Optimal scheduling of critically loaded multiclass $GI/M/n+M$ queues in an alternating renewal environment

ARI ARAPOSTATHIS[†], GUODONG PANG[‡], AND YI ZHENG[‡]

Abstract. In this paper, we study optimal control problems for multiclass $GI/M/n+M$ queues in an alternating renewal (up-down) random environment in the Halfin–Whitt regime. Assuming that the downtimes are asymptotically negligible and only the service processes are affected, we show that the limits of the diffusion-scaled state processes under non-anticipative, preemptive, work-conserving scheduling policies, are controlled jump diffusions driven by a compound Poisson jump process. We establish the asymptotic optimality of the infinite-horizon discounted and long-run average (ergodic) problems for the queueing dynamics.

Since the process counting the number of customers in each class is not Markov, the usual martingale arguments for convergence of mean empirical measures cannot be applied. We surmount this obstacle by demonstrating the convergence of the generators of an augmented Markovian model which incorporates the age processes of the renewal interarrival times and downtimes. We also establish long-run average moment bounds of the diffusion-scaled queueing processes under some (modified) priority scheduling policies. This is accomplished via Foster–Lyapunov equations for the augmented Markovian model.

1. Introduction

There has been a lot of research activity on scheduling control problems for queueing networks in the Halfin–Whitt regime. The discounted problem for multiclass many-server queues was first studied in [1]. See also the work in [2,3]. For the ergodic control problem in the case of Markovian queueing networks see [4–6]. Scheduling control problems for queueing networks in random environments have also attracted much attention recently [7–10]. It is worth noting that in the study of asymptotic optimality in Markov-modulated environments, the scaling parameter depends on the rate of the underlying Markov process; see, for example, [7,10,11].

In this paper we consider queueing networks operating in alternating renewal (up-down) random environments, modeling service interruptions, and with renewal arrivals. It is well known that for large-scale service systems, service interruptions can have a dramatic impact on system performance [12]. For single class queues and networks in an alternating renewal environment, limit theorems have been studied in [12–16]. To the best of our knowledge, there are no studies on optimal scheduling control for multiclass many-server queues in alternating renewal environments, or even ergodic control in the Halfin–Whitt regime with arrivals that are renewal processes.

Specifically, we consider multiclass (d classes) $GI/M/n+M$ queues with service interruptions in the Halfin–Whitt regime, where the arrival rate in each class and the number of servers in the pool are large, with a scaling parameter n, and the service interruptions are asymptotically negligible of order $n^{-1/2}$. The service interruption is modeled as an alternating renewal process constructed

[†] Department of Electrical and Computer Engineering, The University of Texas at Austin, 2501 Speedway, EERC 7.824, Austin, TX 78712

[‡] The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, College of Engineering, Pennsylvania State University, University Park, PA 16802

E-mail addresses: ari@utexas.edu, {gup3,yxz282}@psu.edu.

2000 *Mathematics Subject Classification.* Primary: 90B22, 90B36, 60K37. Secondary: 60K25, 60J75, 60F17.

Key words and phrases. multiclass many-server queues, Halfin–Whitt (QED) regime, service interruptions, renewal arrivals, alternating renewal process, jump diffusions, discounted cost, ergodic control, asymptotic optimality.

by regenerative 'up' and 'down' cycles. In the 'down' state, all servers stop functioning, and new customers arrive, which may abandon the queue. In the 'up' state, the queueing system functions normally. We assume that at least one class of customers has a strictly positive abandonment rate. The scheduling policy determines the allocation of servers to different classes of customers. We approximate the scheduling problem via the corresponding control problem of the limiting jump diffusion in the heavy-traffic regime, for which a sharp characterization of optimal Markov controls is available [17], and use this to exhibit matching upper and lower bounds on the optimal scheduling performance for the queueing dynamics.

In Proposition 3.1, we establish a functional central limit theorem (FCLT) for the d-dimensional diffusion-scaled state processes under work-conserving scheduling policies. The limiting controlled processes are jump diffusions with piecewise linear drift and compound Poisson jumps. The proof of weak convergence relies on the construction of a modified diffusion-scaled state process, where we add the cumulative downtime to a diffusion-scaled state process without interruptions. We show that the modified and original diffusion-scaled state processes have the same weak limits, which are governed by the jump diffusions described above.

The discounted and ergodic control problems for a large class of jump diffusions arising from queueing networks in the Halfin–Whitt regime have been studied in [17], and these results are essential for establishing asymptotic optimality in the present paper. In Theorem 3.1, we show that the optimal value functions of the discounted problem for the diffusion-scaled processes converge to the corresponding function for the limiting jump diffusion. The proof of asymptotic optimality for the discounted problem follows the approach in [1], which deals with the discounted problem for multiclass $GI/M/n + M$ queues. An essential part of this proof involves moment bounds for the diffusion-scaled state process, and the cumulative downtime process.

Asymptotic optimality for the ergodic control problem is more challenging. The result is stated in Theorem 3.2. Here, long-run average moment bounds for the diffusion-scaled state processes play a crucial role (see Proposition 4.2). Typically, such bounds are obtained in the literature via Foster–Lyapunov inequalities [4–6, 10, 18]. However, since the process counting the number of customers in each class, referred to as the queueing process, or state process, is not Markov, we first construct a sequence of auxiliary diffusion-scaled processes by adding the scaled residual time process of the alternating renewal process in the 'down' state to the original process, taking advantage of the fact that the long-run average moments of the scaled residual time process are negligible as the scaling parameter n tends to infinity (see equation (4.25)). We then consider the joint Markov process comprised of the auxiliary diffusion-scaled state process and the age processes of renewal arrival and alternating renewal processes, and construct Foster–Lyapunov functions, which bear a resemblance to the Lyapunov functions in [19]. In this part, we assume that the mean residual life functions are bounded, and use the criterion in [20, Theorem 4.2] to show that the joint Markov processes are positive Harris recurrent for all large enough n under some (modified) priority scheduling policy. We apply a two-step scheduling: first, the servers are allocated to the classes of customers with zero abandonment rate in such a manner that the servers used for each class do not exceed a certain proportion dictated by the traffic intensity; second, a static priority rule is applied to allocate the remaining servers. We show that the long-run average moments of the auxiliary diffusion-scaled state processes are bounded under this scheduling policy. We then establish a moment estimate for the difference between the auxiliary and original diffusion-scaled processes, and proceed to show that the analogous moment bounds hold for the original diffusion-scaled processes.

To prove asymptotic optimality for the ergodic control problem, we establish lower and upper bounds for the limits of the value functions (see equations (5.10) and (5.28)). For the proof of the lower bound, we show that the sequence of mean empirical measures of the diffusion-scaled state processes is tight (see Lemma 5.2), and any limit of mean empirical measures is an ergodic occupation measure for the limiting jump diffusion. This is analogous to the technique used in

[4–6, 10]. However, characterizing the limits of mean empirical measures (see Theorem 5.2) is quite challenging here. Since we consider the diffusion-scaled processes with renewal arrivals in an alternating renewal environment, the martingale arguments in the above papers cannot be applied here. Instead, we develop a new approach. Following the technique of the proof of ergodicity under the specific scheduling policy described in the preceding paragraph, we consider the generator of the joint Markov process of the auxiliary diffusion-scaled state process, which incorporates the residual time process, and the associated age processes of the renewal arrivals and the alternating renewal environment. We construct suitable test functions (see (5.12)) which involve the coefficients of variation of interarrival times, and proceed to show the convergence of generators.

For the proof of the upper bound, we adopt the spatial truncation technique developed in [4], which is also used in [5,6,10], and is extended to jump diffusions in [17]. This involves a concatenated scheduling policy. We first construct a continuous precise ϵ-optimal control for the ergodic control problem for the limiting jump diffusion (see Proposition 5.1). Then, inside a compact set, we map this control to a scheduling policy for the diffusion-scaled process. On the complement of this set, we apply the (modified) priority scheduling policy. We show that the long run average moments of the diffusion-scaled state process are bounded under this concatenated scheduling policy (see Proposition 4.3), and the limit of mean empirical measures is the ergodic occupation measure of the limiting jump diffusion governed by the ϵ-optimal control (see Lemma 5.3). Here, the techniques used in establishing the long-run average moment bounds under the (modified) priority scheduling policy, and the convergence of mean empirical measures, play an important role.

1.1. **Organization of the paper.** The notation used in the paper is summarized in the next subsection. In Section 2, we describe the model of multiclass many-server queues with service interruptions. In Section 3, we define the diffusion-scaled processes and associated control problems, and state the main results on weak convergence and asymptotic optimality. In Section 4, we summarize the ergodic properties of the limiting controlled jump diffusion, and state the results concerning long-run average moment bounds for the diffusion-scaled processes. The proofs of Theorems 3.1 and 3.2 are given in Section 5. Appendix A is devoted to the proofs of Lemma 3.1 and Proposition 3.1. Appendix B contains the proofs of Lemmas 4.1 and 5.2.

1.2. **Notation.** We let $|\cdot|$ and $\langle\cdot,\cdot\rangle$ denote the standard Euclidean norm and the inner product in $\mathbb{R}^d$, respectively. For $x\in\mathbb{R}^d$, we let $\|x\| := \sum_i |x_i|$, and x' denote the transpose of x. The symbols $\mathbb{R}_+$, $\mathbb{Z}_+$, $\mathbb{N}$, denote the set of nonnegative real numbers, nonnegative integers, and the set of natural numbers, respectively. The indicator function of a set $A\in\mathbb{R}^d$ is denoted by $\mathbb{1}_A$. Given $a,b\in\mathbb{R}$, the minimum (maximum) is denoted by $a\wedge b$ ($a\vee b$), respectively, $\lfloor a\rfloor$ denotes the integer part of a, and $a^{\pm} := (\pm a)\vee 0$. The complement and closure of a set $A\subset\mathbb{R}^d$ are denoted by A^c and $\bar{A}$, respectively. We use the notation e_i to denote the vector with i-th entry equal to 1 and all other entries equal to 0. We also let $e := (1,\dots,1)^{\mathsf{T}}$. We let B_r denote the open ball of radius r in $\mathbb{R}^d$, centered at the origin. For a process $\{X_t\}_{t\ge 0}$, $\uptau(A)$ denotes the first exit time from the set $A\subset\mathbb{R}^d$, defined by $\uptau(A) := \inf\{t>0 : X_t\notin A\}$, and we let $\uptau_r := \uptau(B_r)$.

For a domain $D\subset\mathbb{R}^d$, the space $\mathcal{C}^k(D)$ ($\mathcal{C}^\infty(D)$), $k\ge 0$, stands for the class of all real-valued functions on D whose partial derivatives up to order k (of any order) exist and are continuous. $\mathcal{C}^{k,r}(D)$ stands for the set of functions that are k-times continuously differentiable and whose k^{th} derivatives are locally Hölder continuous with exponent r. We let $\mathcal{C}^k_c(D)$ denote the space of functions in $\mathcal{C}^k(D)$ with compact support, and $\mathcal{C}^k_b$ the set of functions in $\mathcal{C}^k(D)$ whose partial derivatives up to order k are bounded. For a nonnegative function $g\in\mathcal{C}(\mathbb{R}^d)$, $\mathcal{O}(g)$ denotes the space of functions $f\in\mathcal{C}(\mathbb{R}^d)$ satisfying $\sup_{x\in\mathbb{R}^d}\frac{|f(x)|}{1+g(x)}<\infty$. By a slight abuse of notation, $\mathcal{O}(g)$ also denotes a generic member of these spaces.

For $k \in \mathbb{N}$, we let $\mathbb{D}^k := \mathbb{D}(\mathbb{R}_+, \mathbb{R}^k)$ denote the space of $\mathbb{R}^k$-valued cádlág functions on $\mathbb{R}_+$. When $k = 1$, we write $\mathbb{D}$ for $\mathbb{D}^k$. Given a Polish space E, by $\mathcal{P}(E)$ we denote the space of probability measures on E, endowed with the Prokhorov metric.

2. Multiclass $GI/M/N + M$ queues with service interruptions

2.1. The model and assumptions.

We consider a sequence of $GI/M/n + M$ queueing models with d classes of customers. Let $\mathcal{I} := \{1, \dots, d\}$. For the n^{th} system, let $\{A^n_i(t)\}_{t\ge0}$ denote the arrival process of class-i customers. We assume that the arrivals are mutually independent renewal processes defined as follows. Let $\{G_{i,j} \colon j \in \mathbb{N}\}$, $i \in \mathcal{I}$, be an i.i.d. sequence of strictly positive random variables with mean $\mathbb{E}[G_i] = 1$ and finite (squared) coefficient of variation $c^2_{a,i} := {}^{\mathrm{Var}(G_i)}/_{(\mathbb{E}[G_i])^2}$, where $G_i \equiv G_{i,1}$. Then, we define

$$A^n_i(t) := \max\left\{m \ge 0\colon \sum_{j=1}^{m} G_{i,j} \le \lambda^n_i t\right\}, \quad t \ge 0,\ i \in \mathcal{I}, \tag{2.1}$$

where $\lambda^n_i > 0$ denotes the arrival rate. For each $n \in \mathbb{N}$, the service and patience times of the class-i customers are exponentially distributed with parameters μ^n_i and γ^n_i, respectively.

We adopt the following standard assumption on the parameters (see [1,4,13]).

Assumption 2.1. (*The Halfin–Whitt regime*) The parameters satisfy the following limits for each $i \in \mathcal{I}$ as $n \to \infty$:

$$\begin{aligned}
&n^{-1}\lambda^n_i \to \lambda_i > 0, \quad \mu^n_i \to \mu_i > 0, \quad \gamma^n_i \to \gamma_i \ge 0,\\
&n^{-1/2}(\lambda^n_i - n\lambda_i) \to \hat\lambda_i, \quad n^{1/2}(\mu^n_i - \mu_i) \to \hat\mu_i,\\
&\frac{\lambda^n_i}{n\mu^n_i} \to \rho_i := \frac{\lambda_i}{\mu_i} < 1, \quad \sum_{i=1}^{d}\rho_i = 1.
\end{aligned}$$

We assume that $\inf_{n\in\mathbb{N}} \gamma^n_d > 0$. Assumption 2.1, which is also known as the Quality-and-Efficiency-Driven regime, implies that the system is critically loaded and

$$\rho^n \to \hat\rho := \sum_{i=1}^{d} \frac{\rho_i\hat\mu_i - \hat\lambda_i}{\mu_i} \in \mathbb{R}, \quad \text{where} \quad \rho^n := \sqrt{n}\left(1 - \sum_{i=1}^{d}\frac{\lambda^n_i}{n\mu^n_i}\right).$$

All queues are in the same up-down alternating renewal random environment. Waiting customers may abandon at any time. In the 'up' state, the system functions normally, and in the 'down' state all servers stop, while customers keep joining the queues and any jobs that have started service will wait for the system to resume. For this reason, we also refer to this model as multiclass queues with service interruptions. Let $\bigl\{(u^n_k, d^n_k)\colon k \in \mathbb{N}\bigr\}$ be a sequence of i.i.d. positive random vectors denoting the up-down cycles, and define the *counting process of downtimes* by

$$N^n(t) := \max\bigl\{k \ge 0\colon T^n_k \le t\bigr\}, \quad \text{with } T^n_k := \sum_{i=1}^{k}(u^n_i + d^n_i), \quad k \in \mathbb{N}, \tag{2.2}$$

and $T^n_0 \equiv 0$. At time 0, the system is in the 'up' state.

Assumption 2.2. For each n and k in $\mathbb{N}$, u^n_k and d^n_k are independent, u^n_k is exponentially distributed with parameter β^n_u, which converges to $\beta > 0$ as $n \to \infty$. We assume that $d^n_1 = \frac{1}{\vartheta^n} d_1$, with d_1 some nonnegative random variable satisfying $\mathbb{E}[d_1] = 1$, and $\frac{\vartheta^n}{\sqrt{n}} \to \vartheta > 0$ as $n \to \infty$.

For $k \in \mathbb{N}$, we let $(\mathbb{D}^k, M_1)$ and $(\mathbb{D}^k, J_1)$ denote the space $\mathbb{D}^k$ endowed with the Skorokhod M_1 and J_1 topologies, respectively (see, for example, [21,22]). Assumption 2.2 implies that the service interruptions are asymptotically negligible, and

$$N^n \Rightarrow N \quad \text{in} \quad (\mathbb{D}, J_1) \quad \text{as } n \to \infty,$$

where the limiting process N is a Poisson process with rate β. Define the *server availability process* $\Psi^n := \{\Psi^n(t)\colon t \ge 0\}$ by

$$\Psi^n(t) \;=\; \begin{cases} 1, & T_k^n \le t < T_k^n + u_{k+1}^n\,, \\ 0, & T_k^n + u_{k+1}^n \le t < T_{k+1}^n\,, \end{cases} \tag{2.3}$$

for $k \in \mathbb{N}$. We also define the cumulative up-time process $C_{\mathsf{u}}^n = \{C_{\mathsf{u}}^n(t)\}_{t\ge 0}$ by $C_{\mathsf{u}}^n(t) := \int_0^t \Psi^n(s)\,\mathrm{d}s$, and the cumulative down-time process by $C_{\mathsf{d}}^n(t) := t - C_{\mathsf{u}}^n(t)$. Let F^{d_1} denote the distribution function of d_1. By Lemma 2.2 in [13], we have

$$\sqrt{n}C_{\mathsf{d}}^n \;\Rightarrow\; L \quad \text{in } (\mathbb{D}, M_1) \quad \text{as } n \to \infty\,, \tag{2.4}$$

where $\{L_t\}_{t\ge0}$ is a compound Poisson process with intensity $\Pi_L(\mathrm{d}x)\mathrm{d}t = \beta\, F^{d_1}(\vartheta \mathrm{d}x)\mathrm{d}t$, where β is given in Assumption 2.2.

For the n^{th} system, we denote the processes counting the total number of customers, those in queue, and those in service, by $X^n = (X_1^n, \dots, X_d^n)'$, $Q^n = (Q_1^n, \dots, Q_d^n)'$, and $Z^n = (Z_1^n, \dots, Z_d^n)'$, respectively. These processes satisfy the following constraints:

$$X_i^n(t) \;=\; Q_i^n(t) + Z_i^n(t)\,, \quad Q_i^n(t) \;\ge\; 0\,, \quad Z_i^n(t) \;\ge\; 0\,, \quad \text{and} \quad \langle e, Z^n(t)\rangle \;\le\; n \tag{2.5}$$

for each $t \ge 0$ and $i \in \mathcal{I}$. We let

$$\begin{aligned} S_i^n(t,r) \;&:=\; S_{*,i}^n\biggl(\mu_i^n \int_0^t Z_i^n(s)\Psi^n(s)\,\mathrm{d}s + \mu_i^n r\biggr)\,, \\ R_i^n(t,r) \;&:=\; R_{*,i}^n\biggl(\gamma_i^n \int_0^t Q_i^n(s)\,\mathrm{d}s + \gamma_i^n r\biggr)\,, \end{aligned} \tag{2.6}$$

for $i \in \mathcal{I}$, $t \ge 0$, and $r \ge 0$, where $\{S_{*,i}^n, R_{*,i}^n\colon i \in \mathcal{I}, n \in \mathbb{N}\}$ are Poisson processes with rate one. We assume that for each $n \in \mathbb{N}$, $\bigl\{X_i^n(0), A_i^n, S_{*,i}^n, R_{*,i}^n\colon i \in \mathcal{I}\bigr\}$ are mutually independent. These processes are governed by the equation

$$X_i^n(t) \;=\; X_i^n(0) + A_i^n(t) - S_i^n(t) - R_i^n(t) \tag{2.7}$$

for each $t \ge 0$, $n \in \mathbb{N}$, and $i \in \mathcal{I}$, where $S_i^n(t) := S_i^n(t,0)$ and $R_i^n(t) := R_i^n(t,0)$.

2.2. **Scheduling policies.** A scheduling policy is identified with a $\mathbb{Z}_+^d$-valued stochastic process Z^n with cádlág sample paths, which satisfies (2.5). Let

$$\tilde{\tau}_i^n(t) \;:=\; \inf\{r \ge t\colon A_i^n(r) - A_i^n(r-) > 0\}\,, \quad \text{and} \quad \breve{\tau}^n(t) \;:=\; \inf\{r \ge t\colon \Psi^n(r) = 1\}\,, \tag{2.8}$$

for $i \in \mathcal{I}$. Recall the definitions of C_{d}^n in (2.4), and S^n and R^n in (2.6). Define the σ-fields

$$\begin{aligned} \mathcal{F}_t^n \;&:=\; \sigma\bigl\{X^n(0), A_i^n(t), S_i^n(s), R_i^n(s), X_i^n(s), Z_i^n(s), \Psi^n(s), N^n(s)\colon i \in \mathcal{I}, 0 \le s \le t\bigr\} \vee \mathcal{N}\,, \\ \mathcal{G}_t^n \;&:=\; \sigma\bigl\{A_i^n(\tilde{\tau}_i^n(t) + r) - A_i^n\bigl(\tilde{\tau}_i^n(t)\bigr), S_i^n(\breve{\tau}^n(t), r) - S_i^n\bigl(\breve{\tau}^n(t)\bigr)\,, \\ &\qquad R_i^n(\breve{\tau}^n(t), r) - R_i^n\bigl(\breve{\tau}^n(t)\bigr), C_{\mathsf{d}}^n(\breve{\tau}^n(t) + r) - C_{\mathsf{d}}^n\bigl(\breve{\tau}^n(t)\bigr)\colon i \in \mathcal{I}, r \ge 0\bigr\} \vee \mathcal{N}\,, \end{aligned} \tag{2.9}$$

for $t \ge 0$, where $\mathcal{N}$ is the collection of all $\mathbb{P}$-null sets. We say that a scheduling policy Z^n is *non-anticipative* if

(i) $Z^n(t)$ is adapted to $\mathcal{F}_t^n$,
(ii) $\mathcal{F}_t^n$ and $\mathcal{G}_t^n$ are independent at each time $t \ge 0$,
(iii) for each $i \in \mathcal{I}$, and $t \ge 0$, the process $S_i^n(\breve{\tau}^n(t), \cdot) - S_i^n(\breve{\tau}^n(t))$ agrees in law with $S_{*,i}^n(\mu_i^n \cdot)$, and the process $R_i^n(\breve{\tau}^n(t), \cdot) - R_i^n(\breve{\tau}^n(t))$ agrees in law with $R_{*,i}^n(\gamma_i^n \cdot)$.

The information at time t is contained in $\mathcal{F}_t^n$, while $\mathcal{G}_t^n$ represents the information about future increments. The renewal arrivals A_i^n, $i \in \mathcal{I}$, and the alternative renewal process Ψ^n are regenerative processes. So in $\mathcal{G}_t^n$, we use $\tilde{\tau}_i^n(t)$ and $\breve{\tau}^n(t)$, respectively, instead of t. Note that parts (ii) and (iii) in the definition of non-anticipative scheduling policy are required so that the any limit of scheduling policies corresponds to a non-anticipative control for the limiting controlled jump diffusion. See part (iii) of Proposition 3.1 for details.

Let $\tau^n_{i,k}$ denote the k^{th} jump time of $A^n_i - S^n_i - R^n_i$, for each $n \in \mathbb{N}$ and $i \in \mathcal{I}$. Equation (2.7) implies that $X^n_i(t) = X^n_i(0)$ for $0 \le t \le \tau^n_{i,1}$, $X^n_i(t) = X^n_i(0) + \epsilon_1$ for $\tau^n_{i,1} \le t \le \tau^n_{i,2}$ and so forth, where ϵ_k denotes the jump size which takes values in a bounded set. Note that the integrals in (2.6) are finite by the definition of Ψ^n in (2.3) and (2.5). Thus, given any non-anticipative scheduling policy Z^n, and initial condition $X^n(0)$, there exists a unique solution to (2.7).

For $x \in \mathbb{Z}^d_+$, we define the action set $\mathcal{Z}^n(x)$ by

$$\mathcal{Z}^n(x) \,:=\, \bigl\{z \in \mathbb{Z}^d_+ : z \le x\,,\ \langle e, z\rangle = \langle e, x\rangle \wedge n\bigr\}\,.$$

A scheduling policy Z^n is called *admissible* if $Z^n(t)$ takes values in $\mathcal{Z}^n\bigl(X^n(t)\bigr)$ at each t, and is non-anticipative. The set of admissible scheduling policies is denoted by $\mathfrak{Z}^n$. Note that an admissible policy allows preemption, that is, a server can interrupt service of a customer at any time to serve some other class of customers. In summary, given an admissible scheduling policy Z^n, the process X^n in (2.7) is well defined, and we say that X^n is governed by Z^n.

Next, we describe a well-known equivalent parameterization of the set of admissible policies. Let

$$\mathcal{S} \,:=\, \{u \in \mathbb{R}^d{}_+ : \langle e, u\rangle = 1\}\,.$$

We also define

$$\mathcal{S}^n(x) \,:=\, \Bigl\{v \in \mathbb{Z}^d_+ : v = \frac{y}{\langle e, x\rangle - n} \in \mathcal{S}\,,\ y \le x\,,\ y \in \mathbb{Z}^d_+\Bigr\}\,, \quad \text{if}\ \ \langle e, x\rangle > n\,,$$

and $\mathcal{S}^n(x) = \{e_d\}$, if $\langle e, x\rangle \le n$. Let $\mathfrak{U}^n$ denote the class of processes $\{U^n(t)\}_{t\ge 0}$ which are non-anticipative, in the sense of the definition given above, and $U^n(t)$ takes values in $\mathcal{S}^n\bigl(X^n(t)\bigr)$. Then, each $U^n \in \mathfrak{U}^n$ determines a policy $Z^n \in \mathfrak{Z}^n$ via

$$Z^n(t) \,=\, X^n(t) - Q^n(t)\,, \quad \text{with}\ \ Q^n(t) \,=\, \bigl(\langle e, X^n(t)\rangle - n\bigr)^+ U^n(t)\,.$$

This map is invertible, and its inverse is given by

$$U^n(t) \,:=\, \begin{cases} \frac{X^n(t) - Z^n(t)}{\langle e, X^n(t)\rangle - n} & \text{for}\ \langle e, X^n(t)\rangle > n\,,\\ e_d & \text{for}\ \langle e, X^n(t)\rangle \le n\,. \end{cases}$$

Therefore, as far as control problems are concerned, we can use policies in $\mathfrak{U}^n$ or $\mathfrak{Z}^n$ interchangeably. Note that U^n_i can be considered as the proportion of class-i customers in the queue when there are waiting customers in the system.

Next, we augment the state space, and define the class of stationary Markov scheduling policies. Recall the definitions of A^n, N^n, and Ψ^n in (2.1)–(2.3), respectively.

Definition 2.1. Let $H^n_i(t)$ denote the age process for the class-i customers, that is,

$$H^n_i(t) \,:=\, t - \frac{1}{\lambda^n_i}\sum_{j=1}^{A^n_i(t)} G_{i,j}\,, \qquad t \ge 0\,, \qquad i \in \mathcal{I}\,, \tag{2.10}$$

and define the age process K^n for the alternating renewal process in the 'down' state by

$$K^n(t) \,:=\, \left(t - \sum_{k=1}^{N^n(t)} (u^n_k + d^n_k) - u^n_{N^n(t)+1}\right)^+, \qquad t \ge 0\,. \tag{2.11}$$

Then, (A^n_i, H^n_i), $i \in \mathcal{I}$, and (Ψ^n, K^n) are strong Markov processes (see, e.g., [23]). We say that a scheduling policy $Z^n \in \mathfrak{Z}^n$ is (stationary) Markov if

$$Z^n(t) \,=\, z^n\bigl(X^n(t), H^n(t), \Psi^n(t), K^n(t)\bigr)$$

for some $z^n\colon \mathbb{Z}^d_+\times\mathbb{R}^d_+\times\{0,1\}\times\mathbb{R}_+\to\mathbb{Z}^d_+$, and we let $\mathfrak{Z}^n_{\mathrm{sm}}$ denote the class of these policies. Under a policy $Z^n\in\mathfrak{Z}^n_{\mathrm{sm}}$, the process (X^n,H^n,Ψ^n,K^n) is Markov with state space

$$\bigl\{(x,h,\psi,k)\in\mathbb{Z}^d_+\times\mathbb{R}^d_+\times\{0,1\}\times\mathbb{R}_+\colon k\equiv 0 \text{ if } \psi=1\bigr\}\,.$$

Abusing the notation, when z^n depends only on its first argument, we simply write $Z^n(t)=z^n\bigl(X^n(t)\bigr)$.

3. Diffusion-scaled processes and control problems

Let $\hat X^n$, $\hat Q^n$, and $\hat Z^n$ denote the diffusion-scaled processes defined by

$$\hat X^n_i(t)\,:=\,n^{-1/2}(X^n_i(t)-\rho_i n)\,,\quad \hat Q^n_i(t)\,:=\,n^{-1/2}Q^n_i(t)\,,\quad \hat Z^n_i(t)\,:=\,n^{-1/2}(Z^n_i(t)-\rho_i n)\,,$$

respectively, for $t\ge 0$ and $i\in\mathcal{I}$. It follows by (2.7) that the process $\hat X^n_i$ takes the form

$$\begin{aligned}\hat X^n_i(t) \;=\; &\hat X^n_i(0)+\ell^n_i t+\hat A^n_i(t)-\hat S^n_i(t)-\hat R^n_i(t)\\ &-\mu^n_i\int_0^t \hat Z^n_i(s)\Psi^n(s)\,\mathrm{d}s-\gamma^n_i\int_0^t\hat Q^n_i(s)\,\mathrm{d}s+\hat L^n_i(t)\,,\quad t\ge 0\,,\end{aligned}\tag{3.1}$$

where $\ell^n_i:=n^{-1/2}(\lambda^n_i-n\mu^n_i\rho_i)$,

$$\hat A^n_i(t)\,:=\,n^{-1/2}\bigl(A^n_i(t)-\lambda^n_i t\bigr)\,,\qquad \hat S^n_i(t)\,:=\,n^{-1/2}\biggl(S^n_i(t)-\mu^n_i\int_0^t Z^n_i(s)\Psi^n(s)\,\mathrm{d}s\biggr)\,,$$

$$\hat R^n_i(t)\,:=\,n^{-1/2}\biggl(R^n_i(t)-\gamma^n_i\int_0^t Q^n_i(s)\,\mathrm{d}s\biggr)\,,\quad\text{and}\quad \hat L^n_i(t)\,:=\,\sqrt{n}\mu^n_i\rho_i C^n_{\mathsf{d}}(t)\,.$$

Let $\hat W^n$ and $\hat Y^n$, $n\in\mathbb{N}$, be d-dimensional processes defined by

$$\hat W^n_i\,:=\,\hat A^n_i-\hat S^n_i-\hat R^n_i\qquad\text{for } i\in\mathcal{I}\,,\tag{3.2}$$

and

$$\hat Y^n_i(t)\,:=\,\ell^n_i t-\mu^n_i\int_0^t\hat Z^n_i(s)\Psi^n(s)\,\mathrm{d}s-\gamma^n_i\int_0^t\hat Q^n_i(s)\,\mathrm{d}s\qquad\text{for } i\in\mathcal{I}\,,\ t\ge 0\,,$$

respectively. Then, $\hat X^n_i$ in (3.1) has the representation

$$\hat X^n_i(t)\;=\;\hat X^n_i(0)+\hat Y^n_i(t)+\hat W^n_i(t)+\hat L^n_i(t)\,.$$

The initial condition $\hat X^n(0)$, $n\in\mathbb{N}$, is assumed to be deterministic throughout the paper.

3.1. The limiting controlled diffusion with compound Poisson jumps.

In Lemma 3.1 and Proposition 3.1 which follow, products or powers of the spaces $(\mathbb{D}^d,J_1)$ and $(\mathbb{D}^d,M_1)$ are viewed as metric spaces endowed with the maximum metric. The proofs of these results are given in Appendix A.

Lemma 3.1. *Suppose that Assumptions 2.1 and 2.2 hold, and that $\{\hat X^n(0)\colon n\in\mathbb{N}\}$ is bounded. Then, under any sequence of $U^n\in\mathfrak{U}^n$, we have*

$$(n^{-1}Q^n,n^{-1}Z^n)\;\Rightarrow\;(\mathfrak{e}_0,\mathfrak{e}_\rho)\quad\textit{in}\quad(\mathbb{D}^d,M_1)^2\,,$$

where $\mathfrak{e}_0(t)\equiv(0,\dots,0)'$ for all $t\ge 0$, and $\mathfrak{e}_\rho(t)\equiv(\rho_1,\dots,\rho_d)'$.

Proposition 3.1. *Grant the assumptions in Lemma 3.1. Then, the following hold.*

(i) *As $n\to\infty$,*

$$(\hat W^n,\hat L^n)\;\Rightarrow\;(\Sigma W,\lambda L)\quad\textit{in}\quad(\mathbb{D}^d,J_1)\times(\mathbb{D}^d,M_1)\,,$$

where the matrix Σ is given by $\Sigma:=\operatorname{diag}\Bigl(\sqrt{\lambda_1(1+c^2_{a,1})},\dots,\sqrt{\lambda_d(1+c^2_{a,d})}\Bigr)$, W is a d-dimensional standard Wiener process, $\lambda:=(\lambda_1,\dots,\lambda_d)'$, and $\{L_t\}_{t\ge0}$ is the one-dimensional Lévy process in (2.4), and is independent of W.

(ii) *The sequence* $(\hat{X}^n, \hat{Y}^n, \hat{W}^n, \hat{L}^n)$ *is tight in* $(\mathbb{D}^d, M_1) \times (\mathbb{D}^d, J_1)^2 \times (\mathbb{D}^d, M_1)$.

(iii) *Provided* U^n *is tight in* $(\mathbb{D}^d, J_1)$, *any limit* X *of* $\hat{X}^n$ *is a strong solution to the stochastic differential equation*

$$\mathrm{d}X_t \;=\; b(X_t, U_t)\,\mathrm{d}t + \Sigma\,\mathrm{d}W_t + \lambda\,\mathrm{d}L_t\,, \tag{3.3}$$

with initial condition $X_0 = x \in \mathbb{R}^d$, *where* U *is a limit of* U^n, *and* $b(x,u)\colon \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}^d$ *takes the form*

$$b(x,u) \;=\; \ell - M(x - \langle e, x\rangle^+ u) - \langle e, x\rangle^+ \Gamma u\,, \tag{3.4}$$

with $\ell := (\ell_1, \dots, \ell_d)'$, $M := \mathrm{diag}(\mu_1, \dots, \mu_d)$, *and* $\Gamma := \mathrm{diag}(\gamma_1, \dots, \gamma_d)$. *Moreover, any such limit* U *is non-anticipative, that is, for* $s < t$, $(W_t - W_s, L_t - L_s)$ *is independent of*

$$\mathcal{F}_s \;:=\; \textit{the completion of } \sigma\{X_0, U_r, W_r, L_r \colon r \le s\}\,.$$

Throughout the paper, the time variable appears as a subscript in the processes governing the limiting controlled jump diffusion in order to distinguish them from the processes associated with the n^{th} system.

3.2. The control problems.

Define $\widetilde{\mathcal{R}}\colon \mathbb{R}^d_+ \to \mathbb{R}_+$ by

$$\widetilde{\mathcal{R}}(x) \;:=\; c|x|^m \tag{3.5}$$

for some $c > 0$ and $m \ge 1$. The running cost function $\mathcal{R}\colon \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}_+$ is defined by

$$\mathcal{R}(x,u) \;:=\; \widetilde{\mathcal{R}}\bigl(\langle e, x\rangle^+ u\bigr)\,.$$

Remark 3.1. We only choose a running cost function as in (3.5) to simplify the exposition. One may replace (3.5) with a function $\widetilde{\mathcal{R}}$, which is locally Lipschitz continuous, and satisfies

$$c_1|x|^m \;\le\; \widetilde{\mathcal{R}}(x) \;\le\; c_2|x|^m \qquad \forall\, x \in \mathbb{R}^d\,, \tag{3.6}$$

for some positive constants c_1, c_2, and $m \ge 1$. All the results still hold with (3.6). Moreover, the lower bound in (3.6) is not needed for the discounted problem (see, e.g., [1]).

The α-discounted control problem for the n^{th} system is given by

$$\hat{V}^n_\alpha\bigl(\hat{X}^n(0)\bigr) \;:=\; \inf_{U^n \in \mathfrak{U}^n} \hat{J}_\alpha(\hat{X}^n(0), U^n) \qquad \alpha > 0\,,\; n \in \mathbb{N}\,,$$

where the cost criterion is defined by

$$\hat{J}_\alpha(\hat{X}^n(0), U^n) \;:=\; \mathbb{E}\biggl[\int_0^\infty \mathrm{e}^{-\alpha t}\, \mathcal{R}\bigl(\hat{X}^n(s), U^n(s)\bigr)\,\mathrm{d}s\biggr] \qquad \forall\, \alpha > 0\,.$$

For the controlled (jump) diffusion X in (3.3), we say that a control U is admissible if it takes values in $\mathcal{S}$, and non-anticipative (see [17]). We denote the set of all admissible controls by $\mathfrak{U}$. The corresponding α-discounted cost criterion for the diffusion takes the form

$$J_\alpha(x, U) \;:=\; \mathbb{E}^U_x\biggl[\int_0^\infty \mathrm{e}^{-\alpha t}\, \mathcal{R}(X_s, U_s)\,\mathrm{d}s\biggr] \qquad \forall\, \alpha > 0\,,$$

and the optimal α-discounted value function is given by

$$V_\alpha(x) \;:=\; \inf_{U \in \mathfrak{U}} J_\alpha(x, U) \qquad \forall\, \alpha > 0\,, \tag{3.7}$$

where $\mathbb{E}^U_x$ denotes the expectation operator corresponding to the process under the control U, with initial condition $x \in \mathbb{R}^d$. We introduce the following assumption for the discounted problem.

Assumption 3.1. There exists a constant $m_A \ge m \vee 2$ with m as in (3.5) such that $\mathbb{E}[(G_i)^{m_A}] < \infty$, for all $i \in \mathcal{I}$, and $\mathbb{E}[(d_1)^{m_A \vee (m+1)}] < \infty$.

We state the main result for the discounted problem in the next theorem, whose proof is given in Section 5.2.

Theorem 3.1. *Grant the hypotheses in Assumptions 2.1, 2.2, and 3.1, and suppose that $\hat{X}^n(0) \to x \in \mathbb{R}^d$ as $n \to \infty$. Then*

$$\lim_{n\to\infty} \hat{V}^n_\alpha\bigl(\hat{X}^n(0)\bigr) \;=\; V_\alpha(x)\,. \tag{3.8}$$

Remark 3.2. Note that in Theorem 3.1, we do not need to impose any restrictions on the limiting abandonment rates $\{\gamma_i\colon i \in \mathcal{I}\}$.

We define the ergodic control problem for the diffusion-scaled process by

$$\varrho^n\bigl(\hat{X}^n(0)\bigr) \;\coloneqq\; \inf_{Z^n\in\mathfrak{Z}^n_{\rm sm}} \hat{J}(\hat{X}^n(0), Z^n)\,,$$

where the cost criterion $\hat{J}$ is given by

$$\hat{J}(\hat{X}^n(0), Z^n) \;\coloneqq\; \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}^{Z^n}\biggl[\int_0^T \widetilde{\mathcal{R}}\bigl(\hat{Q}^n(s)\bigr)\,\mathrm{d}s\biggr].$$

Here, the infimum is over all Markov scheduling policies, since for the ergodic control problem, we work with Markov processes. For the controlled jump diffusion in (3.3), the ergodic cost criterion, and the optimal ergodic value are defined by

$$J(x,U) \;\coloneqq\; \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}^U_x\biggl[\int_0^T \mathcal{R}(X_s, U_s)\,\mathrm{d}s\biggr],$$

and

$$\varrho_*(x) \;\coloneqq\; \inf_{U\in\mathfrak{U}} J(x,U)\,, \tag{3.9}$$

respectively. By Theorem 4.1 in [17], it follows that ϱ_* is independent of x, and optimality is attained by a stationary Markov control.

We introduce the following assumption on G_i and d_1 for the ergodic control problem.

Assumption 3.2. The following hold.

(i) The right derivative of $F_i(t)$ is finite, and $F_i(t) < 1$, for all $t \ge 0$ and $i \in \mathcal{I}$. The distribution function F^{d_1} of d_1 satisfies the same property.

(ii) The mean residual life functions of G_i and d_1 are bounded, that is, there exists some positive constant $\widehat{C}$ such that

$$\frac{\int_t^\infty \bigl(1 - F^{d_1}(y)\bigr)\,\mathrm{d}y}{1 - F^{d_1}(t)} \le \widehat{C}\,, \quad \text{and} \quad \frac{\int_t^\infty \bigl(1 - F_i(y)\bigr)\,\mathrm{d}y}{1 - F_i(t)} \le \widehat{C} \quad \forall\, i \in \mathcal{I}\,, \tag{3.10}$$

and for all $t \ge 0$.

Assumption 3.2 implies that all absolute moments of G_i, $i \in \mathcal{I}$, and d_1 are finite. The main result of the ergodic control problem is stated in the next theorem, whose proof is given in Section 5.3.

Theorem 3.2. *Grant Assumptions 2.1, 2.2, and 3.2. In addition, suppose that m in (3.5) is larger than 1, and that $\hat{X}^n(0) \to x \in \mathbb{R}^d$ as $n \to \infty$. Then, we have*

$$\lim_{n\to\infty} \varrho^n\bigl(\hat{X}^n(0)\bigr) \;=\; \varrho_*\,.$$

4. Ergodic properties

In this section, we present some ergodicity results for the limiting jump diffusion and the diffusion-scaled processes. These results are used to prove Theorem 3.2 in Section 5.3.

4.1. **The limiting controlled diffusion with compound Poisson jumps.** The controlled generator of the controlled limiting jump diffusion in (3.3) is given by

$$\mathcal{A}\varphi(x,u) \;=\; \sum_{i\in\mathcal{I}} b_i(x,u)\partial_i\varphi(x) + \frac{1}{2}\sum_{i\in\mathcal{I}} \lambda_i(1+c_{a,i}^2)\partial_{ii}\varphi(x) + \int_{\mathbb{R}^d}\bigl(\varphi(x+y)-\varphi(x)\bigr)\nu_L(\mathrm{d}y) \qquad (4.1)$$

for $\varphi\in\mathcal{C}^2(\mathbb{R}^d)$, where the drift b satisfies (3.4), and $\nu_L(A) \coloneqq \Pi_L\bigl(\{z\in\mathbb{R}_*\colon \lambda z\in A\}\bigr)$ for any Borel measurable set A, with Π_L as in (2.4). We refer the reader to [20, Section 6] for the definition of exponential ergodicity. The following proposition is a direct consequence of [24, Theorem 3.5].

Proposition 4.1. *Under any constant control v such that $\Gamma v\neq 0$, the controlled limiting jump diffusion in (3.3) is exponentially ergodic.*

Remark 4.1. It is shown in [25, Theorem 5] that the limiting controlled jump diffusion is exponentially ergodic uniformly over all stationary Markov controls resulting in a locally Lipschitz continuous drift, if $\Gamma>0$.

Proposition 4.1 implies that the optimal control problems for the limiting jump diffusion are well-posed.

4.2. **Preliminaries.** We denote the scaled hazard rate function of G_i by r_i^n. This is defined by

$$r_i^n(h_i) \;\coloneqq\; \frac{\lambda_i^n\dot{F}_i(\lambda_i^n h_i)}{1-F_i(\lambda_i^n h_i)}\,,\quad \forall\, h_i\in\mathbb{R}_+\,,\quad \forall\, i\in\mathcal{I}\,,$$

where $\dot{F}_i$ denotes the right derivative of F_i. Recall H^n in (2.10). The extended generator of (A^n,H^n) associated with the renewal arrival processes, denoted by $\mathcal{H}^n$, is given by

$$\mathcal{H}^n f(x,h) \;=\; \sum_{i\in\mathcal{I}}\frac{\partial f(x,h)}{\partial h_i} + \sum_{i\in\mathcal{I}} r_i^n(h_i)\bigl(f(x+e_i,h-h_ie_i)-f(x,h)\bigr) \qquad (4.2)$$

for $f\in\mathcal{C}_b(\mathbb{R}^d\times\mathbb{R}^d_+)$.

Remark 4.2. We sketch the derivation of (4.2); see also [26, Theorem 5.5]. It is enough to consider one component (A_i^n,H_i^n), $i\in\mathcal{I}$. We obtain

$$\begin{aligned}
&\mathbb{E}_{x,h}\bigl[f\bigl(A_i^n(t+s),H_i^n(t+s)\bigr)\bigr]-f(x,h)\\
&\quad= \mathbb{E}_{x,h}\bigl[f\bigl(A_i^n(t+s),H_i^n(t+s)\bigr)\bigr]-\mathbb{E}_{x,h}\bigl[f\bigl(A_i^n(t+s),h\bigr)\bigr]+\mathbb{E}_{x,h}\bigl[f\bigl(A_i^n(t+s),h\bigr)\bigr]-f(x,h)\\
&\quad= r_{i,0,s}^n(h)\bigl(f(x,h+s)-f(x,h)\bigr)+r_{i,1,s}^n(h)\bigl(f(x+1,h)-f(x,h)\bigr)\\
&\qquad+\sum_{j\in\mathbb{N}} r_{i,j,s}^n(h)\,\mathbb{E}_{x,h}\bigl[f\bigl(x+j,H_i^n(t+s)\bigr)-f(x+j,h)\;\big|\;A_i^n(t+s)=x+j\bigr]\\
&\qquad+\sum_{j\in\mathbb{N},j\geq 2} r_{i,j,s}^n(h)\bigl(f(x+j,h)-f(x,h)\bigr)\quad \forall\, f\in\mathcal{C}_b(\mathbb{R}\times\mathbb{R})\,,\ \forall\,(x,h)\in\mathbb{R}\times\mathbb{R}_+\,,
\end{aligned}$$

where

$$r_{i,j,s}^n(h) \;\coloneqq\; \mathbb{P}\bigl(A_i^n(t+s)=x+j\,|\,A_i^n(t)=x,H_i^n(t)\;=\;h\bigr) \;=\; \mathbb{P}\bigl(A_i^n(s+h)=j\,|\,G_i\geq\lambda_i^n h\bigr)$$

by the regenerative property of renewal process. Since $\dot{F}_i(t)$ is finite for all $t\geq 0$, it follows that

$$r_i^n(h) \;\equiv\; \lim_{s\searrow 0}\frac{1}{s}\,r_{i,1,s}^n(h) \;=\; \frac{\lambda_i^n\dot{F}_i(\lambda_i^n h_i)}{1-F_i(\lambda_i^n h_i)}\,,\quad\text{and}\quad \lim_{s\searrow 0}\frac{1}{s}\,r_{i,j,s}^n(h)=0\quad\text{for } j\geq 2\,.$$

It is evident that $\lim_{s\searrow 0} r_{i,0,s}^n=1$ and $\lim_{s\searrow 0} r_{i,j,s}^n=0$ for $j\in\mathbb{N}$. Thus, we obtain (4.2).

We define (compare this with [19])

$$\eta_i^n(h_i) \coloneqq 1 - \frac{\int_{\lambda_i^n h_i}^{\infty} \bigl(1 - F_i(y)\bigr)\,\mathrm{d}y}{1 - F_i(\lambda_i^n h_i)}\,, \quad h_i \in \mathbb{R}_+\,,\ i \in \mathcal{I}\,. \tag{4.3}$$

Note that η_i^n is bounded by (3.10). The following identity is frequently used throughout the paper.

$$\dot{\eta}_i^n(h_i) - \eta_i^n(h_i) r_i^n(h_i) \;=\; \lambda_i^n - r_i^n(h_i)\,, \quad \forall\, h_i \in \mathbb{R}_+\,, \quad \forall\, i \in \mathcal{I}\,. \tag{4.4}$$

Recall that $c_{a,i}^2$ denotes the squared coefficient of variation of G_i. Let

$$\kappa_i^n(h_i) \coloneqq \frac{\int_{\lambda_i^n h_i}^{\infty} \int_t^{\infty} \bigl(1 - F_i(x)\bigr)\,\mathrm{d}x\,\mathrm{d}t}{1 - F_i(\lambda_i^n h_i)} - \frac{c_{a,i}^2 + 1}{2} \frac{\int_{\lambda_i^n h_i}^{\infty} \bigl(1 - F_i(x)\bigr)\,\mathrm{d}x}{1 - F_i(\lambda_i^n h_i)} \tag{4.5}$$

for $h_i \in \mathbb{R}_+$ and $i \in \mathcal{I}$. Note that the first term on the right-hand side of (4.5) is the second order residual life function. It follows by (3.10) that κ_i^n is bounded. Using (4.5), we obtain $\kappa_i^n(0) = 0$, and

$$\dot{\kappa}_i^n(h_i) - r_i^n(h_i)\kappa_i^n(h_i) \;=\; \biggl(\eta_i^n(h_i) + \frac{c_{a,i}^2 - 1}{2}\biggr)\lambda_i^n\,, \quad h_i \in \mathbb{R}_+\,,\ i \in \mathcal{I}\,. \tag{4.6}$$

The scaled hazard rate function of d_1 is defined by

$$\beta_{\mathsf{d}}^n(k) \coloneqq \frac{\vartheta^n \dot{F}^{d_1}(\vartheta^n k)}{1 - F^{d_1}(\vartheta^n k)}\,, \quad k \in \mathbb{R}_+\,.$$

Recall K^n in (2.11). The extended generator of (Ψ^n, K^n) associated with the alternating renewal process, denoted by $\mathcal{K}^n$, is given by

$$\mathcal{K}^n f(\psi, k) \;=\; \psi\, \beta_{\mathsf{u}}^n\bigl(f(0,0) - f(1,0)\bigr) + (1-\psi)\biggl(\beta_{\mathsf{d}}^n(k)\bigl(f(1,0) - f(0,k)\bigr) + \frac{\partial f(0,k)}{\partial k}\biggr) \tag{4.7}$$

for $f \in \mathcal{C}_b(\{0,1\} \times \mathbb{R}_+)$, with β_{u}^n as in Assumption 2.2. In analogy to (4.4), we define

$$\alpha^n(k) \coloneqq 1 - \frac{\int_{\vartheta^n k}^{\infty} \bigl(1 - F^{d_1}(x)\bigr)\,\mathrm{d}x}{1 - F^{d_1}(\vartheta^n k)} \qquad \forall\, k \in \mathbb{R}_+\,. \tag{4.8}$$

The following identities hold: $\alpha^n(0) = 0$, and

$$\dot{\alpha}^n(k) - \beta_{\mathsf{d}}^n(k)\alpha^n(k) \;=\; \vartheta^n - \beta_{\mathsf{d}}^n(k) \qquad \forall\, k \in \mathbb{R}_+\,. \tag{4.9}$$

Let $\tilde{\alpha}^n(\psi, k) \coloneqq \bigl(\psi + \alpha^n(k)\bigr)(\vartheta^n)^{-1}$. It follows by (4.9) that

$$\mathcal{K}^n \tilde{\alpha}^n(\psi, k) \;=\; -\frac{\beta_{\mathsf{u}}^n}{\vartheta^n}\psi + (1 - \psi)\,. \tag{4.10}$$

Note that $\tilde{\alpha}^n$ is bounded by (3.10).

4.3. **Diffusion-scaled processes.** To prove Theorem 3.2, we need to establish long-run average moment bounds for the diffusion-scaled processes under a class of scheduling policies, which agree with a proposed policy outside a compact set. We make this formal in Proposition 4.3. The proposed policy is given in the next definition.

Let $\mathcal{I}_0 \coloneqq \{i \in \mathcal{I} \colon \gamma_i = 0\}$. If $\mathcal{I}_0 \neq \varnothing$, then, without loss of generality, we assume that $\mathcal{I}_0 = \{1, \dots, |\mathcal{I}_0|\}$, where $|\mathcal{I}_0|$ denotes the cardinality of the set $\mathcal{I}_0$. In Definition 4.1 below, we introduce a modified priority scheduling policy which can be described as follows: First, $\lfloor n\rho_i / \sum_{i \in \mathcal{I}_0} \rho_i \rfloor \wedge x_i$ servers are allocated to each class $i \in \mathcal{I}_0$. Then, the remaining servers are allocated following the static priority rule.

Definition 4.1. The Markov policy $\check{z}^n$ is defined by

$$\check{z}_i^n(x) \;=\; \left\lfloor \frac{n\rho_i}{\sum_{i\in\mathcal{J}_0}\rho_i} + \left(n - \sum_{j\in\mathcal{J}_0}\left(x_j \wedge \left\lfloor \frac{n\rho_j}{\sum_{i\in\mathcal{J}_0}\rho_i}\right\rfloor\right) - \sum_{j=1}^{i-1}\left(x_j - \left\lfloor \frac{n\rho_j}{\sum_{i\in\mathcal{J}_0}\rho_i}\right\rfloor\right)^+\right)^+\right\rfloor \wedge x_i\,, \quad \forall\, i\in\mathcal{J}_0\,,$$

and

$$\check{z}_i^n(x) \;\coloneqq\; x_i \wedge \left(n - \sum_{j=1}^{i-1} x_j\right)^+, \quad \forall\, i \in \mathcal{J}\setminus\mathcal{J}_0\,.$$

We let $\check{q}_i^n(x) \coloneqq x_i - \check{z}_i^n(x)$, $i\in\mathcal{J}$.

In obtaining long-run average moment bounds, since the queueing system is in an alternative renewal environment, we do not work with the diffusion-scaled processes directly. To utilize the fact that (Ψ^n, K^n) is a Markov process, we introduce the following auxiliary process. We define the 'unscaled' process $\breve{X}^n$ by

$$\begin{aligned}\breve{X}_i^n(t) \;&\coloneqq\; X_i^n(0) + A_i^n(t) - S_i^n(t)\\ &\quad - R_{*,i}^n\left(\gamma_i^n \int_0^t \big(\breve{X}_i^n(s) - n\mu_i^n\rho_i\mathcal{R}^n(s) - Z_i^n(s)\big)\,\mathrm{d}s\right) + n\mu_i^n\rho_i\mathcal{R}^n(t)\\ &=\; X_i^n(t) + n\mu_i^n\rho_i\mathcal{R}^n(t) \quad \text{a.s.}\end{aligned} \tag{4.11}$$

for $i\in\mathcal{J}$ and $t\ge 0$, where $\mathcal{R}^n(t)$ is the residual time process for the system in the 'down' state given by

$$\mathcal{R}^n(t) \;=\; \sum_{k=1}^{N_{\mathsf{u}}^n(t)} d_k^n - \int_0^t \big(1-\Psi^n(s)\big)\,\mathrm{d}s\,,$$

and $N_{\mathsf{u}}^n(t)$ is the process counting the number of completed 'up' periods by time t. Here, the second equality in (4.11) follows by the fact that given $X^n(0)$, Ψ^n and Z^n, the evolution equation in (2.7) admits a unique solution. Also, if $\Psi^n(t)=1$, then $\mathcal{R}^n(t)=0$ and thus $\breve{X}^n(t) = X^n(t)$ a.s. Note that under a Markov policy $z^n\in\mathfrak{Z}_{\mathrm{sm}}^n$, the process $(\breve{X}^n, H^n, \Psi^n, K^n)$ is Markov with state space

$$\mathfrak{D} \;\coloneqq\; \big\{(\breve{x}, h, \psi, k)\in\mathbb{R}_+^d\times\mathbb{R}_+^d\times\{0,1\}\times\mathbb{R}_+\colon k\equiv 0 \text{ if } \psi=1\big\}\,,$$

and

$$Z^n(t) \;=\; z^n\big(\breve{X}^n(t) - n\mu_i^n\rho_i\mathcal{R}^n(t), H^n(t), \Psi^n(t), K^n(t)\big)\,.$$

Under $z^n\in\mathfrak{Z}_{\mathrm{sm}}^n$, the generator of $(\breve{X}^n, H^n, \Psi^n, K^n)$ denoted by $\breve{\mathcal{L}}_n^{z^n}$ is given by

$$\breve{\mathcal{L}}_n^{z^n} f(\breve{x}, h,\psi,k) \;=\; \overline{\mathcal{L}}_{n,\psi}^{z^n} f(\breve{x}, h,\psi,k) + \mathcal{I}_{n,\psi} f(\breve{x}, h,\psi,k) + \mathcal{Q}_{n,\psi} f(\breve{x}, h,\psi,k) \tag{4.12}$$

for $(\breve{x}, h,\psi,k)\in\mathfrak{D}$ and $f\in\mathcal{C}_b(\mathbb{R}^d\times\mathbb{R}_+^d\times\{0,1\}\times\mathbb{R}_+)$. The operators on the right-hand side of (4.12) are defined by

$$\begin{aligned}\overline{\mathcal{L}}_{n,\psi}^{z^n} f(\breve{x}, h,\psi,k) \;&\coloneqq\; \sum_{i\in\mathcal{J}} \frac{\partial f(\breve{x}, h,\psi,k)}{\partial h_i} + \sum_{i\in\mathcal{J}} r_i^n(h_i)\big(f(\breve{x}+e_i, h-h_ie_i,\psi,k) - f(\breve{x}, h,\psi,k)\big)\\ &\quad + \psi\sum_{i\in\mathcal{J}}\big(\mu_i^n z_i^n(\breve{x}, h, 1, 0) + \gamma_i^n q_i^n(\breve{x}, z^n)\big)\big(f(\breve{x}-e_i, h, 1, 0) - f(\breve{x}, h, 1, 0)\big)\\ &\quad + (1-\psi)\sum_{i\in\mathcal{J}}\gamma_i^n\big(f(\breve{x}-e_i, h,0,k) - f(\breve{x},h,0,k)\big)\int_{\mathbb{R}_*} q_i^n\big(\breve{x} - n\mu^n(y-k), z^n\big)\,\tilde{F}_{\breve{x},k}^{d_1^n}(\mathrm{d}y)\\ &\quad - (1-\psi)\sum_{i\in\mathcal{J}} n\rho_i\mu_i^n\frac{\partial f(\breve{x},h,0,k)}{\partial\breve{x}_i}\end{aligned} \tag{4.13}$$

with $q^n(\breve{x}, z^n) = \breve{x} - z^n$,

$$\mathcal{I}_{n,\psi} f(\breve{x}, h, \psi, k) \;\coloneqq\; \psi\, \beta^n_{\mathsf{u}} \int_{\mathbb{R}_*} \biggl(f\Bigl(\breve{x} + \frac{n}{\vartheta^n}\upmu^n y, h, 0, 0\Bigr) - f(\breve{x}, h, 1, 0)\biggr)\, F^{d_1}(\mathrm{d}y)\,, \tag{4.14}$$

and

$$\mathcal{Q}_{n,\psi} f(\breve{x}, h, \psi, k) \;\coloneqq\; (1-\psi)\biggl(\beta^n_{\mathsf{d}}(k)\bigl(f(\breve{x}, h, 1, 0) - f(\breve{x}, h, 0, k)\bigr) + \frac{\partial f(\breve{x}, h, 0, k)}{\partial k}\biggr)\,. \tag{4.15}$$

In (4.13), $\upmu^n \coloneqq (\mu^n_1\rho_1, \dots, \mu^n_d\rho_d)'$, $\tilde{F}^{d^n_1}_{\breve{x},k}$ denotes the conditional distribution of d^n_1 given $\{d^n_1 > k\}$, and $\{n\mu^n_i\rho_i(d^n_1 - k) \le \breve{x}_i \colon i \in \mathcal{I}\}$.

The first two terms on the right-hand side of (4.13) correspond to the extended generator associated with the renewal arrival processes. Compare this with (4.2). Conditioning on the alternative renewal process Ψ^n in the 'up' state, the third term on the right-hand side of (4.13) corresponds to the service and abandonment processes, and $\mathcal{I}_{n,\psi}$ corresponds to the residual time process $\mathcal{R}^n$ together with Ψ^n. Similarly, conditioning on the alternative renewal process in the 'down' state, the last two terms on the right-hand side of (4.13) correspond to the abandonment process and $\mathcal{R}^n$, respectively, and $\mathcal{Q}_{n,\psi}$ corresponds to (Ψ^n, K^n). The generators in (4.14) and (4.15) are analogous to the extended generator associated with the alternating renewal process in (4.7).

Remark 4.3. We sketch the derivation of $\mathcal{I}_{n,\psi}$. The rest of the terms in (4.12) follow by the calculation below and Remark 4.2. To simplify the calculation, we assume that the arrival processes are Poisson, and only consider the i^{th} component $(\breve{X}^n_i, \Psi^n, K^n)$, $i \in \mathcal{I}$. Note that $K^n(t) = 0$ when $\Psi^n(t) = 1$. Since there are no simultaneous jumps w.p.1., here we only consider the jumps caused by Ψ^n, that is, we consider

$$\sum_{j\in\mathbb{N}} \Bigl(\mathbb{E}_{\breve{x},1,0}\bigl[f(\breve{X}^n_i(t+s), \Psi^n(t+s), K^n(t+s)) \bigm| \breve{N}^n(t+s) - \breve{N}^n(t) = j\bigr] - f(\breve{x}, 1, 0)\Bigr) p^n_j(t,s)\,,$$

for $s, t \ge 0$, where $\breve{N}^n(t)$ denotes the number of jumps of Ψ^n up to time t, and $p^n_j(t,s) = \mathbb{P}\bigl(\breve{N}^n(t+s) - \breve{N}^n(t) = j\bigr)$, $j \in \mathbb{N}$. By the memoryless property of 'up' times, and using the same calculation as in Remark 4.2 for 'down' times, it is straightforward to check that

$$\lim_{s\searrow 0} \frac{1}{s}\, p^n_1(t,s) \;=\; \beta^n_{\mathsf{u}}\,, \quad \text{and} \quad \lim_{s\searrow 0} \frac{1}{s}\, p^n_j(t,s) \;=\; 0 \quad \text{for } j \ge 2\,,$$

and for any $t \ge 0$. By the continuity of K^n, we have

$$\lim_{s\searrow 0} \mathbb{P}\bigl(\breve{N}^n(t+s) - \breve{N}^n(t) = 1, K^n(t+s) = 0 \bigm| K^n(t) = 0\bigr) \;=\; 1\,.$$

Thus,

$$\begin{aligned}\lim_{s\searrow 0} \mathbb{E}_{\breve{x},1,0}\bigl[f(\breve{X}^n_i(t+s), \Psi^n(t+s), K^n(t+s)) \bigm| \breve{N}^n(t+s) - \breve{N}^n(t) = 1\bigr]& \\ = \mathbb{E}_{\breve{x},1,0}\biggl[f\Bigl(\breve{x} + n\mu^n_i\rho_i\frac{1}{\vartheta^n}d_1, 0, 0\Bigr)\biggr]&\,.\end{aligned}$$

This proves (4.14).

Definition 4.2. We define $\bar{x}^n_i(\breve{x}) \coloneqq \breve{x}_i - \rho_i n$, $i \in \mathcal{I}$,

$$\bar{x} \;=\; \bar{x}^n(\breve{x}) \;\coloneqq\; (\bar{x}^n_1(\breve{x}), \dots, \bar{x}^n_d(\breve{x}))'\,, \quad \tilde{x} \;=\; \tilde{x}^n(\breve{x}) \;\coloneqq\; n^{-1/2}\bar{x}^n(\breve{x})\,, \quad \breve{x} \in \mathbb{R}^d\,,$$

and

$$\mathfrak{A}^n_R \;\coloneqq\; \bigl\{x \in \mathbb{R}^d \colon |x - \rho n| \;\le\; R\sqrt{n}\bigr\}$$

for a positive constant R.

Let $\widetilde{\mathcal{L}}_n^{z_n}$ denote the generator of the scaled joint process $\widetilde{\Xi}^n := (\widetilde{X}^n, H^n, \Psi^n, K^n)$ with $\widetilde{X}^n := n^{-1/2}(\breve{X}^n - n\rho)$. The state space of $\widetilde{\Xi}^n$ is given by

$$\widetilde{\mathfrak{D}}^n := \bigl\{(\tilde{x}^n(\breve{x}), h, \psi, k) \in \mathbb{R}^d \times \mathbb{R}_+^d \times \{0,1\} \times \mathbb{R}_+ \colon \breve{x} \in \mathbb{R}_+^d,\ k \equiv 0 \text{ if } \psi = 1\bigr\}.$$

Then, under any $z^n \in \mathfrak{Z}_{\mathrm{sm}}^n$, we have

$$\widetilde{\mathcal{L}}_n^{z_n} f(\tilde{x}, h, \psi, k) = \breve{\mathcal{L}}_n^{z_n} f(\tilde{x}^n(\breve{x}), h, \psi, k)\,, \tag{4.16}$$

for $f \in \mathcal{C}_b(\mathbb{R}^d \times \mathbb{R}_+^d \times \{0,1\} \times \mathbb{R}_+)$.

The next lemma concerns the ergodicity of the process $\widetilde{\Xi}^n$ under the modified priority policy in Definition 4.1. Let $\mathcal{V}_{\kappa,\xi}(x) := \sum_{i\in\mathcal{I}} \xi_i |x_i|^\kappa$ for $x \in \mathbb{R}^d$, where $\kappa > 0$, and ξ is a positive vector. Define the function $\widetilde{\mathcal{V}}_{\kappa,\xi}^n \colon \mathbb{R}^d \times \mathbb{R}_+^d \times \{0,1\} \times \mathbb{R}_+ \to \mathbb{R}$ by

$$\begin{aligned}\widetilde{\mathcal{V}}_{\kappa,\xi}^n(x,h,\psi,k) &:= \mathcal{V}_{\kappa,\xi}(x) + \sum_{i\in\mathcal{I}} \eta_i^n(h_i)\bigl(\mathcal{V}_{\kappa,\xi}(x + n^{-1/2}e_i) - \mathcal{V}_{\kappa,\xi}(x)\bigr)\\ &\quad + \frac{\psi + \alpha^n(k)}{\vartheta^n} \sum_{i\in\mathcal{I}} \mu_i^n \xi_i \Bigl(\tilde{\mathcal{V}}_{\kappa,i}^n(x_i) + \eta_i^n(h_i)\bigl(\tilde{\mathcal{V}}_{\kappa,i}^n(x_i + n^{-1/2}) - \tilde{\mathcal{V}}_{\kappa,i}^n(x_i)\bigr)\Bigr)\,,\end{aligned} \tag{4.17}$$

where η_i^n and α^n are as in (4.3) and (4.8), respectively, and $\tilde{\mathcal{V}}_{\kappa,i}^n(x_i) := -|x_i|^\kappa$ for $x_i \in \mathbb{R}_+$ and $i \in \mathcal{I} \setminus \mathcal{I}_0$, and

$$\tilde{\mathcal{V}}_{\kappa,i}^n(x_i) := \begin{cases} -|x_i|^\kappa\,, & \text{for } x_i < \dfrac{\sqrt{n}\rho_i \sum_{j\in\mathcal{I}\setminus\mathcal{I}_0} \rho_j}{\sum_{j\in\mathcal{I}_0} \rho_j}\,,\\ -\dfrac{\sqrt{n}\rho_i \sum_{j\in\mathcal{I}\setminus\mathcal{I}_0} \rho_j}{\sum_{j\in\mathcal{I}_0} \rho_j} |x_i|^{\kappa-1}\,, & \text{for } x_i \ge \dfrac{\sqrt{n}\rho_i \sum_{j\in\mathcal{I}\setminus\mathcal{I}_0} \rho_j}{\sum_{j\in\mathcal{I}_0} \rho_j}\,, \end{cases} \quad \forall\, i \in \mathcal{I}_0\,.$$

The function $\widetilde{\mathcal{V}}_{\kappa,\xi}^n$ is constructed in such a manner as to allow us to take advantage of the identities in (4.4) and (4.10). We define the set

$$\mathcal{K}_n(x) := \biggl\{i \in \mathcal{I}_0 \colon x_i \ge \frac{\sqrt{n}\rho_i \sum_{j\in\mathcal{I}\setminus\mathcal{I}_0} \rho_j}{\sum_{j\in\mathcal{I}_0} \rho_j}\biggr\}.$$

Note that $\widetilde{\mathcal{L}}_n^{\check{z}^n}$ denotes the generator of $\widetilde{\Xi}^n$ under the modified priority scheduling policy in Definition 4.1. We have the following lemma.

Lemma 4.1. *Grant Assumptions 2.1, 2.2, and 3.2. For any even integer $\kappa \ge 2$, there exist positive constants $\widetilde{C}_0$ and $\widetilde{C}_1$, a positive vector $\xi \in \mathbb{R}_+^d$, and $\tilde{n} \in \mathbb{N}$ such that:*

$$\widetilde{\mathcal{L}}_n^{\check{z}^n} \widetilde{\mathcal{V}}_{\kappa,\xi}^n(\tilde{x}, h, \psi, k) \le \widetilde{C}_0 - \widetilde{C}_1 \sum_{i\in\mathcal{I}\setminus\mathcal{K}_n(\tilde{x})} \mathcal{V}_{\kappa,\xi}(\tilde{x}) - \widetilde{C}_1 \sum_{i\in\mathcal{K}_n(\tilde{x})} \mathcal{V}_{\kappa-1,\xi}(\tilde{x}) \tag{4.18}$$

for all $n > \tilde{n}$, and $(\tilde{x}, h, y, k) \in \widetilde{\mathfrak{D}}^n$. As a consequence, for all large enough n, $\widetilde{\Xi}^n$ is positive Harris recurrent under the modified priority scheduling policy $\check{z}^n$.

The proof of Lemma 4.1 is given in Appendix B. We continue with the following proposition, which plays a crucial role in proving Proposition 4.3. In its proof, especially, equation (4.26), we show the relationship between the processes $\hat{X}^n$ and $\widetilde{X}^n$.

Proposition 4.2. *Grant Assumptions 2.1, 2.2, and 3.2. Under the scheduling policy $\check{z}^n$ in Definition 4.1, and for any $\kappa > 0$, there exists $\check{n} \in \mathbb{N}$ such that*

$$\sup_{n>\check{n}} \limsup_{T\to\infty} \frac{1}{T}\, \mathbb{E}^{\check{z}^n}\biggl[\int_0^T |\hat{X}^n(s)|^\kappa\, \mathrm{d}s\biggr] < \infty\,. \tag{4.19}$$

Proof. Let $\kappa \ge 2$ be an arbitrary even integer. By (4.18), we have

$$\begin{aligned} \mathbb{E}^{\check{z}^n}\bigl[\widetilde{\mathcal{V}}^n_{\kappa,\xi}\bigl(\widetilde{\Xi}^n(T)\bigr)\bigr] - \mathbb{E}^{\check{z}^n}\bigl[\widetilde{\mathcal{V}}^n_{\kappa,\xi}\bigl(\widetilde{\Xi}^n(0)\bigr)\bigr] \;&=\; \mathbb{E}^{\check{z}^n}\biggl[\int_0^T \widetilde{\mathcal{L}}^{\check{z}^n}_n \widetilde{\mathcal{V}}^n_{\kappa,\xi}\bigl(\widetilde{\Xi}^n(s)\bigr)\,\mathrm{d}s\biggr] \\ &\le\; \widetilde{C}_0 T - \widetilde{C}_1\,\mathbb{E}^{\check{z}^n}\biggl[\int_0^T \mathcal{V}_{\kappa-1,\xi}\bigl(\widetilde{X}^n(s)\bigr)\,\mathrm{d}s\biggr]. \end{aligned} \tag{4.20}$$

Since $(\vartheta^n)^{-1}$ is of order $n^{-1/2}$ by Assumption 2.2, it follows by Young's inequality together with (3.10) that there exist some positive constants c_0 and c_1 such that $c_0(\mathcal{V}_{\kappa,\xi}-1) \le \widetilde{\mathcal{V}}^n_{\kappa,\xi} \le c_1(1+\mathcal{V}_{\kappa,\xi})$ for all large n. Note that $\hat{X}^n(0) = \widetilde{X}^n(0)$. Thus, by (4.20), we obtain

$$\widetilde{C}_1\,\mathbb{E}^{\check{z}^n}\biggl[\int_0^T \mathcal{V}_{\kappa-1,\xi}\bigl(\widetilde{X}^n(s)\bigr)\,\mathrm{d}s\biggr] \;\le\; (\widetilde{C}_0 + c_0)T + c_1\bigl(1 + \mathcal{V}_{\kappa,\xi}\bigl(\hat{X}^n(0)\bigr)\bigr) \tag{4.21}$$

for some positive constants C_3 and C_4. By dividing both sides of (4.21) by T, and taking $T\to\infty$, we have

$$\sup_{n>\check{n}}\,\limsup_{T\to\infty}\,\frac{1}{T}\,\mathbb{E}^{\check{z}^n}\biggl[\int_0^T |\widetilde{X}^n(s)|^{\kappa-1}\,\mathrm{d}s\biggr] \;<\; \infty\,. \tag{4.22}$$

Let $\mathbb{E} \equiv \mathbb{E}^{U^n}$ for some admissible scheduling policy U^n. We have

$$\frac{1}{T}\,\mathbb{E}\biggl[\int_0^T |\hat{X}^n_i(s) - \widetilde{X}^n_i(s)|^{\kappa-1}\,\mathrm{d}s\biggr] \;=\; (\mu^n_i\rho_i)^{\kappa-1}\frac{1}{T}\,\mathbb{E}\biggl[\int_0^T \bigl(\sqrt{n}\mathcal{R}^n(s)\bigr)^{\kappa-1}\,\mathrm{d}s\biggr] \quad \forall\, i\in\mathcal{I}\,. \tag{4.23}$$

We use the identity

$$\mathbb{E}\bigl[\bigl(\sqrt{n}\mathcal{R}^n(s)\bigr)^{\kappa-1}\bigr] \;=\; \mathbb{E}\bigl[\bigl(\sqrt{n}\mathcal{R}^n(s)\bigr)^{\kappa-1}\,|\,\mathcal{R}^n(s)>0\bigr]\,\mathbb{P}(\mathcal{R}^n(s)>0) \tag{4.24}$$

for any $s \ge 0$. Here $\mathcal{R}^n(s)$ is the residual time of the system in the 'down' state, and thus $\mathbb{E}[(\sqrt{n}\mathcal{R}^n(s))^{\kappa-1}|\mathcal{R}^n(s)>0] \le \mathbb{E}[(\sqrt{n}d^n_1)^{\kappa-1}] \le c_2$ for some positive constant c_2, by Assumption 2.2 and (3.10). Also, $\mathbb{P}(\mathcal{R}^n(s)>0) = \mathbb{P}(\Psi^n(s)=0)$, and it follows by [27, Theorem 3.4.4] that

$$\lim_{s\to\infty}\,\mathbb{P}(\Psi^n(s)=0) \;=\; \frac{(\vartheta^n)^{-1}}{(\beta^n_{\mathsf{u}})^{-1} + (\vartheta^n)^{-1}}\,,$$

which is of order $n^{-1/2}$ by Assumption 2.2. Therefore, applying (4.24), we obtain

$$\lim_{(n,T)\to\infty}\,\frac{1}{T}\,\mathbb{E}\biggl[\int_0^T \bigl(\sqrt{n}\mathcal{R}^n(s)\bigr)^{\kappa-1}\,\mathrm{d}s\biggr] \;=\; 0\,. \tag{4.25}$$

It follows by (4.23) and (4.25) that

$$\lim_{(n,T)\to\infty}\,\frac{1}{T}\,\mathbb{E}\biggl[\int_0^T \|\hat{X}^n(s) - \widetilde{X}^n(s)\|^{\kappa-1}\,\mathrm{d}s\biggr] \;=\; 0\,. \tag{4.26}$$

Thus (4.19) follows by (4.22) and (4.26). This completes the proof. □

The next proposition is used to prove the upper bound for the ergodic control problem in Section 5.3.2, where we adopt the spatial truncation technique developed in [4]. We first introduce a class of concatenated scheduling policies.

Definition 4.3. We define the *quantization function* $\varpi\colon \mathbb{R}^d_+ \to \mathbb{Z}^d_+$ by

$$\varpi(x) \;:=\; \biggl(\lfloor x_1\rfloor,\dots,\lfloor x_{d-1}\rfloor,\lfloor x_d\rfloor + \sum_{i=1}^d (x_i - \lfloor x_i\rfloor)\biggr)\,.$$

For a sequence $v^n\colon \mathbb{R}^d \to \mathcal{S}$, $n \in \mathbb{N}$, of continuous functions satisfying $v^n\bigl(\tilde{x}^n(x)\bigr) = e_d$ if $x \notin \mathfrak{A}_R^n$, $R > 1$, with $\mathfrak{A}_R^n$ as in Definition 4.2, we define the map

$$q^n[v^n](x) \;\coloneqq\; \begin{cases} \varpi\bigl(\bigl(\langle e, x\rangle - n\bigr)^+ v^n\bigl(\tilde{x}^n(x)\bigr)\bigr) & \text{for } \sup_{i\in\mathcal{I}}|\tilde{x}^n(x)| \;\le\; \frac{1}{2d}\sqrt{n}\bigl(\min_i\, \rho_i\bigr)\,, \\ \check{q}^n(x) & \text{for } \sup_{i\in\mathcal{I}}|\tilde{x}^n(x)| \;>\; \frac{1}{2d}\sqrt{n}\bigl(\min_i\, \rho_i\bigr)\,, \end{cases}$$

and the scheduling policy $z^n[v^n](x) \coloneqq x - q^n[v^n](x)$

Proposition 4.3. *Under the scheduling policy $z^n[v^n]$ in Definition 4.3, the conclusions in Lemma 4.1 and Proposition 4.2 hold.*

Proof. For all sufficiently large n, we have $q_i^n[v^n](\breve{x}) \le 2dR\sqrt{n}$ for $\breve{x} \in \mathfrak{A}_R^n$ (see also the proof of [4, Lemma 5.1]). If $\sup_{i\in\mathcal{I}}|\tilde{x}_i^n(\breve{x})| \le \frac{1}{d}\sqrt{n}\bigl(\min_i\, \rho_i\bigr)$, it is evident that $\sum_{i=1}^{d-1} \breve{x}_i \le n$, and thus $z^n[e_d]$ is equivalent to the modified priority policy on this set. Therefore, the result follows by the argument in Lemma 4.1 and Proposition 4.2. ☐

5. Asymptotic Optimality

5.1. **Results concerning the limiting jump diffusion.** In this subsection, we present some optimality results for the limiting jump diffusion. These results are used in proving asymptotic optimality.

Recall that a stationary Markov control v is called stable if the process under v is positive recurrent, and the set of such controls is denoted by $\mathfrak{U}_{\mathrm{ssm}}$. Let $\mathcal{G}$ denote the set of ergodic occupation measures, that is,

$$\mathcal{G} \;\coloneqq\; \biggl\{\uppi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{U})\colon \int_{\mathbb{R}^d\times\mathbb{U}} \mathcal{A}f(x,u)\,\uppi(\mathrm{d}x,\mathrm{d}u) \;=\; 0 \quad \forall\, f \in \mathcal{C}_c^\infty(\mathbb{R}^d)\biggr\}\,. \tag{5.1}$$

See Section 2.1 in [28] for more details.

We summarize the characterization of optimal controls for the limiting jump diffusion in the following theorem. Recall the definition of d_1 in Assumption 2.2.

Theorem 5.1. *Assume that $\mathbb{E}[(d_1)^{m+1}] < \infty$ with m as in (3.5). The following hold:*

(i) *For $\alpha > 0$, V_α in (3.7) is the minimal nonnegative solution in $\mathcal{C}^{2,r}(\mathbb{R}^d)$, $r \in (0,1)$, to the HJB equation*

$$\min_{u\in\mathbb{U}}\bigl[\mathcal{A}V_\alpha(x,u) + \mathcal{R}(x,u)\bigr] \;=\; \alpha V_\alpha(x) \quad \text{a.e. in } \mathbb{R}^d\,. \tag{5.2}$$

In addition, V_α has at most polynomial growth with degree m. Moreover, a stationary Markov control v is optimal for the α-discounted problem if and only if it is an a.e. measurable selector from the minimizer in (5.2).

(ii) *There exists a solution $V \in \mathcal{C}^{2,r}(\mathbb{R}^d)$, $r \in (0,1)$, to the HJB equation*

$$\min_{u\in\mathbb{U}}\bigl[\mathcal{A}V(x,u) + \mathcal{R}(x,u)\bigr] \;=\; \varrho_* \quad \text{a.e. in } \mathbb{R}^d\,. \tag{5.3}$$

Moreover, a stationary Markov control v is optimal for the ergodic control problem if and only if it is an a.e. measurable selector from the minimizer (5.3).

Proof. We first consider (i). It follows by Remark 5.1 in [24] and Proposition 4.1 that Assumptions 2.1 and 2.2 in [17] hold with $\mathcal{V}_\circ$ and $\mathcal{V}$ having at most polynomial growth of degree m. Since $\mathbb{E}[(d_1)^{m+1}] < \infty$, then (4.1) satisfies Assumption 5.1 in [17]. Therefore, the results in part (i) follow by Theorems 5.1 and 5.3 in [17]. Note that by (5.4) in [17], V_α has at most polynomial growth of degree m. Similarly, the claim in part (ii) follows by Theorems 5.2 and 5.3 of [17]. ☐

Remark 5.1. If there is no jump part in (4.1), then it corresponds to the controlled limiting diffusion for $GI/M/n+M$ queues. If we define the optimal control problems for the limiting diffusion in the same way as in (3.7) and (3.9), then the results in Theorem 5.1 still hold when $\mathcal{A}$ in (4.1) does not contain the jump component. As a consequence, part (i) of Theorem 5.1 corresponds to [1, Theorem 3].

If we consider (3.9) over all stable Markov controls, then the ergodic control problem is equivalent to $\min_{\pi\in\mathcal{G}} \int_{\mathbb{R}^d\times\mathbb{U}} \mathcal{R}(x,u)\,\pi(\mathrm{d}x,\mathrm{d}u)$, see, for example, [17, Section 4]. We summarize a result on ϵ-optimal controls for the ergodic problem in the next proposition, which follows directly by Corollary 7.1 in [17]. Note that the constant control $v\equiv e_d$ also satisfies Proposition 4.1. Recall that a stationary Markov control v is called precise if it is a measurable map from $\mathbb{R}^d$ to $\mathbb{U}$.

Proposition 5.1. *Assume that $\mathbb{E}[(d_1)^m]<\infty$, with m as in (3.5). For any $\epsilon>0$, there exist a continuous precise control $v_\epsilon\in\mathfrak{U}_{\mathrm{ssm}}$, and $R\equiv R(\epsilon)\in\mathbb{N}$ such that $v_\epsilon\equiv e_d$ on $\bar{B}_R^c$, and v_ϵ is ϵ-optimal, that is,*

$$\int_{\mathbb{R}^d\times\mathbb{U}} \mathcal{R}(x,u)\,\pi_{v_\epsilon}(\mathrm{d}x,\mathrm{d}u) \;\le\; \varrho_*+\epsilon\,.$$

5.2. **Proof of Theorem 3.1.** To prove Theorem 3.1, we use the approach developed in [1]. We first establish a key moment estimate for the diffusion-scaled process $\hat{X}^n$, whose proof is similar to that of [1, Lemma 3].

Lemma 5.1. *Grant the hypotheses in Theorem 3.1. Then*

$$\mathbb{E}\bigl[\|\hat{X}^n(t)\|^{m_A}\bigr] \;\le\; c_1(1+t^{m_1})(1+\|x\|^{m_1}) \quad \forall\, t\ge 0\,, \tag{5.4}$$

where c_1 and m_1 are some positive constants independent of n, x and t.

Proof. Recall $\hat{L}^n$ and $\hat{X}^n$ in (3.1), and $\hat{W}^n$ in (3.2). Let $\hat{\Phi}^n$ be a d-dimensional process defined by $\hat{\Phi}^n_i(\cdot) := \mu^n_i\int_0^\cdot \hat{Z}^n_i(s)\bigl(1-\Psi^n(s)\bigr)\,\mathrm{d}s$, for $i\in\mathcal{I}$. Then,

$$\mu^n_i\int_0^t \hat{Z}^n_i(s)\Psi^n(s)\,\mathrm{d}s \;=\; -\hat{\Phi}^n_i(t)+\mu^n_i\int_0^t \hat{Z}^n_i(s)\,\mathrm{d}s \qquad \forall\, t\ge 0\,.$$

Thus, we obtain

$$\hat{X}^n_i(t) \;=\; \hat{X}^n_i(0)+\ell^n_i t+\hat{W}^n_i(t)+\hat{\Phi}^n_i(t)+\hat{L}^n_i(t)-\mu^n_i\int_0^t \hat{Z}^n_i(s)\,\mathrm{d}s-\gamma^n_i\int_0^t \hat{Q}^n_i(s)\,\mathrm{d}s$$

for all $t\ge 0$ and $i\in\mathcal{I}$. Following the same method as in [1, Lemma 3], we have

$$\begin{aligned}\|\hat{X}^n(t)\| \;\le\; C\biggl[&1+t^2+\|\hat{X}^n(0)\|+\|\hat{W}^n(t)+\hat{L}^n(t)+\hat{\Phi}^n(t)\|\\ &+\int_0^t\|\hat{W}^n(s)+\hat{L}^n(s)+\hat{\Phi}^n(s)\|\,\mathrm{d}s+\int_0^t\int_0^s\|\hat{W}^n(r)+\hat{L}^n(r)+\hat{\Phi}^n(r)\|\,\mathrm{d}r\,\mathrm{d}s\biggr]\end{aligned} \tag{5.5}$$

for some positive constant C. Let

$$\widehat{N}^n(t) \;:=\; \max\biggl\{k\ge 0\colon \sum_{i=1}^k u^n_i\le t\biggr\}$$

with u^n as in (2.2). By Assumption 2.2, $\widehat{N}^n(t)$ is a Poisson process with rate β^n_{u}. Then, we obtain

$$\mathbb{E}\Bigl[\|\hat{L}^n(t)\|^{m_A}\Bigr] \;\le\; C_1\,\mathbb{E}\bigl[\bigl(\sqrt{n}C^n_{\mathsf{d}}(t)\bigr)^{m_A}\bigr] \;\le\; C_1\biggl(\frac{\sqrt{n}}{\vartheta^n}\biggr)^{m_A}\mathbb{E}\Biggl[\Biggl(\sum_{i=1}^{\widehat{N}^n(t)+1} d_i\Biggr)^{m_A}\Biggr] \;\le\; C_2(1+t^{m_2}) \tag{5.6}$$

for some positive constants $C_1 = \sup\{\mu_i^n \rho_i \colon n \in \mathbb{N}, i \in \mathcal{I}\}$, C_2, and m_2. The third inequality in (5.6) follows by the independence of $\widehat{N}^n$ and d_i, and Assumption 3.1. On the other hand, for some positive constant C_3, we have

$$|n^{-1/2} \hat{Z}_i^n(s)| \,\le\, C_3\bigl(1 + n^{-1} A_i^n(s)\bigr) \quad \text{a.s.} \quad \forall\, s \ge 0\,. \tag{5.7}$$

Thus,

$$\begin{aligned}\mathbb{E}\Bigl[\bigl|\hat{\Phi}_i^n(t)\bigr|^{m_A}\Bigr] \,&\le\, \mu_i^n\, \mathbb{E}\biggl[\biggl(\int_0^t \bigl|n^{-1/2}\hat{Z}_i^n(s)\bigr|\bigl|\sqrt{n}\bigl(1-\Psi^n(s)\bigr)\bigr|\,\mathrm{d}s\biggr)^{m_A}\biggr]\\ &\le\, \mu_i^n (C_3)^{m_A}\Bigl(1 + \sup_{s\le t}\mathbb{E}\bigl[n^{-1}A_i^n(s)\bigr]\Bigr)^{m_A}\,\mathbb{E}\Bigl[\bigl(\sqrt{n}C_{\mathsf{d}}^n(t)\bigr)^{m_A}\Bigr]\\ &\le\, C_4(1+t^{m_3})\end{aligned} \tag{5.8}$$

for some positive constant C_4, where the second inequality follows by (5.7) and the independence of A^n and Ψ^n, and the third inequality follows by [29, Theorem 4] and (5.6). Therefore, following the argument in the proof of [1, Lemma 3], and using (5.5), (5.6), and (5.8), we establish (5.4). This completes the proof. □

Proof of Theorem 3.1. We first prove the lower bound:

$$\liminf_{n\to\infty}\, \hat{V}_\alpha^n\bigl(\hat{X}^n(0)\bigr) \,\ge\, V_\alpha(x)\,.$$

By Theorem 5.1, the partial derivatives of $V_\alpha(x)$ up to order two are locally Hölder continuous. Let $V_\alpha^l := \chi_l \circ V_\alpha = \chi_l(V_\alpha)$, where $\chi_l \in \mathcal{C}^2(\mathbb{R})$ satisfies $\chi_l(x) = x$ for $x \le l$ and $\chi_l(x) = l+1$ for $x \ge l+2$. Let $\mathcal{L}\colon \mathcal{C}^2(\mathbb{R}^d) \to \mathcal{C}^2(\mathbb{R}^d \times \mathcal{S})$ be the local operator defined by

$$\mathcal{L}\varphi(x,u) \,:=\, \langle b(x,u), \nabla\varphi(x)\rangle + \frac{1}{2}\sum_{i\in\mathcal{I}} \lambda_i(1+c_{a,i}^2)\,\partial_{ii}\varphi(x)\,, \qquad \varphi \in \mathcal{C}^2(\mathbb{R}^d)\,.$$

Compare this with (4.1). We define $\mathcal{H}(x,p) := \min_{u\in\mathbb{U}}[\langle b(x,u), p\rangle + \mathcal{R}(x,u)]$, for $(x,p) \in \mathbb{R}^d \times \mathbb{R}^d$. By Itô's formula, for any $l > \sup_{B_R} V_\alpha$, it follows that

$$\begin{aligned}\mathrm{e}^{-\alpha(t\wedge\tau_R)}V_\alpha^l(X_{t\wedge\tau_R}) \,=\, V_\alpha^l(x) - \int_0^{t\wedge\tau_R} \alpha\mathrm{e}^{-\alpha s}\, V_\alpha(X_s)\,\mathrm{d}s + \int_0^{t\wedge\tau_R} \mathrm{e}^{-\alpha s}\,\mathcal{L}V_\alpha(X_s,U_s)\,\mathrm{d}s\\ + \int_0^{t\wedge\tau_R}\langle \mathrm{e}^{-\alpha s}\,\nabla V_\alpha(X_s), \Sigma\,\mathrm{d}W_s\rangle + \int_0^{t\wedge\tau_R}\int_{\mathbb{R}_*} \mathrm{e}^{-\alpha s}\,\bigl(V_\alpha^l(X_{s-}+\lambda y) - V_\alpha(X_{s-})\bigr)\mathcal{N}_L(\mathrm{d}s,\mathrm{d}y)\,,\end{aligned}$$

where $\mathcal{N}_L$ is the Poisson random measure of $\{L_t \colon t \ge 0\}$ with the intensity Π_L. Thus, applying (5.2), we obtain

$$\begin{aligned}\mathrm{e}^{-\alpha(t\wedge\tau_R)}V_\alpha^l(X_{t\wedge\tau_R}) \,&= V_\alpha^l(x) + \int_0^{t\wedge\tau_R} \mathrm{e}^{-\alpha s}\,\langle b(X_s,U_s), \nabla V_\alpha(X_s)\rangle\,\mathrm{d}s\\ &\quad + \int_0^{t\wedge\tau_R}\langle \mathrm{e}^{-\alpha s}\,\nabla V_\alpha(X_s), \Sigma\,\mathrm{d}W_s\rangle - \int_0^{t\wedge\tau_R} \mathrm{e}^{-\alpha s}\,\mathcal{H}\bigl(X_s, \nabla V_\alpha(X_s)\bigr)\,\mathrm{d}s\\ &\quad + \int_0^{t\wedge\tau_R}\int_{\mathbb{R}_*} \mathrm{e}^{-\alpha s}\,\bigl(V_\alpha^l(X_{s-}+\lambda y) - V_\alpha(X_{s-})\bigr)\widetilde{\mathcal{N}}_L(\mathrm{d}s,\mathrm{d}y)\\ &\quad + \int_0^{t\wedge\tau_R}\int_{\mathbb{R}_*} \mathrm{e}^{-\alpha s}\,\bigl(V_\alpha^l(X_{s-}+\lambda y) - V_\alpha(X_{s-}+\lambda y)\bigr)\Pi_L(\mathrm{d}s,\mathrm{d}y)\,,\end{aligned}$$

where $\widetilde{\mathcal{N}}_L(t,A) = \mathcal{N}_L(t,A) - t\,\Pi_L(A)$ for any Borel set $A \subset \mathbb{R}$. Repeating the same calculation as for the claim (71) in [1], we obtain

$$\begin{aligned}
\mathrm{e}^{-\alpha(t\wedge\tau_R)} V^l_\alpha(X_t) \;\ge\; V^l_\alpha(x) &+ \int_0^{t\wedge\tau_R} \langle \mathrm{e}^{-\alpha s}\,\nabla V^l_\alpha(X_s), \Sigma\,\mathrm{d}W_s\rangle - \int_0^{t\wedge\tau_R} \mathrm{e}^{-\alpha s}\,\mathcal{R}(X_s,U_s)\,\mathrm{d}s \\
&+ \int_0^{t\wedge\tau_R}\int_{\mathbb{R}_*} \mathrm{e}^{-\alpha s}\,\bigl(V^l_\alpha(X_{s-}+\lambda y) - V_\alpha(X_{s-})\bigr)\widetilde{\mathcal{N}}_L(\mathrm{d}s,\mathrm{d}y) \\
&+ \int_0^{t\wedge\tau_R}\int_{\mathbb{R}_*} \mathrm{e}^{-\alpha s}\,\bigl(V^l_\alpha(X_{s-}+\lambda y) - V_\alpha(X_{s-}+\lambda y)\bigr)\Pi_L(\mathrm{d}s,\mathrm{d}y)\,.
\end{aligned} \tag{5.9}$$

Note that $\widetilde{\mathcal{N}}_L$ is a martingale measure and V_α is nonnegative. Taking expectations on both sides of (5.9), the second and fourth terms on the right-hand side of (5.9) vanish. Thus, first taking limits as $l \to \infty$, and then as $R \to \infty$, it follows by the monotone convergence theorem that

$$\mathbb{E}\biggl[\int_0^t \mathrm{e}^{-\alpha s}\,\mathcal{R}(X_s,U_s)\,\mathrm{d}s\biggr] \;\ge\; V_\alpha(x) - \mathbb{E}\bigl[\mathrm{e}^{-\alpha t}V_\alpha(X_t)\bigr]\,.$$

Applying Theorem 5.1 it follows that solutions of (5.2) have at most polynomial growth of degree m, which corresponds to [1, Proposition 5 (i)]. Note that Lemma 5.1 corresponds to Lemma 3 in [1]. The rest of the proof of the lower bound follows exactly the proof of [1, Theorem 4 (i)].

To prove (3.8), we construct a sequence of asymptotically optimal scheduling policies U^n. Let v_α be an optimal control to (5.2). Recall the quantization function in Definition 4.3. We define a sequence of scheduling policies

$$\bar{z}^n[v_\alpha](\hat{x}) \;:=\; \begin{cases} \varpi\bigl(\langle e,\hat{x}\rangle^+ v_\alpha(\hat{x})\bigr)\,, & \text{if } \hat{x} \in \hat{\mathfrak{X}}^n\,,\\ \breve{z}^n(\sqrt{n}\hat{x} + n\rho) & \text{if } \hat{x} \notin \hat{\mathfrak{X}}^n\,,\end{cases}$$

where $\breve{z}^n$ is the modified priority policy in Definition 4.1, and

$$\hat{\mathfrak{X}}^n \;:=\; \bigl\{n^{-1/2}(x - n\rho)\colon x \in \mathbb{R}^d,\ \langle e,x\rangle \le x_i\ \forall\, i \in \mathcal{I}\bigr\}\,.$$

Here the policy on $(\hat{\mathfrak{X}}^n)^c$ may be chosen arbitrarily. Let $U^n[v_\alpha]$ be the equivalent parameterization of $\bar{z}^n[v_\alpha]$. Following the proof of [1, Theorem 2 (i)], we obtain

$$\int_0^{\cdot} \mathrm{e}^{-\alpha s}\,\Upsilon^n(s)\,\mathrm{d}s \;\Rightarrow\; 0\,,$$

where

$$\Upsilon^n(s) \;:=\; \bigl\langle b\bigl(\hat{X}^n(s), U^n[v_\alpha](s)\bigr), \nabla V_\alpha\bigl(\hat{X}^n(s)\bigr)\bigr\rangle + \mathcal{R}\bigl(\hat{X}^n(s), U^n[v_\alpha](s)\bigr) - \mathcal{H}\bigl(\hat{X}^n(s), \nabla V_\alpha\bigl(\hat{X}^n(s)\bigr)\bigr)\,.$$

Thus, by using the method in [1, Theorem 4 (ii)], and repeating the above calculation, we obtain

$$\limsup_{n\to\infty}\, \hat{V}^n_\alpha\bigl(\hat{X}^n(0)\bigr) \;\le\; V_\alpha(x)\,.$$

This completes the proof. □

5.3. **Proof of Theorem 3.2.** In this section, we prove Theorem 3.2 by establishing lower and upper bounds.

5.3.1. *The lower bound.* We show that

$$\liminf_{n\to\infty}\, \varrho^n\bigl(\hat{X}^n(0)\bigr) \;\ge\; \varrho_*\,. \tag{5.10}$$

The proof is given at the end of this subsection.

We need the following lemma whose proof is similar to that of Proposition 4.2, and is given in Appendix B.

Lemma 5.2. *Grant the hypotheses in Assumptions 2.1, 2.2, and 3.2. For any $m > 1$, and any sequence $\{z^n \in \mathfrak{Z}^n_{\mathrm{sm}} \colon n \in \mathbb{N}\}$ with $\sup_n \hat{J}(\hat{X}^n(0), z^n) < \infty$, there exists $n_\circ > 0$ such that*

$$\sup_{n>n_\circ} \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}^{z^n}\left[\int_0^T |\hat{X}^n(s)|^m\,\mathrm{d}s\right] \;<\; \infty\,. \tag{5.11}$$

The main challenge in the proof lies in approximating the generator of the diffusion-scaled process with the generator of the limiting jump diffusion. Recall the extended generator $\mathcal{H}^n$ of (A^n, H^n) in (4.2). We define the function $\phi^n[f]$ by

$$\begin{aligned}\phi^n[f](x,h) \;\coloneqq\; f(x) &+ \sum_{j\in\mathcal{J}} \hat{\phi}^n_{1,j}[f](x,h) + \sum_{j\in\mathcal{J}} \frac{c^2_{a,j}-1}{2\sqrt{n}}\partial_j f(x)\\ &+ \sum_{j\in\mathcal{J}} \hat{\phi}^n_{2,j}[f](x,h) + \sum_{j\in\mathcal{J}} \frac{\kappa^n_j(h_j)}{n}\partial_{jj} f(x) + \sum_{j=1}^{d-1} \hat{\phi}^n_{3,j}[f](x,h)\end{aligned} \tag{5.12}$$

for any $f \in \mathcal{C}^\infty_c(\mathbb{R}^d)$, and $n \in \mathbb{N}$, where

$$\hat{\phi}^n_{1,j}[f](x,h) \;\coloneqq\; \frac{1}{j\,!}\sum_{i_j\in\mathcal{J}}\sum_{i_{j-1}\neq i_j}\cdots\sum_{i_1\notin\{i_l\colon l>1\}}\prod_{r=1}^{j}\eta^n_{i_r}(h_{i_r})\,[f]^{1,n}_{i_1\cdots i_j}(x)\,,$$

with

$$\begin{aligned}[f]^{1,n}_{i_1\cdots i_j}(x) \;&\coloneqq\; [f]^{1,n}_{i_1\cdots i_{j-1}}(x+n^{-1/2}e_{i_j}) - [f]^{1,n}_{i_1\cdots i_{j-1}}(x)\,,\\ [f]^{1,n}_{i_1}(x) \;&\coloneqq\; f(x+n^{-1/2}e_{i_1}) - f(x)\,.\end{aligned} \tag{5.13}$$

The function $\hat{\phi}^n_{2,j}[f]$ is defined analogously to (5.13) with $[f]^{1,n}_{i_1\cdots i_j}$ and $[f]^{1,n}_{i_1}$ replaced by $[f]^{2,n}_{i_1\cdots i_j}$ and

$$[f]^{2,n}_{i_1}(x) \;\coloneqq\; \sum_{j\in\mathcal{J}}\frac{c^2_{a,j}-1}{2\sqrt{n}}\bigl(\partial_j f(x+n^{-1/2}e_{i_1}) - \partial_j f(x)\bigr)\,,$$

respectively. Also,

$$\hat{\phi}^n_{3,j}[f](x,h) \;\coloneqq\; \frac{1}{j\,!}\sum_{i_j\in\mathcal{J}}\sum_{i_{j-1}\neq i_j}\cdots\sum_{i_1\notin\{i_l\colon l>1\}}\prod_{r=2}^{j+1}\eta^n_{i_r}(h_{i_r})\frac{\kappa^n_{i_1}(h_{i_1})}{n}\,[f]^{3,n}_{i_1\cdots i_{j+1}}(x)$$

with $[f]^{3,n}_{i_1\cdots i_{j+1}}(x)$ defined analogously to (5.13), and

$$[f]^{3,n}_{i_1 i_2}(x) \;\coloneqq\; \partial_{i_1 i_1} f(x+n^{-1/2}e_{i_2}) - \partial_{i_1 i_1} f(x) \quad \text{for } i_1, i_2, \dots, i_j,\ j\in\mathcal{J}\,.$$

Note that $\phi^n[f]$ is bounded by Assumption 3.2 (i).

The extended generator $\widetilde{\mathcal{H}}^n$ of the scaled process $(\hat{A}^n, H^n)$ is given by $\widetilde{\mathcal{H}}^n f(\tilde{x},h) = \mathcal{H}^n f(\tilde{x}^n(x),h)$, for $f \in \mathcal{C}_b(\mathbb{R}^d \times \mathbb{R}^d_+)$. We have the following lemma.

Lemma 5.3. *Grant Assumption 2.1 and Assumption 3.2 (i). Then,*

$$\begin{aligned}\widetilde{\mathcal{H}}^n\phi^n[f](\tilde{x},h) \;=\; &\sum_{i\in\mathcal{J}}\frac{\lambda^n_i}{\sqrt{n}}\partial_i f(\tilde{x}) + \sum_{i\in\mathcal{J}}\frac{\lambda^n_i c^2_{a,i}}{2n}\partial_{ii} f(\tilde{x})\\ &+ \sum_{i\in\mathcal{J}}\frac{\lambda^n_i}{n}\sum_{j\in\mathcal{J}}\Bigl(\eta^n_j(h_j) + \frac{c^2_{a,j}-1}{2}\Bigr)\partial_{ij} f(\tilde{x}) + \mathcal{O}\Bigl(\frac{1}{\sqrt{n}}\Bigr)\end{aligned} \tag{5.14}$$

for all $f \in \mathcal{C}^\infty_c(\mathbb{R}^d)$ and $(\tilde{x},h) \in \mathbb{R}^d \times \mathbb{R}^d_+$.

Proof. Note that

$$\hat{\phi}^n_{1,1}[f] \,=\, \sum_{i\in\mathcal{I}} \eta^n_i(h_i)\bigl(f(\tilde{x}+n^{-1/2}e_i)-f(\tilde{x})\bigr)\,,$$

$$\hat{\phi}^n_{2,1}[f] \,=\, \sum_{i\in\mathcal{I}} \eta^n_i(h_i)\sum_{j\in\mathcal{I}}\frac{c^2_{a,j}-1}{2\sqrt{n}}\bigl(\partial_j f(\tilde{x}+n^{-1/2}e_{i_1})-\partial_j f(\tilde{x})\bigr)\,.$$

Using (4.4) and (4.6), and the Taylor expansion, we have

$$\begin{aligned}
&\widehat{\mathcal{H}}^n\biggl(f+\hat{\phi}^n_{1,1}[f]+\sum_{j\in\mathcal{I}}\frac{c^2_{a,j}-1}{2\sqrt{n}}\partial_j f+\hat{\phi}^n_{2,1}[f]+\sum_{j\in\mathcal{I}}\frac{\kappa^n_j(h_j)}{n}\partial_{jj}f\biggr)(\tilde{x},h)\\
&\quad= \sum_{i\in\mathcal{I}}\frac{\lambda^n_i}{\sqrt{n}}\partial_i f(\tilde{x})+\sum_{i\in\mathcal{I}}\frac{\lambda^n_i c^2_{a,i}}{2n}\partial_{ii}f(\tilde{x})+\sum_{i\in\mathcal{I}}\frac{\lambda^n_i}{n}\sum_{j\neq i}\frac{c^2_{a,j}-1}{2}\partial_{ij}f(\tilde{x})+\mathcal{O}\Bigl(\frac{1}{\sqrt{n}}\Bigr)\\
&\qquad+\sum_{i\in\mathcal{I}} r^n_i(h_i)\sum_{j\neq i}\eta^n_j(h_j)\Bigl([f]^{1,n}_{ij}(\tilde{x})+[f]^{2,n}_{ij}(\tilde{x})\Bigr)\\
&\qquad\qquad+\sum_{i\in\mathcal{I}}\frac{\lambda^n_i}{n}\biggl(\eta^n_i(h_i)+\frac{c^2_{a,i}-1}{2}\biggr)\partial_{ii}f(\tilde{x})+\sum_{i\in\mathcal{I}} r^n_i(h_i)\sum_{j\neq i}\frac{\kappa^n_j(h_j)}{n}[f]^{3,n}_{ij}(\tilde{x})\,.
\end{aligned}\tag{5.15}$$

It is straightforward to verify that

$$\begin{aligned}
&\widehat{\mathcal{H}}^n\bigl(\hat{\phi}^n_{1,2}[f]+\hat{\phi}^n_{2,2}[f]+\hat{\phi}^n_{3,1}[f]\bigr)(\tilde{x},h)\\
&\quad=\sum_{i\in\mathcal{I}}\bigl(\dot{\eta}^n_i(h_i)-\eta^n_i(h_i)r^n_i(h_i)\bigr)\sum_{j\neq i}\eta^n_j(h_j)\Bigl([f]^{1,n}_{ij}(\tilde{x})+[f]^{2,n}_{ij}(\tilde{x})\Bigr)\\
&\quad+\frac{1}{2}\sum_{i\in\mathcal{I}} r^n_i(h_i)\sum_{j\neq i}\sum_{k\neq i,j}\eta^n_j(h_j)\eta^n_k(h_k)\Bigl([f]^{1,n}_{ijk}(\tilde{x})+[f]^{2,n}_{ijk}(\tilde{x})\Bigr)\\
&\quad+\sum_{i\in\mathcal{I}}\biggl(\bigl(\dot{\eta}^n_i(h_i)-\eta^n_i(h_i)r^n_i(h_i)\bigr)\sum_{j\neq i}\frac{\kappa^n_j(h_j)}{n}+\bigl(\dot{\kappa}^n_i-r^n_i(h_i)\kappa^n_i(h_i)\bigr)\sum_{j\neq i}\frac{\eta^n_j(h_j)}{n}\biggr)[f]^{3,n}_{ij}(\tilde{x})\\
&\quad+\sum_{i\in\mathcal{I}} r^n_i(h_i)\sum_{j\neq i}\eta^n_j(h_j)\sum_{k\neq i,j}\frac{\kappa^n_k(h_k)}{n}[f]^{3,n}_{ijk}(\tilde{x})
\end{aligned}\tag{5.16}$$

for any $(\tilde{x},h)\in\mathbb{R}^d\times\mathbb{R}^d_+$. Applying (4.4) and (4.6), and combining the first term on the right-hand side of (5.16) with the third, fifth and sixth terms on the right-hand side of (5.15), we obtain the third term on the right-hand side of (5.14). We repeat this procedure until all the terms r^n_i are canceled. This proves (5.14). □

Definition 5.1. We define the operator $\hat{\mathcal{A}}^n\colon \mathcal{C}^2(\mathbb{R}^d)\to\mathcal{C}^2(\mathbb{R}^d\times\mathcal{S})$ by

$$\hat{\mathcal{A}}^n f(x,u) \,\coloneqq\, \sum_{i\in\mathcal{I}}\Bigl(\mathcal{A}^n_{1,i}(x,u)\partial_i f(x)+\frac{1}{2}\mathcal{A}^n_{2,i}(x,u)\partial_{ii}f(x)\Bigr)\,,$$

where $\mathcal{A}^n_{1,i},\mathcal{A}^n_{2,i}\colon\mathbb{R}^d\times\mathcal{S}\to\mathbb{R}$, $i\in\mathcal{I}$, are given by

$$\begin{aligned}
\mathcal{A}^n_{1,i}(x,u) &\,\coloneqq\, \ell^n_i-\mu^n_i(x_i-\langle e,x\rangle^+u_i)-\gamma^n_i\langle e,x\rangle^+u_i\,,\\
\mathcal{A}^n_{2,i}(x,u) &\,\coloneqq\, \frac{\lambda^n_i}{n}c^2_{a,i}+\rho_i\mu^n_i+\frac{\mu^n_i(x_i-\langle e,x\rangle^+u_i)+\gamma^n_i\langle e,x\rangle^+u_i}{\sqrt{n}}\,,
\end{aligned}$$

respectively. Define the operator $\hat{\mathcal{I}}^n$ by

$$\hat{\mathcal{I}}^n f(x) \,\coloneqq\, \int_{\mathbb{R}^d}\bigl(f(x+y)-f(x)\bigr)\,\nu^n_{d_1}(\mathrm{d}y)\,,$$

where

$$\nu^n_{d_1}(A) \;:=\; \Pi^n_{d_1}\Bigl(\bigl\{y \in \mathbb{R}_*\colon \bigl(\tfrac{\sqrt{n}}{\vartheta^n}\mu^n_1\rho_1 y, \dots, \tfrac{\sqrt{n}}{\vartheta^n}\mu^n_d\rho_d y\bigr) \in A\bigr\}\Bigr)\,,$$

with $\Pi^n_{d_1}(\mathrm{d}y) := \beta^n_{\mathsf{u}} F^{d_1}(\mathrm{d}y)$, and β^n_{u} as in Assumption 2.2.

Recall the generator $\widetilde{\mathcal{L}}^{z^n}_n$ of $\widetilde{\Xi}^n$ given in (4.16). The next lemma establishes the relation between the generator of the diffusion-scaled process and the operator in Definition 5.1.

Lemma 5.4. *Grant Assumptions 2.1, 2.2, and 3.2. Then,*

$$\begin{aligned}\widetilde{\mathcal{L}}^{z^n}_n \phi^n[f](\tilde{x}, h, \psi, k) \;=\;& \hat{\mathcal{A}}^n f\bigl(\tilde{x}, v^n(\tilde{x}, h, \psi, k)\bigr) + \hat{\mathcal{I}}^n f(\tilde{x}) \\ &+ \mathcal{O}\Bigl(\frac{1}{\sqrt{n}}\Bigr)\bigl(\|\tilde{x}\| + \|\tilde{q}^n\|\bigr) + \mathcal{O}(1)(1-\psi)\bigl(\|\tilde{x}\| + \|\tilde{q}^n\| + 1\bigr)\,,\end{aligned} \tag{5.17}$$

for any $f \in \mathcal{C}^\infty_c(\mathbb{R}^d)$ and $z^n \in \mathfrak{Z}^n_{\mathrm{sm}}$, where $\tilde{q}^n = n^{-1/2} q^n$, and

$$v^n(\tilde{x}, h, \psi, k) = \begin{cases} \frac{\tilde{x} - \tilde{z}^n(\sqrt{n}\tilde{x} + n\rho, h, \psi, k)}{\langle e, \tilde{x}\rangle}\,, & \text{if } \langle e, \tilde{x}\rangle > 0\,, \\ e_d\,, & \text{if } \langle e, \tilde{x}\rangle \le 0\,, \end{cases} \tag{5.18}$$

for $(\tilde{x}, h, \psi, k) \in \widetilde{\mathfrak{D}}^n$, with $\tilde{z}^n := n^{-1/2}(z^n - n\rho)$.

Proof. Note that Lemma 5.3 concerns the renewal arrival process in the diffusion-scale. Recall that $z^n_i = \sqrt{n}(\tilde{x}_i - \tilde{q}^n_i) + n\rho_i$ for $i \in \mathcal{J}$, and $\breve{x} = \sqrt{n}\tilde{x} + n\rho$. We let $q^n \equiv q^n(\sqrt{n}\tilde{x} + n\rho, z^n)$ and $z^n \equiv z^n(\sqrt{n}\tilde{x} + n\rho, h, \psi, k)$. Applying Lemma 5.3 and the Taylor expansion, it follows by the definition of $\widetilde{\mathcal{L}}^{z^n}_n$ that

$$\begin{aligned}\widetilde{\mathcal{L}}^{z^n}_n \phi^n[f](\tilde{x}, h, \psi, k) \;=\;& \sum_{i\in\mathcal{J}} \biggl[\biggl(\frac{(\lambda^n_i - n\rho_i\mu^n_i)}{\sqrt{n}} - \mu^n_i(\tilde{x}_i - \tilde{q}^n_i) - \gamma^n_i \tilde{q}^n_i\biggr)\partial_i f(\tilde{x}) \\ &+ \frac{1}{2}\biggl(\frac{\lambda^n_i c^2_{a,i}}{n} + \rho_i\mu^n_i + \frac{\tilde{x}_i + (\mu^n_i - \gamma^n_i)\tilde{q}^n_i}{\sqrt{n}}\biggr)\partial_{ii} f(\tilde{x}) + \frac{\lambda^n_i - n\rho_i\mu^n_i}{n}\sum_{j\in\mathcal{J}}\biggl(\eta^n_j(h_j) + \frac{c^2_{a,j} - 1}{2}\biggr)\partial_{ij} f(\tilde{x}) \\ &+ (1-\psi)\gamma^n_i\bigl(\phi^n[f](\tilde{x} - n^{-1/2}e_i, h) - \phi^n[f](\tilde{x}, h)\bigr)\int_{\mathbb{R}_*} q^n_i\bigl(\sqrt{n}\tilde{x} + n\rho - n\upmu^n(y-k), z^n\bigr)\tilde{F}^{d^n_1}_{\breve{x},k}(\mathrm{d}y) \\ &+ (\psi - 1)(\mu^n_i z^n_i + \gamma^n_i q^n_i)\bigl(\phi^n[f](\tilde{x} - n^{-1/2}e_i, h) - \phi^n[f](\tilde{x}, h)\bigr) \\ &- (1-\psi)\sqrt{n}\mu^n_i\rho_i\frac{\partial\phi^n[f](\tilde{x}, h)}{\partial\tilde{x}_i}\biggr] + \psi\,\hat{\mathcal{I}}^n\phi^n[f](\tilde{x}, h) + \mathcal{O}\Bigl(\frac{1}{\sqrt{n}}\Bigr)(\|\tilde{x}\| + \|\tilde{q}^n\|)\end{aligned} \tag{5.19}$$

for any $f \in \mathcal{C}^\infty_c(\mathbb{R}^d)$, where

$$\hat{\mathcal{I}}^n\phi^n[f](\tilde{x}, h) \;=\; \int_{\mathbb{R}^d}\bigl(\phi^n[f](\tilde{x} + y, h) - \phi^n[f](\tilde{x}, h)\bigr)\,\nu^n_{d_1}(\mathrm{d}y)$$

by a slight abuse of notation. It is clear that

$$\lambda^n_i - n\mu^n_i\rho_i \;=\; \mathcal{O}(\sqrt{n}) \tag{5.20}$$

by Assumption 2.1, and thus the third term in the sum on the right-hand side of (5.19) is of order $n^{-1/2}$. We next consider the fifth and sixth terms in the sum on the right-hand side of (5.19). Using the fact that

$$\phi^n[f](\tilde{x} - n^{-1/2}e_i, h) - \phi^n[f](\tilde{x}, h) \;=\; -\frac{1}{\sqrt{n}}\frac{\partial\phi^n[f](\tilde{x}, h)}{\partial\tilde{x}_i} + \mathcal{O}\biggl(\frac{1}{n}\biggr)\,,$$

and $z_i^n = \sqrt{n}\tilde{x}_i + n\rho_i - \sqrt{n}\tilde{q}_i^n$, we obtain

$$(\psi-1)(\mu_i^n z_i^n + \gamma_i^n q_i^n)\bigl(\phi^n[f](\tilde{x}-n^{-1/2}e_i,h) - \phi^n[f](\tilde{x},h)\bigr) - (1-\psi)\sqrt{n}\mu_i^n\rho_i\frac{\partial\phi^n[f](x,h)}{\partial\tilde{x}_i}$$
$$= (\psi-1)\bigl(\mu_i^n\tilde{x}_i + (\mu_i^n-\gamma_i^n)\tilde{q}_i^n\bigr)\biggl(-\frac{\partial\phi^n[f](\tilde{x},h)}{\partial x_i} + \mathcal{O}\Bigl(\frac{1}{\sqrt{n}}\Bigr)\biggr).$$

Recall the definition of $\tilde{F}^{d_1^n}_{\breve{x},k}$ in (4.13). Note that

$$\int_{\mathbb{R}_*} n\mu_i^n\rho_i(y-k)\,\tilde{F}^{d_1^n}_{\breve{x},k}(\mathrm{d}y) \;\le\; \frac{n}{\vartheta^n}\mu_i^n\rho_i\,\mathbb{E}\bigl[d_1-\vartheta^n k \,|\, d_1>\vartheta^n k\bigr] \;\in\; \mathcal{O}(\sqrt{n})\,, \tag{5.21}$$

where the second equality follows by Assumption 2.2 and (3.10). Note that $\tilde{q}_i^n \le \langle e,\tilde{x}\rangle^+$ for $i\in\mathcal{I}$ and $(\tilde{x},h,\psi,k)\in\widetilde{\mathfrak{D}}^n$. Thus, the fourth term in the sum on the right-hand side of (5.19) is bounded by $C(1-\psi)(1+\langle e,\tilde{x}\rangle^+)$ for some positive constant C. It is evident that $\phi^n[f]-f\in\mathcal{O}(n^{-1/2})$, and

$$\psi\,\hat{\mathcal{I}}^n\phi^n[f](\tilde{x},h) \;=\; \hat{\mathcal{I}}^n f(\tilde{x}) + (\psi-1)\,\hat{\mathcal{I}}^n f(\tilde{x}) + \psi\,\hat{\mathcal{I}}^n(\phi^n[f]-f)(\tilde{x},h)\,.$$

Therefore, (5.17) follows by the boundedness of $\phi^n[f]$ and (5.19). This completes the proof. □

Definition 5.2. The mean empirical measure $\hat{\zeta}_T^{z^n}\in\mathcal{P}(\mathbb{R}^d\times\mathcal{S})$ associated with $\hat{X}^n$ and a stationary Markov policy $z^n\in\mathfrak{Z}^n_{\mathrm{sm}}$ is defined by

$$\hat{\zeta}_T^{z^n}(A\times B) \;\coloneqq\; \frac{1}{T}\,\mathbb{E}\biggl[\int_0^T \mathbb{1}_{A\times B}\bigl(\hat{X}^n(s), v^n\bigl(\hat{X}^n(s),H^n(s),\Psi^n(s),K^n(s)\bigr)\bigr)\,\mathrm{d}s\biggr]$$

for any Borel sets $A\subset\mathbb{R}^d$ and $B\subset\mathcal{S}$, and with v^n as in (5.18).

The following theorem characterizes the limit points of mean empirical measures.

Theorem 5.2. *Grant the hypotheses in Theorem 3.2. Let $\{z^n\in\mathfrak{Z}^n_{\mathrm{sm}}\colon n\in\mathbb{N}\}$ be a sequence of policies satisfying (5.11). Then any limit point $\uppi\in\mathcal{P}(\mathbb{R}^d\times\mathcal{S})$ of $\hat{\zeta}_T^{z^n}$ as $(n,T)\to\infty$ lies in $\mathcal{G}$.*

Proof. It follows directly by Assumptions 2.1 and 2.2 that, for any $f\in\mathcal{C}_c^\infty(\mathbb{R}^d)$, we have

$$\hat{\mathcal{A}}^n f(\hat{x},u) + \hat{\mathcal{I}}^n f(\hat{x}) \;\to\; \mathcal{A}f(\hat{x},u) \quad \text{as } n\to\infty \tag{5.22}$$

uniformly over compact sets of $\mathbb{R}^d\times\mathcal{S}$. Thus, in view of (5.1) and (5.22), in order to prove the theorem, it is enough to show that

$$\lim_{(n,T)\to\infty}\int_{\mathbb{R}^d\times\mathcal{S}}\bigl(\hat{\mathcal{A}}^n f(\hat{x},u) + \hat{\mathcal{I}}^n f(\hat{x})\bigr)\,\hat{\zeta}_T^{z^n}(\mathrm{d}\hat{x},\mathrm{d}u) \;=\; 0 \qquad \forall\, f\in\mathcal{C}_c^\infty(\mathbb{R}^d)\,. \tag{5.23}$$

Applying (4.26) and (5.11), we obtain

$$\sup_{n>n_\circ}\,\limsup_{T\to\infty}\,\frac{1}{T}\,\mathbb{E}^{z^n}\biggl[\int_0^T|\widetilde{X}^n(s)|^m\,\mathrm{d}s\biggr] \;<\; \infty\,. \tag{5.24}$$

It follows by the same calculation as in (5.6) that, for some positive constant C_1, we have

$$\mathbb{E}^{z^n}\biggl[\int_0^T\sqrt{n}(1-\Psi^n(s))\,\mathrm{d}s\biggr] \;\le\; C_1(1+T) \quad \forall\, T\ge 0\,. \tag{5.25}$$

Using the facts that $\tilde{q}^n_i \le \langle e, x\rangle^+$ and $\Psi^n(s) \in \{0,1\}$, and Young's inequality, we obtain

$$\begin{aligned}
&\frac{1}{T}\,\mathbb{E}^{z^n}\biggl[\int_0^T n^{\frac{m-1}{4m}}\bigl(1-\Psi^n(s)\bigr)n^{\frac{1-m}{4m}}\Bigl(\|\widetilde{X}^n(s)\| + \|\tilde{q}^n\bigl(\sqrt{n}\widetilde{X}^n(s)+n\rho, z^n\bigr)\|\Bigr)\,\mathrm{d}s\biggr]\\
&\qquad\le\; \frac{1}{T}\,\mathbb{E}^{z^n}\biggl[\int_0^T n^{\frac{1}{4}}\bigl(1-\Psi^n(s)\bigr)\,\mathrm{d}s\biggr] + \frac{C_2}{T}\,\mathbb{E}^{z^n}\biggl[\int_0^T n^{\frac{1-m}{4}}|\widetilde{X}^n(s)|^m\,\mathrm{d}s\biggr]\\
&\qquad\le\; \frac{1}{Tn^{\frac{1}{4}}}C_1(1+T) + n^{\frac{1-m}{4}}\frac{C_2}{T}\,\mathbb{E}^{z^n}\biggl[\int_0^T |\widetilde{X}^n(s)|^m\,\mathrm{d}s\biggr] \;\longrightarrow\; 0 \quad \text{as } (n,T)\to\infty\,,
\end{aligned}\tag{5.26}$$

where C_2 is a positive constant. In (5.26), the second inequality follows by (5.25), and the convergence follows by (5.24) and the fact that $m > 1$. Applying Itô's formula to $\phi^n[f]$, and using Lemma 5.4 and (5.24) and (5.26), it follows by the boundedness of $\phi^n[f]$ that

$$\lim_{(n,T)\to\infty}\,\frac{1}{T}\,\mathbb{E}^{z^n}\biggl[\int_0^T \hat{\mathcal{A}}^n f\bigl(\widetilde{X}^n(s), v^n\bigl(\widetilde{\Xi}^n(s)\bigr)\bigr) + \hat{\mathcal{I}}^n f\bigl(\widetilde{X}^n(s)\bigr)\,\mathrm{d}s\biggr] \;=\; 0\,.$$

Therefore, using (4.26) again, we obtain (5.23). This completes the proof. □

Proof of (5.10). Without loss of generality, suppose $\{n_j\} \subset \mathbb{N}$ is an increasing sequence such that $z^{n_j} \in \mathfrak{Z}_{\mathrm{sm}}$ and $\sup_j \hat{J}(\hat{X}^{n_j}(0), z^{n_j}) < \infty$. Recall $\hat{\zeta}^{z^n}_T$ in Definition 5.2. There exists a subsequence of $\{n_j\}$, denoted as $\{n_l\}$, such that $T_l \to \infty$ as $l \to \infty$, and

$$\liminf_{j\to\infty}\,\hat{J}(\hat{X}^{n_j}(0), z^{n_j}) + \frac{1}{l} \;\ge\; \int_{\mathbb{R}^d\times\mathbb{U}} \mathcal{R}(\hat{x}, u)\,\hat{\zeta}^{z^{n_l}}_{T_l}(\mathrm{d}\hat{x}, \mathrm{d}u)\,. \tag{5.27}$$

Applying Lemma 5.2 and Theorem 5.2, any limit of $\hat{\zeta}^{z^{n_l}}_{T_l}$ along some subsequence is in $\mathcal{G}$. Choose any further subsequence of (T_l, n_l), also denoted by (T_l, n_l), such that $(T_l, n_l) \to \infty$ as $l \to \infty$, and $\hat{\zeta}^{z^{n_l}}_{T_l} \to \uppi \in \mathcal{G}$. By letting $l \to \infty$ and using (5.27), we obtain

$$\liminf_{j\to\infty}\,\hat{J}(\hat{X}^{n_j}(0), z^{n_j}) \;\ge\; \int_{\mathbb{R}^d\times\mathbb{U}} \mathcal{R}(\hat{x}, u)\,\uppi(\mathrm{d}\hat{x}, \mathrm{d}u) \;\ge\; \varrho_*\,.$$

This completes the proof. □

5.3.2. *The upper bound.* In this subsection, we show that

$$\limsup_{n\to\infty}\,\varrho^n\bigl(\hat{X}^n(0)\bigr) \;\le\; \varrho_*\,. \tag{5.28}$$

The following lemma concerns the convergence of mean empirical measures for the diffusion-scaled state processes under the scheduling policies in Definition 4.3. Recall $\mathfrak{A}^n_R$ in Definition 4.2 and $\hat{\zeta}^{z^n}_T$ in Definition 5.2.

Lemma 5.5. *Grant the hypotheses in Theorem 3.2. For $\epsilon > 0$, let v_ϵ be a continuous ϵ-optimal precise control, whose existence is asserted in Proposition 5.1, and $\{z^n[v^n] \colon n \in \mathbb{N}\}$ be as in Definition 4.3, and such that $R \equiv R(\epsilon)$ and v^n agrees with v_ϵ on $\mathfrak{A}^n_R$. Then, the ergodic occupation measure $\uppi_{v_\epsilon}$ of the controlled jump diffusion in (3.3) under the control v_ϵ is the unique limit point in $\mathcal{P}(\mathbb{R}^d \times \mathcal{S})$ of $\hat{\zeta}^{z^n[v^n]}_T$ as $(n,T)\to\infty$.*

Proof. Using Proposition 4.3 and Theorem 5.2, the proof of this lemma is the same as that of Lemma 7.2 in [5]. □

Proof of (5.28). Let $\upkappa = 2\lfloor m \rfloor$ with m as in (3.5), and $z^n[v^n]$ be the scheduling policy in Lemma 5.5. By Proposition 4.3, there exist $\tilde{n}_\circ \in \mathbb{N}$, and positive constants $\widetilde{C}_0$ and $\widetilde{C}_1$ such that

$$\widetilde{\mathcal{L}}^{z^n[v^n]}_n \widetilde{\mathcal{V}}^n_{\upkappa,\xi}(\tilde{x}, h, \psi, k) \;\le\; \widetilde{C}_0 - \widetilde{C}_1 \mathcal{V}_{\upkappa-1,\xi}(\tilde{x}) \qquad \forall\,(\tilde{x}, h, \psi, k) \in \widetilde{\mathfrak{D}}^n\,, \quad \forall\, n > \tilde{n}_\circ\,. \tag{5.29}$$

Recall the definition of $\widetilde{\mathcal{R}}$ in (3.5), and let $\hat{z}^n[v^n] = n^{-1/2}(z^n[v^n] - n\rho)$. Applying (4.26) and (5.29), we may select an increasing sequence T_n such that

$$\sup_{n\ge\tilde{n}_\circ}\ \sup_{T\ge T_n} \int_{\mathbb{R}^d\times\mathbb{U}} \mathcal{V}_{\kappa-1,\xi}(\hat{x})\, \hat{\zeta}_T^{z^n[v^n]}(\mathrm{d}\hat{x},\mathrm{d}u) \,<\, \infty\,.$$

This implies that $\widetilde{\mathcal{R}}\big(\hat{x} - \hat{z}^n[v](\sqrt{n}\hat{x} + n\rho)\big)$ is uniformly integrable. By Lemma 5.5, $\hat{\zeta}_T^{z^n[v^n]}$ converges in $\mathcal{P}(\mathbb{R}^d\times\mathcal{S})$ to π_{v_ϵ} as $(n,T)\to\infty$. Applying Proposition 5.1, we deduce that v_ϵ is an ϵ-optimal control for the running cost function. Since ϵ is arbitrary, (5.28) follows. □

Appendix A. Proofs of Lemma 3.1 and Proposition 3.1

Proof of Lemma 3.1. By [13, Lemma 5.1], $\hat{S}_i^n(t)$ and $\hat{R}_i^n(t)$ in (3.1) are martingales with respect to the filtration $\mathcal{F}_t^n$ in (2.9), having predictable quadratic variation processes given by

$$\langle \hat{S}_i^n\rangle(t) \,=\, \mu_i^n\int_0^t n^{-1}Z_i^n(s)\Psi^n(s)\,\mathrm{d}s \quad\text{and}\quad \langle \hat{R}_i^n\rangle(t) \,=\, \gamma_i^n\int_0^t n^{-1}Q_i^n(s)\,\mathrm{d}s\,,\quad t\ge 0\,,$$

respectively. By (2.7), we have the crude inequality

$$0 \,\le\, n^{-1}X_i^n(t) \,\le\, n^{-1}X_i^n(0) + n^{-1}A_i^n(t)\,,\quad t\ge 0\,.$$

Using the balance equation in (2.5), we see that the same inequalities hold for $n^{-1}Z_i^n$ and $n^{-1}Q_i^n$. Since $\Psi^n(s)\in\{0,1\}$, it follows by Lemma 5.8 in [30] that $\{\hat{W}_i^n\colon n\in\mathbb{N}\}$ is stochastically bounded in $(\mathbb{D}^d, J_1)$. Also, $\{\hat{L}_i^n\colon n\in\mathbb{N}\}$ is stochastically bounded in $(\mathbb{D}^d, M_1)$ by (2.4). On the other hand, it is evident that

$$\hat{Y}_i^n(t) \,\le\, C\int_0^t (1 + \|n^{-1}X^n(s)\|)\,\mathrm{d}s\,,\quad t\ge 0\,,$$

where C is some positive constant. Thus, we obtain

$$\|\hat{X}^n(t)\| \,\le\, \|\hat{X}^n(0)\| + \|\hat{W}^n(t)\| + \|\hat{L}^n(t)\| + C\int_0^t (1 + \|\hat{X}^n(s)\|)\,\mathrm{d}s \quad \forall\, t\ge 0\,. \tag{A.1}$$

Since $\hat{X}^n(0)$ is uniformly bounded, applying Lemma 5.3 in [30] and Gronwall's inequality, we deduce that $\{\hat{X}^n\colon n\in\mathbb{N}\}$ is stochastically bounded in $(\mathbb{D}^d, M_1)$. Using Lemma 5.9 in [30], we see that

$$n^{-1/2}\hat{X}^n \,=\, n^{-1}X^n - \rho \,\Rightarrow\, \mathfrak{e}_0 \quad\text{in}\quad (\mathbb{D}^d, M_1) \quad \text{as } n\to\infty\,,$$

which implies that $n^{-1}X^n \Rightarrow \mathfrak{e}_\rho$ in $(\mathbb{D}^d, M_1)$. By (2.5), and the fact $\langle e, n^{-1}Q^n\rangle = (\langle e, n^{-1}X^n\rangle - 1)^+ \Rightarrow \mathfrak{e}_0$, we have $n^{-1}Q^n\Rightarrow\mathfrak{e}_0$, and thus $n^{-1}Z^n\Rightarrow\mathfrak{e}_\rho$. This completes the proof. □

To prove Proposition 3.1, we first consider a modified process. Let $\check{X}^n = (\check{X}_1^n,\dotsc,\check{X}_d^n)'$ be the d-dimensional process defined by

$$\begin{aligned}\check{X}_i^n(t) \,:=\, \hat{X}^n(0) + \ell_i^n t + \hat{W}_i^n(t) + \hat{L}_i^n(t) - \int_0^t \mu_i^n\big(\check{X}_i^n(s) - \langle e, \check{X}^n(s)\rangle^+ U_i^n(s)\big)\,\mathrm{d}s \\ - \int_0^t \gamma_i^n \langle e, \check{X}^n(s)\rangle^+ U_i^n(s)\,\mathrm{d}s\,,\quad \text{for } i\in\mathcal{I}\,.\end{aligned} \tag{A.2}$$

Lemma A.1. *As $n\to\infty$, $\check{X}^n$ and $\hat{X}^n$ are asymptotically equivalent, that is, if either of them converges in distribution as $n\to\infty$, then so does the other, and both of them have the same limit.*

Proof. Let $K = K(\epsilon_1) > 0$ be the constant satisfying $\mathbb{P}(\|\hat{X}^n\|_T > K) < \epsilon_1$ for $T > 0$ and any $\epsilon_1 > 0$, where $\|\hat{X}^n\|_T := \sup_{0\le t\le T}\|\hat{X}^n(t)\|$. Since $\hat{U}^n(s) \in \mathcal{S}$ for $s \ge 0$, on the event $\{\|\hat{X}^n\|_T \le K\}$, we obtain

$$\begin{aligned}\|\check{X}^n(t) - \hat{X}^n(t)\| \;&\le\; C_1\int_0^t \|\hat{X}^n(s)\|\bigl(1-\Psi^n(s)\bigr)\,\mathrm{d}s + C_2\int_0^t \|\check{X}^n(s) - \hat{X}^n(s)\|\,\mathrm{d}s \\ &\le\; C_1 K C^n_{\mathsf{d}}(t) + C_2\int_0^t \|\check{X}^n(s) - \hat{X}^n(s)\|\,\mathrm{d}s \quad \forall\, t\in[0,T]\,,\end{aligned}$$

where C_1 and C_2 are some positive constants. Then, by Gronwall's inequality, on the event $\{\|\hat{X}^n\|_T \le K\}$, we have

$$\|\check{X}^n(t) - \hat{X}^n(t)\| \;\le\; C_1 K C^n_{\mathsf{d}}(t)\mathrm{e}^{C_2 T} \quad \forall\, t\in[0,T]\,.$$

Thus, applying Lemma 2.2 in [13], we deduce that for any $\epsilon_2 > 0$, there exist $\epsilon_3 > 0$ and $n_\circ = n_\circ(\epsilon_1,\epsilon_2,\epsilon_3,T)$ such that

$$\|\check{X}^n - \hat{X}^n\|_T \;\le\; \epsilon_2$$

on the event $\{\|\hat{X}^n\|_T \le K\} \cap \{\|C^n_{\mathsf{d}}\|_T \le \epsilon_3\}$, for all $n \ge n_\circ$, which implies that

$$\mathbb{P}(\|\check{X}^n - \hat{X}^n\|_T > \epsilon_2) < \epsilon_1\,, \quad \forall\, n\ge n_\circ\,.$$

As a consequence, $\|\check{X}^n - \hat{X}^n\|_T \Rightarrow 0$, as $n\to\infty$, and this completes the proof. □

Proof of Proposition 3.1. We first prove (i). Define the processes

$$\uptau^n_{1,i}(t) \;:=\; \frac{\mu^n_i}{n}\int_0^t Z^n(s)\Psi^n(s)\,\mathrm{d}s\,, \quad \uptau^n_{2,i}(t) \;:=\; \frac{\gamma^n_i}{n}\int_0^t Q^n(s)\,\mathrm{d}s\,,$$

$\tilde{S}^n_i(t) := n^{-1/2}(S^n(nt) - nt)$, and $\tilde{R}^n_i(t) := n^{-1/2}(R^n(nt) - nt)$, for $i\in\mathcal{J}$. Then, since $\Psi^n(s)\in\{0,1\}$ for $s \ge 0$, applying Lemma 3.1 and Lemma 2.2 in [13], we have

$$\uptau^n_{1,i}(\cdot) \;=\; \mu^n_i\int_0^{\cdot}(n^{-1}Z^n_i(s) - \rho_i)\Psi^n(s)\,\mathrm{d}s + \mu^n_i\int_0^{\cdot}\rho_i\Psi^n(s)\,\mathrm{d}s \;\Rightarrow\; \lambda_i\mathfrak{e}(\cdot)\,.$$

in $(\mathbb{D}, M_1)$, as $n\to\infty$, and that $\uptau^n_{2,i}$ weakly converges to the zero process. Since $\{A^n_i, S^n_i, R^n_i, \Psi^n : i\in\mathcal{J}, n\in\mathbb{N}\}$ are independent processes, and $\tau^n_{1,i}$ and $\tau^n_{2,i}$ converge to deterministic functions, we have joint weak convergence of $(\hat{A}^n, \hat{S}^n, \hat{R}^n, \hat{L}^n, \uptau^n_1, \uptau^n_2)$, where $\uptau^n_1 := (\uptau^n_{1,1},\dots,\uptau^n_{1,d})'$, and $\uptau^n_2$ is defined analogously. On the other hand, since the second moment of A^n is finite, it follows that $\hat{A}^n$ converges weakly to a d-dimensional Wiener process with mean 0 and covariance matrix $\operatorname{diag}\bigl(\sqrt{\lambda_1 c^2_{a,1}},\dots,\sqrt{\lambda_d c^2_{a,d}}\bigr)$ (see, e.g., [31]). Therefore, by the FCLT for the Poisson processes $\tilde{S}^n$ and $\tilde{R}^n$, and using the random time change lemma in [21, Page 151], we obtain (i).

Using (A.1) and Proposition 3.1 (i), the proof of (ii) is same as the proof of [1, Lemma 4 (iii)].

To prove (iii), we first show any limit of $\check{X}^n$ in (A.2) satisfies (3.3). Following an argument similar to the proof of Lemma 5.2 in [13], one can easily show that the d-dimensional integral mapping $x = \Lambda(y,u)\colon \mathbb{D}^d\times\mathbb{D}^d \to \mathbb{D}^d$ defined by

$$x(t) = y(t) + \int_0^t h\bigl(x(s), u(s)\bigr)\,\mathrm{d}s$$

is continuous in $(\mathbb{D}^d, M_1)$, provided that the function $h\colon \mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}^d$ is Lipschitz continuous in each coordinate. Since

$$\check{X}^n \;=\; \Lambda(\hat{X}^n(0) + \hat{W}^n + \hat{L}^n, U^n)\,,$$

then, by the tightness of U^n and the continuous mapping theorem, any limit of $\check{X}^n$ satisfies (A.2), and the same result holds for $\hat{X}^n$ by Lemma A.1.

Recall the definition of $\breve{\tau}^n$ in (2.8). It is evident that

$$\begin{aligned}\hat{L}^n_i(t+r)-\hat{L}^n_i(t) \;=\; &\hat{L}^n_i(\breve{\tau}^n(t)+r)-\hat{L}^n_i\bigl(\breve{\tau}^n(t)\bigr)\\ &+\hat{L}^n_i(t+r)-\hat{L}^n_i(\breve{\tau}^n(t)+r)+\hat{L}^n_i\bigl(\breve{\tau}^n(t)\bigr)-\hat{L}^n_i(t)\,.\end{aligned} \tag{A.3}$$

for all $t,r\geq 0$ and $i\in\mathcal{I}$. By Assumption 2.2, we have $\breve{\tau}^n(t)\Rightarrow t$ as $n\to\infty$, for $t\geq 0$. Then, by the random time change lemma in [21, Page 151], we deduce that the last four terms on the right-hand side of (A.3) converge to 0 in distribution. It follows by Proposition 3.1 (i) and (A.3) that

$$\hat{L}^n(\breve{\tau}^n(t)+r)-\hat{L}^n\bigl(\breve{\tau}^n(t)\bigr)\;\Rightarrow\;\lambda L_{t+r}-\lambda L_t\quad \text{in } \mathbb{R}^d\,.$$

Repeating the same argument we establish convergence of $\hat{S}^n$ and $\hat{R}^n$. Proving that U is non-anticipative follows exactly as in [1, Lemma 6]. This completes the proof of (iii). ☐

Appendix B. Proofs of Lemmas 4.1 and 5.2

In this section, we construct two functions, which are used to show the ergodicity of $\widetilde{\Xi}^n$. We provide two lemmas concerning the properties of these functions, respectively. The proofs of Lemmas 4.1 and 5.2 are given at the end of this section.

Definition B.1. For $z^n\in\mathfrak{Z}^n_{\mathrm{sm}}$, define the operator $\mathcal{L}^{z^n}_n\colon\mathcal{C}_b(\mathbb{R}^d\times\mathbb{R}^d)\to\mathcal{C}_b(\mathbb{R}^d\times\mathbb{R}^d)$ by

$$\begin{aligned}\mathcal{L}^{z^n}_n f(\breve{x},h)\;\coloneqq\;&\sum_{i\in\mathcal{I}}\frac{\partial f(\breve{x},h)}{\partial h_i}+\sum_{i\in\mathcal{I}}r^n_i(h_i)\bigl(f(\breve{x}+e_i,h-h_i\,e_i)-f(\breve{x},h)\bigr)\\ &+\sum_{i\in\mathcal{I}}\mu^n_i z^n_i\bigl(f(\breve{x}-e_i,h)-f(\breve{x},h)\bigr)+\sum_{i\in\mathcal{I}}\gamma^n_i q^n_i\bigl(f(\breve{x}-e_i,h)-f(\breve{x},h)\bigr)\end{aligned} \tag{B.1}$$

for $f\in\mathcal{C}_b(\mathbb{R}^d\times\mathbb{R}^d)$ and any $(\breve{x},h)\in\mathbb{R}^d_+\times\mathbb{R}^d_+$, with $q^n\coloneqq\breve{x}-z^n$.

Note that if $d^n_1\equiv 0$ for all n, the queueing system has no interruptions. In this situation, under a Markov scheduling policy, the (infinitesimal) generator of (X^n,H^n) takes the form of (B.1). Recall the scheduling policies $\breve{z}^n$ in Definition 4.1, and $\bar{x}=\breve{x}-n\rho$ in Definition 4.2. We define the sets

$$\breve{\mathcal{K}}_n(\breve{x})\;\coloneqq\;\biggl\{i\in\mathcal{I}_0\colon \breve{x}_i\;\geq\;\frac{n\rho_i}{\sum_{j\in\mathcal{I}_0}\rho_j}\biggr\}\;=\;\biggl\{i\in\mathcal{I}_0\colon \bar{x}_i\;\geq\;\frac{n\rho_i\sum_{j\in\mathcal{I}\setminus\mathcal{I}_0}\rho_j}{\sum_{j\in\mathcal{I}_0}\rho_j}\biggr\}.$$

We have the following lemma.

Lemma B.1. *Grant Assumptions 2.1, 2.2, and 3.2. For any even integer $\kappa\geq 2$, there exist a positive vector $\xi\in\mathbb{R}^d_+$, $\breve{n}\in\mathbb{N}$, and positive constants $\breve{C}_0$ and $\breve{C}_1$, such that the functions f_n, $n\in\mathbb{N}$, defined by*

$$f_n(\breve{x},h)\;\coloneqq\;\sum_{i\in\mathcal{I}}\xi_i|\bar{x}_i|^\kappa+\sum_{i\in\mathcal{I}}\eta^n_i(h_i)\xi_i\bigl(|\bar{x}_i+1|^\kappa-|\bar{x}_i|^\kappa\bigr)\quad\forall\,(\breve{x},h)\in\mathbb{R}^d_+\times\mathbb{R}^d_+\,, \tag{B.2}$$

with η^n_i as defined in (4.3), satisfy

$$\begin{aligned}\mathcal{L}^{\breve{z}^n}_n f_n(\breve{x},h)\;\leq\;&\breve{C}_0 n^{\kappa/2}-\breve{C}_1\sum_{i\in\mathcal{I}\setminus\breve{\mathcal{K}}_n(\breve{x})}\xi_i|\bar{x}_i|^\kappa-\breve{C}_1\sum_{i\in\breve{\mathcal{K}}_n(\breve{x})}\bigl(\mu^n_i(\breve{z}^n_i-n\rho_i)+\gamma^n_i\breve{q}^n_i\bigr)|\bar{x}_i|^{\kappa-1}\\ &+\sum_{i\in\mathcal{I}}\bigl(\mathcal{O}(\sqrt{n})\mathcal{O}\bigl(|\bar{x}_i|^{\kappa-1}\bigr)+\mathcal{O}(n)\mathcal{O}\bigl(|\bar{x}_i|^{\kappa-2}\bigr)\bigr)\end{aligned} \tag{B.3}$$

for all $n\geq\breve{n}$ and $(\breve{x},h)\in\mathbb{R}^d_+\times\mathbb{R}^d_+$.

Proof. Using the estimate

$$(a \pm 1)^m - a^\kappa \;=\; \pm \kappa a^{\kappa-1} + \mathcal{O}(a^{\kappa-2}) \qquad \forall\, a \in \mathbb{R}\,, \tag{B.4}$$

an easy calculation shows that

$$\begin{aligned} \mathcal{L}_n^{\check{z}^n} f_n(\check{x},h) \;=\; & \sum_{i\in\mathcal{I}} \dot{\eta}_i^n(h_i)\xi_i\bigl(|\bar{x}_i+1|^\kappa - |\bar{x}_i|^\kappa\bigr) + \sum_{i\in\mathcal{I}} r_i^n(h_i)\eta_i^n(0)\xi_i\bigl((\bar{x}_i+2)^\kappa - (\bar{x}_i+1)^\kappa\bigr) \\ & - \sum_{i\in\mathcal{I}} r_i^n(h_i)\eta_i^n(h_i)\xi_i\bigl(|\bar{x}_i+1|^\kappa - |\bar{x}_i|^\kappa\bigr) \\ & + \sum_{i\in\mathcal{I}} \eta_i^n(h_i)(\mu_i^n\check{z}_i^n + \gamma_i^n\check{q}_i^n)\mathcal{O}(|\bar{x}_i|^{\kappa-2}) + \sum_{i\in\mathcal{I}} r_i^n(h_i)\xi_i(|\bar{x}_i+1|^\kappa - |\bar{x}_i|^\kappa) \\ & + \sum_{i\in\mathcal{I}} (\mu_i^n\check{z}_i^n + \gamma_i^n\check{q}_i^n)\xi_i(|\bar{x}_i-1|^\kappa - |\bar{x}_i|^\kappa)\,, \end{aligned} \tag{B.5}$$

where for the fourth term on the right-hand side we also used the fact that

$$\bigl(|\bar{x}_i|^\kappa - |\bar{x}_i-1|^\kappa\bigr) - \bigl(|\bar{x}_i+1|^\kappa - |\bar{x}_i|^\kappa\bigr) \;=\; \mathcal{O}(|\bar{x}_i|^{\kappa-2})\,.$$

It is clear that $\eta_i^n(0) = 0$, since $F_i(0) = 0$ and $\mathbb{E}[G_i] = 1$. On the other hand, $\eta_i^n(t)$ is bounded for all $n \in \mathbb{N}$ and $t \ge 0$ by Assumption 3.2. Thus, applying (4.4), (B.4), and (B.5), it follows that

$$\begin{aligned} \mathcal{L}_n^{\check{z}^n} f_n(\check{x},h) \;=\; & \sum_{i\in\mathcal{I}} \bigl[\xi_i(\lambda_i^n - \mu_i^n\check{z}_i^n - \gamma_i^n\check{q}_i^n)\bigl(\kappa(\bar{x}_i)^{\kappa-1} + \mathcal{O}(|\bar{x}_i|^{\kappa-2})\bigr) \\ & \qquad + \eta_i^n(h_i)(\mu_i^n\check{z}_i^n + \gamma_i^n\check{q}_i^n)\mathcal{O}(|\bar{x}_i|^{\kappa-2})\bigr]\,. \end{aligned} \tag{B.6}$$

Since $\eta_i^n(h_i)$ is uniformly bounded, and $\check{z}_i^n, \check{q}_i^n \le \bar{x}_i + n\rho_i$, it follows that the last term in (B.6) is equal to $\mathcal{O}(n)\mathcal{O}(|\bar{x}_i|^{\kappa-2}) + \mathcal{O}(|\bar{x}_i|^{\kappa-1})$. Note that for $i \in \mathcal{I}\setminus\mathcal{I}_0$, $\check{z}_i^n$ is equivalent to the static priority scheduling policy. Note also, that

$$\bar{x}_i \;\ge\; \check{z}_i^n - n\rho_i \;\ge\; \frac{n\rho_i\sum_{j\in\mathcal{I}\setminus\mathcal{I}_0}\rho_j}{\sum_{j\in\mathcal{I}_0}\rho_j} \;>\; 0 \qquad \forall\, i \in \tilde{\mathcal{K}}_n(\check{x})\,, \tag{B.7}$$

and for $i \in \mathcal{I}_0 \setminus \tilde{\mathcal{K}}_n(\check{x})$, we have $\check{z}_i^n - n\rho_i = \bar{x}_i$ and $\check{q}_i^n = 0$. By using (B.6), and the identity in (5.20), we obtain

$$\begin{aligned} \mathcal{L}_n^{\check{z}^n} f_n(\check{x},h) \;\le\; & \sum_{i\in\mathcal{I}\setminus\mathcal{I}_0} \xi_i\bigl(-\mu_i^n\bar{x}_i + (\mu_i^n - \gamma_i^n)\check{q}_i^n\bigr)m(\bar{x}_i)^{\kappa-1} \\ & - \sum_{i\in\tilde{\mathcal{K}}_n(\check{x})} \xi_i\bigl(\mu_i^n(\check{z}_i^n - n\rho_i) + \gamma_i^n\check{q}_i^n\bigr)|\bar{x}_i|^{\kappa-1} \\ & - \sum_{i\in\mathcal{I}_0\setminus\tilde{\mathcal{K}}_n(\check{x})} \xi_i\mu_i^n|\bar{x}_i|^\kappa + \sum_{i\in\mathcal{I}}\bigl(\mathcal{O}(\sqrt{n})\mathcal{O}(|\bar{x}_i|^{\kappa-1}) + \mathcal{O}(n)\mathcal{O}(|\bar{x}_i|^{\kappa-2})\bigr)\,. \end{aligned} \tag{B.8}$$

Let $\check{c}_1 := \sup_{i,n}\{\gamma_i^n, \mu_i^n\}$, and $\check{c}_2$ be some constant such that $\inf\{\mu_i^n, \gamma_j^n \colon i \in \mathcal{I}, j \in \mathcal{I}\setminus\mathcal{I}_0, n \in \mathbb{N}\} \ge \check{c}_2 > 0$. We select a positive vector $\xi \in \mathbb{R}_+^d$ such that $\xi_1 := 1$, $\xi_i := \frac{\kappa_1^m}{d^\kappa}\min_{i'\le i-1}\xi_{i'}$, $i \ge 2$, with $\kappa_1 := \frac{\check{c}_1}{8\check{c}_2}$. Compared with [4, Lemma 5.1], the important difference here is that, for $i \in \mathcal{I}\setminus\mathcal{I}_0$, we have

$$\check{q}_i^n \;=\; \Biggl(\check{x}_i - \Biggl(n - \sum_{j\in\tilde{\mathcal{K}}_n(\check{x})} \check{z}_j^n - \sum_{j\in\mathcal{I}_0\setminus\tilde{\mathcal{K}}_n(\check{x})} x_j - \sum_{j=|\mathcal{I}_0|+1}^{i-1} x_j\Biggr)^+\Biggr)^+\,.$$

Repeating the argument in the proof of [4, Lemma 5.1], it follows by (B.8) that

$$\begin{aligned}\mathcal{L}_n^{\check{z}^n} f_n(\breve{x},h) \;\le\; c_3 n^{\kappa/2} - c_4 \sum_{i\in\mathcal{I}\setminus\tilde{\mathcal{K}}_n(\breve{x})} \xi_i|\bar{x}_i|^{\kappa} - c_5 \sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})} \xi_i\bigl(\mu_i^n(\check{z}_i^n - n\rho_i) + \gamma_i^n \check{q}_i^n\bigr)|\bar{x}_i|^{\kappa-1} \\ + \frac{c_5}{2}\sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})} \xi_i\mu_i^n\bigl(\check{z}_i^n - n\rho_i\bigr)^{\kappa} + \sum_{i\in\mathcal{I}}\bigl(\mathcal{O}(\sqrt{n})\mathcal{O}(|\bar{x}_i|^{\kappa-1}) + \mathcal{O}(n)\mathcal{O}(|\bar{x}_i|^{\kappa-2})\bigr)\end{aligned} \tag{B.9}$$

for some positive constants c_3, c_4 and c_5. Therefore, (B.3) follows by (B.7) and (B.9), and this completes the proof. □

Let

$$\tilde{g}_n(\breve{x},h,\psi,k) \;:=\; \frac{\psi+\alpha^n(k)}{\vartheta^n}\sum_{i\in\mathcal{I}}\mu_i^n\xi_i\Bigl(\tilde{g}_{n,i}(\breve{x}_i) + \eta_i^n(h_i)\bigl(\tilde{g}_{n,i}(\breve{x}_i+1) - \tilde{g}_{n,i}(\breve{x}_i)\bigr)\Bigr) \tag{B.10}$$

for $(\breve{x},h,\psi,k)\in\mathfrak{D}$, where $\tilde{g}_{n,i}(\breve{x}_i) := -|\bar{x}_i|^{\kappa}$ for $i\in\mathcal{I}\setminus\mathcal{I}_0$, and

$$\tilde{g}_{n,i}(\breve{x}_i) \;:=\; \begin{cases} -|\bar{x}_i|^{\kappa}\,, & \text{if } \bar{x}_i \;<\; \dfrac{n\rho_i\sum_{j\in\mathcal{I}\setminus\mathcal{I}_0}\rho_j}{\sum_{j\in\mathcal{I}_0}\rho_j}\,,\\ -\dfrac{n\rho_i\sum_{j\in\mathcal{I}\setminus\mathcal{I}_0}\rho_j}{\sum_{j\in\mathcal{I}_0}\rho_j}|\bar{x}_i|^{\kappa-1}\,, & \text{if } \bar{x}_i \;\ge\; \dfrac{n\rho_i\sum_{j\in\mathcal{I}\setminus\mathcal{I}_0}\rho_j}{\sum_{j\in\mathcal{I}_0}\rho_j}\,.\end{cases} \quad \forall\, i\in\mathcal{I}_0\,.$$

Recall $\overline{\mathcal{L}}_{n,\psi}^{z^n}$ in (4.13). We also define

$$\overline{q}_i^{n,k}(\breve{x},z^n) \;=\; \int_{\mathbb{R}_*} q_i^n\bigl(\breve{x} - n\upmu^n(y-k), z^n\bigr)\,\tilde{F}_{\breve{x},k}^{d_1^n}(\mathrm{d}y)\,.$$

Lemma B.2. *Grant Assumptions 2.1, 2.2, and 3.2, and let $\xi\in\mathbb{R}_+^d$ be as in (B.2). Then, for any even integer $\kappa\ge 2$ and any $\varepsilon>0$, there exist a positive constant $\overline{C}$, and $\bar{n}\in\mathbb{N}$, such that*

$$\begin{aligned}\overline{\mathcal{L}}_{n,\psi}^{z^n}\,\tilde{g}_n(\breve{x},h,\psi,k) \;\le\; \overline{C}n^{\kappa/2} + \varepsilon\sum_{i\in\mathcal{I}\setminus\tilde{\mathcal{K}}_n(\breve{x})}|\bar{x}_i|^{\kappa} + \sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})}\mathcal{O}\bigl(|\bar{x}_i|^{\kappa-1}\bigr) \\ + \frac{1}{\sqrt{n}}\sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})}\bigl(\psi\mu_i^n(|z_i^n - n\rho_i|) + \psi\gamma_i^n q_i^n + (1-\psi)\gamma_i^n\overline{q}_i^{n,k}\bigr)\mathcal{O}\bigl(|\bar{x}_i|^{\kappa-1}\bigr)\end{aligned} \tag{B.11}$$

for any $z^n\in\mathfrak{Z}_{\mathrm{sm}}^n$, and all $(\breve{x},h,\psi,k)\in\mathfrak{D}$ and $n>\bar{n}$.

Proof. It is straightforward to verify that

$$\begin{aligned}|g_{n,i}(\breve{x}_i\pm 1) - g_{n,i}(\breve{x}_i)| \;=\; \mathcal{O}(|\bar{x}_i|^{\kappa-1})\,,\\ |\bigl(g_{n,i}(\breve{x}_i) - g_{n,i}(\breve{x}_i-1)\bigr) - \bigl(g_{n,i}(\breve{x}_i+1) - g_{n,i}(\breve{x}_i)\bigr)| \;=\; \mathcal{O}(|\bar{x}_i|^{\kappa-2})\,,\end{aligned} \tag{B.12}$$

for $i\in\mathcal{I}$. Repeating the calculation in (B.5) and (B.6), and applying (B.4) and (B.12), we have

$$\begin{aligned}\overline{\mathcal{L}}_{n,\psi}^{z^n}\tilde{g}_n(\breve{x},h,\psi,k) \;\le\; \frac{\psi+\alpha^n(k)}{\vartheta^n} \\ \Biggl[\sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})}\mu_i^n\xi_i\Bigl[\bigl(|\lambda_i^n - n\mu_i^n\rho_i| + \psi\mu_i^n|z_i^n - n\rho_i| + \psi\gamma_i^n q_i^n + (1-\psi)\gamma_i^n\overline{q}_i^{n,k}\bigr)\mathcal{O}(|\bar{x}_i|^{\kappa-1}) \\ + \eta_i^n(h_i)\bigl(\psi\mu_i^n z_i^n + \psi\gamma_i^n q_i^n + (1-\psi)\gamma_i^n\overline{q}_i^{n,k}\bigr)\mathcal{O}(|\bar{x}_i|^{\kappa-2})\Bigr] \\ + \sum_{i\in\mathcal{I}\setminus\tilde{\mathcal{K}}_n(\breve{x})}\mu_i^n\xi_i\Bigl[\bigl(\lambda_i^n + (1-\psi)n\mu_i^n\rho_i \\ + \bigl(1+\eta_i^n(h_i)\bigr)(\psi\mu_i^n z_i^n + \psi\gamma_i^n q_i^n + (1-\psi)\gamma_i^n\overline{q}_i^{n,k}\bigr)\mathcal{O}\bigl(|\bar{x}_i|^{\kappa-1}\bigr)\Bigr]\Biggr]\,.\end{aligned} \tag{B.13}$$

Note that $\overline{q}_i^{n,k} \le c(1+\langle e,\bar{x}\rangle^+)$ for some positive constant c, by (5.21). Since $z_i^n, q_i^n \le \bar{x}_i + n\rho_i$, $(\vartheta^n)^{-1}$ is of order $n^{-1/2}$ by Assumption 2.2, and η_i^n and α^n are bounded, it follows by (5.20) and (B.13) that

$$\overline{\mathcal{L}}^{z^n}_{n,\psi}\tilde{g}_n(\breve{x},h,\psi,k) \le \sum_{i\in\mathcal{I}\setminus\tilde{\mathcal{K}}_n(\breve{x})} \frac{1}{\sqrt{n}}\big(\mathcal{O}(n)\mathcal{O}(|\bar{x}_i|^{\kappa-1}) + \mathcal{O}(|\bar{x}_i|^{\kappa})\big) + \sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})} \mathcal{O}(\sqrt{n})\mathcal{O}(|\bar{x}_i|^{\kappa-2})$$
$$+ \sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})} \frac{1}{\sqrt{n}}\big(\mathcal{O}(\sqrt{n}) + \psi\mu_i^n|z_i^n - n\rho_i| + \psi\gamma_i^n q_i^n + (1-\psi)\gamma_i^n\overline{q}_i^{n,k}\big)\mathcal{O}(|\bar{x}_i|^{\kappa-1})\,.$$

Thus, applying Young's inequality, we obtain (B.11), and this completes the proof. □

Proof of Lemma 4.1. We define the function $\tilde{f}_n \in \mathcal{C}(\mathbb{R}^d\times\mathbb{R}^d_+\times\{0,1\}\times\mathbb{R}_+)$ by

$$\tilde{f}_n(\breve{x},h,\psi,k) \coloneqq f_n(\breve{x},h) + \tilde{g}_n(\breve{x},h,\psi,k)\,,$$

with f_n and $\tilde{g}_n$ in (B.2) and (B.10), respectively. Recall $\widetilde{\mathcal{V}}^n_{\kappa,\xi}$ in (4.17). With $\xi\in\mathbb{R}^d_+$ as in (B.2), we have

$$n^{\kappa/2}\widetilde{\mathcal{V}}^n_{\kappa,\xi}(\tilde{x}^n(\breve{x}),h,\psi,k) = \tilde{f}_n(\breve{x},h,\psi,k) \qquad \forall\,(\breve{x},h,\psi,k)\in\mathfrak{D}\,.$$

Hence, to prove (4.18), it suffices to show that

$$\breve{\mathcal{L}}^{\breve{z}^n}_n \tilde{f}_n(\breve{x},h,\psi,k) \le \widetilde{C}_0 n^{\kappa/2} - \widetilde{C}_1 \sum_{i\in\mathcal{I}\setminus\tilde{\mathcal{K}}_n(x)} \xi_i|\bar{x}_i|^{\kappa} - \widetilde{C}_1\sqrt{n}\sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})}\xi_i|\bar{x}_i|^{\kappa-1} \qquad \forall\, n>\breve{n}\,, \tag{B.14}$$

and all $(\breve{x},h,\psi,k)\in\mathfrak{D}$, where the generator $\breve{\mathcal{L}}^{\breve{z}^n}_n$ is given in (4.12). It is clear that $\mathcal{Q}_{n,\psi}f_n(\breve{x},h)=0$. Since $(\vartheta^n)^{-1}$ is of order $n^{-1/2}$, it follows by (4.10) and (4.15) that

$$\begin{aligned}\mathcal{Q}_{n,0}\tilde{g}_n(\breve{x},h,0,k) \le& \sum_{i\in\mathcal{I}\setminus\tilde{\mathcal{K}}_n(\breve{x})} -\mu_i^n\xi_i|\bar{x}_i|^{\kappa} + \sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})} -\mu_i^n\xi_i\frac{n\rho_i\sum_{j\in\mathcal{I}\setminus\mathcal{I}_0}\rho_j}{\sum_{j\in\mathcal{I}_0}\rho_j}|\bar{x}_i|^{\kappa-1}\\ &+\epsilon_n\sum_{i\in\mathcal{I}\setminus\tilde{\mathcal{K}}_n(\breve{x})}\mathcal{O}(|\bar{x}_i|^{\kappa}) + \sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})}\mathcal{O}(\sqrt{n})\mathcal{O}(|\bar{x}_i|^{\kappa-1})\,,\end{aligned}\tag{B.15}$$

where C is some positive constant and $\epsilon_n\to0$ as $n\to\infty$. Since all the moments of d_1 are finite by (3.10) and $(a+z)^{\kappa}-a^{\kappa} = \mathcal{O}(z)\mathcal{O}(a^{\kappa-1}) + \mathcal{O}(z^2)\mathcal{O}(a^{\kappa-2}) + \cdots + \mathcal{O}(z^{\kappa})$ for any $a,z\in\mathbb{R}$, it is easy to verify that

$$\mathcal{I}_{n,1}\hat{f}_n(\breve{x},h,1,0) = \sum_{i\in\mathcal{I}}\sum_{j=1}^{\kappa}\mathcal{O}(n^{j/2})\mathcal{O}(|\bar{x}_i|^{\kappa-j})\,,\tag{B.16}$$

using also the fact that

$$\beta^n_{\mathsf{u}}\int_{R_*}\left(\frac{n}{\vartheta^n}\mu_i^n\rho_i z\right)^j F^{d_1}(\mathrm{d}z) = \beta^n_{\mathsf{u}}\left(\frac{n}{\vartheta^n}\right)^j(\mu_i^n\rho_i)^j\,\mathbb{E}\big[(d_1)^j\big] = \mathcal{O}(n^{j/2}) \quad \forall\, j>0\,,$$

which follows by by Assumptions 2.1 and 2.2 and (3.10). Then, for $\psi=1$, it follows by (B.16) and Young's inequality that

$$\begin{aligned}\breve{\mathcal{L}}^{\breve{z}^n}_n\tilde{f}_n(\breve{x},h,1,0) \le& \mathcal{L}^{\breve{z}^n}_n f_n(\breve{x},h) + \overline{\mathcal{L}}^{\breve{z}^n}_{n,1}\tilde{g}_n(\breve{x},h,1,0)\\ &+ Cn^{\kappa/2} + \epsilon_n\sum_{i\in\mathcal{I}\setminus\tilde{\mathcal{K}}_n(\breve{x})}\mathcal{O}(|\bar{x}_i|^{\kappa}) + \sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})}\mathcal{O}(\sqrt{n})\mathcal{O}(|\bar{x}_i|^{\kappa-1})\,.\end{aligned}\tag{B.17}$$

Note that the last two terms in (B.3) and the last term in (B.11) are of smaller order than the second and third terms on the right-hand side of (B.3), respectively. Thus, applying Lemmas B.1

and B.2, and using (B.17), we obtain

$$n^{-\kappa/2}\breve{\mathcal{L}}_n^{\breve{z}^n}\tilde{f}_n(\breve{x},h,1,0) \le \widetilde{C}_0-\widetilde{C}_1\sum_{i\in\mathcal{I}\setminus\tilde{\mathcal{K}}_n(\tilde{x})}|\bar{x}_i|^{\kappa}-\widetilde{C}_1\sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})}n^{-1/2}\big(\mu_i^n(\breve{z}_i^n-n\rho_i)+\gamma_i^n\breve{q}_i^n\big)|\tilde{x}_i|^{\kappa-1} \quad (B.18)$$

for all large enough n, where $\tilde{x}$ is defined in Definition 4.2. Since $\breve{q}_i^n\ge 0$ and $\breve{z}_i^n-n\rho_i>0$ for $i\in\tilde{\mathcal{K}}_n(\breve{x})$, then by using (B.7) and (B.18), we see that (B.14) holds when $y=1$.

For $\psi=0$, using (B.15), Young's inequality, and the fact that for $i\in\tilde{\mathcal{K}}_n(\breve{x})$, $\bar{x}_i>0$, we obtain

$$\begin{aligned}\breve{\mathcal{L}}_n^{\breve{z}^n}\tilde{f}_n(\breve{x},h,0,k) &\le \sum_{i\in\mathcal{I}}\mathcal{O}(\sqrt{n})\mathcal{O}(|\bar{x}_i|^{\kappa-1})+\sum_{i\in\mathcal{I}}\mathcal{O}(n)\mathcal{O}(|\bar{x}_i|^{\kappa-2})+Cn^{\kappa/2}\\&\quad+(\epsilon+\epsilon_n)\sum_{i\in\mathcal{I}\setminus\tilde{\mathcal{K}}_n(\breve{x})}\xi_i|\bar{x}_i|^{\kappa}+\sum_{i\in\mathcal{I}\setminus\tilde{\mathcal{K}}_n(\breve{x})}\Big(-\mu_i^n\xi_i|\bar{x}_i|^{\kappa}+\gamma_i^n\xi_i\overline{q}_i^{n,k}\big(-\kappa(\bar{x}_i)^{\kappa-1}+\mathcal{O}(|\bar{x}_i|^{\kappa-2})\big)\Big)\\&\qquad+\sum_{i\in\tilde{\mathcal{K}}_n(\breve{x})}-\frac{n\rho_i\sum_{j\in\mathcal{I}\setminus\mathcal{I}_0}\rho_j}{\sum_{j\in\mathcal{I}_0}\rho_j}\mu_i^n\xi_i|\bar{x}_i|^{\kappa-1}+\overline{\mathcal{L}}_{n,0}^{\breve{z}^n}\tilde{g}_n(\breve{x},h,0,k)\end{aligned}$$

for some positive constant C and sufficiently small $\epsilon>0$. We proceed by invoking the argument in the proof of [4, Lemma 5.1]. The important difference here is that

$$\breve{q}_i^n\big(\breve{x}-n\upmu^n(z-k)\big) = \tilde{\epsilon}_i\big(\breve{x}-n\upmu^n(z-k)\big)\big(\bar{x}_i-n\mu_i\rho_i(z-k)\big)+\bar{\epsilon}_i\big(\breve{x}-n\upmu^n(z-k)\big)\sum_{j=1}^{i-1}\big(\bar{x}_j-n\mu_j\rho_j(z-k)\big)\,,$$

where the functions $\tilde{\epsilon}_i,\bar{\epsilon}_i\colon\mathbb{R}^d\to[0,1]$, for $i\in\mathcal{I}$. Since $\tilde{\epsilon}_i$ and $\bar{\epsilon}_i$ are bounded, we have some additional terms which are bounded by $C\int_{\mathbb{R}_*}n\mu_i\rho_i(y-k)\,\tilde{F}_{\breve{x},k}^{d_1^n}(\mathrm{d}y)$ for some positive constant C. Therefore, these are of order $\sqrt{n}$ by (5.21). Thus, repeating the argument in the proof of Lemma B.1, and applying Lemma B.2, we deduce that (B.14) holds with $\psi=0$. This completes the proof. ∎

Proof of Lemma 5.2. The proof mimics that of Proposition 4.2. We sketch the proof when $\mathcal{I}_0$ is empty. Using the estimate

$$\mathcal{O}(q_i^n)\mathcal{O}(|\bar{x}_i|^{m-1}) \le \epsilon^{1-m}\big(\mathcal{O}(q_i^n)\big)^m+\epsilon\big(\mathcal{O}(|\bar{x}_i|^{m-1})\big)^{m/m-1} \quad (B.19)$$

for any $\epsilon>0$, which follows by Young's inequality, we deduce that, for some positive constants $\{c_k\colon k=1,2,3\}$, we have

$$\mathcal{L}_n^{z^n}f_n(\breve{x},h) \le c_1n^{m/2}+c_2(\langle e,q^n\rangle)^m-c_3\sum_{i\in\mathcal{I}}\xi_i|\bar{x}_i|^m \quad \forall\,(\breve{x},h)\in\mathbb{R}_+^d\times\mathbb{R}_+^d\,, \quad (B.20)$$

and all large enough n. Note that Lemma B.2 holds for all $z^n\in\mathfrak{Z}_{\mathrm{sm}}^n$. Then, we may repeat the steps in the proof of Lemma 4.1, except that here we use

$$\begin{aligned}(\tilde{x}_i)^{m-1}\int_{\mathbb{R}_*}\hat{q}_i^n\big(\breve{x}-n\upmu^n(y-k),z^n\big)\,\tilde{F}_{\breve{x},k}^{d_1^n}(\mathrm{d}y)&\\ \le \epsilon|\bar{x}_i|^m+\epsilon^{1-m}\Big(\mathbb{E}\big[\hat{q}_i^n\big(\breve{x}-n\upmu^n(d_1^n-k),z^n\big)\,|\,d_1^n>k\big]\Big)^m&\,,\end{aligned} \quad (B.21)$$

where $\hat{q}^n=n^{-1/2}q^n$, with $\epsilon>0$ chosen sufficiently small. Since $\hat{q}_i^n(\breve{x},z^n)\le\langle e,\tilde{x}\rangle^+$, it follows by (5.21) that

$$\mathbb{E}\big[\hat{q}_i^n\big(\breve{x}-n\upmu^n(d_1^n-k),z^n\big)\;\big|\;d_1^n>k\big] \le c_4(1+\langle e,\tilde{x}\rangle^+)\,. \quad (B.22)$$

Thus, by the same calculation in Proposition 4.2, and using (B.19)–(B.22), we obtain

$$\mathbb{E}^{z^n}\bigg[\int_0^T|\widetilde{X}^n(s)|^m\bigg] \le C_1(T+|\hat{X}^n(0)|^m)+C_2\,\mathbb{E}^{z^n}\bigg[\int_0^T\big(1+\langle e,\widetilde{X}^n(s)\rangle^+\big)^m\,\mathrm{d}s\bigg] \quad (B.23)$$

for all large enough n, and $\{z^n \in \mathfrak{Z}^n_{\mathrm{sm}} \colon n \in \mathbb{N}\}$. Since $\sup_n \hat{J}(\hat{X}^n(0), z^n) < \infty$, it follows by (4.26) that

$$\sup_n \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}\biggl[\int_0^T \bigl(\langle e, \widetilde{X}^n(s)\rangle^+\bigr)^m \,\mathrm{d}s\biggr] < \infty\,.$$

Therefore, dividing both sides of (B.23) by T, taking $T \to \infty$ and using (4.26) again, we obtain (5.11). We may show that the result also holds when $\mathcal{I}_0$ is nonempty by repeating the above argument and applying Lemma B.2. This completes the proof. □

Acknowledgments

This research was supported in part by the Army Research Office through grant W911NF-17-1-001, in part by the National Science Foundation through grants DMS-1715210, CMMI-1635410 and DMS-1715875, and in part by the Office of Naval Research through grant N00014-16-1-2956 and was approved for public release under DCN #43-5442-19.

References

[1] R. Atar, A. Mandelbaum, and M. I. Reiman, *Scheduling a multi class queue with many exponential servers: asymptotic optimality in heavy traffic*, Ann. Appl. Probab. **14** (2004), no. 3, 1084–1134. MR2071417

[2] R. Atar, *Scheduling control for queueing systems with many servers: asymptotic optimality in heavy traffic*, Ann. Appl. Probab. **15** (2005), no. 4, 2606–2650. MR2187306

[3] R. Atar, A. Mandelbaum, and G. Shaikhet, *Simplified control problems for multiclass many-server queueing systems*, Math. Oper. Res. **34** (2009), no. 4, 795–812. MR2573496

[4] A. Arapostathis, A. Biswas, and G. Pang, *Ergodic control of multi-class $M/M/N+M$ queues in the Halfin-Whitt regime*, Ann. Appl. Probab. **25** (2015), no. 6, 3511–3570. MR3404643

[5] A. Arapostathis and G. Pang, *Infinite-horizon average optimality of the N-network in the Halfin-Whitt regime*, Math. Oper. Res. **43** (2018), no. 3, 838–866. MR3846075

[6] A. Arapostathis and G. Pang, *Infinite horizon asymptotic average optimality for large-scale parallel server networks*, Stochastic Process. Appl. **129** (2019), no. 1, 283–322. MR3906999

[7] A. Budhiraja, A. Ghosh, and X. Liu, *Scheduling control for Markov-modulated single-server multiclass queueing systems in heavy traffic*, Queueing Syst. **78** (2014), no. 1, 57–97. MR3238008

[8] R. Kumar, M. E. Lewis, and H. Topaloglu, *Dynamic service rate control for a single-server queue with Markov-modulated arrivals*, Naval Res. Logist. **60** (2013), no. 8, 661–677. MR3146992

[9] L. Xia, Q. He, and A. S. Alfa, *Optimal control of state-dependent service rates in a MAP/M/1 queue*, IEEE Trans. Automat. Control **62** (2017), no. 10, 4965–4979. MR3708873

[10] A. Arapostathis, A. Das, G. Pang, and Y. Zheng, *Optimal control of Markov-modulated multiclass many-server queues*, Stochastic Systems **9** (2019), no. 2, 155–181.

[11] H. M. Jansen, M. Mandjes, K. De Turck, and S. Wittevrongel, *Diffusion limits for networks of Markov-modulated infinite-server queues*, Performance Evaluation **135** (2019), online.

[12] G. Pang and W. Whitt, *Service interruptions in large-scale service systems*, Management Science **55** (2009), no. 9, 1499–1512.

[13] G. Pang and W. Whitt, *Heavy-traffic limits for many-server queues with service interruptions*, Queueing Syst. **61** (2009), no. 2-3, 167–202. MR2485887

[14] H. Lu, G. Pang, and Y. Zhou, *$G/GI/N(+GI)$ queues with service interruptions in the Halfin-Whitt regime*, Math. Methods Oper. Res. **83** (2016), no. 1, 127–160. MR3464192

[15] H. Lu and G. Pang, *Heavy-traffic limits for an infinite-server fork-join queueing system with dependent and disruptive services*, Queueing Syst. **85** (2017), no. 1-2, 67–115. MR3604118

[16] G. Pang and Y. Zhou, *$G/G/\infty$ queues with renewal alternating interruptions*, Adv. in Appl. Probab. **48** (2016), no. 3, 812–831. MR3568893

[17] A. Arapostathis, G. Pang, and Y. Zheng, *Ergodic control of diffusions with compound Poisson jumps under a general structural hypothesis*, ArXiv e-prints **1908.01068** (2019), available at https://arxiv.org/abs/1908.01068.

[18] R. Atar, C. Giat, and N. Shimkin, *On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost*, Queueing Syst. **67** (2011), no. 2, 127–144. MR2771197

[19] T. Konstantopoulos and G. Last, *On the use of Lyapunov function methods in renewal theory*, Stochastic Process. Appl. **79** (1999), no. 1, 165–178. MR1670534

[20] S. P. Meyn and R. L. Tweedie, *Stability of Markovian processes. III. Foster-Lyapunov criteria for continuous-time processes*, Adv. in Appl. Probab. **25** (1993), no. 3, 518–548. MR1234295

[21] P. Billingsley, *Convergence of probability measures*, Second, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, Inc., New York, 1999. A Wiley-Interscience Publication. MR1700749
[22] W. Whitt, *Stochastic-process limits*, Springer Series in Operations Research, Springer-Verlag, New York, 2002. An introduction to stochastic-process limits and their application to queues. MR1876437
[23] J. G. Dai, *On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models*, Ann. Appl. Probab. **5** (1995), no. 1, 49–77. MR1325041
[24] A. Arapostathis, G. Pang, and N. Sandrić, *Ergodicity of a Lévy-driven SDE arising from multiclass many-server queues*, Ann. Appl. Probab. **29** (2019), no. 2, 1070–1126. MR3910024
[25] A. Arapostathis, H. Hmedi, G. Pang, and N. Sandrić, *Uniform polynomial rates of convergence for a class of Lévy-driven controlled SDEs arising in multiclass many-server queues* (G. Yin and Q. Zhang, eds.), Modeling, Stochastic Control, Optimization, and Applications. The IMA Volumes in Mathematics and its Applications, vol. 164, Springer, Cham, 2019. MR3970163
[26] M. H. A. Davis, *Piecewise-deterministic Markov processes: a general class of nondiffusion stochastic models*, J. Roy. Statist. Soc. Ser. B **46** (1984), no. 3, 353–388. With discussion. MR790622
[27] S. M. Ross, *Stochastic processes*, Second, John Wiley & Sons, Inc., New York, 1996. MR1373653
[28] A. Arapostathis, L. Caffarelli, G. Pang, and Y. Zheng, *Ergodic control of a class of jump diffusions with finite Lévy measures and rough kernels*, SIAM J. Control Optim. **57** (2019), no. 2, 1516–1540. MR3942851
[29] E. V. Krichagina and M. I. Taksar, *Diffusion approximation for $GI/G/1$ controlled queues*, Queueing Systems Theory Appl. **12** (1992), no. 3-4, 333–367. MR1200872
[30] G. Pang, R. Talreja, and W. Whitt, *Martingale proofs of many-server heavy-traffic limits for Markovian queues*, Probab. Surv. **4** (2007), 193–267. MR2368951
[31] D. L. Iglehart and W. Whitt, *The equivalence of functional central limit theorems for counting processes and associated partial sums*, Ann. Math. Statist. **42** (1971), 1372–1378. MR0310941