but also provides some new types of scalings as follows:

(a) *No scaling.* Let $F^n(x) = F(x)$ for $x \geq 0$, where $F(\cdot)$ is a probability distribution function with $F(0) = 0$ and $F'(0+) = \alpha$. In this case, $f(x) = \alpha x$ for $x \geq 0$.

(b) *Hazard-rate scaling.* Let $F^n(x) = 1 - \exp(-\int_0^x h(\sqrt{n}s)\, ds)$ for $x \geq 0$, for some locally Lipschitz-continuous hazard-rate function $h(\cdot)$. In this case, $f(x) = \int_0^x h(s)\, ds$ for $x \geq 0$.

(c) *Mixture of hazard-rate scaling and no scaling.* For any give $p \in (0, 1)$, let $F(\cdot)$ be a distribution function and $h(\cdot)$ be a locally Lipschitz-continuous hazard-rate function. Let $F^n(x) = pF(x) + (1-p)[1 - \exp(-\int_0^x h(\sqrt{n}s)\, ds)]$, $x \geq 0$. In this case, $f(x) = pF'(0+)x + (1-p)\int_0^x h(s)\, ds$, $x \geq 0$.

(d) *Delayed hazard-rate scaling.* Let $h_1(\cdot)$ and $h_2(\cdot)$ be two locally Lipschitz-continuous hazard-rate functions, and let

$$F^n(x) = \begin{cases} 1 - \exp\left(-\displaystyle\int_0^x h_1(s)\, ds\right) & \text{if } x \in \left[0, \dfrac{x_0}{\sqrt{n}}\right], \\ 1 - \exp\left(-\displaystyle\int_0^{x_0/\sqrt{n}} h_1(s)\, ds - \int_{x_0/\sqrt{n}}^x h_2(\sqrt{n}s)\, ds\right) & \text{if } x \in \left(\dfrac{x_0}{\sqrt{n}}, \infty\right), \end{cases}$$

where x_0 is a positive constant, is usually called delayed time point. Then,

$$f(x) = \begin{cases} h_1(0)x & \text{if } 0 \leq x \leq x_0; \\ h_1(0)x_0 + \displaystyle\int_{x_0}^x h_2(s)\, ds & \text{if } x > x_0. \end{cases}$$

To obtain the asymptotic relationship (Theorem 1), the key assumption is that the sequence of diffusion-scaled queue length processes $\{\tilde{Q}^n, n \in \mathbb{Z}_+\}$ is C-tight. That is, on any finite interval $[0, T]$, the sequence is stochastically bounded, i.e.,

$$\lim_{\Gamma\to\infty} \limsup_{n\to\infty} \mathbb{P}\left\{\sup_{0\leq t\leq T} \tilde{Q}^n(t) > \Gamma\right\} = 0, \tag{6}$$

and the modulus of continuity is asymptotically small, i.e., for any $\varepsilon > 0$,

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}\left\{\sup_{s,\, t\in[0,\, T],\, |s-t|<\delta} |\tilde{Q}^n(s) - \tilde{Q}^n(t)| > \varepsilon\right\} = 0. \tag{7}$$

THEOREM 1. *If a sequence of $G/GI/N_n + GI$ queues satisfies* (2)–(4) *and* (6), *then the sequence* $\{\tilde{G}^n, n \in \mathbb{Z}_+\}$ *is* C*-tight. Moreover, when* (5) *and* (7) *also hold, we have that for each* $T > 0$,

$$\sup_{0\le t\le T}\left|\tilde{G}^n(t) - \mu\int_0^t f\left(\frac{1}{\mu}\tilde{Q}^n(s)\right)ds\right| \Rightarrow 0 \quad \text{as } n\to\infty. \tag{8}$$

REMARK 1. Note that C-tightness of $\{\tilde{G}^n, n\in\mathbb{Z}_+\}$ implies that the fluid-scaled process $(1/n)G^n(\cdot)$ converges to zero in probability, which is the fluid limit result for the abandonment process. Due to this result, the abandonment process is negligible in fluid scaling, hence analyzing the fluid limit of the system with abandonment is essentially the same as analyzing the fluid limit of the system without abandonment.

Theorem 1 does not need any condition on the service process as long as the queue length processes satisfy (6)–(7). Whether the patience times have hazard-rate scaling or no scaling, the theorem yields the following result.

COROLLARY 1. *Assume that the sequence of $G/GI/N_n + GI$ queues satisfies* (2)–(7). (i) *If the patience-time distribution has a hazard-rate scaling, namely,* $F^n(x) = 1 - \exp(-\int_0^x h(\sqrt{n}s)\,ds)$ *for some locally bounded hazard-rate function* $h(\cdot)$, *then*

$$\sup_{0\le t\le T}\left|\tilde{G}^n(t) - \int_0^t\int_0^{\tilde{Q}^n(s)} h\left(\frac{u}{\mu}\right)du\,ds\right| \Rightarrow 0 \quad \text{as } n\to\infty; \tag{9}$$

(ii) *If the patience-time distribution has no scaling, that is,* $F^n(x) = F(x)$ *with derivative* $\alpha = F'(0+)$, *then*

$$\sup_{0\le t\le T}\left|\tilde{G}^n(t) - \alpha\int_0^t \tilde{Q}^n(s)\,ds\right| \Rightarrow 0 \quad \text{as } n\to\infty. \tag{10}$$

Corollary 1 (ii) is the same as Theorem 2.1 of Dai and He [7] who obtained such asymptotic relationship when the patience time is not scaled and only (2)–(6) hold. However, due to the general scaling (4) for patience-time distributions, we need the additional condition (7) to deal with the nonlinearity of the function $f(\cdot)$.

The independence of specific queueing models for Theorem 1 enables us to develop a unified approach to diffusion analysis. Note that among conditions required by Theorem 1, (2)–(5) are standard for the system parameters. The applicability of Theorem 1 often depends on the verification of conditions (6)–(7), in particular (7), which is often a major difficulty in most queueing analysis. So we now provide a continuous mapping technique as the second step of our unified approach to overcome the difficulty associated with the verification of condition (7) on the queue length processes. This, consequently, leads to the diffusion approximations by the continuous mapping theorem.

To establish (7) on the modulus of continuity for $\{\tilde{Q}^n, n\in\mathbb{Z}_+\}$, in view of $\tilde{Q}^n = (\tilde{X}^n)^+$, it is sufficient to consider the modulus of continuity for $\{\tilde{X}^n, n\in\mathbb{Z}_+\}$. To this end, in view of Theorem 1, we define the centered abandonment process $\tilde{G}^n_c = \{\tilde{G}^n_c(t)\colon t\ge 0\}$ as

$$\tilde{G}^n_c(t) = \tilde{G}^n(t) - \mu\int_0^t f\left(\frac{1}{\mu}(\tilde{X}^n(s))^+\right)ds. \tag{11}$$

Suppose there exists a sequence of processes $\tilde{Y}^n = \{\tilde{Y}^n(t)\colon t\ge 0\}$ and a mapping $\Phi\colon \mathbf{D}(\mathbb{R}_+,\mathbb{R}) \to \mathbf{D}(\mathbb{R}_+,\mathbb{R})$ such that

$$\tilde{X}^n = \Phi(\tilde{Y}^n - \tilde{G}^n_c). \tag{12}$$

Roughly speaking, $\tilde{Y}^n$ is the centered diffusion-scaled process of the system primitives. Its exact form depends on the specific queueing system under examination. See (18), (30), and (46) for the expressions of $\tilde{Y}^n$ in the three concrete models studied in §3. The following result characterizes the asymptotic behavior of $\tilde{X}^n$.

THEOREM 2. *Assume that* (i) *condition* (6) *on* $\{\tilde{Q}^n, n\in\mathbb{Z}_+\}$ *holds*; (ii) *there exists* $\tilde{Y}\in\mathbf{C}(\mathbb{R}_+,\mathbb{R})$ *such that* $\tilde{Y}^n \Rightarrow \tilde{Y}$ *as* $n\to\infty$; (iii) *the mapping* $\Phi(\cdot)$ *is Lipschitz continuous in the topology of uniform convergence over bounded intervals, measurable with respect to the Borel σ-field generated by the Skorohod J_1-topology, and* $\Phi(\mathbf{C}(\mathbb{R}_+,\mathbb{R}))\subseteq\mathbf{C}(\mathbb{R}_+,\mathbb{R})$. *Then,*

$$\tilde{X}^n \Rightarrow \tilde{X} = \Phi(\tilde{Y}) \quad \text{as } n\to\infty. \tag{13}$$

The proofs of Theorems 1–2, and Corollary 1 are postponed to §4. In their proofs, we can see that Theorem 1 is a key step to prove Theorem 2. Theorem 2 outlines our unified approach in more detail. We first obtain the stochastic boundedness for $\{\tilde{Q}^n, n\in\mathbb{Z}_+\}$ by a comparison with the systems without customer abandonment. Then, we construct the continuous mapping $\Phi(\cdot)$. The principle of the construction of $\Phi(\cdot)$ is to make the weak convergence of $\{\tilde{Y}^n, n\in\mathbb{Z}_+\}$ to be tractable, which can usually be established by going along the same way as in the systems without abandonment. Hence the approach developed here makes the diffusion analysis of queues with customer abandonment to be essentially the same as the diffusion analysis of queues without customer abandonment.

Next, we apply this approach to the diffusion analysis for the $G/GI/N_n + GI$ queue.

3. Diffusion analysis for $G/GI/N_n+GI$.

The setup for the sequence of the $G/GI/N_n+GI$ systems is as follows. For the nth system, let v_i^n, $i=1,2,\dots$ be the service requirement of the ith customer who arrives at the system after time 0 and will not abandon. For $i=-X^n(0)+1,\dots,-Q^n(0)$, v_i^n denotes the remaining service requirement of the ith customer initially in service. For $i=-Q^n(0)+1,\dots,0$, v_i^n denotes the service requirement of the ith customer initially waiting in the queue. Customer $-Q^n(0)+1$ is the first in the queue, customer $-Q^n(0)+2$ is the second, and so on. We assume $\{v_i^n, i\ge -X^n(0)+1\}$ is a sequence of independent random variables, and is independent of the patience times $\{\gamma_i^n, i\in\mathbb{Z}_+\}$ and the arrival process E^n given in §2 for each n. We assume the convergence of initial states,

$$\tilde{X}^n(0)\Rightarrow \xi \quad \text{as } n\to\infty \tag{14}$$

for some random variable ξ. We also assume the following *heavy-traffic condition*,

$$\beta^n := \sqrt{n}\left(\frac{\lambda^n}{n\mu}-1\right)\to\beta \quad \text{as } n\to\infty \tag{15}$$

for some $\beta\in\mathbb{R}$. In particular, the heavy-traffic condition implies (2). Our diffusion approximation results will be established in the heavy-traffic regime specified by (15) with assumption (3) on the arrival processes, assumptions (4)–(5) on the patience-time distribution, and the initial condition (14). The relationship between μ in the heavy-traffic condition (15) and the means of the customer service requirements $\{v_i^n, i\ge -X^n(0)+1\}$ will be characterized through the system service speed of processing the service requirements in the concrete models, see Assumptions 1–3.

3.1. $G/GI/1+GI$ in the traditional heavy-traffic regime.

In this subsection, we study a sequence of single-server queues in the traditional heavy-traffic regime. We adopt a general model to allow the service speed in the nth system to be modulated by a continuous-time Markov chain $\Delta^n=\{\Delta^n(t)\colon t\ge 0\}$ with a finite-state space $\mathcal{S}=\{1,\dots,l\}$. At time t, the server will process customer service requirements at speed $n\mu_i$ when $\Delta^n(t)=i\in\mathcal{S}$. This is a general model as the classical single-server queue is a special case where the state space has only a single state, i.e., the service speed is constant. The following setup for the model is standard (see Dorsman et al. [10]).

Assumption 1. *The customer service requirements $\{v_i^n, i\ge -X^n(0)+1\}$ are independent and identically distributed with mean 1 and variance θ^2. The Markov chain $\{\Delta^n(t)\colon t\ge 0\}$ is given by $\Delta^n(t)=\Delta(nt)$ for $t\ge 0$, where $\Delta=\{\Delta(t)\colon t\ge 0\}$ is an irreducible and stationary continuous-time Markov chain with state space $\mathcal{S}$, generator $\mathcal{G}$ ($(l\times l)$-matrix), and stationary distribution $\pi:=(\pi_1,\dots,\pi_l)$.*

We assume μ in (2) is equal to $\sum_{i\in\mathcal{S}}\pi_i\mu_i$, which can be considered as the long-run average speed at which the server processes service requests. The following preliminary result on continuous Markov chains will be needed in establishing the diffusion approximation for the $G/GI/1+GI$ with Markov-modulated service speeds. For its proof, see Yin and Zhang [34].

Lemma 1. *Under Assumption 1, for any $T\ge 0$, as $n\to\infty$,*

$$\sup_{0\le t\le T}\left|\int_0^t \mu_{\Delta^n(s)}\,ds-\sum_{i\in\mathcal{S}}\mu_i\pi_i t\right|\Rightarrow 0 \qquad \text{and} \qquad \tilde{\Delta}^n\Rightarrow\tilde{\Delta},$$

where $\tilde{\Delta}^n=\{\tilde{\Delta}^n(t)\colon t\ge 0\}$ given by

$$\tilde{\Delta}^n(t)=\sqrt{n}\left(\int_0^t \mu_{\Delta^n(s)}\,ds-\sum_{i\in\mathcal{S}}\mu_i\pi_i t\right),$$

and $\tilde{\Delta}=\{\tilde{\Delta}(t)\colon t\ge 0\}$ is a Brownian motion with zero drift, and variance $\theta_{\mathcal{S}}^2$ given by

$$\theta_{\mathcal{S}}^2=\sum_{i,\,j\in\mathcal{S}}\mu_i\mu_j\left(\pi_i\int_0^\infty\varphi_{ij}(s)\,ds+\pi_j\int_0^\infty\varphi_{ji}(s)\,ds\right)$$

with $(\varphi_{ij}(s))_{l\times l}=(I-(1,\dots,1)'\cdot\pi)\cdot\exp(\mathcal{G}s)$, where I is an l-by-l identity matrix.

The total amount of customer service requests processed by the server during $[0,t)$ is $\int_0^t n\mu_{\Delta^n(s)}\cdot(X^n(s)\wedge 1)\,ds$. Define

$$S^n(t)=\max\{k\colon v^n_{-X^n(0)+1}+\cdots+v^n_{-X^n(0)+k}\le t\}$$

as the renewal process associated with the sequence of service requirements. Then, the number of customers served by time t is

$$S^n\left(n\int_0^t \mu_{\Delta^n(s)}\cdot(X^n(s)\wedge 1)\,ds\right).$$

The evolution of the process X^n can be characterized by the system dynamics equation

$$X^n(t)=X^n(0)+E^n(t)-S^n\left(n\int_0^t \mu_{\Delta^n(s)}\cdot(X^n(s)\wedge 1)\,ds\right)-G^n(t). \tag{16}$$

In view of (1) and (11), we rewrite (16) as

$$\tilde{X}^n(t)=\tilde{Y}^n(t)-\tilde{G}^n_c(t)-\mu\int_0^t f\left(\frac{1}{\mu}(\tilde{X}^n(s))^+\right)ds+n\int_0^t \mu_{\Delta^n(s)}(\tilde{X}^n(s))^-\,ds, \tag{17}$$

where

$$\tilde{Y}^n(t)=\tilde{X}^n(0)+\tilde{E}^n(t)-\tilde{S}^n\left(\int_0^t \mu_{\Delta^n(s)}\cdot(X^n(s)\wedge 1)\,ds\right)+\mu\sqrt{n}\left(\frac{\lambda^n}{n\mu}-1\right)t-\tilde{\Delta}^n(t), \tag{18}$$

$$\tilde{S}^n(t)=\frac{S^n(nt)-nt}{\sqrt{n}}.$$

To use Theorem 2, we first establish the following lemma.

Lemma 2. *Assume that $g(\cdot)$ is a locally Lipschitz-continuous function defined on $\mathbb{R}_+$ with $g(0)=0$. For any $y(\cdot)\in\mathbf{D}(\mathbb{R}_+,\mathbb{R})$, there exists a unique solution $(x(\cdot),z(\cdot))$ to the following set of equations*:

$$x(t)=y(t)+\int_0^t g((x(s))^+)\,ds+z(t), \tag{19}$$

$$x(t)\ge 0,$$

$$z(\cdot)\ \textit{is nondecreasing and}\ z(0)=0,$$

$$\int_0^\infty x(s)\,dz(s)=0.$$

Moreover, the mapping $\Phi_g(\cdot)\colon \mathbf{D}(\mathbb{R}_+,\mathbb{R})\to\mathbf{D}(\mathbb{R}_+,\mathbb{R})$ defined by $x=\Phi_g(y)$ is Lipschitz continuous in the topology of uniform convergence over bounded intervals, measurable with respect to the Borel σ-field generated by the Skorohod J_1-topology, and $\Phi_g(\mathbf{C}(\mathbb{R}_+,\mathbb{R}))\subseteq\mathbf{C}(\mathbb{R}_+,\mathbb{R})$.

It is worth pointing out that the mapping is continuous in the Skorohod J_1-topology according to Proposition 4.9 of Lee and Weerasinghe [16]. Their proof is based on the earlier work of Reed and Ward [24] and some classical results of the Skorohod J_1-topology. In Appendix A, we provide a simple and direct way to show the Lipschitz continuity under the uniform topology and demonstrate that this is sufficient for our reflection mapping approach.

Theorem 3. *Assume that conditions (3)–(5) and (14)–(15) hold. For the stochastic processes $\{\tilde{X}^n, n\in\mathbb{Z}_+\}$ associated with the sequence of $G/GI/1+GI$ systems, if Assumption 1 holds, then $\tilde{X}^n\Rightarrow\tilde{X}$ as $n\to\infty$. Here, $\tilde{X}=\{\tilde{X}(t)\colon t\ge 0\}$ is given by $\tilde{X}=\Phi_g(\tilde{Y})$ with*

$$g(t)=-\mu f\left(\frac{1}{\mu}t\right),\qquad t\ge 0;$$

$$\tilde{Y}=\{\tilde{Y}(t)\colon t\ge 0\},\qquad \tilde{Y}(t)=\xi+\tilde{E}(t)-\sqrt{\mu}\theta\tilde{S}(t)-\tilde{\Delta}(t)+\beta\mu t,$$

where $\tilde{S}=\{\tilde{S}(t)\colon t\ge 0\}$ is a standard Brownian motion independent of ξ, $\tilde{E}$, and $\tilde{\Delta}$. Moreover, $\tilde{Q}^n\Rightarrow\tilde{X}$ as $n\to\infty$.

Proof. By (17), we have

$$(\tilde{X}^n(t))^+=\tilde{Y}^n(t)+(\tilde{X}^n(t))^--\tilde{G}^n(t)+n\int_0^t \mu_{\Delta^n(s)}(\tilde{X}^n(s))^-\,ds. \tag{20}$$

Note that

$$\sup_{0\le t\le T}\left|\tilde{S}^n\left(\int_0^t \mu_{\Delta^n(s)}\cdot(X^n(s)\wedge 1)\,ds\right)\right|\le\sup_{0\le t\le\max_{i\in\mathcal{S}}\{\mu_i T\}}|\tilde{S}^n(t)|.$$

By (3), (14)–(15), Lemma 1 and the definition of $\tilde{Y}^n$ in (18), $\{\tilde{Y}^n, n\in\mathbb{Z}_+\}$ is stochastically bounded. It is clear, by $(\tilde{X}^n(t))^-\le 1/\sqrt{n}$, that $\{(\tilde{X}^n)^-, n\in\mathbb{Z}_+\}$ is also stochastically bounded. In view of $\tilde{Y}^n(t)+(\tilde{X}^n(t))^--\tilde{G}^n(t)\le$

$\tilde{Y}^n(t)+(\tilde{X}^n(t))^-$ and (20), it follows from Lemma 4.1 in Kruk et al. [15] that $(\tilde{X}^n)^+$ can be bounded by the one-dimensional Skorohod mapping of $\tilde{Y}^n+(\tilde{X}^n)^-$. Hence, $\{(\tilde{X}^n)^+, n\in\mathbb{Z}_+\}$ is stochastically bounded. This, consequently, implies that for any $T\ge 0$, as $n\to\infty$,

$$\sup_{0\le t\le T}\frac{(X^n(t)-1)^+}{n}\Rightarrow 0. \tag{21}$$

We now prove the convergence of the sequence $\{\tilde{Y}^n, n\in\mathbb{Z}_+\}$. It follows from the above stochastic boundedness analysis that for any $T\ge 0$ as $n\to\infty$,

$$\sup_{0\le t\le T}\frac{\tilde{Y}^n(t)+(\tilde{X}^n(t))^- - \tilde{G}^n(t)}{\sqrt{n}}\Rightarrow 0. \tag{22}$$

By (20)–(22), as $n\to\infty$,

$$\sup_{0\le t\le T}\int_0^t \mu_{\Delta^n(s)}(X^n(s)-1)^-\,ds\Rightarrow 0.$$

The above limit together with Lemma 1 implies that as $n\to\infty$,

$$\sup_{0\le t\le T}\left|\int_0^t \mu_{\Delta^n(s)}\cdot(X^n(s)\wedge 1)\,ds-\mu t\right|\Rightarrow 0. \tag{23}$$

Applying the above limit and the random-time-change theorem (Corollary 1 of Whitt [32]), the sequence of processes $\{\tilde{S}^n(\int_0^t \mu_{\Delta^n(s)}\cdot(X^n(s)\wedge 1)\,ds)\colon t\ge 0\}$ converges in distribution to $\sqrt{\mu}\theta\tilde{S}$ with $\{\tilde{S}(t)\colon t\ge 0\}$ being a Brownian motion. It follows from conditions (3), (14)–(15), and Lemma 1 that

$$\tilde{Y}^n\Rightarrow\tilde{Y}, \tag{24}$$

where $\tilde{Y}=\{\tilde{Y}(t)\colon t\ge 0\}$ with $\tilde{Y}(t)=\xi+\tilde{E}(t)-\sqrt{\mu}\theta\tilde{S}(t)-\tilde{\Delta}(t)+\beta\mu t$. We have so far verified conditions (i) and (ii) in Theorem 2. Condition (iii) follows directly from (17) and Lemma 2. This completes the proof. □

Remark 2. For the classical $G/GI/1+GI$, the limit $\tilde{\Delta}$ in Lemma 1 becomes 0 since $\mathscr{S}$ contains only a single state. Theorem 3, consequently, gives the weak convergence for the queue length process of the classical $G/GI/1+GI$ considered by Ward and Glynn [30] and Reed and Ward [24].

3.2. $G/GI/n+GI$ in the Halfin-Whitt regime.

In this subsection, we apply our unified approach to establishing the diffusion approximation for $G/GI/n+GI$, where customer service requests are assumed to be processed by each server at speed 1 without loss of generality. Since we use the general scaling (4), our result covers the case in Mandelbaum and Momčilović [18] where patience times are not scaled, and the case in Reed and Tezcan [23] where hazard-rate scaling is applied to the patience times. We also generalize the latter work to a generally distributed service requirement.

Let $H_e(\cdot)$ denote the equilibrium distribution associated with the distribution $H(\cdot)$ of customer service requirements, i.e.,

$$H_e(x)=\mu\int_0^x(1-H(s))\,ds,\quad x\ge 0.$$

Thus the renewal function of the delayed renewal process with initial distribution $H_e(\cdot)$ and inter-renewal distribution $H(\cdot)$ is μt. The following assumption on the service process is required for this example.

Assumption 2. *The customer service requirements $\{v_i^n, i\ge -Q^n(0)+1\}$ are independent and identically distributed with distribution function $H(\cdot)$, which has mean $1/\mu$ and variance θ^2. The remaining service requirements of the customers who are initially in service, $\{v_i^n, -X^n(0)+1\le i\le -Q^n(0)\}$, are independent and identically distributed with distribution function $H_e(\cdot)$. Moreover, the two sequences are independent.*

Let $D^n(t)$ be the number of customers whose service requirements have been completed by time t in the nth system. We then have the following simple balance equation for the total number of customers in the nth system at time t:

$$X^n(t)=X^n(0)+E^n(t)-D^n(t)-G^n(t). \tag{25}$$

Let $M(\cdot)$ denote the renewal function associated with $\{v_i^n, i\ge -Q^n(0)+1\}$, i.e., $M(\cdot)$ satisfies the following renewal equation:

$$M(t)=H(t)+\int_0^t H(t-s)\,dM(s). \tag{26}$$

Define

$$D_c^n(t) = D^n(t) - n\mu t - (X^n(0)-n)^- \cdot (M(t)-\mu t) + \int_0^t (X^n(t-s)-n)^- \, dM(s), \tag{27}$$

$$\tilde{D}_c^n(t) = \frac{D_c^n(t)}{\sqrt{n}}.$$

The idea of (27), which follows Equation (33) in Reed and Shaki [22], is to center the service completion process using the renewal function $M(\cdot)$. Then, (25) becomes

$$X^n(t) = X^n(0) + E^n(t) - G^n(t) - D_c^n(t) - n\mu t - (X^n(0)-n)^- \cdot (M(t)-\mu t) + \int_0^t (X^n(t-s)-n)^- \, dM(s). \tag{28}$$

Applying diffusion scaling (1) to (28) implies that

$$\tilde{X}^n(t) = \tilde{Y}^n(t) - \tilde{G}_c^n(t) + \int_0^t (\tilde{X}^n(t-s))^- \, dM(s) - \mu \int_0^t f\left(\frac{1}{\mu}(\tilde{X}^n(s))^+\right) ds, \tag{29}$$

where $\tilde{G}_c^n = \{\tilde{G}_c^n(t)\colon t \geq 0\}$ is defined as in (11), and

$$\tilde{Y}^n(t) = \tilde{X}^n(0) + \tilde{E}^n(t) - \tilde{D}_c^n(t) + \beta^n \mu t + (\tilde{X}^n(0))^- \cdot (\mu t - M(t)). \tag{30}$$

The following proposition yields the weak convergence for $\{\tilde{Y}^n, n \in \mathbb{Z}_+\}$.

Proposition 1. *Assume that conditions (3)–(5), (14)–(15), and Assumption 2 hold. For the sequence of $G/GI/n+GI$ systems, $\tilde{Y}^n \Rightarrow \tilde{Y}$ with*

$$\tilde{Y}(t) = \xi + \tilde{E}(t) - \tilde{D}(t) + \beta\mu t + \xi^- \cdot (\mu t - M(t)),$$

where $\tilde{D} = \{\tilde{D}(t)\colon t \geq 0\}$ is a zero mean Gaussian process, which is independent of $\tilde{E}$ and ξ, with the covariance given by

$$\mathbb{E}[\tilde{D}(s)\tilde{D}(t)] = 2\int_0^s \left(M(u) - u + \frac{1}{2}\right) du + \int_0^s \int_0^t M(s-u)M(t-v) \, dH(u+v) \tag{31}$$

for any $0 \leq s \leq t$.

The proof of this proposition is postponed until after we have established the diffusion approximation Theorem 4. To use Theorem 2, we introduce a regulator mapping in the following lemma.

Lemma 3. *Assume that $g(\cdot)$ is a locally Lipschitz-continuous function with $g(0)=0$. For any $y(\cdot) \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$ and $M(\cdot)$ given by (26), there exists a unique solution $x(\cdot)$ to the following equation:*

$$x(t) = y(t) + \int_0^t (x(t-s))^- \, dM(s) + \int_0^t g((x(s))^+) \, ds. \tag{32}$$

Moreover, the mapping $\Phi_{M,g}(\cdot)\colon \mathbf{D}(\mathbb{R}_+, \mathbb{R}) \to \mathbf{D}(\mathbb{R}_+, \mathbb{R})$ defined by $x = \Phi_{M,g}(y)$ is Lipschitz continuous in the topology of uniform convergence over bounded intervals, measurable with respect to the Borel σ-field generated by the Skorohod J_1-topology, and $\Phi_{M,g}(\mathbf{C}(\mathbb{R}_+, \mathbb{R})) \subseteq \mathbf{C}(\mathbb{R}_+, \mathbb{R})$.

This lemma is a generalization of Proposition 7 in Reed [20] in which $g(\cdot) \equiv 0$ is assumed. The proof of this lemma is presented in Appendix A. Following this lemma and (29), we have $\tilde{X}^n = \Phi_{M,g}(\tilde{Y}^n - \tilde{G}_c^n)$ with $g(t) = -\mu f(t/\mu)$. Theorem 2 can now be applied to obtain the following diffusion approximation.

Theorem 4. *Assume that conditions (3)–(5) and (14)–(15) hold. For the stochastic processes $\{\tilde{X}^n, n \in \mathbb{Z}_+\}$ associated with the sequence of $G/GI/n+GI$ systems, if Assumption 2 holds, then $\tilde{X}^n \Rightarrow \tilde{X}$ as $n \to \infty$, where $\tilde{X} = \{\tilde{X}(t)\colon t \geq 0\}$ is the solution to the following:*

$$\tilde{X}(t) = \tilde{Y}(t) + \int_0^t (\tilde{X}(t-s))^- \, dM(s) - \mu \int_0^t f\left(\frac{1}{\mu}(\tilde{X}(s))^+\right) ds, \tag{33}$$

and $\tilde{Y}$ is given by Proposition 1. Moreover, $\tilde{Q}^n \Rightarrow \tilde{X}^+$ as $n \to \infty$.

Proof. In view of Lemma 3, Proposition 1, and (30), we just need to verify Theorem 2(i). Let $\tilde{Q}_0^n$ denote the queue length process of the many-server system without abandonment. It is proved in Reed [21] that $\{\tilde{Q}_0^n, n \in \mathbb{Z}_+\}$ is stochastically bounded. By Theorem 2.2 of Dai and He [7] with probability one, $\tilde{Q}^n(t) \leq \tilde{Q}_0^n(t)$ for all $t \geq 0$. This implies that $\{\tilde{Q}^n, n \in \mathbb{Z}_+\}$ is stochastically bounded. □

Remark 3. For a given n-server system with patience-time distribution $F(\cdot)$, from Theorem 4 and Corollary 1, we can use

$$\mu\int_0^t \sqrt{n}F\left(\frac{1}{\mu\sqrt{n}}\frac{Q(s)}{\sqrt{n}}\right)ds \tag{34}$$

to approximate $G(t)/\sqrt{n}$. In particular, if $F(x)=1-\exp(-\int_0^x h(s)\,ds)$, then

$$\begin{aligned}\sqrt{n}\left[1-\exp\left(-\int_0^{x/(\mu\sqrt{n})} h(s)\,ds\right)\right] &= \sqrt{n}\left[1-\exp\left(-\frac{1}{\mu\sqrt{n}}\int_0^x h\left(\frac{1}{\mu\sqrt{n}}s\right)ds\right)\right]\\ &\approx \frac{1}{\mu}\int_0^x h\left(\frac{1}{\mu\sqrt{n}}s\right)ds \approx \frac{1}{\mu}\int_0^x h\left(\frac{\sqrt{n}}{\lambda^n}s\right)ds,\end{aligned} \tag{35}$$

which implies that $\int_0^t\int_0^{Q(s)/\sqrt{n}} h((\sqrt{n}u)/\lambda^n)\,du\,ds$ can approximate $G(t)/\sqrt{n}$ well. Dai and He [8] proposed this approximation for the scaled abandonment process $G(t)/\sqrt{n}$ when the patience-time distribution $F(x)=1-\exp(-\int_0^x h(s)\,ds)$. Numerical experiments showed that their approximations are very accurate. Hence our Corollary 1 and Theorem 4 theoretically validate their approximations from the perspective of the diffusion approximations.

Remark 4. When $f(x)=\alpha x$, that is, there is no hazard rate scaling of the patience-time distribution, Theorem 4 gives the diffusion approximation of the queue length for $G/GI/n+G$, which is obtained by Mandelbaum and Momčilović [18]. If the service times are independent and exponentially distributed (Assumption 2 holds by the memoryless property of the exponential distributions), then Theorem 4 gives the diffusion approximations for $G/M/n+G$ with the hazard rate scaling, which is studied by Reed and Tezcan [23].

To obtain Proposition 1, we introduce the following lemma that is related to the weak convergence of the pure empirical processes, and is of independent interest itself. Its proof is presented in Appendix A. To describe the lemma, let $C^n=\{C^n(t)\colon t\geq 0\}$ be a sequence of counting processes and τ_i^n be its ith jump point. Furthermore, for each $n\in\mathbb{Z}_+$, let $\{u_i^n, i\in\mathbb{Z}_+\}$ be a sequence of i.i.d. random variables with a finite mean and some distribution function $H_\star(\cdot)$. Define $\tilde{\mathcal{T}}^n=\{\tilde{\mathcal{T}}^n(t)\colon t\geq 0\}$ and $\tilde{U}^n=\{\tilde{U}^n(t)\colon t\geq 0\}$ with

$$\tilde{\mathcal{T}}^n(t)=\frac{1}{\sqrt{n}}\sum_{i=1}^{C^n(t)}(\mathbf{1}_{\{\tau_i^n+u_i^n>t\}}-(1-H_\star(t-\tau_i^n))),$$

$$\tilde{U}^n(t)=\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor n\mu t\rfloor}\left(\mathbf{1}_{\{i/(n\mu)+u_i^n>t\}}-\left(1-H_\star\left(t-\frac{i}{n\mu}\right)\right)\right).$$

Lemma 4. *Assume that for each $k\in\mathbb{Z}_+$, $\{\tau_1^n,\ldots,\tau_k^n\}$ and $\{u_i^n, i\geq k\}$ are independent, and as $n\to\infty$*

$$\bar{C}^n\Rightarrow\bar{e}, \tag{36}$$

where $\bar{e}(t)=\mu t$. Then, for any $T>0$,

$$\sup_{0\leq t\leq T}|\tilde{\mathcal{T}}^n(t)-\tilde{U}^n(t)|\Rightarrow 0, \tag{37}$$

and $\tilde{\mathcal{T}}^n\Rightarrow\tilde{\mathcal{T}}$, where $\tilde{\mathcal{T}}=\{\tilde{\mathcal{T}}(t)\colon t\geq 0\}$ is a Gaussian process with continuous sample paths, zero mean, and covariance function given by

$$\mathbb{E}[\tilde{\mathcal{T}}(s)\tilde{\mathcal{T}}(t)]=\mu\int_0^s H_\star(s-u)[1-H_\star(t-u)]\,du,\quad 0\leq s\leq t. \tag{38}$$

Proof of Proposition 1. The asymptotic analysis, in particular, that of $\{D^n(t)\colon t\geq 0\}$, follows the idea of Reed [21] and Krichagina and Puhalskii [14]. For completeness, we include the proof here.

Let $K^n(t)$ be the number of customers who have entered service by time t, and denote by κ_i^n the ith jump time of the counting process $\{K^n(t)\colon t\geq 0\}$. Define

$$\tilde{M}_1^n(t)=\frac{1}{\sqrt{n}}\sum_{i=-Q^n(0)+1}^{K^n(t)-Q^n(0)}(\mathbf{1}_{\{\kappa_i^n+v_i^n>t\}}-(1-H(t-\kappa_i^n))),$$

$$\tilde{N}_1^n(t)=\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor\mu nt\rfloor}\left(\mathbf{1}_{\{i/(n\mu)+v^n_{-Q^n(0)+i}>t\}}-\left(1-H\left(t-\frac{i}{n\mu}\right)\right)\right),$$

and

$$\tilde{M}_0^n(t) = \frac{1}{\sqrt{n}} \sum_{i=-X^n(0)+1}^{-Q^n(0)} (\mathbf{1}_{\{v_i^n > t\}} - (1 - H_e(t))),$$

$$\tilde{N}_0^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{1}_{\{v^n_{-Q^n(0)-(n-i)} > t\}} - (1 - H_e(t))),$$

where $\{v_i^n, i \le -Q^n(0)\}$ are independent and identically distributed with distribution function $H_e(\cdot)$, and independent of $\{v_i^n, i \ge -Q^n(0)+1\}$. Hence $\tilde{N}_1^n = \{\tilde{N}_1^n(t)\colon t \ge 0\}$ and $\tilde{N}_0^n = \{\tilde{N}_0^n(t)\colon t \ge 0\}$ are two independent processes. Let $\tilde{M}_1^n = \{\tilde{M}_1^n(t)\colon t \ge 0\}$. Similarly, we define the process $\tilde{M}_0^n$. By the weak convergence of the empirical processes (see Chapter 3 in Shorack and Wellner [25]), we have

$$\tilde{N}_1^n \Rightarrow \tilde{N}_1 \qquad \text{and} \qquad \tilde{N}_0^n \Rightarrow \tilde{N}_0, \tag{39}$$

where, by the independence of $\tilde{N}_1^n$ and $\tilde{N}_0^n$, $\tilde{N}_1 = \{\tilde{N}_1(t)\colon t \ge 0\}$ and $\tilde{N}_0 = \{\tilde{N}_0(t)\colon t \ge 0\}$ are two independent Gaussian processes with continuous sample paths, zero mean and covariance functions given by (38), and $H_e(s) \wedge H_e(t) - H_e(s)H_e(t)$, respectively. By the assumption on the independence between the service times and arrival process $\{E^n(t)\colon t \ge 0\}$, and in view of Theorem 2.8 in Billingsley [3], we have

$$(\tilde{E}^n, \tilde{N}_1^n, \tilde{N}_0^n) \Rightarrow (\tilde{E}, \tilde{N}_1, \tilde{N}_0).$$

Using (14) and Theorem 3.9 in Billingsley [3], we obtain

$$(\tilde{X}^n(0), \tilde{E}^n, \tilde{N}_1^n, \tilde{N}_0^n) \Rightarrow (\xi, \tilde{E}, \tilde{N}_1, \tilde{N}_0). \tag{40}$$

Again, by (14), we have

$$\frac{X^n(0)}{n} \Rightarrow 1.$$

In view of the definitions of $\{\tilde{M}_0^n(t)\colon t \ge 0\}$ and $\{\tilde{N}_0^n(t)\colon t \ge 0\}$,

$$\sup_{0 \le t \le T} |\tilde{M}_0^n(t) - \tilde{N}_0^n(t)| \Rightarrow 0. \tag{41}$$

Let $\bar{K}^n = \{K^n(t)/n\colon t \ge 0\}$. Recall that in the proof of Theorem 4 of Reed [21] and Theorem 2.2 of Dai and He [7], the sequence of queue length processes $\{\tilde{Q}^n, n \in \mathbb{Z}_+\}$ is stochastically bounded. It follows from Theorem 1 and

$$Q^n(t) = Q^n(0) + A^n(t) - K^n(t) - G^n(t)$$

that $\bar{K}^n \Rightarrow \bar{e}$ as $n \to \infty$. By Lemma 4, we have

$$\sup_{0 \le t \le T} |\tilde{M}_1^n(t) - \tilde{N}_1^n(t)| \Rightarrow 0. \tag{42}$$

Therefore, it follows from (40)–(42) that

$$(\tilde{X}^n(0), \tilde{E}^n, \tilde{M}_1^n, \tilde{M}_0^n) \Rightarrow (\xi, \tilde{E}, \tilde{N}_1, \tilde{N}_0). \tag{43}$$

By Proposition 2.1 in Reed [21], similar to Proposition 4.4 of Reed and Shaki [22], and in view of (27), we have

$$\tilde{D}^n(t) = -(\tilde{M}_0^n(t) + \tilde{M}_1^n(t)) - \int_0^t (\tilde{M}_0^n(t-s) + \tilde{M}_1^n(t-s))\, dM(s).$$

Thus the proposition follows directly from (43) and the continuous mapping theorem. □

3.3. $G/M/n^\alpha + GI$ in the NDS regime.

In this subsection, we consider the sequence of $G/M/n^\alpha + G$ queues with $\alpha \in (0, 1)$ in the NDS regime considered in Atar [1] and Atar and Gurvich [2]. Again, the speed for each server to process customer service requirements is assumed to be one. The following assumption on the customer service times is needed.

Assumption 3. *For the nth system $G/M/n^\alpha + G$, the customers' remaining service requirements and service requirements $\{v_i^n, i \ge -X^n(0)+1\}$ are independent and exponentially distributed with parameter $\mu^n = n^{1-\alpha}\mu$.*

By the memoryless property of customer service times, as usual, the evolution of the process X^n can be characterized by the system dynamics equation

$$X^n(t) = X^n(0) + E^n(t) - S_p\left(\mu^n \int_0^t (X^n(s) \wedge n^\alpha)\, ds\right) - G^n(t),$$

where $S_p(\cdot)$ is a Poisson process with rate one. Since

$$n\mu t - \mu^n \int_0^t (X^n(s) \wedge n^\alpha)\, ds = \mu^n \int_0^t (X^n(s) - n^\alpha)^-\, ds,$$

we have

$$\begin{aligned} X^n(t) - n^\alpha &= X^n(0) - n^\alpha + E^n(t) - \lambda^n t \\ &\quad - \left[S_p\left(\mu^n \int_0^t (X^n(s) \wedge n^\alpha)\, ds\right) - \mu^n \int_0^t (X^n(s) \wedge n^\alpha)\, ds\right] - G^n(t) \\ &\quad + (\lambda^n - n\mu)t + \mu^n \int_0^t (X^n(s) - n^\alpha)^-\, ds. \end{aligned} \tag{44}$$

Applying the diffusion scaling for X^n, E^n, and G^n and the definition of β^n in (15), we obtain

$$(\tilde{X}^n(t))^+ = \tilde{Y}^n(t) + (\tilde{X}^n(t))^- - \tilde{G}^n(t) + \mu^n \int_0^t (\tilde{X}^n(s))^-\, ds, \tag{45}$$

where

$$\tilde{Y}^n(t) = \tilde{X}^n(0) + \tilde{E}^n(t) - \tilde{S}_p^n(t) + \beta^n \mu t, \tag{46}$$

$$\tilde{S}_p^n(t) = \frac{1}{\sqrt{n}}\left[S_p\left(\mu^n \int_0^t (X^n(s) \wedge n^\alpha)\, ds\right) - \mu^n \int_0^t (X^n(s) \wedge n^\alpha)\, ds\right]. \tag{47}$$

We can see that $(\tilde{X}^n(t))^+$, by (45), is related to the solution of the Skorohod equation. This observation is useful in establishing the stochastic boundedness of queue length processes, see the proof of Proposition 3. First, we prove the following result.

Proposition 2. *Assume that conditions* (3) *and* (15), *and Assumption* 3 *hold. If condition* (14) *holds with* $\mathbb{P}(\xi \ge 0) = 1$, *then* $(\tilde{X}^n)^- \Rightarrow 0$ *as* $n \to \infty$.

Proof. The proof is similar to the one in Atar [1]. For any fixed $\varepsilon > 0$, we will prove that

$$\mathbb{P}\left(\sup_{0 \le t \le T} (\tilde{X}^n(t))^- \ge \varepsilon\right) \to 0 \quad \text{as } n \to \infty. \tag{48}$$

Define

$$\Omega_0^n(\varepsilon) = \left\{(\tilde{X}^n(0))^- \le \frac{\varepsilon}{4}\right\}, \qquad \Omega^n(\varepsilon, T) = \left\{\sup_{0 \le t \le T} (\tilde{X}^n(t))^- \ge \varepsilon\right\}, \qquad t_1^n = \inf\{t \ge 0\colon (\tilde{X}^n(t))^- \ge \varepsilon\}.$$

Because of $\xi \ge 0$ with probability one, by (14), it is sufficient to prove that the probabilities of the event $\Omega^n(\varepsilon, T) \cap \Omega_0^n(\varepsilon)$ vanishes as n converges to infinity. On the set $\Omega^n(\varepsilon, T) \cap \Omega_0^n(\varepsilon)$, define $t_2^n = \sup\{0 \le t \le \eta^n\colon (\tilde{X}^n(t))^- \le \varepsilon/3\} \vee 0$. By the definitions of t_1^n and t_2^n, we clearly have that

$$(\tilde{X}^n(t_1^n))^- \ge \varepsilon \qquad \text{and} \qquad (\tilde{X}^n(t_2^n-))^- \le \frac{\varepsilon}{3}.$$

Note that $(\tilde{X}^n(t))^- \ge \varepsilon/3$ for all $t \in [t_2^n, t_1^n]$. As a result, $\tilde{X}^n(t) = -(\tilde{X}^n(t))^-$ and there is no abandonment during this interval. From Equation (45), we have that on $\Omega^n(\varepsilon, T) \cap \Omega_0^n(\varepsilon)$,

$$\begin{aligned} \tilde{Y}^n(t_1^n) - \tilde{Y}^n(t_2^n-) &= (\tilde{X}^n(t_2^n-))^- - (\tilde{X}^n(t_1^n))^- - \mu^n \int_{t_2^n}^{t_1^n} (\tilde{X}^n(s))^-\, ds \\ &\le -\frac{2\varepsilon}{3} - \frac{\varepsilon\mu^n(t_1^n - t_2^n)}{3}. \end{aligned}$$

For fixed δ, depending on whether $t_1^n - t_2^n > \delta$ or $t_1^n - t_2^n \le \delta$, we get

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{P}(\Omega^n(\varepsilon, T)\cap \Omega_0^n(\varepsilon))\\ &\quad\le \lim_{n\to\infty} \mathbb{P}\left(\tilde{Y}^n(t_1^n) - \tilde{Y}^n(t_2^n-) \le -\frac{2\varepsilon}{3} - \frac{\varepsilon\mu^n(t_1^n - t_2^n)}{3}\right)\\ &\quad\le \lim_{n\to\infty} \mathbb{P}\left(\sup_{\substack{0\le s,\, t\le T\\ |s-t|\le\delta}} |\tilde{Y}^n(t) - \tilde{Y}^n(s)| \ge \frac{2\varepsilon}{3}\right) + \lim_{n\to\infty} \mathbb{P}\left(\sup_{0\le t\le T} |\tilde{Y}^n(t)| \ge \frac{\varepsilon\mu^n\delta}{6}\right).\end{aligned} \tag{49}$$

Noting that for $u, v\in[0,\infty)$ with $u<v$,

$$\begin{aligned}0 &\le \mu^n \int_0^v (X^n(s)\wedge n^\alpha)\,ds - \mu^n \int_0^u (X^n(s)\wedge n^\alpha)\,ds\\ &= \mu^n \int_u^v (X^n(s)\wedge n^\alpha)\,ds\\ &\le n(v-u),\end{aligned}$$

and $\mu^n \int_0^t (X^n(s)\wedge n^\alpha)\,ds \le nt$, we know $\{\tilde{S}_p^n, n\in\mathbb{Z}_+\}$ with $\tilde{S}_p^n = \{\tilde{S}_p^n(t)\colon t\ge 0\}$ given by (47) is C-tight. Hence, by (3) and (15), we have $\{\tilde{Y}^n, n\in\mathbb{Z}_+\}$ is C-tight. Letting $n\to\infty$ and then $\delta\to 0$, the term in (49) then converges to 0. This completes the proof. □

To get the diffusion approximation for the queue length processes, we need the following proposition.

Proposition 3. *Under the conditions required by Proposition 2, $\tilde{S}_p^n \Rightarrow \sqrt{\mu}\tilde{S}_p$ as $n\to\infty$, where $\tilde{S}_p = \{\tilde{S}_p(t)\colon t\ge 0\}$ is a standard Brownian motion, which is independent of the limit of the arrival processes ($\tilde{E}$ given by (3)) as well as of the initial states (ξ given by (14)).*

Proof. Consider the solution $(\tilde{Z}^n(\cdot), \tilde{Z}_r^n(\cdot))$ to the following Skorohod equation with probability one,

$$\begin{aligned}&\tilde{Z}^n(t) = \tilde{Y}^n(t) + (\tilde{X}^n(t))^- + \tilde{Z}_r^n(t), \quad t\ge 0,\\ &\tilde{Z}^n(t)\ge 0, \quad t\ge 0;\\ &\tilde{Z}_r^n(\cdot) \text{ is nondecreasing;}\\ &\int_0^\infty \mathbf{1}_{\{\tilde{Z}^n(t)>0\}}\, d\tilde{Z}_r^n(t) = 0.\end{aligned}$$

By C-tightness of $\{\tilde{Y}^n, n\in\mathbb{Z}_+\}$ given by the proof of Proposition 2, and the Lipschitz continuity of the Skorohod mapping, it follows from Proposition 2 that $\{\tilde{Z}^n, n\in\mathbb{Z}_+\}$ is stochastically bounded. By $\tilde{Y}^n(t) + (\tilde{X}^n(t))^- \ge \tilde{Y}^n(t) + (\tilde{X}^n(t))^- - \tilde{G}^n(t)$ with probability one, and (45), with the help of Lemma 4.1 in Kruk et al. [15], we know that $(\tilde{X}^n(t))^+$ $(=\tilde{Q}^n(t))$ can be bounded by $\tilde{Z}^n(t)$. Therefore $\{\tilde{Q}^n, n\in\mathbb{Z}_+\}$ is also stochastically bounded. Thus, by Theorem 1, we have that as $n\to\infty$

$$\bar{G}^n \Rightarrow 0 \quad \text{with } \bar{G}^n = \left\{\frac{1}{n}G^n(t)\colon t\ge 0\right\}. \tag{50}$$

On the other hand, by (45),

$$\frac{\mu^n}{\sqrt{n}} \int_0^t (\tilde{X}^n(s))^-\,ds = \frac{1}{\sqrt{n}}((\tilde{X}^n(t))^+ - \tilde{Y}^n(t) - (\tilde{X}^n(t))^- + \tilde{G}^n(t)). \tag{51}$$

Combining (50)–(51) yields that

$$\frac{1}{n}\int_0^\cdot \mu^n (X^n(s) - n^\alpha)^-\,ds \Rightarrow 0 \quad \text{as } n\to\infty,$$

Which, consequently, implies that

$$\frac{1}{n}\int_0^\cdot \mu^n (X^n(s)\wedge n^\alpha)\,ds \Rightarrow \bar{e}(\cdot) \quad \text{as } n\to\infty, \tag{52}$$

where $\bar{e}(t) = \mu t$. The proposition directly follows from (47) and the random-time-change theorem (Corollary 1 of Whitt [32]). □

Now, we are ready to state the diffusion approximation for the queue length processes.

Theorem 5. *Assume that conditions (3)–(5) and (14)–(15) hold. If Assumption 3 holds and $\xi \geq 0$ with probability one, then $\tilde{X}^n \Rightarrow \tilde{X}$ as $n \to \infty$. Here, $\tilde{X} = \{\tilde{X}(t): t \geq 0\}$ is given by $\tilde{X} = \Phi_g(\tilde{Y})$ (recall that $\Phi_g(\cdot)$ is defined in Lemma 2) with*

$$g(t) = -\mu f\left(\frac{1}{\mu}t\right), \qquad t \geq 0;$$
$$\tilde{Y} = \{\tilde{Y}(t): t \geq 0\}, \qquad \tilde{Y}(t) = \xi + \tilde{E}(t) - \sqrt{\mu}\tilde{S}_p(t) + \beta\mu t,$$

where $\tilde{S}_p$ given by Proposition 3 is a standard Brownian motion independent of ξ and $\tilde{E}$. Moreover, $\tilde{Q}^n \Rightarrow \tilde{X}$ as $n \to \infty$.

Proof of Theorem 5. First, from condition (14) on the initial states, the condition (3) on the arrival process, (15) on the traffic condition, Proposition 2 on $(\tilde{X}^n(t))^-$, and Proposition 3 for $\{\tilde{S}_p^n, n \in \mathbb{Z}_+\}$, we have that as $n \to \infty$,

$$\tilde{Y}^n + (\tilde{X}^n)^- \quad \Rightarrow \quad \xi + \tilde{E} - \sqrt{\mu}\tilde{S}_p + \beta\bar{e}.$$

Note that by (45),

$$(\tilde{X}^n(t))^+ = \tilde{Y}^n(t) + (\tilde{X}^n(t))^- - \tilde{G}_c^n(t) - \mu\int_0^t f\left(\frac{1}{\mu}(\tilde{X}^n(s))^+\right) ds + \mu^n\int_0^t (\tilde{X}^n(s))^- \, ds.$$

Recall that the stochastic boundedness of the queue length processes is proved in the proof of Proposition 3. With $\tilde{Y}^n + (\tilde{X}^n)^-$ playing the role of $\tilde{Y}^n$ in Theorem 2, it follows from Lemma 2 and Theorem 2 that $\tilde{Q}^n = (\tilde{X}^n)^+ \Rightarrow \tilde{X}$ as $n \to \infty$. It then follows from Proposition 2 that $\tilde{X}^n \Rightarrow \tilde{X}$ as $n \to \infty$. Hence the proof of the theorem is completed. □

Remark 5. Note that $\tilde{X}$ in Theorem 5 has the similar structure as the one in Theorem 3.

4. Proofs of Theorems 1–2, and Corollary 1.

In this section, we give the proofs of the theorems and corollary given in §2. First, we look at the first theorem, Theorem 1. The proof of Theorem 1 is based on three properties of such queueing systems, namely, Propositions 4–6, which are of independent interest themselves. We will first state these properties and then apply them to prove Theorem 1. The proofs of the three propositions are given in Appendix B.2.

To describe these three propositions, following Dai and He [7], we introduce two notions. The first one is the *offered waiting time* ω_i^n, which denotes the time that the ith arriving customer in the nth system after time 0 has to wait before receiving service for each $i \geq 1$. When $Q^n(0) > 0$, we index the initial customer in the queue by $0, -1, \ldots, -Q^n(0)+1$, with customer $-Q^n(0)+1$ being the first one in the queue. Each ω_i^n denotes the remaining waiting time of the ith customer for $i = -Q^n(0)+1, \ldots, 0$. The second notion is the *virtual waiting time* $\omega^n(t)$, which is the amount of time a hypothetical customer with infinite patience would have to wait before receiving service upon arriving at time t in the nth system. We introduce the diffusion-scaled virtual waiting-time process $\tilde{\omega}^n = \{\tilde{\omega}^n(t): t \geq 0\}$ as

$$\tilde{\omega}^n(t) = \sqrt{n}\omega^n(t).$$

The first property of interest is the stochastic boundedness of the scaled virtual waiting time and abandonment probability.

Proposition 4. *Under assumptions (2)–(4) and (6), the sequences of the scaled virtual waiting times $\{\tilde{\omega}^n, n \in \mathbb{Z}_+\}$ and the scaled abandonment probabilities $\{\tilde{F}_\omega^n, n \in \mathbb{Z}_+\}$ given by*

$$\tilde{F}_\omega^n = \left\{\sup_{0 \leq i \leq E^n(t)} \sqrt{n}F^n(\omega_i^n): t \geq 0\right\}$$

are stochastically bounded for any given $T > 0$.

The second proposition reveals an asymptotic relationship between the abandonment process and the offered waiting time.

Proposition 5. *Under assumptions (2)–(4) and (6), for each $T > 0$,*

$$\sup_{0 \leq t \leq T}\left|\tilde{G}^n(t) - \frac{1}{\sqrt{n}}\sum_{j=1}^{E^n(t)} F^n(\omega_j^n)\right| \Rightarrow 0 \quad \text{as } n \to \infty.$$

Note that neither of the above two propositions needs the modulus of continuity to asymptotically vanish as in (7). The next proposition establishes the relationship between the virtual waiting time and the queue length. For this, condition (7) is required.

Proposition 6. *Under assumptions (2)–(4) and (6)–(7), for each $T>0$,*

$$\sup_{0\le t\le T}|\mu\tilde{\omega}^n(t)-\tilde{Q}^n(t)|\Rightarrow 0 \quad \text{as } n\to\infty.$$

Remark 6. By the above proposition and the triangle inequality, for any $s,t\in[0,T]$,

$$|\tilde{\omega}^n(t)-\tilde{\omega}^n(s)|\le 2\sup_{0\le t\le T}\left|\tilde{\omega}^n(t)-\frac{1}{\mu}\tilde{Q}^n(t)\right|+\frac{1}{\mu}|\tilde{Q}^n(t)-\tilde{Q}^n(s)|.$$

Thus the C-tightness of $\{\tilde{Q}^n, n\in\mathbb{Z}_+\}$ implies the C-tightness of $\{\tilde{w}^n, n\in\mathbb{Z}_+\}$.

Remark 7. The same result has been proved by Talreja and Whitt [27] under different assumptions. Theorem 3.1 in Talreja and Whitt [27] requires the convergence of several scaled processes, including those describing arrival, service completion, abandonment, and total number of customers in the system, whereas our result only needs the convergence of the arrival processes and C-tightness of the queue length processes.

Proof of Theorem 1. First, consider C-tightness of $\{\tilde{G}^n, n\in\mathbb{Z}_+\}$. According to Proposition 5, it is enough to show the C-tightness for $\{(1/\sqrt{n})\sum_{i=1}^{E^n(t)}F^n(\omega_i^n), n\in\mathbb{Z}_+\}$. Define the fluid-scaled arrival process $\bar{E}^n=\{\bar{E}^n(t)\colon t\ge 0\}$ by

$$\bar{E}^n(t)=\frac{E^n(t)}{n}.$$

Condition (3) implies that as $n\to\infty$,

$$\bar{E}^n\Rightarrow\bar{e}, \tag{53}$$

where $\bar{e}(t)=\mu t$. Note that for any $0\le s\le t\le T$,

$$\frac{1}{\sqrt{n}}\sum_{i=E^n(s)+1}^{E^n(t)}F^n(\omega_i^n)\le[\bar{E}^n(t)-\bar{E}^n(s)]\cdot\sup_{0\le i\le E^n(T)}\sqrt{n}F^n(\omega_i^n).$$

Therefore the C-tightness follows from the C-tightness of $\bar{E}^n$ (due to (53)) and the stochastic boundedness of $\sup_{0\le i\le E^n(T)}\sqrt{n}F^n(\omega_i^n)$ (due to Proposition 4).

Next, we look at (8). According to Proposition 5, it is enough to prove that as $n\to\infty$,

$$\sup_{0\le t\le T}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^{E^n(t)}F^n(\omega_j^n)-\mu\int_0^t f\left(\frac{1}{\mu}\tilde{Q}^n(s)\right)ds\right|\Rightarrow 0. \tag{54}$$

After adding and subtracting a new term, we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{j=1}^{E^n(t)}F^n(\omega_j^n)-\mu\int_0^t f\left(\frac{1}{\mu}\tilde{Q}^n(s)\right)ds\\&\quad=\frac{1}{\sqrt{n}}\sum_{j=1}^{E^n(t)}F^n(\omega_j^n)-\mu\int_0^t f(\tilde{\omega}^n(s))\,ds+\mu\int_0^t f(\tilde{\omega}^n(s))\,ds-\mu\int_0^t f\left(\frac{1}{\mu}\tilde{Q}^n(s)\right)ds.\end{aligned}$$

Thus, it is enough to prove that when $n\to\infty$,

$$\sup_{0\le t\le T}\left|\int_0^t f(\tilde{\omega}^n(s))\,ds-\int_0^t f\left(\frac{1}{\mu}\tilde{Q}^n(s)\right)ds\right|\Rightarrow 0, \tag{55}$$

$$\sup_{0\le t\le T}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^{E^n(t)}F^n(\omega_j^n)-\mu\int_0^t f(\tilde{\omega}^n(s))\,ds\right|\Rightarrow 0. \tag{56}$$

We first prove (55). By assumption (6) and Proposition 4, for any $\varepsilon>0$, there exists Γ large enough such that for all large enough n,

$$\mathbb{P}\left\{\sup_{0\le t\le T}\frac{1}{\mu}\tilde{Q}^n(t)\ge\Gamma\right\}+\mathbb{P}\left\{\sup_{0\le t\le T}\tilde{\omega}^n(t)\ge\Gamma\right\}\le\frac{\varepsilon}{2}. \tag{57}$$

For any $\delta>0$, by the local Lipschitz continuity of $f(\cdot)$ given by (5), we have

$$\begin{aligned}&\mathbb{P}\left\{\sup_{0\le t\le T}\left|\int_0^t f(\tilde{\omega}^n(s))\,ds-\int_0^t f\left(\frac{1}{\mu}\tilde{Q}^n(s)\right)ds\right|\ge\delta\right\}\\&\quad\le\mathbb{P}\left\{\sup_{0\le t\le T}\frac{1}{\mu}\tilde{Q}^n(t)\ge\Gamma\right\}+\mathbb{P}\left\{\sup_{0\le t\le T}\tilde{\omega}^n(t)\ge\Gamma\right\}+\mathbb{P}\left\{\sup_{0\le t\le T}\left|\tilde{\omega}^n(t)-\frac{1}{\mu}\tilde{Q}^n(t)\right|\ge\frac{\delta}{T\Lambda_\Gamma}\right\}.\end{aligned}$$

By Proposition 6, the third term on the right-hand side in the above can be less than $\varepsilon/2$ for all large enough n. Thus (55) is proved by (57).

Next, we prove (56). According to Lemma 3.2 of Dai and He [7] and the monotonicity of the distribution function $F^n(\cdot)$,

$$\int_0^t \sqrt{n}F^n\left(\frac{1}{\sqrt{n}}\tilde{\omega}^n(s-)\right)d\bar{E}^n(s) \leq \frac{1}{\sqrt{n}}\sum_{j=1}^{E^n(t)} F^n(\omega_j^n) \leq \int_0^t \sqrt{n}F^n\left(\frac{1}{\sqrt{n}}\tilde{\omega}^n(s)\right)d\bar{E}^n(s).$$

As a result, it is sufficient to prove the following convergence as $n \to \infty$,

$$\sup_{0\leq t\leq T}\left|\int_0^t \sqrt{n}F^n\left(\frac{1}{\sqrt{n}}\tilde{\omega}^n(s)\right)d\bar{E}^n(s) - \mu\int_0^t f(\tilde{\omega}^n(s))\,ds\right| \Rightarrow 0, \tag{58a}$$

$$\sup_{0\leq t\leq T}\left|\int_0^t \sqrt{n}F^n\left(\frac{1}{\sqrt{n}}\tilde{\omega}^n(s-)\right)d\bar{E}^n(s) - \mu\int_0^t f(\tilde{\omega}^n(s))\,ds\right| \Rightarrow 0. \tag{58b}$$

We only prove (58a) since (58b) can be proved similarly. The idea is similar to the one proposed by Ward and Glynn [30]. By Remark 6 and (53), we have that $\{(\tilde{\omega}^n, \bar{E}^n), n \in \mathbb{Z}_+\}$ is C-tight. Therefore for every convergent subsequence indexed by n_k,

$$(\tilde{\omega}^{n_k}, \bar{E}^{n_k}) \Rightarrow (\tilde{\omega}, \bar{e}) \quad \text{as } n_k \to \infty,$$

for some process $\tilde{\omega} \in \mathbf{C}(\mathbb{R}_+, \mathbb{R})$. By the Skorohod representation theorem, there exists another probability space $(\breve{\Omega}, \breve{\mathcal{F}}, \breve{\mathbb{P}})$, as well as a sequence of processes $(\breve{\omega}^{n_k}, \breve{E}^{n_k})$ and $(\breve{\omega}, \bar{e})$ defined on it, such that

$$(\breve{\omega}^{n_k}, \breve{E}^{n_k}) \overset{d}{=} (\tilde{\omega}^{n_k}, \bar{E}^{n_k}),$$
$$(\breve{\omega}, \bar{e}) \overset{d}{=} (\tilde{\omega}, \bar{e}),$$

and with probability one, $\breve{\omega}^{n_k}$ converges to $\breve{\omega}$ in $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ and $\breve{E}^{n_k}$ converges to $\bar{e}$ in $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$. We have that for any $T \geq 0$,

$$\begin{aligned}
&\sup_{0\leq t\leq T}\left|\int_0^t \sqrt{n_k}F^{n_k}\left(\frac{1}{\sqrt{n_k}}\tilde{\omega}^{n_k}(s)\right)d\bar{E}^{n_k}(s) - \mu\int_0^t f(\tilde{\omega}^{n_k}(s))\,ds\right| \\
&\quad\overset{d}{=} \sup_{0\leq t\leq T}\left|\int_0^t\left(\sqrt{n_k}F^{n_k}\left(\frac{1}{\sqrt{n_k}}\breve{\omega}^{n_k}(s)\right) - f(\breve{\omega}^{n_k}(s))\right)d\breve{E}^{n_k}(s)\right| \\
&\qquad + \sup_{0\leq t\leq T}\left|\int_0^t f(\breve{\omega}^{n_k}(s))\,d\breve{E}^{n_k}(s) - \mu\int_0^t f(\breve{\omega}(s))\,ds\right| + \mu\sup_{0\leq t\leq T}\left|\int_0^t f(\breve{\omega}^{n_k}(s))\,ds - \int_0^t f(\breve{\omega}(s))\,ds\right|.
\end{aligned} \tag{59}$$

There exist N_1 and M such that when $n_k \geq N_1$,

$$\sup_{0\leq t\leq T}|\breve{E}^{n_k}(t)| \leq 2\mu T, \qquad \sup_{0\leq t\leq T}|\breve{\omega}^{n_k}(t)| \leq M.$$

As a result of Lemma 4.1 of Dai [5] and Condition (4), $\sqrt{n_k}F^{n_k}(x/\sqrt{n_k})$ converge to $f(x)$ uniformly on compact sets. Thus, for any given $\varepsilon > 0$, we can find an N_2 such that when $n_k \geq N_2$,

$$\sup_{0\leq x\leq M}\left|\sqrt{n_k}F^{n_k}\left(\frac{x}{\sqrt{n_k}}\right) - f(x)\right| \leq \frac{\varepsilon}{2\mu T}.$$

Therefore we can conclude that

$$\sup_{0\leq t\leq T}\left|\int_0^t\left(\sqrt{n_k}F^{n_k}\left(\frac{1}{\sqrt{n_k}}\breve{\omega}^{n_k}(s)\right) - f(\breve{\omega}^{n_k}(s))\right)d\breve{E}^{n_k}(s)\right| \leq \varepsilon$$

for all $n_k \geq \max(N_1, N_2)$. This proves that the first term in (59) converges to 0. By the continuous mapping theorem and (4), we also know that with probability one, $f(\breve{\omega}^{n_k})$ converges to $f(\breve{\omega})$ as $n_k \to \infty$ in $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$. By Lemma 8.3 of Dai and Dai [6], we know that with probability one, as $n_k \to \infty$,

$$\sup_{0\leq t\leq T}\left|\int_0^t f(\breve{\omega}^{n_k}(s))\,d\breve{E}^{n_k}(s) - \mu\int_0^t f(\breve{\omega}(s))\,ds\right| \to 0,$$

$$\sup_{0\leq t\leq T}\left|\int_0^t f(\breve{\omega}^{n_k}(s))\,ds - \int_0^t f(\breve{\omega}(s))\,ds\right| \to 0.$$

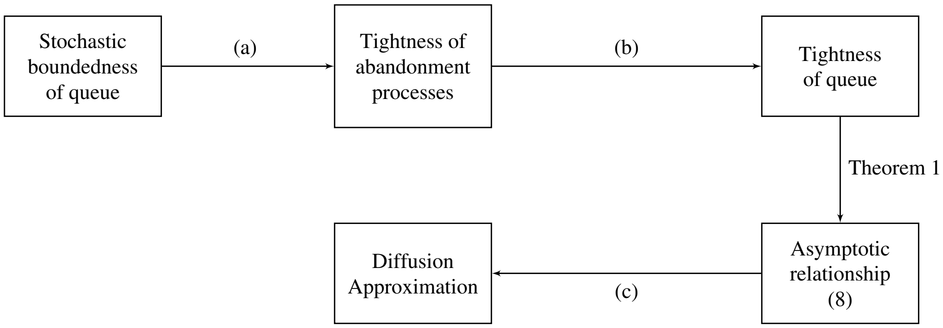
Stochastic boundedness of queue
(a)
Tightness of abandonment processes
(b)
Tightness of queue
Theorem 1
Asymptotic relationship (8)
(c)
Diffusion Approximation

Figure 1. The approach to the diffusion approximation.

As a result, with probability one as $n_k \to \infty$,

$$\sup_{0\le t\le T}\left|\int_0^t \sqrt{n_k}F^{n_k}\left(\frac{1}{\sqrt{n_k}}\breve{\omega}^{n_k}(s)\right)d\breve{E}^{n_k}(s)-\mu\int_0^t f(\breve{\omega}^{n_k}(s))\,ds\right|\to 0. \tag{60}$$

Since $(\breve{\omega}^{n_k},\breve{E}^{n_k})\overset{d}{=}(\tilde{\omega}^{n_k},\bar{E}^{n_k})$, we have

$$\begin{aligned}&\sqrt{n_k}\int_0^t F^{n_k}\left(\frac{1}{\sqrt{n_k}}\breve{\omega}^{n_k}(s)\right)d\breve{E}^{n_k}(s)-\mu\int_0^t f(\breve{\omega}^{n_k}(s))\,ds\\ &\quad\overset{d}{=}\sqrt{n_k}\int_0^t F^{n_k}\left(\frac{1}{\sqrt{n_k}}\tilde{\omega}^{n_k}(s)\right)d\bar{E}^{n_k}(s)-\mu\int_0^t f(\tilde{\omega}^{n_k}(s))\,ds.\end{aligned}$$

Hence (60) implies that as $k\to\infty$,

$$\sup_{0\le t\le T}\left|\sqrt{n_k}\int_0^t F^{n_k}\left(\frac{1}{\sqrt{n_k}}\tilde{\omega}^{n_k}(s)\right)d\bar{E}^{n_k}(s)-\mu\int_0^t f(\tilde{\omega}^{n_k}(s))\,ds\right|\Rightarrow 0.$$

Since the above convergence to zero holds for all convergent subsequences, (58a) is established. □

Proof of Corollary 1. If $F^n(x)=1-\exp(-\int_0^x h(\sqrt{n}s)\,ds)$, we have $f(x)=\int_0^x h(s)\,ds$. Hence

$$\mu\int_0^t f\left(\frac{1}{\mu}\tilde{Q}^n(s)\right)ds=\mu\int_0^t\int_0^{\tilde{Q}^n(s)/\mu}h(u)\,du\,ds=\int_0^t\int_0^{\tilde{Q}^n(s)}h\left(\frac{u}{\mu}\right)du\,ds.$$

This, by Theorem 1, implies (i).

Now, we prove (ii). Note that if $F^n(x)=F(x)$ with derivative $F'(0+)$ at $x=0$, then $f(x)=F'(0+)x$. It follows from Theorem 1 that

$$\sup_{0\le t\le T}\left|\tilde{G}^n(t)-F'(0+)\int_0^t\tilde{Q}^n(s)\,ds\right|\Rightarrow 0\quad\text{as } n\to\infty.\quad\square$$

Proof of Theorem 2. To better understand the proof, we depict its logic in Figure 1.

Condition (i), together with Theorem 1, implies that the abandonment process $\{\tilde{G}^n, n\in\mathbb{Z}_+\}$ is C-tight (see arrow (a) in Figure 1). Further, the continuity of $f(\cdot)$ implies that the sequence of processes given by the second term on the right-hand side of (11) is also C-tight. Therefore according to the definition given by (11), $\{\tilde{G}^n_c, n\in\mathbb{Z}_+\}$ is C-tight. By condition (ii), $\{\tilde{Y}^n-\tilde{G}^n_c, n\in\mathbb{Z}_+\}$ is also C-tight. Then, for any subsequence $\{n_k, k\in\mathbb{Z}_+\}$, we can find another subsequence with indices $\{n'_k, k\in\mathbb{Z}_+\}\subseteq\{n_k, k\in\mathbb{Z}_+\}$ such that

$$\tilde{Y}^{n'_k}-\tilde{G}^{n'_k}_c\quad\Rightarrow\quad\tilde{Y}'\quad\text{as } n'_k\to\infty \tag{61}$$

in the Skorohod J_1-topology for some limit $\tilde{Y}'\in\mathbf{C}(\mathbb{R}_+,\mathbb{R})$. Recalling that if $x(\cdot)\in\mathbf{C}(\mathbb{R}_+,\mathbb{R})$, then $x_n\to x$ in the Skorohod J_1-topology is equivalent to $x_n\to x$ in the uniform topology. Thus $\Phi(\cdot)$ in condition (iii) is continuous on $\mathbf{C}(\mathbb{R}_+,\mathbb{R})$ under the Skorohod J_1-topology. Denote by D_Φ the set of discontinuous points of $\Phi(\cdot)$ in the Skorohod J_1-topology. Then, $\mathbf{D}(\mathbb{R}_+,\mathbb{R})\setminus\mathbf{C}(\mathbb{R}_+,\mathbb{R})\supseteq D_\Phi$. By the condition that $\tilde{Y}'\in\mathbf{C}(\mathbb{R}_+,\mathbb{R})$, we have $\mathbb{P}(\tilde{Y}'\in D_\Phi)=0$. Therefore the measurability of $\Phi(\cdot)$, together with the continuous mapping theorem (see Theorem 2.7 in Billingsley [3], page 21) and (61), yields

$$\tilde{X}^{n'_k}=\Phi(\tilde{Y}^{n'_k}-\tilde{G}^{n'_k}_c)\quad\Rightarrow\quad\Phi(\tilde{Y}')\quad\text{as } n'_k\to\infty \tag{62}$$

in the Skorohod J_1-topology. This and $\Phi(\tilde{Y}') \subseteq \mathbf{C}(\mathbb{R}_+, \mathbb{R})$ (see condition (iii)) show that $\{\tilde{X}^{n'_k}, k \in \mathbb{Z}_+\}$ is C-tight. A direct consequence is the C-tightness of the queue length processes $\{\tilde{Q}^{n'_k}, k \in \mathbb{Z}_+\}$ (see arrow (b) in Figure 1). In other words, $\{\tilde{Q}^{n'_k}, k \in \mathbb{Z}_+\}$ satisfies condition (7). From Theorem 1 and (ii), we know $\tilde{Y}' = \tilde{Y}$. Therefore, in view of the arbitrariness of the subsequence of $\{n_k, k \in \mathbb{Z}_+\}$, we have

$$\tilde{X}^n \Rightarrow \Phi(\tilde{Y}) \quad \text{as } n \to \infty$$

in the Skorohod J_1-topology (see arrow (c) in Figure 1). This completes the proof. □

Acknowledgments. The authors thank Jim Dai and Shuangchi He for suggesting the proof for Proposition 4. The authors also thank the associate editor and two anonymous referees for their valuable suggestions, which lead to §3.3 and the simplification of the proof of Lemma 6. The research is supported by a start-up grant from NUS Business School and GRF grants (Projects 24500314, 622110 and 622411) from the Hong Kong Research Grants Council.

Appendix A. Regulator mappings.

In this section, we prove Lemmas 2 and 3 based on the following result.

Lemma 5. *Assume that Υ: $\mathbf{D}(\mathbb{R}_+, \mathbb{R}) \to \mathbf{D}(\mathbb{R}_+, \mathbb{R})$ is measurable under the Skorohod J_1-topology, Lipschitz continuous under the topology of uniform convergence over bounded intervals, and $\Upsilon(0) = 0$, and $h(\cdot)$ is a Lipschitz-continuous function on $\mathbb{R}$ with $h(0) = 0$. Then, for $y(\cdot) \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$,*

$$x(t) = y(t) + \int_0^t h(\Upsilon(x)(s))\, ds \tag{A1}$$

has a unique solution (denoted by $x = \Xi_{\Upsilon, h}(y)$). The mapping $\Xi_{\Upsilon, h}$: $\mathbf{D}(\mathbb{R}_+, \mathbb{R}) \to \mathbf{D}(\mathbb{R}_+, \mathbb{R})$ is Lipschitz continuous under the topology of uniform convergence over bounded intervals, and measurable under the Skorohod J_1-topology, and $\Xi_{\Upsilon, h}(\mathbf{C}(\mathbb{R}_+, \mathbb{R})) \subseteq \mathbf{C}(\mathbb{R}_+, \mathbb{R})$.

Proof. We will prove this lemma in the following three steps:
(a) the existence and uniqueness of the solution to (A1);
(b) $\Xi_{\Upsilon, h}$ is Lipschitz continuous with the topology of uniform convergence over bounded intervals;
(c) $\Xi_{\Upsilon, h}$ is measurable with respect to the Borel σ-field generated by the Skorohod J_1-topology.

We focus our analysis on the bounded interval $[0, T]$ for some $T > 0$. Let Λ^h be the Lipschitz constant of $h(\cdot)$ and Λ_T^{Υ} be the Lipschitz constant of $\Upsilon(\cdot)$ on the interval $[0, T]$. Let $\delta = 2/(3\Lambda_T^{\Upsilon}\Lambda^h)$.

Proof of (a): We first show the existence of a solution. Define $u_0(\cdot) = 0$ and $u_n(\cdot)$ iteratively by

$$u_{n+1}(t) = y(t) + \int_0^t h(\Upsilon(u_n)(s))\, ds$$

for all $n \geq 0$. Then,

$$u_{n+1}(t) - u_n(t) = \int_0^t [h(\Upsilon(u_n)(s)) - h(\Upsilon(u_{n-1})(s))]\, ds. \tag{A2}$$

Now, we will show that

$$\|u_{n+1} - u_n\|_{j\delta} \leq j^j n^j \left(\frac{2}{3}\right)^n \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta} \quad \text{for } j = 1, 2, \ldots, \lfloor \delta^{-1}T \rfloor + 1. \tag{A3}$$

For $j = 1$,

$$\|u_{n+1} - u_n\|_{\delta} \leq \Lambda^h \Lambda_T^{\Upsilon} \|u_n - u_{n-1}\|_{\delta} \times \delta \leq \frac{2}{3}\|u_n - u_{n-1}\|_{\delta}.$$

Since $h(0) = 0$ and $\Upsilon(0) = 0$, we have $\|u_1 - u_0\|_{\delta} = \|y\|_{\delta} \leq \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta}$. As a result,

$$\|u_{n+1} - u_n\|_{\delta} \leq \left(\frac{2}{3}\right)^n \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta} \leq n\left(\frac{2}{3}\right)^n \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta}.$$

Now, assume that we have proved (A3) for $j \leq k$. Then, for $j = k+1$, by (A2),

$$\begin{aligned}
\|u_{n+1} - u_n\|_{(k+1)\delta} &\leq \sum_{j=1}^k \Lambda^h \Lambda_T^{\Upsilon} \delta \|u_n - u_{n-1}\|_{j\delta} + \Lambda^h \Lambda_T^{\Upsilon} \delta \|u_n - u_{n-1}\|_{(k+1)\delta} \\
&= \sum_{j=1}^k \frac{2}{3}\|u_n - u_{n-1}\|_{j\delta} + \frac{2}{3}\|u_n - u_{n-1}\|_{(k+1)\delta} \\
&\leq \sum_{j=1}^k \frac{2}{3} j^j (n-1)^j \left(\frac{2}{3}\right)^{n-1} \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta} + \frac{2}{3}\|u_n - u_{n-1}\|_{(k+1)\delta} \\
&\leq k^{k+1} n^k \left(\frac{2}{3}\right)^n \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta} + \frac{2}{3}\|u_n - u_{n-1}\|_{(k+1)\delta}.
\end{aligned}$$

Since $\|u_1 - u_0\|_{(k+1)\delta} \le \|y\|_{(\lfloor \delta^{-1}T\rfloor+1)\delta}$, we have

$$\begin{aligned}\|u_{n+1}-u_n\|_{(k+1)\delta} &\le k^{k+1}\left(\sum_{i=0}^{n} i^k\right)\left(\frac{2}{3}\right)^n \|y\|_{(\lfloor \delta^{-1}T\rfloor+1)\delta}\\ &\le (k+1)^{k+1} n^{k+1}\left(\frac{2}{3}\right)^n \|y\|_{(\lfloor \delta^{-1}T\rfloor+1)\delta}.\end{aligned}$$

Hence we have proved (A3), which implies

$$\begin{aligned}\sum_{n=1}^{\infty}\|u_{n+1}-u_n\|_T &\le \sum_{n=1}^{\infty}\|u_{n+1}-u_n\|_{(\lfloor \delta^{-1}T\rfloor+1)\delta}\\ &\le \sum_{n=1}^{\infty}(\lfloor \delta^{-1}T\rfloor+1)^{\lfloor \delta^{-1}T\rfloor+1} n^{\lfloor \delta^{-1}T\rfloor+1}\left(\frac{2}{3}\right)^n \|y\|_{(\lfloor \delta^{-1}T\rfloor+1)\delta}\\ &< \infty.\end{aligned}$$

Thus $\{u_n(\cdot), n\in\mathbb{Z}_+\}$ is a Cauchy sequence. As $\mathbf{D}(\mathbb{R}_+,\mathbb{R})$ is a Banach space in the uniform metric, the sequence $\{u_n(\cdot), n\in\mathbb{Z}_+\}$ converges to the limit $u(\cdot)$, which is a solution to (A1).

The uniqueness of the solution is an immediate consequence of the Lipschitz continuity of $\Xi_{\Upsilon,h}$, which we will prove next.

Proof of (b): For any $y_1(\cdot), y_2(\cdot)\in\mathbf{D}(\mathbb{R}_+,\mathbb{R})$, the definition of δ and (A1) also imply that

$$\|\Xi_{\Upsilon,h}(y_2)-\Xi_{\Upsilon,h}(y_1)\|_\delta \le \|y_2-y_1\|_\delta + \frac{2}{3}\|\Xi_{\Upsilon,h}(y_2)-\Xi_{\Upsilon,h}(y_1)\|_\delta,$$

Hence $\|\Xi_{\Upsilon,h}(y_2)-\Xi_{\Upsilon,h}(y_1)\|_\delta \le 3\|y_2-y_1\|_\delta$. Suppose, for $i=0,1,\ldots,k$,

$$\|\Xi_{\Upsilon,h}(y_2)-\Xi_{\Upsilon,h}(y_1)\|_{i\delta} \le (3i)^i\|y_2-y_1\|_{i\delta}. \tag{A4}$$

We now show that (A4) holds for $i=k+1$. For any $t\in[0,(k+1)\delta]$, by the induction assumption and the definition of δ, we have

$$\|\Xi_{\Upsilon,h}(y_2)-\Xi_{\Upsilon,h}(y_1)\|_t \le \|y_2-y_1\|_t + \sum_{i=1}^{k}\frac{2}{3}(3i)^i\|y_2-y_1\|_t + \frac{2}{3}\|\Xi_{\Upsilon,h}(y_2)-\Xi_{\Upsilon,h}(y_1)\|_t.$$

This implies that (A4) holds for $i=k+1$. Continuing the induction until $k=\lfloor\delta^{-1}T\rfloor$ yields the Lipschitz continuity property of $\Xi_{\Upsilon,h}(\cdot)$.

Proof of (c): Define

$$\Theta(y,u)(t) = y(t) + \int_0^t h(\Upsilon(u)(s))\,ds.$$

First, we prove the function $\Theta(\cdot)$ is measurable with respect to the Borel σ-field generated by the Skorohod J_1-topology in $\mathbf{D}^2(\mathbb{R}_+,\mathbb{R})$ and $\mathbf{D}(\mathbb{R}_+,\mathbb{R})$. Define $\Pi(y,u)(t)=y(t)+\int_0^t h(u(s))\,ds$. It is clear that $\Pi(\cdot)$ is measurable (in fact, continuous) in the Skorohod J_1-topology. Since $\Theta(y,u)=\Pi(y,\Upsilon(u))$ and $\Upsilon(\cdot)$ are measurable under the Skorohod J_1-topology, the measurability of $\Theta(\cdot)$ is proved.

We know that $\Phi_g(y)=\lim_{n\to\infty}\Theta^n(y,0)$, where $\Theta^n(\cdot)$ is iteratively defined by

$$\Theta^n(y,u)=\Theta(y,\Theta^{n-1}(y,u)),\quad n=1,2,\ldots$$

with $\Theta^0(y,u)=u$. According to Theorem 2 on page 14 of Chow and Teicher [4], we can prove by induction that $\Theta^n(y,0)$ is measurable for each n. Since $\Xi_{\Upsilon,h}(y)$ is the limit of $\Theta^n(y,0)$ under the topology of uniform convergence over bounded intervals, it is also the limit of $\Theta^n(y,0)$ under the Skorohod J_1-topology. By Theorem 4.2.2 of Dudley [11], we know that $\Xi_{\Upsilon,h}(\cdot)$ is measurable with respect to the Borel σ-field generated by the Skorohod J_1-topology. □

Proof of Lemma 2. Note the special case where $g\equiv 0$ gives us the conventional Skorohod mapping. In other words, for any $y\in\mathbf{D}(\mathbb{R}_+,\mathbb{R})$ with $y(0)\ge 0$, there exists a unique $(a,b)\in\mathbf{D}(\mathbb{R}_+,\mathbb{R})$ such that

$$\begin{aligned}&a(t)=y(t)+b(t),\\ &\int_0^\infty a(t)\,db(t)=0,\\ &a(t)\ge 0,\quad \forall\, t\ge 0.\end{aligned}$$

Define the mapping $(\Phi,\Psi)\colon \mathbf{D}(\mathbb{R}_+,\mathbb{R})\to\mathbf{D}(\mathbb{R}_+,\mathbb{R}^2)$ by $(\Phi,\Psi)(y)=(a,b)$. Then, (Φ,Ψ) is Lipschitz continuous in the topology of uniform convergence over bounded intervals and the Skorohod J_1-topology.

To deal with the integral Equation (19), we use Lemma 5 with $h(\cdot)=g(\cdot)$ and $\Upsilon(\cdot)=\Phi(\cdot)$. Then, we can obtain a mapping $\bar\Phi(\cdot)$ given by $u=\bar\Phi(y)$ with

$$u(t)=y(t)+\int_0^t g(\Phi(u)(s))\,ds\quad \text{for } y\in\mathbf{D}(\mathbb{R}_+,\mathbb{R}). \tag{A5}$$

Clearly, $(x,z)=(\Phi(\bar\Phi(y)),\Psi(\bar\Phi(y)))$ (that is, $\Phi_g(\cdot)=(\Phi,\Psi)\circ\bar\Phi(\cdot)$) is a solution to (19). Other properties (measurability, Lipschitz continuity and $\Phi_g(\mathbf{C}(\mathbb{R}_+,\mathbb{R}))\subseteq\mathbf{C}(\mathbb{R}_+,\mathbb{R})$) can easily be proved from the properties of $(\Phi,\Psi)(\cdot)$ and $\bar\Phi(\cdot)$. □

Proof of Lemma 3. Note the special case where $g(\cdot)\equiv 0$ was proved by Reed [20] (cf. Proposition 7 there). In other words, for any $y(\cdot)\in\mathbf{D}(\mathbb{R}_+,\mathbb{R})$, there exists a unique $u(\cdot)\in\mathbf{D}(\mathbb{R}_+,\mathbb{R})$ such that

$$u(t)=y(t)+\int_0^t (u(t-s))^-\,dM(s).$$

Define the mapping $\Phi_M(\cdot)$ by $u(\cdot)=\Phi_M(y)(\cdot)$. Then $\Phi_M(\cdot)$ is Lipschitz continuous under the topology of uniform convergence over bounded intervals, measurable with respect to the Borel σ-field generated by the Skorohod J_1-topology.

To deal with the integral equation (32), we use Lemma 5 with $h(t)=g(t^+)$ for $t\in\mathbb{R}$, and $\Upsilon(\cdot)=\Phi_M(\cdot)$. Then, we can obtain a mapping $\Psi_g(\cdot)$ given by $a(\cdot)=\Psi_g(y)(\cdot)$ with

$$a(t)=y(t)+\int_0^t g((\Phi_M(a)(s))^+)\,ds \quad \text{for } y\in\mathbf{D}(\mathbb{R}_+,\mathbb{R}). \tag{A6}$$

Clearly, $x(\cdot)=\Phi_M(\Psi_g(y))(\cdot)$ (that is, $\Phi_{M,g}(\cdot)=\Phi_M\circ\Psi_g(\cdot)$) is a solution to (32). The other properties (measurability, Lipschitz continuity, and $\Phi_{M,g}(\mathbf{C}(\mathbb{R}_+,\mathbb{R}))\subseteq\mathbf{C}(\mathbb{R}_+,\mathbb{R})$) can easily be proved from the properties of $\Phi_M(\cdot)$ and $\Psi_g(\cdot)$. □

Appendix B. Technical proofs.

B.1. Proof of Lemma 4. First, introduce auxiliary processes $\tilde{\mathcal{T}}_0^n=\{\tilde{\mathcal{T}}_0^n(t)\colon t\ge 0\}$ and $\tilde{U}_0^n=\{\tilde{U}_0^n(t)\colon t\ge 0\}$

$$\tilde{\mathcal{T}}_0^n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor n\mu(T+1)\rfloor}\left(\mathbf{1}_{\{u_i^n+\tau_i^n>t\}}-(1-H_\star(t-\tau_i^n))\right),$$

$$\tilde{U}_0^n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor n\mu(T+1)\rfloor}\left(\mathbf{1}_{\{u_i^n+i/(n\mu)>t\}}-\left(1-H_\star\left(t-\frac{i}{n\mu}\right)\right)\right).$$

By the weak convergence of the empirical processes (see Chapter 3 in Shorack and Wellner [25]), as $n\to\infty$,

$$\tilde{U}_0^n\Rightarrow\tilde{U}, \tag{B1}$$

where $\tilde{U}=\{\tilde{U}(t)\colon t\ge 0\}$ is a Gaussian process with continuous sample paths, zero mean, and the same covariance function as $\tilde{\mathcal{T}}$. By (36), we have

$$\sup_{0\le t\le T}|\tilde{\mathcal{T}}^n(t)-\tilde{\mathcal{T}}_0^n(t)|\Rightarrow 0 \qquad\text{and}\qquad \sup_{0\le t\le T}|\tilde{U}^n(t)-\tilde{U}_0^n(t)|\Rightarrow 0. \tag{B2}$$

In view of (B1) and (B2), to prove the lemma, it remains to be shown that

$$\sup_{0\le t\le T}|\tilde{\mathcal{T}}_0^n(t)-\tilde{U}_0^n(t)|\Rightarrow 0. \tag{B3}$$

According to (B1) and Theorem 13.1 of Billingsley [3], the proof of (B3) is derived in two steps.

Step 1. Establish the convergence of all the finite-dimensional distributions of $\{\tilde{\mathcal{T}}_0^n, n\in\mathbb{Z}_+\}$ with the same limit as $\{\tilde{\mathcal{U}}_0^n, n\in\mathbb{Z}_+\}$.

By (36), for any $T>0$,

$$\max_{1\le i\le\lfloor n\mu T\rfloor}\left|\tau_i^n-\frac{i}{n\mu}\right|\Rightarrow 0.$$

This implies that there is a positive sequence $\{\varepsilon^n, n\in\mathbb{Z}_+\}$ with $\varepsilon^n\downarrow 0$ such that

$$\sup_{0\le t\le T}|\tilde{\mathcal{T}}_0^n(t)+\tilde{\mathcal{T}}_1^n(t,\varepsilon^n)|\Rightarrow 0 \quad\text{and}\quad \sup_{0\le t\le T}|\tilde{U}_0^n(t)+\tilde{U}_1^n(t,\varepsilon^n)|\Rightarrow 0, \tag{B4}$$

where

$$\tilde{\mathcal{T}}_1^n(t,\varepsilon^n) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor n\mu(T+1)\rfloor}\mathbf{1}_{\{(i-1)/(n\mu)-\varepsilon^n\le\tau_i^n\le i/(n\mu)+\varepsilon^n\}}(\mathbf{1}_{\{\tau_i^n+u_i^n\le t\}}-H_\star(t-\tau_i^n)),$$

$$\tilde{U}_1^n(t,\varepsilon^n) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor n\mu(T+1)\rfloor}\mathbf{1}_{\{(i-1)/(n\mu)-\varepsilon^n\le\tau_i^n\le i/(n\mu)+\varepsilon^n\}}\left(\mathbf{1}_{\{i/(n\mu)+u_i^n\le t\}}-H_\star\left(t-\frac{i}{n\mu}\right)\right).$$

Thus, for this step, we just need to show that for each $t\le T$, as $n\to\infty$,

$$\tilde{\mathcal{T}}_1^n(t,\varepsilon^n)-\tilde{U}_1^n(t,\varepsilon^n)\Rightarrow 0. \tag{B5}$$

Note that

$$
\begin{aligned}
&\mathbb{P}\{|\tilde{\mathcal{T}}_1^n(t,\varepsilon^n)-\tilde{U}_1^n(t,\varepsilon^n)|>\delta\}\\
&\quad\le \frac{1}{n\delta^2}\mathbb{E}\Bigg(\sum_{i=1}^{\lfloor n\mu(T+1)\rfloor}\mathbf{1}_{\{(i-1)/(n\mu)-\varepsilon^n\le\tau_i^n\le i/(n\mu)+\varepsilon^n\}}\Bigg[(\mathbf{1}_{\{\tau_i^n+u_i^n\le t\}}-H_\star(t-\tau_i^n))-\bigg(\mathbf{1}_{\{i/(n\mu)+u_i^n\le t\}}-H_\star\bigg(t-\frac{i}{n\mu}\bigg)\bigg)\Bigg]\Bigg)^2\\
&\quad\le \frac{1}{n\delta^2}\sum_{i=1}^{\lfloor n\mu(T+1)\rfloor}\mathbb{E}\Bigg(\mathbf{1}_{\{(i-1)/(n\mu)-\varepsilon^n\le\tau_i^n\le i/(n\mu)+\varepsilon^n\}}\Bigg[(\mathbf{1}_{\{\tau_i^n+u_i^n\le t\}}-H_\star(t-\tau_i^n))-\bigg(\mathbf{1}_{\{i/(n\mu)+u_i^n\le t\}}-H_\star\bigg(t-\frac{i}{n\mu}\bigg)\bigg)\Bigg]^2\Bigg)\\
&\quad\le \frac{4}{n\delta^2}\sum_{i=1}^{\lfloor n\mu(T+1)\rfloor}\bigg(H_\star\bigg(t-\frac{i-1}{n\mu}+\varepsilon^n\bigg)-H_\star\bigg(t-\frac{i}{n\mu}-\varepsilon^n\bigg)\bigg). \qquad \text{(B6)}
\end{aligned}
$$

Consider the set of intervals given by

$$
\bigg\{\bigg(\bigg(t-\frac{i}{n\mu}-\varepsilon^n\bigg)^+,\bigg(t-\frac{i-1}{n\mu}+\varepsilon^n\bigg)^+\bigg],\ i=1,\ldots,\lfloor n\mu(T+1)\rfloor\bigg\}.
$$

Note that the intervals $(t-i/(n\mu)-\varepsilon^n,\ t-(i-1)/(n\mu)+\varepsilon^n]$ and $(t-j/(n\mu)-\varepsilon^n,\ t-(j-1)/(n\mu)+\varepsilon^n]$ with $i<j$ are disjoint if $j\ge i+1+\lceil 2n\mu\varepsilon^n\rceil$. Thus the set can be partitioned into $2+\lceil 2n\mu\varepsilon^n\rceil$ groups such that any two intervals in each group are disjoint. Therefore

$$
\begin{aligned}
&\frac{4}{n\delta^2}\sum_{i=1}^{\lfloor n\mu(T+1)\rfloor}\bigg(H_\star\bigg(t-\frac{i-1}{n\mu}+\varepsilon^n\bigg)-H_\star\bigg(t-\frac{i}{n\mu}-\varepsilon^n\bigg)\bigg)\\
&\quad\le\frac{4}{n\delta^2}\times(2+\lceil 2n\mu\varepsilon^n\rceil)\to 0\quad\text{as } n\to\infty.
\end{aligned}
$$

Thus (B5) is proved by (B6).

Step 2. Establish the tightness of $\{\tilde{\mathcal{T}}_0^n, n\in\mathbb{Z}_+\}$.

We first present a simple proof of the tightness, which is itself interesting and requires that function $H_\star(\cdot)$ is locally Lipschitz continuous. Note that for any $t_1\le t\le t_2$,

$$
\begin{aligned}
&\mathbb{E}[(\tilde{\mathcal{T}}_0^n(t)-\tilde{\mathcal{T}}_0^n(t_1))^2\times(\tilde{\mathcal{T}}_0^n(t_2)-\tilde{\mathcal{T}}_0^n(t))^2]\\
&\quad=\frac{1}{n^2}\sum_{i,\,j=1}^{\lfloor n\mu(T+1)\rfloor}\mathbb{E}[(\mathbf{1}_{\{t_1<\tau_i^n+u_i^n\le t\}}-[H_\star(t-\tau_i^n)-H_\star(t_1-\tau_i^n)])^2\\
&\qquad\qquad\times(\mathbf{1}_{\{t<\tau_j^n+u_j^n\le t_2\}}-[H_\star(t_2-\tau_j^n)-H_\star(t-\tau_j^n)])^2]\\
&\qquad+\frac{2}{n^2}\sum_{i\ne j}\mathbb{E}[(\mathbf{1}_{\{t_1<\tau_i^n+u_i^n\le t\}}-[H_\star(t-\tau_i^n)-H_\star(t_1-\tau_i^n)])\\
&\qquad\qquad\times(\mathbf{1}_{\{t<\tau_i^n+u_i^n\le t_2\}}-[H_\star(t_2-\tau_i^n)-H_\star(t-\tau_i^n)])\\
&\qquad\qquad\times(\mathbf{1}_{\{t_1<\tau_j^n+u_j^n\le t\}}-[H_\star(t-\tau_j^n)-H_\star(t_1-\tau_j^n)])\\
&\qquad\qquad\times(\mathbf{1}_{\{t<\tau_j^n+u_j^n\le t_2\}}-[H_\star(t_2-\tau_j^n)-H_\star(t-\tau_j^n)])]\\
&\quad\le 3\mu^2(T+1)^2\sup_{0\le s\le T}(H_\star(t_2-s)-H_\star(t_1-s))^2.
\end{aligned}
$$

When $H_\star(\cdot)$ is locally Lipschitz continuous, the right-hand side of the above inequality can be bounded by $\Lambda(t_2-t_1)^2$ for some constant Λ. Therefore the tightness of $\{\tilde{\mathcal{T}}_0^n, n\in\mathbb{Z}_+\}$ follows from Theorem 13.5 of Billingsley [3].

Now, consider the case without the local Lipschitz continuity of $H_\star(\cdot)$. Note that by (B2), the tightness of $\{\tilde{\mathcal{T}}_0^n, n\in\mathbb{Z}_+\}$ and the tightness of $\{\tilde{\mathcal{T}}^n, n\in\mathbb{Z}_+\}$ are equivalent. Therefore it is sufficient to prove the tightness of $\{\tilde{\mathcal{T}}^n, n\in\mathbb{Z}_+\}$. Let

$$
\begin{aligned}
\tilde{U}^n(t,x)&=\frac{1}{\sqrt{n}}\sum_{i=1}^{C^n(t)}(\mathbf{1}_{\{u_i^n\le x\}}-H_\star(x)),\qquad t\ge 0,\ \ x\ge 0,\\
\tilde{\mathcal{T}}_c^n(t)&=\frac{1}{\sqrt{n}}\sum_{i=1}^{C^n(t)}\bigg(\mathbf{1}_{\{0<u_i^n\le t-\tau_i^n\}}-\int_{0+}^{u_i^n\wedge(t-\tau_i^n)^+}\frac{dH_\star(s)}{1-H_\star(s-)}\bigg).
\end{aligned}
$$

Then, we have

$$
\tilde{\mathcal{T}}^n(t)=\int_0^t\frac{\tilde{U}^n(t-s,s-)}{1-H_\star(s-)}\,dH_\star(s-)-\frac{1}{\sqrt{n}}\sum_{i=1}^{C^n(t)}(\mathbf{1}_{\{u_i^n=0\}}-H_\star(0))-\tilde{\mathcal{T}}_c^n(t). \qquad \text{(B7)}
$$

Thus the tightness of $\{\tilde{\mathscr{T}}^n, n \in \mathbb{Z}_+\}$ follows from the tightness of the three terms on the right-hand side of (B7). The tightness of the second term in (B7) follows directly from (36). For the first term in (B7), we can divide it into two parts (for any $\varepsilon > 0$):

$$\int_0^t \frac{\tilde{U}^n(t-s, s-)}{1-H_\star(s-)} \mathbf{1}_{\{H_\star(s-)>1-\varepsilon\}}\, dH_\star(s-) + \int_0^t \frac{\tilde{U}^n(t-s, s-)}{1-H_\star(s-)} \mathbf{1}_{\{H_\star(s-)\le 1-\varepsilon\}}\, dH_\star(s-). \tag{B8}$$

In the same way Krichagina and Puhalskii [14] proved their Lemma 3.4, we obtain the tightness of the second term in (B8) and

$$\lim_{\varepsilon\downarrow 0} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{t\le T} \left| \int_0^t \frac{\tilde{V}^n(t-s, s-)}{1-H_\star(s-)} \mathbf{1}_{\{H_\star(s-)>1-\varepsilon\}}\, dH_\star(s-) \right| > \delta \right) = 0.$$

Finally, we prove the tightness of the third term on the right-hand side of (B7). Let

$$\mathscr{F}_t^n = \sigma\{C^n(s)\colon s \le t\} \vee \sigma\{\mathbf{1}_{\{\tau_i^n + u_i^n \le s\}}\colon s \le t,\ i = 1, \ldots, C^n(t)\}.$$

Then, for each positive integer k,

$$\tilde{\mathscr{T}}^n_{c,k}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{C^n(t)\wedge k} \left(\mathbf{1}_{\{0<u_i^n \le t-\tau_i^n\}} - \int_{0+}^{u_i^n \wedge (t-\tau_i^n)^+} \frac{dH_\star(s)}{1-H_\star(s-)} \right)$$

is an $\mathscr{F}_t^n$-square integrable martingale with the predictable quadratic variation process

$$\langle \tilde{\mathscr{T}}^n_{c,k} \rangle (t) = \frac{1}{n} \sum_{i=1}^{C^n(t)\wedge k} \int_{0+}^{u_i^n \wedge (t-\tau_i^n)^+} \frac{1-H_\star(s)}{(1-H_\star(s-))^2}\, dH_\star(s).$$

Then, the tightness of the third term follows the same argument used by Krichagina and Puhalskii [14] in the proof of their Lemma 3.7.

Combining Steps 1–2 yields the convergence of $\{\tilde{\mathscr{T}}^n_0, n \in \mathbb{Z}_+\}$ with the same limit as that of $\{\tilde{\mathscr{U}}^n_0, n \in \mathbb{Z}_+\}$. Therefore we have (B3), which implies the lemma. □

B.2. Proofs of Propositions 4–6. In this subsection, we provide the proofs for Propositions 4–6. To prove the propositions, we need the following lemma. For each $\delta > 0$, let

$$\tilde{L}^n_\delta(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \left(\mathbf{1}_{\{\gamma_i^n \le \delta/\sqrt{n}\}} - F^n\left(\frac{\delta}{\sqrt{n}}\right) \right).$$

Lemma 6. *For a fixed $\delta > 0$, if $F^n(\delta/\sqrt{n}) \to 0$ as $n \to \infty$ (which is implied by (4)), then for any $T > 0$,*

$$\sup_{0\le t\le T} |\tilde{L}^n_\delta(t)| \Rightarrow 0 \quad \textit{as } n \to \infty.$$

Proof. Denote $p_n = F^n(\delta/\sqrt{n})$ and $X_{ni} = \mathbf{1}_{\{\gamma_i^n \le \delta/\sqrt{n}\}} - p_n$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}\left\{ \sup_{0\le t\le T} |\tilde{L}^n_\delta(t)| > \varepsilon \right\} = \mathbb{P}\left\{ \max_{1\le k\le \lfloor nT \rfloor} \left| \tilde{L}^n_\delta\left(\frac{k}{n}\right) \right| > \varepsilon \right\}. \tag{B9}$$

Note that for each n, $\{X_{ni}, i \in \mathbb{Z}_+\}$ are independent and identically distributed random variables with $\mathbb{E}(X_{ni})^2 = p_n(1-p_n) < \infty$. By the Kolmogorov's inequality (see page 133 of Chow and Teicher [4])

$$\begin{aligned} \mathbb{P}\left\{ \max_{1\le k\le \lfloor nT \rfloor} \left| \tilde{L}^n_\delta\left(\frac{k}{n}\right) \right| > \varepsilon \right\} &\le \frac{1}{\varepsilon^2} \mathbb{E}\, |\tilde{L}^n_\delta(\lfloor T \rfloor)|^2 \\ &= \frac{1}{\varepsilon^2 n} \sum_{i=1}^{\lfloor nT \rfloor} \mathbb{E}(X_{ni})^2 \\ &\le \frac{1}{\varepsilon^2} T p_n(1-p_n) \to 0, \end{aligned}$$

according to the assumption that $p_n \to 0$ as $n \to \infty$. Thus the lemma follows from (B9). □

Proof of Proposition 4. First, we look at the sequence of the scaled virtual waiting times $\{\tilde{\omega}^n, n \in \mathbb{Z}_+\}$. According to FCFS, for any $s \in (0, \tilde{\omega}^n(t))$, customers who arrive during the interval $(t, t + s/\sqrt{n})$ will not receive service until $t + s/\sqrt{n}$, hence they either stay in the queue or have abandoned by $t + s/\sqrt{n}$. Therefore for any $s \in (0, \tilde{\omega}^n(t))$,

$$E^n\left(t + \frac{s}{\sqrt{n}}\right) - E^n(t) \le Q^n\left(t + \frac{s}{\sqrt{n}}\right) + \sum_{i=E^n(t)+1}^{E^n(t+s/\sqrt{n})} \mathbf{1}_{\{\gamma_i^n \le s/\sqrt{n}\}}. \tag{B10}$$

The above inequality (B10) implies that

$$\begin{aligned}\tilde{E}^n\left(t+\frac{s}{\sqrt{n}}\right)-\tilde{E}^n(t)+\frac{\lambda^n s}{n} \le \tilde{Q}^n\left(t+\frac{s}{\sqrt{n}}\right)+\tilde{L}^n_s\left(\bar{E}^n\left(t+\frac{s}{\sqrt{n}}\right)\right)-\tilde{L}^n_s(\bar{E}^n(t))\\ +\sqrt{n}\cdot F^n\left(\frac{s}{\sqrt{n}}\right)\cdot\left(\bar{E}^n\left(t+\frac{s}{\sqrt{n}}\right)-\bar{E}^n(t)\right).\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{P}\left\{\sup_{0\le t\le T}\tilde{\omega}^n(t)>s\right\} &\le \mathbb{P}\left\{\inf_{0\le t\le T}\left[\tilde{E}^n\left(t+\frac{s}{\sqrt{n}}\right)-\tilde{E}^n(t)+\frac{\lambda^n s}{n}-\tilde{Q}^n\left(t+\frac{s}{\sqrt{n}}\right)\right.\right.\\ &\qquad -\tilde{L}^n_s\left(\bar{E}^n\left(t+\frac{s}{\sqrt{n}}\right)\right)+\tilde{L}^n_s(\bar{E}^n(t))\\ &\qquad \left.\left.-\sqrt{n}\cdot F^n\left(\frac{s}{\sqrt{n}}\right)\cdot\left(\bar{E}^n\left(t+\frac{s}{\sqrt{n}}\right)-\bar{E}^n(t)\right)\right]\le 0\right\}\\ &\le \mathbb{P}\left\{\inf_{0\le t\le T}\left[\tilde{E}^n\left(t+\frac{s}{\sqrt{n}}\right)-\tilde{E}^n(t)+\frac{\lambda^n s}{n}\right]\le\frac{\mu s}{2}\right\}\\ &\quad +\mathbb{P}\left\{\sup_{0\le t\le T}\tilde{Q}^n\left(t+\frac{s}{\sqrt{n}}\right)\ge\frac{\mu s}{12}\right\}\\ &\quad +\mathbb{P}\left\{\sup_{0\le t\le T}\left|\tilde{L}^n_s\left(\bar{E}^n\left(t+\frac{s}{\sqrt{n}}\right)\right)-\tilde{L}^n_s(\bar{E}^n(t))\right|\ge\frac{\mu s}{12}\right\}\\ &\quad +\mathbb{P}\left\{\sup_{0\le t\le T}\sqrt{n}\cdot F^n\left(\frac{s}{\sqrt{n}}\right)\cdot\left(\bar{E}^n\left(t+\frac{s}{\sqrt{n}}\right)-\bar{E}^n(t)\right)\ge\frac{\mu s}{12}\right\},\end{aligned}$$

where μ is given by (2). It follows from (2)–(3) that as $n\to\infty$,

$$\mathbb{P}\left\{\inf_{0\le t\le T}\left[\tilde{E}^n\left(t+\frac{s}{\sqrt{n}}\right)-\tilde{E}^n(t)+\frac{\lambda^n s}{n}\right]\le\frac{\mu s}{2}\right\}\to 0, \tag{B11}$$

Lemma 6 and (53) imply that as $n\to\infty$,

$$\mathbb{P}\left\{\sup_{0\le t\le T}\left|\tilde{L}^n_s\left(\bar{E}^n\left(t+\frac{s}{\sqrt{n}}\right)\right)-\tilde{L}^n_s(\bar{E}^n(t))\right|\ge\frac{\mu s}{12}\right\}\to 0. \tag{B12}$$

By (4) and (53), as $n\to\infty$,

$$\mathbb{P}\left\{\sup_{0\le t\le T}\sqrt{n}\cdot F^n\left(\frac{s}{\sqrt{n}}\right)\cdot\left(\bar{E}^n\left(t+\frac{s}{\sqrt{n}}\right)-\bar{E}^n(t)\right)\ge\frac{\mu s}{12}\right\}\to 0. \tag{B13}$$

Hence the stochastic boundedness of $\{\tilde{\omega}^n, n\in\mathbb{Z}_+\}$ follows from assumption (6) and (B11)–(B13).

Next, we look at $\{\tilde{F}^n_\omega, n\in\mathbb{Z}_+\}$. It is sufficient to show that for any given $T>0$,

$$\lim_{\Gamma\to\infty}\limsup_{n\to\infty}\mathbb{P}\left\{\sup_{0\le i\le E^n(T)}\sqrt{n}F^n(\omega^n_i)\ge\Gamma\right\}=0, \tag{B14}$$

where ω^n_i is the offered waiting time. According to Lemma 3.2 of Dai and He [7] and the monotonicity of the distribution function $F^n(\cdot)$,

$$\mathbb{P}\left\{\sup_{0\le i\le E^n(T)}\sqrt{n}F^n(\omega^n_i)\ge\Gamma\right\}\le\mathbb{P}\left\{\sup_{0\le t\le T}\sqrt{n}F^n\left(\frac{1}{\sqrt{n}}\tilde{\omega}^n(t)\right)\ge\Gamma\right\}. \tag{B15}$$

Thus, it is enough to prove

$$\lim_{\Gamma\to\infty}\limsup_{n\to\infty}\mathbb{P}\left\{\sup_{0\le t\le T}\sqrt{n}F^n\left(\frac{1}{\sqrt{n}}\tilde{\omega}^n(t)\right)\ge\Gamma\right\}=0. \tag{B16}$$

Next, we note

$$\mathbb{P}\left\{\sup_{0\le t\le T}\sqrt{n}F^n\left(\frac{1}{\sqrt{n}}\tilde{\omega}^n(t)\right)\ge\Gamma\right\}\le\mathbb{P}\left\{\sqrt{n}F^n\left(\frac{1}{\sqrt{n}}\Gamma_1\right)\ge\Gamma\right\}+\mathbb{P}\left\{\sup_{0\le t\le T}\tilde{\omega}^n(t)\ge\Gamma_1\right\}.$$

For any given $\varepsilon>0$, by stochastic boundedness of $\{\tilde{\omega}^n, n\in\mathbb{Z}_+\}$, we can choose Γ_1 such that

$$\limsup_{n\to\infty}\mathbb{P}\left\{\sup_{0\le t\le T}\tilde{\omega}^n(t)\ge\Gamma_1\right\}\le\frac{\varepsilon}{2}.$$

Now, from (4), for the Γ_1 fixed above,

$$\lim_{\Gamma\to\infty}\limsup_{n\to\infty}\mathbb{P}\left\{\sqrt{n}F^n\left(\frac{1}{\sqrt{n}}\Gamma_1\right)\ge\Gamma\right\}=0.$$

Thus, for any given $\varepsilon > 0$, there is a Γ_0 such that when $\Gamma \ge \Gamma_0$,

$$\limsup_{n\to\infty} \mathbb{P}\left\{\sup_{0\le t\le T} \sqrt{n}F^n\left(\frac{1}{\sqrt{n}}\tilde{\omega}^n(t)\right) \ge \Gamma\right\} \le \varepsilon. \tag{B17}$$

This completes the proof of (B16). Thus (B14) is proved due to (B15). Hence the proof of the proposition is completed. □

An immediate consequence of (B14) is that

$$\mathbb{E}\left[\sup_{0\le i\le E^n(T)} F^n(\omega_i^n)\right] \to 0 \quad \text{as } n\to\infty. \tag{B18}$$

This will help to prove Lemma 7 below, which is an extension of Proposition 4.2 of Dai and He [7], where $F^n(\cdot) = F(\cdot)$. The general approach of the proof is the same whether $F^n(\cdot)$'s are the same or vary with n. That is, we need to use the martingale convergence theorem (cf. Lemma 4.3 of Dai and He [7] and Whitt [33]). The key condition for applying the theorem is (B18). We thus present the result without repeating the proof.

Lemma 7. *Under assumptions* (2)–(4) *and* (6),

$$\sup_{0\le t\le T}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor}(\mathbf{1}_{\{\gamma_i^n\le\omega_i^n\}} - F^n(\omega_i^n))\cdot g(\omega_i^n)\right| \Rightarrow 0 \quad \textit{as } n\to\infty,$$

where $g(\cdot)\colon \mathbb{R}_+\to\mathbb{R}_+$ *is a Borel measurable function such that* $0\le g(t)\le 1$ *for all* $t\in\mathbb{R}_+$.

Similar to Dai et al. [9], we define the process

$$\zeta^n(t) = \inf\{s\ge 0\colon s+\omega^n(s)\ge t\}.$$

It is clear that $\zeta^n \in \mathbf{D}(\mathbb{R}_+,\mathbb{R})$ and is nondecreasing for each $n\in\mathbb{Z}_+$.

Lemma 8. *Under assumptions* (2)–(4) *and* (6), *as* $n\to\infty$

$$\sup_{0\le t\le T}|\zeta^n(t)-t| \Rightarrow 0.$$

Proof. By the definition of $\zeta^n(t)$, for any $t\ge 0$,

$$0\le t-\zeta^n(t)\le \omega^n(\zeta^n(t)).$$

Hence

$$\sup_{0\le t\le T}|\zeta^n(t)-t| \le \sup_{0\le t\le T}\omega^n(\zeta^n(t)) \le \sup_{0\le t\le T}\omega^n(t).$$

Thus the result follows from Proposition 4. □

Proof of Proposition 5. According to Lemma 7, it suffices to show that as $n\to\infty$,

$$\sup_{0\le t\le T}\left|\tilde{G}^n(t) - \frac{1}{\sqrt{n}}\sum_{i=1}^{E^n(t)}\mathbf{1}_{\{\gamma_i^n\le\omega_i^n\}}\right| \Rightarrow 0.$$

As a customer arriving at the system before time $\zeta^n(t)$ must have either entered service or abandoned the queue by time t, we have the following relationship:

$$\sum_{i=1}^{E^n(\zeta^n(t)-)}\mathbf{1}_{\{\gamma_i^n\le\omega_i^n\}} \le G^n(t) \le \sum_{i=1}^{E^n(t)}\mathbf{1}_{\{\gamma_i^n\le\omega_i^n\}}.$$

Hence, it is enough to prove that as $n\to\infty$,

$$\sup_{0\le t\le T}\frac{1}{\sqrt{n}}\sum_{i=E^n(\zeta^n(t)-)+1}^{E^n(t)}\mathbf{1}_{\{\gamma_i^n\le\omega_i^n\}} \Rightarrow 0. \tag{B19}$$

Note that

$$\begin{aligned}\sup_{0\le t\le T}\frac{1}{\sqrt{n}}\sum_{i=E^n(\zeta^n(t)-)+1}^{E^n(t)}\mathbf{1}_{\{\gamma_i^n\le\omega_i^n\}} &= \sup_{0\le t\le T}\frac{1}{\sqrt{n}}\sum_{i=E^n(\zeta^n(t)-)+1}^{E^n(t)}(\mathbf{1}_{\{\gamma_i^n\le\omega_i^n\}} - F^n(\omega_i^n))\\ &\quad + \sup_{0\le t\le T}\frac{1}{\sqrt{n}}\sum_{i=E^n(\zeta^n(t)-)+1}^{E^n(t)}F^n(\omega_i^n).\end{aligned} \tag{B20}$$

By Lemma 7, the first term on the right-hand side of (B20) will converge to 0. For the second term,

$$\sup_{0\le t\le T}\frac{1}{\sqrt{n}}\sum_{i=E^n(\zeta^n(t)-)+1}^{E^n(t)}F^n(\omega_i^n) \le \left(\sup_{0\le t\le T}[\bar{E}^n(t)-\bar{E}^n(\zeta^n(t)-)]\right)\cdot\left(\sup_{0\le i\le E^n(T)}\sqrt{n}F^n(\omega_i^n)\right), \tag{B21}$$

which weakly converges to 0 due to (53), Proposition 4 and Lemma 8. Thus (B19) holds and the proof is completed. □

Proof of Proposition 6. First, note that $\omega^n(t) = \tilde{\omega}^n(t)/\sqrt{n}$. It directly follows from the stochastic boundedness of $\{\tilde{\omega}^n, n \in \mathbb{Z}_+\}$ (Proposition 4) that

$$\sup_{0\le t\le T} \omega^n(t) \Rightarrow 0. \tag{B22}$$

By the definition of $\omega^n(t)$, we have

$$\begin{aligned} Q^n(t+\omega^n(t)) &\le E^n(t+\omega^n(t)) - E^n(t) \\ &\le Q^n((t+\omega^n(t))-) + \left(E^n(t+\omega^n(t)) - E^n\left(t+\omega^n(t)-\frac{1}{n}\right)\right) + \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} \mathbf{1}_{\{\gamma_i^n \le \omega_i^n\}}. \end{aligned} \tag{B23}$$

Note that

$$\frac{1}{\sqrt{n}}(E^n(t+\omega^n(t)) - E^n(t)) = \tilde{E}^n(t+\omega^n(t)) - \tilde{E}^n(t) + \frac{\lambda^n}{n}\cdot\sqrt{n}\cdot\omega^n(t), \tag{B24}$$

$$\frac{1}{\sqrt{n}} \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} \mathbf{1}_{\{\gamma_i^n \le \omega_i^n\}} = \frac{1}{\sqrt{n}} \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} (\mathbf{1}_{\{\gamma_i^n \le \omega_i^n\}} - F^n(\omega_i^n)) + \frac{1}{\sqrt{n}} \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} F^n(\omega_i^n). \tag{B25}$$

By (3) and (B22),

$$\sup_{0\le t\le T} |\tilde{E}^n(t+\omega^n(t)) - \tilde{E}^n(t)| \Rightarrow 0 \quad \text{as } n\to\infty. \tag{B26}$$

By (2) and Proposition 4,

$$\left|\frac{\lambda^n}{n}\cdot\sqrt{n}\cdot\omega^n(t) - \mu\tilde{\omega}^n(t)\right| \Rightarrow 0 \quad \text{as } n\to\infty. \tag{B27}$$

It follows from (2) and (3) that

$$\begin{aligned} &\sup_{0\le t\le T} \frac{1}{\sqrt{n}}\left|E^n(t+\omega^n(t)) - E^n\left(t+\omega^n(t)-\frac{1}{n}\right)\right| \\ &\quad\le \sup_{0\le t\le T}\left|\tilde{E}^n(t+\omega^n(t)) - \tilde{E}^n\left(t+\omega^n(t)-\frac{1}{n}\right)\right| + \frac{\lambda^n}{\sqrt{n^3}} \Rightarrow 0 \quad \text{as } n\to\infty. \end{aligned} \tag{B28}$$

Note that the inequality (B21) also holds with $(\zeta^n(t)-, t)$ replaced by $(t, t+\omega^n(t))$, therefore by (B22) as $n\to\infty$,

$$\sup_{0\le t\le T} \frac{1}{\sqrt{n}} \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} F^n(\omega_i^n) \Rightarrow 0. \tag{B29}$$

Lemma 7, (B25), and (B29) imply that as $n\to\infty$,

$$\sup_{0\le t\le T} \frac{1}{\sqrt{n}} \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} \mathbf{1}_{\{\gamma_i^n \le \omega_i^n\}} \Rightarrow 0. \tag{B30}$$

By condition (7), as $n\to\infty$,

$$\sup_{0\le t\le T} |\tilde{Q}^n(t+\omega^n(t)) - \tilde{Q}^n((t+\omega^n(t))-)| \Rightarrow 0. \tag{B31}$$

Applying the above convergence (B26)–(B31) to the inequality (B23) yields that as $n\to\infty$,

$$\sup_{0\le t\le T} |\tilde{Q}^n(t+\omega^n(t)) - \mu\tilde{\omega}^n(t)| \Rightarrow 0.$$

By condition (7) and (B22), as $n\to\infty$,

$$\sup_{0\le t\le T} |\tilde{Q}^n(t+\omega^n(t)) - \tilde{Q}^n(t)| \Rightarrow 0.$$

Thus the result of this proposition follows. □

References

[1] Atar R (2012) A diffusion regime with non-degenerate slowdown. *Oper. Res.* 60(2):490–500.

[2] Atar R, Gurvich I (2014) Scheduling parallel servers in the nondegenerate slowdown diffusion regime: Asymptotic optimality results. *Ann. Appl. Probab.* 24(2):760–810.

[3] Billingsley P (1999) *Convergence of Probability Measures*, Wiley Series in Probability and Statistics, 2nd ed. (John Wiley & Sons, New York).

[4] Chow YS, Teicher H (2003) *Probability Theory: Independence, Interchangeability, Martingales* (Springer, New York).

[5] Dai JG (1995) On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* 5(1):49–77.

[6] Dai JG, Dai W (1999) A heavy traffic limit theorem for a class of open queueing networks with finite buffers. *Queueing Syst.* 32(1–3):5–40.

[7] Dai JG, He S (2010) Customer abandonment in many-server queues. *Math. Oper. Res.* 35(2):347–362.

[8] Dai JG, He S (2013) Many-server queues with customer abandonment: Numerical analysis of their diffusion model. *Stochastic Systems* 3(1):96–146.

[9] Dai JG, He S, Tezcan T (2010) Many-server diffusion limits for $G/Ph/n+GI$ queues. *Ann. Appl. Probab.* 20(5):1854–1890.

[10] Dorsman J, Vlasiou M, Zwart B (2015) Heavy-traffic asymptotics for networks of parallel queues with Markov-modulated service speeds. *Queueing Syst.* 79(3):293–319.

[11] Dudley RM (2002) *Real Analysis and Probability*, Cambridge Studies in Advanced Mathematics (Cambridge University Press, Cambridge, UK).

[12] Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.

[13] Katsuda T (2015) General hazard-type scaling of abandonment time distribution for a $G/Ph/n+GI$ queue in the Halfin-Whitt heavy-traffic regime. *Queueing Syst.* 80(1–2):155–195.

[14] Krichagina EV, Puhalskii AA (1997) A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Syst.* 25(1–4):235–280.

[15] Kruk L, Lehoczky J, Ramanan K, Shreve S (2007) An explicit formula for the Skorokhod map on $[0,a]$. *Ann. Probab.* 35(5):1740–1768.

[16] Lee C, Weerasinghe A (2011) Convergence of a queueing system in heavy traffic with general patience-time distributions. *Stochastic Processes and Their Appl.* 121(11):2507–2552.

[17] Mahabhashyam SR, Gautam N (2005) On queues with Markov modulated service rates. *Queueing Syst.* 51(1–2):89–113.

[18] Mandelbaum A, Momčilović P (2012) Queues with many servers and impatient customers. *Math. Oper. Res.* 37(1):41–65.

[19] Palm C (1937) Etude des delais d'attente. *Ericson Technics* 5:37–56.

[20] Reed JE (2007) The $G/GI/N$ queue in the Halfin-Whitt regime II: Idle time system equation. Technical report, New York University, New York).

[21] Reed JE (2009) The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Probab.* 19(6):2211–2269.

[22] Reed JE, Shaki Y (2015) A fair policy for the $G/GI/N$ queue with multiple server pools. *Math. Oper. Res.* 40(3):558–595.

[23] Reed JE, Tezcan T (2012) Hazard rate scaling of the abandonment distribution for the $GI/M/n+GI$ queue in heavy traffic. *Oper. Res.* 60(4):981–995.

[24] Reed JE, Ward AR (2008) Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Math. Oper. Res.* 33(3):606–644.

[25] Shorack GR, Wellner JA (2009) *Empirical Processes with Applications to Statistics*, Wiley Series in Probability and Mathematical Statistics (John Wiley & Sons, New York).

[26] Takine T (2005) Single-server queues with Markov-modulated arrivals and service speed. *Queueing Syst.* 49(1):7–22.

[27] Talreja R, Whitt W (2009) Heavy-traffic limits for waiting times in many-server queues with abandonment. *Ann. Appl. Probab.* 19(6):2137–2175.

[28] Ward AR (2012) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys Oper. Res. Management Sci.* 17(1):1–14.

[29] Ward AR, Glynn PW (2003) A diffusion approximation for a Markovian queue with reneging. *Queueing Syst.* 43(1–2):103–128.

[30] Ward AR, Glynn PW (2005) A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. *Queueing Syst.* 50(4):371–400.

[31] Weerasinghe A (2014) Diffusion approximations for $G/M/n$+GI queues with state-dependent service rates. *Math. Oper. Res.* 39(1):207–228.

[32] Whitt W (1980) Some useful functions for functional limit theorems. *Math. Oper. Res.* 5(1):67–85.

[33] Whitt W (2007) Proofs of the martingale FCLT. *Probab. Surveys* 4:268–302.

[34] Yin GG, Zhang Q (2013) *Continuous-Time Markov Chains and Applications*, Stochastic Modelling and Applied Probability, 2nd ed., Vol. 37 (Springer, New York).

[35] Zeltyn S, Mandelbaum A (2005) Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Syst.* 51(3–4):361–402.