Sample-Path Moderate Deviation Principle for GI/GI/1+GI Queues in the Nearly Critically Loaded Regime

CHANG FENG*, JOHN J. HASENBEIN*, AND GUODONG PANG†

Abstract. This paper establishes sample-path moderate deviation principles (MDP) for GI/GI/1+GI queues in the nearly critically loaded regime (or near-heavy traffic regime). The processes of interest including queue-length process and offered waiting time process are scaled appropriately, with the space scaled in between the order of the time scaling and its square root. The rate functions in the sample-path MDPs for these processes can be explicitly expressed. We employ the method of exponential tightness, exponential equivalence and the contraction principle in large deviation theory and apply to these MDP-scaled processes.

1. Introduction

The single-server queue with abandonment has become a fundamental model in queueing theory and in many applications such as customer service systems, manufacturing and production processes. Many useful results have been derived for GI/GI/1+GI queues under the first-come first-served (FCFS) discipline in heavy traffic. Ward and Glynn [25] proved that the queue-length process converges to a reflected Ornstein–Uhlenbeck (OU) process in the usual heavy-traffic regime, extending the diffusion approximation of the Markovian M/M/1+M model in [24]. In [22], diffusion approximations for GI/GI/1+GI queues are established under the hazard scaling of the patience times in heavy traffic. In [15], the authors studied the convergence of the stationary distributions for the offered waiting processes for GI/GI/1+ GI queues in heavy traffic. On the other hand, very little is known about the large deviations behavior of single-server queues with abandonment. Recently, a sample-path large deviation principle (LDP) was established for the M/M/1+M queue in [4]. It remains open to understand the LDP behavior for more general GI/GI/1+GI queues.

In this paper, we establish a sample-path moderate deviation principle (MDP) for GI/GI/1+GI queues in a near-heavy traffic regime (nearly critically loaded regime). This extends the sample-path MDP result for GI/GI/1 queues in a near-heavy traffic regime by Puhalskii [18] to models with abandonment. For large deviations of queues, the scaling of the processes is the same as the scaling in the functional law of large numbers (FLLN), that is, the same scale of time and space. However, for moderate deviations, the scaling is in between the FLLN and the functional central limit theorem (FCLT), that is, the scaling of space is in between the scale of time and its square root and a centering term is also required as in FCLTs (unless the FLLN limit is zero). For this purpose, the usual heavy traffic condition ($\sqrt{n}(1-\rho_n) \to \beta \in \mathbb{R}$ as $n \to \infty$ with ρ_n

*The University of Texas at Austin, Dept. of Mechanical Engineering, University Station C2200, Austin, TX, 78712-0292

† Department of Computational Applied Mathematics and Operations Research, George R. Brown School of Engineering, Rice University, Houston, TX 77005

E-mail addresses: chang.feng@utexas.edu, has@me.utexas.edu, gdpang@rice.edu.

Date: February 13, 2025.

Key words and phrases. GI/GI/1+GI queues, near-heavy traffic regime (nearly critically loaded regime), sample-path moderate deviation principles, exponential tightness.

being the traffic intensity and n being the scaling parameter) needs to be modified, namely, to a *near-heavy traffic condition*, as it is called in [18] (with $\sqrt{n}$ being modified accordingly, say $n^{1/2-\epsilon}$ for $\epsilon \in (0, 1/2)$). As a result, although FCLT results in diffusion approximations and sample-path MDPs are characterized similarly as sample-path LDPs with a rate function, they are also related in that the rate function in the sample-path MDP only involves the first two orders of the primitives, as in diffusion approximations. The rate function for GI/GI/1+GI queues generalizes that for GI/GI/1 queues in that the dependence on the state of the queue is involved, which is similar to the comparison between the reflected OU diffusion limit for GI/GI/1+GI queues and the reflected Brownian motion limit for GI/GI/1 queues in heavy traffic.

To prove the sample-path MDP for GI/GI/1+GI queues, we adapt the techniques that are often used to prove sample-path LDPs (see, e.g., [11,13,17,20]) by establishing exponential tightness and exponential equivalence and applying the contraction principle for the MDP-scaled processes (see the relevant definitions and discussions in Section 1.2 and Appendix A). Under the MDP scaling, the queue-length process and offered waiting time process resemble the expressions in the FCLT scaling in [25], which we take advantage of. However, to establish exponential tightness and exponential equivalence properties, we need to prove finer estimates than those required in the proofs of the FCLT. We start with the sample-path MDPs for random walks and renewal processes as given in [20] (see also the recent survey paper on random walks [1]), since they are the building blocks for the processes of interest. We then establish finer estimates in the moderate deviation scaling for the various components in the representations of the queueing and workload processes. In the proofs, we also prove stochastic equivalence properties for processes with random time change (see Theorem A.4) and stochastic integrals in $\mathcal{D}$ (see Theorem A.5). These results are new technical contributions to the literature of sample-path MDP theory.

This paper contributes to the very limited literature on moderate deviations in queueing theory. Wischik [27] gave an overview of the moderate deviations in traffic processes and discussed how they are related to the large deviations and central limit theorems (see also [14, 23]). Majewski [16] established sample-path moderate deviations for the cumulative fluid in a queueing system with an increasing number of exponential on-off sources. In [9], large and moderate deviations were established for the workload process at a given time in a stochastic fluid queue model with long-range dependent input. Sample-path moderate deviations for infinite-server queues with general time-varying service times as the rate of the renewal arrival process increases were recently established in [2] as a special case of general shot noise processes in the large intensity regime. Puhalski [19] proved sample-path moderate deviations for GI/GI/N queues in the near Halfin–Whitt regime. There are also several recent works on optimal control of multiclass single-server queues [3,5,6] and multi-server queues [8] in the moderate deviation regime.

1.1. **Organization of the paper.** For the rest of this section, we introduce the basic definitions necessary for our study and some frequently used notation. In Section 2, we set up the model for the GI/GI/1+GI queue under the nearly critically loaded heavy-traffic regime and outline the model assumptions. Section 3 contains the main results of this paper, which include establishing the sample-path MDP for the offered waiting time and queue length processes, as well as solving the rate functions explicitly for the queue-length process. Section 4 presents the proof of Theorem 3.2 and Section 5 presents the technical supporting lemmas and their proofs. In Appendix A, we expand upon the definitions in Section 1.2 by providing several general results in sample-path MDP theory.

1.2. **Notation and definitions.** Let (S, ρ) be a metric space. A function $I : S \to [0, \infty]$ is a *rate function* if its level set $\{x : I(x) \le a\}$ is compact for all $a \ge 0$. We say a sequence $\{P_n,\ n \ge 1\}$ of

probability measures on the Borel σ-algebra of S satisfies a large deviation principle (LDP) with rate a_n and rate function $I : S \to [0, \infty]$ if

$$\limsup_{n\to\infty} \frac{1}{a_n} \log P_n(F) \leq - \inf_{x\in F} I(x), \tag{1.1}$$

for all closed $F \subset S$, and

$$\liminf_{n\to\infty} \frac{1}{a_n} \log P_n(G) \geq - \inf_{x\in G} I(x), \tag{1.2}$$

for all open $G \subset S$.

Throughout the paper, we work in the function space $\mathcal{D} \equiv \mathcal{D}([0, \infty), \mathbb{R}^d)$ of càdlàg processes with sample paths from $[0, \infty)$ to $\mathbb{R}^d$, endowed with the Skorokhod J_1 topology. We note that it is a Polish space. The space $\mathcal{C} \equiv \mathcal{C}([0, \infty), \mathbb{R}^d)$ consists of processes in $\mathcal{D}$ that are continuous. In addition, let $\mathcal{AC}$ be the subset of $\mathcal{C}([0, \infty), \mathbb{R})$ that are absolutely continuous, and $\mathcal{AC}_0 \subset \mathcal{AC}$ be the subset of functions taking value 0 at 0.

When dealing with the limiting phenomenons of stochastic processes, one typically needs to perform certain scaling of time and space. Let X be a process in space $\mathcal{D}$. In the functional law of large numbers (FLLN) setting, one considers $\{\bar{X}^n(t) \equiv n^{-1}X(nt),\ t \geq 0\}$, which usually converges to some deterministic function as $n \to \infty$. This is commonly known in the queueing literature as the fluid limit or fluid approximation. Sample-path LDP usually refers to establishing an LDP with rate $a_n = n$ for the laws of the family $\{\bar{X}^n,\ n \geq 1\}$. In the functional central limit theorem (FCLT) setting, the scaling involves $\{n^{-1/2}(X(nt) - n\bar{X}(t)),\ t \geq 0\}$ with the centering term $\bar{X}(t)$ being the limit of $\bar{X}^n(t)$, which usually converges to some diffusion process as $n \to \infty$. This is commonly known in the queueing literature as the diffusion limit or diffusion approximation.

The study of moderate deviations fills the gap in between. Specifically, the *moderate deviation scaling* concerns a family of the following scaled processes $\{\tilde{X}^n(t) \equiv n^{-1/2-\beta}(X(nt) - n\bar{X}(t)), t \geq 0\}$, where $\beta \in (0, 1/2)$. In this case, one aims to establish an LDP with rate $a_n = n^{2\beta}$ (which we shall refer to as an MDP) for the laws of the family $\{\tilde{X}^n,\ n \geq 1\}$.

We say a family of processes $\{X_n,\ n \geq 1\}$ in $\mathcal{D}$ is *exponentially tight* with rate a_n if for all $\alpha \geq 0$, there exists a compact set $K_\alpha \subset \mathcal{D}$ such that

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}(X_n \notin K_\alpha) < -\alpha. \tag{1.3}$$

Here we note that on Polish spaces, exponential tightness is implied by having an MDP. See Appendix A for a characterization of exponential tightness and some further discussions. A detailed study can be found in Puhalskii [17].

Another concept that frequently shows up in the applications of large deviation theory is exponential equivalence. We say two families of processes $\{X_n,\ n \geq 1\}$ and $\{Y_n,\ n \geq 1\}$ in $\mathcal{D}$ are *exponentially equivalent* with rate a_n if for all $\delta > 0$,

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}(d_{J_1}(X_n, Y_n) > \delta) = -\infty, \tag{1.4}$$

where $d_{J_1}(\cdot, \cdot)$ is the metric induced by the Skorokhod J_1 topology. A special case is when we substitute $\{Y_n,\ n \geq 1\}$ for a fixed $x_0 \in \mathcal{D}$. We say $\{X_n,\ n \geq 1\}$ *converges super-exponentially in probability* to x_0 with rate a_n if for all $\delta > 0$,

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}(d_{J_1}(X_n, x_0) > \delta) = -\infty, \tag{1.5}$$

and write $X_n \overset{P^{1/a_n}}{\longrightarrow} x_0$. Note that this is equivalent to saying that $\{X_n,\ n \geq 1\}$ satisfies an MDP with rate a_n and rate function $I(x) = 0$ when $x = x_0$ and infinity otherwise.

A common tool used in large deviation analysis is the *contraction principle*, see Dembo and Zeitouni [11]. In our setting, it states that if $\{X_n,\ n \geq 1\} \subset \mathcal{D}$ satisfies an LDP with rate a_n and rate function I and if $f : \mathcal{D} \to \mathcal{D}$ is continuous, then $\{f(X_n),\ n \geq 1\}$ satisfies an LDP with rate a_n and rate function

$$I'(y) = \inf_{x:f(x)=y} I(x). \tag{1.6}$$

We mention that the continuity of f can be relaxed to having f continuous on the set where the rate function I is finite. This is sometimes referred to as the extended contraction principle. See Ganesh, O'Connell and Wischik [14, Theorem 4.6] or Puhalskii and Whitt [20, Section 3] for details.

Here is some notation frequently used in this paper. For $X \in \mathcal{D}$, we denote $X(t-)$ as its left limit at time t, where $X(0-) \equiv 0$. We use $X^{-1}(t) \equiv \inf\{s \geq 0 : X(s) > t\}$ to denote the inverse process of X at time t. For any $T > 0$, we denote the supremum of jumps of X on interval $[0, T]$ by $j_T(x) \equiv \sup_{t\in[0,T]} |X(t) - X(t-)|$. Additionally, we use $\mathfrak{e}$ to denote the identity process, that is $\mathfrak{e}(t) = t$, for all $t \geq 0$. For reflection mappings, we use $(\mathcal{R}, \bar{\mathcal{R}})$ to denote the conventional reflection mapping and $(\mathcal{R}_\Gamma, \bar{\mathcal{R}}_\Gamma)$ for the linearly generalized reflection mapping. See Appendix A.1 for details.

2. Model Description and Assumptions

2.1. **The Model.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space that supports all of the random variables and stochastic processes defined in this section. For the model primitives, we take three independent i.i.d. sequences of random variables $\{u_i : i \geq 1\}$, $\{v_i : i \geq 1\}$, and $\{w_i : i \geq 1\}$ that are non-negative with mean 1.

We consider a sequence of GI/GI/1+GI queues indexed by n with renewal arrivals, i.i.d. service times and i.i.d. patience times, under the first-come first-served (FCFS) discipline. For the n-th queue, it has inter-arrival rate ρ_n, service rate 1 and mean patience time m_n. The i-th customer will arrive at the queue at time $t_i^n \equiv \rho_n^{-1} \sum_{k=1}^{i} u_i$, with service time v_i and patience time $d_i^n \equiv m_n w_i$. The patience time d_i^n is interpreted to be the maximum time that the customer will wait in queue before reneging. It has the cumulative distribution function

$$F^n(x) \equiv \mathbb{P}(d_i^n \leq x) = \mathbb{P}\left(w_i \leq \frac{x}{m_n}\right) = F\left(\frac{x}{m_n}\right),$$

where F is the cumulative distribution function (cdf) of w_1.

Next, we define the renewal processes

$$A^n(t) \equiv \max\{i \geq 1 : u_1 + \cdots + u_i \leq \rho_n t\}, \quad t \geq 0,$$

which represents the cumulative number of arrivals up to time t and

$$S^n(t) \equiv \max\{i \geq 1 : v_1 + \cdots + v_i \leq t\}, \quad t \geq 0,$$

which represents the cumulative number of services in the first t units of server's busy time. The maximum over the empty set is to be interpreted as 0.

We shall track the waiting time that a customer arriving at time t experiences. Define the *offered waiting time process* by

$$V^n(t) \equiv \sum_{i=1}^{A^n(t)} v_i 1\{V^n(t_i^n-) < d_i^n\} - B^n(t), \quad t \geq 0, \tag{2.1}$$

where

$$B^n \equiv \int_0^{\cdot} 1\{V^n(s) > 0\} ds,$$

which is the cumulative busy time process. Then the *queue length process* is given by

$$Q^n(t) = \sum_{i=1}^{A^n(t)} 1\{V^n(t_i^n-) < d_i^n\} + \sum_{i=1}^{A^n(t)} 1\{V^n(t_i^n-) \geq d_i^n \text{ and } t_i^n \leq t < t_i^n + d_i^n\} - D^n(t), \quad t \geq 0, \tag{2.2}$$

where $D^n \equiv S^n \circ B^n$ is the departure process.

Consider the filtration $\{\mathcal{F}_i\} \equiv \{\mathcal{F}_i,\ i \geq 0\}$ where $\mathcal{F}_i \equiv \sigma((u_1, v_1, w_1), \ldots, (u_i, v_i, w_i), u_{i+1})$ for all $i \in \mathbb{N}$ and $\mathcal{F}_0 \equiv \sigma(u_1)$. We can define two discrete-time martingales with respect to $\{\mathcal{F}_i\}$, namely,

$$M_v^n(i) \equiv \sum_{j=1}^{i} (v_j - 1) 1\{V^n(t_j^n-) < d_j^n\}, \qquad \forall i \in \mathbb{N}, \tag{2.3}$$

$$M_d^n(i) \equiv \sum_{j=1}^{i} \left(1\{V^n(t_j^n-) \geq d_j^n\} - \mathbb{E}\left[1\{V^n(t_j^n-) \geq d_j^n\} \mid \mathcal{F}_{j-1}\right]\right), \qquad \forall i \in \mathbb{N}, \tag{2.4}$$

with $M_v^n(0) \equiv 0$ and $M_d^n(0) \equiv 0$. After some algebra, we obtain the following relationship, which we call the *evolution equation*:

$$V^n(t) + \sum_{i=1}^{A^n(t)} F\left(\frac{V^n(t_i^n-)}{m_n}\right) = A^n(t) + M_v^n(A^n(t)) - M_d^n(A^n(t)) - t + I^n(t), \quad \forall t \geq 0, \tag{2.5}$$

where $I^n \equiv \mathfrak{e} - B^n$ is the idle time process.

Remark 2.1. Consider a GI/GI/1 queue (without reneging) sharing the same inter-arrival times $\{\rho_n^{-1} u_i,\ i \geq 1\}$ and service times $\{v_i,\ i \geq 1\}$ with the GI/GI/1+GI queue. Denote W^n as its workload process and B_W^n its busy time process. We shall make several observations on their relationship with their respective counterparts in the GI/GI/1+GI queue.

Both workload processes V^n and W^n increase exclusively upon job arrivals, resulting in jumps in their respective paths. When V^n experiences a jump, the process W^n also has a jump of equivalent magnitude at the same time. However, the converse need not hold, as a newly arriving job might opt to renege, thereby excluding it from the V^n workload count. In the absence of arrivals, these processes either diminish at a unit rate when the server is occupied or remain static at zero until the next arrival. Therefore, it follows that

$$V^n(t) \leq W^n(t), \quad \forall t \geq 0,$$

and

$$\sup_{s \leq u, v < t} |V^n(u) - V^n(v)| \leq \sup_{s \leq u, v < t} |W^n(u) - W^n(v)|, \quad \forall s, t \geq 0.$$

Since the GI/GI/1 queue handles the jobs that would otherwise renege in the GI/GI/1+GI queue, we have the following relationship between the busy time processes

$$0 \le B_W^n(t) - B^n(t) \le \sum_{i=1}^{A^n(t)} v_i 1\{V(t_i^n-) \ge d_i\}, \quad \forall t \ge 0.$$

2.2. **Near-Heavy Traffic Condition and MDP Scalings.** We shall assume the GI/GI/1+GI queues satisfy the *near-heavy traffic condition* (also called the nearly critically loaded heavy-traffic regime), see Puhalskii [18, section 1] for a discussion on the nomenclature. Some requirements on the distribution of the primitive random variables are also needed. We gather all the assumptions of our model below:

Assumption 2.2 (Near-Heavy Traffic Requirements)**.**

(a). The sequence $\{b_n, n \ge 1\}$ satisfies $b_n = n^\epsilon$ for some $0 < \epsilon < 1/2$.
(b). As $n \to \infty$, $\frac{1}{b_n}\sqrt{n}(1-\rho_n) \to r$, where r is a finite constant.
(c). $\mathbb{E}[\exp(\rho u_1)] < \infty$ and $\mathbb{E}[\exp(\rho v_1)] < \infty$ for some $\rho > 0$.
(d). The cdf F of w_1 is differentiable at 0, and $F(0) = 0$.
(e). The average patience time $m_n = n$.

Remark 2.3. Some of the above assumptions can be relaxed to a certain degree. There exist different versions in the literature for Assumption 2.2(a). Ours is the same as in Ganesh, O'Connell and Wischik [14, Chapter 9]. Puhalskii [18] instead only requires that $b_n = o(\sqrt{n})$. This is insufficient for our proposes, see the proof of (5.37). Minimally, we require $b_n = o(\sqrt{n})$ and $\lim_{n\to\infty} \log(n)/b_n^2 < \infty$.

Assumption 2.2(c) can be relaxed to having u_1 and v_1 satisfy

$$\limsup_{n\to\infty} \frac{1}{b_n^2} \log[n\mathbb{P}(|u_1| > b_n\sqrt{n})] = -\infty \quad \text{and} \quad \limsup_{n\to\infty} \frac{1}{b_n^2} \log[n\mathbb{P}(|v_1| > b_n\sqrt{n})] = -\infty.$$

This was used in Eichelsbacher and Löwe [12] to establish MDP for sums of i.i.d. random variables. By the contraction principle, we can then establish the MDP for renewal processes and the offered waiting time process. However, the proof strategy involves a truncation argument, see the proof of [12, Theorem 2.2]. This will complicate our proofs, particularly for Lemmas 5.3 and 5.4. For clarity of presentation, we opt for the stronger assumption of finite moment generating functions around a neighborhood of zero.

We postpone the discussions on Assumption 2.2(e) until after presenting the main results. See Remark 4.1.

To conclude this section, we introduce various types of scaling for the above processes:

$$\bar{A}^n(t) \equiv \frac{1}{n}A(nt), \qquad \tilde{A}^n(t) \equiv \frac{1}{b_n\sqrt{n}}\left(A^n(nt) - \rho_n nt\right),$$
$$\bar{S}^n(t) \equiv \frac{1}{n}S(nt), \qquad \tilde{S}^n(t) \equiv \frac{1}{b_n\sqrt{n}}\left(S^n(nt) - nt\right),$$
$$\bar{B}^n(t) \equiv \frac{1}{n}B(nt) \qquad \tilde{B}^n(t) \equiv \frac{1}{b_n\sqrt{n}}(B^n(nt) - nt),$$
$$\tilde{V}^n(t) \equiv \frac{1}{b_n\sqrt{n}}V^n(nt), \qquad \tilde{Q}^n(t) \equiv \frac{1}{b_n\sqrt{n}}Q^n(nt),$$
$$\tilde{D}^n(t) \equiv \frac{1}{b_n\sqrt{n}}(D^n(nt) - nt), \qquad \tilde{I}^n(t) \equiv \frac{1}{b_n\sqrt{n}}I^n(nt),$$

$$\tilde{M}_v^n(t) \equiv \frac{1}{b_n\sqrt{n}} M_v^n(\lfloor nt \rfloor), \qquad \tilde{M}_d^n(t) \equiv \frac{1}{b_n\sqrt{n}} M_d^n(\lfloor nt \rfloor).$$

3. Main results

In this section we present our main results. First, we shall need a well-known result on the MDP for renewal processes. The general version can be found in Puhalskii and Whitt [20, Theorem 6.2]. To obtain the MDP for the product measure, see Dembo and Zeitouni [11, page 129].

Lemma 3.1. *Under Assumption 2.2, $\{(\tilde{A}^n, \tilde{S}^n), n \geq 1\}$ satisfies an MDP in $\mathcal{D}([0,\infty), \mathbb{R}^2)$ with rate b_n^2 and rate function*

$$I_{A,S}(a,s) = I_A(a) + I_S(s), \tag{3.1}$$

where

$$I_A(a) = \begin{cases} \frac{1}{2\sigma_A^2} \int_0^\infty \dot{a}(t)^2 dt, & \text{if } a \in \mathcal{AC}_0, \\ \infty, & \text{otherwise}, \end{cases} \tag{3.2}$$

and

$$I_S(s) = \begin{cases} \frac{1}{2\sigma_S^2} \int_0^\infty \dot{s}(t)^2 dt, & \text{if } s \in \mathcal{AC}_0, \\ \infty, & \text{otherwise}. \end{cases} \tag{3.3}$$

In the following theorem, we establish an MDP for the tuple $(\tilde{V}^n, \tilde{D}^n, \tilde{B}^n)$ by using the contraction principle and Lemma 3.1. See Section 4 for the proof.

Theorem 3.2. *Under Assumption 2.2, $\{(\tilde{V}^n, \tilde{D}^n, \tilde{B}^n), n \geq 1\}$ satisfies an MDP in $\mathcal{D}([0,\infty), \mathbb{R}^3)$ with rate b_n^2 and rate function*

$$I_{V,D,B}(v,d,b) = \inf_{\substack{(a,s)\in\mathcal{D}([0,\infty),\mathbb{R}^2):\\ v=\mathcal{R}_\Gamma(a-s+r\mathfrak{e}),\\ d=s+b,\\ b=a-s+r\mathfrak{e}-v-F'(0)\int_0^\cdot v(s)ds.}} I_{A,S}(a,s), \tag{3.4}$$

where $I_{A,S}$ is given in (3.1).

We next establish an MDP for the tuple $(\tilde{Q}^n, \tilde{D}^n, \tilde{B}^n)$ using the previous theorem.

Theorem 3.3. *Under Assumption 2.2, $\{(\tilde{Q}^n, \tilde{D}^n, \tilde{B}^n), n \geq 1\}$ satisfies an MDP in $\mathcal{D}([0,\infty), \mathbb{R}^3)$ with rate b_n^2 and rate function*

$$I_{Q,D,B}(q,d,b) = \inf_{\substack{(a,s)\in\mathcal{D}([0,\infty),\mathbb{R}^2):\\ q=\mathcal{R}_\Gamma(a-s+r\mathfrak{e}),\\ d=s+b,\\ b=a-s+r\mathfrak{e}-q-F'(0)\int_0^\cdot q(s)ds.}} I_{A,S}(a,s), \tag{3.5}$$

where $I_{A,S}$ is given in (3.1).

Proof. By Lemma 5.7, $\{\tilde{Q}^n,\ n \geq 1\}$ and $\{\tilde{V}^n,\ n \geq 1\}$ are exponentially equivalent. The MDP for $\{\tilde{V}^n,\ n \geq 1\}$ is established in Theorem 3.2. By Dembo and Zeitouni [11, Theorem 4.2.13], $\{(\tilde{Q}^n, \tilde{D}^n, \tilde{B}^n), n \geq 1\}$ satisfies an MDP with the same rate function. □

In the following result, we solve the optimization problem in (3.5) and provide the explicit rate functions. This leads to an MDP for the queue length processes $\{\tilde{Q}^n,\ n \geq 1\}$. The rate function obtained is similar to that of the GI/GI/1 queue, see Puhalskii [18, Theorem 3.2]. However, we note

that our rate function has an additional term showing state dependency. This is a characteristic behavior of the Ornstein–Uhlenbeck process (with its drift depending on the state, in contrast to Brownian motion), to which the queue length process converges under the FCLT setting, as shown in Ward and Glynn [25].

Theorem 3.4. *Under Assumption 2.2, we have the following:*

(a). The sequence $\{(\tilde{Q}^n, \tilde{D}^n, \tilde{B}^n), n \geq 1\}$ *obeys an MDP in* $\mathcal{D}([0,\infty), \mathbb{R}^3)$ *with rate* b_n^2 *and rate function*

$$I_{Q,D,B}(q,d,b) = \int_0^\infty 1\{q(t) > 0\} \left[\frac{1}{2\sigma_A^2}(\dot{q}(t) + F'(0)q(t) + \dot{d}(t) - r)^2 + \frac{1}{2\sigma_S^2}\dot{d}(t)^2\right] dt + \int_0^\infty 1\{q(t) = 0\} \left[\frac{1}{2\sigma_A^2}(\dot{d}(t) - r)^2 + \frac{1}{2\sigma_S^2}(\dot{d}(t) - \dot{b}(t))^2\right] dt,$$

when $q, d, b \in \mathcal{AC}_0$, *$q$ is non-negative, b is non-positive and non-increasing,* $\dot{b}(t) = 0$ *a.e. on the set* $\{q(t) > 0\}$ *and* $I_{Q,D,B}(q,d,b) = \infty$ *otherwise.*

(b). The sequence $\{(\tilde{Q}^n, \tilde{D}^n), n \geq 1\}$ *obeys an MDP in* $\mathcal{D}([0,\infty), \mathbb{R}^2)$ *with rate* b_n^2 *and rate function*

$$I_{Q,D}(q,d) = \int_0^\infty 1\{q(t) > 0\} \left[\frac{1}{2\sigma_A^2}(\dot{q}(t) + F'(0)q(t) + \dot{d}(t) - r)^2 + \frac{1}{2\sigma_S^2}\dot{d}(t)^2\right] dt + \int_0^\infty 1\{q(t) = 0\} \left[\frac{1}{2\sigma_A^2}(\dot{d}(t) - r)^2 + \frac{1}{2\sigma_S^2}\dot{d}(t)^2\right] dt,$$

when $q, d \in \mathcal{AC}_0$, *q is non-negative, and* $I_{Q,D}(q,d) = \infty$ *otherwise.*

(c). The sequence $\{(\tilde{Q}^n, \tilde{B}^n), n \geq 1\}$ *obeys an MDP in* $\mathcal{D}([0,\infty), \mathbb{R}^2)$ *with rate* b_n^2 *and rate function*

$$I_{Q,B}(q,b) = \frac{1}{2(\sigma_S^2 + \sigma_A^2)} \int_0^\infty 1\{q(t) > 0\} \Big(\dot{q}(t) + F'(0)q(t) - r\Big)^2 dt + \frac{1}{2(\sigma_S^2 + \sigma_A^2)} \int_0^\infty 1\{q(t) = 0\} \Big(\dot{b}(t) - r\Big)^2 dt,$$

when $q, b \in \mathcal{AC}_0$, *q is non-negative, b is non-positive and non-increasing,* $\dot{b}(t) = 0$ *a.e. on the set* $\{q(t) > 0\}$ *and* $I_{Q,B}(q,b) = \infty$ *otherwise.*

(d). The sequence $\{\tilde{Q}^n, n \geq 1\}$ *obeys an MDP in* $\mathcal{D}([0,\infty), \mathbb{R})$ *with rate* b_n^2 *and rate function*

$$I_Q(q) = \frac{1}{2(\sigma_S^2 + \sigma_A^2)} \int_0^\infty 1\{q(t) > 0\} \Big(\dot{q}(t) + F'(0)q(t) - r\Big)^2 dt + \frac{r^2}{2(\sigma_S^2 + \sigma_A^2)} \int_0^\infty 1\{q(t) = 0\} dt,$$

when $q \in \mathcal{AC}_0$, *q is non-negative, and* $I_Q(q) = \infty$ *otherwise.*

Proof. For part (a), by Theorem 3.2, the rate function (3.4) at a tuple $(q,d,b) \in D([0,\infty), \mathbb{R}^3)$ is solved by minimizing $I_{A,S}$ over $(a,s) \in D([0,\infty), \mathbb{R}^2)$ that satisfies

$$\begin{cases} q = \mathcal{R}_\Gamma(a - s + r\mathfrak{e}), \\ d = s + b, \\ b = a - s + r\mathfrak{e} - q - \int_0^\cdot F'(0)q(s)ds. \end{cases} \tag{3.6}$$

By Lemma A.1, for any given (q,d,b) and (a,s) that satisfy (3.6), there exists a unique $y \in \mathcal{D}$ such that

$$\begin{cases} \dot{q} + F'(0)q = \dot{a} - \dot{s} + r + \dot{y}, \\ \dot{d} = \dot{s} + \dot{b} = \dot{a} + r - \dot{q} - F'(0)q, \\ \dot{b} = -\dot{y}. \end{cases}$$

Therefore, it follows that

$$\begin{cases} \dot{a} = \dot{d} - r + \dot{q} + F'(0)q, \\ \dot{s} = \dot{d} - \dot{b}. \end{cases} \tag{3.7}$$

Plugging (3.7) into (3.1) and branching on the events $\{q(t) = 0\}$ and $\{q(t) > 0\}$ gives the result.

For parts (b), (c) and (d), note that the projection mapping is continuous under the product topology. We can then otain the results by using the rate function in part (a) and applying the contraction principle. □

4. Proof of Theorem 3.2

In this section we prove the main result Theorem 3.2, with the technical supporting lemmas to be proved in the next section. We also discuss the condition of Assumption 2.2(e) in Remark 4.1 after the proof.

Proof of Theorem 3.2. Scaling the evolution equation (2.5) under the moderate deviations setting yields for all $t \geq 0$,

$$\begin{aligned} &\frac{1}{b_n\sqrt{n}} V(nt) + \frac{1}{b_n\sqrt{n}} \int_0^{nt} F\left(\frac{V(s-)}{m_n}\right) dA(s) \\ = \;& \frac{1}{b_n\sqrt{n}}(A(nt) - \rho_n nt) + \frac{1}{b_n\sqrt{n}}\rho_n nt + \frac{1}{b_n\sqrt{n}} M_v(A(nt)) - \frac{1}{b_n\sqrt{n}} M_d(A(nt)) \\ & - \frac{1}{b_n\sqrt{n}} nt + \frac{1}{b_n\sqrt{n}} I(nt). \end{aligned} \tag{4.1}$$

By a change of variable and pluging in $m_n = n$, we have

$$\frac{1}{b_n\sqrt{n}} \int_0^{nt} F\left(\frac{V(s-)}{m_n}\right) dA(s) = \int_0^t \frac{\sqrt{n}}{b_n} F\left(\frac{b_n}{\sqrt{n}} \tilde{V}^n(s)\right) d\bar{A}(s). \tag{4.2}$$

Substituting the definitions in Section 2 into (4.1), we obtain the following key relationship:

$$\tilde{V}^n(t) + F'(0) \int_0^t \tilde{V}^n(s) ds = \tilde{X}^n(t) + \tilde{I}^n(t), \quad \forall t \geq 0, \tag{4.3}$$

where

$$\begin{aligned} \tilde{X}^n(t) = \tilde{A}^n(t) + \tilde{M}_\nu^n(\bar{A}^n(t)) + \frac{\sqrt{n}}{b_n}(\rho_n - 1)t - \tilde{M}_d^n(\bar{A}^n(t)) \\ + F'(0) \int_0^t \tilde{V}^n(s) ds - \frac{\sqrt{n}}{b_n} \int_0^t F(\frac{b_n}{\sqrt{n}} \tilde{V}^n(s)) d\bar{A}^n(s), \end{aligned} \tag{4.4}$$

and

$$\tilde{I}^n(t) = \frac{1}{b_n\sqrt{n}} \int_0^{nt} 1\{V(s) = 0\} ds = \frac{\sqrt{n}}{b_n} \int_0^t 1\{\tilde{V}^n(s) = 0\} ds. \tag{4.5}$$

Observe that $\tilde{V}^n(t)$ is non-negative and $\int_0^t 1\{\tilde{V}^n(s) = 0\}ds$ increases only when $\tilde{V}^n(t) = 0$. Then by Appendix A.1, we conclude that $\tilde{V}^n(t)$ is a linearly generalized reflection mapping of $\tilde{X}^n(t)$. Specifically,

$$(\tilde{V}^n, \tilde{I}^n) = (\mathcal{R}_\Gamma, \bar{\mathcal{R}}_\Gamma)(\tilde{X}^n), \tag{4.6}$$

where $\Gamma = F'(0)$.

Next, we shall analyze the terms in (4.4). For all $t \geq 0$, we can write

$$\begin{aligned}
\tilde{M}_v^n(t) &= \frac{1}{b_n\sqrt{n}} M_v(nt) \\
&= \frac{1}{b_n\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} (v_j - 1) 1\{V(t_j^{n-}) < d_j^n\} \\
&= \frac{1}{b_n\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} (v_j - 1) \; - \; \frac{1}{b_n\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} (v_j - 1) 1\{V(t_j^{n-}) \geq d_j^n\}.
\end{aligned} \tag{4.7}$$

Let

$$Y_n(\cdot) \equiv \frac{1}{n} \sum_{j=1}^{\lfloor n\cdot \rfloor} v_j.$$

For the first term in (4.7), it is straightforward to check that the process

$$\frac{1}{b_n\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} (v_j - 1) = \frac{\sqrt{n}}{b_n} \left(Y_n(t) - t\right) + \frac{nt - \lfloor nt \rfloor}{b_n\sqrt{n}}, \quad t \geq 0, \tag{4.8}$$

is exponentially equivalent to $\frac{\sqrt{n}}{b_n}(Y_n - \mathfrak{e})$. Similarly, we can check that the process

$$\frac{\sqrt{n}}{b_n}\left(Y_n^{-1}(t) - t\right) = \frac{\sqrt{n}}{b_n}\left(\frac{1}{n} S^n(nt) + \frac{1}{n} - t\right) = \tilde{S}^n(t) + \frac{1}{b_n\sqrt{n}}, \quad t \geq 0, \tag{4.9}$$

is exponentially equivalent to $\tilde{S}^n$. By Theorem 3.1 and Dembo and Zeitouni [11, Lemma 4.2.13], $\frac{\sqrt{n}}{b_n}\left(Y_n^{-1} - \mathfrak{e}\right)$ satisfies an MDP in $\mathcal{D}$ with rate b_n^2 and rate function I_S given in (3.3). By Puhalskii and Whitt [20, Theorem 5.4], this implies that $\{\left(b_n^{-1}\sqrt{n}(Y_n^{-1} - \mathfrak{e}), b_n^{-1}\sqrt{n}(Y_n - \mathfrak{e})\right), \, n \geq 1\}$ satisfies an MDP in $\mathcal{D}([0,\infty), \mathbb{R}^2)$ with rate b_n^2 and rate function

$$I'(x,y) = \begin{cases} I_S(x), & \text{when } y = -x, \\ \infty, & \text{otherwise.} \end{cases} \tag{4.10}$$

For the second term in (4.7), Lemma 5.4 gives

$$\frac{1}{b_n\sqrt{n}} \sum_{j=1}^{\lfloor n\cdot \rfloor} (v_j - 1) 1\{V(t_j^n-) \geq d_j^n\} \stackrel{P^{1/b_n^2}}{\longrightarrow} 0. \tag{4.11}$$

By the MDP of $\{\tilde{A}^n, \; n \geq 1\}$ in Theorem 3.1 and the assumption that $\sqrt{n}/b_n \to \infty$, we can apply Puhalskii and Whitt [20, Lemma 4.2(b)] and obtain

$$\bar{A}^n \stackrel{P^{1/b_n^2}}{\longrightarrow} \mathfrak{e}. \tag{4.12}$$

Combining (4.7)-(4.11) and Puhalskii and Whitt [20, lemma 4.3], it follows that $\{(\tilde{S}^n, \, \tilde{M}_v^n \circ \bar{A}^n), \, n \geq 1\}$ satisfies an MDP in $\mathcal{D}([0,\infty), \mathbb{R}^2)$ with rate b_n^2 and rate function I' given by (4.10).

By Lemma 5.5, (4.12) and Theorem A.4, we have

$$\tilde{M}_d^n \circ \bar{A}^n \overset{P^{1/b_n^2}}{\longrightarrow} 0, \tag{4.13}$$

and Lemma 5.6 shows that

$$F'(0)\int_0^{\cdot} \tilde{V}^n(s)ds - \frac{\sqrt{n}}{b_n}\int_0^{\cdot} F(\frac{b_n}{\sqrt{n}}\tilde{V}^n(s))d\bar{A}^n(s) \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{4.14}$$

Therefore, by (4.4), (4.13) and (4.14), we see that

$$\tilde{X}^n - (\tilde{A}^n + \tilde{M}_v^n \circ \bar{A}^n + \frac{\sqrt{n}}{b_n}(\rho_n - 1)\mathfrak{e}) \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{4.15}$$

By (4.3) and some algebra, we have

$$\begin{aligned}
\tilde{V}^n &= \mathcal{R}_\Gamma(\tilde{X}^n),\\
\tilde{D}^n &= \tilde{S}^n \circ \bar{B}^n + \tilde{B}^n,\\
\tilde{B}^n &= -\frac{\sqrt{n}}{b_n}\int_0^{\cdot} 1\{\tilde{V}^n(s) = 0\}ds = -\tilde{I}^n,\\
\bar{B}^n &= \int_0^{\cdot} 1\{\tilde{V}^n(s) > 0\}ds.
\end{aligned}$$

This indicates the following relationship:

$$\tilde{B}^n = \tilde{X}^n - \mathcal{R}_\Gamma(\tilde{X}^n). \tag{4.16}$$

Note that the linearly generalized reflection mapping is Lipschitz continuous under the local uniform topology. See the discussions in Appendix A.1 and the references within. Therefore we obtain that for any $T > 0$, there exists $K(T) > 0$ such that

$$\sup_{t\in[0,T]} |\tilde{B}^n(t)| \le K(T) \sup_{t\in[0,T]} |\tilde{X}^n(t)|. \tag{4.17}$$

By (4.15), (4.17), the MDP for $\{\tilde{A}^n,\ n \ge 1\}$, $\{\tilde{M}_v^n \circ \bar{A}^n,\ n \ge 1\}$ and Assumption 2.2 (b), we obtain

$$\begin{aligned}
&\lim_{a\to\infty}\limsup_{n\to\infty} \frac{1}{b_n^2}\log\mathbb{P}\Big(\sup_{t\in[0,T]} |\tilde{B}^n(t)| > a\Big)\\
\le\ &\lim_{a\to\infty}\limsup_{n\to\infty} \frac{1}{b_n^2}\log\mathbb{P}\Big(\sup_{t\in[0,T]} |\tilde{X}^n(t)| > \frac{a}{K(T)}\Big)\\
=\ &\lim_{a\to\infty}\limsup_{n\to\infty} \frac{1}{b_n^2}\log\mathbb{P}\Big(\sup_{t\in[0,T]} |\tilde{A}^n(t) + \tilde{M}_v^n \circ \bar{A}^n(t) + \frac{\sqrt{n}}{b_n}(\rho_n - 1)t| > \frac{a}{K(T)}\Big)\\
=\ &-\infty.
\end{aligned}$$

Then, it follows that

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{b_n^2}\log\mathbb{P}\Big(\sup_{t\in[0,T]} \Big|\bar{B}^n(t) - t\Big| > \delta\Big)\\
=\ &\limsup_{n\to\infty} \frac{1}{b_n^2}\log\mathbb{P}\Big(\sup_{t\in[0,T]} \Big|\int_0^t 1\{\tilde{V}^n(s) = 0\}ds\Big| > \delta\Big)\\
=\ &\limsup_{n\to\infty} \frac{1}{b_n^2}\log\mathbb{P}\Big(\sup_{t\in[0,T]} \Big|\tilde{B}^n(t)\Big| > \frac{\sqrt{n}}{b_n}\delta\Big)\\
=\ &-\infty,
\end{aligned}$$

which implies

$$\bar{B}^n \xrightarrow{P^{1/b_n^2}} \mathfrak{e}. \tag{4.18}$$

Therefore, $\tilde{S}^n \circ \bar{B}^n$ satisfies an MDP with rate b_n^2 and rate function I_S given in Theorem 3.1.

Finally, using (4.15), we define exponentially equivalent versions of $\tilde{V}^n$ and $\tilde{B}^n$. For $n \geq 1$, let

$$\tilde{V}'^n \equiv \mathcal{R}_\Gamma(\tilde{A}^n + \tilde{M}_v^n \circ \bar{A}^n + (\sqrt{n}/b_n)(\rho_n - 1)\mathfrak{e}), \tag{4.19}$$

$$\tilde{B}'^n \equiv \tilde{A}^n + \tilde{M}_v^n \circ \bar{A}^n + (\sqrt{n}/b_n)(\rho_n - 1)\mathfrak{e} - \mathcal{R}_\Gamma(\tilde{A}^n + \tilde{M}_v^n \circ \bar{A}^n + (\sqrt{n}/b_n)(\rho_n - 1)\mathfrak{e}). \tag{4.20}$$

We observe that $(\tilde{V}'^n, \tilde{D}^n, \tilde{B}'^n)$ is a continuous function of $(\tilde{A}^n, \tilde{S} \circ \bar{B}^n, \tilde{M}_v^n \circ \bar{A}^n, (\sqrt{n}/b_n)(\rho_n - 1)\mathfrak{e})$. The latter satisfies an MDP with rate b_n^2 and rate function $I''(a, s, m, \rho) = I_{A,S}(a, s)$ when $m = -x$, $\rho = r$, and infinity otherwise. Then the MDP for $(\tilde{V}^n, \tilde{D}^n, \tilde{B}^n)$ follows from the contraction principle and exponential equivalence. □

Remark 4.1. In Assumption 2.2(e), the requirement $m_n = n$ can be relaxed to having m_n grow at the order of n, with little change to the proofs. Further, Ward and Glynn [25] show that in the case when $m_n = n^{1+\epsilon}$, the customers do not renege fast enough and the reneging model behaves like a regular GI/GI/1 queue in the diffusion limit. This observation still holds in the moderate deviations setting. To see this, we point out the necessary changes to our proofs below.

In proof of Theorem 3.2, equation (4.3) becomes

$$\tilde{V}^n(t) = \tilde{\Xi}^n(t) + \tilde{I}^n(t),$$

where

$$\begin{aligned}\tilde{\Xi}^n(t) = \tilde{A}^n(t) + \tilde{M}_v^n(\bar{A}^n(t)) + \frac{\sqrt{n}}{b_n}(\rho_n - 1)t \\ - \tilde{M}_d^n(\bar{A}^n(t)) - \frac{\sqrt{n}}{b_n}\int_0^t F(\frac{b_n}{n^\epsilon\sqrt{n}}\tilde{V}^n(s))d\bar{A}^n(s).\end{aligned}$$

Therefore, we have

$$(\tilde{V}^n, \tilde{I}^n) = (\Phi, \Psi)(\tilde{\Xi}^n).$$

Lemmas 5.4 and 5.5 remain essentially the same. The statement of Lemma 5.6 becomes

$$\frac{\sqrt{n}}{b_n}\int_0^\cdot F(\frac{b_n}{n^\epsilon\sqrt{n}}\tilde{V}^n(s))d\bar{A}^n(s) \xrightarrow{P^{1/b_n^2}} 0.$$

We can show this by following the steps in the proof of (5.32) and observing that for all $t \geq 0$,

$$\frac{\sqrt{n}}{b_n}F(\frac{b_n}{n^\epsilon\sqrt{n}}\tilde{V}^n(t)) = \frac{\tilde{V}^n(t)}{n^\epsilon}\left(F'(0) + \frac{R(\frac{b_n}{n^\epsilon\sqrt{n}}\tilde{V}^n(t))}{\frac{b_n}{n^\epsilon\sqrt{n}}\tilde{V}^n(t)}\right),$$

where $R(x) = o(x)$ as $x \to 0$.

Then in the statement of Theorem 3.2, we would instead have the rate function

$$I_{V,D,B}(v, d, b) = \inf_{\substack{(a,s)\in D([0,\infty),\mathbb{R}^2):\\ v=\mathcal{R}(a-s+r\mathfrak{e}),\\ d=s+b,\\ b=a-s+r\mathfrak{e}-v.}} I_{A,S}(a, s). \tag{4.21}$$

For the queue length process, the proof for Lemma 5.7 remains essentially the same, so the rate function in Theorem 3.3 becomes (4.21) after replacing V with Q. Note that this is the same as the MDP result for the GI/GI/1 queue (cf. Puhalskii [18]).

5. Technical Lemmas and their Proofs

In this section we state and prove the technical lemmas that are used in the proof of Theorems 3.2 and 3.3.

Lemma 5.1. *The family* $\{\tilde{V}^n,\ n \geq 1\}$ *is exponentially tight in* $\mathcal{D}$. *Specifically, for any* $T > 0$,

$$\lim_{K\to\infty} \limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\sup_{0\leq t\leq T} \tilde{V}^n(t) > K \right) = -\infty, \tag{5.1}$$

and for any $\eta > 0$,

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\Big(w'_T(\tilde{V}^n, \delta) \geq \eta \Big) = -\infty, \tag{5.2}$$

where w'_T *is defined in* (A.4). *Further, for any* $\epsilon > 0$,

$$\limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\Big(j_T(\tilde{V}^n) > \epsilon \Big) = -\infty. \tag{5.3}$$

Proof. Consider a GI/GI/1 queue (without reneging) sharing the same inter-arrival times $\{\rho_n^{-1} u_i,\ i \geq 1\}$ and service times $\{v_i,\ i \geq 1\}$ with the GI/GI/1+GI queue. Let W^n be its workload process and $\tilde{W}^n(t) \equiv (b_n\sqrt{n})^{-1} W^n(nt)$ for all $t \geq 0$. By Remark 2.1 and (A.4), we have

$$\tilde{V}^n(t) \leq \tilde{W}^n(t), \quad \forall t \geq 0, \tag{5.4}$$

and

$$w'_T(\tilde{V}^n, \delta) \leq w'_T(\tilde{W}^n, \delta), \quad \forall \delta > 0. \tag{5.5}$$

By Puhalskii [18, Corollary 3.1], $\{\tilde{W}^n,\ n \geq 1\}$ satisfies an MDP, which implies that it is an exponentially tight family. Then (5.4), (5.5) and Lemma A.2 imply (5.1) and (5.2), which are equivalent to the exponential tightness of the family $\{\tilde{V}^n,\ n \geq 1\}$.

Finally, Remark 2.1 implies $j_T(\tilde{W}^n) \geq j_T(\tilde{V}^n)$, which, together with Puhalskii [18, (3.29)], yields (5.3). □

Remark 5.2. We frequently use several facts in the proofs, and we collect them here for clarity. Note that for any $x, y \in \mathbb{R}$, $\log(x + y) \leq \log(2) + \log(x \vee y)$. Also, we have from real analysis that for any sequences $\{x_n,\ n \geq 1\}$, $\{y_n,\ n \geq 1\}$ and $\{a_n,\ n \geq 1\}$,

$$\limsup_{n\to\infty} \frac{1}{a_n} \log(x_n \vee y_n) \leq \left(\limsup_{n\to\infty} \frac{1}{a_n} \log x_n \right) \vee \left(\limsup_{n\to\infty} \frac{1}{a_n} \log y_n \right).$$

Then it follows that if $a_n \to \infty$ as $n \to \infty$,

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{a_n} \log(x_n + y_n) &\leq \limsup_{n\to\infty} \frac{\log(2)}{a_n} + \left(\limsup_{n\to\infty} \frac{1}{a_n} \log x_n \right) \vee \left(\limsup_{n\to\infty} \frac{1}{a_n} \log y_n \right) \\ &= \left(\limsup_{n\to\infty} \frac{1}{a_n} \log x_n \right) \vee \left(\limsup_{n\to\infty} \frac{1}{a_n} \log y_n \right). \end{aligned} \tag{5.6}$$

Sometimes, we use a symmetry argument which relies on the following fact. Let $x \equiv \{x(t),\ t \geq 0\}$ be a process in $\mathcal{D}$. Then, for any $T > 0$, $\delta > 0$,

$$\begin{aligned} \mathbb{P}\Big(\sup_{t\in[0,T]} |x(t)| > \delta \Big) &\leq \mathbb{P}\Big(\Big\{ \sup_{t\in[0,T]} x(t) > \delta \Big\} \cup \Big\{ \sup_{t\in[0,T]} -x(t) > \delta \Big\} \Big) \\ &\leq \mathbb{P}\Big(\sup_{t\in[0,T]} x(t) > \delta \Big) + \mathbb{P}\Big(\sup_{t\in[0,T]} -x(t) > \delta \Big). \end{aligned} \tag{5.7}$$

Lemma 5.3. *Under Assumption 2.2,*

$$\frac{1}{n}\sum_{i=1}^{\lfloor n\cdot\rfloor} u_i \overset{P^{1/b_n^2}}{\longrightarrow} \mathfrak{e}. \tag{5.8}$$

Proof. Let $T>0$ be fixed. Note that for any $\delta>0$ and $\rho>0$, and for n large enough,

$$\begin{aligned}
&\frac{1}{b_n^2}\log\mathbb{P}\left(\sup_{t\in[0,T]}\left|\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}(u_i-1)\right|>\delta\right)\\
\leq\ &-\rho\delta+\frac{1}{b_n^2}\log\mathbb{E}\left[\exp\left(\rho\frac{b_n^2}{n}\sum_{i=1}^{\lfloor nT\rfloor}|u_i-1|\right)\right] &(5.9)\\
\leq\ &-\rho\delta+\frac{\lfloor nT\rfloor}{b_n^2}\log\mathbb{E}\left[\exp\left(\rho\frac{b_n^2}{n}|u_i-1|\right)\right]\\
\leq\ &-\rho\delta+\frac{\lfloor nT\rfloor}{b_n^2}\left(\rho^2\left(\frac{b_n^2}{n}\right)^2\mathbb{E}|u_i-1|^2+\mathcal{O}\left(\rho^3\left(\frac{b_n^2}{n}\right)^3\mathbb{E}|u_i-1|^3\right)\right) &(5.10)\\
\leq\ &-\rho\delta+\frac{\lfloor nT\rfloor}{n}\left(\rho^2\left(\frac{b_n^2}{n}\right)\mathbb{E}|u_i-1|^2+\mathcal{O}\left(\rho^3\left(\frac{b_n^2}{n}\right)^2\mathbb{E}|u_i-1|^3\right)\right).
\end{aligned}$$

The inequality (5.9) is obtained by first applying the triangle inequality and then Doob's maximal inequality for submartingales; (5.10) comes from performing the Taylor expansion, using $\log(x)\leq x-1$ and noticing $\mathbb{E}[\exp\left(\rho b_n^2 n^{-1}|u_i-1|\right)]<\infty$ for n large enough by Assumption 2.2(c). Since $b_n^2/n\to 0$ as $n\to\infty$, we obtain

$$\limsup_{n\to\infty}\frac{1}{b_n^2}\log\mathbb{P}\left(\sup_{t\in[0,T]}\left|\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}(u_i-1)\right|>\delta\right)\leq-\rho\delta.$$

By taking $\rho\to\infty$ and Lemma A.3, it follows that $\frac{1}{n}\sum_{i=1}^{\lfloor n\cdot\rfloor}(u_i-1)\overset{P^{1/b_n^2}}{\longrightarrow}0$. It is easy to see that $\frac{1}{n}\sum_{i=1}^{\lfloor n\cdot\rfloor}u_i-\mathfrak{e}$ is exponentially equivalent to $\frac{1}{n}\sum_{i=1}^{\lfloor n\cdot\rfloor}(u_i-1)$, which implies (5.8). □

Lemma 5.4. *Under Assumption 2.2,*

$$\frac{1}{b_n\sqrt{n}}\sum_{j=1}^{\lfloor n\cdot\rfloor}(v_j-1)1\{V(t_j^n-)\geq d_j^n\}\overset{P^{1/b_n^2}}{\longrightarrow}0. \tag{5.11}$$

Proof. We shall first show that for any $T>0$ and $\delta>0$,

$$\limsup_{n\to\infty}\frac{1}{b_n^2}\log\mathbb{P}\left(\sup_{t\in[0,T]}\frac{1}{b_n\sqrt{n}}\sum_{j=1}^{\lfloor nt\rfloor}(v_j-1)1\{V(t_j^n-)\geq d_j^n\}>\delta\right)=-\infty. \tag{5.12}$$

Fix any $T>0$, $\delta>0$. Observe that

$$\mathbb{P}\left(\sup_{0\leq t\leq T}\frac{1}{b_n\sqrt{n}}\sum_{j=1}^{\lfloor nt\rfloor}(v_j-1)1\{V^n(t_j^{n,-})\geq d_j^n\}>\delta\right)$$

$$\leq \mathbb{P}\left(\left\{\sup_{0\leq t\leq T}\frac{1}{b_n\sqrt{n}}\sum_{j=1}^{\lfloor nt\rfloor}(v_j-1)1\{V^n(t_j^{n,-})\geq d_j^n\}>\delta\right\}\cap\left\{\max_{j=1,\ldots,\lfloor nT\rfloor}\frac{1}{b_n\sqrt{n}}V^n(t_j^{n,-})\leq K\right\}\right)$$
$$+\mathbb{P}\left(\max_{j=1,\ldots,\lfloor nT\rfloor}\frac{1}{b_n\sqrt{n}}V^n(t_j^{n,-})>K\right)$$
$$\leq \mathbb{P}\left(\sup_{0\leq t\leq T}\frac{1}{b_n\sqrt{n}}\sum_{j=1}^{\lfloor nt\rfloor}(v_j-1)1\{d_j^n\leq b_n\sqrt{n}K\}>\delta\right)+\mathbb{P}\left(\max_{j=1,\ldots,\lfloor nT\rfloor}\frac{1}{b_n\sqrt{n}}V^n(t_j^{n,-})>K\right).$$

We claim that for any $K>0$,

$$\limsup_{n\to\infty}\frac{1}{b_n^2}\log\mathbb{P}\left(\sup_{0\leq t\leq T}\frac{1}{b_n\sqrt{n}}\sum_{j=1}^{\lfloor nt\rfloor}(v_j-1)1\{d_j^n\leq b_n\sqrt{n}K\}>\delta\right)=-\infty, \tag{5.13}$$

and that for any $\alpha>0$, there exists an $K_\alpha>0$ large enough such that

$$\limsup_{n\to\infty}\frac{1}{b_n^2}\log\mathbb{P}\left(\max_{j=1,\ldots,\lfloor nT\rfloor}\frac{1}{b_n\sqrt{n}}V^n(t_j^n-)>K_\alpha\right)<-\alpha. \tag{5.14}$$

Then by (5.6) in Remark 5.2 and taking $\alpha\to\infty$, (5.12) is obtained.

To show (5.13), first denote $p_n\equiv\mathbb{P}\big(w_1\leq\frac{b_n}{\sqrt{n}}K\big)$. By Assumption 2.2, $p_n\to 0$ as $n\to\infty$. Take any $\rho>0$. Then, for n large enough, we have the following:

$$\frac{1}{b_n^2}\log\mathbb{P}\left(\sup_{0\leq t\leq T}\frac{1}{b_n\sqrt{n}}\sum_{j=1}^{\lfloor nt\rfloor}(v_j-1)1\{d_j^n\leq b_n\sqrt{n}K\}>\delta\right)$$
$$\leq -\rho\delta+\frac{1}{b_n^2}\log\mathbb{E}\left[\exp\left(\rho\frac{b_n}{\sqrt{n}}\sum_{j=1}^{\lfloor nT\rfloor}(v_j-1)1\{d_j^n\leq b_n\sqrt{n}K\}\right)\right] \tag{5.15}$$
$$= -\rho\delta+\frac{\lfloor nT\rfloor}{b_n^2}\log\mathbb{E}\left[\exp\left(\rho\frac{b_n}{\sqrt{n}}(v_1-1)1\Big\{w_1\leq\frac{b_n}{\sqrt{n}}K\Big\}\right)\right] \tag{5.16}$$
$$= -\rho\delta+\frac{\lfloor nT\rfloor}{b_n^2}\log\left(1+\frac{1}{2}\rho^2p_n\Big(\frac{b_n}{\sqrt{n}}\Big)^2\mathbb{E}[(v_1-1)^2]+\mathcal{O}\left(\Big(\frac{b_n}{\sqrt{n}}\Big)^3\rho^3p_n\mathbb{E}[(v_1-1)^3]\right)\right) \tag{5.17}$$
$$\leq -\rho\delta+\frac{\lfloor nT\rfloor}{n}\left(\frac{1}{2}\rho^2p_n\mathbb{E}[(v_1-1)^2]+\mathcal{O}\left(\frac{b_n}{\sqrt{n}}\rho^3p_n\mathbb{E}[(v_1-1)^3]\right)\right), \tag{5.18}$$

where (5.15) is by applying Markov's inequality and Doob's maximal inequality for submartingales; (5.16) uses the assumptions that $\{v_i,\ i\in\mathbb{N}\}$, $\{w_i,\ i\in\mathbb{N}\}$ are mutually independent i.i.d. sequences and $d_j^n=m_nw_j=nw_j$; (5.17) follows from

$$\exp\left(\rho\frac{b_n}{\sqrt{n}}(v_1-1)1\Big\{w_1\leq\frac{b_n}{\sqrt{n}}K\Big\}\right)\leq\exp\left(\rho\frac{b_n}{\sqrt{n}}|v_1-1|\right)<\infty,$$

for n large by Assumption 2.2; and (5.18) uses the fact that $\log x\leq x-1$.

In (5.18), observe that as $n\to\infty$, $p_n\to 0$ and $b_n/\sqrt{n}\to 0$. Therefore, by Assumption 2.2 (c), $b_nn^{-1/2}\rho^3p_n\mathbb{E}[(v_1-1)^3]\to 0$. Taking $\rho\to\infty$ gives (5.13).

To show (5.14), observe that for any $K>0$,

$$\mathbb{P}\left(\max_{j=1,\ldots,\lfloor nT\rfloor}\frac{1}{b_n\sqrt{n}}V^n(t_j^n-)>K\right)\leq\mathbb{P}\left(\sup_{t\in[0,T]}\frac{1}{b_n\sqrt{n}}V^n\left(n\cdot\frac{1}{\rho_n}\cdot\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}u_i\right)>K\right)$$

$$= \mathbb{P}\left(\sup_{t\in[0,T]} \tilde{V}^n\left(\frac{1}{\rho_n} \cdot \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} u_i \right) > K \right).$$

Assumption 2.2 (b) implies that $\rho_n \to 1$ as $n \to \infty$. It is easy to check that $\frac{1}{\rho_n} \cdot \frac{1}{n} \sum_{i=1}^{\lfloor n\cdot \rfloor} u_i$ is exponentially equivalent to $\frac{1}{n} \sum_{i=1}^{\lfloor n\cdot \rfloor} u_i$. Then by Lemma 5.3,

$$\frac{1}{\rho_n} \cdot \frac{1}{n} \sum_{i=1}^{\lfloor n\cdot \rfloor} u_i \overset{P^{1/b_n^2}}{\longrightarrow} \mathfrak{e}.$$

Therefore, by (5.1) in Lemma 5.1 and Lemma A.4, we obtain (5.14).

By the same argument used for showing (5.12), for any $T > 0$ and $\delta > 0$,

$$\limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\sup_{t\in[0,T]} -\frac{1}{b_n\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} (v_j - 1) 1\{V(t_j^n-) \geq d_j^n\} > \delta \right) = -\infty. \tag{5.19}$$

Then by (5.12), (5.19), and Remark 5.2, we obtain

$$\limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\sup_{t\in[0,T]} \left| \frac{1}{b_n\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} (v_j - 1) 1\{V(t_j^n-) \geq d_j^n\} \right| > \delta \right) = -\infty. \tag{5.20}$$

Finally, (5.11) follows from Lemma A.3. □

Lemma 5.5. *Under Assumption 2.2,*

$$\tilde{M}_d^n \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{5.21}$$

Proof. Similar to the proof of Lemma 5.4, we first show that for any $T > 0$ and $\delta > 0$,

$$\limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\sup_{t\in[0,T]} \tilde{M}_d(t) > \delta \right) = -\infty. \tag{5.22}$$

Observe that for any $K > 0$,

$$\begin{aligned}
&\frac{1}{b_n^2} \log \mathbb{P}\left(\sup_{t\in[0,T]} \tilde{M}_d(t) > \delta \right) \\
&\leq \frac{1}{b_n^2} \log \left[\mathbb{P}\left(\left\{ \sup_{t\in[0,T]} \tilde{M}_d(t) > \delta \right\} \cap \left\{ \max_{j=1,\dots,\lfloor nT \rfloor} \frac{1}{b_n\sqrt{n}} V^n(t_j^n-) \leq K \right\} \right) \right. \\
&\qquad \left. + \mathbb{P}\left(\max_{j=1,\dots,\lfloor nT \rfloor} \frac{1}{b_n\sqrt{n}} V^n(t_j^n-) > K \right) \right] \\
&\leq \frac{1}{b_n^2} \log \left[\mathbb{P}\left(\sup_{t\in[0,T]} \frac{1}{b_n\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} 1\{d_j^n \leq b_n\sqrt{n}K\} - \mathbb{E}\left[1\{d_j^n \leq b_n\sqrt{n}K\} | \mathcal{F}_{j-1} \right] > \delta \right) \right. \\
&\qquad \left. + \mathbb{P}\left(\max_{j=1,\dots,\lfloor nT \rfloor} \frac{1}{b_n\sqrt{n}} V^n(t_j^n-) > K \right) \right] \\
&= \frac{1}{b_n^2} \log \left[\mathbb{P}\left(\max_{i=1,\dots,\lfloor nT \rfloor} \frac{1}{b_n\sqrt{n}} \sum_{j=1}^{i} 1\left\{ w_j \leq \frac{b_n}{\sqrt{n}} K \right\} - \mathbb{P}\left(w_j \leq \frac{b_n}{\sqrt{n}} K \right) > \delta \right) \right.
\end{aligned}$$

$$+\mathbb{P}\left(\max_{j=1,\dots,\lfloor nT\rfloor}\frac{1}{b_n\sqrt{n}}V^n(t_j^n-)>K\right)\Bigg],$$

where in the last step we substitute in $d_j^n = w_j m_n = w_j n$ and use w_j's independence from $\mathcal{F}_{j-1}$. Therefore to show (5.22), by Remark 5.2, it suffices to show that for any $\delta > 0$,

$$\limsup_{n\to\infty}\frac{1}{b_n^2}\log\left[\mathbb{P}\left(\max_{i=1,\dots,\lfloor nT\rfloor}\frac{1}{b_n\sqrt{n}}\sum_{j=1}^{i}1\Big\{w_j\le\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_j\le\frac{b_n}{\sqrt{n}}K\Big)>\delta\right)\right]=-\infty, \quad (5.23)$$

and for any $\alpha > 0$, there exists $K_\alpha > 0$ such that

$$\limsup_{n\to\infty}\frac{1}{b_n^2}\log\left(\max_{j=1,\dots,\lfloor nT\rfloor}\frac{1}{b_n\sqrt{n}}V^n(t_j^n-)>K_\alpha\right)<-\alpha. \quad (5.24)$$

Note that (5.24) is already shown in the proof of Lemma 5.4. For (5.23), observe that for any $\rho > 0$,

$$\begin{aligned}
&\frac{1}{b_n^2}\log\left[\mathbb{P}\left(\max_{i=1,\dots,\lfloor nT\rfloor}\frac{1}{b_n\sqrt{n}}\sum_{j=1}^{i}1\Big\{w_j\le\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_j\le\frac{b_n}{\sqrt{n}}K\Big)>\delta\right)\right]\\
\le\ &-\rho\delta+\frac{1}{b_n^2}\log\mathbb{E}\left[\exp\left(\rho\frac{b_n}{\sqrt{n}}\sum_{j=1}^{\lfloor nT\rfloor}1\Big\{w_j\le\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_j\le\frac{b_n}{\sqrt{n}}K\Big)\right)\right] \quad (5.25)\\
=\ &-\rho\delta+\frac{\lfloor nT\rfloor}{b_n^2}\log\mathbb{E}\left[\exp\left(\rho\frac{b_n}{\sqrt{n}}1\Big\{w_1\le\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_j\le\frac{b_n}{\sqrt{n}}K\Big)\right)\right] \quad (5.26)\\
=\ &-\rho\delta+\frac{\lfloor nT\rfloor}{b_n^2}\log\Bigg[1+\frac{1}{2}\rho^2\frac{b_n^2}{n}\mathbb{E}\left[\left(1\Big\{w_1\le\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_1\le\frac{b_n}{\sqrt{n}}K\Big)\right)^2\right]\\
&\qquad+\mathcal{O}\left(\rho^3(\frac{b_n}{\sqrt{n}})^3\mathbb{E}\left[\left(1\Big\{w_1\le\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_1\le\frac{b_n}{\sqrt{n}}K\Big)\right)^3\right]\right)\Bigg] \quad (5.27)\\
\le\ &-\rho\delta+\frac{\lfloor nT\rfloor}{n}\Bigg(\frac{1}{2}\rho^2\mathbb{E}\left[\left(1\Big\{w_1\le\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_1\le\frac{b_n}{\sqrt{n}}K\Big)\right)^2\right]\\
&\qquad+\mathcal{O}\left(\rho^3\frac{b_n}{\sqrt{n}}\mathbb{E}\left[\left(1\Big\{w_1\le\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_1\le\frac{b_n}{\sqrt{n}}K\Big)\right)^3\right]\right)\Bigg), \quad (5.28)
\end{aligned}$$

where (5.25) uses Doob's maximal inequality for submartingales; (5.26) uses the i.i.d. assumption of sequence $\{w_i,\ i\in\mathbb{N}\}$; (5.27) follows by applying the Taylor expansion and then bounded convergence theorem, and (5.28) uses the fact that $\log x \le x - 1$.

By Assumption 2.2, $\mathbb{P}\{w_1 \le b_n n^{-1/2}K\} \to 0$ and $b_n n^{-1/2} \to 0$ as $n \to \infty$. Also note $\mathbb{E}\big[\big(1\{w_1 \le b_n n^{-1/2}K\} - \mathbb{P}(w_1 \le b_n n^{-1/2}K)\big)^3\big]$ is bounded. Then in (5.28), as $n \to \infty$,

$$\mathbb{E}\left[\left(1\Big\{w_1\le\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_1\le\frac{b_n}{\sqrt{n}}K\Big)\right)^2\right]=\mathbb{P}\Big(w_1\le\frac{b_n}{\sqrt{n}}K\Big)-\mathbb{P}\Big(w_1\le\frac{b_n}{\sqrt{n}}K\Big)^2\to 0,$$

and

$$\rho^3\frac{b_n}{\sqrt{n}}\mathbb{E}\left[\left(1\Big\{w_1\le\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_1\le\frac{b_n}{\sqrt{n}}K\Big)\right)^3\right]\to 0.$$

So it follows that

$$\limsup_{n\to\infty}\frac{1}{b_n^2}\log\left[\mathbb{P}\left(\max_{i=1,\dots,\lfloor nT\rfloor}\frac{1}{b_n\sqrt{n}}\sum_{j=1}^{i}1\Big\{w_j\le\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_j\le\frac{b_n}{\sqrt{n}}K\Big)>\delta\right)\right]<-\rho\delta.$$

Taking $\rho\to\infty$, we obtain (5.23) and hence (5.22).

By the same arguments used for (5.22), for any $T>0$ and $\delta>0$,

$$\limsup_{n\to\infty}\frac{1}{b_n^2}\log\mathbb{P}\left(\sup_{t\in[0,T]}-\tilde{M}_d(t)>\delta\right)=-\infty. \tag{5.29}$$

Then by Remark 5.2, together with (5.22) and (5.29), we have

$$\limsup_{n\to\infty}\frac{1}{b_n^2}\log\mathbb{P}\left(\sup_{t\in[0,T]}|\tilde{M}_d^n(t)|>\delta\right)=-\infty.$$

Finally, (5.21) follows from Lemma A.3. □

Lemma 5.6. *Under Assumption 2.2,*

$$\int_0^{\cdot}F'(0)\tilde{V}^n(s)ds-\int_0^{\cdot}\frac{\sqrt{n}}{b_n}F\Big(\frac{b_n}{\sqrt{n}}\tilde{V}^n(s)\Big)d\bar{A}^n(s)\overset{P^{1/b_n^2}}{\longrightarrow}0. \tag{5.30}$$

Proof. By (4.12), Lemma 5.1 and Theorem A.5, we have

$$\int_0^{\cdot}F'(0)\tilde{V}^n(s)ds-\int_0^{\cdot}F'(0)\tilde{V}^n(s)d\bar{A}^n(s)\overset{P^{1/b_n^2}}{\longrightarrow}0. \tag{5.31}$$

Next, we show that

$$\frac{\sqrt{n}}{b_n}F\Big(\frac{b_n}{\sqrt{n}}\tilde{V}^n\Big)-F'(0)\tilde{V}^n\overset{P^{1/b_n^2}}{\longrightarrow}0. \tag{5.32}$$

By a Taylor expansion and the assumption that $F(0)=0$ and $F\in C^1$ in a neighborhood of 0, we have

$$\frac{\sqrt{n}}{b_n}F\Big(\frac{b_n}{\sqrt{n}}\tilde{V}^n(t)\Big)-F'(0)\tilde{V}^n(t)=\frac{\sqrt{n}}{b_n}R\Big(\frac{b_n}{\sqrt{n}}\tilde{V}^n(t)\Big),\quad\forall t\ge0,$$

where $R(x)=o(x)$ as $x\to0$. It follows that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,T]}\Big|\frac{\sqrt{n}}{b_n}F\Big(\frac{b_n}{\sqrt{n}}\tilde{V}^n(t)\Big)-F'(0)\tilde{V}^n(t)\Big|>\delta\right)\\
&=\mathbb{P}\left(\sup_{t\in[0,T]}\Big|\frac{\sqrt{n}}{b_n}R\Big(\frac{b_n}{\sqrt{n}}\tilde{V}^n(t)\Big)\Big|>\delta\right)\\
&=\mathbb{P}\left(\sup_{t\in[0,T]}\tilde{V}^n(t)\frac{\Big|R\Big(b_nn^{-1/2}\tilde{V}^n(t)\Big)\Big|}{b_nn^{-1/2}\tilde{V}^n(t)}>\delta\right)\\
&\le\mathbb{P}\left(\omega:\sup_{t\in[0,T]}\tilde{V}^n(t,\omega)>\frac{\delta}{c_n(\omega)}\right)+\mathbb{P}\left(\sup_{t\in[0,T]}\tilde{V}^n(t)>K\right),
\end{aligned}$$

where in the last inequality, we once again separate over the event $\{\sup_{t\in[0,T]} \tilde{V}^n(t) > K\}$ and its complement. Notice on $\{\sup_{t\in[0,T]} \tilde{V}^n(t) \le K\}$, we have $b_n n^{-1/2}\tilde{V}^n(t) \to 0$. Therefore we can replace the ratio $R(b_n n^{-1/2}\tilde{V}^n(t))/b_n n^{-1/2}\tilde{V}^n(t)$ by a sequence $c_n(\omega)$ such that

$$\lim_{n\to\infty} c_n(\omega) = 0, \quad \forall \omega \in \Omega.$$

Then by Lemma 5.1, for any $\alpha > 0$, there exists $K_\alpha > 0$ such that

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\sup_{t\in[0,T]} \left| \frac{\sqrt{n}}{b_n} F\Big(\frac{b_n}{\sqrt{n}} \tilde{V}^n(t)\Big) - F'(0)\tilde{V}^n(t) \right| > \delta \right) \\
\le\ &\limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\omega : \sup_{t\in[0,T]} \tilde{V}^n(t,\omega) > \frac{\delta}{c_n(\omega)} \right) \vee \limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\sup_{t\in[0,T]} \tilde{V}^n(t) > K_\alpha \right) \\
\le\ &-\alpha.
\end{aligned}$$

Taking $\alpha \to \infty$ gives (5.32).

By (5.32), (4.12), Remark A.6 and Theorem A.5, we have

$$\int_0^\cdot \left[F'(0)\tilde{V}^n(s) - \frac{\sqrt{n}}{b_n} F\Big(\frac{b_n}{\sqrt{n}}\tilde{V}^n(s)\Big) \right] ds - \int_0^\cdot \left[F'(0)\tilde{V}^n(s) - \frac{\sqrt{n}}{b_n} F\Big(\frac{b_n}{\sqrt{n}}\tilde{V}^n(s)\Big) \right] d\bar{A}^n(s) \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{5.33}$$

By (5.31), it follows from (5.33) that

$$\int_0^\cdot \frac{\sqrt{n}}{b_n} F\Big(\frac{b_n}{\sqrt{n}}\tilde{V}^n(s)\Big) ds - \int_0^\cdot \frac{\sqrt{n}}{b_n} F\Big(\frac{b_n}{\sqrt{n}}\tilde{V}^n(s)\Big) d\bar{A}^n(s) \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{5.34}$$

We note that for $x \in \mathcal{D}$, the integral mapping $x \mapsto \int_0^\cdot x(s)ds$ is continuous. See Whitt [26, Theorem 11.5.1]. So by applying the contraction principle to (5.32), we have

$$\int_0^\cdot F'(0)\tilde{V}^n(s)ds - \int_0^\cdot \frac{\sqrt{n}}{b_n} F\Big(\frac{b_n}{\sqrt{n}}\tilde{V}^n(s)\Big) ds \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{5.35}$$

Finally, combining (5.34) and (5.35) yields (5.30), as desired. □

Lemma 5.7. *The MDP-scaled queue length process $\tilde{Q}^n$ and offered waiting time process $\tilde{V}^n$ are exponentially equivalent.*

Proof. Let $a^n(t)$ denote the arrival time of the job in service at time t. Following the same logic as in Ward and Glynn [25, theorem 3], we have

$$\begin{aligned}
\left|\tilde{Q}^n(t) - \tilde{V}^n(t)\right| \le\ &\left|\tilde{A}^n(t) - \tilde{A}^n(\bar{a}^n(t))\right| + \left|\rho_n \left[\frac{\sqrt{n}}{b_n}(t - \bar{a}^n(t)) - \tilde{V}^n(\bar{a}^n(t)-) \right]\right| \\
&+ \left|\rho_n \left[\tilde{V}^n(\bar{a}^n(t)-) - \tilde{V}^n(t) \right]\right| + \left|\tilde{V}^n(t)(\rho_n - 1)\right| + \frac{1}{b_n\sqrt{n}} \\
&+ \frac{1}{b_n\sqrt{n}} \sum_{i=A^n(a^n(nt))}^{A^n(nt)} 1\{V^n(t_i^n-) \ge d_i^n\}, \quad \forall t \ge 0,
\end{aligned} \tag{5.36}$$

where $\bar{a}^n(t) = n^{-1}a^n(nt)$. We will show that every term on the right hand side is exponentially equivalent to 0.

First observe that for any $t \ge 0$,

$$V^n(a^n(t)-) \le t - a^n(t) \le V^n(a^n(t)-) + v_{A^n(a^n(t))}.$$

After some algebra, we have

$$\tilde{V}^n(\bar{a}^n(t)-) \le \frac{\sqrt{n}}{b_n}(t-\bar{a}^n(t)) \le \tilde{V}^n(\bar{a}^n(t)-) + \frac{1}{b_n\sqrt{n}} v_{A^n(a^n(t))}.$$

We claim that

$$\sup_{k=1,\dots,\lfloor nt\rfloor} \frac{1}{b_n\sqrt{n}} v_k \overset{P^{1/b_n^2}}{\longrightarrow} 0, \tag{5.37}$$

which then implies

$$\left\{ \frac{\sqrt{n}}{b_n}(t-\bar{a}^n(t)) - \tilde{V}^n(\bar{a}^n(t)-),\ t\ge 0 \right\} \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{5.38}$$

To see (5.37), note that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{k=1,\dots,\lfloor nt\rfloor} \frac{1}{b_n\sqrt{n}} v_k > \delta \right) \\
&= \mathbb{P}\left(\left\{ \frac{1}{b_n\sqrt{n}} v_1 > \delta \right\} \cup \cdots \cup \left\{ \frac{1}{b_n\sqrt{n}} v_{\lfloor nt\rfloor} > \delta \right\} \right) \\
&\le \lfloor nt\rfloor \cdot \mathbb{P}\left(\frac{1}{b_n\sqrt{n}} v_1 > \delta \right).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\sup_{k=1,\dots,\lfloor nt\rfloor} \frac{1}{b_n\sqrt{n}} v_k > \delta \right) \\
&\le \limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\frac{1}{b_n\sqrt{n}} v_1 > \delta \right) + \frac{\log(\lfloor nt\rfloor)}{b_n^2}.
\end{aligned} \tag{5.39}$$

For the first term in (5.39), note that for any $\delta>0$, $\rho>0$ and n large enough,

$$\begin{aligned}
&\frac{1}{b_n^2} \log \mathbb{P}\left(\frac{1}{b_n\sqrt{n}} v_1 > \delta \right) \\
&\le -\rho\delta + \frac{1}{b_n^2} \log \mathbb{E}\left[\exp\left(\frac{b_n}{\sqrt{n}} \rho v_1 \right) \right] \\
&\le -\rho\delta + \frac{1}{b_n\sqrt{n}} \rho \mathbb{E}[v_1] + \frac{1}{2n}\rho^2 \mathbb{E}[v_1^2] + \mathcal{O}\left(\frac{b_n}{n\sqrt{n}} \rho^3 \mathbb{E}[v_1^3] \right).
\end{aligned}$$

By taking $\rho\to\infty$, we obtain

$$\limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\frac{1}{b_n\sqrt{n}} v_1 > \delta \right) = -\infty.$$

For the second term in (5.39), Assumption 2.2 (a) implies that $\log(n)/b_n^2 \to 0$ as $n\to\infty$. Therefore, we obtain (5.37).

Assumption 2.2 (b) implies that $\rho_n \to 1$ as $n\to\infty$. Then by Lemma 5.1, we have that for all $\delta>0$,

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\sup_{t\in[0,T]} (1-\rho_n)\tilde{V}^n(t) > \delta \right) \\
&= \limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\sup_{t\in[0,T]} \tilde{V}^n(t) > \frac{\delta}{1-\rho_n} \right) \\
&= -\infty.
\end{aligned}$$

By Lemma A.3, this is equivalent to

$$(1-\rho_n)\tilde{V}^n \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{5.40}$$

In the proof of Theorem 3.2, we have shown that $\{\tilde{V}^n,\ n\geq 1\}$ satisfies an MDP. By Puhalskii and Whitt [20, Lemma 4.2 (b)], the assumption $\sqrt{n}/b_n \to \infty$ as $n\to\infty$, and the fact that $\bar{a}^n(t)\leq t$ for all $t\geq 0$, we have

$$\frac{b_n}{\sqrt{n}}\tilde{V}^n\circ\bar{a}^n \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{5.41}$$

The same reasoning applied to (5.38) yields

$$\left\{\big(t-\bar{a}^n(t)\big)-\frac{b_n}{\sqrt{n}}\tilde{V}^n\big(\bar{a}^n(t)-\big),\ t\geq 0\right\} \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{5.42}$$

Combining (5.41) and (5.42), it follows that

$$\bar{a}^n \overset{P^{1/b_n^2}}{\longrightarrow} \mathfrak{e}. \tag{5.43}$$

Therefore, we can apply Theorem A.4 and obtain

$$\tilde{A}^n(t)-\tilde{A}^n(\bar{a}^n(t)) \overset{P^{1/b_n^2}}{\longrightarrow} 0, \tag{5.44}$$

and

$$\tilde{V}^n(\bar{a}^n(t)-)-\tilde{V}^n(t) \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{5.45}$$

Lastly we shall show

$$\frac{1}{b_n\sqrt{n}}\sum_{i=A^n(a^n(n\cdot))}^{A^n(n\cdot)} 1\{V^n(t_i^n-)\geq d_i^n\} \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{5.46}$$

For any $K>0$ and $t\geq 0$, we can write

$$\begin{aligned}
&\frac{1}{b_n\sqrt{n}}\sum_{i=A^n(a^n(nt))}^{A^n(nt)} 1\{V^n(t_i^n-)\geq d_i^n\}\\
&=\frac{1}{b_n\sqrt{n}}\sum_{i=A^n(a^n(nt))}^{A^n(nt)} 1\{V^n(t_i^n-)\geq d_i^n\}-\mathbb{P}\Big(w_1\leq\frac{b_n}{\sqrt{n}}K\Big)\\
&\qquad+\frac{1}{b_n\sqrt{n}}\sum_{i=A^n(a^n(nt))}^{A^n(nt)}\mathbb{P}\Big(w_1\leq\frac{b_n}{\sqrt{n}}K\Big).
\end{aligned} \tag{5.47}$$

For the first term in (5.47), splitting over the event $\big\{\max_{j=1,\dots,A^n(nT)}V^n(t_j^n-)\leq b_n\sqrt{n}K\big\}$ and its complement, we obtain for any $\delta>0$,

$$\begin{aligned}
&\mathbb{P}\Bigg(\sup_{t\in[0,T]}\Bigg|\frac{1}{b_n\sqrt{n}}\sum_{i=A^n(a^n(nt))}^{A^n(nt)} 1\{V^n(t_i^n-)\geq d_i^n\}-\mathbb{P}\Big(w_1\leq\frac{b_n}{\sqrt{n}}K\Big)\Bigg|>\delta\Bigg)\\
&\leq\mathbb{P}\Bigg(\sup_{t\in[0,T]}\Bigg|\frac{1}{b_n\sqrt{n}}\sum_{i=A^n(a^n(nt))}^{A^n(nt)} 1\Big\{w_i\leq\frac{b_n}{\sqrt{n}}K\Big\}-\mathbb{P}\Big(w_1\leq\frac{b_n}{\sqrt{n}}K\Big)\Bigg|>\delta\Bigg)
\end{aligned}$$

$$+\mathbb{P}\Big(\max_{i=1,\dots,A^n(nT)} V^n(t_i^n-) > b_n\sqrt{n}K\Big). \tag{5.48}$$

Equation (5.23) implies

$$\frac{1}{b_n\sqrt{n}}\sum_{i=1}^{\lfloor n\cdot\rfloor} 1\Big\{w_i \le \frac{b_n}{\sqrt{n}}K\Big\} - \mathbb{P}\Big(w_1 \le \frac{b_n}{\sqrt{n}}K\Big) \overset{P^{1/b_n^2}}{\longrightarrow} 0.$$

Thanks to (4.12), we can apply Puhalskii and Whitt [20, lemma 4.3] and obtain

$$\frac{1}{b_n\sqrt{n}}\sum_{i=1}^{A^n(n\cdot)} 1\Big\{w_i \le \frac{b_n}{\sqrt{n}}K\Big\} - \mathbb{P}\Big(w_1 \le \frac{b_n}{\sqrt{n}}K\Big) \overset{P^{1/b_n^2}}{\longrightarrow} 0.$$

Similarly, by (4.12) and (5.43),

$$\bar{A}^n \circ \bar{a}^n \overset{P^{1/b_n^2}}{\longrightarrow} \mathfrak{e}, \tag{5.49}$$

and hence, we obtain

$$\frac{1}{b_n\sqrt{n}}\sum_{i=1}^{A^n(a^n(n\cdot))} 1\Big\{w_i \le \frac{b_n}{\sqrt{n}}K\Big\} - \mathbb{P}\Big(w_1 \le \frac{b_n}{\sqrt{n}}K\Big) \overset{P^{1/b_n^2}}{\longrightarrow} 0.$$

Taking the difference, we have

$$\frac{1}{b_n\sqrt{n}}\sum_{i=A^n(a^n(n\cdot))}^{A^n(n\cdot)} 1\Big\{w_i \le \frac{b_n}{\sqrt{n}}K\Big\} - \mathbb{P}\Big(w_1 \le \frac{b_n}{\sqrt{n}}K\Big) \overset{P^{1/b_n^2}}{\longrightarrow} 0. \tag{5.50}$$

By Remark 5.2, the bound in (5.48) combined with (5.50) and (5.14) yields that for any $\delta > 0$ and $\alpha > 0$, there exists $K_\alpha > 0$ such that

$$\limsup_{n\to\infty}\frac{1}{b_n^2}\log\mathbb{P}\Big(\sup_{t\in[0,T]}\Big|\frac{1}{b_n\sqrt{n}}\sum_{i=A^n(a^n(nt))}^{A^n(nt)} 1\{V^n(t_i^n-)\ge d_i^n\} - \mathbb{P}\Big(w_1 \le \frac{b_n}{\sqrt{n}}K_\alpha\Big)\Big| > \delta\Big) < -\alpha. \tag{5.51}$$

For the second term in (5.46), we obtain that for any $K > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,T]}\frac{1}{b_n\sqrt{n}}\sum_{i=A^n(a^n(nt))}^{A^n(nt)}\mathbb{P}\Big(w_1 \le \frac{b_n}{\sqrt{n}}K\Big) > \delta\right)\\
&= \mathbb{P}\left(\sup_{t\in[0,T]}\frac{1}{b_n\sqrt{n}}F\Big(\frac{b_n}{\sqrt{n}}K\Big)\Big(A^n(nt) - A^n(a^n(nt))\Big) > \delta\right)\\
&= \mathbb{P}\left(\sup_{t\in[0,T]}\frac{\sqrt{n}}{b_n}F\Big(\frac{b_n}{\sqrt{n}}K\Big)\Big(\bar{A}^n(t) - \bar{A}^n\circ\bar{a}^n(t)\Big) > \delta\right).
\end{aligned}$$

Note by (4.12) and (5.49),

$$\bar{A}^n - \bar{A}^n\circ\bar{a}^n \overset{P^{1/b_n^2}}{\longrightarrow} 0.$$

By similar arguments for (5.32), the sequence $(\sqrt{n}/b_n)F((b_n/\sqrt{n})K) \to F'(0)K$ as $n\to\infty$. Then it is easy to see that for any $K > 0$,

$$\limsup_{n\to\infty}\frac{1}{b_n^2}\log\mathbb{P}\left(\sup_{t\in[0,T]}\frac{1}{b_n\sqrt{n}}\sum_{i=A^n(a^n(nt))}^{A^n(nt)}\mathbb{P}\Big(w_1 \le \frac{b_n}{\sqrt{n}}K\Big) > \delta\right)$$

$$= \limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\left(\sup_{t\in[0,T]} \frac{\sqrt{n}}{b_n} F\Big(\frac{b_n}{\sqrt{n}} K\Big)\Big(\bar{A}^n(t) - \bar{A}^n \circ \bar{a}^n(t)\Big) > \delta \right)$$
$$= -\infty. \tag{5.52}$$

By (5.47), (5.51), (5.52) and taking $\alpha \to \infty$, we obtain (5.46). Finally, combining (5.38)–(5.46), we have shown that

$$|\tilde{Q}^n - \tilde{V}^n| \overset{P^{1/b_n^2}}{\longrightarrow} 0,$$

which completes our proof. □

Appendix A. Some preliminary results and background on sample-path MDP theory

A.1. **Linearly generalized reflection mapping.** We use $(\mathcal{R}, \bar{\mathcal{R}})(x)$ to denote the conventional reflection mapping discussed in Chen and Yao [10, Section 7.2]. In dealing with reflected Ornstein-Uhlenbeck process, the linearly generalized reflection mapping is used, see Reed and Ward [21, Appendix], which establishes its existence, uniqueness and continuity properties. We reproduce here the relevant results under our notation for the reader's convenience.

For a positive integer d, let $x \in \mathcal{D}([0,\infty), \mathbb{R}^d)$ with $x(0) = 0$, and Γ, R be $d \times d$ square matrices, the linearly generalized regulator mapping

$$(\mathcal{R}_\Gamma, \bar{\mathcal{R}}_\Gamma)(x) : \mathcal{D}([0,\infty), \mathbb{R}^d) \to \mathcal{D}([0,\infty), \mathbb{R}^{2d}),$$

is defined by

$$(\mathcal{R}_\Gamma, \bar{\mathcal{R}}_\Gamma)(x) = (\mathcal{R}, \bar{\mathcal{R}})(\mathcal{M}(x)) = (z, l), \tag{A.1}$$

which satisfies

(C1). $z(t) + \int_0^t \Gamma z(s) ds = x(t) + R\, l(t)$, for all $t \geq 0$,
(C2). $l(0) = 0$, l is non-decreasing, and $\int_0^\infty z_j(t) dl_j(t) = 0$, for $j = 1, \ldots, d$,

with $\mathcal{M}(x) = u$ being the solution to the integral equation

$$u(t) = x(t) - \int_0^t \Gamma \mathcal{R}(u)(s) ds. \tag{A.2}$$

It is shown in the appendix of [21] that the solution to (A.2) exists uniquely and is Lipschitz continuous in the local uniform topology. Due to the representation (A.1), it is immediate that the mappings $\mathcal{R}_\Gamma$ and $\bar{\mathcal{R}}_\Gamma$ are Lipschitz continuous. Further, we have an extension of Puhalskii [18, Lemma 2.5], which can be used to explicitly compute the rate functions for moderate deviations.

Lemma A.1. *Let $z \in \mathcal{D}([0,\infty), \mathbb{R}^d)$ be component-wise non-negative and $x \in \mathcal{D}([0,\infty), \mathbb{R}^d)$ be component-wise absolutely continuous. Then $z = \mathcal{R}_\Gamma(x)$ if and only if z is absolutely continuous and there exists an absolutely continuous function $y \in \mathcal{D}([0,\infty), \mathbb{R}^d)$ with the properties*

$$\dot{z}(t) + \Gamma z(t) = \dot{x}(t) + R\dot{y}(t) \quad a.e.,$$

and

$$y_j(0) = 0, \quad \dot{y}_j(t) \geq 0 \ a.e., \quad z_j(t)\dot{y}_j(t) = 0 \ a.e., \quad 1 \leq j \leq d.$$

Further, $\dot{z}_j(t) = 0$ a.e. on the set $\{t : z_j(t) = 0\}$, $j = 1, 2, \ldots, d$.

A.2. **A Criterion for exponential tightness.** Recall that for any $x \in \mathcal{C}[0,T]$ and $\delta \in [0,T]$, we can define the modulus of continuity

$$w(x,\delta) \equiv \sup_{|s-t|<\delta} |x(s)-x(t)|,$$

which is used to characterize tightness in $\mathcal{C}$. For a function $x = \{x(t), t \geq 0\} \in \mathcal{D}$, let

$$w_x[s,t) \equiv \sup_{s \leq u,v < t} |x(u)-x(v)|, \quad s < t, \tag{A.3}$$

and then, for $T > 0$, $\delta > 0$, define the following notion of "modulus of continuity":

$$w'_T(x,\delta) \equiv \inf_{\{t_j\}} \max_{0<j\leq k} w_x[t_{j-1},t_j), \tag{A.4}$$

where $\{t_j\}_{j=0,1,\ldots,k}$ are finite partitions of $[0,T]$ such that $t_j - t_{j-1} > \delta$, for all $j = 1,\ldots,k$.

We restate the following necessary and sufficient condition from Puhalskii [17, Theorem 4.2] for exponential tightness of probability measures in space $\mathcal{D}$ with Skorokhod J_1 topology.

Lemma A.2. *A sequence of probability measures (P_n) on $(\mathcal{D}, J_1)$ is exponentially tight with rate ϵ if and only if:*

(i) For any $T > 0$,

$$\lim_{A\to\infty} \limsup_{\epsilon\to 0} \epsilon \log P_\epsilon(x : \sup_{t\leq T} |x(t)| \geq A) = -\infty. \tag{A.5}$$

(ii) For any $\eta > 0$, $T > 0$,

$$\lim_{\delta\to 0} \limsup_{\epsilon\to 0} \epsilon \log P_\epsilon(x : w'_T(x,\delta) \geq \eta) = -\infty. \tag{A.6}$$

A.3. **Super-exponential Convergence in Probability.** A detailed study can be found in Puhalskii and Whitt [20]. We first state a useful result taken from there, which is a characterization of super-exponential convergence in probability when the limit is deterministic and continuous.

Lemma A.3. *Let $x_0 \equiv (x_0(t),\ t \geq 0)$ be continuous. Then $X_n \overset{P^{1/a_n}}{\longrightarrow} x_0$ if and only if*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\left(\sup_{t\in[0,T]} |X_n(t) - x_0(t)| > \epsilon \right) = -\infty, \tag{A.7}$$

for all $\epsilon > 0$, $T > 0$.

The following result can be seen as an analog of the random time-change theorem in Chen and Yao [10, Theorem 5.3]. It describes when a process is exponentially equivalent to itself after performing a random time-change.

Theorem A.4. *Suppose that the processes $\{y^n,\ n \geq 1\} \subset \mathcal{D}$ satisfy $y^n \overset{P^{1/a_n}}{\longrightarrow} \mathfrak{e}$ and the family of processes $\{X^n,\ n \geq 1\} \subset \mathcal{D}$ is exponentially tight with rate a_n and*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\Big(j_T(X^n) > \epsilon \Big) = -\infty.$$

Then the following holds:

$$X^n - X^n \circ y^n \overset{P^{1/a_n}}{\longrightarrow} 0.$$

Proof. It suffices to show that for any $T > 0$ and $\epsilon > 0$,

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\left(\sup_{t\in[0,T]} \left| X^n(t) - X^n \circ y^n(t) \right| > \epsilon \right) = -\infty. \tag{A.8}$$

Notice that for any $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,T]} \left| X^n(t) - X^n \circ y^n(t) \right| > \epsilon \right) \\
&\le \mathbb{P}\left(\left\{ \sup_{t\in[0,T]} \left| X^n(t) - X^n \circ y^n(t) \right| > \epsilon \right\} \cup \left\{ \sup_{t\in[0,T]} \left| y_n(t) - t \right| < \delta \right\} \right) + \mathbb{P}\left(\sup_{t\in[0,T]} \left| y_n(t) - t \right| \ge \delta \right) \\
&\le \mathbb{P}\Big(w(X^n, \delta) > \epsilon \Big) + \mathbb{P}\Big(\sup_{t\in[0,T]} \left| y_n(t) - t \right| \ge \delta \Big) \\
&\le \mathbb{P}\Big(2w'_T(X^n, \delta) + j_T(X^n) > \epsilon \Big) + \mathbb{P}\Big(\sup_{t\in[0,T]} \left| y_n(t) - t \right| \ge \delta \Big) \\
&\le \mathbb{P}\Big(w'_T(X^n, \delta) > \frac{\epsilon}{3} \Big) + \mathbb{P}\Big(j_T(X^n) > \frac{\epsilon}{3} \Big) + \mathbb{P}\Big(\sup_{t\in[0,T]} \left| y_n(t) - t \right| \ge \delta \Big),
\end{aligned}$$

where in the second to last inequality, we use the fact that $w(x, \delta) \le 2w'_T(x, \delta) + j_T(x)$ (see Billingsley [7, (12.9)]).

By the assumptions on X^n, for any $a > 0$, we can find $\delta_a > 0$ small enough, such that

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\left(w'_T(X^n, \delta_a) > \frac{\epsilon}{3} \right) < -a, \tag{A.9}$$

and that

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\left(j_T(X^n) > \frac{\epsilon}{3} \right) = -\infty. \tag{A.10}$$

By the assumption that $y_n \overset{P^{1/a_n}}{\longrightarrow} \mathfrak{e}$, we have

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\left(\sup_{t\in[0,T]} |y_n(t) - t| > \delta \right) = -\infty. \tag{A.11}$$

Combining (A.9), (A.10) and (A.11), we have

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\left(\sup_{t\in[0,T]} \left| X^n(t) - X^n \circ y^n(t) \right| > \epsilon \right) < -a.$$

Finally, we can take $a \to \infty$, and use Lemma A.3 to conclude the proof. □

The next theorem gives a sufficient condition to the exponential equivalence of stochastic integrals in the space $\mathcal{D}$.

Theorem A.5. *Suppose that $\{x_n,\ n \ge 1\}$ is exponentially tight with rate a_n and for any $\epsilon > 0$*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}(j_T(x_n) > \epsilon) = -\infty.$$

Let $y_n \overset{P^{1/a_n}}{\longrightarrow} y_0$ with $y_0 \in \mathcal{C}$ and each y_n being non-decreasing. Then,

$$\int_0^{\cdot} x_n(s) dy_n(s) - \int_0^{\cdot} x_n(s) dy_0(s) \overset{P^{1/a_n}}{\longrightarrow} 0.$$

Proof. By Lemma A.2, for any $\alpha > 0$, one can choose $K_\alpha > 0$ and $\eta > 0$ such that

$$\limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\Big(\sup_{t\in[0,T]} x_n(t) > K_\alpha \Big) < -\alpha, \tag{A.12}$$

and for any $\epsilon_1 > 0$,

$$\limsup_{n\to\infty} \frac{1}{b_n^2} \log \mathbb{P}\big(w'(x_n, \eta) > \epsilon_1 \big) < -\alpha. \tag{A.13}$$

Take any $\epsilon_2 > 0$, $\epsilon_3 > 0$ and consider a family $\{\Gamma_n,\ n \geq 1\}$ of events defined by

$$\Gamma_n \equiv \big\{ w'(x_n, \eta) \leq \epsilon_1 \big\} \cap \big\{ j_T(x_n) \leq \epsilon_2 \big\} \cap \Big\{ \sup_{t\in[0,T]} |y_n(t) - y_0(t)| \leq \epsilon_3 \Big\} \cap \Big\{ \sup_{t\in[0,T]} x_n(t) \leq K_\alpha \Big\}. \tag{A.14}$$

On the event Γ_n, we can approximate x_n by a step function g_n. Specifically, we partition the time interval $[0, T]$ into intervals of size η by setting $t_i = i\eta$ for $i = 0, 1, \ldots, \lfloor T/\eta \rfloor$ and $t_{\lfloor T/\eta \rfloor + 1} = T$, and then define

$$g_n(t) = \sum_{i=1}^{\lfloor T/n \rfloor} x_n(t_i) 1_{[t_i, t_{i+1})}(t) + x_n(T) 1_T(t), \quad t \in [0, T]. \tag{A.15}$$

By Billingsley [7, (12,9)] and (A.14), it follows that

$$w(x_n, \eta) \leq 2w'(x_n, \eta) + j_T(x_n) \leq 2\epsilon_1 + \epsilon_2,$$

which implies

$$\sup_{t\in[0,T]} |x_n(t) - g_n(t)| \leq 2\epsilon_1 + \epsilon_2. \tag{A.16}$$

Next, observe that for any $u \in [0, T]$ and any $k \geq 1$,

$$\begin{aligned}
&\left| \int_0^u x_n(s) d(y_n - y_0)(s) \right| \\
\leq\ & \left| \int_0^u \Big(x_n(s) - g_n(s) \Big) d(y_n - y_0)(s) \right| + \left| \int_0^u g^n(s) d(y_n - y_0)(s) \right| \\
\leq\ & \int_0^u \big| x_n(s) - g_n(s) \big| d(y_n + y_0)(s) + \left| \int_0^u g_n(s) d(y_n - y_0)(s) \right| \\
\leq\ & \sup_{t\in[0,T]} \big| x_n(t) - g_n(t) \big| \big(y_n(T) + y_0(T) \big) \\
&\qquad + \sup_{t\in[0,T]} \sum_{i=1}^{\lfloor T/\eta \rfloor} |g_n(t_i \wedge t)| \cdot |(y_n - y_0)(t_{i+1} \wedge t) - (y_n - y_0)(t_i \wedge t)| \\
\leq\ & (2\epsilon_1 + \epsilon_2)(2y_0(T) + \epsilon_3) + (K_\alpha + 2\epsilon_1 + \epsilon_2) \cdot \sup_{t\in[0,T]} \sum_{i=1}^{\lfloor T/\eta \rfloor} |(y_n - y_0)(t_{i+1} \wedge t) - (y_n - y_0)(t_i \wedge t)| \\
\leq\ & (2\epsilon_1 + \epsilon_2)(2y_0(T) + \epsilon_3) + (K_\alpha + 2\epsilon_1 + \epsilon_2) \cdot 2\lfloor T/\eta \rfloor \cdot \sup_{t\in[0,T]} |y_n - y_0| \\
\leq\ & (2\epsilon_1 + \epsilon_2)(2y_0(T) + \epsilon_3) + (K_\alpha + 2\epsilon_1 + \epsilon_2) \cdot 2\lfloor T/\eta \rfloor \cdot \epsilon_3.
\end{aligned}$$

Since $\epsilon_1, \epsilon_2, \epsilon_3$ are arbitrary, by picking their values small enough, we obtain that for all $n \geq 1$ and any $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left\{ \sup_{t\in[0,T]} \left| \int_0^t x_n(s) d(y_n - y_0)(s) \right| > \delta \right\} \cap \Gamma_n \right) \\
&\leq \mathbb{P}\Big((2\epsilon_1 + \epsilon_2)(2y_0(T) + \epsilon_3) + (K_\alpha + 2\epsilon_1 + \epsilon_2) \cdot 2\lfloor T/\eta \rfloor \cdot \epsilon_3 > \delta \Big) = 0.
\end{aligned} \tag{A.17}$$

Remark 5.2 yields that for any $\delta > 0$ and $\alpha > 0$,

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\left(\sup_{t\in[0,T]} \left| \int_0^t x_n(s) d(y_n - y_0)(s) \right| > \delta \right) \\
\leq\ & \limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\left(\sup_{t\in[0,T]} |y_n(t) - y_0(t)| > \epsilon_3 \right) \vee \limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\left(\sup_{t\in[0,T]} x_n(t) > K_\alpha \right) \\
&\qquad \vee \limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\big(w'(x_n, \eta) > \epsilon_1\big) \vee \limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\big(j_T(x_n) > \epsilon_2\big) \\
\leq\ & -\alpha.
\end{aligned}$$

We conclude the proof by taking $\alpha \to \infty$ and using Lemma A.3. □

Remark A.6. We note that in Theorems A.4 and A.5, the conditions that $\{x_n,\ n \geq 1\}$ is exponentially tight and

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\big(j_T(x_n) > \epsilon\big) = -\infty, \quad \forall \epsilon > 0,$$

are automatically satisfied if $\{x_n,\ n \geq 1\} \subset \mathcal{D}$ obeys an MDP with a rate function that equals infinity at elements of $\mathcal{D}$ that are either discontinuous or not equal to 0 at 0, see Puhalskii [18, Lemma 2.3].

Acknowledgment We thank the anonymous referee for helpful comments that improved the exposition of the paper. C. Feng and J. Hasenbein are supported by the NSF grant DMS 2108682 and G. Pang is supported by the NSF grant DMS 2216765.

References

[1] Sumith Reddy Anugu and Guodong Pang, *On sample-path moderate deviation principles for random walks*, Working paper (2024).
[2] ———, *Sample path moderate deviations for shot noise processes in the high intensity regime*, Stochastic Processes and their Applications (2024), to appear.
[3] Rami Atar and Anup Biswas, *Control of the multiclass G/G/1 queue in the moderate deviation regime* **24** (2014), no. 5, 2033–2069.
[4] Rami Atar, Amarjit Budhiraja, Paul Dupuis, and Ruoyu Wu, *Large deviations for the single-server queue and the reneging paradox*, Mathematics of Operations Research **47** (2022), no. 1, 232–258.
[5] Rami Atar and Asaf Cohen, *Asymptotically optimal control for a multiclass queueing model in the moderate deviation heavy traffic regime* **27** (2017), no. 5, 2862–2906.
[6] Rami Atar and Subhamay Saha, *Optimality of the generalized cμ rule in the moderate deviation regime*, Queueing Systems **87** (2017), no. 1, 113–130.
[7] Patrick Billingsley, *Convergence of probability measures*, 2nd ed, Wiley series in probability and statistics, John Wiley & Sons, New York, 1999.
[8] Anup Biswas, *Risk-sensitive control for the multiclass many-server queues in the moderate deviation regime*, Mathematics of Operations Research **39** (2014), no. 3, 908–929.
[9] Cheng-Shang Chang, David D Yao, and Tim Zajic, *Large deviations, moderate deviations, and queues with long-range dependent input*, Advances in Applied Probability **31** (1999), no. 1, 254–278.
[10] Hong Chen and David D. Yao, *Fundamentals of queuing networks: performance, asymptotics, and optimization*, Applications of mathematics, Springer, New York, 2001.
[11] Amir Dembo and Ofer Zeitouni, *Large Deviations Techniques and Applications*, Springer New York, New York, NY, 1998.
[12] Peter Eichelsbacher and Matthias Lö"we, *Moderate deviations for i.i.d. random variables*, ESAIM: PS **7** (2003), 209–218.
[13] Jin Feng and Thomas G Kurtz, *Large deviations for stochastic processes*, American Mathematical Society, 2006.
[14] Ayalvadi Ganesh, Neil O'Connell, and Damon Wischik, *Big Queues* (J.-M. Morel, F. Takens, and B. Teissier, eds.), Lecture Notes in Mathematics, vol. 1838, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[15] Chihoon Lee, Amy R Ward, and Heng-Qing Ye, *Stationary distribution convergence of the offered waiting processes for GI/GI/1+ GI queues in heavy traffic*, Queueing Systems **94** (2020), no. 1, 147–173.
[16] Kurt Majewski, *Sample path moderate deviations for the cumulative fluid produced by an increasing number of exponential on-off sources*, Queueing Systems **56** (2007), 9–26.
[17] Anatolii A. Puhalskii, *On functional principle of large deviations*, Vol. 1 Proceedings of the Bakuriani Colloquium in honour of Yu.V. Prohorov, 1991, pp. 198–218.
[18] ———, *Moderate deviations for queues in critical loading*, Queueing Systems **31** (1999), no. 3, 359–392.
[19] ———, *Moderate deviations of many-server queues in the Halfin-Whitt regime and weak convergence methods*, arXiv preprint arXiv:2305.01612 (2023).
[20] Anatolii A. Puhalskii and Ward Whitt, *Functional large deviation principles for first-passage-time processes*, The Annals of Applied Probability (1997), 362–381.
[21] Josh Reed and Amy R Ward, *A diffusion approximation for a generalized Jackson network with reneging*, Proceedings of the 42nd annual Allerton conference on communication, control, and computing, 2004.
[22] Josh E Reed and Amy R Ward, *Approximating the GI/GI/1+ GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic*, Mathematics of Operations Research **33** (2008), no. 3, 606–644.
[23] Adam Shwartz and Alan Weiss, *Large deviations for performance analysis: queues, communications, and computing*, 1st ed, Stochastic modeling series, Chapman & Hall, London ; New York, 1995.
[24] Amy R Ward and Peter W Glynn, *A diffusion approximation for a Markovian queue with reneging*, Queueing Systems **43** (2003), 103–128.
[25] Amy R. Ward and Peter W. Glynn, *A diffusion approximation for a GI/GI/1 queue with balking or reneging*, Queueing Systems **50** (2005), 371–400.
[26] Ward Whitt, *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues* (Peter W. Glynn and Stephen M. Robinson, eds.), Springer Series in Operations Research and Financial Engineering, Springer New York, New York, NY, 2002 (en).
[27] Damon Wischik, *Moderate deviations in queueing theory*, preprint (2001).