Queueing Systems (2020) 94:357–392
https://doi.org/10.1007/s11134-019-09644-9

Stationary distributions and convergence for $M/M/1$ queues in interactive random environment

Guodong Pang[1] · **Andrey Sarantsev**[2] · **Yana Belopolskaya**[3] · **Yuri Suhov**[4,5]

Received: 11 February 2019 / Revised: 1 November 2019 / Published online: 8 January 2020

Abstract
A Markovian single-server queue is studied in an interactive random environment. The arrival and service rates of the queue depend on the environment, while the transition dynamics of the random environment depend on the queue length. We consider in detail two types of Markov random environments: a pure jump process and a reflected jump diffusion. In both cases, the joint dynamics are constructed so that the stationary distribution can be explicitly found in a simple form (weighted geometric). We also derive an explicit estimate for the exponential rate of convergence to the stationary distribution via coupling.

Keywords Queues in interactive random environment · Stationary distribution · Rate of convergence to stationarity · Coupling

Mathematics Subject Classification 60K25 · 60K30 · 60K35 · 60K37 · 90B22 · 60J60 · 60J65 · 37A25

✉ Guodong Pang
gup3@psu.edu

Andrey Sarantsev
asarantsev@unr.edu

Yana Belopolskaya
yana@yb1569.spb.edu

Yuri Suhov
yms@statslab.cam.ac.uk; ims14@psu.edu

1 Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA

2 Department of Mathematics and Statistics, University of Nevada in Reno, Reno, USA

3 Petersburg Department of Mathematical Institute of Russian Academy of Science, Saint Petersburg State University of Architecture and Civil Engineering, Saint Petersburg, Russia

4 Statistical Laboratory, University of Cambridge, Cambridge, UK

5 Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, USA

1 Introduction

In this paper, we propose a tractable modeling approach to studying queues in an interactive random environment, where the arrival and/or service rates are modulated by a Markov process and the dynamics of the environment also depend on the state of the queue. Such models may be used in the following setting: In a service system (for example, on-demand service platforms), the demand may be affected by the service quality as indicated by dynamic "ratings" which may be modeled by a Markov chain, while the ratings dynamics may depend on the congestion level in the system.

For an $M/M/1$ queue in an interactive random environment, let $N(t)$ be the queue length process (the number of customers in the system) and $Z(t)$ be the random process for the environment. The joint process $(N(t), Z(t))$ can be modeled as a continuous-time Markov process on $\mathbb{N} \times \mathcal{Z}$ ($\mathcal{Z}$ representing the range of $Z(t)$), with a generator

$$\mathcal{L}f(n,z) = \mathcal{M}_z f(n,z) + \mathcal{A}_n f(n,z), \tag{1.1}$$

where $\mathcal{M}_z$ describes the queueing dynamics depending on the environment state z and $\mathcal{A}_n$ describes the environment dynamics depending on the queueing state n. Specifically, given the arrival and service rates $\lambda(z)$ and $\mu(z)$, we can write

$$\mathcal{M}_z f(n,z) = \lambda(z)(f(n+1,z) - f(n,z)) + 1_{\{n\neq 0\}}\mu(z)(f(n-1,z) - f(n,z)).$$

On the other hand, the generator $\mathcal{A}_n$ can, for a general Markov process, depend on the queue length n. For example, for a given n, $\mathcal{A}_n$ may represent the generator of a diffusion process

$$\mathcal{A}_n f(n,z) = b_n(z) \cdot \nabla_z f(n,z) + \frac{1}{2}\operatorname{tr}\left(\Sigma_n(z)\nabla_z^2 f(n,z)\right)$$

or a continuous-time jump Markov chain with a transition rate matrix depending on the queueing state n. In the utmost generality, one can impose mild conditions on the generators $\mathcal{M}_z$ and $\mathcal{A}_n$ to guarantee the existence of an invariant measure for the joint process $(N(t), Z(t))$. However, it seems difficult to go beyond that without any structural assumptions on the joint generator, especially $\mathcal{A}_n$. In many applications, it is convenient to have an explicit invariant measure to work with. In general, it is hard to find an explicit form for stationary distributions of multidimensional Markov processes. (For example, in [42], it is shown that an obliquely reflected Brownian motion (RBM) in a polyhedral domain in $\mathbb{R}^d$ has a product-of-exponentials stationary distribution under the *skew symmetry* condition, the only case with an explicit stationary measure.)

Therefore, in order to provide an *explicit* expression for the invariant measure of the joint process, we study a particular *multiplicative* (scaled) form in the generator component $\mathcal{A}_n$, that is,

$$\mathcal{A}_n f(n,z) = \beta_n \rho^{-n}(z)\mathcal{A}f(n,z),$$

where β_n is a positive constant, $\rho(z) = \lambda(z)/\mu(z)$ is the traffic intensity in the queue, and $\mathcal{A}f(n, z)$ is a generator corresponding to a Markov process whose transition dynamics do not depend on n. (In the case of reflected processes, the boundary conditions should be treated carefully; see Sect. 3 for details.) The scaling factors not only depend on the queue length n, but also include the traffic intensity $\rho(z)$. For an environment state z, $\rho^{-n}(z) > 1$ for all queue states n, but the factor β_n gives more flexibility (slowing down or speeding up) to the scaling of the generator $\mathcal{A}$. Our approach is motivated by applications where the environment dynamics may be sped up or slowed down by the congestion. For example, in on-demand service systems, the transitions among the different service quality "ratings" may simultaneously change faster when many customers experience more congestion due to higher response rates.

We discuss two types of random environment: a pure jump Markov chain taking values in a discrete state space D (finite or countable) and a reflected (jump) diffusion in a piecewise smooth domain, also denoted by D. Each type of environment is of interest. Under certain assumptions, we prove the existence of the joint invariant measure, derive its explicit expression, and establish the exponential rate of convergence to the steady state (in the total variation norm). The explicit expression of the invariant measure can be regarded as a weighted geometric form (or some "product form," although not exactly in the same sense as in the literature on stochastic networks [10,22,23]). Specifically, we have the joint invariant measure for $(N(t), Z(t))$ of the form $\pi(\{n\}, \mathrm{d}z) = \Xi^{-1}\rho^n(z)\nu(\mathrm{d}z)$, where Ξ is some normalization constant, and $\nu(\cdot)$ is the invariant measure associated with the generator $\mathcal{A}$. Recall that the steady-state distribution of the $M/M/1$ queue itself given an environment state z is geometric ($P(N(\infty) = n) = (1-\rho(z))\rho^n(z)$). The product of the terms "$\rho^n(z)$" and "$\nu(\mathrm{d}z)$" mixes the invariant measures for the queue and the environment, despite $\rho(z)$ depending on z. Here, the scaling factor $\rho^{-n}(z)$ in $\mathcal{A}_n$ is critical. For the two types of environment processes, we are able to establish the exponential rate of convergence.

With a diffusive environment, our work introduces new stochastic models. The simple models include: (a) an $M/M/1$ queue with an interactive diffusive arrival rate: The arrival rate is a one-dimensional reflected (jump) diffusion process in $[0, 1]$ under a fixed service rate 1; (b) an $M/M/1$ queue with an interactive diffusive service rate: The service rate is a one-dimensional reflected (jump) diffusion process in $[1, \infty)$ under a fixed arrival rate 1; and (c) the arrival and service rates form a two-dimensional RBM in an open convex cone (with arrival rate strictly lower than service rate). RBMs have been extensively studied in the queueing (network) literature as scaling limits. However, RBMs as arrival and/or service rates have not been carefully studied. When there is no interactive behavior, the $M/M/1$ queue with a RBM arrival rate can be regarded as a special case of queues of the so-called doubly stochastic Poisson arrival processes with the arrival rate being an independent stochastic process. (See, for example, [3–5].) Our first model extends such existing interesting studies to include feedback from queue to environment. The second and third models with RBM being the service rate or the RBM in the wedge for both arrival and service rates are new, even in the setting of no interactive behavior. Such models are worth further careful investigation. Of course, our models go beyond RBMs, to general reflected (jump) diffusion models.

We aim to find the explicit rate of convergence to the stationary distribution in these models. For standard $M/M/1$ queues, it is well known that the rate of convergence is

exponential; see, for example, [36, Proposition 5.8]. However, for diffusion processes (solutions of SDEs), reflected diffusions, and their versions with jumps, the characterization of an explicit rate of convergence to steady state (as opposed to simply proving that there exists an exponential rate of convergence) is quite a challenging problem. See, for example, [11,20,37,38]. Thus, it is a much more difficult problem to study the rate of convergence for the joint Markov process with a generator in the general form in (1.1) due to the complicated interactive behavior of the two processes (one being discrete and the other being continuous). We attempt to solve this problem via a *coupling* technique for the joint process (N, Z). We provide a novel way to construct the coupling time for the joint process in order to prove that the convergence rate is exponential and, more importantly, provide good estimates of the rate of convergence via careful studies of the exponential bounds for the coupling time. This appears to be the first work in the literature to carefully find estimates of the coupling times of joint processes for queueing processes in random environments.

Although our main focus is on the multiplicative (scaling) form in the generator $\mathcal{A}_n$, we have also considered a setup where the environment jump diffusion described above depends on the queue length n via its domain $D_n \subseteq D$. In particular, the drift vector field, covariance matrix field and the jump measure remain the same for all n, but reflection vector fields may depend on the queueing state n. The entire domain D is the union of these D_n, $n = 0, 1, 2, \ldots$. We assume that this reflected jump diffusion in D_n has a unique invariant probability measure $\nu_{D_n}^{(n)}$ inside the domain D_n, which is the projection of a certain finite measure on D to D_n. (The corresponding boundary measures may depend on n.) See Assumptions 3.3–3.6. We prove similar results as above in this setting. We construct two special examples: an $M/M/1$ queue with a fixed service rate and a reflected diffusion arrival rate, controlled based on a threshold of queue length (Example 3.1), and an $M/M/1$ queue with a fixed arrival rate and a diffusive service rate, controlled similarly (Example 3.2).

When the random environment is a Markov chain taking discrete values, our results also extend to the generator $\mathcal{A}_n$ of the form $\rho^{-n}(z)\tau_n(z, z')$, where the generator rate τ_n may depend on the queueing state n unlike the multiplicative case. However, it is assumed that an invariant measure associated with the transition rate $\tau_n(z, z')$ exists such that it is independent of the queueing state n (Assumption 2.1). This is slightly more general than the multiplicative case, so we state the model and results in Sect. 2 in this setup. We also give an example to illustrate how this slightly more general setup is used. (See Examples 2.1 and 2.2.)

1.1 Literature review on queues in interactive environments

Queues in random environments (for example, Markov-modulated models) have been extensively studied in the literature. Most of the literature assumes that the queueing dynamics are affected by the environment, but not interactive. For example, the paper [35] studies Markov-modulated arrival and service rates with finite environment space and finds expressions for waiting times. The paper [41] deals with similar questions by comparing this queue with an appropriate $M/M/1$ queue. Optimization of the service rate for the case when the arrival rate is a Markov process is studied in [26]. See also

a birth–death process in random environment [13] and a Markov chain in Markov environment, studied in [12,16,33]. A particular case of a Markov-modulated setting is when the service dynamics are subject to interruptions. In this case, the random environment only affects service rate μ. The survey [25] summarizes the existing literature on this topic.

In the Markov-modulated queueing literature, the arrival or service rates under modulation take a finite or countable number of values. However, in practice, the rates under modulation can possibly take continuous values. Our work thus goes beyond the existing frameworks and develops new queueing models.

In [18], the authors study a random particle (a distinguished customer) walking randomly over the sites of a symmetric Jackson network (open or closed), where the arrival rate of a station/node or the transition of customers from it to other stations/nodes is affected if the particle occupies it, while the jump rate of the particle depends on the state of the station/node it currently occupies. An explicit steady-state distribution for the joint process is derived. In [24], Jackson networks in interactive random environments are studied, where the service capacities are affected by the environment, while customer departure may force the environment to jump immediately. An explicit expression of the product form is derived for the joint queueing and environment processes. Inspired by [18], a different construction of Markov processes in random environments resulting in product-form invariant measure is provided. In [6], various Markov processes with interactive random environments are constructed. This paper is of the same flavor as [6]. The paper [17] deals with the feedback loop created by blocking some channels in a multi-server queue and finds a product-form stationary distribution for the joint process. None of these papers investigate the rate of convergence to stationarity. Our model of a single-server queue is also constructed in a more general manner.

The papers [14,44] study birth–death processes in a random environment with feedback. This is a more general setup than in our paper, because an $M/M/1$ queue is a particular case of a birth–death process. However, [14] is concerned with explosion questions, rather than stationary distributions and convergence rates, and [44] focuses on the generating function approach and achieves only partial results for the steady-state distribution.

1.2 Notation

The integral with respect to the measure ν applied to the function f is written as $\langle \nu, f \rangle$. An exponential distribution with rate α is denoted by $\text{Exp}(\alpha)$. The arrow $\Rightarrow$ indicates weak convergence. The dot product of two vectors a and b is denoted by $a \cdot b$. We say two finite measures μ, ν on $\mathbb{R}$ satisfy $\mu \preceq \nu$ if, for all $u \in \mathbb{R}$, we have $\mu(-\infty, u] \leq \nu(-\infty, u]$, but $\mu(\mathbb{R}) = \nu(\mathbb{R})$. We say that μ is *stochastically dominated by* ν. We transfer this concept to random variables: X is stochastically dominated by Y if the distribution of X is stochastically dominated by the distribution of Y. Let $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$ and $\mathbb{R}_+ := [0, \infty)$. Define the *total variation norm:* For a signed measure ν, let $\|\nu\|_{\text{TV}} := \sup_A |\nu(A)|$. Throughout this article, we consider continuous-time random processes (unless otherwise noted) on a filtered

probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ with the filtration satisfying the usual conditions.

1.3 Organization of the paper

In Sect. 2, we study the model in an interactive jump environment. In Sect. 3, we study the single-server queue with a reflected jump diffusion environment. In Sect. 4, we estimate the explicit rate of exponential convergence for the case of a compact environment state space, for both models in Sects. 2 and 3. In Sect. 6, we state and prove some auxiliary lemmata. We make some concluding remarks in Sect. 5.

2 $M/M/1$ Queue in an interactive jump environment

Consider an $M/M/1$ queue with an infinite waiting space operating in an interactive jump environment described as follows: Let D be a finite or countable state space. For every $n \in \mathbb{Z}_+$, let $\mathbf{T}_n = (\tau_n(z, z'))_{z,z\in D}$ be the generator of an irreducible continuous-time Markov chain on D; this (finite or countable-sized) matrix is called the *nominal jump intensity matrix* for the jump process Z in the queueing state n. We define a two-component Markov process (N, Z) taking values in the countable state space $\mathbb{Z}_+ \times D$ with the following generator matrix $\mathbf{R} = \big(R[(n, z), (n', z')]\big)$:

$$\begin{aligned} R[(n, z), (n+1, z)] = \lambda(z), \quad R[(n, z), (n-1, z)] = \mu(z), \\ R[(n, z), (n, z')] = \rho^{-n}(z)\tau_n(z, z'), \quad R[(n, z), (n', z')] = 0, \quad n \neq n',\ z \neq z', \end{aligned} \tag{2.1}$$

where $\rho(z) := \lambda(z)/\mu(z)$ for each $z \in D$. Here, $N = \{N(t) : t \geq 0\}$ represents the number of jobs in the system (including those in the queue and in service), taking values in $\mathbb{Z}_+$, and $Z = \{Z(t) : t \geq 0\}$ represents a jump process taking values in D. When the environment is in state z, the arrival and service rates for the queueing process are $\lambda(z)$ and $\mu(z)$, respectively, both depending on the state z.

When the queue size is in state n, the transition of the environment Z from state z to state z' occurs at the rate $\rho^{-n}(z)\tau_n(z, z')$. Note that the fourth equation in (2.1) does not allow simultaneous jumps for N and Z. It is evident that the pair (N, Z) is a well-defined Markov process in $\mathbb{Z}_+ \times D$ with the generator $\mathbf{R}$.

Remark 2.1 We do not multiply this transition rate τ_n by a factor β_n: Dependence on n is already enshrined in the rate τ_n. We impose a condition (2.2) to guarantee the product form of the steady state.

We first make the following assumption on the nominal jump intensity matrix $\mathbf{T}_n$.

Assumption 2.1 For each $n \in \mathbb{Z}_+$, $z \in D$, and for some function $v : D \to \mathbb{R}_+$,

$$v(z) \sum_{z'\in D} \tau_n(z, z') = \sum_{z'\in D} v(z')\tau_n(z', z). \tag{2.2}$$

For fixed $n \in \mathbb{Z}_+$, if we define a Markov process $\tilde{Z}_n := \{\tilde{Z}_n(t) : t \geq 0\}$ on D with the nominal jump intensity matrix $\mathbf{T}_n$ as the generator, then (2.2) implies that $v(\cdot)$ defines an invariant measure for $\tilde{Z}_n$. If $\sum_{z\in D} v(z) < \infty$, then this measure can be normalized to a probability distribution. If

$$\sum_{z'\in D} \tau_n(z, z') = \sum_{z'\in D} \tau_n(z', z),$$

then the counting measure is invariant for $\tilde{Z}_n$; if D is a finite set, then it is normalized to a uniform distribution on D. It is important to note that the invariant measure $v(\cdot)$ does not depend on n, although the jump intensity matrix $\mathbf{T}_n$ depends on n.

Remark 2.2 A simple example is when $\tau_n(z, z')$ has a multiplicative form:

$$\tau_n(z, z') = \beta_n \tau(z, z')$$

for some transition rate matrix $\tau(z, z')$ satisfying $v(z)\sum_{z'\in D} \tau(z, z') = \sum_{z'\in D} v(z')\tau(z', z)$. However, we provide examples below in which $\tau_n(z, z')$ depends on n in a nontrivial manner, while the existence of v independent of n is guaranteed. See Examples 2.1 and 2.2.

Assumption 2.2 The functions ρ, v satisfy

$$\rho(z) < 1 \quad \text{for} \quad z \in D, \tag{2.3}$$

$$\Xi := \sum_{z\in D} \frac{v(z)}{1-\rho(z)} = \sum_{n=0}^{\infty}\sum_{z\in D} \rho^n(z)v(z) < \infty. \tag{2.4}$$

Note that the constant Ξ is the normalization constant in the joint invariant measure π in (2.5).

Theorem 2.1 *Under Assumptions 2.1 and 2.2 , the Markov process (N, Z) is irreducible, aperiodic, and positive recurrent. It has an invariant probability measure*

$$\pi(n, z) := \eta(n, z)/\Xi, \quad \forall (n, z) \in \mathbb{Z}_+ \times D, \tag{2.5}$$

where Ξ is given in (2.4), and

$$\eta(n, z) := \rho^n(z)v(z), \quad \forall (n, z) \in \mathbb{Z}_+ \times D. \tag{2.6}$$

This process has transition kernel $P_t(x, \cdot)$ which converges to this invariant measure:

$$\|P_t(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \to 0 \text{ as } t \to \infty, \text{ for all } x \in \mathbb{Z}_+ \times D. \tag{2.7}$$

Proof We first show that the process (N, Z) is irreducible and aperiodic. It follows from the observation that for every $t > 0$, (n, z), $(n', z') \in \mathbb{Z}_+ \times D$, one can with

positive probability get from (n, z) to (n', z') in time t. For the measure η from (2.6) to be finite, we need

$$\sum_{(n,z)} \eta(n,z) = \sum_{(n,z)} \rho^n(z)v(z) = \sum_z \frac{v(z)}{1-\rho(z)} < \infty,$$

which is implied by (2.3)–(2.4) in Assumption 2.2. If we prove that η from (2.6) is indeed an invariant measure, the positive recurrent property follows from [34, Theorem 3.5.3], [40, Theorem 2.7.18], and then the ergodicity, as in (2.7), follows from [32]. To verify that $\eta(n, z)$ in (2.6) is an invariant measure, we show that $\eta'\mathbf{R} = 0$. Let us show that for all $n = 1, 2, \ldots$ and $z \in D$,

$$\begin{aligned}
-\eta(n,z)R[(n,z),(n,z)] &= \eta(n-1,z)R[(n-1,z),(n,z)] + \eta(n+1,z)R[(n+1,z),(n,z)] \\
&\quad + \sum_{z' \neq z} \eta(n,z')R[(n,z'),(n,z)], \\
-\eta(0,z)R[(0,z),(0,z)] &= \eta(1,z)R[(1,z),(0,z)] + \sum\nolimits_{z' \neq z} \eta(0,z')R[(0,z'),(0,z)].
\end{aligned} \tag{2.8}$$

By (2.1), the left- and right-hand sides of the first equation in (2.8) are equal to, respectively,

$$\begin{aligned}
&\eta(n,z) \sum_{(n',z') \neq (n,z)} R[(n',z),(n',z)] \\
&= \rho^n(z)v(z) \left(R[(n,z),(n+1,z)] + R[(n,z),(n-1,z)] + \sum_{z' \neq z} R[(n,z),(n,z')] \right) \\
&= \rho^n(z)v(z) \left(\lambda(z) + \mu(z) + \sum_{z' \neq z} \rho^{-n}(z)\tau_n(z,z') \right) \\
&= \rho^n(z)v(z)(\lambda(z) + \mu(z)) + v(z) \sum_{z' \neq z} \tau_n(z,z'); \\
&\rho^{n-1}(z)v(z)\lambda(z) + \rho^{n+1}(z)v(z)\mu(z) + \sum_{z' \neq z} \rho^n(z')v(z')\rho^{-n}(z')\tau_n(z',z) \\
&= \lambda(z)v(z)\rho^n(z)(\lambda(z) + \mu(z)) + \sum_{z' \neq z} v(z')\tau_n(z',z).
\end{aligned}$$

From (2.2) in Assumption 2.1, the last terms on the right-hand sides of these two last equations are equal. This proves the first equation in (2.8); the second one is similar. This completes the proof. □

Example 2.1 (*D as a union of finite sets*) In Examples 2.1 and 2.2, $\delta(i, j)$ stands for the Kronecker delta. Given $n \in \mathbb{Z}_+$, let D_n be a finite set in $(0, 1)$ with cardinality m_n. For definiteness, assume that $1 < m_n < M$, where $M \in \mathbb{Z}_+$ is a fixed value. Introduce an enumeration of points in each D_n: $D_n = \{z(1), \ldots, z(m_n)\}$ (say, in an increasing order) and make a convention that $z(0) = z(m_n)$, $z(m_n + 1) = z(1)$. The sets D_n

can have common points for different n or be pairwise disjoint. Sets $D = \bigcup_n D_n$ and $\upsilon(z) = 1$ for $z \in D$. The set D can be finite or countable.

Next, take a subset $\mathbb{L} \subseteq \mathbb{Z}_+$ ($\mathbb{L}$ or $\mathbb{Z}_+ \setminus \mathbb{L}$ can be empty). For $n \in \mathbb{L}$, set

$$\tau_n(z, z') = \frac{\beta_n}{m_n - 1}, \quad \forall\, z, z' \in D_n \ \text{ with } \ z \neq z'.$$

For $n \in \mathbb{Z}_+ \setminus \mathbb{L}$, set

$$\tau_n(z(i), z(j)) = \frac{1}{2}\delta(j, i \pm 1), \quad \forall\, i, j \in \{1, \ldots, m_n\}.$$

Here, $\beta_n \in (0, \infty)$ are scaling constants depending on n (which is irrelevant for the invariant measure of the process (N, Z)). Pictorially, τ_n for $n \in \mathbb{L}$ describes uniform jumps on D_n, while for $n \in \mathbb{Z}_+ \setminus \mathbb{L}$, τ_n yields a "nearest-neighbor" walk with cyclic (periodic) boundary condition. Either way, the counting measure υ is invariant; cf. cf. Assumption 2.1.Thus, (2.2) holds true.

Then, $\mathbf{T}_n = \big(\tau_n(z, z')\big)$ generates a Markov chain $\tilde{Z}_n$ with an invariant probability measure $\mathbf{1}_{D_n}(z)/m_n$, $z \in D$. The invariant measure η is then given in (2.6) with $\eta(n, z) = z^n$.

Example 2.2 (D *as a countable set,* τ_n *as a null-recurrent jump chain.*) Assume that $D \subset (0, 1)$ is countable and can be enumerated by $i = 0, \pm 1, \pm 2$, so that $\rho_i := \rho_{z_i}$ for $i \geq 0$ satisfies $\rho_0 < \rho_1 < \cdots < 1$ and $\lim_{i\to\infty} \rho_i = 1$. (Enumeration with labels $i = -1, -2, \ldots$ does not matter.) Set $\upsilon(z) = 1$ and

$$\tau_n(z_0, z_j) = \tau_n(z_i, z_{i+j}) = \beta_n \delta(j, n), \quad \forall i, j \in \mathbb{Z}.$$

Here, as earlier, β_n is a scaling constant depending on n (again irrelevant for the invariant measure of the process (N, Z)). Then, $\mathbf{T}_n = \big(\tau_n(z, z')\big)$ generates a null-recurrent Markov chain $\tilde{Z}_n$ with the invariant measure $\upsilon(z) = 1$, $z \in D$. Thus, the random traffic intensity $\rho_{\tilde{Z}_n}$, depending on both the state of the queue and the environment, will approach the critical value 1 infinitely often. However, under the condition (2.4), the resulting Markov process (N, Z) is positive recurrent, with an invariant measure $\eta(n, z) = \rho^n(z)$ for $(n, z) \in \mathbb{Z}_+ \times D$.

3 *M*/*M*/1 Queue in an interactive diffusive environment

3.1 Reflected jump diffusions

In this section, we consider the queue with λ and μ dependent on a diffusive environment process $Z(t)$. First, let us define the dynamics of this environment process as a reflected (jump) diffusion in a certain domain in $\mathbb{R}^d$.

It is instrumental to recapitulate some basic notion. A *domain* in $\mathbb{R}^d$ is the closure of an open connected subset. A domain D is called *smooth* if its boundary ∂D is

a $(d-1)$-dimensional C^2 manifold. Take m smooth domains $D_1, \ldots, D_m$ in $\mathbb{R}^d$. Assume $D = \cap_{i=1}^m D_i$ has boundary ∂D with m *faces* $F_i := \partial D \cap \partial D_i$ which are $(d-1)$-dimensional manifolds with an edge, and such that all m domains are essential: Removal from the intersection of any domain will change the result. Then, D is called a *piecewise smooth domain* in $\mathbb{R}^d$. Define by $\mathbf{n}_i(z)$ the inward unit normal vector to ∂D_i at $z \in F_i$. *Inward* in this case is defined as pointing inside D_i, even if this is not inside D. An important example is a *convex polyhedron* with D_i being half-spaces. Of particular interest is the positive orthant $D = \mathbb{R}^d_+$. Of course, smooth domains also belong to this class of domains, with $m = 1$.

Take continuous functions $g : D \to \mathbb{R}^d$ and $\Sigma : D \to \mathbb{R}^{d\times d}$ such that the matrix $\Sigma(z) = (a_{ij}(z))$ is symmetric and positive definite for all $z \in D$, and there exists a $\delta > 0$ such that $\Sigma(z)v \cdot v \geq \delta\|v\|^2$ for all $v \in \mathbb{R}^d$ and $z \in D$. For every $z \in D$, define a finite measure $\varpi(z, \cdot)$ on D such that $\varpi(z, \cdot) \Rightarrow \varpi(z^0, \cdot)$ as $z \to z^0$ in D. Recall that $\Rightarrow$ denotes weak convergence. Take $r_i : F_i \to \mathbb{R}^d$: continuous functions, pointing inside D; that is, $r_i(z) \cdot \mathbf{n}_i(z) > 0$ for $i = 1, \ldots, m,\ z \in F_i$. Let us define a *reflected jump diffusion*: a process $Z = \{Z(t) : t \geq 0\}$ in D with *drift vector field* g, *diffusion matrix field* Σ, *jump measures* $\varpi(z, \cdot)$, and *reflection vector fields* $r_1, \ldots, r_m$.

This process will be adapted and right continuous with left limits. Take a d-dimensional Brownian motion $B = \{B(t) : t \geq 0\}$, adapted to the filtration. Take continuous nondecreasing processes $\ell_i = \{\ell_i(t) : t \geq 0\}$ for $i = 1, \ldots, m$ such that ℓ_i can grow only when $Z(t) \in F_i$, another right-continuous process with left limits $Z = (Z(t),\ t \geq 0)$ with values in D, and yet another process $\mathcal{N} = \{\mathcal{N}(t) : t \geq 0\}$ which is right-continuous and piecewise constant, with jump measure $\varpi(Z(t-), \cdot)$, and such that

$$\mathrm{d}Z(t) = g(Z(t))\,\mathrm{d}t + \Sigma^{1/2}(Z(t))\,\mathrm{d}B(t) + \mathcal{N}(t) + \sum_{i=1}^{m} r_i(Z(t))\,\mathrm{d}\ell_i(t), \quad t \geq 0. \tag{3.1}$$

We assume the Eq. (3.1) has a well-defined unique weak solution and forms a Feller continuous strong Markov semigroup, with generator

$$\mathcal{A}f(z) = g(z) \cdot \nabla f(z) + \frac{1}{2}\operatorname{tr}(\Sigma(z)\nabla^2 f(z)) + \int_D (f(z') - f(z))\varpi(z, \mathrm{d}z'), \tag{3.2}$$

which consists of a nondegenerate uniformly elliptic diffusion and a state-dependent finite jump measure. This existence and uniqueness were proved under Lipschitz conditions on vector field $g(\cdot)$ and the matrix $(a_{ij}(\cdot))$, as well as continuity of $r_i(\cdot)$ for each $i = 1, \ldots, m$, and some additional technical conditions. The case without jumps was proved in [28]; the general case follows from the standard construction by *piecing out* [39]. The reflection at the boundary translates into boundary conditions for (3.2):

$$r_i(z) \cdot \nabla f(z) = 0, \quad z \in F_i, \quad i = 1, \ldots, m. \tag{3.3}$$

The dynamics of this process can be described as follows:

- As long as it is strictly inside D, this process behaves as a jump diffusion in d dimensions with drift vector field g, diffusion matrix field Σ and the family ϖ of jump measures. These jump measures are such that the process does not jump out of D.
- At a point $z \in F_i, i = 1, \ldots, m$, it is reflected back inside the domain D, according to the vector $r_i(z)$.
- If it hits the lower-dimensional edges, intersections of two or more faces $F_1, \ldots, F_m$, it is reflected back inside D according to a positive linear combination of reflection vectors corresponding to these intersecting faces.

Normal reflection corresponds to the case when $r_i(z) = \mathbf{n}_i(z)$, where $z \in F_i$ and $i = 1, \ldots, m$.

Remark 3.1 In the case of a diffusion without reflection, the state space may be $\mathbb{R}^d$, or still some subset D. The latter happens if the drift coefficient is sufficiently large to compel the process to stay in a certain domain. An example of this is the drift for a Bessel process on the half-line; see [21, Chapter 3, Problem 3.23].

In the case $d = 1$, for a reflection on $[a, b]$, we have normal reflection and the boundaries consisting of two pieces $\{a\}$ and $\{b\}$. For a reflection on $[a, \infty)$, we have normal reflection again, with the boundary $\{a\}$.

3.2 Construction of the joint Markov process

Let us now use symbol z for a point in D (instead of x). Take continuous functions $\lambda, \mu : D \to (0, \infty)$ with $\lambda(z) \le \mu(z)$ for $z \in D$. Define the *traffic intensity*

$$\rho(z) := \frac{\lambda(z)}{\mu(z)} \le 1. \tag{3.4}$$

For every $z \in D$, consider an $M/M/1$ queue with arrival intensity $\lambda(z)$ and service intensity $\mu(z)$, where n is the state of this queue. The process $\tilde{N}$ counting the number of jobs in the system, called the *queueing process* in the sequel, is a continuous-time Markov process on $\mathbb{Z}_+$ with generator

$$\mathcal{M}_z f(n) = \lambda(z)(f(n+1) - f(n)) + 1_{\{n \neq 0\}}\mu(z)(f(n-1) - f(n)). \tag{3.5}$$

We now consider a $(1+d)$-dimensional Markov process $(N, Z) = \{(N(t), Z(t)) : t \ge 0\}$ with values in $\mathbb{Z}_+ \times D$ which evolves as follows:

(a) If $N(t) = n \in \mathbb{Z}_+$, then Z behaves as a reflected jump diffusion in D with generator $\rho^{-n}(z)\beta_n \mathcal{A}$ and reflection fields $r_1, \ldots, r_m$.
(b) If $Z(t) = z$, then $N(t)$ jumps from n to $n+1$ with intensity $\lambda(z)$, and (if $n \neq 0$) to $n-1$ with intensity $\mu(z)$.

Here, β_n is the *variability coefficient* for the diffusive environment, depending on the queueing state n, and $\rho^{-n}(z)$ is the *queueing impact* factor, capturing the impact from

the traffic intensity (congestion) from the queueing process. The component N can be informally described as the queueing process of an $M/M/1$ queue with arrival and service rates, $\lambda(z)$ and $\mu(z)$, respectively. These rates depend on an auxiliary process Z. The dynamics of Z, however, depend on the current position of this queueing process. Therefore, we call such a system *an $M/M/1$ queue in an interactive diffusive environment.*

The joint dynamics are described via a combined Markov process (N, Z) with the following generator:

$$\mathcal{L}f(n,z) = \mathcal{M}_z f(n,z) + \beta_n \rho^{-n}(z)\mathcal{A}f(n,z), \quad f \in \mathcal{D}. \tag{3.6}$$

Here, $\mathcal{D}$ stands for the following subspace of the domain of $\mathcal{L}$:

$$\begin{aligned} \mathcal{D} &:= \{f : \mathbb{Z}_+ \times D \to \mathbb{R} \mid \forall n \in \mathbb{Z}_+,\ f(n,\cdot) \in \mathcal{D}_D\}, \\ \mathcal{D}_D &:= \{f \in C_b^2(D) \mid r_i(z)\cdot\nabla f(z) = 0,\ z \in F_i,\ i = 1,\dots,m\}. \end{aligned} \tag{3.7}$$

(Note that we were intentionally loose on the domains of f in (3.2) and (3.5), but they are clear from this definition.) From the general theory of *piecing out*, it follows that this is a Feller process; see [39]. We denote by $C_b^2(D)$ the set of twice continuously differentiable functions $D \to \mathbb{R}$ which are bounded with their first and second derivatives (the last condition is automatically fulfilled for bounded D). This is a separable Banach space with the norm

$$\|f\|_{D,2} := \sup_{z\in D}\left(|f(z)| + \|\nabla f(z)\| + \|\nabla^2 f(z)\|\right).$$

Denote by $P^t(y,\cdot)$ the transition kernel of (N, Z), where $y = (n, z) \in \mathbb{Z}_+ \times D$.

We give three special cases to illustrate the construction above.

(a) *$M/M/1$ queue with an interactive diffusive arrival rate.*
Assume that $\lambda(z) = z$ and $\mu \equiv 1$. Let $D = [0, 1]$ and the generator $\mathcal{A}$ in (3.2) be that of a reflected diffusion in $(0, 1)$ without jumps. The reflections at 0 and 1 correspond to the Neumann boundary conditions

$$\frac{\partial}{\partial z} f(n, 0+) = 0, \quad \frac{\partial}{\partial z} f(n, 1-) = 0, \quad \forall n \in \mathbb{Z}_+.$$

(b) *$M/M/1$ queue with an interactive diffusive service rate.*
Assume that $\lambda \equiv 1$ and $\mu(z) = z$. Let $D = [\mu_0, \infty)$ for some $\mu_0 \geq 1$ and the generator $\mathcal{A}$ in (3.2) be that of a simple RBM on $[\mu_0, \infty)$ without jumps. The reflection at μ_0 satisfies the Neumann boundary condition.

(c) *$M/M/1$ queue with both diffusive arrival and service rates.*
Take $D = \{(z_1, z_2) \in \mathbb{R}_+^2 : z_2 \leq z_1\}$ be a cone in the positive orthant and the generator $\mathcal{A}$ in (3.2) be that of a two-dimensional Brownian motion in D with normal reflections at the boundary. Let $(\lambda(z), \mu(z)) = z$. Then, the arrival and service rates of the $M/M/1$ queue follow the dynamics of a reflected RBM in D in the interactive manner described above.

3.3 Invariant measures

We need the following assumptions on some properties of the reflected jump-diffusion process. The first assumption states that, for each level, there exists a steady-state distribution. The second assumption ensures that the whole process has a steady-state distribution.

Assumption 3.1 Assume the (reflected) jump diffusion with generator $\mathcal{A}$ is positive recurrent and has a unique stationary/invariant measure ν_D, together with *boundary measures* ν_{F_i}, $i = 1, \ldots, m$. This means that the stationary copy of this process $\tilde{Z}^* = \{\tilde{Z}^*(t) : t \geq 0\}$ with $\tilde{Z}^*(t) \sim \nu_D$ for $t \geq 0$ satisfies the following condition: For every $t \geq 0$, each $i = 1, \ldots, m$, and every bounded function $f : F_i \to \mathbb{R}$,

$$\mathbb{E} \int_0^t f(\tilde{Z}^*(s)) \, \mathrm{d}\ell_i(s) = t \int_{F_i} f(z) \, \nu_{F_i}(\mathrm{d}z), \tag{3.8}$$

where $\ell_i(s)$ is the nondecreasing process in (3.1).

Assumption 3.2 The measure $\nu_D(\cdot)$ satisfies

$$\Xi := \int_D \frac{\nu_D(\mathrm{d}z)}{1 - \rho(z)} = \sum_{n=0}^{\infty} \int_D \rho^n(z) \nu_D(\mathrm{d}z) \; < \infty. \tag{3.9}$$

Note that Ξ is the normalization constant in the joint invariant measure of (N, Z) in (3.10). This invariant measure on each boundary F_i has value zero.

Remark 3.2 Similar to (3.8), we can define the concept of *boundary measures* for the joint process (N, Z). First, construct the boundary process $\ell_i = (\ell_i(t), \, t \geq 0)$ for the component Z and face F_i of the boundary ∂D. Assume $0 = \rho_0 < \rho_1 < \ldots$ are jump times for N. Then, $Z(\rho_k + t)$ for $t \in [0, \rho_{k+1} - \rho_k]$ behaves as a reflected jump diffusion on D with generator $\rho^{-n_k}(z)\beta_{n_k}\mathcal{A}$ and reflection fields $r_1, \ldots, r_m$, with $N(t) = n_k$ for $t \in [\rho_k, \rho_{k+1})$. Thus, there exists a continuous nondecreasing process $\ell_i^{(k)}(t)$, $t \in [0, \rho_{k+1} - \rho_k]$, such that (3.1) holds with adjusted drift vector field, diffusion matrix field, and jump measure family. Define

$$\ell_i(t) = \ell_i(\rho_k) + \ell_i^{(k)}(t - \rho_k), \quad t \in [\rho_k, \rho_{k+1}],$$

using induction over k. This defines $\ell_i = (\ell_i(t), \, t \geq 0)$ for $i = 1, \ldots, m$. Next, define a *boundary measure* ν_{F_i} on the face F_i corresponding to a stationary distribution π for this joint process (N, Z): Take the corresponding stationary copy (N^*, Z^*) with $(N^*(t), Z^*(t)) \sim \pi$ for $t \geq 0$. For a bounded function $f : \mathbb{Z}_+ \times F_i \to \mathbb{R}$ and $t \geq 0$,

$$\mathbb{E} \int_0^t f(\tilde{N}^*(s), \tilde{Z}^*(s)) \, \mathrm{d}\ell_i(s) = t \sum_{n=0}^{\infty} \int_{F_i} f(n, z) \, \nu_{F_i}(\{n\} \times \mathrm{d}z).$$

Now, we are ready to state and prove the main result of this section.

Theorem 3.1 *Under Assumptions 3.1 and 3.2, there is a unique invariant measure for* (N, Z)*:*

$$\pi(\{n\}, \mathrm{d}z) = \Xi^{-1}\rho^n(z)\nu_D(\mathrm{d}z). \tag{3.10}$$

The corresponding boundary measures π_i *for* F_i *(if there is reflection) are given by*

$$\pi_i(\{n\}, \mathrm{d}z) = \Xi^{-1}\rho^n(z)\nu_{F_i}(\mathrm{d}z), \quad i = 1, \ldots, m. \tag{3.11}$$

Finally, this Markov process is ergodic: For every $y \in \mathbb{Z}_+ \times D$*,*

$$\|P^t(y, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \to 0 \quad as \quad t \to \infty. \tag{3.12}$$

Proof From stationarity, we immediately get that, for all $f \in C_b^2(D)$,

$$\int_D \mathcal{A}f(z)\, \nu_D(\mathrm{d}z) + \sum_{i=1}^{m} \int_{F_i} r_i \cdot \nabla f(z)\, \nu_{F_i}(\mathrm{d}z) = 0. \tag{3.13}$$

This is called the *basic adjoint relationship* in the literature. We refer to [43] for its deduction in the case of a convex polyhedron; the same is true for a general piecewise smooth domain D, as in our case. Apply [27, Theorem 1.7, Theorem 2.2, Lemma 2.4, Remark 2.5] using their notation, with the state space $E = \mathbb{Z}_+ \times D$; $U = \{0, 1, \ldots, m\}$, where the point 0 corresponds to the domain D itself, and $i = 1, \ldots, m$ correspond to faces $F_1, \ldots, F_m$ of the boundary; for all $z \in D$, $n \in \mathbb{Z}_+$, and $u \in U$,

$$\begin{aligned}
\mu_0(\{u\} \times \{n\} \times \mathrm{d}z) &= 1(u = 0)\, \rho^n(z)\nu_D(\mathrm{d}z),\\
\mu_1(\{u\} \times \{n\} \times \mathrm{d}z) &= 1(u \neq 0)\, \rho^n(z)\nu_{F_u}(\mathrm{d}z);\\
\mu_0^E(\{n\} \times \mathrm{d}z) &= \rho^n(z)\nu_D(\mathrm{d}z),\\
\mu_1^E(\{n\} \times \mathrm{d}z) &= \rho^n(z)\left[\nu_{F_1}(\mathrm{d}z) + \cdots + \nu_{F_m}(\mathrm{d}z)\right];\\
\eta_0((n, z), \{u\}) &= 1(u = 0),\\
\eta_1((n, z), \{u\}) &= 1(u \neq 0);\\
Af((n, z), u) &:= \mathcal{L}f(n, z), \quad \text{cf. (3.6)},\\
Bf((n, z), u) &:= 1(u \neq 0,\ z \in \partial D)\, \beta_n \rho^{-n}(z) r_u(z) \cdot \nabla f(z).
\end{aligned} \tag{3.14}$$

We need to check [27, Condition 1.2] on the absolutely continuous generator A and the singular generator B. Let

$$\mathcal{D} := \{f : \mathbb{Z}_+ \times D \to \mathbb{R} \mid \forall n \in \mathbb{Z}_+,\ f(n, \cdot) \in C_b^2(D)\}.$$

Part (i) requires that $A, B : \mathcal{D} \subset C_b(E) \to C(E \times U)$, and the unity function $\mathbf{1}(n, z) = 1$ for $(n, z) \in E$ satisfies $\mathbf{1} \in \mathcal{D}$, $A\mathbf{1} = 0$, and $B\mathbf{1} = 0$. This is trivially satisfied.

Part (ii) requires that there exist $\psi_A(n,z)$ and $\psi_B(n,z)$ in $C(E\times U)$, $\psi_A,\psi_B\geq 1$, and constants a_f, b_f, $f\in\mathcal{D}$ such that

$$|Af(x,u)|\leq a_f\psi_A(x,u),\quad |Bf(x,u)|\leq b_f\psi_B(x,u),\quad \forall (x,u)\in\mathcal{U},$$

where $\mathcal{U}$ is any closed set of $E\times U$. We can take $a_f=b_f:=\|f\|_{D,2}$, and

$$\psi_A=\psi_B=\|A(z)\|+\sum_{i=1}^{m}\|r_i(z)\|+\rho^n(z).$$

Part (iii) requires the following: Defining $(A_0,B_0)=\{(f,\psi_A^{-1}Af,\psi_B^{-1}Bf):f\in\mathcal{D}\}$, (A_0,B_0) is separable in the sense that there exists a countable collection $\{g_k\}\subset\mathcal{D}$ such that (A_0,B_0) is contained in the bounded, pointwise closure of the linear span of $\{(g_k,A_0g_k,B_0g_k)=(g_k,\psi_A^{-1}Ag_k,\psi_B^{-1}g_k)\}$. This is proved by taking a dense countable subset Υ of $C_b^2(D)$ in the norm $\|\cdot\|_2$, and then taking a countable subset

$$\bigcup_{n=0}^{\infty}[\Upsilon]^n\subseteq\mathcal{D}\simeq\left[C_b^2(D)\right]^{\mathbb{Z}_+}.$$

This subset is dense in the sense of pointwise convergence.

Part (iv) requires that, for each $u\in U$, the operators A_u and B_u defined by $A_uf(x)=Af(x,u)$ and $B_uf(x)=Bf(x,u)$ are pre-generators. This follows from [27, Remark 1.1], because all these operators satisfy the positive maximum principle.

Part (v) requires that $\mathcal{D}$ is closed under multiplication and separates points. This follows directly from the definition.

Finally, we need to prove the main condition as in [27, Theorem 1.7, (1.17)]:

$$\int_{E\times U}Af(x,u)\,\mu_0(\mathrm{d}x\times\mathrm{d}u)+\int_{E\times U}Bf(x,u)\,\mu_1(\mathrm{d}x\times\mathrm{d}u)=0. \tag{3.15}$$

From (3.14) and (3.6), canceling β_n and $\rho^n(z)$ when appropriate, we rewrite the left-hand side of (3.15) as follows:

$$\begin{aligned}\sum_{n=0}^{\infty}\beta_n\left[\int_D\mathcal{A}f(n,z)\,\nu_D(\mathrm{d}z)+\sum_{i=1}^{m}\int_{F_i}r_i(z)\cdot\nabla f(n,z)\,\nu_{F_i}(\mathrm{d}z)\right]\\ +\int_D\sum_{n=0}^{\infty}\rho^n(z)\mathcal{M}_zf(n,\cdot)\,\nu_D(\mathrm{d}z).\end{aligned} \tag{3.16}$$

The first line in (3.16) is equal to zero; this follows from (3.13). Let us show that the second line in (3.16) is equal to zero, too. For every $z\in D$, $\mathcal{M}_z$ is the generator of the $M/M/1$ queue with arrival and service rates $\lambda(z)$ and $\mu(z)$, respectively. This queue has a geometric stationary distribution $(1-\rho(z))\rho^n(z)$, $n\in\mathbb{Z}_+$. Thus,

$$\sum_{n=0}^{\infty}\rho^n(z)\mathcal{M}_zf(n,\cdot)=0,\quad z\in D. \tag{3.17}$$

Integrating (3.17) with respect to $\mu_D(\mathrm{d}z)$, we get that the second line in (3.16) is equal to zero. We interchanged integration and series, which we can do by uniform boundedness of f combined with Assumption 3.2. This completes the proof of (3.15), and with it [27, (1.17)]. Next, $K_1 := \partial D$ is the closed support for μ_1^E. By [27, Remark 2.5], the results of [27, Lemma 2.4] hold, and we can apply [27, Theorem 2.2 (f)] and obtain the stationary copy of our process (N, Z).

We have written the proof for reflected diffusions. For nonreflected ones, it is simpler: We can simply verify (3.13), which in our case then becomes

$$\int_D \mathcal{A}f(z)\,\nu_D(\mathrm{d}z) = 0, \quad f \in C^2(D). \tag{3.18}$$

This is done similarly to the computation above, but without all boundary terms. The lack of reflection obviates the need to apply results cited above from [27].

Finally, ergodicity follows from [32, Theorem 6.1] in the following way. (For terminology, we refer the reader to [32].) Our process is positive Harris recurrent, since the invariant measure is finite. Meanwhile, every skeleton chain is irreducible, because of the following *irreducibility property:* Define a Lebesgue measure on $\mathbb{Z}_+ \times D$ as a sum of Lebesgue measures on each layer of this set. □

Lemma 3.2 *For every $n \in \mathbb{Z}_+$, $z \in D$, and a subset $G \subseteq \mathbb{Z}_+ \times D$ of positive Lebesgue measure,*

$$P^t((n, z), G) > 0. \tag{3.19}$$

Proof Without loss of generality, assume $G = \{m\} \times E$ for a subset $E \subseteq D$ of positive Lebesgue measure, and $m \geq n$. We prove the statement (3.19) by induction over m.

Induction base $m = n$. Consider the probability

$$P^t(y, G) = \mathbb{P}_{(n,z)}(N(t) = n,\ Z(t) \in E)$$

that, starting from $y = (n, z)$, the joint process (N, Z) at time t will be in $\{n\} \times E$. This probability is bounded from below by

$$P^t(y, G) \geq Q_n^t(z, E) := \mathbb{P}_{(n,z)}\big(Z(t) \in E,\ N(s) = n,\ \forall\, s \in [0, t]\big). \tag{3.20}$$

This probability $Q_n^t(z, E)$, in turn, is estimated from below by (with $z_* > 0$ fixed later)

$$\begin{aligned} Q_n^t(z, E) &\geq \tilde{Q}_n^t(z, z_*, E) := \mathbb{P}_{(n,z)}\big(Z(t) \in E,\ \|Z(t)\| \leq z_*;\ N(s){=}n,\ \forall\, s{\in}[0, t]\big) \\ &\geq \exp\left(-t \max_{\|z\| \leq z_*} (\lambda(z) + \mu(z))\right) \cdot q_*. \end{aligned} \tag{3.21}$$

Here, q_* is the probability that, starting from $Z_n(0) = z$, the reflected jump diffusion Z_n in D with generator $\rho^{-n}(z)\mathcal{L}$ and reflection vector fields $r_1, \ldots, r_m$ ends at $Z_n(t) \in E$ and $\|Z_n(s)\| \le z_*$ for $s \in [0, t]$. It follows from known properties of a reflected jump diffusion with nonsingular covariance matrix $\Sigma(\cdot)$ that $q_* > 0$ for large enough $z_* > 0$. This, together with (3.20) and (3.21), proves that

$$P^t(y, G) \ge Q_n^t(z, E) \ge \tilde{Q}_n^t(z, z_*, E) > 0. \tag{3.22}$$

Thus, we have proved the statement (3.19) for $m = n$.

Induction step First, consider the case $m = n + 1$. This probability $P^t(y, G)$ is estimated from below by the probability that, for some time $\tau \in [0, t]$, the process N will stay an level n, then jump at time τ to level $n + 1$, and stay there until time t, and $Z(t) \in E$. If $\hat{\mu}$ is the distribution of τ (which is a positive measure on $[0, t]$), then

$$P^t(y, G) \ge \int_0^t \int_D Q_n^s(y, \mathrm{d}w)\, Q_{n+1}^{t-s}(w, E)\, \hat{\mu}(\mathrm{d}s). \tag{3.23}$$

It suffices to show that the double integral on the right-hand side of (3.23) is positive. Indeed, from (3.22), we get that $Q_n^s(y, E') > 0$ for $E' \subseteq D$ of positive Lebesgue measure, and $Q_{n+1}^{t-s}(w, E) > 0$. In addition, $\hat{\mu}$ is a positive measure on $[0, t]$. Use twice the observation that the integral of a positive function over a positive measure is positive, and complete the proof that the right-hand side (and therefore the left-hand side) in (3.23) is positive.

Assuming we proved (3.19) for $m = n+k, k \ge 0$, let us prove this for $m = n+k+1$:

$$P^t(y, G) \ge \int_D P^{t/2}(y, (n + k, \mathrm{d}w))\, P^{t/2}((n + k, w), \{n + k + 1\} \times E) > 0. \tag{3.24}$$

This follows from the same logic: The function $P^{t/2}((n + k, w), \{n + k + 1\} \times E)$ is positive by the previous part of the induction step, applied to $n + k$ instead of n, and to $n + k + 1$ instead of $m = n + 1$. The measure $P^{t/2}(y, (n + k, \mathrm{d}w))$ is positive by the induction hypothesis. This completes the proof of this lemma. □

Using Lemma 3.2, we have shown ergodicity as in (3.12). Earlier, we have proved (3.10) and (3.11). Thus, we have completed the proof of Theorem 3.1. □

Remark 2.3 The crucial property is that for each $n \in \mathbb{Z}_+$, $z \in D$, $t > 0$, $V, V' \subseteq D$,

$$\begin{aligned} &\int_{D\times D} \mathbf{1}(z \in V, z' \in V')\mathfrak{p}^t(z, z')\nu_D(\mathrm{d}z)\nu_D(\mathrm{d}z') \\ &\quad = \textstyle\int_{D\times D} \mathbf{1}(z \in V, z' \in V')\mathfrak{p}^t(z', z)\nu_D(\mathrm{d}z)\nu_D(\mathrm{d}z'). \end{aligned} \tag{3.25}$$

In fact, further generalizations depend on whether an analog of this equality can be established. Here, $\mathfrak{p}^t$ stands for the transition density for the diffusion with generator $\mathcal{A}$ in (3.2).

3.4 A more general setup

We offer a similar result under a more general feedback scheme. For $n = 1, 2, \ldots$, fix a piecewise smooth domain $D_n \subseteq D$ with m_n faces of the boundary ∂D_n,

$$F_1^{(n)}, \ldots, F_{m_n}^{(n)}, \tag{3.26}$$

and corresponding reflection vector fields

$$r_i^{(n)} : F_i^{(n)} \to \mathbb{R}^d. \tag{3.27}$$

For each $n \in \mathbb{Z}_+$, this domain D_n, its boundary ∂D_n with faces (3.26), and reflection vector fields (3.27) satisfy the same assumptions enunciated at the very beginning of Sect. 3, as the original domain D and reflection vector fields $r_1, \ldots, r_m$. In addition, we impose the following assumptions on the domains D and D_n.

Assumption 3.3 For all $n \in \mathbb{Z}_+$, $D_n \cap D_{n+1}$ contains an open subset of D; and $D = \cup_{n\in\mathbb{Z}_+} D_n$.

For every level $N(t) = n$ of the queue-size component, the environment variable $z \in D$ is kept fixed when $z \in D \setminus D_n$ and follows a reflected jump-diffusion process in D_n as in (3.1) where parameters vary with n. In other words, the process $\widetilde{Z}_n$ lives in D, but its mechanism depends on n. The generator $\mathcal{A}_n$ of $\widetilde{Z}_n$ has the form

$$\mathcal{A}_n f(z) = g(z) \cdot \nabla f(z) + \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} a_{ij}(z) \frac{\partial^2 f(z)}{\partial z_i \partial z_j} + \int_{D_n} (f(z') - f(z)) \varpi(z, \mathrm{d}z'), \tag{3.28}$$

for $z \in D_n$, and $\mathcal{A}_n f(z) = 0$ for other z. The generator $\mathcal{L}$ of the joint process, instead of (3.6), has the following form:

$$\mathcal{L} f(n, z) = \mathcal{M}_z f(n, z) + \beta_n \rho^{-n}(z) \mathcal{A}_n f(n, z), \quad f \in \widetilde{\mathcal{D}}. \tag{3.29}$$

Here, $\widetilde{\mathcal{D}}$ is the following domain, defined similarly to (3.7):

$$\widetilde{\mathcal{D}} := \{f : \mathbb{Z}_+ \times D \to \mathbb{R} \mid \forall n \in \mathbb{Z}_+, \ f(n, \cdot) \in \mathcal{D}^{(n)}\}, \tag{3.30}$$

$$\mathcal{D}^{(n)} := \{f : D \to \mathbb{R} \mid f \in C_b(\overline{D}), \ f|_{D_n} \in C_b^2(D_n) \cap C_b^1(\overline{D}_n), \\ r_i^{(n)}(z) \cdot \nabla f(z) = 0, \ z \in F_i^{(n)}, \ i = 1, \ldots, m_n\}. \tag{3.31}$$

Note that in this setup, the dependence of the generator $\mathcal{A}_n$ on the queueing state n is only through the domain D_n, while the drift vector field $g(\cdot)$, covariance matrix field $\Sigma(\cdot)$, and jump measure family $\varpi(\cdot, \cdot)$ are all independent of n; see also Examples 3.1 and 3.2 .

Let us impose assumptions on $\mathcal{A}_n$, similar to Assumptions 3.1 and 3.2 .

Assumption 3.4 For every $n \in \mathbb{Z}_+$, the above (reflected) jump diffusion in D_n has a unique invariant distribution $\nu_{D_n}^{(n)}$, with corresponding boundary measures $\nu_{F_i^{(n)}}^{(n)}$, $i = 1, \ldots, m_n$.

Assumption 3.5 There exists a finite measure υ on D whose restriction $\nu_{D_n}^{(n)}$ on D_n is a stationary measure for $\widetilde{Z}_n$, for every n.

This independence of the invariant measure υ of n is similar to Assumption 2.1 in Sect. 3.

Assumption 3.6 We have

$$\Xi := \sum_{n=0}^{\infty} \int_{D_n} \rho^n(z) \upsilon(\mathrm{d}z) \; < \infty. \tag{3.32}$$

Under these assumptions, we obtain the following theorem, analogous to Theorem 3.1.

Theorem 3.3 *Under Assumptions 3.3–3.6, the combined process (N, Z) with the generator $\mathcal{L}$ from (3.28) has a unique invariant probability distribution π given by*

$$\pi(\{n\}, \mathrm{d}z) = \Xi^{-1} \rho^n(z) \upsilon(\mathrm{d}z). \tag{3.33}$$

The corresponding boundary measures ν_{F_i} for F_i (if there is reflection) are given by

$$\nu_{F_i}(\{n\}, \mathrm{d}z) = \Xi^{-1} \rho^n(z) \nu_{F_i}^{(n)}(\mathrm{d}z), \quad i = 1, \ldots, m_n. \tag{3.34}$$

Finally, this process is ergodic in the sense of (3.12).

Proof For the proof of the stationary measure, we proceed very similarly to the proof of Theorem 3.1, except that we change (3.14):

$$\begin{aligned}
\mu_0(\{u\} \times \{n\} \times \mathrm{d}z) &= 1(u = 0)\, \rho^n(z)\, \upsilon(\mathrm{d}z), \\
\mu_1(\{u\} \times \{n\} \times \mathrm{d}z) &= 1(u \neq 0)\, \rho^n(z)\, \nu_{F_u}^{(n)}(\mathrm{d}z); \\
\mu_0^E(\{n\} \times \mathrm{d}z) &= \rho^n(z)\, \upsilon(\mathrm{d}z), \\
\mu_1^E(\{n\} \times \mathrm{d}z) &= \rho^n(z) \nu_{F_i}^{(n)}(\mathrm{d}z), \quad z \in F_i,\ i = 1, \ldots, m; \\
\eta_0((n, z), \{u\}) &= 1(u = 0), \\
\eta_1((n, z), \{u\}) &= 1(u \neq 0); \\
Af((n, z), u) &:= \mathcal{L} f(n, z), \quad \text{cf. (3.29)}, \\
Bf((n, z), u) &:= 1(u \neq 0,\ z \in \partial D)\, \rho^{-n}(z)\, r_u(z) \cdot \nabla f(z).
\end{aligned} \tag{3.35}$$

To prove ergodicity as in (3.12), similarly to Theorem 3.1, we show an analog of Lemma 3.2:

Lemma 3.4 *For all $n, m \in \mathbb{Z}_+$, $z \in D$, and a subset $G \subseteq \mathbb{Z}_+ \times D$ of positive Lebesgue measure*

$$P^t((n,z),G) > 0. \tag{3.36}$$

Proof Similarly to Lemma 3.4, without loss of generality, assume $G = \{m\} \times E$ for a subset $E \subseteq D$ of positive Lebesgue measure, and $m \geq n$.

Case (a) $z \in D_n$, $E \subseteq D_m$. We prove this statement similarly to Lemma 3.2, using induction over m. The induction base ($m = n$) can be shown as in (3.20). Induction step: for $m = n+1$, we prove this as in (3.23) (using the same notation), but we integrate over $D_n \cap D_{n+1}$ instead of D:

$$P^t((n,z),\{m\}\times E) \geq \int_0^t \int_{D_n \cap D_{n+1}} Q_n^s(y,\mathrm{d}w)\, Q_{n+1}^{t-s}(w,E)\, \hat{\mu}(\mathrm{d}s).$$

Assuming we proved this for $m = n+k$, let us prove this for $m = n+k+1$. Similarly to (3.24), but integrating over $D_{n+k} \cap D_{n+k+1}$, we get

$$\begin{aligned} &P^t((n,z),\{m\}\times E) \\ &\geq \int_{D_{n+k}} P^{t/2}(y,(n+k,\mathrm{d}w))\, P^{t/2}((n+k,w),\{n+k+1\}\times E) > 0. \end{aligned}$$

This completes the proof of the induction step, and with it the proof of (3.36) in case (a).

Case (b) $z \in D_n$, $E \cap D_m = \varnothing$. (Clearly, we can reduce the case of a general E to these two cases (a) and (b).) Since $E \subseteq D = \cup_k D_k$, there exists a k such that $E \cap D_k$ has positive Lebesgue measure. Take the k with such a property which is closest to m. The process can get from (n,z) to $\{k\} \times (E \cap D_k)$ with positive probability in time $t/2$, using the path described in case (a) above. Afterward, for every $z \in E \cap D_k$, the process (N,Z) can jump from (k,z) to (m,z) in time $t/2$ with positive probability. Indeed, for l between k and m, we have $z \notin D_l$; thus, the component N will jump from k to m, and the environment component Z will stay constant at z.

Case (c) $z \notin D_n$. There exists a k such that $z \in D_k$, since $z \in D = \cup_k D_k$. Find such k which is closest to n. The process (N,Z) can get from (n,z) to (k,z) in time $t/2$ with positive probability: The queue component N will jump from n to k, and the environment component Z will stay constant at z, since $z \notin D_l$ for l between n and k. Starting the process from (k,z) instead of (n,z) now, we are back to cases (a) and (b). Applying results from these cases for $t/2$ instead of t, we prove (3.36) for $z \notin D_n$. □

We proved Lemma 3.4, and with it, we proved ergodicity (3.12) and thus Theorem 3.3. □

Now, we provide examples in which the generator of the diffusive component in the joint process depends on the queueing state in a nontrivial manner.

Example 3.1 Assume $D = [0,1]$, $D_n = [0, \alpha_n]$, $\lambda(z) = z$, $\mu(z) = 1$. Assume $\mathcal{A}_n$ is a reflected diffusion (without jumps):

$$\mathcal{A}_n f(z) = \frac{a^2(z)}{2} f''(z) + b(z) f'(z), \quad z \in D_n. \tag{3.37}$$

The functions $a, b \in C_b^2([0,1])$ are given, describing the local diffusion coefficient and the local drift of the processes $\widetilde{Z}_n$ in D_n, with $a(z) > 0$ for $z \in (0,1)$. Standard formulas from [8] guarantee that the measure υ on D has Lebesgue density

$$q(z) = \frac{2}{a^2(z)} \exp\left(\int_0^z \frac{2b(y)}{a^2(y)} \, \mathrm{d}y \right) \tag{3.38}$$

assuming that

$$\int_0^1 \frac{|b(y)|}{a^2(y)} \, \mathrm{d}y < \infty. \tag{3.39}$$

Assumption 3.6 becomes

$$\sum_{n=0}^{\infty} \int_0^{\alpha_n} \rho^n(z) q(z) \, \mathrm{d}z < \infty. \tag{3.40}$$

In particular, for $a(z) \equiv 1$ and $b(z) = \theta/(z-1)$ with $\theta > 0$, we get $q(z) = 2(1-z)^{-2\theta}$. If we choose

$$\alpha_n = \begin{cases} 1, & n < n_0, \\ \alpha_*, & n \geq n_0, \end{cases} \tag{3.41}$$

for some $\alpha_* \in (0,1)$ and $n_0 \in \mathbb{N}$, then (3.40) holds for $\theta \in (0, 1/2)$. If $a \equiv 1$ and $b \equiv 0$, then the driving process for the environment is a reflected Brownian motion, with υ being the Lebesgue measure, and $q(z) \equiv 1$.

This example can be interpreted as follows: We keep the service rate fixed, $\mu = 1$, while the arrival rate λ varies as a reflected diffusion on $[0,1]$ if the queue size n is less than an agreed threshold n_0. However, if n reaches level n_0, while $\lambda < \alpha^*$, we allow λ to vary only in a "safety range" $[0, \alpha^*]$. If n attains the level n_0, while $\lambda \geq \alpha^*$, we simply "freeze" λ until the queue size becomes $n_0 - 1$, at which time λ is again allowed to follow the diffusion on $[0,1]$.

Example 3.2 Fix the arrival rate $\lambda = 1$, while the service rate μ_n is subject to a reflected diffusion on the interval $D_n := [\alpha_n, \alpha^*] \subset [1, \alpha^*] =: D$ and kept unchanged in $D \setminus D_n$. Here, $\alpha_* > 1$ is a fixed constant. Here again, the generator $\mathcal{A}_n$ is given by (3.37), with $a, b \in C_b^2([1, \alpha^*])$; this operator from (3.37) acts on $f \in C^2([\alpha_n, \alpha^*])$ with boundary conditions $f'(\alpha_n) = f'(\alpha^*) = 0$. Instead of (3.38), we have

$$q(z) = \frac{2}{a^2(z)} \exp\left(\int_1^z \frac{2b(y)}{a^2(y)} \, \mathrm{d}y \right), \tag{3.42}$$

and instead of assumption (3.39), we have

$$\int_1^{\alpha^*} \frac{|b(y)|}{a^2(y)}\,\mathrm{d}y < \infty. \tag{3.43}$$

Assumption 3.6 becomes

$$\sum_{n=0}^{\infty} \int_{\alpha_n}^{\alpha^*} \rho^n(z) q(z)\,\mathrm{d}z < \infty. \tag{3.44}$$

As in Example 3.1, if $a \equiv 1$, $b \equiv 0$, then the driving process for the environment is a reflected Brownian motion, with υ being the Lebesgue measure, and $q(z) \equiv 1$.

4 Explicit rates of exponential convergence

4.1 A brief summary of results and methods

In this section, we prove (for both discrete-space and reflected diffusion environments) that for some constants $C, \varkappa > 0$, we have

$$\|P^t(x,\cdot) - \pi(\cdot)\|_{\mathrm{TV}} \le C(x) e^{-\varkappa t},\ x \in D,\ t \ge 0, \tag{4.1}$$

and estimate the constant $\varkappa$. We do this by *coupling:* Take two copies (N_1, Z_1) and (N_2, Z_2) of this process starting from $x_1 = (n_1, z_1)$ and $x_2 = (n_2, z_2)$. Couple them (that is, construct them on the same probability space) such that the *coupling time*

$$\tau := \inf\{t \ge 0 \mid N_1(t) = N_2(t),\ Z_1(t) = Z_2(t)\}$$

satisfies $\mathbb{E}[e^{\varkappa\tau}] < \infty$ for some constant $\varkappa > 0$. By the standard Lindvall inequality, we get

$$\|P^t(x_1,\cdot) - P^t(x_2,\cdot)\|_{\mathrm{TV}} \le \mathbb{E}\left[e^{\varkappa\tau}\right] e^{-\varkappa t},\ x \in D,\ t \ge 0. \tag{4.2}$$

We need only to integrate (4.2) with respect to $x_2 \sim \pi$ to get (4.1). To obtain such a coupling, we apply the following method: We wait until the queue component hits 0 for both copies. Thus, these queue components become coupled, that is, they are at the same point. Then, we wait until: (a) either one of these queue components jumps back to 1, or (b) the environment components become coupled. In case (b), we have coupled both copies. In case (a), we have failed and need to repeat this procedure. Each time, we succeed with positive probability (bounded from below). Thus, the number of tries is dominated by a geometric distribution.

To couple the environment components, we use the results of [37]; however, it is well known how to find the hitting time of zero by the $M/M/1$ queue [36]. Note that assuming exponential rates of convergence of $\mathcal{A}_n$ given each queue state n does not immediately imply the exponential rate of convergence of the joint process (N, Z). The particular multiplicative structure we consider in $\mathcal{A}_n$ enables us to obtain exponential

estimates for the coupling time constructed for the joint processes (N, Z) under the mild conditions imposed on $\mathcal{A}$ as well as the arrival and service rates.

4.2 Main statements

We impose two assumptions. The first assumes exponential bounds on the coupling time (uniform in state variables) associated with the generator $\mathcal{A}$.

Assumption 4.1 The domain $D \subseteq \mathbb{R}^d$ is bounded. There exist constants $\alpha > 1$ and $\gamma > 0$ such that, for all $z_1, z_2 \in D$, we can couple two processes Z_1, Z_2 with generator $\mathcal{A}$, starting from $Z_1(0) = z_1$ and $Z_2(0) = z_2$, in time $\tau_{z_1,z_2} := \inf\{t \geq 0 \mid Z_1(t) = Z_2(t)\}$, with

$$\mathbb{P}(\tau_{z_1,z_2} \geq t) \leq \alpha e^{-\gamma t}. \tag{4.3}$$

The other assumption is a stronger condition on the traffic intensity: In previous sections, we assumed it is less than 1, but now it has to be uniformly bounded away from 1.

Assumption 4.2 There exist constants $\overline{\lambda}, \overline{\mu} > 0$ which satisfy

$$\lambda(z) \leq \overline{\lambda} < \overline{\mu} \leq \mu(z), \quad z \in D.$$

From this Assumption 4.2,

$$\rho(z) \leq \overline{\rho} := \frac{\overline{\lambda}}{\overline{\mu}} < 1, \quad z \in D. \tag{4.4}$$

Next, define the function

$$m(c) := -\overline{\lambda}c - \overline{\mu}c^{-1} + (\overline{\lambda} + \overline{\mu}), \quad c \geq 1. \tag{4.5}$$

This function is concave, increasing on $[1, c^*]$ and decreasing on $[c^*, \infty)$, with $c^* := \overline{\rho}^{-1/2}$, and

$$m(1) = 0, \quad m(c^*) = \left(\sqrt{\overline{\mu}} - \sqrt{\overline{\lambda}}\right)^2.$$

Finally, define the function

$$\theta(\alpha, \beta, \gamma, a) := \frac{a\gamma}{(a-\beta)(\beta+\gamma-a)}\alpha^{(a-\beta)/\gamma} + \frac{\beta}{\beta - a}, \tag{4.6}$$

for any $\alpha > 1$, $\beta, \gamma > 0$, and $a \geq 0$.

Theorem 4.1 *Fix an initial condition $x_0 = (n_0, z_0) \in \mathbb{Z}_+ \times \mathbb{R}_+$. Under Assumptions 4.1 and 4.2 , for some constants $C > 0$ and $c \in (1, c_*)$,*

$$\|P^t((n_0, z_0), \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \le C\left(1 + c^{n_0}\right) e^{-\varkappa t}, \ t \ge 0. \tag{4.7}$$

where we can take any $\varkappa = (1-\varepsilon)m(c)$ *for* $\varepsilon \in (0,1)$ *and* $c \in (1, c^*)$ *such that*

$$c\theta(\alpha, \overline{\lambda}, \gamma, m(c)) < \left(1 - \alpha^{-\overline{\lambda}/\gamma} \frac{\gamma}{\overline{\lambda} + \gamma}\right)^{-\varepsilon/(1-\varepsilon)}. \tag{4.8}$$

The proof of the theorem is given at the end of this section. The only condition on the environment process is Assumption 4.1 on the coupling time with (uniformly) exponential tail for the environment process corresponding to $N(t) = 0$. It is natural to assume this condition also holds for the finite environment space.

Note that there exists a $c \in (1, c^*)$ such that (4.8) is satisfied. Indeed, the left-hand side of (4.8) is continuous with respect to c and is equal to 1 for $c = 1$, whereas the right-hand side of (4.8) is larger than one for any $\epsilon \in (0,1)$. However, to find a maximal rate of convergence, one needs to maximize $\varkappa$ over the space of two parameters (ε, c) which satisfy (4.8). Possible values of $\varkappa$ form an interval $[0, \varkappa_*)$, which does not contain its upper endpoint; therefore, we cannot claim that $\varkappa_*$ is itself a rate of convergence.

Compare this with the simple $M/M/1$ queue with constant rates, arrival rate $\overline{\lambda}$ and service rate $\overline{\mu}$, which has an exact rate of convergence $e^{-m(c)t}$ for $c > 0$ such that $m(c) > 0$ from (4.5). [36, Proposition 5.8] states that the upper bound, restricting to only the queueing process, is

$$\left(1 + \overline{\rho}^{-n/2}\right) \exp\left[-\left(\overline{\lambda}^{1/2} - \overline{\mu}^{1/2}\right)^2 t\right].$$

The constant in the exponent does not depend on n. Our result matches this rate.

After some modifications, this theorem is applicable not only to reflected diffusions from Sect. 2, but for the discrete environment space from Sect. 3. Here is its version:

Assumption 4.3 There exist constants $\alpha > 1$ and $\gamma > 0$ such that, for all $z_1, z_2 \in D$, we can couple two continuous-time Markov chains Z_1, Z_2 with common generator $\sigma(\cdot)\mathbf{T}_0$, starting from $Z_1(0) = z_1$ and $Z_2(0) = z_2$, in time τ_{z_1,z_2}, such that (4.3) holds.

Theorem 4.2 *Under Assumptions 4.2 and 4.3, the result (4.7) for* (c, ε) *satisfying (4.8) holds.*

4.3 On Assumptions 4.1 or 4.3

Below, we give examples of discrete and continuous environment processes which satisfy Assumptions 4.1 or 4.3.

4.3.1 Coupling of jump processes

First, let us start with discrete-space Markov chains. The relation between coupling times and mixing times (for $P^t(x, \cdot)$ to converge within a fixed TV distance from the

stationary distribution) is partially explored in [19]. There is a lot of existing literature on mixing times. For example, an extensive treatment of mixing times is given by Levin et al. [30]. The literature on coupling times is sparse. Much of the existing research is focused on γ from Assumption 4.1; see, for example, [9], but we need to know both α and γ. We could not find articles which estimate both of them. Thus, we present an elementary result, which we hope will be useful. The proof is in the Appendix.

Lemma 4.3 *Take a pure jump Markov process on the state space D (finite, countable, or a domain in $\mathbb{R}^d$) such that the family of jump measures $(\nu(x,\cdot))_{x\in D}$ obeys*

$$\Lambda := \sup_{x\in D} \lambda(x), \quad \lambda(x) := \nu(x, D), \quad x \in D,$$

and the family of probability measures

$$\overline{\nu}(x,\cdot) := \frac{1}{\Lambda}\nu(x,\cdot) + \frac{\Lambda - \lambda(x)}{\Lambda}\delta_{\{x\}}, \quad x \in D,$$

satisfies the following condition:

$$q := \sup_{x,y\in D} \|\overline{\nu}(x,\cdot) - \overline{\nu}(y,\cdot)\|_{\mathrm{TV}} < 1. \tag{4.9}$$

Then, the coupling times $\tau_{x,y}$ satisfy the following uniform estimate:

$$\mathbb{P}(\tau_{x,y} \geq t) \leq \exp\big(-(1-q)\Lambda t\big).$$

Remark 4.1 The same result is true if the process is a reflected jump diffusion with jump measures satisfying the conditions of Lemma 4.3.

Example 4.1 The condition (4.9) is not true if at least two measures $\overline{\nu}(x,\cdot)$ and $\overline{\nu}(y,\cdot)$ are mutually singular; that is, there exists a set $D_0 \subseteq D$ such that $\overline{\nu}(x, D_0) = 0$ but $\overline{\nu}(y, D_0) = 1$. Indeed, we then have

$$\|\nu(x,\cdot) - \nu(y,\cdot)\|_{\mathrm{TV}} \geq |\nu(x, D_0) - \nu(y, D_0)| = 1.$$

Example 4.2 Assume that, for all $x \in D$, $\nu(x,\cdot) \ll \mu(\cdot)$ for some σ-finite Borel measure μ on D. It can be the Lebesgue measure if D is a domain in $\mathbb{R}^d$, or the counting measure for discrete D. Define the Radon–Nikodym derivative

$$f(x,z) := \frac{\mathrm{d}\nu(x,\cdot)}{\mathrm{d}\mu(\cdot)}(z).$$

Then, condition (4.9) is equivalent to

$$\sup_{x,y\in D} \int_D |f(x,z) - f(y,z)|\mathrm{d}\mu(z) = q < 1.$$

For example, take a finite D (with m elements). Let μ be the counting measure, then $\nu(\cdot,\cdot)$ can be given by an $m \times m$ matrix (ν_{ij}) (with zero diagonal elements). The ith row gives the Radon–Nikodym derivative of $\nu(i,\cdot)$ with respect to μ. Thus, we obtain

$$q := \max_{i,j=1,\dots,m} \sum_{k=1}^{m} |\nu_{ik} - \nu_{jk}|.$$

4.3.2 Coupling of reflected diffusions

Now, consider a reflected diffusion on $[0,a]$. It is stochastically ordered, so every $\tau_{x,y}$ is stochastically dominated by $\mathcal{T}$, the hitting time of a starting from 0. Thus,

$$\mathbb{P}(\tau_{x,y} \geq t) \leq \mathbb{P}(\mathcal{T} \geq t).$$

Let us estimate the tail of $\mathcal{T}$. Take a nonreflected diffusion $Z^* = \{Z^*(t) : t \geq 0\}$ on the real line, with drift and diffusion coefficients

$$g^*(x) = \begin{cases} g(x), & x \geq 0, \\ -g(-x), & x < 0, \end{cases} \qquad \sigma^*(x) = \sigma(|x|), \quad x \in \mathbb{R}.$$

Let $\mathcal{T}^* := \inf\{t \geq 0 : |Z^*(t)| = a\}$. Then, the laws of $Z(\cdot \wedge \mathcal{T})$ and $Z^*(\cdot \wedge \mathcal{T}^*)$ are the same, and the laws of $\mathcal{T}$ and $\mathcal{T}^*$ are the same. Thus, we have reduced this to tail estimation for an exit time of a diffusion process from a strip $[-a,a]$.

Denote by $u^*(t,x)$ the probability that Z^* stays in $(-a,a)$ until at least time t, if $Z^*(0) = x$. Denote by $G(t,x,y)$ the transition density of this diffusion killed at $\pm a$, otherwise known as a Green's function (or heat kernel) of the infinitesimal generator $\mathcal{A}^*$ of Z^*. Then, the function u^* satisfies the initial-boundary value problem

$$\frac{\partial u^*}{\partial t} = \mathcal{A}^* u^*, \quad t \geq 0, \quad -a < x < a,$$

with initial and boundary conditions $u^*|_{t=0} = 1$ and $u|_{x=\pm a} = 0$. Thus, we can express

$$u^*(t,x) = \int_{-a}^{a} G(t,x,y)\,\mathrm{d}y.$$

Knowing the spectral decomposition of G gives us the exponent in (4.3). To find the constant A is a little harder, since it requires some information on the function G itself, or its eigenvalues. In some simple cases, however, it can be found explicitly. For example, for a RBM Z on $[0,a]$, the process Z^* is also a Brownian motion, and [21, Chapter 2, Problem 8.2] gives us an exact estimate.

4.4 Proof of Theorem 4.2

We proceed in seven steps. *Step 1*. It suffices to prove the following version of (4.2): For $(n_1,z_1),(n_2,z_2) \in \mathbb{Z}_+ \times D$,

$$\|P^t((n_1,z_1),\cdot)-P^t((n_2,z_2),\cdot)\|_{\mathrm{TV}}\le C_*\left(c^{n_1}+c^{n_2}\right)e^{-\varkappa t},\quad t\ge 0,\quad (4.10)$$

for some constant C_* (which will be determined below). Indeed, then we can rewrite (4.10) as follows: For every Borel subset $A\subseteq\mathbb{Z}_+\times D$,

$$\left|P^t((n_1,z_1),A)-P^t((n_2,z_2),A)\right|\le C_*\left(c^{n_1}+c^{n_2}\right)e^{-\varkappa t}.\quad (4.11)$$

Integrate (4.11) with respect to $(n_2,z_2)\sim\pi$. Note that the function $(n,z)\mapsto c^n$ is integrable with respect to π. Indeed, this integral is equal to

$$\Xi^{-1}\sum_{n=0}^{\infty}\int_D c^n\rho^n(z)\nu_D(\mathrm{d}z).$$

From (4.4), $\nu_D(D)=1$, and $c<c_*=\overline{\rho}^{-1/2}$,

$$\sum_{n=0}^{\infty}\int_D c^n\rho^n(z)\nu_D(\mathrm{d}z)\le\sum_{n=0}^{\infty}\overline{\rho}^{n/2}=(1-\overline{\rho}^{1/2})^{-1}<\infty.\quad (4.12)$$

Combining (4.11) and (4.12), we get (4.7).

Step 2. To get (4.10), we use *coupling:* As explained in the beginning of this section, we take on the same filtered probability space two copies $X_1=(N_1,Z_1)$ and $X_2=(N_2,Z_2)$ of this queue, starting from $x_1=(n_1,z_1)$ and $x_2=(n_2,z_2)$. Assume $\tau\equiv\tau(x_1,x_2)$ is a stopping time such that $X_1(t)=X_2(t)$ for $t\ge\tau$ a.s. Then, τ is called a *coupling time*. For every $t\ge 0$ and a function $f:\mathbb{Z}_+\times D\to\mathbb{R}$ with $|f|\le 1$, we can write

$$\begin{aligned}|\mathbb{E}f(X_1(t))-\mathbb{E}f(X_2(t))|&\le\left|\mathbb{E}\left[f(X_1(t))1_{\{\tau\le t\}}\right]-\mathbb{E}\left[f(X_2(t))1_{\{\tau\le t\}}\right]\right|\\&+\left|\mathbb{E}\left[f(X_1(t))1_{\{\tau>t\}}\right]-\mathbb{E}\left[f(X_2(t))1_{\{\tau>t\}}\right]\right|\le 2\mathbb{P}(\tau>t).\end{aligned}\quad (4.13)$$

In other words, we get the classic Lindvall inequality

$$|\mathbb{E}f(X_1(t))-\mathbb{E}f(X_2(t))|\le 2\mathbb{P}(\tau>t).\quad (4.14)$$

Next, assuming that we prove that $\mathbb{E}e^{\varkappa\tau}<\infty$, then

$$\mathbb{P}(\tau>t)\le e^{-\varkappa t}\cdot\mathbb{E}e^{\varkappa\tau}.\quad (4.15)$$

Combining (4.14) with (4.15), we get (4.11). In the proof below, we shall see that the constant before $e^{-\varkappa t}$ turns out to be of the same form as required in (4.11).

Step 3. Let us now describe the coupling in detail.

(a) First, we couple the queue components. Both N_1 and N_2 are stochastically dominated by $\overline{N}$, which is defined as the $M/M/1$ queue with arrival rate $\overline{\lambda}$ and service

rate $\overline{\mu}$, starting from $\overline{N}(0) = n_1 \vee n_2$. Therefore, we can take copies of N_1, N_2, $\overline{N}$ such that

$$N_1(t) \le \overline{N}(t) \text{ and } N_2(t) \le \overline{N}(t), \quad t \ge 0. \tag{4.16}$$

From (4.16), it follows that for $\tau_0 := \inf\{t \ge 0 \mid \overline{N}(t) = 0\}$, we have $N_1(\tau_0) = N_2(\tau_0) = 0$.

(b) At τ_0, we start two competing clocks. The first one is an exponential clock $\eta_0 \sim \text{Exp}(\overline{\lambda})$, which measures the time until arrival of the process $\overline{N}$ to 1 from 0. The second one is ζ_0, a coupling time of $Z_1(\tau_0+\cdot)$ and $Z_2(\tau_0+\cdot)$. This time ζ_0 exists by Assumption 4.1, since these two processes are copies of the environment process with generator $\mathcal{A}$ (recall $\beta_0 = 1$) starting from $Z_1(\tau_0)$ and $Z_2(\tau_0)$, respectively. At least (importantly for us here), this is true until η_0, when those drift and diffusion coefficients change.

(c) If $\zeta_0 < \eta_0$, then Z_1 and Z_2 have time to couple, while $\overline{N}(t) = 0$. By stochastic domination, $N_1(t) = N_2(t) = 0$. Thus, $S_0 := \tau_0 + \zeta_0$ is a coupling time for X_1 and X_2.

(d) If, however, $\zeta_0 \ge \eta_0$, then the coupling did not work. The process $\overline{N}$ has jumped at time $\tau_0 + \eta_0$ back to 1, and we need to repeat this procedure. Let

$$\tau_1 := \inf\{t \ge 0 \mid \overline{N}(t + \tau_0 + \eta_0) = 0\}, \quad \eta_1 \sim \text{Exp}(\overline{\lambda}).$$

Let ζ_1 be a coupling time of $Z_1(\tau_1 + \tau_0 + \eta_0 + \cdot)$ and $Z_2(\tau_1 + \tau_0 + \eta_0 + \cdot)$. If $\zeta_1 < \eta_1$, then for $S_1 := \tau_0 + \eta_0 + \tau_1 + \zeta_1$, we have $\overline{N}(S_1) = 0$, and thus, $N_1(S_1) = N_2(S_1) = 0$. But since ζ_1 is also a coupling time for the environment components, $Z_1(S_1) = Z_2(S_1)$. Thus, S_1 is a coupling time for (N_1, Z_1) and (N_2, Z_2).

(e) If $\zeta_1 \ge \eta_1$, then this coupling did not work, and we need to repeat this procedure, with ζ_2, η_2, S_2, and so on. Let $\mathcal{J} := \min\{j \ge 0 \mid \zeta_j < \eta_j\}$. Then, the ultimate coupling time is

$$\tau := \sum_{j=0}^{\mathcal{J}-1} (\tau_j + \eta_j) + \tau_{\mathcal{J}} + \zeta_{\mathcal{J}} = \sum_{j=0}^{\mathcal{J}} (\tau_j + \eta_j \wedge \zeta_j) = S_{\mathcal{J}}, \tag{4.17}$$

where we define the following random times:

$$S_k := \sum_{j=0}^{k} \xi_j, \quad \xi_k := \tau_k + \zeta_k \wedge \eta_k, \quad k \in \mathbb{Z}_+. \tag{4.18}$$

Next, we estimate the MGF of τ from (4.17).

Step 4. First, we estimate the MGF for each τ_k. The generator of $\overline{N}$ is

$$\overline{\mathcal{M}} f(n) = \overline{\lambda}(f(n+1) - f(n)) + \overline{\mu} 1_{\{n \ne 0\}} (f(n-1) - f(n)).$$

Therefore, letting $f(n) = c^n$ for a constant $c > 1$, we get

$$\overline{\mathcal{M}} f(n) = -m(c) f(n), \quad n \geq 1,$$

with the constant $m(c)$ defined in (4.5). The process

$$L(t) := c^{\overline{N}(t\wedge\tau_0)} + m(c) \int_0^{t\wedge\tau_0} c^{\overline{N}(s)} \, \mathrm{d}s, \quad t \geq 0,$$

is a local supermartingale, because the function $W_N : n \mapsto c^n$ satisfies

$$\overline{\mathcal{M}} W_N(n) \leq -m(c) W_N(n), \quad n = 1, 2, \ldots.$$

In the terminology of [37, Section 4], this is a *modified Lyapunov function* for $\overline{N}$. Then, the derivation is similar to [37, Section 5]. By Fatou's lemma, L is a true supermartingale. Let

$$L_*(t) := \int_0^t e^{m(c)s} \, \mathrm{d}L(s), \quad t \geq 0.$$

Because $e^{ms} \geq 0$, this process is also a supermartingale. Consider the process

$$L^*(t) := e^{m(c)(t\wedge\tau_0)} c^{\overline{N}(t\wedge\tau_0)}, \quad t \geq 0.$$

By an elementary calculation, $\mathrm{d}L^*(t) = \mathrm{d}L_*(t)$. Therefore $L^*(t) = L_*(t) + \text{const}$, and L^* is itself a supermartingale. Thus, for every $t \geq 0$,

$$\mathbb{E}\left[e^{m(c)(t\wedge\tau_0)} c^{\overline{N}(t\wedge\tau_0)} \right] \leq \mathbb{E} c^{\overline{N}(0)}. \tag{4.19}$$

Let $t \to \infty$ in (4.19). By Fatou's lemma with the observation that $\overline{N}(\tau_0) = 0$, we get

$$\mathbb{E} e^{m(c)\tau_0} \leq c^{n_1 \vee n_2}. \tag{4.20}$$

Similarly to (4.20), we get estimates for the MGFs of $\tau_1, \tau_2, \ldots$, with the difference that the initial state becomes 1 instead of $n_1 \vee n_2$. Therefore,

$$\mathbb{E} e^{m(c)\tau_k} \leq c, \quad k = 1, 2, \ldots. \tag{4.21}$$

Step 5. By Assumption 4.1, we have $P(\zeta_k > t) \leq \alpha e^{-\gamma t}$ for $t > 0$, and we recall that $\eta_k \sim \mathrm{Exp}(\bar{\lambda})$. Also, ζ_k and η_k are independent. Thus, by Lemma 6.1, we have, for all $k \in \mathbb{Z}_+$,

$$\mathbb{P}(\zeta_k \leq \eta_k) \leq \frac{\gamma}{\bar{\lambda} + \gamma} \alpha^{-\bar{\lambda}/\gamma} =: p.$$

Thus, the number of "tries," $\mathcal{J}$, is stochastically dominated by a geometric random variable $\widetilde{\mathcal{J}}$, which is the number of trials that one needs to get to the first success if the probability of success of each trial is p. It has the distribution and generating function (with $q := 1 - p$)

$$\mathbb{P}(\widetilde{\mathcal{J}} = n) = pq^{n-1},\ n = 1, 2, \ldots, \quad \text{and} \quad \mathbb{E}\left[s^{\widetilde{\mathcal{J}}}\right] = \frac{ps}{1 - qs},\ s \in [0, q^{-1}). \tag{4.22}$$

Step 6. Let us estimate the MGF of ξ_k, defined in (4.18). By Assumption 4.1 and Lemma 6.1 applied to $a := m(c)$ for $c \in [1, c_*]$,

$$\mathbb{E}\left[e^{m(c)(\zeta_k \wedge \eta_k)}\right] \le \theta(\alpha, \overline{\lambda}, \gamma, m(c)). \tag{4.23}$$

The expression for $\theta(\alpha, \beta, \gamma, a)$ is given in (4.6). Combining (4.21) and (4.23), we get

$$\mathbb{E}\left[e^{m(c)\xi_k}\right] \le c\theta(\alpha, \overline{\lambda}, \gamma, m(c)) =: \kappa(c), \quad k = 1, 2, \ldots$$

The same holds if we take conditional expectation

$$\mathbb{E}\left[e^{m(c)\xi_k} \mid \mathcal{F}_{S_{k-1}}\right] \le c\theta(\alpha, \overline{\lambda}, \gamma, m(c)), \quad k = 1, 2, \ldots \tag{4.24}$$

Combining (4.20) and (4.23), we get

$$\mathbb{E}\left[e^{m(c)\xi_k}\right] \le c^{n_1 \vee n_2}\theta(\alpha, \overline{\lambda}, \gamma, m(c)). \tag{4.25}$$

Step 7. Finally, recall (4.17). We estimate from above the MGF for appropriate $\varkappa > 0$: $\mathbb{E}\left[e^{\varkappa S_{\mathcal{J}}}\right] = \mathbb{E}\left[e^{\varkappa\tau}\right]$. By (4.24), the process $(M_k)_{k\in\mathbb{Z}_+}$ defined by

$$M_k := \exp\left(m(c)S_k - k\ln\kappa(c)\right), \quad k \in \mathbb{Z}_+,$$

is an $(\mathcal{F}_{S_k})_{k\in\mathbb{Z}_+}$-supermartingale. It is positive, and $\mathcal{J}$ is an $(\mathcal{F}_{S_k})_{k\in\mathbb{Z}_+}$-stopping time. Applying the optional stopping theorem and using (4.25), we obtain

$$\mathbb{E}\left[M_{\mathcal{J}}\right] \le \mathbb{E}[M_0] = \mathbb{E}\left[e^{m(c)\xi_0}\right] = c^{n_1 \vee n_2}\theta(\alpha, \overline{\lambda}, \gamma, m(c)). \tag{4.26}$$

By Hölder's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left((1-\varepsilon)m(c)S_{\mathcal{J}}\right)\right] \\
&\quad \le \left(\mathbb{E}\left[e^{m(c)S_{\mathcal{J}} - \mathcal{J}\ln\kappa(c)}\right]\right)^{1-\varepsilon} \cdot \left(\mathbb{E}\left[\exp\left(\mathcal{J}((1-\varepsilon)/\varepsilon)\ln\kappa(c)\right)\right]\right)^{\varepsilon} \\
&\quad = \left(\mathbb{E}\left[M_{\mathcal{J}}\right]\right)^{1-\varepsilon}\mathbb{E}\left[\kappa(c)^{(1-\varepsilon)\mathcal{J}/\varepsilon}\right].
\end{aligned} \tag{4.27}$$

Since $\kappa(c) > 0$ for $c \in [1, c_*]$, and $\mathcal{J}$ is stochastically dominated by a geometric random variable $\widetilde{\mathcal{J}}$ as in (4.22), we have

$$\mathbb{E}\left[\kappa(c)^{(1-\varepsilon)\mathcal{J}/\varepsilon}\right] \le \mathbb{E}\left[\kappa(c)^{(1-\varepsilon)\widetilde{\mathcal{J}}/\varepsilon}\right] = \frac{p\kappa(c)^{(1-\varepsilon)/\varepsilon}}{1-\kappa(c)^{(1-\varepsilon)/\varepsilon}q}. \tag{4.28}$$

Here, we require that $\kappa(c)^{(1-\varepsilon)/\varepsilon} < q^{-1}$, which is exactly the condition for c in (4.8). Combining (4.26), (4.27) and (4.28), we get

$$\begin{aligned}\mathbb{E}\left[\exp\left((1-\varepsilon)m(c)S_{\mathcal{J}}\right)\right] &\le c^{(1-\epsilon)(n_1\vee n_2)}\theta(\alpha,\overline{\lambda},\gamma,m(c))^{1-\epsilon}\frac{p\kappa(c)^{(1-\varepsilon)/\varepsilon}}{1-\kappa(c)^{(1-\varepsilon)/\varepsilon}q}\\ &= C_* c^{(1-\epsilon)[(n_1\vee n_2)-1]} < C_* c^{n_1\vee n_2} \le C_*(c^{n_1}+c^{n_2}),\\ C_* &:= \frac{p\kappa(c)^{(1-\varepsilon)(1/\varepsilon+1)}}{1-\kappa(c)^{(1-\varepsilon)/\varepsilon}q}.\end{aligned} \tag{4.29}$$

From (4.17), this completes the proof of (4.10) for $\varkappa := (1-\varepsilon)m(c)$ and Theorem 4.1.

5 Concluding remarks

We have found the explicit invariant measure for the joint interactive queueing and environment process and estimated the exponential rate of convergence for the compact environment case. One interesting question would be to consider unbounded environment domains, but with environment process being exponentially ergodic. This will require much finer estimates, because Assumption 4.1 will hold only with α dependent on z_1 and z_2. One way to find such coupling was developed in [7,20,31,37] via Lyapunov functions. Subgeometric rates of convergence seem interesting. Some work was done in [15,29] for general Markov processes and in [1,2] for some SDEs arising from many-server queues, but to the best of our knowledge none for our setup.

6 Appendix

6.1 Proof of Lemma 4.3

Alternatively, we can describe such a pure jump process $X = (X(t),\, t \ge 0)$ as follows: Run an exponential clock $\eta_1 \sim \mathrm{Exp}(\Lambda)$, and then let $X(t) = X(0)$ for $t < \eta_1$, and $X(\eta_1) \sim \overline{\nu}(X(0), \cdot)$ (independently of η_1). Run another exponential clock $\eta_2 \sim \mathrm{Exp}(\Lambda)$ independent of those random variables, then $X(S_2), Y(S_2)$ with $S_2 := \eta_1 + \eta_2$, and repeat the process. Thus, we couple these processes $X = \{X(t) : t \ge 0\}$ and $Y = \{Y(t) : t \ge 0\}$ starting from $X(0) = x$ and $Y(0) = y$ as follows: We use the same exponential clocks $\eta_1, \eta_2, \ldots$, and couple $X(S_k)$ and $Y(S_k)$ with $S_k := \eta_1 + \ldots + \eta_k$, using the *maximal coupling* from [30, Proposition 4.7]:

$$\mathbb{P}\left(X(\eta_k) \neq Y(\eta_k) \mid \mathcal{F}_{S_{k-1}}\right) = \|\overline{\nu}(X(\eta_{k-1}, \cdot) - \overline{\nu}(Y(\eta_{k-1}, \cdot))\|_{\mathrm{TV}}, \quad k = 1, 2, \ldots \tag{6.1}$$

The coupling time then becomes

$$\tau_{x,y} := S_{\mathcal{J}}, \quad \mathcal{J} := \min\{k \geq 1 : X(\eta_k) = Y(\eta_k)\}. \tag{6.2}$$

Combining (4.9) and (6.1), we get

$$\mathbb{P}\left(X(\eta_k) = Y(\eta_k) \mid \mathcal{F}_{S_{k-1}}\right) \geq p := 1 - q, \quad k = 1, 2, \ldots. \tag{6.3}$$

Therefore, $\mathcal{J}$ is stochastically dominated by a geometric random variable $\tilde{\mathcal{J}}$ (the number of tries until the first success in a sequence of independent Bernoulli trials with individual success probability p), independent of $\eta_1, \eta_2, \ldots$. From (6.2), we get

$$\tau_{x,y} \preceq \eta_1 + \cdots + \eta_{\tilde{\mathcal{J}}} =: \tilde{S}. \tag{6.4}$$

The MGF of each of these exponential random variables is

$$\mathbb{E}\left[e^{u\eta_k}\right] = \frac{\Lambda}{\Lambda - u}, \quad u < \Lambda,$$

and the generating function for this geometric random variable is

$$\mathbb{E}\left[s^{\tilde{\mathcal{J}}}\right] = \frac{ps}{1 - qs}, \quad s < q^{-1}.$$

Therefore, the MGF for $\tilde{S}$ from the right-hand side of (6.4) is the composition

$$\mathbb{E}\left[e^{u\tilde{S}}\right] = \frac{p\frac{\Lambda}{\Lambda - u}}{1 - q\frac{\Lambda}{\Lambda - u}} = \frac{p\Lambda}{p\Lambda - u}.$$

Thus, $\tilde{S} \sim \mathrm{Exp}(p\Lambda)$, and it satisfies $\mathbb{P}(\tilde{S} \geq t) \leq e^{-p\Lambda t}$. The rest is trivial.

6.2 A technical comparison lemma

Lemma 6.1 *Fix constants $\alpha > 1$, $\beta, \gamma > 0$. Take two independent random variables $\xi \sim \mathrm{Exp}(\beta)$ and $\eta > 0$ which satisfies $\mathbb{P}(\eta > u) \leq \alpha e^{-\gamma u}$ for $u \geq 0$. Then,*

$$\mathbb{P}(\eta < \xi) \geq \alpha^{-\beta/\gamma} \frac{\gamma}{\beta + \gamma}. \tag{6.5}$$

For $a \in [0, \beta + \gamma)$, the moment generating function for $\xi \wedge \eta$ satisfies

$$\mathbb{E}\left[e^{a(\xi \wedge \eta)}\right] \leq \theta(\alpha, \beta, \gamma, a), \tag{6.6}$$

where the function θ is defined in (4.6).

Proof Let us first show (6.5). We have $\alpha e^{-\gamma u} < 1$ for $u > u_0 := \gamma^{-1} \ln(\alpha)$. Then, we can rewrite our tail estimate for η as follows:

$$\mathbb{P}(\eta \geq u) \leq \begin{cases} \alpha e^{-\gamma u}, & u \geq u_0, \\ 1, & u < u_0. \end{cases}$$

Therefore, we have

$$\begin{aligned} \mathbb{P}(\xi \leq \eta) &= \int_0^\infty \beta e^{-\beta u} \mathbb{P}(u \leq \eta)\, \mathrm{d}u \\ &\leq \int_{u_0}^\infty \alpha\beta e^{-\beta u} e^{-\gamma u}\, \mathrm{d}u + \int_0^{u_0} \beta e^{-\beta u}\, \mathrm{d}u \\ &= \frac{\alpha\beta}{\beta+\gamma} e^{-(\beta+\gamma)u_0} + (1 - e^{-\beta u_0}) = 1 - \frac{\gamma}{\beta+\gamma}\alpha^{-\beta/\gamma}. \end{aligned}$$

From here, (6.5) immediately follows. Next, let us show (6.6). For every $u \geq 0$,

$$\begin{aligned} \mathbb{E}\left[e^{a(u\wedge\eta)}\right] &= e^{au}\, \mathbb{P}(\eta > u) + \int_0^u e^{av}\, \mathbb{P}(\eta \in \mathrm{d}v) \\ &\leq e^{au}\, \mathbb{P}(\eta > u) - \int_0^u e^{av}\, \mathrm{d}\mathbb{P}(\eta > v) \\ &= e^{au}\, \mathbb{P}(\eta > u) - e^{av}\, \mathbb{P}(\eta > v)\big|_{v=0}^{v=u} + \int_0^u \mathbb{P}(\eta > v)\, \mathrm{d}e^{av} \\ &\leq 1 + \int_0^u (\alpha e^{-\gamma v} \wedge 1)\, a e^{av}\, \mathrm{d}v. \end{aligned}$$

Calculate the integral on the right-hand side by splitting it into two integrals: from 0 to u_0 (where u_0 is defined above), and from u_0 to u. If $u \in [0, u_0]$, this integral is equal to

$$\int_0^u (\alpha e^{-\gamma v} \wedge 1)\, a e^{av}\, \mathrm{d}v = \int_0^u a e^{av}\, \mathrm{d}v = e^{au} - 1.$$

If $u > u_0$, then this integral is equal to

$$\begin{aligned} \int_0^u (\alpha e^{-\gamma v} \wedge 1)\, a e^{av}\, \mathrm{d}v &= \int_0^{u_0} a e^{av}\, \mathrm{d}v + \int_{u_0}^u \alpha a e^{(a-\gamma)v}\, \mathrm{d}v \\ &= e^{au_0} - 1 + \frac{\alpha a}{a-\gamma}\left[e^{(a-\gamma)u} - e^{(a-\gamma)u_0}\right] \\ &= \alpha^{a/\gamma} + \frac{\alpha a}{a-\gamma}\left[e^{(a-\gamma)u} - \alpha^{(a-\gamma)/\gamma}\right] - 1 \\ &= \frac{\gamma}{\gamma-a}\alpha^{a/\gamma} + \frac{\alpha a}{a-\gamma} e^{(a-\gamma)u} - 1. \end{aligned}$$

Combining all these computations, we get

$$\mathbb{E}\left[e^{a(u\wedge\eta)}\right] \le \begin{cases} \frac{\gamma}{\gamma-a}\alpha^{a/\gamma} + \frac{\alpha a}{a-\gamma}e^{(a-\gamma)u}, & u > u_0, \\ e^{au}, & u \in [0, u_0]. \end{cases} \tag{6.7}$$

Now, integrate (6.7) with respect to the exponential distribution of ξ, $\beta e^{-\beta u}\,\mathrm{d}u$:

$$\begin{aligned}\mathbb{E}\left[e^{a(\xi\wedge\eta)}\right] &\le \int_0^{u_0} e^{au}\,\beta e^{-\beta u}\,\mathrm{d}u + \int_{u_0}^{\infty}\left[\frac{\gamma}{\gamma-a}\alpha^{a/\gamma} + \frac{\alpha a}{a-\gamma}e^{(a-\gamma)u}\right]\beta e^{-\beta u}\,\mathrm{d}u \\ &= \frac{\beta}{a-\beta}\left[e^{(a-\beta)u_0} - 1\right] + \frac{\gamma}{\gamma-a}\alpha^{a/\gamma}e^{-\beta u_0} + \frac{\alpha\beta a}{(a-\gamma)(\gamma+\beta-a)}e^{-(\gamma+\beta-a)u_0} \\ &= \frac{a(\beta-\gamma)}{(a-\beta)(a-\gamma)}\alpha^{(a-\beta)/\gamma} - \frac{\beta}{a-\beta} + \frac{\beta}{(a-\gamma)(\beta+\gamma-a)}\alpha^{1-(\gamma+\beta-a)/\gamma} \\ &= \frac{a\gamma}{(a-\beta)(\beta+\gamma-a)}\alpha^{(a-\beta)/\gamma} + \frac{\beta}{\beta-a}.\end{aligned}$$

This completes the proof. □

Acknowledgements G. Pang was supported in part by NSF grants CMMI-1635410, and DMS/CMMI-1715875 and in part by an Army Research Office Grant W911NF-17-1-0019. Y. Belopolskaya was supported in part by RSF 17-11-01136. Y. Suhov thanks Department of Mathematics at Pennsylvania State University for hospitality and support.

References

1. Arapostathis, A., Pang, G., Sandrić, N.: Ergodicity of Lévy-driven SDEs arising from multiclass many-server queues. Ann. Appl. Probab. **29**(2), 1070–1126 (2019)
2. Arapostathis, A., Hmedi, H., Pang, G., Sandrić, N.: Uniform polynomial rates of convergence for a class of Lévy-driven controlled SDEs arising in multiclass many-server queues. In: Yin, G., Zhang, Q. (eds.) Modeling, Stochastic Control, Optimization, and Applications. The IMA Volumes in Mathematics and its Applications, vol. 164, pp. 1–20. Springer, New York (2019)
3. Ata, B., Peng, X.: An optimal callback policy for general arrival processes: a pathwise analysis. Working paper (2018)
4. Bassamboo, A., Harrison, J.M., Zeevi, A.: Dynamic routing and admission control in high-volume service systems: asymptotic analysis via multi-scale fluid limits. Queueing Syst. **51**(3–4), 249–285 (2005)
5. Bassamboo, A., Harrison, J.M., Zeevi, A.: Design and control of a large call center: asymptotic analysis of an LP-based method. Oper. Res. **54**(3), 419–435 (2006)
6. Belopolskaya, Y., Suhov, Y.: Models of Markov processes with a random transition mechanism (2015). arXiv:1508.05598
7. Blanchet, J., Chen, X.: Rates of convergence to stationarity for multidimensional RBM (2016). arXiv:1601.04111
8. Borodin, A., Salminen, P.: Handbook of Brownian Motion. Facts and Formulae, 2nd edn. Birkhauser, Basel (2002)
9. Burdzy, K., Kendall, W.S.: Efficient Markovian couplings: examples and counterexamples. Ann. Appl. Probab. **10**(2), 362–409 (2000)
10. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks. Stochastic Modeling and Applied Probability, vol. 46. Springer, Berlin (2001)
11. Chen, M.-F., Li, S.-F.: Coupling methods for multidimensional diffusion processes. Ann. Probab. **17**(1), 151–177 (1989)

12. Cogburn, R.: Markov chains in random environments: the case of Markovian environments. Ann. Appl. Probab. **8**(5), 908–916 (1980)
13. Cogburn, R., Torrez, W.C.: Birth and death processes with random environments in continuous time. J. Appl. Probab. **18**(1), 19–30 (1981)
14. Cornez, R.: Birth and death processes in random environments with feedback. J. Appl. Probab. **24**(1), 25–34 (1987)
15. Douc, R., Fort, G., Guillin, A.: Subgeometric rates of convergence of f-ergodic strong Markov processes. Stoch. Proc. Appl. **119**(3), 897–923 (2009)
16. Economou, A.: Generalized product-form stationary distributions for Markov chains in random environments with queueing applications. Adv. Appl. Probab. **37**(1), 185–211 (2005)
17. Falin, G.: A heterogeneous blocking system in a random environment. J. Appl. Probab. **33**(1), 211–216 (1996)
18. Gannon, M., Pechersky, E., Suhov, Y., Yambartsev, A.: A random walk in a queueing network environment. J. Appl. Probab. **53**(2), 448–462 (2016)
19. Hunter, J.J.: Coupling and mixing times in a Markov chain. Lin. Algebra Appl. **430**(10), 2607–2621 (2009)
20. Ichiba, T., Sarantsev, A.: Stationary distributions and convergence for Walsh diffusions. Bernoulli **25**(4A), 2439–2478 (2019)
21. Karatzas, I., Shreve, S.E.: Brownian Motion and Stochastic Calculus. Graduate Texts in Mathematics, vol. 113. Springer, Berlin (1991)
22. Kelly, F.: Reversibility and Stochastic Networks. Wiley, Chichester (1979)
23. Kelly, F., Yudovina, E.: Stochastic Networks. Cambridge University Press, Cambridge (2014)
24. Krenzler, R., Daduna, H., Otton, S.: Jackson networks in nonautonomous random environments. Adv. Appl. Probab. **48**(2), 315–331 (2016)
25. Krishnamoorthy, A., Pramod, P.K., Chakravarthy, S.R.: Queues with interruptions: a survey. TOP **22**(1), 290–320 (2014)
26. Kumar, R., Lewis, M.E., Topaloglu, H.: Dynamic service rate control for a single-server queue with Markov-modulated arrivals. Naval Res. Logist. **60**(8), 661–677 (2013)
27. Kurtz, T.G., Stockbridge, R.H.: Stationary solutions and forward equations for controlled and singular martingale problems. Electron. J. Probab. **6**(17), 1–52 (2001)
28. Lions, P.-L., Sznitman, A.-S.: Stochastic differential equations with reflecting boundary conditions. Commun. Pure Appl. Math. **37**(4), 511–537 (1984)
29. Liu, Y., Zhang, H., Zhao, Y.: Subgeometric ergodicity for continuous-time Markov chains. J. Math. Anal. Appl. **368**(1), 178–189 (2010)
30. Levin, D.A., Perez, Y., Wilmer, E.L.: Markov Chains and Mixing Times, 2nd edn. American Mathematical Society, Providence (2017)
31. Lund, R., Meyn, S.P., Tweedie, R.L.: Computable exponential convergence rates for stochastically ordered Markov processes. Ann. Appl. Probab. **6**(1), 218–237 (1996)
32. Meyn, S.P., Tweedie, R.L.: Stability of Markovian processes II. Continuous-time processes and sampled chains. Adv. Appl. Probab. **25**(3), 487–517 (1993)
33. Neuts, R.F.: Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Dover, Mineola (1995)
34. Norris, J.R.: Markov chains. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (1997)
35. Regterschot, G.J.K., de Smit, J.H.A.: The queue $M/G/1$ with Markov modulated arrivals and services. Math. Oper. Res. **11**(3), 465–483 (1986)
36. Robert, P.: Stochastic Networks and Queues. Stochastic Modeling and Applied Probability, vol. 52. Springer, Berlin (2003)
37. Sarantsev, A.: Explicit rates of exponential convergence for reflected jump-diffusions on the positive half-line. ALEA Lat. Am. J. Probab. Math. Stat. **13**(2), 1069–1093 (2016)
38. Sarantsev, A.: Reflected Brownian motion in a convex polyhedral cone: tail estimates for the stationary distribution. J. Theor. Probab. **32**(3), 545–585 (2019)
39. Sawyer, S.A.: A formula for semigroups, with an application to branching diffusion processes. Trans. Am. Math. Soc. **152**(1), 1–38 (1970)
40. Suhov, Y., Kelbert, M.: Probability and statistics by example. Markov chains: a primer in random processes and their applications. Cambridge University Press, Cambridge (2008)

41. Takine, T.: Single-server queues with Markov-modulated arrival and service speed. Queueing Syst. **49**(1), 7–22 (2005)
42. Williams, R.J.: Reflected Brownian motion with skew symmetric data in a polyhedral domain. Probab. Theor. Relat. Fields **75**, 459–485 (1987)
43. Williams, R.J.: Semimartingale reflecting Brownian motions in the orthant: a survey. IMA Vol. Math. Appl. **71**, 125–137 (1995)
44. Yechiali, U.: A queueing-type birth-and-death process defined in a continuous-time Markov chain. Oper. Res. **21**(2), 604–609 (1973)