The Annals of Applied Probability
2019, Vol. 29, No. 1, 613–651
https://doi.org/10.1214/18-AAP1432

SERVE THE SHORTEST QUEUE AND WALSH BROWNIAN MOTION

BY RAMI ATAR[1] AND ASAF COHEN

Technion—Israel Institute of Technology and University of Haifa

We study a single-server Markovian queueing model with N customer classes in which priority is given to the shortest queue. Under a critical load condition, we establish the diffusion limit of the nominal workload and queue length processes in the form of a Walsh Brownian motion (WBM) living in the union of the N nonnegative coordinate axes in $\mathbb{R}^N$ and a linear transformation thereof. This reveals the following asymptotic behavior. Each time that queues begin to build starting from an empty system, one of them becomes dominant in the sense that it contains nearly all the workload in the system, and it remains so until the system becomes (nearly) empty again. The radial part of the WBM, given as a reflected Brownian motion (RBM) on the half-line, captures the total workload asymptotics, whereas its angular distribution expresses how likely it is for each class to become dominant on excursions.

As a heavy traffic result, it is nonstandard in three ways: (i) In the terminology of Harrison (In *Stochastic Networks* (1995) 1–20 Springer), it is *unconventional*, in that the limit is not a RBM. (ii) It does not constitute an *invariance principle*, in that the limit law (specifically, the angular distribution) is not determined solely by the first two moments of the data, and is sensitive even to tie breaking rules. (iii) The proof method does not fully characterize the limit law (specifically, it gives no information on the angular distribution).

1. Introduction. We consider a multiclass single-server queueing system operating under *serve the shortest queue* (SSQ) (also referred to in the literature as *shortest queue first*) regime, where service is offered to the customer class in which the queue is shortest. The practical significance of this policy has been recognized [3, 5, 6, 8–10, 19, 20], and analytic results have been obtained [6, 8–10]. Briefly, our probabilistic assumptions are that both arrival and potential service processes are Poisson, which makes the model Markovian, and that arrival and service rates are class-dependent. The diffusion scale behavior of the model in heavy traffic has not been studied before. The main result of this paper addresses the N-dimensional nominal workload (a term adopted from [21], expressing conditional expectation

Received February 2018; revised July 2018.
[1]Supported in part by ISF Grant 1184/16.
MSC2010 subject classifications. 60F05, 93E03, 60K25, 60J65, 60J70.
Key words and phrases. Serve the shortest queue, heavy traffic, diffusion limits, Walsh Brownian motion.

of workload given the state) and queue length processes, where N denotes the number of classes. It asserts that, under a critical load condition, the diffusion scale versions of both these processes converge to processes living in the set $\mathcal{S}_0$, which consists of the union of the N coordinate axes in $\mathbb{R}_+^N$. Specifically, the rescaled nominal workload converges to a *Walsh Brownian motion* (WBM) on $\mathcal{S}_0$, and the rescaled queue length converges to a certain diagonal transformation of the same process.

WBM was introduced by Walsh in [27] as a planar diffusion that has a singular behavior at the origin. Away from the origin, it evolves as a one-dimensional Brownian motion (BM) along a ray connecting its position to the origin, and its excursions into rays emanating from the origin follow a fixed angular distribution. Some early results on this process, including its special case referred to as *skew BM*, where the state space consists of exactly two rays, are [1, 2, 13, 23, 24, 26]. Intriguing aspects related to the natural filtration of this process were addressed in [25]. Recently, vast extensions of this model have been proposed and thoroughly studied. The reader is referred to [16] and the references therein for this development.

In the terminology of Harrison [12], an *unconventional* limit theorem for a queueing system in heavy traffic is one for which the limit process is not given as a reflected Brownian motion (RBM). Our result thus belongs to a family of unconventional heavy traffic limits, starting from [14] and including the more recent [17] as well as several other results surveyed in [29] and [17]. Moreover, our heavy traffic result is nonstandard in that it does not constitute an *invariance principle*. That is, it is observed in simulations that the limit law (specifically, the angular distribution of the limit WBM) is not determined solely by the first two moments of the data. The simulations also indicate that it is sensitive even to tie breaking rules. A third nonstandard aspect of the result is that the proof method does not provide an explicit expression or a characterization of the limit law. Whereas the modulus is given as a RBM with specified drift and diffusion coefficients, no information on the angular distribution is available from the proof. In fact, it appears unlikely to the authors that an explicit expression can be attained except under some special symmetry.

Some further details on the policy are as follows. In the literature, there are two variants, distinguished by the interpretation given to the selection of jobs from the shortest queue: that may refer to the one having least nominal workload or the one having least number of jobs. We adopt here the convention of [8–10] and work with the former. However, for all other purposes, the term *queue length* refers in this paper to job count. Next, the service rule is assumed to follow a preemptive priority. Finally, the tie breaking rule is a part of the model description. We allow for a rather general choice, by assuming that when the collection, $\mathcal{K}$, of classes having shortest queue consists of more than one class, the server's effort is split according to a specified probability measure $p^{\mathcal{K}}$ supported on the set $\mathcal{K}$.

Under static priority it is well known since Whitt's result [28] that in heavy traffic, the queue which has least priority is always *dominant*, where this term means that nearly all the workload in the system is contained in this queue. Under SSQ, heuristically, one may imagine that very soon after each time the system (nearly) empties, a competition takes place among the queues, where the one that loses ends up with most workload, and consequently least priority. Thus it is reasonable, in view of the aforementioned result on static priority, to expect that the losing class actually becomes dominant and remains so until again the system becomes empty (or nearly empty). This heuristic suggests, moreover, that the choice of the class to become dominant during an excursion (the outcome of the competition, one may say) is random, and is highly sensitive to the dynamics of the Markov process as the queues just start to build. The result of this paper reveals an asymptotic behavior with exactly these elements. One of the most significant and least obvious aspects of it is that the probabilities of each class becoming dominant starting from an empty system do converge in the scaling limit. Indeed, their limit is given by the WBM's angular distribution.

An important feature of SSQ is that when two streams of arrivals have similar first-order characteristics but one is more variable than the other, or has greater tendency to exhibit bursts, the policy tends to prioritize the former over the latter. This is due to the fact that a burst of traffic is likely to cause a long queue, resulting in lower priority. For this reason, SSQ has been referred to in the literature as *"implicit service differentiation"* [5, 6, 20] and *"self prioritization"* [5]. Quoting from [9], *"... priority is thus implicitly given to smooth flows over data traffic... sending packets in bursts"*. The policy has gained interest in technological uses, specifically in the context of packet scheduling [3, 5, 6, 8–10, 20]. For example, in [3, 5, 20] SSQ (referred to there as *shortest queue first*) is compared with *first-in first-out* and *stochastic fairness queueing*, via experimental tests, and is argued to be the best candidate solution for *quality of service* on ADSL internet access in various tests (web browsing, file download, peer-to-peer file sharing, VoIP and video calls, audio streaming and video streaming). It is also found experimentally that the policy prioritizes TCP acknowledgment and delay- and loss-sensitive applications (voice, audio and video streaming), which leads to lower loss counts and delays. For further advantages and additional uses of this policy, see [6] and the references therein, as well as [19].

The policy has been theoretically analyzed in several papers. Guillemin and Simonian [9] study the case of two buffers with Poisson arrivals and general service time distributions, establishing functional equations for the Laplace transform of the workload processes at stationarity. They also specialize to the symmetric, exponential service time case, where they are able to derive empty queue probabilities and tail behavior for the distribution of the workload. In [10], the authors study the same features in the asymmetric case, again for $N = 2$ at stationarity, where service times are exponentially distributed. The paper by Carofiglio and Muscariello [6] studies instantaneous throughput and buffer occupancy of $N \geq 2$ long-lived

TCP sources, using a deterministic fluid model, under three per-flow scheduling disciplines: fair queuing, longest queue first and shortest queue first, assuming longest queue drop buffer management. They obtain closed form expressions for the stationary throughput and the buffer occupancy.

We now make some comments about the proof. To this end, we introduce $\hat{X}^r(t)$, $t \in [0,\infty)$, that are $[0,\infty)^N$-valued processes indexed by the scaling parameter $r \in [1,\infty)$. The component $\hat{X}^r_i(t)$ represents the nominal workload in buffer i at time t, rescaled diffusively; the precise definition appears in Section 2. We start by treating the rescaled total nominal workload, $\sum_i \hat{X}^r_i(t)$, and recall the well-known fact that it converges to an RBM under any work conserving policy, to which SSQ is no exception. This result is required in a slightly extended form, stated in Lemma 3.1, which asserts that convergence holds uniformly with respect to initial conditions. The remainder of the proof has three main ingredients. The first is concerned with showing that $\hat{X}^r$ resides close to $\mathcal{S}_0$ as r gets large. The aforementioned term "dominant queue" is treated mathematically by considering tubes of width $\varepsilon > 0$ about each of the N positive coordinate axes. In terms of these tubes, queue i is dominant at time t if $\hat{X}^r_i(t)$ resides in an ε-tube about axis i, for arbitrarily small ε and large r. Thus the first main ingredient of the proof is to show that the probability of exiting the collection of N tubes tends to zero as $r \to \infty$. This is the content of Lemma 3.2(i). Note that this element, along with the weak convergence of the total nominal workload to a RBM, immediately provides the convergence of the modulus process to the same RBM.

The second main ingredient is concerned with the angular behavior. It is to show that the entrance law into tubes converges in the scaling limit. We consider first a special case of the model, that we call the *homogeneous* case, in which the transition intensities of the underlying Markov process corresponding to $r > 1$ are rescaled version of those for $r = 1$. This trick buys us the ability to transform the double limit problem of entrance law into ε-tubes (involving $\varepsilon \to 0$ and $r \to \infty$) to a single limit (involving $r \to \infty$ only). The existence of a limit of the entrance law is shown by arguing that, starting at the origin, the probabilities of entering $r^{-\kappa_0}$-tubes form a Cauchy sequence, where $\kappa_0 > 0$ is a suitable constant. The tools used to establish this argument are the martingale property of the total nominal workload (that also owes to homogeneity), and a strengthening of Lemma 3.2(i) which improves the $o(1)$ exit probability estimates to polynomial estimates. Relying on the homogeneous case, the general case is then treated by means of a change of measure. The homogeneous case is stated in Lemma 3.5. The double limit assertion is stated as Proposition 3.3, and the reduced version in the form of a single limit is given in (3.43). The polynomial exit probability measure is proved by means of construction of a Lyapunov function for the distance of the state from $\mathcal{S}_0$, that may be interpreted as the nominal workload included in all but the dominant class. This tool is stated in Lemma 3.4. Finally, the change of measure argument is provided within the proof of Proposition 3.3 in Section 3.5.

The third main ingredient is the asymptotic independence of modulus and angle. This relies, first and foremost, on the second ingredient alluded to above, as well as on strong Markovity of the prelimit process and some estimates on the heat kernel associated with RBM on the half-line. This asymptotic independence property is stated in (3.14). These ingredients are finally combined in the proof of the main result, building on the characterization of WBM via its semigroup [1], and using crucially strong Markovity of the prelimit.

Some earlier results on the convergence of discrete processes to WBM appear in [13] and [11]. The paper [13] studies the case of a skew BM. The convergence result included within this paper addresses a suitably defined random walk on the integers observed at the diffusion scale, and establishes its weak convergence to a skew BM. The focus of [11] is the stochastic flow associated to WBM, and for this model, discrete approximations to the flow are obtained. In both these references, the pre-limit processes already live in a collection of N rays ($N = 2$ in the former, $N \geq 2$, finite, in the latter), forming a symmetric random walk everywhere on the state space except at the origin. Consequently, the three main issues alluded to above in the description of our proof (estimates on exiting tubes, existence of a limit for the entrance probability into tubes, asymptotic independence) are all trivial in the cases studied in [13] and [11].

A general method was introduced in [18] for obtaining convergence of regenerative processes from a certain notion of convergence of their excursions. The regenerative processes we treat do fall into the category of those addressed in [18]. However, in the setting considered here, proving the convergence of excursions amounts, roughly speaking, to establishing the three ingredients alluded to above, and so it seems that as far as our result is concerned, this method does not provide a significant shortcut.

The paper is organized as follows. Section 2 presents the model and the main result. Section 3 is devoted to the proof. First, in Section 3.1, the result is proved based on Lemma 3.1, Lemma 3.2 and Proposition 3.3, stated in the beginning of the section. The convergence of the total nominal workload to a RBM is proved in Section 3.2. Section 3.3 provides estimates on probabilities to exit the tubes. Section 3.4 and Section 3.5 establish the limit result regarding the angular distribution, dealing with the homogeneous case and the general case, respectively. Finally, some concluding remarks are included in Section 4.

Notation. For $x, y \in \mathbb{R}^N$ (N a positive integer), let $x \cdot y$ and $\|x\|$ denote the usual scalar product and ℓ_2 norm, respectively. Denote $[N] = \{1, 2, \ldots, N\}$ and let $\{e_i : i \in [N]\}$ denote the standard basis in $\mathbb{R}^N$. Let $\mathbb{1}$ denote the N-dimensional vector whose all entries equal 1. For $x \in \mathbb{R}^N$ and $A \subset \mathbb{R}^N$, let $\operatorname{dist}(x, A) = \inf\{\|x - y\| : y \in A\}$. Let $B(x, r) = \{y \in \mathbb{R}^N : \|y - x\| \leq r\}$ denote the closed ball. Denote $\mathbb{R}_+ = [0, \infty)$. For $f : \mathbb{R}_+ \to \mathbb{R}^N$ and $T \in \mathbb{R}_+$, let $\|f\|_T = \sup_{t \in [0,T]} \|f(t)\|$, and, for $\theta > 0$, $w_T(f, \theta) = \sup_{0 \leq s < u \leq s + \theta \leq T} \|f(u) - f(s)\|$. For

a Polish space E, let $\mathcal{C}_E[0,T]$ and $\mathcal{D}_E[0,T]$ denote the set of continuous and, respectively, càdlàg functions $[0,T] \to E$. Let $\mathcal{C}_E[0,\infty)$ and $\mathcal{D}_E[0,\infty)$ denote the respective sets of functions $[0,\infty) \to E$. Endow $\mathcal{D}_E[0,\infty)$ with the Skorohod J_1 topology. A sequence of processes $\{X_n\}_n$ with sample paths in $\mathcal{D}_E[0,\infty)$ is said to be *$\mathcal{C}$-tight* if it is tight and every subsequential limit has, with probability 1, sample paths in $\mathcal{C}_E[0,\infty)$. Write $X_n \Rightarrow X$ for convergence in law. Let $\mathcal{C}_0(E)$ denote the set of continuous, compactly supported functions on E. For $b \in \mathbb{R}$ and $\sigma \in (0,\infty)$, a *(b,σ)-BM starting from $x \in \mathbb{R}$* is a 1-dimensional BM having drift b, infinitesimal covariance σ^2 and initial condition x. A (b,σ)-RBM starting from $x \in \mathbb{R}_+$ is an RBM in $\mathbb{R}_+$ with reflection at zero, with the corresponding parameters and initial condition x. Denote by $\mathcal{M}_1$ the collection of N-dimensional probability vectors, namely $\mathcal{M}_1 = \{x \in \mathbb{R}_+^N : \sum_i x_i = 1\}$. Throughout, we use the letter c to denote a positive deterministic constant whose value may change from one appearance to another.

2. Setting and result.

2.1. *Serve-the-shortest-queue in heavy traffic.* Consider a sequence of queueing models indexed by $r \in [1,\infty)$, defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. A server operates to serve customers of $N \geq 2$ classes. Each customer class has a dedicated buffer with infinite room. Upon arrival, a class-i customer is queued in buffer $i \in [N]$. The process representing the number of customers in buffer i is called the *ith queue length* and is denoted by $Q^r = (Q_1^r, \ldots, Q_N^r)$. The $\mathbb{Z}_+^N$-valued random variable (RV) $Q^r(0) = (Q_1^r(0), \ldots, Q_N^r(0))$ is referred to as the *initial queue length*. The arrivals are Poissonian and the service times are exponential. To model these, let $\{A_i^r\}_{i\in[N]}$, $\{S_i^r\}_{i\in[N]}$ be a collection of $2N$ mutually independent Poisson processes, with right-continuous sample paths, independent of the initial queue length, where A_i^r (resp., S_i^r) has rate λ_i^r (resp., μ_i^r). The processes A_i^r and S_i^r represent the arrival and potential service processes for class i, respectively. More precisely, $A_i^r(t)$ is the number of class-i customers to arrive (to buffer i) until time t, and $S_i^r(t)$ gives the number of class-i service completions by the time the server has dedicated t units of time to class-i customers.

The process $X^r = (X_1^r, \ldots, X_N^r)$ defined by

$$X_i^r = (\mu_i^r)^{-1} Q_i^r \tag{2.1}$$

is referred to as the *nominal workload process*. This term, borrowed from [21], expresses the fact that $X_i^r(t)$ represents the conditional expectation of the time it takes to serve the $Q_i^r(t)$ customers present in buffer i at time t, conditioned on $Q_i^r(t)$ (assuming that the server works exclusively on this class).

Within each class, only one customer may be served at a time (and for concreteness, we may assume it is the oldest one present in the system), although service effort is sometimes split among classes (see below). The priority rule among classes

is to always serve the shortest queue as measured in terms of nominal workload. To make this statement precise, some additional notation is required. We say that *buffer i contains the shortest queue* at time t, if

$$0 < X_i^r(t) = \min\{X_j^r(t) : X_j^r(t) > 0,\ j \in [N]\}.$$

When there is exactly one buffer containing the shortest queue, the server serves it at full capacity (thus, service is preemptive). When there is more than one such buffer, the server's effort is split among the buffers containing the shortest queue according to predetermined fractions in a head-of-the-line form. To model these fractions, it is assumed that for any $\varnothing \neq \mathcal{K} \subseteq [N]$ we are given a vector $p^{\mathcal{K}} \in \mathbb{R}_+^{\mathcal{K}}$, such that $\sum_{i\in\mathcal{K}} p_i^{\mathcal{K}} = 1$. When the collection of shortest queues is $\mathcal{K}$, the fraction of effort dedicated to class i is given by $p_i^{\mathcal{K}}$. If we denote by $T_i^r(t)$ the total effort dedicated to class i by time t (measured in units of time), then it is given by

$$T_i^r(t) = \int_0^t p_i^{\mathcal{K}(X^r(s))}\, ds, \tag{2.2}$$

where, for $x \in \mathbb{R}_+^N$, we denote

$$\mathcal{K}(x) = \{i \in [N] : 0 < x_i \le x_j \text{ for all } j \in [N]\}. \tag{2.3}$$

The departure process $D^r = (D_1^r, \dots, D_N^r)$ consists of N counting processes, where for each i, D_i^r gives the number of class-i job completions. It thus satisfies

$$D_i^r(t) = S_i^r\big(T_i^r(t)\big). \tag{2.4}$$

Clearly, Q^r satisfies the balance equation

$$Q_i^r(t) = Q_i^r(0) + A_i^r(t) - D_i^r(t). \tag{2.5}$$

This completes the description of the model. Note that according to this description, the queue length process Q^r is a Markov process on $\mathbb{Z}_+^N$, whereas X^r is a Markov process on

$$\mathcal{S}_{\mathrm{u}}^r = \frac{1}{\mu_1^r}\mathbb{Z}_+ \times \cdots \times \frac{1}{\mu_N^r}\mathbb{Z}_+ \tag{2.6}$$

(where "u" is mnemonic for *unscaled*). Thus an alternative, concise description of the model is via the generator of the process X^r, denoted by $\mathcal{L}_{\mathrm{u}}^r$. It is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{u}}^r f(x) = &\sum_{i\in[N]} \lambda_i^r \left(f\left(x + \frac{e_i}{\mu_i^r}\right) - f(x)\right) \\ &+ \sum_{i\in\mathcal{K}(x)} p_i^{\mathcal{K}(x)} \mu_i^r \left(f\left(x - \frac{e_i}{\mu_i^r}\right) - f(x)\right),\end{aligned} \tag{2.7}$$

for any bounded $f : \mathcal{S}_{\mathrm{u}}^r \to \mathbb{R}$. Note that $\mathcal{K}(0) = \varnothing$ and that, by the assumptions on $p^{\mathcal{K}}$, $p_i^{\mathcal{K}} = 1$ whenever $\mathcal{K}$ consists of the singleton $\{i\}$, $i \in [N]$.

The parameters λ_i^r and μ_i^r are assumed to scale like r^2. The precise assumption is that there exist constants $\lambda_i, \mu_i \in (0,\infty)$ and $\hat{\lambda}_i, \hat{\mu}_i \in \mathbb{R}$, such that for $i \in [N]$, as $r \to \infty$,

$$\begin{aligned} r^{-1}\big(\lambda_i^r - r^2\lambda_i\big) &\to \hat{\lambda}_i, \\ r^{-1}\big(\mu_i^r - r^2\mu_i\big) &\to \hat{\mu}_i. \end{aligned} \tag{2.8}$$

The system is assumed to be critically loaded in the sense that the overall traffic intensity equals 1. This is expressed as a condition on the first-order parameters as follows:

$$\sum_{i\in[N]} \frac{\lambda_i}{\mu_i} = 1. \tag{2.9}$$

Our main result regards rescaled versions of the nominal workload and queue length processes, defined as

$$\hat{X}^r(t) = rX^r(t), \qquad \hat{Q}^r(t) = r^{-1}Q^r(t), \qquad t \in \mathbb{R}_+. \tag{2.10}$$

Both these processes are obtained from Q^r via invertible transformations, and are therefore Markov processes on discrete spaces. The one to which most of the analysis is devoted in this paper is $\hat{X}^r$. Recalling (2.1), it follows that $\hat{X}^r$ is a Markov process with state space

$$\mathcal{S}^r = \frac{r}{\mu_1^r}\mathbb{Z}_+ \times \cdots \times \frac{r}{\mu_N^r}\mathbb{Z}_+. \tag{2.11}$$

Specifically, the jump rates of both X^r and $\hat{Q}^r$ are of order r^2 and their jump sizes are of order r^{-1}, confirming that (2.10) gives the usual heavy traffic scaling.

2.2. *Walsh Brownian motion.* In [27], Walsh introduced a diffusion process in the plane that can informally be described as follows. Let $\xi(t) = (\rho(t), \theta(t))$, $t \in \mathbb{R}_+$, be the representation of the process in polar coordinates. Then the radial part $\rho(t)$ is an RBM, and on each excursion of $\xi(t)$ away from the origin, the angular part $\theta(t)$ remains fixed. Moreover, the constant value which $\theta(t)$ takes on each such excursion has a fixed distribution, independent for the different excursions. The precise definition that we shall work with is the one given by Barlow, Pitman and Yor [1], via its semigroup. However, rather than working with a planar diffusion we work with what is more natural for our purposes, namely a process in $\mathcal{S} := \mathbb{R}_+^N$. Also, it is not necessary for our purposes to consider general angular measures, and so the presentation below is restricted to angular measures supported on the N vectors $\{e_i : i \in [N]\}$.

Let $b \in \mathbb{R}$, $\sigma \in (0,\infty)$, and $q \in \mathcal{M}_1$ be given. Let Π_t^+, $t \in \mathbb{R}_+$ and Π_t^0, $t \in \mathbb{R}_+$ denote the semigroups of a (b,σ)-RBM and a (b,σ)-BM killed at 0, respectively. That is, for $f \in \mathcal{C}_0(\mathbb{R}_+)$,

$$\Pi_t^+ f(x) = \mathbb{E}_x\big[f\big(\rho(t)\big)\big], \qquad \Pi_t^0 f(x) = \mathbb{E}_x\big[f\big(\rho(t)\big)\mathbb{1}_{\{t<\zeta\}}\big], \qquad x \in \mathbb{R}_+,$$

where $\rho(t)$ is a (b,σ)-RBM and ζ denotes its hitting time at zero, and, throughout the paper, $\mathbb{P}_x$ (resp., $\mathbb{E}_x$) denotes the law of ρ with $\rho(0)=x$ (resp., the corresponding expectation). Let S^k denote the k-sphere. Use polar coordinates $(\rho,\theta)\in\mathbb{R}_+\times S^{N-1}$ to denote members $x\in\mathcal{S}$ by setting $\rho=\|x\|$ and $\theta=x/\|x\|$ when $x\neq 0$, $\theta=e_1$ when $x=0$. The semigroup Π_t of a (b,σ,q)-WBM is defined as follows. For $f\in\mathcal{C}_0(\mathcal{S})$, Π_t acts on f as

$$\begin{aligned}\Pi_t f(0,\theta)&=\Pi_t^+\bar f(0),\\ \Pi_t f(\rho,\theta)&=\Pi_t^+\bar f(\rho)+\Pi_t^0(f_\theta-\bar f)(\rho),\end{aligned}\tag{2.12}$$

where we denote

$$\begin{aligned}\bar f(\rho)&=\sum_{i\in[N]} q_i f(\rho,e_i),\qquad \rho\geq 0,\\ f_\theta(\rho)&=f(\rho,\theta),\qquad \rho\geq 0,\theta\in S^{N-1}.\end{aligned}\tag{2.13}$$

It is shown in [1] that Π_t is a Feller semigroup on $\mathcal{C}_0(\mathcal{S})$ and that there exists a strong Markov process $\{\xi(t)\}$ with state space $\mathcal{S}$ and semigroup Π_t, that has a.s.-continuous sample paths. Moreover, this process has the properties alluded to above. More precisely, when written in polar coordinates as $\xi(t)=(\rho(t),\theta(t))$, the radial part $\rho(t)$ is a (b,σ)-RBM and the values that the angular part $\theta(t)$ takes are constant on the interval $[0,\zeta]$ (where the constant is determined by the initial condition θ_0) as well as on each excursion away from zero. These constant values on the excursions away from zero are mutually independent with common distribution $\sum_{i\in[N]} q_i\delta_{e_i}(dx)$, where δ_{e_i} is the Dirac measure at e_i. In this paper, we are interested in the case where the initial condition is supported on $\mathcal{S}_0:=\bigcup_{i\in[N]}\{xe_i : x\in\mathbb{R}_+\}$. Note that in this case, $\xi(t)$ takes values in $\mathcal{S}_0$ for all t.

Throughout, let $\mathbb{P}_x^{\mathrm{wbm}}$ and $\mathbb{E}_x^{\mathrm{wbm}}$ denote the law of ξ for $\xi(0)=x$, and respective expectation. Then relations (2.12) can be expressed, for $x=(\rho_0,\theta_0)$, as

$$\mathbb{E}_x^{\mathrm{wbm}}\big[f\big(\xi(t)\big)\big]=\mathbb{E}_{\rho_0}\big[f\big(\rho(t)\theta_0\big)\mathbb{1}_{\{t<\zeta\}}\big]+\sum_{i\in[N]} q_i\mathbb{E}_0\big[f\big(\rho(t)e_i\big)\mathbb{1}_{\{t\geq\zeta\}}\big].\tag{2.14}$$

2.3. *Main result.* The linear relation between X^r and Q^r, the convergence $r^{-2}\mu_i^r\to\mu_i$ that follows from (2.8), and the rescaling defined in (2.10) imply an asymptotic relation between $\hat X^r$ and $\hat Q^r$ which one can express in terms of the $N\times N$ matrix $\hat M=\mathrm{diag}(\mu_i)_{i\in[N]}$. For example, the statement $\hat X^r(0)\Rightarrow\xi(0)$ is equivalent to the statement $\hat Q^r(0)\Rightarrow\hat M\xi(0)$, as $r\to\infty$, where, throughout, the symbol $\Rightarrow$ denotes convergence in law under $\mathbf{P}$. Denote

$$b=\sum_{i\in[N]}\frac{1}{\mu_i}\Big(\hat\lambda_i-\frac{\lambda_i}{\mu_i}\hat\mu_i\Big),\qquad \sigma^2=2\sum_{i\in[N]}\frac{\lambda_i}{\mu_i^2}.$$

THEOREM 2.1. *There exists $q \in \mathcal{M}_1$ such that, if $\{\xi(t)\}$ is a (b, σ, q)-WBM with initial distribution supported on $\mathcal{S}_0$ and $\hat{X}^r(0) \Rightarrow \xi(0)$ then $\hat{X}^r \Rightarrow \xi$ and $\hat{Q}^r \Rightarrow \hat{M}\xi$, as $r \to \infty$.*

REMARK 2.2. (a) Whereas the coefficients b and σ of the process ξ are given explicitly, our approach does not provide a construction or any explicit information of the angular distribution q. However, this much can be said: q does not depend on the second order parameters $(\hat{\lambda}_i, \hat{\mu}_i)$ (where we use standard terminology by which (λ_i, μ_i) and $(\hat{\lambda}_i, \hat{\mu}_i)$ are called first- and second-order parameters, respectively, due to the fact that in most conventional queueing models, LLN limits depend only on the former, whereas CLT limits are also affected by the latter). This statement is a direct consequence of our results of Section 3.5.

(b) Initial conditions which are not asymptotically concentrated on $\mathcal{S}_0$ are excluded from our treatment. For such initial conditions the asymptotic behavior is expected to follow a jump to $\mathcal{S}_0$ at time zero, and then proceed as a WBM. However, the position to which the process jumps is dictated by properties finer than the limiting initial distribution, to the extent that the limit does not exist in general. For example, for $N = 2$, a sequence of initial conditions may converge to a point on the diagonal in such a way that $\hat{Q}_1^r(0) > \hat{Q}_2^r(0) + \varepsilon_r$. It is not hard to see that, due to even a small advantage $\varepsilon_r > 0$ to $\hat{Q}_2^r(0)$, the limiting process will initially jump to a point on the e_1 axis, provided that ε_r tends to zero sufficiently slowly. Interchanging the roles of $\hat{Q}_1^r(0)$ and $\hat{Q}_2^r(0)$ will result in a jump to the e_2 axis.

3. Proof of the main result. Below we present two central lemmas and one central proposition required to prove our main result. The proof of the main result is presented next, in Section 3.1. The proofs of the lemmas and the proposition are then provided in Sections 3.2–3.5.

Some notation used throughout this section is as follows. We use $\sum_i$ as shorthand notation for $\sum_{i \in [N]}$. For $\varphi \in \mathcal{D}_{\mathbb{R}}[0, \infty)$, let $\Gamma[\varphi] = (\Gamma_1[\varphi], \Gamma_2[\varphi])$ be defined by

$$\begin{aligned} &\big(\Gamma_1[\varphi](t), \Gamma_2[\varphi](t)\big) \\ &\quad = \Big(\varphi(t) - \inf_{s \le t}\big(\varphi(s) \wedge 0\big), -\inf_{s \le t}\big(\varphi(s) \wedge 0\big)\Big), \qquad t \in [0, \infty). \end{aligned} \tag{3.1}$$

The *Skorohod map* Γ just introduced transforms a (b, σ)-BM starting from $x \ge 0$, say, W, into a (b, σ)-RBM starting from the same point, via $R = \Gamma_1[W]$. The process given by $\Gamma_2[W]$ gives the corresponding boundary term.

Let $\hat{R}^r(t) = \mathbb{1} \cdot \hat{X}^r(t)$, $t \in \mathbb{R}_+$, and let ρ be a (b, σ)-RBM. In addition to the notation $\mathbb{P}_x^{\mathrm{wbm}}$ and $\mathbb{P}_x$ introduced above, for each r and $x \in \mathcal{S}^r$, we use $\mathbb{P}_x^r$ and $\mathbb{E}_x^r$ for the law of the Markov process $\hat{X}^r$ with $\hat{X}^r(0) = x$, and the respective expectation. Moreover, for each r and $x \in \mathcal{S}^r$, we use $\mathbf{P}_x^r$ and $\mathbf{E}_x^r$ for the law of the tuple (A^r, S^r, X^r) with $\hat{X}^r(0) = x$, and the respective expectation.

Let $\mathcal{S}_\varepsilon = \{x \in \mathcal{S} : \operatorname{dist}(x, \mathcal{S}_0) < \varepsilon\}$. Finally, denote $\zeta^r = \inf\{t \geq 0 : \hat{R}^r(t) = 0\}$, and for $\varepsilon > 0$,

$$\tau^r(\varepsilon) = \inf\{t \geq 0 : \hat{R}^r(t) \geq \varepsilon\}. \tag{3.2}$$

Both ζ^r and $\tau^r(\varepsilon)$ are easily seen to be a.s. finite.

LEMMA 3.1. (i) *The process $\hat{R}^r$ is given as $\hat{R}^r = \Gamma_1[\hat{R}^r(0) + B^r]$, where B^r decomposes as $B^r = \tilde{B}^r + E^r$. For each r, $\tilde{B}^r$ (resp., E^r) is measurable w.r.t. $\sigma\{A^r(t), S^r(t), t \in \mathbb{R}_+\}$ (resp., $\sigma\{A^r(t), S^r(t), X^r(t), t \in \mathbb{R}_+\}$) and $\tilde{B}^r \Rightarrow B$, where B is a (b, σ)-BM starting from zero, whereas*

$$\lim_{v \downarrow 0} \limsup_{r \to \infty} \sup_{x \in \mathcal{S}^r} \mathbf{P}^r_x\big(\|E^r\|_T > v\big) = 0. \tag{3.3}$$

As a consequence, if $\hat{R}^r(0) \Rightarrow \rho(0)$, then $\hat{R}^r \Rightarrow \rho$.

(ii) *For any $t_0 > 0$,*

$$\lim_{v \downarrow 0} \liminf_{r \to \infty} \inf_{x \in \mathcal{S}^r : \mathbb{1} \cdot x < v} \mathbb{P}^r_x\big(\zeta^r \leq t_0\big) = 1.$$

Throughout, let $\mathcal{U}_0$ denote the class of functions $u : [1, \infty) \to (0, \infty)$ for which $u(r) \to 0$ as $r \to \infty$.

LEMMA 3.2. *The processes $\hat{X}^r$ are $\mathcal{C}$-tight under $\mathbf{P}$. Moreover, let ν^r denote the distribution of $\hat{X}^r(0)$. Then there exists $u \in \mathcal{U}_0$ such that for every $T > 0$ one has the following*:

(i) $\mathbb{P}^r_{\nu^r}(\hat{X}^r(t) \in \mathcal{S}_{u(r)}$ *for all* $t \in [0, T]) \to 1$ *as* $r \to \infty$.

(ii) $\inf \mathbb{P}^r_x(\hat{X}^r(t) \in \mathcal{S}_{u(r)}$ *for all* $t \in [\zeta^r \wedge T, T)) \to 1$ *as* $r \to \infty$, *where the infimum extends over $x \in \mathcal{S}^r \cap K$, and $K \subset \mathcal{S}$ is a given compact set.*

(iii) *For $f \in \mathcal{C}_0(\mathcal{S})$, $t \in \mathbb{R}_+$, $i \in [N]$, and $k \in (0, \infty)$,*

$$\lim_{\delta \downarrow 0} \limsup_{r \to \infty} \sup_{y \in \mathcal{S}^r, x \in [0,k] : \|y - x e_i\| < \delta} \big|\mathbb{E}^r_y\big[f\big(\hat{X}^r(t)\big)\mathbb{1}_{\{t < \zeta^r\}}\big] - \mathbb{E}_x\big[f\big(\rho(t)e_i\big)\mathbb{1}_{\{t < \zeta\}}\big]\big| = 0,$$

where we recall that $\zeta = \inf\{t \geq 0 : \rho(t) = 0\}$.

PROPOSITION 3.3. *There exist $q \in \mathcal{M}_1$ and $u \in \mathcal{U}_0$ such that*

$$\lim_{\varepsilon \downarrow 0} \limsup_{r \to \infty} \big|\mathbb{P}^r_0\big(\hat{X}^r\big(\tau^r(\varepsilon)\big) \in B\big(\varepsilon e_i, u(r)\big)\big) - q_i\big| = 0, \qquad i \in [N]. \tag{3.4}$$

3.1. *Proof of Theorem 2.1.* Given $\varepsilon > 0$, define a sequence of hitting times as

$$\begin{aligned} \zeta_0^r &= \inf\{t \geq 0 : \hat{R}^r(t) = 0\}, \\ \tau_m^r &= \inf\{t \geq \zeta_m^r : \hat{R}^r(t) \geq \varepsilon\}, \qquad m = 0, 1, \ldots, \\ \zeta_{m+1}^r &= \inf\{t \geq \tau_m^r : \hat{R}^r(t) = 0\}, \qquad m = 0, 1, \ldots. \end{aligned} \tag{3.5}$$

Let $N_t^r = \sup\{m : \tau_m^r \leq t\}$. When we need to emphasize the dependence on ε we write these RVs as $\zeta_m^r(\varepsilon)$ and $\tau_m^r(\varepsilon)$.

Let $(\xi(t))_{t\in\mathbb{R}_+}$ be a (b, σ, q)-WBM and assume, without loss of generality, that $\rho = \mathbb{1} \cdot \xi$. For this process, we define an analogous sequence of hitting times by

$$\begin{aligned} \zeta_0 &= \inf\{t \geq 0 : \rho(t) = 0\}, \\ \tau_m &= \inf\{t \geq \zeta_m : \rho(t) \geq \varepsilon\}, \qquad m = 0, 1, \ldots, \\ \zeta_{m+1} &= \inf\{t \geq \tau_m : \rho(t) = 0\}, \qquad m = 0, 1, \ldots, \end{aligned}$$

and set $N_t = \sup\{m : \tau_m \leq t\}$.

The weak convergence stated in Lemma 3.1(i) does not directly imply that of the hitting times $\tau^r(\varepsilon)$ of (3.2) to $\tau(\varepsilon) := \inf\{t \geq 0 : \rho(t) \geq \varepsilon\}$ when both $\hat{R}^r$ and ρ start at zero. However, this convergence is clearly valid, as can be seen by using in addition the property of RBM that $\tau(\varepsilon + \delta) \to \tau(\varepsilon)$ in probability as $\delta \downarrow 0$. Moreover, under **P**, it is assumed in Theorem 2.1 that $\hat{X}^r(0)$ converges to $\xi(0)$ in distribution. An inductive use of this fact yields a similar statement for the stopping times $\{\tau_m^r\}_m$. More precisely, for any fixed m, as $r \to \infty$, we have the following uniform convergence: for any compact set $K \subset \mathcal{S}$ and a function $h \in \mathcal{C}_0(\mathbb{R}_+)$,

$$\limsup_{\delta\downarrow 0} \limsup_{r\to\infty} \sup_{x\in K\cap\mathcal{S}_0} \sup_{\substack{y\in\mathcal{S}^r \\ \|x-y\|<\delta}} \left|\mathbb{E}_y^r\big[h(\tau_m^r)\big] - \mathbb{E}_{\|x\|}\big[h(\tau_m)\big]\right| = 0. \tag{3.6}$$

The proof of the main result is based on finite-dimensional convergence and $\mathcal{C}$-tightness. The key ingredient is showing that for any compact set $K \subset \mathcal{S}$, $t \in \mathbb{R}_+$, and a function $f \in \mathcal{C}_0(\mathcal{S})$,

$$\limsup_{\delta\downarrow 0} \limsup_{r\to\infty} \sup_{x\in K\cap\mathcal{S}_0} \sup_{\substack{y\in\mathcal{S}^r \\ \|x-y\|<\delta}} \left|\mathbb{E}_y^r\big[f\big(\hat{X}^r(t)\big)\big] - \mathbb{E}_x^{\mathrm{wbm}}\big[f\big(\xi(t)\big)\big]\right| = 0. \tag{3.7}$$

Before proving this statement, we show, adapting the proof of Theorem 4.2.5 of [7], that it implies the convergence of $\hat{X}^r$ to ξ for finite-dimensional marginals. That is, for every $x \in \mathcal{S}_0$, $\{x^r\}_r$, $x^r \in \mathcal{S}^r$ that converges to x, $m \geq 1$, $0 \leq t_1 < \cdots < t_m$, and functions $h_1, \ldots, h_m \in \mathcal{C}_0(\mathcal{S})$, one has

$$\lim_{r\to\infty} \mathbb{E}_{x^r}^r\big[h_1\big(\hat{X}^r(t_1)\big)\cdots h_m\big(\hat{X}^r(t_m)\big)\big] = \mathbb{E}_x^{\mathrm{wbm}}\big[h_1\big(\xi(t_1)\big)\cdots h_m\big(\xi(t_m)\big)\big]. \tag{3.8}$$

We argue by induction over m. The base case follows from (3.7). Next, assume that (3.8) holds for m. Denote by Π_t^r the semigroup corresponding to $\{\hat{X}^r(t)\}$. Then by

Lemma 3.2(i), there exists $u \in \mathcal{U}_0$, such that

$$\begin{aligned}
&\mathbb{E}^r_{x^r}\big[h_1\big(\hat{X}^r(t_1)\big)\cdots h_m\big(\hat{X}^r(t_m)\big)\cdot h_{m+1}\big(\hat{X}^r(t_{m+1})\big)\big] \\
&\quad = \mathbb{E}^r_{x^r}\big[h_1\big(\hat{X}^r(t_1)\big)\cdots h_m\big(\hat{X}^r(t_m)\big)\cdot \Pi^r_{t_{m+1}-t_m} h_{m+1}\big(\hat{X}^r(t_m)\big)\big] \\
&\quad = \mathbb{E}^r_{x^r}\big[h_1\big(\hat{X}^r(t_1)\big)\cdots h_m\big(\hat{X}^r(t_m)\big)\cdot \Pi^r_{t_{m+1}-t_m} h_{m+1}\big(\hat{X}^r(t_m)\big)\mathbb{1}_{\{\hat{X}^r(t_m)\in\mathcal{S}_{u(r)}\}}\big] \\
&\qquad + o_r(1),
\end{aligned}$$

where here $o_r(1)$ denotes a generic function of r that vanishes as $r \to \infty$. From (3.7) and the Feller property of Π_t proved in [1], it follows that for $h \in \mathcal{C}_0(\mathcal{S})$, $\sup_{x\in\mathcal{S}^r\cap\mathcal{S}_{u(r)}} |\Pi^r_t h(x) - \Pi_t h(x)| \to 0$ as $r \to \infty$. It follows that the expression in the above display equals

$$\begin{aligned}
&\mathbb{E}^r_{x^r}\big[h_1\big(\hat{X}^r(t_1)\big)\cdots h_m\big(\hat{X}^r(t_m)\big)\cdot \Pi_{t_{m+1}-t_m} h_{m+1}\big(\hat{X}^r(t_m)\big)\mathbb{1}_{\{\hat{X}^r(t_m)\in\mathcal{S}_{u(r)}\}}\big] \\
&\qquad + o_r(1) \\
&\quad = \mathbb{E}^r_{x^r}\big[h_1\big(\hat{X}^r(t_1)\big)\cdots h_m\big(\hat{X}^r(t_m)\big)\cdot \Pi_{t_{m+1}-t_m} h_{m+1}\big(\hat{X}^r(t_m)\big)\big] \\
&\qquad + o_r(1).
\end{aligned} \tag{3.9}$$

By the induction hypothesis, the above expression converges to

$$\begin{aligned}
&\mathbb{E}^{\mathrm{wbm}}_x\big[h_1\big(\xi(t_1)\big)\cdots h_m\big(\xi(t_m)\big)\Pi_{t_{m+1}-t_m} h_{m+1}\big(\xi(t_m)\big)\big] \\
&\quad = \mathbb{E}^{\mathrm{wbm}}_x\big[h_1\big(\xi(t_1)\big)\cdots h_{m+1}\big(\xi(t_{m+1})\big)\big].
\end{aligned}$$

This establishes (3.8). In view of the $\mathcal{C}$-tightness of $\hat{X}^r$ stated in Lemma 3.2, this gives the main result $\hat{X}^r \Rightarrow \xi$.

The rest of the proof is devoted to showing that (3.7) holds. Fix $f \in \mathcal{C}_0(\mathcal{S})$. It suffices to prove the result for $f(0) = 0$. Moreover, arguing by approximation, we may, and will assume that f is constant on a ball about the origin. Thus, there exists $\varepsilon > 0$ for which $f(x)$ vanishes for all x with $\mathbb{1}\cdot x \le \varepsilon$. We fix such ε, and let $\tau^r_m = \tau^r_m(\varepsilon)$, and similarly let ζ^r_m, τ_m and ζ_m be defined in terms of the same value of ε.

Fix $t > 0$ and a compact set $K \subset \mathcal{S}_0$. For $u \in \mathcal{U}_0$, we will be concerned with $x \in K$ and $y^r \in \mathcal{S}^r \cap \mathcal{S}_{u(r)}$ such that $\|x - y^r\| < u(r)$. We call such a pair $(x, (y^r)_r)$ a *u-admissible pair*. In what follows, we denote $y = (y^r)$. Since t and K are arbitrary, to prove (3.7), it suffices to show that $\mathbb{E}^r_{y^r}[f(\hat{X}^r(t))] \to \mathbb{E}^{\mathrm{wbm}}_x[f(\xi(t))]$ uniformly over u-admissible pairs (x, y), for an arbitrary $u \in \mathcal{U}_0$. Fix such a function u. Notice that the assertions in Lemma 3.2(i), (ii), and Proposition 3.3 are all monotone in u in the sense that if they hold for some $u \in \mathcal{U}_0$ then they also hold for a function that dominates u and vanishes at infinity. Hence, without loss of generality, we may and will assume that Lemma 3.2(i), (ii), and Proposition 3.3 hold for the function u that we have fixed.

On the intervals $[\zeta_m^r, \tau_m^r)$, one has $\hat{R}^r(t) \le \varepsilon$. As a consequence,

$$\mathbb{E}^r_{y^r}\big[f\big(\hat{X}^r(t)\big)\big] = F_y^{0,r} + F_y^r,$$

where

$$F_y^{0,r} := \mathbb{E}^r_{y^r}\big[f\big(\hat{X}^r(t)\big)\mathbb{1}_{\{0\le t<\zeta_0^r\}}\big], \qquad F_y^r := \sum_{m=0}^{\infty} \mathbb{E}^r_{y^r}\big[f\big(\hat{X}^r(t)\big)\mathbb{1}_{\{\tau_m^r\le t<\zeta_{m+1}^r\}}\big].$$

The above goal will be achieved once we show that, uniformly over u-admissible pairs,

$$F_y^{0,r} \to \mathbb{E}_x\big[f\big(\rho(t)\theta\big)\mathbb{1}_{\{t<\zeta_0\}}\big] \quad \text{and} \quad F_y^r \to \mathbb{E}_x\big[\bar{f}\big(\rho(t)\big)\mathbb{1}_{\{t\ge\zeta_0\}}\big], \tag{3.10}$$

where we recall the definition of $\bar{f}$ from (2.13), that $\theta = x/\|x\|$ for $x \ne 0$ and $\theta = e_1$ for $x = 0$. Note that the first convergence is stated in Lemma 3.2(iii). Thus in what follows we focus on the term F_y^r. Denote

$$\chi_m^r = \hat{X}^r(\tau_m^r).$$

Recall that the jumps of the (unscaled) queue length process Q^r are of size 1. By the way, the scaled nominal workload process $\hat{X}^r$ is defined, it follows that all the jumps of this process are bounded by cr^{-1}, for some positive constant c. As a consequence, one always has $\varepsilon \le \|\chi_m^r\| \le \varepsilon + cr^{-1}$. Denote $B_i^r = B(\varepsilon e_i, u(r))$. It follows from Lemma 3.2(ii) that

$$\mathbb{P}^r_{y^r}\left(\text{for all } m \le N_t^r,\, \chi_m^r \in \bigcup_i B_i^r\right) \to 1,$$

uniformly over u-admissible pairs. As a result,

$$F_y^r = \sum_i \sum_{m=0}^{\infty} \mathbb{E}^r_{y^r}\big[f\big(\hat{X}^r(t)\big)\mathbb{1}_{\{\tau_m^r\le t<\zeta_{m+1}^r\}}\mathbb{1}_{\{\chi_m^r\in B_i^r\}}\big] + o_r(1),$$

where here and in what follows, $o_r(1)$ is a generic function of r, x and y, that converges to zero as $r \to \infty$, uniformly over u-admissible pairs (x, y).

Next, we truncate the sum over m. The tail $\sum_{m>M}$ can be estimated by $\|f\|_\infty \mathbb{P}^r_{y^r}(N_t^r > M)$. For a fixed initial condition x, the $\mathcal{C}$-tightness of $\hat{R}^r$ gives the tightness of the RVs N_t^r. For arbitrary initial conditions y^r, the strong Markovity reduces the same question to that of tightness of N_t^r when starting at 0. Thus

$$F_y^r = \sum_i \sum_{m=0}^{M} \mathbb{E}^r_{y^r}\big[f\big(\hat{X}^r(t)\big)\mathbb{1}_{\{\tau_m^r\le t<\zeta_{m+1}^r\}}\mathbb{1}_{\{\chi_m^r\in B_i^r\}}\big] + o_{M,r}(1), \tag{3.11}$$

where here and in what follows, $o_{M,r}(1)$ refers to any function g of (x, y, r, M) satisfying $\lim_{M\to\infty} \limsup_{r\to\infty} \sup_{(x,y)\ u\text{-admissible}} |g(x, y, r, M)| = 0$.

Our next step is to use the condition $\chi_m^r \in B_i^r$ included in the (i, m)th term in (3.11), to approximate the expression $f(\hat{X}^r(t))$ therein by $f(\hat{R}^r(t)e_i)$. Note

carefully that it is possible for the process to move from B_i^r to B_j^r, $j \neq i$, without exiting $\mathcal{S}_{u(r)}$ or hitting the origin. Thus we must argue that, given any distinct $i, j \in [N]$ and $t > 0$,

$$\begin{aligned}&\mathbb{P}^r_{y^r}\big(\text{there exists } m \in \{0, \dots, M\} \text{ such that}\\ &\qquad \tau_m^r \le t < \zeta_{m+1}^r,\ \chi_m^r \in B_i^r,\ \hat{R}^r(t) > \varepsilon,\\ &\qquad \|\hat{X}^r(t) - \hat{R}^r(t)e_j\| \le u(r)\big) = o_r(1).\end{aligned} \tag{3.12}$$

The proof of this statement, which we now give, is based on the fact that in order for the process to behave as indicated in (3.12) while remaining within $\mathcal{S}_{u(r)}$, it must reach close to the origin. Since we consider only finitely many m's, it is sufficient to show that for every fixed t, m, and $j \neq i$,

$$\begin{aligned}&\mathbb{P}^r_{y^r}\big(\tau_m^r \le t < \zeta_{m+1}^r,\ \chi_m^r \in B_i^r,\ \hat{R}^r(t) > \varepsilon,\ \|\hat{X}^r(t) - \hat{R}^r(t)e_j\| \le u(r)\big)\\ &\qquad = o_r(1).\end{aligned} \tag{3.13}$$

For every $r \in [1, \infty)$, $m \in \mathbb{N}$, and $j \in [N]$ define

$$\pi_m^r = \pi_m^r[j] = \inf\{s \ge \tau_m^r : \hat{R}^r(s) > \varepsilon,\ \|\hat{X}^r(s) - \hat{R}^r(s)e_j\| \le u(r)\}.$$

The probability from (3.13) is bounded above by $\mathbb{P}^r_{y^r}(\tau_m^r \le t \le \pi_m^r < \zeta_{m+1}^r, \chi_m^r \in B_i^r)$. Under this event, if the process does not leave $\mathcal{S}_{u(r)}$ between times τ_m^r and ζ_m^r, then after time τ_m^r, it must reach close to the origin and not hit the origin prior to reaching a small neighborhood of εe_j. Therefore, the LHS of (3.13) is bounded by

$$\mathbb{P}^r_{y^r}\big(\exists s \in [0, t],\ \hat{X}^r(s) \notin \mathcal{S}_{u(r)}\big) + \mathbb{P}^r_{y^r}\big(\phi_m^r < \pi_m^r < \zeta_{m+1}^r\big),$$

where $\phi_m^r = \inf\{s \ge \tau_m^r : \hat{R}^r(s) \le u(r)\sqrt{N}\}$. From Lemma 3.2(ii), the first term is $o_r(1)$. For every r, let $\{\mathcal{F}_s^r\}$ denote the filtration induced by $\{\hat{X}^r(t)\}$. Then for the second term, using the strong Markov property,

$$\begin{aligned}\mathbb{P}^r_{y^r}\big(\phi_m^r < \pi_m^r < \zeta_{m+1}^r\big) &= \mathbb{E}^r_{y^r}\big[\mathbb{E}^r_{y^r}\big[\mathbb{1}_{\{\phi_m^r < \pi_m^r < \zeta_{m+1}^r\}} \mid \mathcal{F}^r_{\phi_m^r}\big]\big]\\ &= \mathbb{E}^r_{y^r}\big[\psi_j^r\big(\hat{X}^r(\phi_m^r)\big)\big],\end{aligned}$$

where $\psi_j^r(z) = \mathbb{E}^r_z[\mathbb{1}_{\{\hat{\pi}^r < \zeta_0^r\}}]$ and $\hat{\pi}^r = \inf\{s \ge 0 : \hat{R}^r(s) > \varepsilon, \|\hat{X}^r(s) - \hat{R}^r(s)e_j\| \le u(r)\}$. It remains to show that $\lim_{r\to\infty} \sup_{\|z\| \le cu(r)} \psi_j^r(z) = 0$, for $c > 0$ a constant. Now, $\psi_j^r(z) \le \mathbb{E}^r_z[\mathbb{1}_{\{\hat{\tau}^r < \zeta_0^r\}}]$, where $\hat{\tau}^r = \inf\{s \ge 0 : \hat{R}^r(s) \ge \varepsilon\}$. The last term goes to zero, uniformly in $\|z\| \le cu(r)$. This shows (3.12).

Equipped with (3.12), we have from (3.11)

$$\begin{aligned}F_y^r = &\sum_i \sum_{m=0}^M \mathbb{E}^r_{y^r}\big[f\big(\hat{X}^r(t)\big)\mathbb{1}_{\{\tau_m^r \le t < \zeta_{m+1}^r\}}\mathbb{1}_{\{\chi_m^r \in B_i^r\}}\mathbb{1}_{\{\|\hat{X}^r(t) - \hat{R}^r(t)e_i\| \le u(r)\}}\big]\\ &+ o_{M,r}(1).\end{aligned}$$

Thus, using the uniform continuity of f and denoting $f_i(z) = f(ze_i)$, $z \in \mathbb{R}_+$,

$$\begin{aligned} F_y^r &= \sum_i \sum_{m=0}^{M} \mathbb{E}_{y^r}^r \big[f_i\big(\hat{R}^r(t)\big) \mathbb{1}_{\{\tau_m^r \le t < \zeta_{m+1}^r\}} \mathbb{1}_{\{\chi_m^r \in B_i^r\}} \mathbb{1}_{\{\|\hat{X}^r(t) - \hat{R}^r(t) e_i\| \le u(r)\}} \big] \\ &\quad + o_{M,r}(1) \\ &= \sum_i \sum_{m=0}^{M} \mathbb{E}_{y^r}^r \big[f_i\big(\hat{R}^r(t)\big) \mathbb{1}_{\{\tau_m^r \le t < \zeta_{m+1}^r\}} \mathbb{1}_{\{\chi_m^r \in B_i^r\}} \big] + o_{M,r}(1), \end{aligned}$$

again using Lemma 3.2. The (i,m)th term can be written as

$$\mathbb{E}_{y^r}^r \big[\mathbb{E}_{y^r}^r \big[f_i\big(\hat{R}^r(t)\big) \mathbb{1}_{\{\tau_m^r \le t < \zeta_{m+1}^r\}} | \mathcal{F}_{\tau_m^r}^r \big] \mathbb{1}_{\{\chi_m^r \in B_i^r\}} \big].$$

By strong Markovity, the conditional expectation above can be written as $\varphi_i^r(\tau_m^r, \chi_m^r)$, where

$$\varphi_i^r(s,z) = \mathbb{E}_z^r \big[f_i(\hat{R}^r(t-s) \mathbb{1}_{\{\zeta_0^r > t-s\}} \big] \mathbb{1}_{\{s \le t\}}.$$

This gives

$$F_y^r = \sum_i \sum_{m=0}^{M} \mathbb{E}_{y^r}^r \big[\varphi_i^r\big(\tau_m^r, \chi_m^r\big) \mathbb{1}_{\{\chi_m^r \in B_i^r\}} \big] + o_{M,r}(1).$$

Define $\varphi_i(s) = \mathbb{E}_\varepsilon [f_i(\rho_{t-s}) \mathbb{1}_{\{\zeta_0 > t-s\}}] \mathbb{1}_{\{s \le t\}}$. Then by Lemma 3.2(iii) one has $\varphi_i^r(s,z) = \varphi_i(s) + o_r(1)$ for $z \in B_i^r$. Hence

$$F_y^r = \sum_i \sum_{m=0}^{M} \mathbb{E}_{y^r}^r \big[\varphi_i\big(\tau_m^r\big) \mathbb{1}_{\{\chi_m^r \in B_i^r\}} \big] + o_{M,r}(1).$$

It will be shown below that, for fixed (i,m), τ_m^r and χ_m^r are asymptotically independent, in the sense that

$$\mathbb{E}_{y^r}^r \big[\varphi_i\big(\tau_m^r\big) \mathbb{1}_{\{\chi_m^r \in B_i^r\}} \big] = \mathbb{E}_{y^r}^r \big[\varphi_i\big(\tau_m^r\big) \big] q_i + o_r(1). \tag{3.14}$$

Hence

$$F_y^r = \sum_i q_i \sum_{m=0}^{M} \mathbb{E}_{y^r}^r \big[\varphi_i\big(\tau_m^r\big) \big] + o_{M,r}(1).$$

Using (3.6), and a similar argument based on strong Markovity,

$$\begin{aligned} F_y^r &= \sum_i q_i \sum_{m=0}^{M} \mathbb{E}_{\|x\|} \big[\varphi_i(\tau_m) \big] + o_{M,r}(1) \\ &= \sum_i q_i \sum_{m=0}^{M} \mathbb{E}_{\|x\|} \big[f_i\big(\rho(t)\big) \mathbb{1}_{\{\tau_m \le t < \zeta_{m+1}\}} \big] + o_{M,r}(1) \end{aligned}$$

$$= \sum_i q_i \mathbb{E}_{\|x\|}\big[f_i\big(\rho(t)\big)\mathbb{1}_{\{t\geq\zeta_0\}}\big] + o_{M,r}(1)$$

$$= \mathbb{E}_{\|x\|}\big[\bar{f}\big(\rho(t)\big)\mathbb{1}_{\{t\geq\zeta_0\}}\big] + o_{M,r}(1).$$

Thus sending $r\to\infty$, then $M\to\infty$ gives the second statement in (3.10).

It remains to prove (3.14). Since φ_i is continuous on $[0,t]$ and the law of τ_m has no atoms, it suffices to prove that for every $s\in[0,t]$ and (i,m),

$$\mathbb{P}^r_{y^r}\big(\tau^r_m\leq s,\,\chi^r_m\in B^r_i\big)\to\mathbb{P}_{\|x\|}(\tau_m\leq s)q_i, \tag{3.15}$$

uniformly over u-admissible pairs (x,y). Toward showing (3.15), we argue that it suffices to establish this assertion for $y\equiv 0$ and $m=0$. Indeed,

$$\mathbb{P}^r_z\big(\tau^r_m\leq s,\,\chi^r_m\in B^r_i\big) = \mathbb{E}^r_z\big[\mathbb{E}^r_z\big[\tau^r_m\leq s,\,\chi^r_m\in B^r_i\,|\mathcal{F}^r_{\zeta^r_m}\big]\big]$$

$$= \mathbb{E}^r_z\big[\tilde{\varphi}^r\big(\zeta^r_m\big)\big],$$

where we used the fact that $\zeta^r_m<\tau^r_m$, and $\hat{X}^r(\zeta^r_m)=0$, and denoted

$$\tilde{\varphi}^r(s) = \mathbb{P}^r_0\big(\tau^r_0<t-s,\,\hat{X}^r\big(\tau^r_0\big)\in B^r_i\big).$$

Similarly, $\mathbb{P}^r_z(\tau^r_m\leq s)=\mathbb{E}^r_z[\varphi^{*,r}(\zeta^r_m)]$, $\varphi^{*,r}(s)=\mathbb{P}^r_0(\tau^r_0<t-s)\to\mathbb{P}_0(\tau_0<t-s)$. Thus if $\mathbb{P}^r_0(\tau^r_0<t-s,\,\hat{X}^r(\tau^r_0)\in B^r_i)\to q_i\mathbb{P}_0(\tau_0<t-s)$ then we obtain

$$\mathbb{P}^r_{y^r}\big(\tau^r_m\leq s,\,\chi^r_m\in B^r_i\big) - q_i\mathbb{P}^r_{y^r}\big(\tau^r_m\leq s\big)\to 0,$$

and since $\mathbb{P}^r_{y^r}(\tau^r_m\leq s)\to\mathbb{P}_{\|x\|}(\tau_m\leq s)$ uniformly over u-admissible pairs (x,y), (3.15) follows.

To prove (3.15) for $y\equiv 0$ and $m=0$, note that under $\mathbb{P}^r_0$, $\zeta^r_0=0$ and so τ^r_0 is a.s. equal to $\tau^r=\tau^r(\varepsilon)=\inf\{s\geq 0:\hat{R}^r(s)\geq\varepsilon\}$ (see (3.2)). Moreover, χ^r_0 that has been defined as $\hat{X}^r(\tau^r_0)$ is a.s. equal to $\chi^r:=\hat{X}^r(\tau^r)$. Hence we aim now at showing

$$\mathbb{P}^r_0\big(\tau^r\leq s,\,\chi^r\in B^r_i\big)\to\mathbb{P}_0(\tau\leq s)q_i. \tag{3.16}$$

Without loss of generality, we take $i=1$. In addition to the parameter ε, that has been fixed, we introduce a new parameter, $a\in(0,\varepsilon)$, that will play the role of the parameter ε in Proposition 3.3. We introduce several pieces of notation associated with a in a way analogous to those defined in terms of ε. Namely, $B^{r,a}_i=B(ae_i,u(r))$, $\tau^{r,a}=\inf\{s\geq 0:\hat{R}^r(s)\geq a\}$ and $\chi^{r,a}=\hat{X}^r(\tau^{r,a})$. In addition, we let $\nu^{r,a}_i$ denote the probability measures supported on $B^{r,a}_i$, given by $\mathbb{P}^r_0(\chi^{r,a}\in\cdot\,|\chi^{r,a}\in B^{r,a}_i)$.

Let

$$g^r(z) = \mathbb{P}^r_z\big(\tau^r<t,\,\chi^r\in B^r_1\big).$$

We analyze $g^r(0)$ by studying its relation to $g^{r,a}_i:=\int g^r(z)\nu^{r,a}_i(dz)$. First,

$$g^r(0) = \mathbb{E}^r_0\big[\mathbb{E}^r_0\big[\mathbb{1}_{\{\tau^r<t,\chi^r\in B^r_1\}}|\mathcal{F}^r_{\tau^{r,a}}\big]\big] = \mathbb{E}^r_0\big[\psi^r\big(\tau^{r,a},\chi^{r,a}\big)\big], \tag{3.17}$$

where

$$\psi^r(s,z) = \mathbb{P}_z^r\big(\tau^r < t-s, \chi^r \in B_1^r\big).$$

Hence

$$g^r(0) = \mathbb{E}_0^r\big[\psi^r\big(0, \chi^{r,a}\big)\big] + \delta_a^r, \tag{3.18}$$

where

$$|\delta_a^r| \le \mathbb{E}_0^r\big[|\psi^r(\tau^{r,a}, \chi^{r,a}) - \psi^r(0, \chi^{r,a})|\big].$$

We denote by $O_a^r(h(a))$ (resp., $o_a^r(h(a))$) any function g of the tuple (x, y, r, a) satisfying $\limsup_{a\downarrow 0} \limsup_{r\to\infty} \sup_{(x,y)u\text{-admissible}} h(a)^{-1}|g(x,y,r,a)| < \infty$ (resp., $= 0$). We argue that $\delta_a^r = O_a^r(a^2)$. To this end, note that

$$\begin{aligned} 0 &\le \psi^r(0,z) - \psi^r(s,z) = \mathbb{P}_z^r\big(\tau^r < t, \chi^r \in B_1^r\big) - \mathbb{P}_z^r\big(\tau^r < t-s, \chi^r \in B_1^r\big) \\ &\le \mathbb{P}_z^r\big(t-s < \tau^r < t\big). \end{aligned}$$

Now, $\tau^r \Rightarrow \tau$ as $t \to \infty$, uniformly for z in $B(0, \varepsilon/2)$. Moreover, for RBM, denoting the density $\frac{d}{d\theta}\mathbb{P}_\eta(\tau \le \theta)$ by $l(\eta, \theta)$ (where, as before, $\tau = \tau(\varepsilon)$), a uniform bound holds in the form

$$\sup_{\eta\in[0,\varepsilon/2], \theta\in[t/2,\infty)} l(\eta,\theta) < \infty. \tag{3.19}$$

Indeed, an explicit eigenfunction expansion of the density l is given in [15]. Using equations (3.15)–(3.19) of [15], one can directly obtain the bound

$$\sup_{x\in[0,\varepsilon], t\ge t_0} l(x,t) < \infty$$

for any constant $t_0 > 0$. This gives (3.19). Using (3.19) for $s \in [0, t/2]$ and the trivial bound 1 for $s \in [t/2, t]$ gives

$$\sup_{\eta\in[0,\varepsilon/2]} \mathbb{P}_\eta(t-s < \tau < t) \le cs,$$

for some constant c (which may depend on t), for all $s \in [0, t]$. In view of this, $\limsup_r |\delta_a^r| \le c \limsup_r \mathbb{E}_0^r[\tau^{r,a}] \le ca^2$, where the last inequality is standard, and follows by Brownian scaling.

Going back to (3.18) and noting that $\psi^r(0,z) = g^r(z)$, we have $g^r(0) = \mathbb{E}_0^r[g^r(\chi^{r,a})] + O_a^r(a^2)$. Therefore, it follows from Lemma 3.2(ii) that the probability of having $\chi^{r,a} \notin \bigcup_i B_i^{r,a}$ is $o_r(1)$, hence

$$\begin{aligned} g^r(0) &= \sum_i \mathbb{P}_0^r\big(\chi^{r,a} \in B_i^{r,a}\big) g_i^{r,a} + O_a^r(a^2) \\ &= \sum_i q_i^{r,a} g_i^{r,a} + O_a^r(a^2), \end{aligned} \tag{3.20}$$

where by Proposition 3.3, $q_i^{r,a} := \mathbb{P}_0^r(\chi^{r,a} \in B_i^{r,a}) = q_i + o_a^r(1)$ (note that $B_i^{r,a}$ agrees with the ball from Proposition 3.3).

Next, consider initial condition $\nu_i^{r,a}$, for which we can write

$$g_i^{r,a} = \mathbb{E}^r_{\nu_i^{r,a}}\big[g^r\big(\hat{X}^r(0)\big)\big] = \mathbb{E}^r_{\nu_i^{r,a}}\big[\hat{\psi}^r\big(\tilde{\tau}^r, \hat{X}^r\big(\tilde{\tau}^r\big)\big)\big],$$

where $\tilde{\tau}^r = \inf\{s \geq 0 : \hat{R}^r(s) \notin (0, \varepsilon)\}$ and $\hat{\psi}^r(s,z) = \mathbb{P}_z^r(\tau^r < t - s, \chi^r \in B_1^r)$. Now,

$$\hat{\psi}^r(s,0) = \mathbb{P}_0^r\big(\tau^r < t - s, \chi^r \in B_1^r\big),$$

and for every s that satisfies $\rho(s) \geq \varepsilon$,

$$\hat{\psi}^r(s,z) = \mathbb{1}_{\{s \leq t\}} \qquad \text{for } z \in B_1^r \quad \text{and} \quad \hat{\psi}^r(s,z) = 0 \qquad \text{for } z \in B_i^r, i \neq 1.$$

Hence

$$\begin{aligned} g_i^{r,a} &= \mathbb{E}^r_{\nu_i^{r,a}}\big[\mathbb{1}_{\{\hat{R}^r(\tilde{\tau}^r)=0\}}\hat{\psi}^r\big(\tilde{\tau}^r, 0\big)\big] + \mathbb{E}^r_{\nu_i^{r,a}}\big[\mathbb{1}_{\{\hat{R}^r(\tilde{\tau}^r)\geq\varepsilon\}}\hat{\psi}^r\big(\tilde{\tau}^r, \hat{X}^r\big(\tilde{\tau}^r\big)\big)\big] \\ &= \mathbb{P}^r_{\nu_i^{r,a}}\big(\hat{R}^r\big(\tilde{\tau}^r\big) = 0\big)\hat{\psi}^r(0,0) + \mathbb{1}_{\{i=1\}}\mathbb{P}^r_{\nu_i^{r,a}}\big(\hat{R}^r\big(\tilde{\tau}^r\big) \geq \varepsilon, \tilde{\tau}^r < t\big) \\ &\quad + \hat{\delta}_a^r + o_r(1), \end{aligned} \tag{3.21}$$

where

$$|\hat{\delta}_a^r| \leq \mathbb{E}^r_{\nu_i^{r,a}}\big[\mathbb{1}_{\{\hat{R}^r(\tilde{\tau}^r)=0\}}\big|\hat{\psi}^r\big(\tilde{\tau}^r, 0\big) - \hat{\psi}^r(0,0)\big|\big],$$

and we used again the bound (3.12). To further bound $\hat{\delta}_a^r$, note that the argument provided earlier for ψ^r can be used also for $\hat{\psi}^r$, and gives $|\hat{\delta}_a^r| \leq c\mathbb{E}^r_{\nu_i^{r,a}}[\mathbb{1}_{\{\hat{R}^r(\tilde{\tau}^r)=0\}}\tilde{\tau}^r]$. On the indicated event, $\tilde{\tau}^r$ is bounded by the exit time of $\hat{R}^r$ from the interval $(0, 2a)$, the expectation of which is $O_a^r(a^2)$. Hence $\hat{\delta}_a^r = O_a^r(a^2)$.

Next, for the RBM ρ denote analogously $\tilde{\tau} = \inf\{s \geq 0 : \rho(s) \notin (0, \varepsilon)\}$. Denote

$$\beta_1(a) = 1 - \mathbb{P}_a\big(\rho(\tilde{\tau}) = 0\big) = \mathbb{P}_a\big(\rho(\tilde{\tau}) = \varepsilon\big), \qquad \beta_2(a) = \mathbb{P}_a\big(\rho(\tilde{\tau}) = \varepsilon, \tilde{\tau} \leq t\big).$$

Then $\mathbb{P}^r_{\nu_i^{r,a}}(\hat{R}^r(\tilde{\tau}^r) = 0) = 1 - \beta_1(a) + o_r(1)$ for all i. Moreover, $\mathbb{P}^r_{\nu_i^{r,a}}(\hat{R}^r(\tilde{\tau}^r) \geq \varepsilon, \tilde{\tau}^r < t) = \beta_2(a) + o_r(1)$. Hence from (3.20) and (3.21), we obtain

$$\begin{cases} g^r(0) = \sum_i (q_i + o_a^r(1))g_i^{r,a} + O_a^r(a^2), \\ g_1^{r,a} = (1 - \beta_1(a) + o_r(1))g^r(0) + \beta_2(a) + O_a^r(a^2) + o_r(1), \\ g_i^{r,a} = (1 - \beta_1(a) + o_r(1))g^r(0) + O_a^r(a^2) + o_r(1), \qquad i \neq 1. \end{cases}$$

Solving this system of equations gives

$$g^r(0) = \frac{(q_1 + o_a^r(1))\beta_2(a) + O_a^r(a^2)}{\beta_1(a) + o_r(1)}.$$

Denote

$$G_a = \frac{(q_1 + o_a(1))\beta_2(a) + O_a(a^2)}{\beta_1(a)}.$$

In order to show that $\lim_r g^r(0) = q_1\mathbb{P}_0(\tau < t)$ it suffices to show that $\lim_{a\downarrow 0} G_a = q_1\mathbb{P}_0(\tau < t)$. Since it is known for RBM (equivalently, for a 1-dimensional BM) that $\beta_1(a) > ca$ for some constant $c > 0$ and all small a, it suffices to show that $\beta_2(a)/\beta_1(a) \to \mathbb{P}_0(\tau < t)$ as $a \downarrow 0$. To this end, use strong Markovity to write

$$\mathbb{P}_a\big(\rho(\tilde{\tau}) = 0, \tau \le t\big) = \mathbb{E}_a\big[\mathbb{1}_{\{\rho(\tilde{\tau})=0\}}\varphi^{\#}\big(\tilde{\tau}, \rho(\tilde{\tau})\big)\big],$$

$\varphi^{\#}(s, x) = \mathbb{P}_x(\tau \le t - s)$ for $x \in \mathbb{R}_+$. Now, $0 \le \varphi^{\#}(0, 0) - \varphi^{\#}(s, 0) \le cs$, and therefore

$$\big|\mathbb{P}_a\big(\tau \le t, \rho(\tilde{\tau}) = 0\big) - \mathbb{P}_0(\tau \le t)\mathbb{P}_a\big(\rho(\tilde{\tau}) = 0\big)\big| \le c\mathbb{E}_a[\mathbb{1}_{\{\rho(\tilde{\tau})=0\}}\tilde{\tau}] \le ca^2.$$

Hence

$$\begin{aligned}\beta_2(a) &= \mathbb{P}_a(\tau \le t) - \mathbb{P}_a\big(\tau \le t, \rho(\tilde{\tau}) = 0\big)\\ &= \mathbb{P}_a(\tau \le t) - \mathbb{P}_0(\tau \le t)\mathbb{P}_a\big(\rho(\tilde{\tau}) = 0\big) + O\big(a^2\big)\\ &= \mathbb{P}_0(\tau \le t)\beta_1(a) + \delta^{\#}(a) + O\big(a^2\big),\end{aligned}$$

where $\delta^{\#}(a) = \mathbb{P}_a(\tau \le t) - \mathbb{P}_0(\tau \le t)$. If we show that $\delta^{\#}(a) = O(a^2)$ then $\beta_2(a)/\beta_1(a) \to \mathbb{P}_0(\tau \le t)$ as $a \downarrow 0$, and the proof is established.

To show that $\delta^{\#}(a) = O(a^2)$, let $\mathcal{L}$ denote the generator of the process ρ, and $v(x, s) = \mathbb{P}_x(\tau > s)$ for $x \in \mathbb{R}_+$. Then $\mathcal{L}$ is given by $\frac{\sigma^2}{2}\frac{d^2}{dx^2} + b\frac{d}{dx}$, with the Neumann boundary condition at 0 and the Dirichlet boundary condition at ε, and v is a smooth function satisfying

$$\begin{aligned}\partial_s v &= \mathcal{L}v, \qquad x \in (0, \varepsilon), s > 0,\\ \partial_x v(0, s) &= v(\varepsilon, s) = 0, \qquad s > 0,\\ v(x, 0) &= 1, \qquad x \in (0, \varepsilon).\end{aligned}$$

In particular, it is a smooth function satisfying $\partial_x v(0, s) = 0$ and, therefore, $v(x, t) = v(0, t) + O(x^2)$, for t fixed and $x \downarrow 0$. This shows that $\delta^{\#}(a) = O(a^2)$.

3.2. *The total nominal workload process.* In this section, we prove Lemma 3.1. Roughly stated, this lemma asserts that the total nominal workload process converges at diffusion scale to an RBM. This is a well-understood fact for an *arbitrary* nonidling policy. However, for completeness and since the statement of the lemma involves uniform convergence, which is perhaps less standard, we provide a proof.

PROOF OF LEMMA 3.1(i). We start the proof with some notation aimed at describing the scaled nominal workload process in terms of scaled arrival and service processes. Let $\bar{T}_i(t) = \frac{\lambda_i}{\mu_i} t$, and

$$\hat{A}_i^r(t) = r^{-1}\big(A_i^r(t) - \lambda_i^r t\big), \qquad \hat{S}_i^r(t) = r^{-1}\big(S_i^r(t) - \mu_i^r t\big),$$

$$\hat{Y}_i^r(t) = \mu_i^r r^{-1}\big(\bar{T}_i(t) - T_i^r(t)\big), \qquad b_i^r = r^{-1}\left(\lambda_i^r - \frac{\lambda_i}{\mu_i}\mu_i^r\right),$$

for $t \in \mathbb{R}_+$. Then by (2.5),

$$\hat{Q}_i^r(t) = \hat{Q}_i^r(0) + \hat{A}_i^r(t) - \hat{S}_i^r\big(T_i^r(t)\big) + b_i^r t + \hat{Y}_i^r(t), \qquad t \in \mathbb{R}_+. \tag{3.22}$$

Note by (2.1) and (2.10) that $\hat{R}^r = \mathbb{1} \cdot \hat{X}^r = \sum_i \frac{r^2}{\mu_i^r} \hat{Q}_i^r$. If we denote

$$B^r(t) = \sum_i \frac{r^2}{\mu_i^r}\big[\hat{A}_i^r(t) - \hat{S}_i^r\big(T_i^r(t)\big) + b_i^r t\big],$$

$$U^r(t) = \sum_i \frac{r^2}{\mu_i^r}\hat{Y}_i^r(t) = r\big[t - \mathbb{1} \cdot T^r(t)\big],$$

then we have the identity $\hat{R}^r = B^r + U^r$. Moreover, by its definition, $\hat{R}^r$ takes values in $\mathbb{R}_+$. Furthermore, by the nonidling property, the right derivative of $\mathbb{1} \cdot T^r$ at t assumes the value 1 if and only if the system is nonempty at this time, that is, $\hat{R}^r(t) > 0$ (it otherwise assumes the value 0). Consequently, $\int_0^\infty \hat{R}^r(t)\, dU^r(t) = 0$. These three properties imply

$$\big(\hat{R}^r, U^r\big) = \Gamma\big[\hat{R}^r(0) + B^r\big]. \tag{3.23}$$

It follows from expression (3.1) for Γ that, for $t > 0$,

$$\big|\hat{R}^r(t) - \hat{R}^r(0)\big| \le 2\big\|B^r\big\|_t. \tag{3.24}$$

Now, the bound $T_i^r(t) \le t$ for all t gives $\|\hat{S}_i^r \circ T_i^r\|_T \le \|\hat{S}_i^r\|_T$. The quantities r^2/μ_i^r as well as b_i^r converge in view of (2.8). Hence $\|B^r\|_T \le c(\|\hat{A}^r\|_T + \|\hat{S}^r\|_T + 1)$. Therefore,

$$\big\|\hat{R}^r - \hat{R}^r(0)\big\|_T \le c\big(\big\|\hat{A}^r\big\|_T + \big\|\hat{S}^r\big\|_T + 1\big).$$

By the functional central limit theorem for renewal processes (see Theorem 14.6 of [4]), $(\hat{A}^r, \hat{S}^r)$ converge to a BM with drift zero and diffusion matrix Ξ, where

$$\hat{A}^r = \big(\hat{A}_i^r\big)_{i=1}^N, \qquad \hat{S}^r = \big(\hat{S}_i^r\big)_{i=1}^N, \qquad \Xi = \operatorname{diag}\big(\lambda_1^{1/2}, \ldots, \lambda_N^{1/2}, \mu_1^{1/2}, \ldots, \mu_N^{1/2}\big).$$

This implies that $\|\hat{A}^r\|_T + \|\hat{S}^r\|_T$ is a tight sequence of RVs (for each fixed T), and in view of the above bound, so is $\|\hat{R}^r\|_T$.

By the discussion preceding Theorem 2.1, $\hat{X}^r$ and $\hat{Q}^r$ are asymptotically related via the matrix $\hat{M}$. Appealing to (3.22) again and recalling that $\mathbb{1} \cdot \hat{X}^r = \hat{R}^r$, we have

$\|\hat{Y}^r\|_T \le c(\|\hat{A}^r\|_T + \|\hat{S}^r\|_T + 1)$, and we get the tightness of $\|\hat{Y}^r\|_T$ (uniformly in the initial state). In view of the definition of $\hat{Y}^r$, we obtain that for every $T, v > 0$,

$$\limsup_{r\to\infty} \sup_{x\in\mathcal{S}^r} \mathbf{P}^r_x\big(\|T^r - \bar{T}\|_T > v\big) = 0. \tag{3.25}$$

Set

$$\tilde{B}^r(t) = \sum_i \frac{r^2}{\mu^r_i}\big[\hat{A}^r_i(t) - \hat{S}^r_i\big(\bar{T}_i(t)\big) + b_i t\big], \qquad E^r = B^r - \tilde{B}^r.$$

Notice that $\tilde{B}^r$ is measurable w.r.t. $\sigma\{A^r(t), S^r(t), t \in \mathbb{R}_+\}$. Now,

$$\|E^r\|_T \le \sum_i \frac{r^2}{\mu^r_i}\|\hat{S}^r_i \circ T^r_i - \hat{S}^r_i \circ \bar{T}_i\|_T \le c\sum_i w_T\big(\hat{S}^r_i, \theta^r\big),$$

where $\theta^t = \|T^r - \bar{T}\|_T$. The $\mathcal{C}$-tightness of $\hat{S}^r$ along with (3.25) give (3.3). Finally, the convergence in law of $(\hat{A}^r, \hat{S}^r)$ gives $\tilde{B}^r \Rightarrow B$, where B is a (b, σ)-BM.

(ii) Fix $v, t_0 > 0$. For every $x \in \mathcal{S}^r$ with $\mathbb{1}\cdot x < v$, one has by the representation $\hat{R}^r = \Gamma_1[\hat{R}^r(0) + B^r]$, noting that B^r is measurable w.r.t. $\sigma\{A^r, S^r, X^r\}$,

$$\begin{aligned}
\mathbb{P}^r_x(\zeta^r \le t_0) &= \mathbb{P}^r_x\Big(\inf_{t\in[0,t_0]} \hat{R}^r(t) = 0\Big) \\
&= \mathbf{P}^r_x\Big(\hat{R}^r(0) + \inf_{t\in[0,t_0]} B^r(t) \le 0\Big) \\
&\ge \mathbf{P}^r_x\Big(\inf_{t\in[0,t_0]} B^r(t) < -v\Big).
\end{aligned}$$

By part (i) of the lemma, specifically, the convergence of $\tilde{B}^r$ to B (indep of x) and the uniform estimate (3.3) on E^r, it follows that for every $\delta > 0$ and all sufficiently large r and $x \in \mathcal{S}^r$ with $\mathbb{1}\cdot x \le v$, the RHS of the above display is bounded below by $\mathbf{P}(\inf_{t\in[0,t_0]} B(t) < -2v) - \delta$. The last expression does not depend on x and, as B is a BM starting at the origin, converges to $1 - \delta$ as $v \downarrow 0$. Therefore,

$$\liminf_{\delta\downarrow 0} \liminf_{r\to\infty} \inf_{x\in\mathcal{S}^r:\mathbb{1}\cdot x<v} \mathbb{P}^r_x(\zeta^r \le t_0) \ge 1 - \delta.$$

Taking $\delta \downarrow 0$ gives the result. □

3.3. *Estimates on exiting the tubes.* In this section, we develop an estimate on the displacement of the prelimit process $\hat{X}^r$ away from $\mathcal{S}_0$. The main use of this estimate is in the argument provided in Section 3.4. In addition, the statement constitutes a strong form of that of Lemma 3.2(i). Thus at the end of the section we provide a proof of Lemma 3.2 based on this estimate.

The proof is based on a Lyapunov function technique. This function is constructed so that it expresses the total nominal workload in all buffers save the one where queue length is greatest. For a precise definition, we need some notation.

Recall from (2.3) the sets $\mathcal{K}(x)$, and note that $\mathcal{K}(\hat{X}^r(t))$ gives the set of shortest nonempty queues at time t. For $x \in \mathbb{R}_+^N$, let

$$\mathcal{M}(x) = \{i \in [N] : x_i \geq x_j \text{ for all } j \in [N]\}.$$

Then $\mathcal{M}(\hat{X}^r)$ gives the set of longest queues. Let $F : \mathbb{R}_+^N \to \mathbb{R}$ be given by

$$F(x) = \sum_i x_i - \max_i x_i. \tag{3.26}$$

Note that F is nonnegative and vanishes on the set $\mathcal{S}_0$ and only there.

LEMMA 3.4. *Given $c_0 > 0$, $\kappa_0 \in (0, 1/2)$ and $0 < \gamma_1 < \gamma_2 < \infty$, there exist constants $r_0, c_1 > 0$ such that for every $r > r_0$ and every initial state $x \in \mathcal{S}^r$ that satisfies $F(x) \leq \gamma_1 r^{-\kappa_0}$,*

$$\mathbb{P}_x^r\big(\|F(\hat{X}^r(\cdot))\|_{c_0 \log r} > \gamma_2 r^{-\kappa_0}\big) \leq r^{-c_1}. \tag{3.27}$$

Lemma 3.4 and the first item of Lemma 3.2 are similar, where the former is concerned with long time intervals as well as rates of convergence. However, the latter is not an immediate consequence of the former. We present their proofs together.

PROOF OF LEMMA 3.4 AND LEMMA 3.2(i). For the proof of Lemma 3.4, fix c_0, κ_0, γ_1 and γ_2 as in the statement of the lemma. Using the expression (2.7) for the generator of X^r write the one for $\hat{X}^r = rX^r$ (see (2.10)), as

$$\begin{aligned}\mathcal{L}^r f(x) = &\sum_i \lambda_i^r \left(f\left(x + \frac{r}{\mu_i^r} e_i\right) - f(x)\right) \\ &+ \sum_{i \in \mathcal{K}(x)} p_i^{\mathcal{K}(x)} \mu_i^r \left(f\left(x - \frac{r}{\mu_i^r} e_i\right) - f(x)\right),\end{aligned} \tag{3.28}$$

for bounded $f : \mathcal{S}^r \to \mathbb{R}$.

Recall that c denotes a generic positive constant that does not depend on r. We begin by showing that there exists a constant c such that for all r sufficiently large,

$$\mathcal{L}^r F(x) < -cr \qquad \text{for all } x \text{ such that } F(x) > 0. \tag{3.29}$$

To this end, note that the first term on the RHS of (3.28), upon substituting F for f, equals

$$\begin{aligned}&\sum_i \lambda_i^r \left(\frac{r}{\mu_i^r} + \max\{x_1, \ldots, x_N\} - \max\{x_1, \ldots, x_{i-1}, x_i + r/\mu_i^r, x_{i+1}, \ldots, x_N\}\right) \\ &\quad \leq r \sum_{i \in [N] \setminus \mathcal{M}(x)} \frac{\lambda_i^r}{\mu_i^r}.\end{aligned}$$

The inequality above is valid since for $i \in \mathcal{M}(x)$ the ith term in the sum is zero, and for $i \notin \mathcal{M}(x)$,

$$\max\{x_1, \dots, x_N\} \le \max\{x_1, \dots, x_{i-1}, x_i + r/\mu_i^r, x_{i+1} \dots, x_N\}.$$

The second term on the RHS of (3.28) (with $f = F$) can be expressed as

$$\begin{aligned} \sum_{i \in \mathcal{K}(x)} p_i^{\mathcal{K}(x)} \mu_i^r \Big(& -\frac{r}{\mu_i^r} + \max\{x_1, \dots, x_N\} \\ & - \max\{x_1, \dots, x_{i-1}, x_i - r/\mu_i^r, x_{i+1}, \dots, x_N\} \Big). \end{aligned} \tag{3.30}$$

We argue that for $i \in \mathcal{K}(x)$,

$$\max\{x_1, \dots, x_N\} = \max\{x_1, \dots, x_{i-1}, x_i - r/\mu_i^r, x_{i+1} \dots, x_N\}.$$

If $\mathcal{K}(x) = \mathcal{M}(x) = \{x_j\}$ for some $j \in [N]$, then $F(x) = 0$. Therefore, if $F(x) > 0$ then either $\mathcal{K}(x) \neq \mathcal{M}(x)$ or $\mathcal{K}(x) = \mathcal{M}(x)$ and $\mathcal{K}(x)$ contains more than one element. In both cases, for every $i \in \mathcal{K}(x)$ there is $j \in \mathcal{M}(x)$, different from i, such that both maxima above equal x_j. This shows that the expression in (3.30) equals $-r$. Combining this with the bound on the first term, and recalling that λ_i^r/μ_i^r is asymptotic to λ_i/μ_i, and that the latter fractions sum to 1, shows (3.29).

We analyze the event $\Omega^r := \{\|F(\hat{X}^r(\cdot))\|_{c_0 \log r} > \gamma_2 r^{-\kappa_0}\}$ under $\mathbb{P}_x^r$ for x such that $F(x) \le \gamma_1 r^{-\kappa_0}$. Recall that the jump sizes of $\hat{X}^r$ are at the scale of r^{-1}; as a result, the same is true for the process $F(\hat{X}^r(\cdot))$. Since $\kappa_0 < 1/2$, $r^{-\kappa_0}$ is at a larger scale than these jumps. Hence there exist random times $0 \le \theta_1^r < \theta_2^r \le c_0 \log r$ such that $\mathbb{P}_x^r$-a.s. on Ω^r,

$$\begin{aligned} F\big(\hat{X}^r(\theta_1^r)\big) \le \gamma_1 r^{-\kappa_0}, \qquad F\big(\hat{X}^r(\theta_2^r)\big) \ge \gamma_2 r^{-\kappa_0} \quad \text{and} \\ 0 < F\big(\hat{X}^r(t)\big) \le \gamma_2 r^{-\kappa_0}, \qquad t \in \big[\theta_1^r, \theta_2^r\big). \end{aligned} \tag{3.31}$$

The process

$$M^r(t) = F\big(\hat{X}^r(t)\big) - F(x) - \int_0^t \mathcal{L}^r F\big(\hat{X}^r(s)\big)\, ds, \qquad t \in \mathbb{R}_+, \tag{3.32}$$

is a local martingale. From (3.29) and (3.31), denoting $\delta = \gamma_2 - \gamma_1 > 0$, one has

$$M^r\big(\theta_2^r\big) - M^r\big(\theta_1^r\big) \ge c r\big(\theta_2^r - \theta_1^r\big) + \delta r^{-\kappa_0}. \tag{3.33}$$

Fix a constant $d \in (2\kappa_0, 1)$ and consider the events

$$\Omega_1^r = \big\{\theta_2^r - \theta_1^r \le r^{-d} \text{ and (3.33) holds}\big\},$$

$$\Omega_2^r = \big\{\theta_2^r - \theta_1^r > r^{-d} \text{ and (3.33) holds}\big\}.$$

Then $\mathbb{P}_x^r(\Omega^r) \le \mathbb{P}_x^r(\Omega_1^r) + \mathbb{P}_x^r(\Omega_2^r)$. We argue separately for the two events.

The event Ω_1^r. Let intervals I_j^r be defined by $I_j^r = [jr^{-d}/2, (j+1)r^{-d}/2]$ for $j \in \{0, 1, \ldots, j_1(r)\}$, where $j_1(r) = \lfloor 2c_0 r^d \log r \rfloor$. On Ω_1^r there must exist j and an interval $J \subset I_j^r \cap (\theta_1^r, \theta_2^r)$ such that

$$\operatorname{osc}_J M^r \geq \delta r^{-\kappa_0}/3,$$

where here and throughout, $\operatorname{osc}_A f = \sup_A f - \inf_A f$. As a result, $\operatorname{osc}_{I_j^r} M^r \geq \delta r^{-\kappa_0}/3$. Therefore, using the Burkholder–Davis–Gundy (BDG) inequality [22], Theorem 48, and denoting by $[M^r]_I$ the quadratic variation of M^r over an interval I,

$$\mathbb{P}_x^r(\Omega_1^r) \leq \sum_{j=0}^{j_1(r)} \mathbb{P}_x^r\left(\operatorname{osc}_{I_j^r} M^r \geq \frac{\delta r^{-\kappa_0}}{3}\right) \leq C_k \sum_{j=0}^{j_1(r)} \left(\frac{6r^{\kappa_0}}{\delta}\right)^{2k} \mathbb{E}_x^r\{[M^r]_{I_j^r}^k\}, \tag{3.34}$$

where k is any number in $[1/2, \infty)$. The quadratic variation process $[M^r]$ has piecewise-constant samples paths with jumps taking values in the set $\{(r/\mu_1^r)^2, \ldots, (r/\mu_N^r)^2\}$. The number of its jumps in an interval $[s, t]$ is stochastically dominated by a Poisson RV with parameter $(t-s)\sum_i(\mu_i^r + \lambda_i^r)$. Since μ_i^r and λ_i^r scale like r^2, $[M^r]_{I_j^r}$ is stochastically dominated by $K^r = cr^{-2}\pi^r$, where π^r is a Poisson RV with parameter cr^{2-d}. Consequently,

$$\mathbb{E}_x^r\{[M^i r]_{I_j^r}^k\} \leq cr^{-dk}.$$

Therefore, the right-hand side of (3.34) is bounded above by $cr^{d-k(d-2\kappa_0)}\log r$ (where c may depend on k). Taking $k > \max\{1/2, d/(d-2\kappa_0)\}$ gives the bound $\mathbb{P}_x^r(\Omega_1^r) \leq r^{-c}$, provided that r is sufficiently large.

The event Ω_2^r. Clearly,

$$\mathbb{P}_x^r(\Omega_2^r) \leq \mathbb{P}_x^r\big(\operatorname{osc}_{[0, c_0 \log r]} M^r \geq cr^{1-d}\big).$$

Using again the BDG inequality (with $k = 1/2$) followed by a domination of the number of jumps in terms of a Poisson RV with parameter $c_0 r^2 \log r$, and the sizes of the jumps by r^{-2}, gives

$$\begin{aligned}\mathbb{P}_x^r(\Omega_2^r) &\leq \mathbb{P}_x^r\big(2\|M^r\|_{c_0 \log r} \geq cr^{1-d}\big) \\ &\leq c\frac{\mathbb{E}_x^r\{[M^r]_{c_0 \log r}\}}{r^{2(1-d)}} \leq \frac{c\log r}{r^{2(1-d)}} < r^{-c}.\end{aligned}$$

This completes the proof of Lemma 3.4.

In order to establish the proof of Lemma 3.2(i) we prove below the following stronger result that also serves us in the proof of Lemma 3.2(iii): for every $u \in U_0$ satisfying $\lim_{r\to\infty} r(u(r))^3 = \infty$, one has

$$\liminf_{r\to\infty} \inf_{x \in \mathcal{S}^r \cap \mathcal{S}_{u(r)/2}} \mathbb{P}_x^r\big(\hat{X}^r(t) \in \mathcal{S}_{u(r)} \text{ for all } t \in [0, T]\big) = 1. \tag{3.35}$$

This statement implies Lemma 3.2(i), since by the assumption on ν^r, there exists $u \in \mathcal{U}_0$, such that $\nu^r(\hat{X}^r(0) \in \mathcal{S}_{u(r)/2}) \to 1$; without loss of generality, we may assume that $\lim_{r\to\infty} r(u(r))^3 = \infty$.

We next show how the details of the above proof are modified in order to prove (3.35). Fix an arbitrary $u \in \mathcal{U}_0$ that satisfies $\lim_{r\to\infty} r(u(r))^3 = \infty$. We claim that

$$\limsup_{r\to\infty} \sup_{x\in\mathcal{S}^r\cap\mathcal{S}_{u(r)/2}} \mathbb{P}^r_x\big(\|F(\hat{X}^r(\cdot))\|_T > u(r)\big) = 0.$$

Unlike in (3.27), we consider here a fixed horizon T, we do not provide a convergence rate, and the polynomial tube widths are replaced by $u(r)$.

Recall (3.28), (3.29) and (3.32). We analyze the event $\bar{\Omega}^r := \{\|F(\hat{X}^r(\cdot))\|_T > u(r)\}$ under $\mathbb{P}^r_x$. Since $ru(r) \to \infty$ and the jump sizes of the process $F(\hat{X}^r(\cdot))$ are at the scale of r^{-1}, there must exist random times $0 \le \bar{\theta}^r_1 < \bar{\theta}^r_2 \le T$ such that $\mathbb{P}^r_x$-a.s. on $\bar{\Omega}^r$,

$$\begin{aligned} F\big(\hat{X}^r(\bar{\theta}^r_1)\big) \le \frac{1}{2}u(r), \qquad F\big(\hat{X}^r(\bar{\theta}^r_2)\big) \ge u(r) \quad \text{and} \\ 0 < F\big(\hat{X}^r(t)\big) \le u(r), \qquad t \in [\bar{\theta}^r_1, \bar{\theta}^r_2). \end{aligned} \tag{3.36}$$

From (3.29) and (3.36), one has

$$M^r(\bar{\theta}^r_2) - M^r(\bar{\theta}^r_1) \ge cr(\bar{\theta}^r_2 - \bar{\theta}^r_1) + \frac{1}{2}u(r). \tag{3.37}$$

Set $u_0(r) = (u(r))^3$ and consider the events

$$\begin{aligned} \bar{\Omega}^r_1 &= \{\bar{\theta}^r_2 - \bar{\theta}^r_1 \le u_0(r) \text{ and (3.37) holds}\}, \\ \bar{\Omega}^r_2 &= \{\bar{\theta}^r_2 - \bar{\theta}^r_1 > u_0(r) \text{ and (3.37) holds}\}. \end{aligned}$$

Then $\mathbb{P}^r_x(\bar{\Omega}^r) \le \mathbb{P}^r_x(\bar{\Omega}^r_1) + \mathbb{P}^r_x(\bar{\Omega}^r_2)$. We argue separately for the two events.

The event $\bar{\Omega}^r_1$. Let intervals $\bar{I}^r_j$ be defined by

$$\bar{I}^r_j = \big[ju_0(r)/2, (j+1)u_0(r)/2\big]$$

for $j \in \{0, 1, \dots, \lfloor 2T/u_0(r)\rfloor\}$. The same arguments given before with the choice of $k = 2$ in BDG inequality lead to the following sequence of inequalities and the uniform limit over $\mathcal{S}^r \cap \mathcal{S}_{u(r)/2}$:

$$\begin{aligned} \mathbb{P}^r_x(\Omega^r_1) &\le \sum_{j=0}^{j_1(r)} \mathbb{P}^r_x\Big(\operatorname{osc}_{I^r_j} M^r \ge \frac{u(r)}{6}\Big) \le C_k \sum_{j=0}^{j_1(r)} \Big(\frac{12}{u(r)}\Big)^4 \mathbb{E}^r_x\{[M^r]^2_{I^r_j}\} \\ &\le \sum_{j=0}^{j_1(r)} \Big(\frac{12}{u(r)}\Big)^4 \big(u_0(r)\big)^2 \le cu^2(r) \to 0. \end{aligned} \tag{3.38}$$

The event $\bar{\Omega}_2^r$. Arguing as before, we obtain

$$\mathbb{P}_x^r(\Omega_2^r) \le \mathbb{P}_x^r\big(2\|M^r\|_T \ge cru_0(r)\big) \le c\frac{\mathbb{E}_x^r\{[M^r]_T\}}{(ru_0(r))^2} \le \frac{c}{(ru_0(r))^2}.$$

By our choice of the function u, the last expression converges to 0 as $r \to \infty$, uniformly for $x \in \mathcal{S}^r \cap \mathcal{S}_{u(r)/2}$. $\square$

PROOF OF LEMMA 3.2 (CONTINUED). First, the assertion regarding $\mathcal{C}$-tightness follows directly from Lemma 3.1(i) and Lemma 3.2(i).

(ii) The statement of this part follows from part (i) with initial condition 0 and strong Markovity.

(iii) It is sufficient to show that for every $u \in \mathcal{U}_0$ satisfying $\lim_{r\to\infty} r(u(r))^3 = \infty$, one has

$$\limsup_{r\to\infty} \sup_{y\in\mathcal{S}^r, x\in[0,k]:\|y-xe_i\|<u(r)/2} \big|\mathbb{E}_y^r\big[f\big(\hat{X}^r(t)\big)\mathbb{1}_{\{t<\zeta^r\}}\big] - \mathbb{E}_x\big[f\big(\rho(t)e_i\big)\mathbb{1}_{\{t<\zeta\}}\big]\big| = 0. \tag{3.39}$$

We first show that for every such u and every $\varepsilon > 0$,

$$\limsup_{r\to\infty} \sup_{y\in\mathcal{S}^r, x\in[0,k]:\|y-xe_i\|<u(r)/2} \mathbb{P}_y^r\big(\zeta^r > t, \hat{R}^r(t) > \varepsilon, \big\|\hat{X}^r(t) - \hat{X}_i^r(t)e_i\big\| > u(r)\big) = 0.$$

Indeed, in order for the process $\hat{X}^r$, starting inside the $u(r)/2$-tube around axis i, to exit the $u(r)$-tube around the same axis by time t and reach $c\varepsilon$ away from the origin, it must either escape $\mathcal{S}_{u(r)}$ before t or pass through a $cu(r)$-neighborhood of the origin without hitting the origin and then move through a different tube and $c\varepsilon$ away from the origin. The probabilities of these two events converge to zero uniformly in the initial conditions; the first convergence follows by (3.35). The second event can be expressed in terms of an atypical behavior of $\hat{R}^r$, as a sequence of processes converging in law to an RBM.

The assertion above along with the uniform continuity of f and a further application of (3.35), imply

$$\limsup_{r\to\infty} \sup_{y\in\mathcal{S}^r, x\in[0,k]:\|y-xe_i\|<u(r)/2} \big|\mathbb{E}_y^r\big[f\big(\hat{X}^r(t)\big)\mathbb{1}_{\{t<\zeta^r\}}\big] - \mathbb{E}_y^r\big[f\big(\hat{R}^r(t)e_i\big)\mathbb{1}_{\{t<\zeta^r\}}\big]\big| = 0.$$

Finally, the statement in Lemma 3.1(i), according to which $\hat{R}^r \Rightarrow \rho$ holds with an error term that converges to zero uniformly in the initial conditions, (3.39) follows, hence the result. $\square$

3.4. *The small ball exit measure.* This section and the next are devoted to the proof of Proposition 3.3. By assumption, the parameters λ_i^r and μ_i^r scale like r^2, as expressed in equation (2.8). The special case where, for all r, $\lambda_i^r = \lambda_i r^2$ and $\mu_i^r = \mu_i r^2$ is referred to as *the homogeneous case*. Our strategy is to first prove the lemma in the homogeneous case, where the processes $\hat{X}^r$ can all be expressed as scaled versions of a *single process*. This is the content of this section. In §3.5, the general case is considered, and by appealing to a change of measure argument, Proposition 3.3 is proved.

LEMMA 3.5. *The statement of Proposition 3.3 holds in the homogeneous case, with* $u(r) = r^{-\kappa_0}$, *for any* $\kappa_0 \in (0, 1/2)$.

PROOF. Let $2N$ mutually independent Poisson processes A_i, S_i, be given, with intensities λ_i and μ_i, respectively. Since by assumption $\mu_i^r = \mu_i r^2$ and $\lambda_i^r = \lambda_i r^2$, the tuple (A_i^r, S_i^r) is equal in law, for each r, to $(A_i(r^2\cdot), S_i(r^2\cdot))$, and without loss of generality we may, and will, assume that $A_i^r(t) = A_i(r^2 t)$ and $S_i^r(t) = S_i(r^2 t)$ for all r, i and t. Let now Q, X, T and D be defined as the processes Q^1, X^1, T^1 and respectively X^1 (that is, $Q = Q^r$ where one sets $r = 1$). Then in particular, equations (2.1), (2.2) and (2.5) are satisfied by Q, X, T and D, and X is a Markov process on $\mathcal{S}_{\mathrm{u}}^1$ (see (2.6)).

Since for each r we have the aforementioned relation between (A_i^r, S_i^r) and (A_i, S_i), one can also express (Q^r, X^r, T^r, D^r) as certain path transformations of (Q, X, T, D), for each r. The most significant aspect of this in the proof is that the rescaled processes $\hat{X}^r$ and $\hat{R}^r$ can be written as rescaled versions of a single process. Denote $R = \mathbb{1} \cdot X$. Then by (2.1), $X_i = (\mu_i)^{-1} Q_i$, whereas $X_i^r = (r^2\mu_i)^{-1} Q_i^r$. Hence $X^r = r^{-2} X(r^2\cdot)$, and thus by (2.10),

$$\hat{X}^r(t) = r^{-1} X\big(r^2 t\big), \qquad \hat{R}^r(t) = r^{-1} R\big(r^2 t\big). \tag{3.40}$$

The state space $\mathcal{S}^r$ for the Markov process $\hat{X}^r$ is given in the case under consideration as $\mathcal{S}^r = r^{-1}\mathcal{S}_{\mathrm{u}}^1 = (r\mu_1)^{-1}\mathbb{Z}_+ \times \cdots \times (r\mu_N)^{-1}\mathbb{Z}_+$.

For $\varepsilon > 0$, set

$$\tau(\varepsilon) = \inf\big\{t \geq 0 : R(t) \geq \varepsilon\big\}. \tag{3.41}$$

Clearly, we have the identity $\tau^r(\varepsilon) = r^{-2}\tau(\varepsilon r)$ (see (3.2)). Let κ_0 be as in the statement of the lemma, that is, $\kappa_0 \in (0, \frac{1}{2})$, and set $\kappa = 1 - \kappa_0 \in (\frac{1}{2}, 1)$. Denote

$$B_i^r = B\big(re_i, r^\kappa\big)$$

and $q^r = (q_i^r)_i$, where

$$q_i^r = \mathbb{P}_0^1\big(X\big(\tau(r)\big) \in B_i^r\big),$$

where $\mathbb{P}_x^1$ (with the corresponding expectation $\mathbb{E}_x^1$) stands for the law of $X = X^1$ with $X(0) = x$. Then

$$\begin{aligned}\mathbb{P}_0^r\big(\hat{X}^r(\tau^r(\varepsilon)) \in B\big(\varepsilon e_i, r^{-\kappa_0}\big)\big) &= \mathbb{P}_0^1\big(X\big(\tau(\varepsilon r)\big) \in B\big(\varepsilon r e_i, r^{1-\kappa_0}\big)\big)\\ &\geq \mathbb{P}_0^1\big(X\big(\tau(\varepsilon r)\big) \in B\big(\varepsilon r e_i, (\varepsilon r)^{1-\kappa_0}\big)\big)\\ &\geq \mathbb{P}_0^1\big(X\big(\tau(\varepsilon r)\big) \in B_i^{\varepsilon r}\big)\\ &= q_i^{\varepsilon r}.\end{aligned} \tag{3.42}$$

Using the fact that the balls $B(\varepsilon e_i, r^{-\kappa_0})$ are disjoint for each ε and sufficiently large r, (3.4) will follow once we show that

$$\text{there exists } q \in \mathcal{M}_1 \text{ such that } \lim_{r\to\infty} q^r = q. \tag{3.43}$$

Note that Proposition 3.3 asserts, moreover, that q does not depend on the choice of κ_0. To address this point, consider $0 < \kappa_0 < \kappa_0' < \frac{1}{2}$ for which q and, respectively q', satisfy (3.43). Then the fact that the LHS of (3.42) is monotone decreasing in κ_0 gives $q_i \geq q_i'$ for all i, and since q and q' are members of $\mathcal{M}_1$, this shows that $q = q'$. Hence the proof will be complete once we show (3.43) for fixed κ_0.

To this end, note that it suffices to show that there exist $\delta \in (0, 1)$ and $K > 0$ such that for every $k \in \mathbb{N}$, $k \geq K$, one has

$$\big|q_i^r - q_i^m\big| \leq \delta^k \qquad \text{for all } i \in [N] \text{ and } r \in \big[2^k, 2^{k+1}\big], \text{ where } m = 2^{k+2} \tag{3.44}$$

and

$$\lim_{r\to\infty} \sum_i q_i^r = 1. \tag{3.45}$$

Indeed, if r and m are both within $[2^k, 2^{k+1}]$ then (3.44) gives $|q_i^r - q_i^m| \leq 2\delta^k$. As a result, for arbitrary $r < m$, denoting $a(\ell) = \lfloor \log_2(\ell) \rfloor$,

$$\big|q_i^r - q_i^m\big| \leq \sum_{j=a(r)}^{a(m)} 2\delta^j \leq 2(1-\delta)^{-1}\delta^{a(r)}.$$

This shows that, for fixed i, any sequence $\{q_i^r\}_r$ is a Cauchy sequence as $r \to \infty$. Along with (3.45), we obtain that (3.43) holds.

In what follows, we prove (3.44) and (3.45). We let $r \in [2^k, 2^{k+1}]$ and $m = 2^{k+2}$, where k is arbitrary, but fixed. Without loss of generality, the proof of (3.44) considers only $i = 1$.

To help explain the main idea and motivate a couple of technical tools, we first consider a highly simplified model, illustrated in Figure 1. Consider a discrete time Markov process on a finite set S that is star shaped. That is, S consists of $2N + 1$ states, denoted by 0, F_i, G_i, $i \in [N]$. For each i, F_i communicates only with 0 and G_i. The state 0 communicates only with the states F_i, while G_i are absorbing. Denoting transition probabilities by $p(s, s')$, we have $p(0, F_i) > 0$, $p(F_i, 0) > 0$,

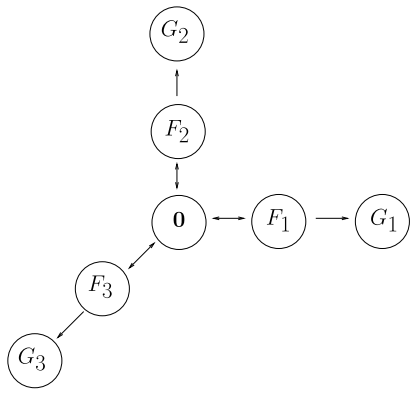

FIG. 1. *In this toy model, denoting transition probabilities by* $p(\cdot,\cdot)$, *the quantity* $\max_{i,j} |p(F_i,0) - p(F_j,0)|$ *controls* $\max_i |p(0,F_i) - P_0(\text{the process is absorbed at } G_i)|$.

$p(F_i, G_i) > 0$, and $p(G_i, G_i) = 1$, while all other transition probabilities are zero. For $s \in S$, let $\bar{p}(s, G_i)$ denote the probability to get absorbed at G_i starting from s. Then

$$\begin{aligned}\bar{p}(0,G_1) &= \sum_{i\geq 1} p(0,F_i)\bar{p}(F_i,G_1) \\ &= p(0,F_1)p(F_1,G_1) + \sum_{i\geq 1} p(0,F_i)p(F_i,0)\bar{p}(0,G_1).\end{aligned}$$

From this, one obtains

$$\frac{\bar{p}(0,G_1)}{p(0,F_1)} = \frac{1-p(F_1,0)}{1-\sum_{i\geq 1} p(0,F_i)p(F_i,0)}.$$

If the transition probabilities starting at F_i depend weakly on i, in the sense that for some $\delta > 0$ one has $|p(F_i,0) - p(F_1,0)| < \delta$ for all i, and if in addition $1 - p(F_1,0) > c > 0$ for some constant c, it follows that

$$\left|\bar{p}(0,G_1) - p(0,F_1)\right| \leq c'\delta,$$

where c' depends on c but not on δ. The relevance to our problem is as follows. Roughly speaking, the states F_i and G_i represent the collections of states within B_i^r and B_i^m, respectively, and 0 represents the origin. The calculation above suggests that if the probability of reaching 0 before reaching B_i^m starting anywhere in B_i^r depends weakly on i then the difference $p_i^r - p_i^m$ is small.

We now consider the process X and the stopping times $\tau(\varepsilon)$, and in addition let

$$\zeta = \inf\{t \geq 0 : X(t) = 0\}.$$

We aim at showing there exist $r_0, c > 0$ such that for every $r > r_0$, one has

$$\text{for every } i \in [N] \text{ and } x \in B_i^r, \qquad \left|\mathbb{P}_x^1\big(\zeta < \tau(m)\big) - \frac{m-r}{m}\right| \leq r^{-c}, \tag{3.46}$$

$$\mathbb{P}_0^1\big(r^{-2}\tau(r) > t\big) \le e^{-ct}, \tag{3.47}$$

$$\mathbb{P}_0^1\Big(X\big(\tau(r)\big) \notin \bigcup_i B_i^r\Big) \le r^{-c}, \tag{3.48}$$

$$\text{for every } i \in [N] \text{ and } x \in B_i^r, \qquad \mathbb{P}_x^1\big(\tau(m) < \zeta, X\big(\tau(m)\big) \notin B_i^m\big) \le r^{-c}. \tag{3.49}$$

For estimate (3.46), note that it follows from the identity $R = \mathbb{1} \cdot X$, the expression (2.7) for the generator of X (with λ_i and μ_i substituted for λ_i^r and μ_i^r), and condition (2.9), that the stopped process $R(\cdot \wedge \zeta \wedge \tau(m))$ is a martingale. By this martingale property and the fact that X lives on the grid $\mathcal{S}_{\mathrm{u}}^1$, there is a constant c such that

$$\frac{m - r - r^\kappa - c}{m + c} \le \mathbb{P}_x^1\big(\zeta < \tau(m)\big) \le \frac{m - r + r^\kappa + c}{m}.$$

Estimate (3.46) follows, using the fact that $m \ge 2r$.

For inequality (3.47), the relation $r^{-2}\tau(r) = \tau^r(1)$ gives $\mathbb{P}_0^1(r^{-2}\tau(r) > 1) = \mathbb{P}_0^r(\|\hat{R}^r\|_1 < 1)$. By Lemma 3.1, $\hat{R}^r$ converges in law to an RBM ρ, and so $\mathbb{P}_0^r(\|\hat{R}^r\|_1 < 1) \to \mathbb{P}_0(\|\rho\|_1 < 1) < 1$. Thus there exists $\gamma \in (0,1)$ such that for all r sufficiently large, $\mathbb{P}_0^1(r^{-2}\tau(r) > 1) \le \gamma$. For other initial conditions x, the probability of this event under $\mathbb{P}_x^1$ is even smaller and, therefore, is still bounded by γ. Markovity thus gives (3.47).

For estimate (3.48), fix $c_0 > 0$ to be a constant c that satisfies (3.47). Recall the definition of F from (3.26). Since X takes values in $\mathcal{S}_{\mathrm{u}}^1$, $\mathbb{1} \cdot X(\tau(r))$ must take a value within $[r, r + c_1]$, for some constant $c_1 > 0$. We claim that if $x \in \mathbb{R}_+^N$, $\mathbb{1} \cdot x \in [r, r + c_1]$ and $x \notin \bigcup_i B_i^r$ then $F(x) \ge r^\kappa/\sqrt{2}$. To this end, assume, without loss of generality, that $x_1 = \max_i x_i$. Since $x \notin B_1^r$, we have $(x_1 - r)^2 + \sum_{i=2}^N x_i^2 \ge r^{2\kappa}$. Therefore, if $r - r^\kappa/\sqrt{2} \le x_1 \le r + c_1$, we have that $F(x)^2 = (\sum_{i=2}^N x_i)^2 \ge r^{2\kappa}/2$, and hence $F(x) \ge r^\kappa/\sqrt{2}$. If on the other hand $x_1 < r - r^\kappa/\sqrt{2}$, then using $\mathbb{1} \cdot x \ge r$, we obtain again $F(x) = \sum_{i=2}^N x_i > r^\kappa/\sqrt{2}$.

As a result of the above claim, for all large r,

$$\begin{aligned}\mathbb{P}_0^1\Big(X\big(\tau(r)\big) \notin \bigcup_i B_i^r\Big) &\le \mathbb{P}_0^1\big(\tau(r) > c_0 r^2 \log r\big) \\ &\quad + \mathbb{P}_0^1\big(\|F\big(X(\cdot)\big)\|_{c_0 r^2 \log r} \ge r^\kappa/\sqrt{2}\big).\end{aligned} \tag{3.50}$$

By (3.47), the first term above is bounded by r^{-c_0}. Since $F(X(0)) = 0$, we have by Lemma 3.4, relation (3.40), and the relation $\kappa = 1 - \kappa_0$, that the second term is bounded by r^{-c}.

For estimate (3.49), define

$$\nu(m) = \inf\big\{t \ge 0 : R(t) \le m^\kappa\big\}.$$

Fix i and consider $x \in B_i^r$. Denote $B_{0,m} = B(0, m^\kappa) \cap \mathcal{S}_{\mathrm{u}}^1$. Then a use of strong Markovity gives

$$\begin{aligned}\mathbb{P}_x^1\big(\tau(m) < \zeta, X\big(\tau(m)\big) \notin B_i^m\big) &\le \mathbb{P}_x^1\big(\tau(m) < \nu(m), X\big(\tau(m)\big) \notin B_i^m\big) \\ &\quad + \max_{z \in B_{0,m}} \mathbb{P}_z^1\big(\tau(m) < \zeta\big).\end{aligned} \tag{3.51}$$

To bound the first term consider the event $\tau(m) < \nu(m)$. If $X(\tau(m)) \in \bigcup_{j \neq i} B_j^m$, then $\|F(X(\cdot))\|_{\tau(m)} > m^\kappa$ holds, whereas if $X(\tau(m)) \notin \bigcup_j B_j^m$, then by the argument provided in the previous paragraph we have $\|F(X(\cdot))\|_{\tau(m)} \ge m^\kappa/\sqrt{2}$. This implies that the first term in (3.51) is bounded by

$$\mathbb{P}_x^1\big(\tau(m) \wedge \nu(m) > c_0 m^2 \log m\big) + \mathbb{P}_x^1\big(\|F\big(X(\cdot)\big)\|_{c_0 m^2 \log m} > m^\kappa/\sqrt{2}\big).$$

This expression can be handled as the RHS of (3.50). Since $x \in B_i^r$, $\|F(X(0))\| \le r^\kappa \le \frac{1}{2^\kappa} m^\kappa$. Thus Lemma 3.4 is applicable with $\gamma_1 = 2^{-\kappa} < \gamma_2 = 1/\sqrt{2}$. This gives the bound r^{-c} on the first term on the RHS of (3.51).

To bound the second term in (3.51), we again use the martingale property of $R(\cdot \wedge \zeta \wedge \tau(m))$. It gives

$$\mathbb{P}_z^1\big(\tau(m) < \zeta\big) \le \frac{m^\kappa}{m - 2m^\kappa}.$$

Since $2r \le m \le 4r$, we obtain that for sufficiently large r, the last term in (3.51) is bounded above by r^{-c}. This completes the proof of (3.46)–(3.49).

We now deduce (3.44) and (3.45) from (3.46)–(3.49). Identity (3.45) follows immediately from (3.48). For $x \in \mathcal{S}_{\mathrm{u}}^r$ (see (2.6)), denote

$$q(x, r, m) = \mathbb{P}_0^1\big(X\big(\tau(r)\big) = x\big)\mathbb{P}_x^1\big(X\big(\tau(m)\big) \in B_1^m\big)$$

and $B_{r,i} = B_i^r \cap \mathcal{S}_{\mathrm{u}}^r$. Then

$$\begin{aligned}q_1^m &= \sum_{x \in \mathcal{S}_{\mathrm{u}}^r : x \notin \bigcup_{i \ge 1} B_i^r} q(x, r, m) + \sum_{i>1} \sum_{x \in B_{r,i}} q(x, r, m) + \sum_{x \in B_{r,1}} q(x, r, m) \\ &=: \beta^{r,m,1} + \beta^{r,m,2} + \beta^{r,m,3}.\end{aligned} \tag{3.52}$$

It follows from (3.48) that $\beta^{r,m,1} \le r^{-c}$. Next, consider the term $\beta^{r,m,2}$. Let $i > 1$ and $x \in B_{r,i}$. Then

$$\begin{aligned}\mathbb{P}_x^1\big(X\big(\tau(m)\big) \in B_1^m\big) &= \mathbb{P}_x^1\big(\zeta < \tau(m), X\big(\zeta(m)\big) \in B_1^m\big) \\ &\quad + \mathbb{P}_x^1\big(\zeta(m) < \zeta, X\big(\tau(m)\big) \in B_1^m\big).\end{aligned}$$

The first term above is equal to $\mathbb{P}_x^1(\zeta < \tau(m)) q_1^m$. By (3.49), the second term bounded by r^{-c}. Combining this with (3.46),

$$\beta^{r,m,2} = \sum_{i>1} q_i^r \frac{m - r}{m} q_1^m + \varepsilon(r, m),$$

where here and in the remainder of this proof, $\varepsilon(r,m)$ denotes a generic function of (r,m) which satisfies $|\varepsilon(r,m)| \le r^{-c}$ for all large r.

As for $\beta^{r,m,3}$, consider $x \in B_1^r$. We have

$$\begin{aligned}
&\mathbb{P}_x^1\big(X\big(\tau(m)\big) \in B_1^m\big) \\
&\quad = \mathbb{P}_x^1\big(\zeta < \tau(m), X\big(\tau(m)\big) \in B_1^m\big) + \mathbb{P}_x^1\big(\tau(m) < \zeta, X\big(\tau(m)\big) \in B_1^m\big) \\
&\quad = \mathbb{P}_x^1\big(\zeta < \tau(m)\big) q_1^m + \mathbb{P}_x^1\big(\tau(m) < \zeta\big) - \mathbb{P}_x^1\big(\tau(m) < \zeta, X\big(\tau(m)\big) \notin B_1^m\big).
\end{aligned}$$

Using (3.49), the last term above is bounded, in absolute value, by r^{-c}. Combined with (3.46), this gives

$$\beta^{r,m,3} = q_1^r\left(\frac{m-r}{m} q_1^m + \frac{r}{m}\right) + \varepsilon(r,m).$$

Combining the three estimates,

$$\begin{aligned}
q_1^m &= \sum_{i=1}^{r} q_i^r \frac{m-r}{m} q_1^m + q_1^r \frac{r}{m} + \varepsilon(r,m) \\
&= \big(1 - \varepsilon(r,m)\big)\frac{m-r}{m} q_1^m + q_1^r \frac{r}{m} + \varepsilon(r,m),
\end{aligned}$$

where (3.48) is used. Hence, using $m/r \le 4$,

$$|q_1^m - q_1^r| \le \frac{m}{r}|\varepsilon(r,m)| \le c r^{-c}.$$

This gives (3.44) and completes the proof of the lemma. □

3.5. *Relaxation of the homogeneity assumption.* In this section, we prove Proposition 3.3 based on Lemma 3.5, by means of a change of measure. Thus the general setting, where λ_i^r and μ_i^r satisfy the hypotheses of Theorem 2.1, is in force. Since the statement of Proposition 3.3 refers to $\mathbb{P}_0^r$, we may and will assume in this section that the initial condition is $Q^r(0) = X^r(0) = 0$ identically. Thus the only stochastic primitives in the model are the processes (A^r, S^r). In particular, as follows from equations (2.1), (2.2), (2.4), (2.5) and (2.10), the processes $X^r(t)$, $t \in [0,T]$ and $\hat{X}^r(t)$, $t \in [0,T]$ are determined by $(A^r(t), S^r(t))$, for $t \in [0,T]$. In addition to the measure $\mathbf{P}$, we introduce below a reference probability measure $\mathbf{Q}$ on $(\Omega, \mathcal{F})$ under which, for all r, the Poisson processes A_i^r and S_i^r have intensities $\lambda_i^{0,r}$ and $\mu_i^{0,r}$, respectively, where we denote $\lambda_i^{0,r} = \lambda_i r^2$ and $\mu_i^{0,r} = \mu_i r^2$. Denote by $\mathbb{E}_{\mathbf{Q}}$ the corresponding expectation. The laws of the driving Poisson processes as well as that of the queue length process Q^r under $\mathbf{P}$ can then be obtained from those under $\mathbf{Q}$ by a change of measure (as shown below). However, this does not apply to the nominal workload process X^r, for which the parameters λ_i^r and μ_i^r determine not only the jump intensities but also the scaling factors in the definition (2.1) of X^r in terms of Q^r. This is reflected also in the formula for the generator

$\mathcal{L}^r_{\mathrm{u}}$ (see (2.7)) where these parameters enter in both the jump rates and the jump sizes. An intermediate transformation is required.

To this end, we define analogously to (2.1) and (2.10), a process $X_i^{0,r}$ and its scaled version by

$$X_i^{0,r} = (\mu_i^{0,r})^{-1} Q_i^r, \qquad \hat{X}_i^{0,r} = r X_i^{0,r}.$$

Similarly, we let $\hat{R}^{0,r} = \mathbb{1} \cdot \hat{X}^{0,r}$ and $\tau^{0,r}(\varepsilon) = \inf\{t \geq 0 : \hat{R}^{0,r}(t) \geq \varepsilon\}$.

The starting point of this section is to notice that Lemma 3.5, proved in the previous section, implies that there exists $q \in \mathcal{M}_1$ such that for $\kappa_0 \in (0, 1/2)$,

$$\lim_{\varepsilon \downarrow 0} \limsup_{r \to \infty} \left| \mathbf{Q}\big(\hat{X}^{0,r}(\tau^{0,r}(\varepsilon)) \in B\big(\varepsilon e_i, r^{-\kappa_0}\big)\big) - q_i \right| = 0, \qquad i \in [N]. \tag{3.53}$$

The proof proceeds in two steps. First, it is shown that a version of (3.53), that refers to $\hat{X}^r(\tau^r(\varepsilon))$ in place of $\hat{X}^{0,r}(\tau^{0,r}(\varepsilon))$, is valid, and then that the same statement remains true under $\mathbf{P}$ (equivalently, under $\mathbb{P}^r_0$).

PROOF OF PROPOSITION 3.3. We first prove that, for q as in (3.53), there exists $u \in \mathcal{U}_0$ such that

$$\lim_{\varepsilon \downarrow 0} \limsup_{r \to \infty} \left| \mathbf{Q}\big(\hat{X}^r(\tau^r(\varepsilon)) \in B\big(\varepsilon e_i, u(r)\big)\big) - q_i \right| = 0, \qquad i \in [N]. \tag{3.54}$$

Based on (3.53), the statement (3.54) is almost an immediate consequence of convergence of $\hat{R}^{0,r}$ to an RBM under $\mathbf{Q}$ and the closeness of $\hat{X}^{0,r}$ and $\hat{X}^r$. Indeed, the relation between $\hat{X}^{0,r}$ and $\hat{X}^r$ is $\hat{X}_i^{0,r} = \beta_i^r \hat{X}_i^r$ where $\beta_i^r = \mu_i^r / \mu_i^{0,r}$. We have $\max_i |\beta_i^r - 1| < cr^{-1}$ by (2.8). Thus for $\varepsilon < 1$, $\|\hat{X}^{0,r}(t) - \hat{X}^r(t)\| < cr^{-1}$ for all $t \leq \tau^r(\varepsilon) \wedge \tau^{0,r}(\varepsilon)$. Hence (3.54) will follow from (3.53) if we show that, as $r \to \infty$, $\sup_{\varepsilon \in (0,1)} \|\hat{X}^{0,r}(\tau^r(\varepsilon)) - \hat{X}^{0,r}(\tau^{0,r}(\varepsilon))\| \to 0$ in probability. Since $\hat{X}^{0,r}$ are $\mathcal{C}$-tight by Lemma 3.2 and $\tau^{0,r}(\varepsilon)$ are dominated by $\tau^{0,r}(1)$, that form a tight sequence of RVs, it suffices to prove that

$$\sup_{\varepsilon \in (0,1)} \left| \tau^r(\varepsilon) - \tau^{0,r}(\varepsilon) \right| \to 0 \qquad \text{in probability.} \tag{3.55}$$

The convergence of $\hat{R}^{0,r}$ to RBM implies that for any $M > 0$ and $\delta > 0$,

$$\lim_{\kappa \downarrow 0} \limsup_{r \to \infty} \mathbf{Q}\Big(\inf_{t \in [0,M]} \sup_{u \in (0,\delta)} \left| \hat{R}^{0,r}(t+u) - \hat{R}^{0,r}(t) \right| < \kappa \Big) = 0.$$

It follows that for any $M > 0$ and $\delta > 0$,

$$\limsup_{r \to \infty} \mathbf{Q}\Big(\sup_{\varepsilon \in (0,1)} \left| \tau^r(\varepsilon) \wedge M - \tau^{0,r}(\varepsilon) \wedge M \right| > \delta \Big) = 0.$$

Using again the tightness of the RVs $\tau^{0,r}(1)$, (3.55) follows, hence also (3.54).

The second and final step is to prove that in (3.54), $\mathbf{Q}$ may be replaced by $\mathbf{P}$. Denote the events of interest by $K^r_{\varepsilon,i} = \{\hat{X}^r(\tau^r(\varepsilon)) \in B(\varepsilon e_i, u(r))\}$. Since $\sum_i q_i =$

1, using the fact that for any ε and all sufficiently large r, $\{K^r_{\varepsilon,i}\}_i$ are disjoint, it suffices to prove for each i the lower bound

$$\liminf_{\varepsilon\downarrow 0}\liminf_{r\to\infty}\mathbf{P}\big(K^r_{\varepsilon,i}\big)\geq q_i. \tag{3.56}$$

Given any $\delta>0$, we clearly have $\lim_{\varepsilon\downarrow 0}\limsup_{r\to\infty}\mathbf{Q}(\tau^r(\varepsilon)>\delta)=0$. Hence by (3.54), denoting $K^r_{\varepsilon,\delta,i}=K^r_{\varepsilon,i}\cap\{\tau^r(\varepsilon)\leq\delta\}$, we have

$$\lim_{\varepsilon\downarrow 0}\limsup_{r\to\infty}\big|\mathbf{Q}\big(K^r_{\varepsilon,\delta,i}\big)-q_i\big|=0,\qquad i\in[N]. \tag{3.57}$$

A change of measure is formulated in terms of the exponential martingale

$$\psi^r_t=\exp\sum_i\bigg[A^r_i(t)\log\bigg(\frac{\lambda^r_i}{\lambda^{0,r}_i}\bigg)-(\lambda^r_i-\lambda^{0,r}_i)t + S^r_i(t)\log\bigg(\frac{\mu^r_i}{\mu^{0,r}_i}\bigg)-(\mu^r_i-\mu^{0,r}_i)t\bigg].$$

Let $\mathcal{A}^r_t=(A^r_i(u),S^r_i(u))_{i\in[N],u\in[0,t]}$. Let also $\mathcal{G}^r_t=\sigma\{\mathcal{A}^r_t\}$. For each r and t, let a probability measure $\mathbf{P}^{r,t}$ on $(\Omega,\mathcal{G}^r_t)$ be defined by $\mathbf{P}^{r,t}(G)=\mathbb{E}_{\mathbf{Q}}[\psi^r_t\mathbb{1}_G]$ for $G\in\mathcal{G}^r_t$. Then, for each r and t, the law of $\mathcal{A}^r_t$ under $\mathbf{P}^{r,t}$ is the same as that under $\mathbf{P}$. Moreover, note that for each i, the event $K^r_{\varepsilon,\delta,i}$ is measurable on $\mathcal{G}^r_\delta$. Hence to establish (3.56), it suffices to prove that for each i,

$$\tilde{q}_i:=\liminf_{\delta\downarrow 0}\liminf_{\varepsilon\downarrow 0}\liminf_{r\to\infty}\mathbb{E}_{\mathbf{Q}}\big[\psi^r_\delta\mathbb{1}_{K^r_{\delta,\varepsilon,i}}\big]\geq q_i. \tag{3.58}$$

For $\eta>0$, denote $G^r_{\delta,\eta}=\{\psi^r_\delta>1-\eta\}$. Suppose we show that for any $\eta>0$,

$$\liminf_{\delta\downarrow 0}\liminf_{r\to\infty}\mathbf{Q}\big(G^r_{\delta,\eta}\big)=1. \tag{3.59}$$

Then we may argue as follows:

$$\mathbb{E}_{\mathbf{Q}}\big[\psi^r_\delta\mathbb{1}_{K^r_{\varepsilon,\delta,i}}\big]\geq\mathbb{E}_{\mathbf{Q}}\big[\psi^r_\delta\mathbb{1}_{K^r_{\varepsilon,\delta,i}\cap G^r_{\delta,\eta}}\big]\geq(1-\eta)\mathbf{Q}\big(K^r_{\varepsilon,\delta,i}\big)-\mathbf{Q}\big(\big(G^r_{\delta,\eta}\big)^c\big).$$

Taking $r\to\infty$ then $\varepsilon\downarrow 0$, then using (3.57), and finally taking $\delta\downarrow 0$, gives

$$\tilde{q}_i\geq(1-\eta)q_i-\limsup_{\delta\downarrow 0}\limsup_{r\to\infty}\mathbf{Q}\big(\big(G^r_{\delta,\eta}\big)^c\big)=(1-\eta)q_i.$$

Since $\eta>0$ is arbitrary, this gives (3.58), and consequently (3.56).

Thus the proof will be complete once (3.59) is shown. To this end, let

$$\tilde{A}^r_i(t)=r^{-1}\big(A^r_i(t)-\lambda^{0,r}_i t\big),\qquad \tilde{S}^r_i(t)=r^{-1}\big(S^r_i(t)-\mu^{0,r}_i t\big).$$

These processes, defined analogously to $\hat{A}^r$ and $\hat{S}^r$, converge under $\mathbf{Q}$ to BMs. Denote $\hat{\lambda}^r_i=r^{-1}(\lambda^r_i-\lambda^{0,r}_i)$ and $\hat{\mu}^r_i=r^{-1}(\mu^r_i-\mu^{0,r}_i)$ and recall by (2.8) that

these sequences converge. Write ψ_t^r in terms of $\tilde{A}^r$ and $\tilde{S}^r$ as

$$\psi_t^r = \exp \sum_i \big[\tilde{A}_i^r(t) r U_i^r + \big(\lambda_i^{0,r} U_i^r - r\hat{\lambda}_i^r\big)t + \tilde{S}_i^r(t) r V_i^r + \big(\mu_i^{0,r} V_i^r - r\hat{\mu}_i^r\big)t\big],$$

where $U_i^r = \log(1 + \frac{\hat{\lambda}_i^r}{r\lambda_i})$, $V_i^r = \log(1 + \frac{\hat{\mu}_i^r}{r\mu_i})$. Denoting $L_t^r = \max_i |\tilde{A}_i^r(t)| \vee |\tilde{S}_i^r(t)|$ and using $|\log(1+x) - x| \le cx^2$ for all $|x| < 1/2$, we have for all large r,

$$\log \psi_t^r \ge -cL_t^r - ct.$$

The aforementioned convergence to BM clearly implies that, for any $\eta > 0$,

$$\liminf_{\delta \downarrow 0} \liminf_{r \to \infty} \mathbf{Q}\big(L_\delta^r < \eta\big) = 1.$$

We thus obtain (3.59) and complete the proof. □

4. Concluding remarks.

1. It is desirable to extend the main result of this paper beyond the Markovian setting, to general service time distributions and renewal arrival distributions, under second moment conditions. Whereas the behavior of the modulus according to an RBM certainly holds in vast generality, and the attraction to the collection of axes $\mathcal{S}_0$ can likely be extended, the existence of limiting entrance laws appears to require different machinery. Indeed, the proof presented here makes crucial use of the strong Markovity of the prelimit processes.
2. The proof presented in this paper sheds no light on the angular distribution q (except that it does not depend on the second order parameters $\hat{\lambda}_i$, $\hat{\mu}_i$). A characterization of q that would be useful and lead to further information about it is desirable.
3. Figure 2 depicts results of Monte Carlo simulations for an SSQ model with $N = 2$ at criticality, aimed at estimating q. It shows the behavior of q_1 as several parameters vary. They all suggest monotone dependence, that one would wish to substantiate mathematically.

   (a) The graph shown at Figure 2(c) is, in particular, relevant to the heuristic mentioned in the Introduction, according to which more variable traffic attains lower priority. In this example, the traffic intensities λ_i/μ_i are kept fixed. As λ_1 increases, the inter-arrival variance increases, which, according to the graph, increases q_1, indicating lower priority for this class.

   (b) Figure 2(d) shows the dependence on the tie breaking parameter, p_1. It exhibits that the tie breaking rule affects the limiting angular distribution. However, we have not aimed at providing a proof of this claim.

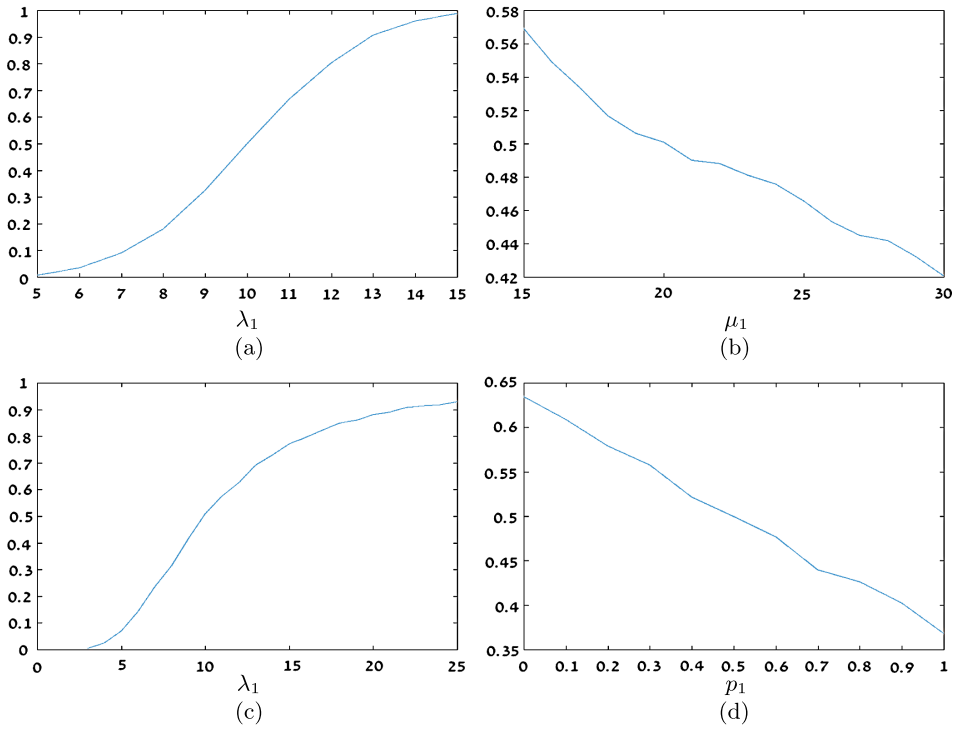

FIG. 2. *Simulation of q_1 as a function of various parameters, for $N = 2$.* (a) *q_1 as a function of λ_1, fixed μ's*: $\mu_1 = \mu_2 = 20$, $\lambda_2 = \mu_2 - \lambda_1$. (b) *$q_1$ as a function of μ_1, fixed λ's*: $\lambda_1 = \lambda_2 = 10$, $\mu_2 = 1/(1/\lambda_1 - 1/\mu_1)$. (c) *$q_1$ as a function of λ_1, fixed ratio λ_1/μ_1, λ_2 and μ_2*: $\lambda_2 = 10$, $\mu_2 = 20$, $\mu_1 = 2\lambda_1$. (d) *q_1 as a function of p_1, fixed λ's and μ's*: $\lambda_1 = \lambda_2 = 10$, $\mu_1 = \mu_2 = 20$, $p_2 = 1 - p_1$.

Acknowledgments. The authors are grateful to Ross Pinsky for useful discussions about heat equation estimates, to Bert Zwart for bringing reference [18] to their attention, and to the referee for careful reading and valuable comments.

REFERENCES

[1] BARLOW, M., PITMAN, J. and YOR, M. (1989). On Walsh's Brownian motions. In *Séminaire de Probabilités, XXIII. Lecture Notes in Math.* **1372** 275–293. Springer, Berlin. MR1022917

[2] BAXTER, J. R. and CHACON, R. V. (1984). The equivalence of diffusions on networks to Brownian motion. In *Conference in Modern Analysis and Probability* (*New Haven, Conn.*, 1982). *Contemp. Math.* **26** 33–48. Amer. Math. Soc., Providence, RI. MR0737386

[3] BENAMEUR, N., GUILLEMIN, F. and MUSCARIELLO, L. (2013). Latency reduction in home access gateways with shortest queue first. In *Proc. ISOC Workshop on Reducing Internet Latency*.

[4] BILLINGSLEY, P. (1999). *Convergence of Probability Measures*, 2nd ed. Wiley, New York. MR1700749

[5] BONALD, T., MUSCARIELLO, L. and OSTALLO, N. (2011). Self-prioritization of audio and video traffic. In *IEEE International Conference on Communications* (*ICC*) 1–6. IEEE, New York.

[6] CAROFIGLIO, G. and MUSCARIELLO, L. (2010). On the impact of TCP and per-flow scheduling on Internet performance. In *INFOCOM*, 2010 *Proceedings IEEE* 1–9. IEEE, New York.

[7] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes*: *Characterization and Convergence*. Wiley, New York. MR0838085

[8] GUILLEMIN, F. and SIMONIAN, A. (2013). Analysis of the shortest queue first service discipline with two classes. In *Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools* 1–10. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Portland.

[9] GUILLEMIN, F. and SIMONIAN, A. (2014). Stationary analysis of the shortest queue first service policy. *Queueing Syst*. **77** 393–426. MR3225817

[10] GUILLEMIN, F. and SIMONIAN, A. (2017). Stationary analysis of the "shortest queue first" service policy: The asymmetric case. *Stoch. Models* **33** 256–296. MR3650557

[11] HAJRI, H. (2012). Discrete approximations to solution flows of Tanaka's SDE related to Walsh Brownian motion. In *Séminaire de Probabilités XLIV*. *Lecture Notes in Math*. **2046** 167–190. Springer, Heidelberg. MR2953347

[12] HARRISON, J. M. (1995). Balanced fluid models of multiclass queueing networks: A heavy traffic conjecture. In *Stochastic Networks*. *IMA Vol. Math. Appl*. **71** 1–20. Springer, New York. MR1381003

[13] HARRISON, J. M. and SHEPP, L. A. (1981). On skew Brownian motion. *Ann. Probab*. **9** 309–313. MR0606993

[14] HARRISON, J. M. and WILLIAMS, R. J. (1996). A multiclass closed queueing network with unconventional heavy traffic behavior. *Ann. Appl. Probab*. **6** 1–47. MR1389830

[15] HU, Q., WANG, Y. and YANG, X. (2012). The hitting time density for a reflected Brownian motion. *Comput. Econ*. **40** 1–18.

[16] ICHIBA, T., KARATZAS, I., PROKAJ, V. and YAN, M. (2018). Stochastic integral equations for Walsh semimartingales. *Ann. Inst. Henri Poincaré Probab. Stat*. **54** 726–756. MR3795064

[17] KRUK, Ł. (2011). An open queueing network with asymptotically stable fluid model and unconventional heavy traffic behavior. *Math. Oper. Res*. **36** 538–551. MR2832406

[18] LAMBERT, A. and SIMATOS, F. (2014). The weak convergence of regenerative processes using some excursion path decompositions. *Ann. Inst. Henri Poincaré Probab. Stat*. **50** 492–511. MR3189081

[19] NASSER, N., AL-MANTHARI, B. and HASSANEIN, H. (2005). A performance comparison of class-based scheduling algorithms in future UMTS access. In *IPCCC* 2005. 24*th IEEE International* 437–441. IEEE, New York.

[20] OSTALLO, N. (2008). Service differentiation by means of packet scheduling. Master's thesis, Institut Eurècom Sophia Antipolis, Biot, France.

[21] PLAMBECK, E., KUMAR, S. and HARRISON, J. M. (2001). A multiclass queue in heavy traffic with throughput time constraints: Asymptotically optimal dynamic controls. *Queueing Syst*. **39** 23–54. MR1865457

[22] PROTTER, P. E. (2004). *Stochastic Integration and Differential Equations*: *Stochastic Modelling and Applied Probability*, 2nd ed. *Applications of Mathematics* (*New York*) **21**. Springer, Berlin. MR2020294

[23] ROGERS, L. C. G. (1983). Itô excursion theory via resolvents. *Z. Wahrsch. Verw. Gebiete* **63** 237–255. MR0701528

[24] SALISBURY, T. S. (1986). Construction of right processes from excursions. *Probab. Theory Related Fields* **73** 351–367. MR0859838

[25] TSIRELSON, B. (1997). Triple points: From non-Brownian filtrations to harmonic measures. *Geom. Funct. Anal.* **7** 1096–1142. MR1487755

[26] VAROPOULOS, N. T. (1985). Long range estimates for Markov chains. *Bull. Sci. Math.* (2) **109** 225–252. MR0822826

[27] WALSH, J. B. (1978). A diffusion with a discontinuous local time. *Astérisque* **52** 37–45.

[28] WHITT, W. (1971). Weak convergence theorems for priority queues: Preemptive-resume discipline. *J. Appl. Probab.* **8** 74–94. MR0307389

[29] WILLIAMS, R. J. (1996). On the approximation of queueing networks in heavy traffic. In *Stochastic Networks*: *Theory and Applications* 35–56. Oxford Univ. Press, Oxford.

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION—ISRAEL INSTITUTE OF TECHNOLOGY
HAIFA 32000
ISRAEL

DEPARTMENT OF STATISTICS
UNIVERSITY OF HAIFA
HAIFA 31905
ISRAEL
E-MAIL: shloshim@gmail.com
URL: https://sites.google.com/site/asafcohentau/