A heavy traffic limit theorem for a class of open queueing networks with finite buffers [*]

J.G. Dai [a] and W. Dai [b,**]

[a] *School of Industrial and Systems Engineering, and School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA*
E-mail: dai@isye.gatech.edu
[b] *School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA*

Submitted 1 February 1998; accepted 1 December 1998

We consider a queueing network of d single server stations. Each station has a finite capacity waiting buffer, and all customers served at a station are homogeneous in terms of service requirements and routing. The routing is assumed to be deterministic and hence feedforward. A server stops working when the downstream buffer is full. We show that a properly normalized d-dimensional queue length process converges in distribution to a d-dimensional semimartingale reflecting Brownian motion (RBM) in a d-dimensional box under a heavy traffic condition. The conventional continuous mapping approach does not apply here because the solution to our Skorohod problem may not be unique. Our proof relies heavily on a uniform oscillation result for solutions to a family of Skorohod problems. The oscillation result is proved in a general form that may be of independent interest. It has the potential to be used as an important ingredient in establishing heavy traffic limit theorems for general finite buffer networks.

Keywords: finite capacity network, blocking probabilities, loss network, semimartingale reflecting Brownian motion, RBM, heavy traffic, limit theorems, oscillation estimates

1. Introduction

This paper proves a heavy traffic limit theorem for an open queueing network with finite buffers. The queueing network has d single server stations. Each station has a finite capacity waiting buffer, and all customers served at a station are *homogeneous* in terms of service requirements and routing. The routing is assumed to be deterministic and hence feedforward. Since there is a single customer class associated with each station, our network is a single class queueing network as opposed to the multiclass queueing networks widely discussed in the literature in recent years (see, e.g., Harrison [24]).

[*] Research supported in part by NSF Grants DMI-9457336 and DMI-9813345, US–Israel Binational Science Foundation Grant 94-00196 and a grant from Harris Semiconductor.

[**] Current address: End-to-End Network Architecture Department, Lucent Technologies, Warren, NJ 07059, USA. E-mail: wdai@lucent.com.

Queueing networks have been used to model telecommunication networks and manufacturing systems. All these networks have finite buffers in practice. See, e.g., Bertsekas and Gallager [3], Buzacott [9], Mitra and Mitrani [31], Perros and Altiok [32], and Yao [40]. In some applications, notably in some manufacturing systems like existing wafer fabrication facilities, buffer constraints have not been a major problem. Therefore, it is safe to ignore buffer constraints in the analysis of these networks. However, in telecommunication networks, more recently in asynchronous transfer mode (ATM) networks, buffer constraints have a major impact on system performances (see, e.g., Elwalid and Mitra [20] or Kroner et al. [29]). Thus, it is imperative to model the finiteness of the buffer sizes in these networks.

In our network the interarrival times and service times at each station are assumed to be independent, identically distributed (iid) sequences with finite first two moments. We show that the normalized d-dimensional queue length process converges in distribution to a d-dimensional reflecting Brownian motion (RBM) under a heavy traffic condition. The RBM lives in a d-dimensional box. The Brownian data, including the drift vector, covariance matrix and reflection matrix, can be calculated explicitly from the moments, network topology and the blocking mechanism employed. There are algorithms to numerically compute the stationary distribution of the RBM. Therefore, one can obtain performance estimates for the queueing network, like blocking probabilities and average queue lengths, from their Brownian counterparts [15].

The normalization of the queue length involves a scaling in time by a factor n and a scaling in space by a factor $1/\sqrt{n}$ for large n. Thus the heavy traffic limit theorem provides qualitative insight for the queueing network when it is operated for a long period of time, and each individual customer's movement is not of primary concern. The heavy traffic condition assumes that the traffic intensity ρ_i at each station i is close to 1 so that $1 - \rho_i$ is of order $1/\sqrt{n}$. In addition, it requires the buffer size at a station is of order $\sqrt{n}$. The limit theorem suggests that this is the magnitude of the buffer size for the network to experience a "moderate level" of blocking.

Although many blocking mechanisms can be employed for a finite buffer network, we will focus on the "block-and-hold-0" mechanism. Under such a blocking mechanism, a server will stop working whenever an immediate downstream buffer is full. Therefore, the number of blocked customers that have completed services is 0. Readers are referred to Cheng and Yao [13] or Cheng [12] for the definition of the general "block-and-hold-k" mechanism. We note that the terms "manufacturing blocking" and "communication blocking" may not have a standard meaning in the literature; see, e.g., Cheng [12], and Konstantopoulos and Walrand [28]. A loss mechanism will be briefly discussed in section 9.

Due to the finiteness of the buffer sizes and the blocking mechanism used, the Skorohod problem associated with the queueing network may not have a unique solution (see the example at the end of section 5). Therefore the conventional continuous mapping approach, as used in Iglehart and Whitt [25,26] for feedforward single class networks, in Reiman [34] for single class networks with feedback and in Peterson [33] for feedforward multiclass queueing networks, does not apply here, although some authors, such as Bardhan and Mithal [1], attempted such an extension. Instead we establish a uniform oscillation result for solutions to a sequence of Skorohod problems. Using this result, one can establish that the sequence of normalized queue length processes is precompact in the space of right continuous paths with left limits. Each limit point of the sequence is shown to be an RBM. Care has been taken to show that the limit satisfies a martingale property which is a defining property of the RBM. (Lemma 7.1 of this paper plays a key role in proving this martingale property. The proof of this lemma is adapted from Williams [38].) Finally, the heavy traffic limit theorem follows from the uniqueness (in distribution) of the RBM [18].

Almost all prior proofs of heavy traffic limit theorems for open networks assume the buffer sizes are infinite. For multiclass queueing networks, the mapping associated with the Skorohod problem is not well defined in general, as illustrated by an example of Dai et al. [17] which is included as appendix A of Williams [38]. The nonuniqueness excludes the usage of the continuous mapping theorem used in Iglehart and Whitt [25, 26], Reiman [34], Johnson [27], Peterson [33], and Chen and Zhang [10] to prove heavy traffic limit theorems. Reiman [35] proved a heavy traffic limit theorem for a multiclass station; see Dai and Kurtz [16] for an alternative proof and extension. Chen and Zhang [11] showed a heavy traffic limit theorem for a multiclass FIFO network with a restrictive spectral radius condition on a certain matrix. Although these three works went beyond the conventional continuous mapping paradigm, until very recently, we have not seen a viable approach to the proof of general heavy traffic limit theorems. The contemporaneous, independent works of Bramson [7] and Williams [38,39] provided sufficient conditions for a heavy traffic limit theorem for multiclass queueing networks under many conventional queueing disciplines, including the FIFO discipline, static buffer priority discipline, and head-of-the-line proportional processor sharing (HLPPS) discipline. These results represent a major breakthrough for proving heavy traffic limit theorems for infinite buffer multiclass queueing networks. In fact, using the sufficient conditions and Bramson [5,6], they established new heavy traffic limit theorems for FIFO networks of Kelly type and open multiclass queueing networks under the HLPPS discipline. The two key ingredients in establishing their heavy traffic limit theorems are oscillation result [38] and “state space collapse” [7].

Although the oscillation result in this paper looks similar to the oscillation result in [38], neither one implies the other. Our oscillation result deals with the Skorohod problem in a general state space and requires some control on the jump sizes of the pushing process, whereas Williams’ result deals with a more general family of perturbed Skorohod problems in an orthant. Our oscillation result, which is proved in a much more general setting than needed in this paper, has the potential to be used as an important ingredient to prove a heavy traffic limit theorem for a general *finite* buffer queueing network, although other important ingredients, like deadlock in feedback networks and “state space collapse” in multiclass networks, have to be dealt with separately.

We now introduce the notation to be used in the paper. The number of stations in the network is assumed to be $d \geqslant 1$. Let $\boldsymbol{I} = \{1, \ldots, d\}$. The set of nonnegative integers is denoted by $\mathbb{Z}_+$, and the k-dimensional nonnegative lattice is denoted by $\mathbb{Z}_+^k$. We use $\mathbb{R}^k$ to denote the k-dimensional Euclidean space. Let $\mathbb{R}_+ = [0, \infty)$. Unless stated otherwise, all vectors are envisioned as column vectors. The prime symbol on a vector or a matrix denotes transpose. For $a = (a_1, \ldots, a_k)' \in \mathbb{R}^k$, $|a| = \max_{i=1}^k |a_i|$. For an $n \times k$ matrix A, $||A|| = \max_{i=1}^n \sum_{j=1}^k |A_{ij}|$. For a vector $a \in \mathbb{R}^k$, we use diag(a) to denote the $k \times k$ diagonal matrix whose diagonal entries are given by the components of a. Vector inequalities are interpreted componentwise. We use e to denote the d-dimensional vector of ones.

For $k \geqslant 1$, the k-dimensional path space $D([0, \infty), \mathbb{R}^k)$ is the set of functions $x : [0, \infty) \to \mathbb{R}^k$ that are right continuous on $[0, \infty)$ and have finite left limits on $(0, \infty)$. For a path $x \in D([0, \infty), \mathbb{R}^k)$, we sometimes use $x(\cdot)$ to denote the path. For a vector $a \in \mathbb{R}^k$ and a path $x \in D([0, \infty), \mathbb{R}^k)$, $x(a\cdot)$ is the path with $x(at) = (x_1(a_1 t), \ldots, x_k(a_k t))'$. More generally, for an $h \in D([0, \infty), \mathbb{R}_+^k)$, $x(h(\cdot))$ is the path with $x(h(t)) = (x_1(h_1(t)), \ldots, x_d(h_d(t)))'$. A path $x \in D([0, \infty), \mathbb{R}^k)$ is nondecreasing if each component is. We use $x(s-)$ to denote the left limit at $s > 0$. The space $D([0, \infty), \mathbb{R}^k)$ is endowed with the Skorohod J_1-topology (see, e.g., Ethier and Kurtz [21]). For a sequence of paths $\{f^n\}$, for each $n \geqslant 1$ the paths $\bar{f}^n$ and $\tilde{f}^n$ are defined by

$$\bar{f}^n(\cdot) = \frac{1}{n} f^n(n\cdot) \quad \text{and} \quad \tilde{f}^n(\cdot) = \frac{1}{\sqrt{n}} f^n(n\cdot).$$

The sequence $\{f^n\}$ is said to converge to f uniformly on compact sets if for each $T > 0$

$$\sup_{0 \leqslant t \leqslant T} |f^n(t) - f(t)| \to 0$$

as $n \to \infty$. We denote such converge by $f^n \to f$ u.o.c.

In section 2, the queueing network model is introduced. The heavy traffic limit theorem is stated in section 3. The Skorohod problem is stated in section 4, where a general oscillation result is established. In section 5, we represent the queue length process as a solution to a Skorohod problem. In section 6 we prove a fluid limit theorem which will be used in the proof of the heavy traffic limit theorem. In section 7 we prove a stopping time property that is needed to prove a martingale property. The proof of the heavy traffic limit theorem is completed in section 8. Extensions will be discussed in section 9.

2. The queueing network model

The queueing network under consideration has d single server stations indexed by $i \in \boldsymbol{I} \equiv \{1, \ldots, d\}$. Customers visiting station i are *homogeneous* in terms of service time distribution and routing. We assume that routing is deterministic. That is, customers leaving station i all go next to station $\sigma(i) \in \boldsymbol{I}$ or leave the system. In the latter case we let $\sigma(i) = 0$. Because all customers leaving station i are deterministically

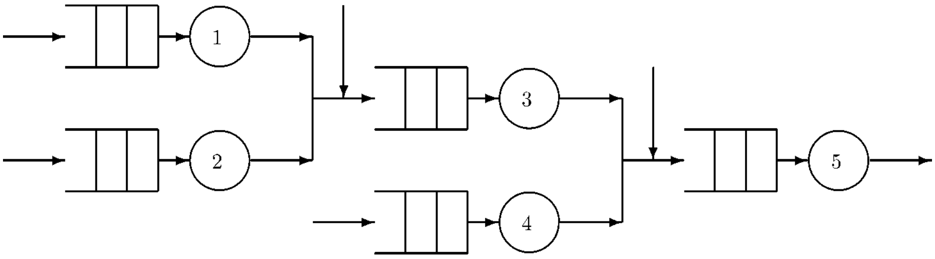

Figure 1. A five station network.

routed to a station, the routing must be feedforward. The network is sometimes called an *in-tree* network. This routing assumption is quite restrictive by conventional standards. An example of such a network is pictured in figure 1. (Other routing assumptions will be discussed in section 9.) We assume that the size b_i of the buffer associated with each station i is *finite*, $i \in \boldsymbol{I}$. Therefore, at each station i there are at most b_i customers, including the one possibly being served. We assume that the network is open. That is, all customers eventually leave the network.

Associated with each station i, there are two sequences of iid positive random variables $\{u_{ik},\ k \geqslant 1\}$ and $\{v_{ik},\ k \geqslant 1\}$, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that

$$\begin{aligned} \mathbb{E}(u_{i1}) &= 1, \quad \mathrm{Var}(u_{i1}) = c_i^a < \infty, \quad i \in \boldsymbol{I}, \\ \mathbb{E}(v_{i1}) &= 1, \quad \mathrm{Var}(v_{i1}) = c_i^s < \infty, \quad i \in \boldsymbol{I}. \end{aligned}$$

Also associated with each station i, there are two numbers: $\alpha_i \geqslant 0$ and $m_i > 0$. The iid random variables $\{v_{ik},\ k \geqslant 1\}$ are the *normalized* service times and the iid random variables $\{u_{ik}, k \geqslant 1\}$ are the *normalized* interarrival times. The actual service times for the kth customer at station i is $m_i v_{ik}$. If $\alpha_i = 0$, there are no external customer arrivals to station i. If $\alpha_i > 0$, the interarrival between the kth and the $(k-1)$th customer is u_{ik}/α_i. Although it is not necessary, for notational convenience, we assume that $\alpha_i > 0$ for each $i \in \boldsymbol{I}$.

An important feature in the network is that the sizes of buffers are *finite*. When the buffer at a downstream station $\sigma(i)$ is full, server i stops working although a customer may still occupy station i. This phenomenon is called the “block-and-hold-0” blocking; see Cheng and Yao [13] for a discussion of general blocking mechanisms. One can envision that when the kth customer enters service at station i, a service time clock (stopwatch) is set to $m_i v_{ik}$. The service is completed when the clock reading reaches zero. During the service period, the clock is turned off or on depending on whether the server is blocked or not. Our blocking mechanism applies to arrivals too. Upon the kth external arrival to station i, an arrival clock at station i is set to $u_{i,k+1}/\alpha_i$. When the clock reading reaches zero, the $k+1$ customer arrives at station i. During this interarrival period, the arrival clock is turned off or on depending on whether buffer i is full or not.

We admit that our blocking mechanism for external arrivals is restrictive for some applications. However, in many manufacturing applications, external arrivals can be controlled. Our blocking mechanism represents one way of modeling arrival processes. In section 9, we will discuss other blocking mechanisms, including loss networks. In heavy traffic analysis, the blocking in our network introduces complications that do not exist in networks with infinite buffers.

For $i \in \boldsymbol{I}$, let $Z_i(t)$ be the number of customers at station i at time t, including possibly the one being served. Note that $Z_i(0)$ is the initial number of customers at station i at time 0. It represents part of an initial network configuration. Let $Y_i(t)$ be the amount of time that server i has been idle while server i is not blocked in time interval $[0,t]$, and let $Y_{i+d}(t)$ be the amount of time that buffer i has been full in time interval $[0,t]$. That is,

$$Y_i(t) = \int_0^t 1_{\{Z_i(s)=0, Z_{\sigma(i)}(s)<b_{\sigma(i)}\}}\,\mathrm{d}s, \qquad Y_{i+d}(t) = \int_0^t 1_{\{Z_i(s)=b_i\}}\,\mathrm{d}s. \tag{2.1}$$

Hereafter, whenever $\sigma(i)=0$ condition $\{a_{\sigma(i)} < b_{\sigma(i)}\}$ always holds for any $a, b \in \mathbb{R}^d$. Let $Z(t) = (Z_1(t), \ldots, Z_d(t))'$ and $Y(t) = (Y_1(t), \ldots, Y_{2d})'$. The process $Z = \{Z(t),\ t \geqslant 0\}$ is called the queue length process and the process $Y = \{Y(t),\ t \geqslant 0\}$ is called the allocation process. Clearly, Y is a nondecreasing, continuous process. Given the iid interarrival time sequences and service time sequences, one can uniquely construct the queue length process and the allocation process. Such detailed construction, though not attempted here, is implicitly assumed in section 7.

For each $i \in \boldsymbol{I}$ and $t \geqslant 0$, let

$$F_i(t) = t - Y_{i+d}(t), \qquad B_i(t) = t - Y_i(t) - Y_{\sigma(i)+d}(t). \tag{2.2}$$

Hereafter, whenever $\sigma(i)=0$, $Y_{\sigma(i)+d}(t)$ is understood to be 0.

It is clear that $B_i(t)$ is the cumulative amount of time that server i has been busy in $[0,t]$ and $F_i(t)$ is the cumulative amount of time that buffer i has not been full in $[0,t]$. That is,

$$F_i(t) = \int_0^t 1_{\{Z_i(s)<b_i\}}\,\mathrm{d}s, \qquad B_i(t) = \int_0^t 1_{\{Z_i(s)>0, Z_{\sigma(i)}(s)<b_{\sigma(i)}\}}\,\mathrm{d}s.$$

3. A heavy traffic limit theorem

To state a heavy traffic limit theorem, we need to consider a sequence of networks indexed by n. The network depends on the index n through the external arrival rates α^n, mean service times m^n and buffer sizes b^n, where

$$\alpha^n = \left(\alpha_1^n, \ldots, \alpha_d^n\right)', \qquad m^n = \left(m_1^n, \ldots, m_d^n\right)', \qquad b^n = \left(b_1^n, \ldots, b_d^n\right)'.$$

We let $\mu_i^n = 1/m_i^n$ be the mean service rate at station i. The normalized interarrival and service times, and the routing do not depend on n. Let $Z^n = \{Z^n(t),\ t \geqslant 0\}$ be the queue length process and $Y^n = \{Y^n(t),\ t \geqslant 0\}$ be the allocation process associated with the nth network. In the following theorem, P is the $d \times d$ routing matrix, i.e., $P_{ij} = 1$ if station i customers go next to station j and $P_{ij} = 0$, otherwise.

Theorem 3.1. Assume that as $n \to \infty$,

$$\alpha^n \to \alpha > 0 \quad \text{and} \quad m^n \to m > 0, \tag{3.1}$$

$$\frac{b^n}{\sqrt{n}} \to b > 0, \tag{3.2}$$

$$\sqrt{n}\left(\alpha^n - \left(I - P'\right)\mu^n\right) \to \theta. \tag{3.3}$$

Assume that for each n, $Z^n(0)$ is defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $Z^n(0)$ is independent of the interarrival and service time sequences such that

$$\frac{1}{\sqrt{n}} Z^n(0) \Longrightarrow \xi, \quad n \to \infty. \tag{3.4}$$

Assume further that

$$\Gamma = \text{diag}\left(\alpha_1 c_1^a, \ldots, \alpha_d c_d^a\right) + \left(I - P'\right) \text{diag}\left(\mu_1 c_1^s, \ldots, \mu_d c_d^s\right)(I - P) \tag{3.5}$$

is (strictly) positive definite. Then

$$\left(\frac{1}{\sqrt{n}} Z^n(n\cdot), \frac{1}{\sqrt{n}} Y^n(n\cdot)\right) \Longrightarrow \left(Z^*(\cdot), Y^*(\cdot)\right), \quad \text{as } n \to \infty, \tag{3.6}$$

where Z^*, together with Y^*, is a semimartingale reflecting Brownian motion (RBM) defined on a filtered probability space $(\Omega^*, \{\mathcal{F}_t^*\}, \mathcal{F}^*, \mathbb{P}^*)$. The process (Z^*, Y^*) is uniquely determined in distribution from the following equations:

$$\mathbb{P}^*\text{-a.s.,} \quad Z^*(t) = Z^*(0) + X^*(t) + RY^*(t) \quad \text{for all } t \geqslant 0, \tag{3.7}$$

$$\mathbb{P}^*\text{-a.s.,} \quad 0 \leqslant Z^*(t) \leqslant b \quad \text{for all } t \geqslant 0, \tag{3.8}$$

$$Z^*(0) \text{ has the same distribution as } \xi, \tag{3.9}$$

$$Z^*(\cdot) \text{ and } Y^*(\cdot) \text{ are} \{\mathcal{F}_t^*\}\text{-adapted}, \tag{3.10}$$

$$\mathbb{P}^*\text{-a.s.,} \quad Y^*(0) = 0,\ Y^*(\cdot) \text{ is continuous and nondecreasing}, \tag{3.11}$$

$$\mathbb{P}^*\text{-a.s.,} \quad \text{for } i \in \boldsymbol{I},\ Y_i^*(\cdot) \text{ increases only at times } t \text{ when } Z_i^*(t) = 0, \tag{3.12}$$

$$\mathbb{P}^*\text{-a.s.,} \quad \text{for } i \in \boldsymbol{I},\ Y_{i+d}^*(\cdot) \text{ increases only at times } t \text{ when } Z_i^*(t) = b_i, \tag{3.13}$$

$$X^* \text{ is a Brownian motion with drift } \theta \text{ and covariance matrix } \Gamma, \tag{3.14}$$

$$\left\{X^*(t) - \theta t\right\} \text{ is an } \left\{\mathcal{F}_t^*\right\}\text{-martingale}, \tag{3.15}$$

where θ is defined in (3.3), Γ is defined in (3.5) and

$$R = \left(\left(I - P'\right) \text{diag}(\mu),\ \left[\left(I - P'\right) \text{diag}(\mu)\right]_\sigma - \text{diag}(\alpha)\right). \tag{3.16}$$

For a $d \times d$ matrix A and a vector $x \in \mathbb{R}^d$,

$$x_\sigma = (x_{\sigma(1)}, \ldots, x_{\sigma(d)})' \quad \text{and} \quad A_\sigma x = A x_\sigma. \tag{3.17}$$

The theorem will be proved in section 8. The vector θ, the matrix Γ and the $d \times 2d$ matrix R are called the drift, the covariance matrix and the reflection matrix of the RBM Z^*, respectively. For $i \in \boldsymbol{I}$, the ith column of R is the direction of reflection used when $Z_i^*(t) = 0$, and the $(i+d)$th column of R is the direction of reflection used when $Z_i^*(t) = b_i$. Because of (3.8), the RBM Z^* lives in the d-dimensional box $\boldsymbol{S}$ defined by

$$\boldsymbol{S} \equiv \bigl\{x = (x_1, \ldots, x_d)' \in \mathbb{R}^d\colon\ 0 \leqslant x_i \leqslant b_i \text{ for } i \in \boldsymbol{I}\bigr\}. \tag{3.18}$$

Therefore, the RBM Z^* in the theorem has state space $\boldsymbol{S}$. From now on, we call the RBM Z^* a $(\Gamma, \theta, R, \boldsymbol{S})$-RBM. The process Y^* is the pushing processes associated with the RBM Z^*. In the stochastic differential equation terminology, the process (Z^*, Y^*) is a *weak* solution to (3.7)–(3.14). Because the corresponding Skorohod problem may not have a unique solution (see the example at the end of section 5), it is not known whether a (strong) solution exists for *each* Brownian motion X^* defined on a probability space. The uniqueness of (Z^*, Y^*) (in distribution) follows from Dai and Williams [18] that generalized an earlier result of Taylor and Williams [37] for RBM's in an orthant.

4. The Skorohod problem and an oscillation theorem

In this section, we define the Skorohod problem and establish an oscillation result for solutions to a family of Skorohod problems. We choose to prove our results in a general polyhedral state space $\boldsymbol{S}$, instead of the d-dimensional box introduced in (3.18). We believe our oscillation result in a general state space is of independent interest.

In this section we follow most of the notation introduced in section 1 of Dai and Williams [18]. Symbols m and F are reused in this section. In the subsequent sections, they retain the original meaning. The polyhedron is defined in terms of m $(m \geqslant 1)$ d-dimensional unit vectors $\{n_i,\ i \in \boldsymbol{J}\}$, $\boldsymbol{J} \equiv \{1, \ldots, m\}$, and an m-dimensional vector $a = (a_1, \ldots, a_m)'$. The state space $\boldsymbol{S}$ is defined by

$$\boldsymbol{S} \equiv \bigl\{x \in \mathbb{R}^d\colon\ n_i \cdot x \geqslant a_i \text{ for all } i \in \boldsymbol{J}\bigr\}, \tag{4.1}$$

where $n_i \cdot x = n_i' x$ denotes the inner product of the vectors n_i and x. It is assumed that the interior of $\boldsymbol{S}$ is non-empty and that the set $\{(n_1, a_1), \ldots, (n_m, a_m)\}$ is minimal in the sense that no proper subset defines $\boldsymbol{S}$. That is, for any strict subset $\boldsymbol{K} \subset \boldsymbol{J}$, the set $\{x \in \mathbb{R}^d\colon\ n_i \cdot x \geqslant a_i\ \forall i \in \boldsymbol{K}\}$ is strictly larger than $\boldsymbol{S}$. This is equivalent to the assumption that each of the faces

$$F_i \equiv \{x \in \boldsymbol{S}\colon\ n_i \cdot x = a_i\}, \quad i \in \boldsymbol{J}, \tag{4.2}$$

has dimension $d - 1$ (cf. [8, theorem 8.2]). As a consequence, n_i is the unit normal to F_i that points into the interior of $\boldsymbol{S}$. Let N denote the $m \times d$ matrix whose ith row is given by the row vector n_i' for each $i \in \boldsymbol{J}$.

For each face $F_i,\ i \in \boldsymbol{J}$, we associate a d-dimensional vector v_i with it. We use R to denote the $d \times m$ matrix whose ith column v_i. Let us first define the Skorohod problem associated with the data $(\boldsymbol{S}, R)$. The matrix R is called the reflection matrix.

In the following, for a Borel set $U \subset \mathbb{R}^k,\ k \geqslant 1$, we define $D([0,T], U) = \{w : [0,T] \to U,\ w$ is right continuous in $[0,T)$ having left limits in $(0,T]\}$.

Definition 4.1 (The Skorohod problem). Given $T > 0$ and $x \in D([0,T], \mathbb{R}^d)$ with $x(0) \in \boldsymbol{S}$, an $(\boldsymbol{S}, R)$-regulation of x over $[0,T]$ is a pair $(z, y) \in D([0,T],\ \boldsymbol{S}) \times D([0,T],\ \mathbb{R}^m_+)$ such that

(i) $z(t) = x(t) + Ry(t)$ for all $t \in [0,T]$,

(ii) $z(t) \in \boldsymbol{S}$ for all $t \in [0,T]$,

(iii) for each $i \in \boldsymbol{J}$,

(a) $y_i(0) = 0$,

(b) y_i is nondecreasing,

(c) $\int_{(0,T]} (n_i \cdot z(t) - a_i)\, \mathrm{d}y_i(t) = 0$.

Remarks. (a) Although in the rest of this paper, y is known to be continuous, we allow y to have jumps in the definition of the Skorohod problem.

(b) The integral $\int_{(0,T]} (n_i \cdot z(t) - a_i)\, \mathrm{d}y_i(t)$ is well defined as a Lebesgue–Stieltjes integral, because any path $z \in D([0,T], \mathbb{R}^d)$ is bounded in $[0,T]$. Loosely speaking, condition (iii)(c) says that y_i can increase only at times $t \in [0,T]$ for which $z(t) \in F_i$. (See lemma 4.4 for a more precise statement.)

The existence and uniqueness of an $(\boldsymbol{S}, R)$-regulation heavily depends on the reflection matrix R.

Definition 4.2. For each $\emptyset \neq \boldsymbol{K} \subset \boldsymbol{J}$, define $F_{\boldsymbol{K}} = \bigcap_{i \in \boldsymbol{K}} F_i$. Let $F_\emptyset = \boldsymbol{S}$. A set $\boldsymbol{K} \subset \boldsymbol{J}$ is *maximal* if $\boldsymbol{K} \neq \emptyset,\ F_{\boldsymbol{K}} \neq \emptyset$, and $F_{\boldsymbol{K}} \neq F_{\widetilde{\boldsymbol{K}}}$ for any $\widetilde{\boldsymbol{K}} \supset \boldsymbol{K}$ such that $\widetilde{\boldsymbol{K}} \neq \boldsymbol{K}$.

Now we introduce an assumption on N and R.

Completely-$\mathcal{S}$ assumption. For each maximal $\boldsymbol{K} \subset \boldsymbol{J}$,

(S.a) there is a positive linear combination $v = \sum_{i \in \boldsymbol{K}} c_i v_i$ ($c_i > 0\ \forall i \in \boldsymbol{K}$) of the $\{v_i,\ i \in \boldsymbol{K}\}$ such that $n_i \cdot v > 0$ for all $i \in \boldsymbol{K}$;

(S.b) there is a positive linear combination $\eta = \sum_{i \in \boldsymbol{K}} c_i n_i$ ($c_i > 0\ \forall i \in \boldsymbol{K}$) of the $\{n_i,\ i \in \boldsymbol{K}\}$ such that $\eta \cdot v_i > 0$ for all $i \in \boldsymbol{K}$.

The labels (S.a) and (S.b) stand for $\mathcal{S}$-condition (a) and (b), respectively. The origin of these labels becomes apparent when the conditions are written in matrix form as below. For a vector $x \in \mathbb{R}^k$, the notation $x > 0$ indicates that all coordinates of x are strictly positive, and the notation $x \geqslant 0$ indicates that all coordinates of x are nonnegative.

Definition 4.3. A matrix A is called an $\mathcal{S}$-matrix if there is a vector $x \geqslant 0$ such that $Ax > 0$.

For an $m \times m$ matrix A and $\boldsymbol{K} \subset \boldsymbol{J}$, let $A_{\boldsymbol{K}}$ denote the $|\boldsymbol{K}| \times |\boldsymbol{K}|$ matrix obtained from A by deleting those rows and columns with indices in $\boldsymbol{J} \backslash \boldsymbol{K}$.

Conditions (S.a) and (S.b) are equivalent to the following:

(S.a) the matrix $(NR)_{\boldsymbol{K}}$ is an $\mathcal{S}$ matrix;

(S.b) the matrix $(NR)'_{\boldsymbol{K}}$ is an $\mathcal{S}$ matrix.

Definition 4.4. The convex polyhedron $\boldsymbol{S}$ is *simple* if for each $\boldsymbol{K} \subset \boldsymbol{J}$ such that $\boldsymbol{K} \neq \emptyset$ and $F_{\boldsymbol{K}} \neq \emptyset$, exactly $|\boldsymbol{K}|$ distinct faces contain $F_{\boldsymbol{K}}$.

The convex polyhedron $\boldsymbol{S}$ is *simple* if and only if for each $\boldsymbol{K} \subset \boldsymbol{J}$, $F_{\boldsymbol{K}} \neq \emptyset$ implies that $\boldsymbol{K}$ is maximal. One can check that the d-dimensional box in (3.18) is a simple polyhedron. The following proposition was proved in Dai and Williams [18, proposition 1.1]. It is a straightforward generalization of Reiman and Williams [36, lemma 3].

Proposition 1. Suppose that $\boldsymbol{S}$ is simple. Then (S.a) holds for all maximal $\boldsymbol{K} \subset \boldsymbol{J}$ if and only if (S.b) holds for all maximal $\boldsymbol{K} \subset \boldsymbol{J}$.

The following oscillation result is concerned with paths in a family of $(\boldsymbol{S}^r, R^r)$-regulations indexed by $r > 0$. In the case that $\boldsymbol{S} = \mathbb{R}^d_+$, and all paths are continuous and from a single $(\boldsymbol{S}, R)$-regulation, this result was proved previously by Bernard and El Kharroubi [2]. Dai and Williams [18] generalized the result to a general polyhedral state space. Our proof here is adapted from [18].

For any $f \in D([t_1, t_2], \mathbb{R}^k)$ with some $k \geqslant 1$, let

$$\begin{aligned} \operatorname{Osc}\bigl(f, [t_1, t_2]\bigr) &= \sup_{t_1 \leqslant s \leqslant t \leqslant t_2} \bigl|f(t) - f(s)\bigr|, \\ \operatorname{Osc}\bigl(f, [t_1, t_2)\bigr) &= \sup_{t_1 \leqslant s \leqslant t < t_2} \bigl|f(t) - f(s)\bigr|, \\ ||\Delta f||_{(t_1, t_2]} &= \sup_{t_1 < s \leqslant t_2} \bigl|\Delta f(s)\bigr|, \end{aligned}$$

where, as before, $\Delta f(s) = f(s) - f(s-)$ and $f(s-)$ is the left limit at s. Note that when f is left continuous at t_2, $\operatorname{Osc}(f, [t_1, t_2]) = \operatorname{Osc}(f, [t_1, t_2))$.

We consider a sequence of state spaces $\boldsymbol{S}^r$ indexed by $r > 0$. The shape of the space state does not change with r. That is, the normal vectors $\{n_i,\ i \in \boldsymbol{J}\}$ do not depend on r. However, the size $(a_1^r, \ldots, a_m^r)'$ of the state space depends on r. Hence,

$$\boldsymbol{S}^r \equiv \bigl\{x \in \mathbb{R}^d\colon\ n_i \cdot x \geqslant a_i^r \text{ for all } i \in \boldsymbol{J}\bigr\}.$$

The reflection matrix associated with each state space $\boldsymbol{S}^r$ is R^r, whose ith column is denoted by v_i^r. Recall that N is a matrix whose ith row is given by n_i'.

Theorem 4.2. Assume that $R^r \to R$ as $r \to \infty$ and (N, R) satisfies the *Completely-$\mathcal{S}$ assumption*. There exist constants $\kappa > 0$ and $\hat{r} > 0$ that depend only on (N, R) such that for any $T > 0$, $r \geqslant \hat{r}$, $x \in D([0,T], \mathbb{R}^d)$ with $x(0) \in \boldsymbol{S}^r$, and an $(\boldsymbol{S}^r, R^r)$-regulation (y, z) of x over $[0, T]$, the following holds for each interval $[t_1, t_2] \subset [0, T]$:

$$\begin{aligned}\mathrm{Osc}\bigl(y, [t_1, t_2]\bigr) &\leqslant \kappa\bigl(\mathrm{Osc}\bigl(x, [t_1, t_2]\bigr) + \|\Delta y\|_{(t_1, t_2]}\bigr),\\ \mathrm{Osc}\bigl(z, [t_1, t_2]\bigr) &\leqslant \kappa\bigl(\mathrm{Osc}\bigl(x, [t_1, t_2]\bigr) + \|\Delta y\|_{(t_1, t_2]}\bigr).\end{aligned}$$

We leave the lengthy proof to the end of this section. To prepare for the proof, we need a few lemmas.

Lemma 4.3. Let $f \in D([0, \infty), \mathbb{R})$. Suppose f is of bounded variation on each finite time interval, and assume that $f(0) = 0$. Then for each $t \geqslant 0$:

$$f^2(t) + \sum_{0 < s \leqslant t} \bigl[\Delta f(s)\bigr]^2 = 2 \int_{(0,t]} f(s)\, \mathrm{d}f(s).$$

Proof. The result is quite standard. See, for example, Last and Brandt [30, theorem A.4.6]. □

Let $g \in D([0, \infty), \mathbb{R})$ be a nondecreasing function. The function g is said to increase at time $t > 0$ if there exists a $\delta > 0$ such that $g(u) < g(v)$ for each $t - \delta < u < t < v < t + \delta$. The following lemma should also be standard. For completeness, we provide a direct proof.

Lemma 4.4. Let $g \in D([0, \infty, \mathbb{R})$ be a nondecreasing function and $f \in D([0, \infty), \mathbb{R})$ be a nonnegative function. For $t > 0$, if $\int_{(0,t]} f(s)\, \mathrm{d}g(s) = 0$ and $f(s) > 0$ for $s \in [0, t)$, then $g(s) = g(0)$ for $s \in [0, t)$.

Proof. Suppose that there is an $s \in (0, t)$ such that $g(s) > g(0)$. If g has jump at a point $t' \in (0, t)$, then

$$\int_{(0,t]} f(s)\, \mathrm{d}g(s) \geqslant f\bigl(t'\bigr)\Delta g\bigl(t'\bigr) > 0,$$

contradicting the fact that $\int_{(0,t]} f(s)\,dg(s) = 0$. Thus g must be continuous on $(0,t)$. Let

$$t' = \inf\big\{s \in (0,t)\colon\ g(s) > g(0)\big\}.$$

By the continuity of g, $g(t') = g(0)$. By the definition of t', for any $s > t'$, $g(s) > g(t')$. Because $f(t') > 0$ and f is right continuous, there is a $\delta > 0$ such that $\inf_{t' \leqslant s \leqslant t'+\delta} f(s) > 0$. Now,

$$\int_{(0,t]} f(s)\,\mathrm{d}g(s) \geqslant \int_{(t',t'+\delta]} f(s)\,\mathrm{d}g(s) \geqslant \inf_{t' \leqslant s \leqslant t'+\delta} f(s)\big(g\big(t'+\delta\big) - g\big(t'\big)\big) > 0,$$

contradicting the fact that $\int_{(0,t]} f(s)\,\mathrm{d}g(s) = 0$. Therefore, $g(s) = g(0)$ for $0 \leqslant s < t$. □

Lemma 4.5. Let $\boldsymbol{S} = [0,\infty)$ and $R = 1$. Then the $(\boldsymbol{S},R)$-regulation of x with $x(0) \geqslant 0$ has a unique solution (z,y) given by

$$\begin{aligned} y(t) &= \sup_{0 \leqslant s \leqslant t} x^-(s) \quad \text{for } 0 \leqslant t \leqslant T, \\ z(t) &= x(t) + y(t), \end{aligned}$$

where $x^-(t) = \max\{-x(t), 0\}$.

Proof. We first show the uniqueness. Suppose there are two solutions (z,y) and $(\hat{z},\hat{y})$ to the $(\boldsymbol{S},R)$-regulation of x. Then $z - \hat{z} = y - \hat{y}$. Now let $f = y - \hat{y}$. By lemma 4.3, we have for each $t \geqslant 0$

$$\begin{aligned} 0 \leqslant f^2(t) + \sum_{0<s\leqslant t} \big[\Delta f(s)\big]^2 &= 2\int_{(0,t]} f(s)\,\mathrm{d}f(s) \\ &= 2\int_{(0,t]} \big(z(s) - \hat{z}(s)\big)\,\mathrm{d}\big(y(s) - \hat{y}(s)\big) \\ &= -2\int_{(0,t]} \hat{z}(s)\,\mathrm{d}y(s) - 2\int_{(0,t]} z(s)\,\mathrm{d}\hat{y}(s) \leqslant 0. \end{aligned}$$

Hence, $f(t) = 0$, thus proving uniqueness.

For existence, let $y(t) = \sup_{0 \leqslant s \leqslant t} x^-(s)$. Since $x(0) \geqslant 0$, $x^-(0) = 0$ and so $y(0) = 0$. Clearly,

$$z(t) \equiv x(t) + y(t) \geqslant x(t) + x^-(t) \geqslant 0 \quad \text{for all } t \geqslant 0,$$

y is nondecreasing, and, hence, it has left limits on $(0,T]$. Since $x(\cdot)$ is right continuous, y is right continuous. It remains to be verified that y satisfies property (iii)(c) in the definition of the Skorohod problem. Suppose y has a jump at time t. Because

$$y\big(t^-\big) = \sup_{0 \leqslant s < t} x^-(s) \quad \text{and} \quad y(t) = \max\big\{y\big(t^-\big), x^-(t)\big\} > y\big(t^-\big),$$

we have $y(t) = x^-(t) = -x(t)$. Thus, $z(t) = x(t) + y(t) = x(t) + x^-(t) = 0$. Therefore, without loss of generality, we assume that y is continuous. If y increases at time t, it follows from the proof of lemma 8.1 in Chung and Williams [14] that $z(t) = 0$. Therefore, by Graves [22, p. 269],

$$\int_0^t z(s)\,\mathrm{d}y(s) = \lim_{n\to\infty} \sum_{k=1}^{2^n t} \Bigl(\inf_{s\in[(k-1)t/2^n,\, kt/2^n]} z(s) \Bigr) \left(y\left(\frac{kt}{2^n}\right) - y\left(\frac{(k-1)t}{2^n}\right)\right) = 0.$$

□

Let C be the constant determined in Dai and Williams [18, lemma B.1]. It depends on $\{n_i,\ i \in J\}$ only, not on $(a_1^r, \dots, a_m^r)'$. For each $\varepsilon \geqslant 0$ and $\boldsymbol{K} \subset \boldsymbol{J}$ (including the empty set), define

$$\begin{aligned} F_{\boldsymbol{K}}^{r,\varepsilon} = \bigl\{ x \in \mathbb{R}^d\colon\ & 0 \leqslant n_i x - a_i^r \leqslant C_\varepsilon \text{ for all } i \in \boldsymbol{K} \\ & \text{and } n_i x - a_i^r > \varepsilon \text{ for all } i \in \boldsymbol{J}\backslash\boldsymbol{K} \bigr\}, \end{aligned} \tag{4.3}$$

where $C_\varepsilon = Cm\varepsilon$. The following lemma, which was proved in [18, lemma 4.1], plays a key role in the proof of the oscillation theorem.

Lemma 4.6. For each $\varepsilon \geqslant 0$,

$$\boldsymbol{S}^r = \bigcup_{\boldsymbol{K}\in\mathcal{C}} F_{\boldsymbol{K}}^{r,\varepsilon}, \tag{4.4}$$

where $\mathcal{C}$ denotes the collection of subsets of $\boldsymbol{J}$ consisting of all maximal sets in $\boldsymbol{J}$ together with the empty set.

Proof of theorem 4.2. Our proof is adapted from that of lemma 4.3 in Dai and Williams [18] who generalized lemma 1 of Bernard and El Kharroubi [2]. We proceed via an induction on the size of $\boldsymbol{J}$, the index set for the faces of $\boldsymbol{S}$. Throughout this proof, $T,\ x,\ y,\ z,\ t_1,\ t_2$ will be as in the statement of the theorem. In general, z and y depend on the index r, but we suppress the dependence in the proof.

First consider the case $|\boldsymbol{J}| = 1$. Then $R^r = v_1^r$ is a vector in $\mathbb{R}^d$ and $v_1^r \to v_1$ as $r \to \infty$. By (S.a), $n_1 \cdot v_1 > 0$. Take r_0 such that

$$n_1 \cdot v_1^r \geqslant \frac{1}{2}(n_1 \cdot v_1) \quad \text{and} \quad \frac{||v_1^r||}{(n_1 \cdot v_1^r)} \leqslant \frac{2||v_1||}{n_1 \cdot v_1}$$

for $r \geqslant r_0$. Fix $r \geqslant r_0$. In this case, y is uniquely given by the one-dimensional regulator mapping for $n_1 \cdot x - a_1^r$ in lemma 4.5:

$$y(t) = \Bigl(- \min_{0\leqslant s\leqslant t} \bigl(n_1 \cdot x - a_1^r\bigr)(s)\Bigr)^+ \big/ \bigl(n_1 \cdot v_1^r\bigr) \quad \text{for all } t \in [0, T]. \tag{4.5}$$

Together with

$$n_1 \cdot z(t) = n_1 \cdot x(t) + n_1 \cdot v_1^r y(t) \quad \text{for all } t \in [0, T],$$

this defines a $([a_1^r,\infty), n_1\cdot v_1^r)$-regulation of $n_1\cdot x$ over $[0,T]$. The oscillation estimates in the theorem then follow easily from (4.5) and the fact that $z = x + v_1^r y$. That is, for $r \geqslant r_0$,

$$\begin{aligned}\operatorname{Osc}\big(y,[t_1,t_2]\big) &\leqslant \frac{1}{n_1\cdot v_1^r}\operatorname{Osc}\big(x,[t_1,t_2]\big) \leqslant \frac{2}{n_1\cdot v_1}\operatorname{Osc}\big(x,[t_1,t_2]\big),\\ \operatorname{Osc}\big(z,[t_1,t_2]\big) &\leqslant 1+\frac{||v_1^r||}{(n_1\cdot v_1^r)}\operatorname{Osc}\big(x,[t_1,t_2]\big) \leqslant 1+\frac{2||v_1||}{(n_1\cdot v_1)}\operatorname{Osc}\big(x,[t_1,t_2]\big).\end{aligned}$$

Thus the theorem holds for $|\boldsymbol{J}| = 1$ with $\hat{r} = r_0$ and

$$\kappa = \max\left\{\left(1+\frac{2||v_1||}{n_1\cdot v_1}\right),\ \frac{2}{n_1\cdot v_1}\right\}.$$

For the induction step, suppose that the theorem is true for $1 \leqslant |\boldsymbol{J}| < m$. Now consider a state space $\boldsymbol{S}$ with $|\boldsymbol{J}| = m$. Our proof of the induction step is separated into several parts.

Part (a). We claim that there exists a constant C_1 that depends only on (N,R) and a constant $r_0 > 0$ such that for $r \geqslant r_0$ and each $\boldsymbol{K} \in \mathcal{C}\backslash\{\boldsymbol{J}\}$ (see lemma 4.6 for the definition of $\mathcal{C}$), if $y_{\boldsymbol{J}\backslash\boldsymbol{K}}$ does not increase on $[t_1,t_2)$, then one has:

$$\operatorname{Osc}\big(y,[t_1,t_2]\big) \leqslant C_1\big(\operatorname{Osc}\big(x,[t_1,t_2]\big) + ||\Delta y||_{(t_1,t_2]}\big), \tag{4.6}$$

$$\operatorname{Osc}\big(z,[t_1,t_2]\big) \leqslant C_1\big(\operatorname{Osc}\big(x,[t_1,t_2]\big) + ||\Delta y||_{(t_1,t_2]}\big). \tag{4.7}$$

To see this, note that under the assumptions of the claim, for $t \in [0, t_2 - t_1)$,

$$z(t+t_1) = z(t_1) + x(t+t_1) - x(t_1) + \sum_{i\in\boldsymbol{K}} v_i^r\big(y_i(t+t_1) - y_i(t_1)\big). \tag{4.8}$$

For any t_2' such that $t_1 \leqslant t_2' < t_2$, it follows that $(z(\cdot+t_1), y_{\boldsymbol{K}}(\cdot+t_1) - y_{\boldsymbol{K}}(t_1))$ is an $(\boldsymbol{S}_{\boldsymbol{K}}^r, R_{\boldsymbol{K}}^r)$-regulation of $z(t_1) + x(\cdot+t_1) - x(t_1)$ over $[0, t_2' - t_1]$. If $\boldsymbol{K} = \emptyset$, then y does not increase on $[t_1, t_2']$ and the oscillation estimate trivially holds with $C_1 = 1$. If $\boldsymbol{K} \neq \emptyset$, then $\boldsymbol{K}$ is maximal and so by Dai and Williams [18, lemma 4.2], (S.a) and (S.b) hold for $(N_{\boldsymbol{K}}, R_{\boldsymbol{K}})$. Then, by the induction assumption, since $|\boldsymbol{K}| < m$, we have that there exist constants $C_{\boldsymbol{K}} \geqslant 1$ and $r_{0,\boldsymbol{K}} > 0$ that depend only on $(N_{\boldsymbol{K}},\ R_{\boldsymbol{K}})$, such that for $r \geqslant r_{0,\boldsymbol{K}}$

$$\begin{aligned}\operatorname{Osc}\big(y,[t_1,t_2']\big) &= \operatorname{Osc}\big(y_{\boldsymbol{K}}(\cdot+t_1),[0,t_2'-t_1]\big)\\ &\leqslant C_{\boldsymbol{K}}\Big(\operatorname{Osc}\big(x(\cdot+t_1)-x(t_1),\big[0,t_2'-t_1\big]\big) + \sup_{t_1<s\leqslant t_2'}\big|\Delta y_{\boldsymbol{K}}(s)\big|\Big)\\ &\leqslant C_{\boldsymbol{K}}\Big(\operatorname{Osc}\big(x,[t_1,t_2]\big) + \sup_{t_1<s\leqslant t_2}\big|\Delta y(s)\big|\Big).\end{aligned}$$

Letting $t_2' \uparrow t_2$,

$$\operatorname{Osc}\big(y,[t_1,t_2)\big) \leqslant C_{\boldsymbol{K}}\Big(\operatorname{Osc}\big(x,[t_1,t_2]\big) + \sup_{t_1<s\leqslant t_2}|\Delta y(s)|\Big).$$

Therefore,

$$\begin{aligned}\mathrm{Osc}\big(y,[t_1,t_2]\big) &\leqslant \mathrm{Osc}\big(y,[t_1,t_2)\big) + \big|\Delta y(t_2)\big| \\ &\leqslant 2C_{\boldsymbol{K}}\Big(\mathrm{Osc}\big(x,[t_1,t_2]\big) + \sup_{t_1<s\leqslant t_2}\big|\Delta y(s)\big|\Big).\end{aligned}$$

It follows from $z(t) = x(t) + R^r y(t)$ that

$$\begin{aligned}\mathrm{Osc}\big(z,[t_1,t_2]\big) &\leqslant \mathrm{Osc}\big(x,[t_1,t_2]\big) + ||R^r||\mathrm{Osc}\big(y,[t_1,t_2]\big) \\ &\leqslant \big(1+\|R^r\|2C_{\boldsymbol{K}}\big)\mathrm{Osc}\big(x,[t_1,t_2]\big).\end{aligned}$$

Because $R^r \to R$ as $r \to \infty$, we can choose r_0 such that r_0 is at least the maximum of the $r_{0,\boldsymbol{K}}$'s for $\boldsymbol{K}$ running through $\mathcal{C}\backslash\{\boldsymbol{J}\}$ and $||R^r|| \leqslant ||R|| + 1$ for $r \geqslant r_0$. Let C_1 be the maximum of $1 + (||R|| + 1)2C_{\boldsymbol{K}}$ for $\boldsymbol{K}$ running through $\mathcal{C}\backslash\{\boldsymbol{J}\}$. Then inequalities (4.6) and (4.7) follow.

For parts (b) and (c), we let

$$\varepsilon = \big(\mathrm{Osc}\big(x,[t_1,t_2]\big) + ||\Delta y||_{(t_1,t_2]}\big).$$

Without loss of generality we assume that $\varepsilon > 0$. By lemma 4.6, $z(t_1) \in F_{\boldsymbol{K}}^{C_1\varepsilon}$ for some $\boldsymbol{K} \in \mathcal{C}$.

Part (b). Suppose that the $\boldsymbol{K}$ found above is not $\boldsymbol{J}$. Then, for all $i \in \boldsymbol{J}\backslash\boldsymbol{K}$,

$$d\big(z(t_1),F_i\big) \geqslant n_i \cdot z(t_1) - a_i^r > C_1\varepsilon,$$

where $d(x,F)$ is the distance from a point x to a set F. We claim that $n_i\cdot z(s) - a_i^r > 0$ for $s \in [t_1,t_2]$ and $i \in \boldsymbol{J}\backslash\boldsymbol{K}$. Assume, on the contrary, that there exist $i \in \boldsymbol{J}\backslash\boldsymbol{K}$ and $s \in [t_1,t_2]$ such that $n_i \cdot z(s) - a_i^r = 0$. Let

$$t_2' = \inf\big\{s \in [t_1,t_2]\colon\ n_i \cdot z(s) - a_i^r = 0\big\}.$$

By the right continuity of z, $n_i \cdot z(t_2') - a_i^r = 0$. From the definition of t_2', $n_i \cdot z(s) - a_i^r > 0$ for $s \in [t_1,t_2')$, and hence y_i does not increase on $[t_1,t_2')$ by lemma 4.4. By part (a), we have

$$\begin{aligned}n_i \cdot z\big(t_2'\big) - a_i^r &= n_i \cdot \big(z\big(t_2'\big) - z(t_1)\big) + n_i \cdot z(t_1) - a_i^r \\ &> -C_1\Big(\mathrm{Osc}\big(x,\big[t_1,t_2'\big]\big) + \sup_{t_1<s\leqslant t_2'}\big|\Delta y(s)\big|\Big) + C_1\varepsilon \geqslant 0\end{aligned}$$

contradicting $n_i \cdot z(t_2') - a_i^r = 0$. Thus, z does not reach F_i^r for any $i \in \boldsymbol{J}\backslash\boldsymbol{K}$ during the interval $[t_1,t_2]$ and therefore $y_{\boldsymbol{J}\backslash\boldsymbol{K}}$ does not increase on $[t_1,t_2]$.

Then part (a) implies that (4.6) holds in this case.

Part (c). Suppose that the $\boldsymbol{K}$ described before part (b) is equal to $\boldsymbol{J}$. Since $z(t_1) \in F_{\boldsymbol{J}}^{C_1\varepsilon}$, by [18, lemma B.1], $d(z(t_1),F_i) \leqslant C_2\varepsilon$, where $C_2 = C_1Cm$. Now one of the following two situations holds.

(i) For every $i \in \boldsymbol{J}$, $d(z(t),F_i) \leqslant 2C_2\,\varepsilon$ for all $t \in [t_1,t_2]$. Then for each $i \in \boldsymbol{J}$,

$$0 \leqslant n_i \cdot z(t) - a_i^r \leqslant d\big(z(t),F_i\big) \leqslant 2C_2\varepsilon \quad \text{for all } t \in [t_1,t_2], \tag{4.9}$$

and so

$$\mathrm{Osc}\big(n_i \cdot z, [t_1, t_2]\big) \leqslant 2C_2\varepsilon. \tag{4.10}$$

Now, since $\boldsymbol{K} = \boldsymbol{J}$ is maximal, there is an $x_0 \in F_{\boldsymbol{J}}$ and by (S.b) there exists a positive linear combination $\eta = \sum_{i \in \boldsymbol{J}} \gamma_i n_i$ ($\gamma_i > 0$ for all i) of the $\{n_i,\ i \in \boldsymbol{J}\}$ such that $\eta \cdot v_i > 0$ for all $i \in \boldsymbol{J}$. Then

$$\eta \cdot \big(z(t) - x_0\big) = \eta \cdot \big(x(t) - x_0\big) + \sum_{i \in \boldsymbol{J}} \big(\eta \cdot v_i^r\big) y_i(t) \quad \text{for all } t \in [0, T]. \tag{4.11}$$

Thus,

$$\begin{aligned} &\min_{i \in \boldsymbol{J}} \big(\eta \cdot v_i^r\big) \mathrm{Osc}\big(y_1 + \cdots + y_m, [t_1, t_2]\big) \\ &\quad \leqslant \mathrm{Osc}\big(\eta \cdot z, [t_1, t_2]\big) + \mathrm{Osc}\big(\eta \cdot x, [t_1, t_2]\big) \\ &\quad \leqslant \sum_{i \in \boldsymbol{J}} \gamma_i \big(\mathrm{Osc}\big(n_i \cdot z, [t_1, t_2]\big) + \mathrm{Osc}\big(n_i \cdot x, [t_1, t_2]\big)\big). \end{aligned} \tag{4.12}$$

Since

$$\min_{i \in J} \big(\eta \cdot v_i^r\big) \to \min_{i \in J} (\eta \cdot v_i) > 0$$

as $r \to \infty$, using (4.10) and $z = x + R^r y$, we see that one can choose a constant C_3 depending only on $(N,\ R)$ and an $r_1 > r_0$ such that

$$\mathrm{Osc}\big(y, [t_1, t_2]\big) \leqslant C_3\varepsilon, \qquad \mathrm{Osc}\big(z, [t_1, t_2]\big) \leqslant C_3\varepsilon.$$

(ii) There is an $i \in \boldsymbol{J}$ and $t_3 \in [t_1, t_2]$ such that $d(z(t_3), F_i) > 2C_2\varepsilon$. Define

$$t_1' = \inf \big\{ t > t_1\colon\ d\big(z(t), F_i\big) > 2C_2\varepsilon \text{ for some } i \in \boldsymbol{J} \big\}.$$

By the definition of t_1', for any $\delta > 0$, over $[t_1, t_1' - \delta]$ we have the situation in part (c)(i) above. That is,

$$\mathrm{Osc}\big(y, \big[t_1, t_1' - \delta\big]\big) \leqslant C_3\varepsilon, \qquad \mathrm{Osc}\big(z, [t_1, t_1' - \delta]\big) \leqslant C_3\varepsilon.$$

Letting $\delta \to 0^+$, we have

$$\mathrm{Osc}\big(y, \big[t_1, t_1'\big)\big) \leqslant C_3\varepsilon, \qquad \mathrm{Osc}\big(z, \big[t_1, t_1'\big)\big) \leqslant C_3\varepsilon.$$

Over $[t_1', t_2]$, by lemma 4.6, we have $z(t_1') \in F_{\boldsymbol{K}}^{C_1\varepsilon}$ for some $\boldsymbol{K} \in \mathcal{C} \backslash \{\boldsymbol{J}\}$, and then we have the situation in part (b). Thus,

$$\mathrm{Osc}\big(y, \big[t_1', t_2\big]\big) \leqslant C_1\varepsilon, \qquad \mathrm{Osc}\big(z, \big[t_1', t_2\big]\big) \leqslant C_1\varepsilon.$$

Therefore,

$$\mathrm{Osc}\bigl(y,[t_1,t_2]\bigr) \leqslant \mathrm{Osc}\bigl(y,\bigl[t_1,t_1'\bigr)\bigr) + \bigl|\Delta y\bigl(t_1'\bigr)\bigr| + \mathrm{Osc}\bigl(y,\bigl[t_1',t_2\bigr]\bigr) \leqslant (1+C_1+C_3)\varepsilon.$$

Hence, there is a constant C_4 depending only on (N, R) such that

$$\mathrm{Osc}\bigl(y,[t_1,t_2]\bigr) \leqslant C_4\varepsilon, \qquad \mathrm{Osc}\bigl(z,[t_1,t_2]\bigr) \leqslant C_4\varepsilon.$$

Thus, the theorem holds for $\kappa = \max\{C_1,C_3,C_4\}$ and $\hat{r} = r_1$. □

5. Network dynamics and preliminaries

For each $t \geqslant 0$, $i \in \boldsymbol{I}$ and $j > 0$ let $U_i(0) = V_i(0) = 0$,

$$\begin{aligned} &U_i(j) = u_{i1} + \cdots + u_{ij}, \qquad V_i(j) = v_{i1} + \cdots + v_{ij},\\ &E_i(t) = \max\{k \geqslant 0\colon\ u_{i1} + \cdots + u_{ik} \leqslant t\},\\ &S_i(t) = \max\{k \geqslant 0\colon\ v_{i1} + \cdots + v_{ik} \leqslant t\}. \end{aligned}$$

Let

$$\widehat{E}_i(t) = E_i(t) - t \quad \text{and} \quad \widehat{S}_i(t) = S_i(t) - t \quad \text{for } i \in \boldsymbol{I}.$$

Let

$$\widehat{E}(t) = \bigl(\widehat{E}_1(t),\dots,\widehat{E}_d(t)\bigr)' \quad \text{and} \quad \widehat{S}(t) = \bigl(\widehat{S}_1(t),\dots,\widehat{S}_d(t)\bigr)'.$$

The two d-dimensional processes $\{\widehat{E}(t),\ t \geqslant 0\}$ and $\{\widehat{S}(t),\ t \geqslant 0\}$ contain all the randomness in the queueing network. It is known that they satisfy the Functional Strong Law of Large Numbers [23, lemma V.2.1]: $\mathbb{P}$-a.s, as $r \to \infty$,

$$\frac{1}{r}\widehat{E}(r\cdot) \to 0 \text{ u.o.c.}, \qquad \frac{1}{r}\widehat{S}(r\cdot) \to 0 \text{ u.o.c.} \tag{5.1}$$

and the Functional Central Limit Theorem [4, section 17]: as $r \to \infty$,

$$\left(\frac{1}{\sqrt{r}}\widehat{E}(r\cdot), \frac{1}{\sqrt{r}}\widehat{S}(r\cdot)\right) \Longrightarrow \bigl(E^*, S^*\bigr), \tag{5.2}$$

where E^* and S^* are two independent, d-dimensional Brownian motions with drift zero and covariance matrices $\mathrm{diag}(c_1^a,\dots,c_d^a)$ and $\mathrm{diag}(c_1^s,\dots,c_d^s)$, respectively.

Recall that we are considering a sequence of networks indexed by n. In particular, α_i^n and μ_i^n are the external arrival rate to station i and the service rate of server i for the nth network. Let

$$E_i^n(t) = E_i\bigl(\alpha_i^n t\bigr), \qquad S_i^n(t) = S_i\bigl(\mu_i^n t\bigr).$$

If server i has been busy all the time in $[0,t]$, $S_i^n(t)$ is the number of services completed by time t at station i. Similarly, if buffer i has never been full in $[0,t]$, $E_i^n(t)$ is the number of arrivals by time t to station i. Recall that $F_i^n(t)$ is the cumulative amount of time that buffer i is not full by time t. From our model assumption, $E_i^n(F_i^n(t))$

is the number of external arrivals to station i by time t in the nth network. Also, $B_i^n(t)$ is the cumulative amount of time that server i has been working by time t and $S_i^n(B_i^n(t))$ is the number of departures from station i by time t in the nth network. Now we can write down the main equation that governs the dynamics of the queue length process. Namely,

$$Z_i^n(t) = Z_i^n(0) + E_i^n\big(F_i^n(t)\big) + \sum_{j\in\boldsymbol{I},\sigma(j)=i} S_j^n\big(B_j^n(t)\big) - S_i^n\big(B_i^n(t)\big), \quad i \in \boldsymbol{I}, \tag{5.3}$$

where $Z_i^n(0)$ is the initial queue length at station i. Let

$$E^n\big(F^n(t)\big) = \big(E_1^n\big(F_1^n(t)\big), \ldots, E_d^n\big(F_d^n(t)\big)\big)'$$

and

$$S^n\big(B^n(t)\big) = \big(S_1^n\big(B_1^n(t)\big), \ldots, S_d^n\big(B_d^n(t)\big)\big)'.$$

Recall that the routing matrix is defined as

$$P_{ij} = \begin{cases} 1 & \text{if station } i \text{ customers go to station } i, \\ 0 & \text{otherwise.} \end{cases}$$

Then we have the vector form of (5.3):

$$Z^n(t) = Z^n(0) + E^n\big(F^n(t)\big) - (I - P')S^n\big(B^n(t)\big). \tag{5.4}$$

Following Harrison [24], we introduce the centered processes

$$\widehat{E}^n(t) = \big(\widehat{E}_1^n(t), \ldots, \widehat{E}_d^n(t)\big)' \quad \text{and} \quad \widehat{S}^n(t) = \big(\widehat{S}_1^n(t), \ldots, \widehat{S}_d^n(t)\big)',$$

where

$$\widehat{E}_i^n(t) = E_i^n(t) - \alpha_i^n t = \widehat{E}_i\big(\alpha_i^n t\big) \quad \text{and} \quad \widehat{S}_i^n(t) = S_i^n(t) - \mu_i^n t = \widehat{S}_i\big(\mu_i^n t\big), \quad i \in I. \tag{5.5}$$

It follows from (5.4) that

$$\begin{aligned} Z^n(t) = {} & Z^n(0) + \widehat{E}^n\big(F^n(t)\big) - \big(I - P'\big)\widehat{S}^n\big(B^n(t)\big) \\ & + \operatorname{diag}\big(\alpha^n\big)F^n(t) - \big(I - P'\big)\operatorname{diag}\big(\mu^n\big)B^n(t). \end{aligned} \tag{5.6}$$

It follows from (5.6) and (2.2) that

$$Z^n(t) = Z^n(0) + X^n(t) + R^n Y^n(t), \tag{5.7}$$

where

$$X^n(t) = \widehat{E}^n\big(F^n(t)\big) - \big(I - P'\big)\widehat{S}^n\big(B^n(t)\big) + \big(\alpha^n - \big(I - P'\big)\mu^n\big)t, \tag{5.8}$$

R^n is the $d \times 2d$ matrix given by

$$R^n = \big(\big(I - P'\big)\operatorname{diag}\big(\mu^n\big), \big[\big(I - P'\big)\operatorname{diag}\big(\mu^n\big)\big]_\sigma - \operatorname{diag}\big(\alpha^n\big)\big)$$

and for a matrix A, A_σ is defined in (3.17). Let $\boldsymbol{S}^n$ be the d-dimensional box defined by

$$\boldsymbol{S}^n = \{x \in \mathbb{R}^d\colon\ 0 \leqslant x_i \leqslant b_i^n\ \forall i \in \boldsymbol{I}\}.$$

One can check that for each sample path:

(i) $Z^n(t) = Z^n(0) + X^n(t) + R^n Y^n(t)$ for all $t \geqslant 0$,

(ii) $Z^n(t) \in \boldsymbol{S}^n$ for all $t \geqslant 0$,

(iii) for each $i = 1, \ldots, 2d$,

(a) $Y_i^n(0) = 0$,

(b) Y_i^n is nondecreasing and continuous,

(c) for $i = 1, \ldots, d$, Y_i increases only when $Z_i^n(t) = 0$ and for $i = d+1, \ldots, 2d$, Y_i^n increases only when $Z_i^n(t) = b_i^n$.

It follows that for each sample path, the pair $(Z^n(\cdot), Y^n(\cdot))$ is an $(\boldsymbol{S}^n, R^n)$-regulation of $Z^n(0) + X^n(\cdot)$.

Using the notion in section 4, for each boundary face of $\boldsymbol{S}^n$, there is a unit vector n_i that is normal to the face. (We number faces such that the ith face is $\{x \in S^n\colon\ x_i = 0\}$ for $i = 1, \ldots, d$ and $\{x \in S^n\colon\ x_i = b_i^n\}$ for $i = d+1, \ldots, 2d$.) Recall that N is a $2d \times d$ matrix whose ith row is the row vector n_i'. It is easy to check that

$$N = (I, -I)',$$

where I is the $d \times d$ identity matrix. Under the assumptions (3.1), as $n \to \infty$, $R^n \to R$ as defined in (3.16).

Lemma 5.1. The completely-$\mathcal{S}$ assumption in theorem 4.2 holds for (N, R).

Proof. Because the state space S^n is simple, by proposition 1, it is enough to show that for each maximal $\boldsymbol{K} \subset \boldsymbol{J} \equiv \{1, \ldots, 2d\}$, $(NR)_{\boldsymbol{K}}$ is an $\mathcal{S}$-matrix. Let R_0 and R_b be two $d \times d$ submatrices of R such that $R = (R_0, R_b)$. It is easy to check that

$$NR = \begin{pmatrix} R_0 & R_b \\ -R_0 & -R_b \end{pmatrix}.$$

A $\boldsymbol{K} \subset \boldsymbol{J}$ is maximal if $\bigcap_{i \in \boldsymbol{K}} F_i^n$ is non-empty. Because F_i^n and F_{i+d}^n are parallel to each other, a non-empty $\boldsymbol{K}$ is maximal if and only if for each $i \in \boldsymbol{K}$, $i + d \notin \boldsymbol{K}$. Let $M = (NR)_{\boldsymbol{K}}$. Then M has the following form:

$$M = \begin{pmatrix} M_1 & M_2 \\ M_3 & M_4 \end{pmatrix} = \begin{pmatrix} M_1 & 0 \\ 0 & M_4 \end{pmatrix} + \begin{pmatrix} 0 & M_2 \\ M_3 & 0 \end{pmatrix},$$

where M_1 is a principal submatrix of R_0, M_4 is a principal submatrix of $-R_b$, M_2 is a submatrix of R_b and M_3 is a submatrix of $-R_0$. Because $\boldsymbol{K}$ is maximal, M_3

does not contain any diagonal elements of $-R_0$. Hence, M_3 is a nonnegative matrix. Similarly, M_2 is a nonnegative matrix. Because R_0 is a completely-$\mathcal{S}$ matrix, hence, M_1 is an $\mathcal{S}$-matrix. Because $-R_b$ is an upper triangular matrix with positive diagonal elements, M_4 is an $\mathcal{S}$-matrix. Thus, M is an $\mathcal{S}$-matrix. □

We end this section by presenting an example in which the associated Skorohod problem does not have a unique solution. Consider, for example, a network of two stations in tandem. The routing matrix

$$P = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Assume that $\alpha^n = (1,0)'$ and $\mu^n = (1,1)'$ for each n. Then the corresponding reflection matrix

$$R = \begin{pmatrix} 1 & 0 & -1 & 1 \\ -1 & 1 & 0 & -1 \end{pmatrix}.$$

We claim that the $(\boldsymbol{S}, R)$-regulation of $x(\cdot)$ is *not* unique for some path $x(\cdot)$. Here the state space $\boldsymbol{S}$ is a box, say, $\{x \in \mathbb{R}^2\colon\ 0 \leqslant x_i \leqslant 1 \text{ for } i = 1,2\}$. Note that the directions of reflection that correspond to the corner $(0,1)'$ are *parallel*, both being $(1,-1)'$. Let $x_1(t) = -t$ and $x_2(t) = 1 + t$ for $t \geqslant 0$. One can check that both (z,y) and $(\hat{z},\hat{y})$ are $(\boldsymbol{S}, R)$-regulations of $x(\cdot)$, where, for $t \geqslant 0$,

$$z_1(t) = 0, \quad z_2(t) = 1, \quad y_1(t) = t, \quad y_2(t) = y_3(t) = y_4(t) = 0,$$

$$\hat{z}_1(t) = 0, \quad \hat{z}_2(t) = 1, \quad \hat{y}_1(t) = \hat{y}_2(t) = \hat{y}_3(t) = 0, \quad \text{and} \quad \hat{y}_4(t) = t.$$

Thus, the conventional continuous mapping approach for proving heavy traffic limit theorems cannot be applied to the network.

6. Fluid limits

Theorem 6.1 (Fluid limit theorem). Assume that (3.1)–(3.3) in theorem 3.1 hold. Then, for each sample path such that (5.1) holds,

$$\frac{1}{n}F^n(n\cdot) \longrightarrow et \text{ u.o.c.,} \quad \text{and} \quad \frac{1}{n}B^n(n\cdot) \longrightarrow et \text{ u.o.c.} \tag{6.1}$$

Proof. Fix a sample path such that (5.1) holds. For a sequence of paths f^n, recall that

$$\bar{f}^n(t) = \frac{1}{n}f^n(nt).$$

By (5.1),

$$\frac{\widehat{E}^n(nt)}{n} = \frac{\widehat{E}(\alpha^n t)}{n} \to 0 \text{ u.o.c.} \quad \text{and} \quad \frac{\widehat{S}^n(nt)}{n} = \frac{\widehat{S}(\mu^n t)}{n} \to 0 \text{ u.o.c.} \tag{6.2}$$

Let ω be a fixed sample path such that (6.2) holds. We claim that as $n \to \infty$,

$$\overline{X}^n(t) \to 0 \text{ u.o.c.}$$

In fact, for $s > 0$,

$$\left|\overline{X}^n(s)\right| \leqslant \frac{1}{n}\left|\widehat{E}^n\left(F^n(ns)\right)\right| + \left\|\left(I-P'\right)\right\|\frac{1}{n}\left|\widehat{S}^n\left(B^n(ns)\right)\right| + \left|\alpha^n - \left(I-P'\right)\mu^n\right|s,$$

where for a matrix A, $||A|| = \max_i \sum_j |A_{ij}|$. Thus,

$$\sup_{0\leqslant s\leqslant t}\left|\overline{X}^n(s)\right| \leqslant \frac{1}{n}\sup_{0\leqslant s\leqslant t}\left|\widehat{E}^n(ns)\right| + \left\|\left(I-P'\right)\right\|\frac{1}{n}\sup_{0\leqslant s\leqslant t}\left|\widehat{S}^n(ns)\right| + \left|\alpha^n - \left(I-P'\right)\mu^n\right|t$$

and $\sup_{0\leqslant s\leqslant t}|\overline{X}^n(s)| \to 0$ as $n \to \infty$.

It is easy to check that $(\overline{Z}^n(\cdot), \overline{Y}^n(\cdot))$ is an $(\overline{S}^n, R^n)$-regulation of $\overline{Z}^n(0) + \overline{X}^n(\cdot)$, where

$$\overline{S}^n = \left\{x \in \mathbb{R}^d \colon\ 0 \leqslant x_i \leqslant b_i^n/n\ \forall i \in \boldsymbol{I}\right\}.$$

By lemma 5.1, theorem 4.2 and the fact that $\overline{Y}^n(\cdot)$ is continuous, there exist constants $\kappa > 0$ and $n_0 > 0$ such that for each $t_1 < t_2$ and $n \geqslant n_0$,

$$\operatorname{Osc}\left(\overline{Y}^n(\cdot,\omega), [t_1, t_2]\right) \leqslant \kappa \operatorname{Osc}\left(\overline{X}^n(\cdot), [t_1, t_2]\right). \tag{6.3}$$

Since $|\overline{Y}^n(t) - \overline{Y}^n(s)| \leqslant 2d(t-s)$ for all n and $t > s > 0$, the sequence $\{\overline{Y}^n(\cdot)\}$ is precompact in $C([0,\infty), \mathbb{R}^{2d})$. Let $\overline{Y}$ be a limit of this sequence. Because $\overline{X}^n(\cdot,\omega) \to 0$ u.o.c. as $n \to \infty$, it follows from (6.3) that $\operatorname{Osc}(\overline{Y}, [t_1, t_2]) = 0$ for any $0 \leqslant t_1 < t_2$. Since $\overline{Y}(0) = 0$, $\overline{Y}(t) = 0$ for all $t \geqslant 0$. Because each limit point $\overline{Y}$ is identically zero, $\overline{Y}^n(\cdot) \to 0$ u.o.c. as $n \to \infty$. The lemma then follows from (2.2). □

7. A stopping time property

In this section we prove a stopping time property that is essential to the proof of the main theorem. Let $p, q \in \mathbb{Z}_+^d$ be d-dimensional indexes. We use $U_i(\cdot \wedge j)$ to denote process $\{U_i(k \wedge j),\ k \geqslant 0\}$. For a d-dimensional index p, let $U(\cdot \wedge p) = (U_1(\cdot \wedge p_1), \ldots, U_d(\cdot \wedge p_d))$. For any $p, q \in \mathbb{Z}_+^d$, let

$$\mathcal{G}_{p,q}^n = \sigma\left\{U\left(\,\cdot \wedge (p+e)\right), V\left(\,\cdot \wedge (q+e)\right), Z^n(0)\right\}, \tag{7.1}$$

where e is the d-dimensional vector of ones. We assume that $\mathcal{G}_{p,q}^n$ has been augmented with all $\mathbb{P}$-null sets. Recall that $E_i^n(F_i^n(t))$ is the number of external arrivals to station i by time t and $S_i^n(B_i^n(t))$ is the number of departures from station i by time t.

Lemma 7.1 (Stopping time property). For any $p, q \in \mathbb{Z}_+^d$ and $t \geqslant 0$,

$$\left\{E^n\left(F^n(t)\right) = p,\ S^n\left(B^n(t)\right) = q\right\} \in \mathcal{G}_{p,q}^n. \tag{7.2}$$

Proof. Since we are going to prove (7.2) is true for each $n \geqslant 1$, we drop the superscript n in this proof. Let

$$A(t) = E\big(F(t)\big) \quad \text{and} \quad D(t) = S\big(B(t)\big).$$

When the event $A_i(t) = p_i$ occurs, the dynamics of the network in $[0, t]$ does *not* depend on the interarrival times $u_{i\ell}$ for $\ell > p_i + 1$. Similarly, when the event $D_i(t) = q_i$ occurs, the dynamics of the network in $[0, t]$ does not depend on the service times $v_{i\ell}$ for $\ell > q_i + 1$. Thus, the lemma is intuitively obvious. However, a rigorous proof is needed to show that $A(t)$ and $D(t)$ *measurably* depend on interarrival and service times. The proof essentially requires us to go through the detailed construction of $A(t)$ and $D(t)$ from the primitive interarrival and service times.

We mimic the proof in Williams [38], where open multiclass queueing networks with *unlimited* buffer size were considered. An event time is the instant when a service completion or an arrival has just occurred. Let $e_0 = 0$ and e_l be the lth event time. Because the mean interarrival times and mean service times are positive, with probability one, $e_l \to \infty$ as $l \to \infty$. Thus, we have, with probability one,

$$\big\{A(t) = p,\ D(t) = q\big\} = \bigcup_{k \geqslant 1} \bigcap_{l \geqslant k} \big\{A(t \wedge e_l) = p,\ D(t \wedge e_l) = q\big\}.$$

Therefore, to show (7.2), it is enough to show

$$\big\{A(t \wedge e_l) = p,\ D(t \wedge e_l) = q\big\} \in \mathcal{G}_{p,q}.$$

for each $t \geqslant 0$, $l \geqslant 0$ and $p, q \in \mathbb{Z}_+^d$. (Here we used the fact that each $\mathcal{G}_{p,q}$ has been augmented with all $\mathbb{P}$-null sets.) For each $t \geqslant 0$ and $i \in \boldsymbol{I}$, let $R_i^a(t)$ be the remaining time (from time t) for the next external arrival to station i to occur if the arrival will never be turned off. Similarly, let $R_i^s(t)$ be the remaining time for the next service at station i to complete if the service will never be interrupted. If there is no customer in service at time t, $R_i^s(t) = \infty$. We adopt the convention that $\infty - a = \infty$ and $\min\{\infty, a\} = a$ for any constant a. We want to use induction to show that for each $l \geqslant 0$

$$C_{l,p,q} \equiv \big\{A(t \wedge e_l) = p,\ D(t \wedge e_l) = q\big\} \in \mathcal{G}_{p,q}, \tag{7.3}$$

$$1_{\{A(t \wedge e_l) = p, D(t \wedge e_l) = q\}} \xi_l \in \mathcal{G}_{p,q} \tag{7.4}$$

hold for each $t \geqslant 0$ and $p, q \in \mathbb{Z}_+^d$, where

$$\xi_l = \big(Z(t \wedge e_l),\ R^a(t \wedge e_l),\ R^s(t \wedge e_l),\ t \wedge e_l\big).$$

From our model assumption, $A_i(0) = 0$ and $D_i(0) = 0$, $R_i^a(0) = u_{i1}/\alpha_i$ and

$$R_i^s(0) = \begin{cases} m_i v_{i1} & \text{if } Z_i(0) > 0, \\ \infty & \text{if } Z_i(0) = 0. \end{cases}$$

Thus, $\xi_0 = (Z(0), R^a(0), R^s(0), 0) \in \mathcal{G}_{0,0}$. For any $(p, q) \neq (0, 0)$,

$$1_{\{A(t \wedge e_0) = p, D(t \wedge e_0) = q\}} \xi_0 = 0 \in \mathcal{G}_{p,q}.$$

Therefore, (7.3) and (7.4) hold for $l = 0$.

We now make the induction assumption that $C_{l,p,q} \in \mathcal{G}_{p,q}$ and $1_{C_{l,p,q}}\xi_l \in \mathcal{G}_{p,q}$ for all $p, q \in \mathbb{Z}_+^d$ and $t \geqslant 0$. We would like to show that $C_{l+1,p,q} \in \mathcal{G}_{p,q}$ and $1_{C_{l+1,p,q}}\xi_{l+1} \in \mathcal{G}_{p,q}$ for all $p, q \in \mathbb{Z}_+^d$ and $t \geqslant 0$. We first show that $1_{C_{l+1,p,q}}\xi_{l+1} \in \mathcal{G}_{p,q}$. Note that

$$1_{C_{l+1,p,q}}\xi_{l+1} = 1_{C_{l+1,p,q}}\xi_{l+1}1_{\{t\leqslant e_l\}} + 1_{C_{l+1,p,q}}\xi_{l+1}1_{\{e_l<t\}}.$$

It is clear that

$$1_{C_{l+1,p,q}}\xi_{l+1}1_{\{t\leqslant e_l\}} = 1_{C_{l,p,q}}\xi_l 1_{\{t=t\wedge e_l\}} \in \mathcal{G}_{p,q}$$

by the induction assumption. It remains to be shown that

$$1_{C_{l+1,p,q}}\xi_{l+1}1_{\{e_l<t\}} \in \mathcal{G}_{p,q}.$$

On $\{t > e_l\}$,

$$e_{l+1} = t \wedge e_l + \min_{i\in \boldsymbol{I}\setminus\boldsymbol{F},\, j\in\boldsymbol{I}\setminus\boldsymbol{B}} \bigl\{R_i^a(t\wedge e_l),\ R_j^s(t\wedge e_l)\bigr\}, \tag{7.5}$$

where $\boldsymbol{F} \subset \boldsymbol{I}$ is the set of buffers that are full at time $t \wedge e_l$, i.e.,

$$\boldsymbol{F} \equiv \bigl\{i \in \boldsymbol{I}:\ Z_i(t\wedge e_l) = b_i\bigr\},$$

and $\boldsymbol{B} \subset \boldsymbol{I}$ is the set of stations are blocked at time $t \wedge e_l$, i.e.,

$$\boldsymbol{B} \equiv \bigl\{i \in \boldsymbol{I}:\ Z_{\sigma(i)}(t\wedge e_l) = b_{\sigma(i)}\bigr\}.$$

It follows from (7.5) and the induction assumption that

$$1_{C_{l,m,n}}1_{\{t>e_l\}}e_{l+1} \in \mathcal{G}_{m,n} \tag{7.6}$$

for all $m, n \in \mathbb{Z}_+^d$.

Now,

$$1_{C_{l+1,p,q}}\xi_{l+1}1_{\{t>e_l\}} = \sum_{(\boldsymbol{a},\boldsymbol{s})} 1_{C_{l,\tilde p,\tilde q}}\xi_{l+1}1_{\{t>e_l\}}1_{B_{\boldsymbol{a},\boldsymbol{s}}},$$

where

$$\begin{aligned}
B_{\boldsymbol{a},\boldsymbol{s}} = {} & \bigcap_{i\in\boldsymbol{a}} \bigl\{R_i^a(t\wedge e_l) = e_{l+1} - t\wedge e_l\bigr\} \\
& \times \bigcap_{i\notin\boldsymbol{a}} \bigl(\bigl\{R_i^a(t\wedge e_l) > e_{l+1} - t\wedge e_l\bigr\} \cup \bigl\{Z_i(t\wedge e_l) = b_i\bigr\}\bigr) \\
& \times \bigcap_{i\in\boldsymbol{s}} \bigl\{R_i^s(t\wedge e_l) = e_{l+1} - t\wedge e_l\bigr\} \\
& \times \bigcap_{i\notin\boldsymbol{s}} \bigl(\bigl\{R_i^s(t\wedge e_l) > e_{l+1} - t\wedge e_l\bigr\} \cup \bigl\{Z_{\sigma(i)}(t\wedge e_l) = b_{\sigma(i)}\bigr\}\bigr),
\end{aligned}$$

$$\tilde{p}_i = \begin{cases} p_i - 1 & \text{if } i \in \boldsymbol{a}, \\ p_i & \text{if } i \notin \boldsymbol{a}, \end{cases} \qquad \tilde{q}_i = \begin{cases} q_i - 1 & \text{if } i \in \boldsymbol{s}, \\ q_i & \text{if } i \notin \boldsymbol{s}, \end{cases}$$

$$Z_i(t \wedge e_{l+1}) = \begin{cases} Z_i(t \wedge e_l) + 1 & \text{if } i \in \boldsymbol{a} \setminus \boldsymbol{s}, \\ Z_i(t \wedge e_l) - 1 & \text{if } i \in \boldsymbol{s} \setminus \boldsymbol{a}, \\ Z_i(t \wedge e_l) & \text{otherwise}, \end{cases}$$

$$R_i^a(t \wedge e_{l+1}) = \begin{cases} u_{i,p_i} & \text{if } i \in \boldsymbol{a}, \\ R_i^a(t \wedge e_l) - (t \wedge e_{l+1} - t \wedge e_l) & \text{if } i \in \boldsymbol{I} \setminus \boldsymbol{a}, \end{cases}$$

$$R_i^s(t \wedge e_{l+1}) = \begin{cases} v_{i,q_i} & \text{if } i \in \boldsymbol{s}, \\ R_i^s(t \wedge e_l) - (t \wedge e_{l+1} - t \wedge e_l) & \text{if } i \in \boldsymbol{I} \setminus \boldsymbol{s}, \end{cases}$$

and the summation is over all pairs $(\boldsymbol{a}, \boldsymbol{s})$ with $\boldsymbol{a} \subset \boldsymbol{I}$ and $\boldsymbol{s} \subset \boldsymbol{I}$. The set $\boldsymbol{a} \cup \boldsymbol{s}$ is the set of indexes whose clocks "expire" exactly at e_{l+1}. If $\boldsymbol{a} \cup \boldsymbol{s} = \emptyset$, then the $(l+1)$th event has not yet happened by time t. It follows from (7.6) and the induction assumption that $1_{C_{l+1,p,q}} \xi_{l+1}$ is $\mathcal{G}_{p,q}$ measurable. Similarly, we can show that $1_{C_{l+1,p,q}} \in \mathcal{G}_{p,q}$. □

8. Proof of the heavy traffic limit theorem

For a sequence of functions f^n, recall that

$$\tilde{f}^n(t) = \frac{1}{\sqrt{n}} f^n(nt).$$

Lemma 8.1. Under the assumptions (3.1)–(3.3) in theorem 3.1, as $n \to \infty$,

$$\big(\widetilde{E}^n, \widetilde{S}^n, \widetilde{X}^n\big) \Longrightarrow \big(E^*, S^*, X^*\big),$$

where E^* and S^* are independent Brownian motions as in (5.2),

$$X^*(t) = E^*(\alpha t) - \big(I - P'\big) S^*(\mu t) + \theta t, \tag{8.1}$$

and X^* is a Brownian motion with drift θ and covariance matrix Γ given in (3.16).

Proof. Let $\widetilde{E}^n(t) = (1/\sqrt{n})\widehat{E}(\alpha^n nt)$ and $\widetilde{S}^n(t) = (1/\sqrt{n})\widehat{S}(\mu^n nt)$. It follows from (5.2), (3.1), (6.1) and the Random Change of Time Theorem [4, section 17] that

$$\big(\widetilde{E}^n\big(\overline{F}^n(\cdot)\big), \widetilde{S}^n\big(\overline{B}^n(\cdot)\big)\big) \Longrightarrow \big(E^*(\alpha \cdot), S^*(\mu \cdot)\big).$$

By the continuous mapping theorem,

$$\begin{aligned} \widetilde{X}^n(\cdot) \; &= \; \widetilde{E}^n\big(\overline{F}^n(\cdot)\big) - \big(I - P'\big)\widetilde{S}^n\big(\overline{B}^n(\cdot)\big) - \big(I - P'\big)\mu^n \cdot \\ &\Longrightarrow E^*(\alpha \cdot) - \big(I - P'\big) S^*(\mu \cdot) + \theta \cdot . \end{aligned}$$

It is easy to check that X^* is a Brownian motion with drift θ and covariance matrix Γ given in (3.16). □

A sequence of stochastic processes $\{X^n\}$ in $D([0,\infty),\mathbb{R}^k)$ is said to be relatively compact if for every sequence $\{n_k\}$, there is a subsequence $\{n_{k_j}\}$ such that $X^{n_{k_j}}$ converges in distribution.

Lemma 8.2. Under the assumptions (3.1)–(3.4) in theorem 3.1, the sequence $\{\widetilde{X}^n, \widetilde{Z}^n, \widetilde{Y}^n\}$ is relatively compact.

Proof. To prove the lemma it suffices to verify conditions (a) and (b) in corollary 7.4 in chapter 3 of Ethier and Kurtz [21]. To state the conditions, we need to define the modulus of continuity of a path $x(\cdot)$. For $T > 0$ and $\delta > 0$, let

$$w\bigl(x(\cdot),\delta,T\bigr) = \inf_{t_i}\max_{i} \operatorname{Osc}\bigl(x(\cdot),[t_{i-1},t_i)\bigr), \tag{8.2}$$

where the infimum extends over the finite sets $\{t_i\}$ of points satisfying $0 = t_0 < t_1 < \cdots < t_r = T$ and $t_j - t_{j-1} > \delta$ for $j = 1,\ldots,r$.

(a) For every $\eta > 0$ and rational $t \geqslant 0$, there exists a constant $c(\eta,t) > 0$ such that

$$\liminf_{n\to\infty} \mathbb{P}\bigl\{\bigl|\bigl(\widetilde{X}^n(t),\widetilde{Z}^n(t),\widetilde{Y}^n(t)\bigr)\bigr| \leqslant c(\eta,t)\bigr\} \geqslant 1-\eta.$$

(b) For every $\eta > 0$ and $T > 0$, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\bigl\{w\bigl((\widetilde{X}^n,\widetilde{Z}^n,\widetilde{Y}^n),\delta,T\bigr) \geqslant \eta\bigr\} \leqslant \eta.$$

To verify condition (a), by lemma 8.1, $\widetilde{X}^n$ converges in distribution. Hence, it follows from remark 7.3 in chapter 3 of Ethier and Kurtz [21] that $\{\widetilde{X}^n\}$ satisfies the following compact containment condition: for every $\eta > 0$ and $T > 0$, there is a constant $M_1 > 0$ such that

$$\inf_n \mathbb{P}\bigl\{\bigl|\widetilde{X}^n(t)\bigr| \leqslant M_1,\ 0 \leqslant t \leqslant T\bigr\} \geqslant 1-\eta/2.$$

By assumption (3.4), there exists a constant $M_2 > 0$ such that $\sup_n \mathbb{P}\{|\widetilde{Z}^n(0)| > M_2\} \leqslant \eta/2$. It is easy to check that for each sample path, $(\widetilde{Z}^n,\widetilde{Y}^n)$ is an $(\widetilde{S}^n,R^n)$-regulation of $\widetilde{Z}^n(0)+\widetilde{X}^n$, where

$$\widetilde{\boldsymbol{S}}^n = \bigl\{x \in \mathbb{R}^d\colon\ 0 \leqslant x_i \leqslant b_i^n/\sqrt{n}\ \forall i \in \boldsymbol{I}\bigr\}.$$

Therefore by theorem 4.2 and the continuity of $\widetilde{Y}^n$, there exist constants $\kappa > 0$ and $n_0 > 0$ such that for all $0 \leqslant t_1 < t_2$ and all $n \geqslant n_0$,

$$\operatorname{Osc}\bigl((\widetilde{X}^n,\widetilde{Z}^n,\widetilde{Y}^n),[t_1,t_2]\bigr) \leqslant \kappa \operatorname{Osc}\bigl(\widetilde{X}^n,[t_1,t_2]\bigr). \tag{8.3}$$

Thus, we have, for $n \geqslant n_0$,

$$\begin{aligned}
&\bigl|\bigl(\widetilde{X}^n(t), \widetilde{Z}^n(t), \widetilde{Y}^n(t)\bigr)\bigr| \\
&\quad \leqslant \bigl|\bigl(\widetilde{X}^n(0), \widetilde{Z}^n(0), \widetilde{Y}^n(0)\bigr)\bigr| + \bigl|\bigl(\widetilde{X}^n(t), \widetilde{Z}^n(t), \widetilde{Y}^n(t)\bigr) - \bigl(\widetilde{X}^n(0), \widetilde{Z}^n(0), \widetilde{Y}^n(0)\bigr)\bigr| \\
&\quad \leqslant \bigl|\widetilde{Z}^n(0)\bigr| + \mathrm{Osc}\bigl(\bigl(\widetilde{X}^n, \widetilde{Z}^n, \widetilde{Y}^n\bigr), [0,t]\bigr) \\
&\quad \leqslant \bigl|\widetilde{Z}^n(0)\bigr| + \kappa\, \mathrm{Osc}\bigl(\widetilde{X}^n, [0,t]\bigr) \leqslant \bigl|\widetilde{Z}^n(0)\bigr| + \kappa \sup_{0 \leqslant t \leqslant T} \bigl|\widetilde{X}^n(t)\bigr|.
\end{aligned}$$

Hence, for $n \geqslant n_0$

$$\begin{aligned}
&\mathbb{P}\bigl\{\bigl|\bigl(\widetilde{X}^n(t), \widetilde{Z}^n(t), \widetilde{Y}^n(t)\bigr)\bigr| > M_2 + \kappa M_1 \text{ for some } t \in [0,T]\bigr\} \\
&\quad \leqslant \mathbb{P}\bigl\{\bigl|\widetilde{Z}^n(0)\bigr| > M_2\bigr\} + \mathbb{P}\bigl\{\bigl|\widetilde{X}^n(t)\bigr| > M_1 \text{ for some } t \in [0,T]\bigr\} \leqslant \eta.
\end{aligned}$$

Therefore, $\{(\widetilde{X}^n, \widetilde{Z}^n, \widetilde{Y}^n)\}$ satisfies the containment condition. Thus, condition (a) in corollary 7.4 holds.

To verify condition (b) in corollary 7.4, because $\{\widetilde{X}^n\}$ is relatively compact, for each $\eta > 0$ and $T > 0$, there exists a $\delta > 0$ such that

$$\limsup_{n \to \infty} \mathbb{P}\biggl\{w\bigl(\widetilde{X}_n, \delta, T\bigr) \geqslant \frac{\eta}{\kappa + 1}\biggr\} \leqslant \frac{\eta}{\kappa + 1}.$$

From (8.3), for $n \geqslant n_0$,

$$w\bigl(\bigl(\widetilde{X}_n, \widetilde{Z}^n, \widetilde{Y}^n\bigr), \delta, T\bigr) \leqslant \kappa w\bigl(\widetilde{X}_n, \delta, T\bigr).$$

Therefore, for $n \geqslant n_0$,

$$\begin{aligned}
\mathbb{P}\bigl\{w\bigl(\bigl((\widetilde{X}_n, \widetilde{Z}^n, \widetilde{Y}^n), \delta, T\bigr) \geqslant \eta\bigr\} &\leqslant \mathbb{P}\bigl\{\kappa w\bigl(\widetilde{X}_n, \delta, T\bigr) \geqslant \eta\bigr\} \\
&\leqslant \mathbb{P}\biggl\{w\bigl(\widetilde{X}_n, \delta, T\bigr) \geqslant \frac{\eta}{\kappa + 1}\biggr\} \leqslant \frac{\eta}{\kappa + 1} \leqslant \eta.
\end{aligned}$$

Thus, condition (b) in corollary 7.4 holds. □

Lemma 8.3. Suppose z^n converges to z in $D([0,\infty), \mathbb{R}^d)$, y^n converges to y in $D([0,\infty), \mathbb{R}_+)$ and y is continuous. Assume that for each n, $y^n(\cdot)$ is nondecreasing. Then, for any $f \in C_b(\mathbb{R}^d)$, we have

$$\int_0^t f\bigl(z^n(s)\bigr)\, \mathrm{d}y^n(s) \to \int_0^t f\bigl(z(s)\bigr)\, \mathrm{d}y(s) \quad \text{as } n \to \infty \tag{8.4}$$

uniformly for t in any compact subset of $[0,\infty)$.

Proof. Noting that $z^n \to z$ in $D([0,\infty), \mathbb{R}^d)$, by proposition 3.5.3 and remark 3.5.4 in Ethier and Kurtz [21] or Billingsley [4, p. 112], there exists a sequence $\{\gamma_n\}$

of continuous, strictly increasing functions from $[0,\infty)$ onto $[0,\infty)$ such that, as $n \to \infty$,

$$z^n\big(\gamma_n(t)\big) \to z(t) \text{ u.o.c.} \quad \text{and} \quad \gamma_n(\cdot) \to t \text{ u.o.c.} \tag{8.5}$$

Now, fix $t > 0$ and observe that for each $u \in [0,t]$,

$$\begin{aligned}
&\int_0^u f\big(z^n(s)\big)\,\mathrm{d}y^n(s) - \int_0^u f\big(z(s)\big)\,\mathrm{d}y(s)\\
&\quad = \int_0^{\gamma_n^{-1}(u)} \Big(f\big(z^n\big(\gamma_n(s)\big)\big) - f\big(z(s)\big)\Big)\,\mathrm{d}y^n\big(\gamma_n(s)\big)\\
&\qquad + \int_u^{\gamma_n^{-1}(u)} f\big(z(s)\big)\,\mathrm{d}y^n\big(\gamma_n(s)\big) + \int_0^u f\big(z(s)\big)\,\mathrm{d}\big(y^n(\gamma_n) - y\big)(s).
\end{aligned} \tag{8.6}$$

The first term on the right side of (8.6) is bounded by

$$\max_{0 \leqslant s \leqslant \gamma_n^{-1}(t)} \left| f\big(z^n\big(\gamma_n(s)\big)\big) - f\big(z(s)\big)\right| y^n(t),$$

which converges to zero as $n \to \infty$ uniformly on $u \in [0,t]$ because $f \in C_b(\mathbb{R}^d)$, $y(t)$ is continuous, and $y^n(t) \to y(t)$.

The second term on the right side of (8.6) is dominated by

$$\begin{aligned}
&\|f\|_\infty \sup_{0 \leqslant u \leqslant t} \left| y^n(u) - y^n\big(\gamma_n(u)\big)\right|\\
&\quad \leqslant \|f\|_\infty \Big(\sup_{0 \leqslant u \leqslant t} \left| y^n(u) - y(u)\right| + \sup_{0 \leqslant u \leqslant t} \left| y(u) - y\big(\gamma_n(u)\big)\right|\\
&\qquad + \sup_{0 \leqslant u \leqslant t} \left| y\big(\gamma_n(u)\big) - y^n\big(\gamma_n(u)\big)\right|\Big),
\end{aligned}$$

which converges to zero because $y(t)$ is continuous, and $y^n(t) \to y(t)$ u.o.c.

Finally, we claim that the third term on the right side of (8.6) converges to zero. In fact, since $f(z(\cdot)) \in D([0,\infty),\mathbb{R})$, by theorem 3.5.6, proposition 3.5.3 and remark 3.5.4 of Ethier and Kurtz [21], there is a sequence of step functions $\{g^k(\cdot)\}_{k=1}^\infty$ of the form

$$g^k(\cdot) = \sum_{i=1}^{l_k} g^k\big(t_i^k\big) I_{[t_i^k, t_{i+1}^k)}(\cdot), \tag{8.7}$$

where $0 = t_1^k < t_2^k < \cdots < t_{l_{k+1}}^k < \infty$, $I_{[s,t)}$ is the indicator function on $[s,t)$, and

$$\sup_{0 \leqslant s \leqslant t} \left| f\big(z(s)\big) - g^k(s)\right| \to 0 \quad \text{as } k \to \infty.$$

Notice that

$$\begin{aligned}
&\left|\int_0^u f\bigl(z(s)\bigr)\,\mathrm{d}\bigl(y^n(\gamma_n)-y\bigr)(s)\right|\\
&\quad\leqslant \left|\int_0^u \bigl(f\bigl(z(s)\bigr)-g^k(s)\bigr)\,\mathrm{d}\bigl(y^n(\gamma_n)-y\bigr)(s)\right| + \left|\int_0^u g^k(s)\,\mathrm{d}\bigl(y^n(\gamma_n)-y\bigr)(s)\right|\\
&\quad\leqslant \sup_{0\leqslant s\leqslant t}\bigl|f\bigl(z(s)\bigr)-g^k(s)\bigr|\bigl(y^n(\gamma_n)(t)+y(t)\bigr)\\
&\qquad + \sup_{0\leqslant u\leqslant t}\sum_{i=1}^{l_k}\bigl|g^k\bigl(t_i^k\wedge u\bigr)\bigr|\\
&\qquad \times\bigl|\bigl(y^n(\gamma_n)-y\bigr)\bigl(t_{i+1}^k\wedge u\bigr)-\bigl(y^n(\gamma_n)-y\bigr)\bigl(t_i^k\wedge u\bigr)\bigr|.
\end{aligned}\tag{8.8}$$

Because $y^n(\cdot)\to y(\cdot)$ u.o.c. and y is continuous, for each $t>0$, there exists $M>0$ such that

$$\limsup_{n\to\infty}\sup_{0\leqslant s\leqslant t}\bigl|y^n(s)\bigr|\leqslant M.$$

Letting $n\to\infty$ in (8.8), noticing that for fixed k, the last term of (8.8) converges to zero, we have

$$\limsup_{n\to\infty}\sup_{0\leqslant u\leqslant t}\left|\int_0^u f\bigl(z(s)\bigr)\mathrm{d}\bigl(y^n(\gamma_n)-y\bigr)(s)\right|\leqslant 2M\sup_{0\leqslant s\leqslant t}\bigl|f\bigl(z(s)\bigr)-g^k(s)\bigr|.\tag{8.9}$$

Let $k\to\infty$, we have

$$\limsup_{n\to\infty}\sup_{0\leqslant u\leqslant t}\left|\int_0^u f\bigl(z(s)\bigr)\mathrm{d}\bigl(y^n(\gamma_n)-y\bigr)(s)\right|=0,\tag{8.10}$$

thus proving the lemma. □

Lemma 8.4. For $i\in\boldsymbol{I}$ and any $t\geqslant 0$,
(a)

$$\mathbb{E}\left[\frac{1}{\sqrt{n}}\max_{1\leqslant j\leqslant E_i(nt)+1}u_{ij}\right]\to 0\quad\text{as } n\to\infty.$$

(b)

$$\left\{\frac{1}{\sqrt{n}}\sup_{0\leqslant s\leqslant 1}\bigl|E_i(ns)-ns\bigr|:\ n\geqslant 1\right\}$$

is uniformly integrable.

Proof. Noting that $E_i(t)+1$ is a stopping time for the discrete filtration $\{\mathcal{G}_j\}$ with

$$\mathcal{G}_j=\sigma\{u_{i1},\dots,u_{ij}\},$$

we can write

$$\frac{E_i(nt) - nt}{\sqrt{n}} = \frac{E_i(nt) + 1 - U_i(E_i(nt) + 1)}{\sqrt{n}} - \frac{1}{\sqrt{n}} + \frac{U_i(E_i(nt) + 1) - t}{\sqrt{n}}. \tag{8.11}$$

The first term on the right, denoted by $M_i^n(t)$, is a square integrable martingale with

$$\mathbb{E}\big[M_i^n(t)^2\big] = c_i^a \frac{\mathbb{E}[E_i(nt) + 1]}{n}. \tag{8.12}$$

Since the right-hand side of (8.12) is bounded in n [23, theorem II.5.1], by [21, corollary 2.2.17] the sequence $\mathbb{E}[\sup_{0\leqslant t\leqslant 1} |M_i^n(t)|]^2$ is bounded, hence,

$$\Big\{ \sup_{0\leqslant t\leqslant 1} \big|M_i^n(t)\big|,\ n \geqslant 1\Big\}$$

is uniformly integrable. Using the fact that

$$\sup_{0\leqslant t\leqslant 1} \big|M_i^n(t) - M_i^n(t-)\big| \leqslant 2 \sup_{0\leqslant t\leqslant 1} \big|M_i^n(t)\big|,$$

for $0 \leqslant t \leqslant 1$, the last term on the right of (8.11) (the overshoot of the renewal process) is bounded by

$$\max_{0\leqslant j\leqslant E_i(n)+1} \frac{u_{ij}}{\sqrt{n}} \leqslant 2 \sup_{0\leqslant t\leqslant 1} \big|M_i^n(t)\big| + \frac{1}{\sqrt{n}}.$$

We then have

$$\sup_{0\leqslant t\leqslant 1} \left|\frac{E_i(nt) - nt}{\sqrt{n}}\right| \leqslant 3 \sup_{0\leqslant t\leqslant 1} \big|M_i^n(t)\big| + \frac{2}{\sqrt{n}},$$

and (a) and (b) follow from the uniform integrability of $\{\sup_{0\leqslant t\leqslant 1} |M_i^n(t)|,\ n \geqslant 1\}$. □

Proof of theorem 3.1. By lemma 8.2 the sequence

$$\big\{\big(\widetilde{Z}^n, \widetilde{X}^n, \widetilde{Y}^n\big),\ n \geqslant n_0\big\}$$

is precompact. Therefore,

$$\big\{\big(\widetilde{E}^n, \widetilde{S}^n, \widetilde{Z}^n, \widetilde{X}^n, \widetilde{Y}^n\big),\ n \geqslant n_0\big\}$$

is precompact. Let $(E^*, S^*, Z^*, X^*, Y^*)$ be a weak limit defined on a probability space $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$. That is, there is a sequence $\{n_k\}$ such that as $n_k \to \infty$

$$\big(\widetilde{E}^{n_k}, \widetilde{S}^{n_k}, \widetilde{Z}^{n_k}, \widetilde{X}^{n_k}, \widetilde{Y}^{n_k}\big) \Longrightarrow \big(E^*, S^*, Z^*, X^*, Y^*\big).$$

By lemma 8.1,

$$X^*(t) = E^*(\alpha t) - \big(I - P'\big)S^*(\mu t) \tag{8.13}$$

is a d-dimensional Brownian motion with drift θ and covariance matrix Γ. We will show that Z^*, together with Y^*, is an RBM associated with the Brownian motion X^*.

Because the $(\Gamma, \theta, R, \boldsymbol{S})$-RBM with initial distribution $\mathbb{P}^* Z^*(0)^{-1}$ is unique in distribution (see Dai and Williams [18]), we have

$$\big(\widetilde{Z}^n, \widetilde{X}^n, \widetilde{Y}^n\big) \Longrightarrow \big(Z^*, X^*, Y^*\big),$$

as $n \to \infty$, thus proving the theorem.

To show Z^* is an RBM, notice that

(i) $\widetilde{Z}^n(t) = \widetilde{Z}^n(0) + \widetilde{X}^n(t) + R^n \widetilde{Y}^n(t)$ for all $t \geqslant 0$,

(ii) $0 \leqslant \widetilde{Z}^n_i(t) \leqslant b^n_i/\sqrt{n}$ for all $t \geqslant 0$ and $i = 1, \ldots, d$,

(iii) for each $i = 1, \ldots, 2d$,

  (a) $\widetilde{Y}^n_i(0) = 0$,

  (b) $\widetilde{Y}^n_i$ is nondecreasing,

  (c) for $i = 1, \ldots, d$, $\widetilde{Y}^n_i$ increases only when $\widetilde{Z}^n_i(t) = 0$ and for $i = d+1, \ldots, 2d$, $\widetilde{Y}^n_i$ increases only when $\widetilde{Z}^n_i(t) = b^n_i/\sqrt{n}$.

To show that the limit process (Z^*, X^*, Y^*) satisfies (3.7)–(3.15), we invoke the Skorohod representation theorem [21, theorem 3.1.8]. Therefore, we assume that $\{(\widetilde{Z}^{n_k}, \widetilde{X}^{n_k}, \widetilde{Y}^{n_k}),\ n \geqslant n_0\}$ and (Z^*, X^*, Y^*) are defined on the same probability space $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$ such that $\mathbb{P}^*$-a.s., (i)–(iii) hold and

$$\big(\widetilde{Z}^{n_k}, \widetilde{X}^{n_k}, \widetilde{Y}^{n_k}\big) \to \big(Z^*, X^*, Y^*\big) \text{ u.o.c.} \quad \text{as } n_k \to \infty. \tag{8.14}$$

It follows from (3.4) that $Z^*(0)$ has the same distribution as ξ. Clearly, (3.10) is satisfied with

$$\mathcal{F}^*_t \equiv \sigma\big\{\big(Z^*(s), X^*(s), Y^*(s)\big),\ 0 \leqslant s \leqslant t\big\}.$$

It easy to check that (3.7), (3.8), and (3.11) follow from (i), (ii), (iii)(a) and (iii)(b). Because $\widetilde{Y}^n_{i+d}(t)$ increases only at times t such that $\widetilde{Z}^n_i(t) = b^n_i/\sqrt{n}$, we have for each $T > 0$

$$\int_0^T \left(\frac{b^n_i}{\sqrt{n}} - \widetilde{Z}^n_i(t)\right) \wedge 1 \, \mathrm{d}\widetilde{Y}^n_{i+d}(t) = 0. \tag{8.15}$$

Let

$$f : (b, z) \in \mathbb{R}^2 \to f(b, z) = (b - z) \wedge 1.$$

Clearly, $f \in C_b(\mathbb{R}^2)$. By lemma 8.3 and (8.14),

$$\int_0^T \big(b_i - Z^*_i(t)\big) \wedge 1 \, \mathrm{d}\widetilde{Y}^*_{i+d}(t) = 0, \quad \text{for all } T > 0.$$

Therefore, $Y^*_{i+d}(\cdot)$ increases only at times t such that $Z^*_i(t) = b_i$, showing (3.13). Similarly, we can show that $Y^*_i(\cdot)$ increases only at times t when $Z^*_i(t) = 0$, i.e., (3.12) holds.

It remains to prove (3.15), i.e., $\{X^*(t) - \theta t,\ t \geqslant 0\}$ is an $\{\mathcal{F}_t^*\}$-martingale. It is enough to show that for each $i \in \boldsymbol{I}$, each $r \geqslant 1$, any $0 \leqslant s_1 < s_2 < \cdots < s_r \leqslant s < t$, and any $f_k, g_k, h_k \in C_b(\mathbb{R}^d)$,

$$\mathbb{E}^*\left[\left(X_i^*(t+s) - X_i^*(s) - \theta_i t\right)\prod_{k=1}^{r} f_j\left(X^*(s_j)\right)g_j\left(Y^*(s_j)\right)h_j\left(Z^*(s_j)\right)\right] = 0. \quad (8.16)$$

Let $p, q \in \mathbb{Z}_+^d$ be d-dimensional indexes. Let

$$\widehat{U}_i^n(p_i) = \sum_{k=2}^{p_i+1} \frac{u_{ik} - 1}{\alpha_i^n}, \qquad \widehat{V}_i^n(q_i) = \sum_{k=2}^{q_i+1} \frac{v_{ik} - 1}{\mu_i^n}.$$

Recall the definition of $\mathcal{G}_{p,q}^n$ in (7.1). Because $Z^n(0)$ is assumed to be independent of the interarrival time and service time sequences, it easy to check that

$$\left\{\left(\widehat{U}^n(p), \widehat{V}^n(q)\right),\ \mathcal{G}_{p,q}^n,\ (p,q) \in \mathbb{Z}_+^d \times \mathbb{Z}_+^d\right\}$$

is a multiparameter martingale (see [21, section 2.8] for the definition). Let

$$A^n(t) = E^n\left(F^n(t)\right), \qquad D^n(t) = S^n\left(B^n(t)\right), \qquad \tau^n(t) = \left(A^n(t), D^n(t)\right).$$

By lemma 7.1, for each fixed t, $\tau^n(t)$ is a multidimensional stopping time with respect to the filtration $\{\mathcal{G}_{p,q}^n\}$. Define

$$\mathcal{G}_{\tau^n(t)}^n \equiv \left\{B \in \mathcal{F},\ B \cap \left\{\tau^n(t) \leqslant (p,q)\right\} \in \mathcal{G}_{p,q}^n \text{ for all } (p,q) \in \mathbb{Z}_+^d \times \mathbb{Z}_+^d\right\}.$$

It is clear that $\tau^n(t) \in \mathcal{F}_{\tau^n(t)}^n$. Because $Z^n(0) \in \mathcal{G}_{0,0}^n$, it follows from (5.4) that $Z^n(t) \in \mathcal{G}_{\tau^n(t)}^n$. From (2.1) $Y^n(t) \in \mathcal{G}_{\tau^n(t)}^n$ and from (5.7) $X^n(t) \in \mathcal{G}_{\tau^n(t)}^n$. Let

$$\begin{aligned}\widehat{U}^{n,k}(p) &= \left(\widehat{U}_1^n(p_1 \wedge k), \ldots, \widehat{U}_d^n(p_d \wedge k)\right),\\ \widehat{V}^{n,k}(q) &= \left(\widehat{V}_1^n(q_1 \wedge k), \ldots, \widehat{V}_d^n(q_d \wedge k)\right).\end{aligned}$$

By the multiparameter optional stopping theorem [21, theorem 2.8.7] we have that for each $n \geqslant n_0$ and $k \geqslant 1$,

$$\left\{\left(\widehat{U}^{n,k}\left(A^n(t)\right), \widehat{V}^{n,k}\left(D^n(t)\right)\right), \mathcal{G}_{\tau^n(t)}^n,\ t \geqslant 0\right\}$$

is a martingale, or

$$\left\{\left(\frac{1}{\sqrt{n}}\widehat{U}^{n,k}\left(A^n(nt)\right), \frac{1}{\sqrt{n}}\widehat{V}^{n,k}\left(D^n(nt)\right)\right), \mathcal{G}_{\tau^n(nt)}^n,\ t \geqslant 0\right\}$$

is a martingale. Therefore, for each $n \geqslant n_0$ and $k \geqslant 1$,

$$\mathbb{E}\biggl[\biggl(\frac{1}{\sqrt{n}}\widehat{U}_i^{n,k}\bigl(A_i^n\bigl(n(t+s)\bigr)\bigr) - \frac{1}{\sqrt{n}}\widehat{U}_i^{n,k}\bigl(A_i^n(ns)\bigr)\biggr) \times \prod_{j=1}^{r} f_j\bigl(\widetilde{X}^n(s_j)\bigr)g_j\bigl(\widetilde{Y}^n(s_j)\bigr)h_j\bigl(\widetilde{Z}^n(s_j)\bigr)\biggr] = 0. \tag{8.17}$$

For a fixed n and for each $k \geqslant 1$,

$$\begin{aligned}\bigl|\widehat{U}_i^{n,k}\bigl(A_i^n(ns)\bigr)\bigr| &\leqslant \sum_{j=2}^{(A_i^n(ns)\wedge k)+1} \frac{u_{ij}}{\alpha_i^n} + \frac{A_i^n(ns)\wedge k}{\alpha_i^n} \\ &\leqslant \sum_{j=1}^{A_i^n(ns)+1} \frac{u_{ij}}{\alpha_i^n} + \frac{A_i^n(ns)}{\alpha_i^n} \leqslant \sum_{j=1}^{E_i^n(ns)+1} \frac{u_{ij}}{\alpha_i^n} + \frac{E_i^n(ns)}{\alpha_i^n}.\end{aligned}$$

Letting $k \to \infty$ in (8.17), by [23, theorem III.3.1],

$$\mathbb{E}\Biggl[\sum_{j=1}^{E_i^n(ns)+1} \frac{u_{ij}}{\alpha_i^n} + \frac{E_i^n(ns)}{\alpha_i^n}\Biggr] < \infty,$$

it follows from the dominated convergence theorem that for each $n \geqslant 1$,

$$\mathbb{E}\biggl[\biggl(\frac{1}{\sqrt{n}}\widehat{U}_i^{n}\bigl(A_i^n\bigl(n(t+s)\bigr)\bigr) - \frac{1}{\sqrt{n}}\widehat{U}_i^{n}\bigl(A_i^n(ns)\bigr)\biggr) \times \prod_{j=1}^{r} f_j\bigl(\widetilde{X}^n(s_j)\bigr)g_j\bigl(\widetilde{Y}^n(s_j)\bigr)h_j\bigl(\widetilde{Z}^n(s_j)\bigr)\biggr] = 0. \tag{8.18}$$

$$\begin{aligned}\widehat{U}_i^n\bigl(A_i^n(ns)\bigr) &= \sum_{j=2}^{A_i^n(ns)+1} \frac{u_{ij}}{\alpha_i^n} - \frac{A_i^n(ns)}{\alpha_i^n} = \sum_{j=2}^{E_i^n(F_i^n(ns))+1} \frac{u_{ij}}{\alpha_i^n} - \frac{E_i^n(F_i^n(ns))}{\alpha_i^n} \\ &= \sum_{j=2}^{E_i^n(F_i^n(ns))+1} \frac{u_{ij}}{\alpha_i^n} - F_i^n(ns) + F_i^n(ns) - \frac{E_i^n(F_i^n(ns))}{\alpha_i^n} \\ &= \varepsilon^n - \frac{\widehat{E}_i^n(F_i^n(ns))}{\alpha_i^n},\end{aligned}$$

where

$$\varepsilon^n = \sum_{j=2}^{E_i^n(F_i^n(ns))+1} \frac{u_{ij}}{\alpha_i^n} - F_i^n(ns).$$

Because

$$\begin{aligned}|\varepsilon^n| &\leqslant \frac{u_{i1}}{\alpha_i^n} + \max_{1\leqslant j\leqslant E_i^n(F_i^n(ns))+1} \frac{u_{i,j}}{\alpha_i^n} \leqslant \frac{u_{i1}}{\alpha_i^n} + \max_{1\leqslant j\leqslant E_i^n(ns)+1} \frac{u_{i,j}}{\alpha_i^n}\\ &\leqslant \frac{u_{i1}}{\alpha_i^n} + \max_{1\leqslant j\leqslant E_i(n\alpha_i^n s)+1} \frac{u_{i,j}}{\alpha_i^n},\end{aligned}$$

it follows from part (a) of lemma 8.4 that as $n \to \infty$,

$$\mathbb{E}\left[\frac{1}{\sqrt{n}}|\varepsilon^n|\right] \to 0.$$

Because $\alpha_i^n \to \alpha_i$, by part (b) of lemma 8.4,

$$\left\{\frac{1}{\sqrt{n}}\sup_{0\leqslant t\leqslant s}\bigl|\widehat{E}_i\bigl(\alpha_i^n nt\bigr)\bigr|,\ n \geqslant 1\right\}$$

is uniformly integrable. Notice that

$$\bigl|\widehat{E}_i^n\bigl(F_i^n(ns)\bigr)\bigr| \leqslant \sup_{0\leqslant t\leqslant s}\bigl|\widehat{E}_i^n(nt)\bigr|$$

and, therefore,

$$\left\{\frac{1}{\sqrt{n}}\bigl|\widehat{E}_i^n\bigl(F_i^n(ns)\bigr)\bigr|,\ n \geqslant 1\right\}$$

is uniformly integrable. Because

$$\left(\frac{1}{\sqrt{n_k}}\widehat{E}_i^{n_k}\bigl(F_i^{n_k}(n_k s)\bigr), \widetilde{X}^{n_k}(\cdot), \widetilde{Z}^{n_k}(\cdot), \widetilde{Y}^{n_k}(\cdot)\right) \Longrightarrow \bigl(E_i^*(\alpha_i s), X^*(\cdot), Z^*(\cdot), Y^*(\cdot)\bigr),$$

we have

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{\sqrt{n_k}}\widehat{U}_i^{n_k}\bigl(A_i^{n_k}(n_k s)\bigr)\right)\prod_{j=1}^r f_j\bigl(\widetilde{X}^{n_k}(s_j)\bigr)g_j\bigl(\widetilde{Y}^{n_k}(s_j)\bigr)h_j\bigl(\widetilde{Y}^{n_k}(s_j)\bigr)\right]\\ &\quad\to \mathbb{E}^*\left[-\frac{E_i^*(\alpha_i s)}{\alpha_i}\prod_{j=1}^r f_j\bigl(X^*(s_j)\bigr)g_j\bigl(Y^*(s_j)\bigr)h_j\bigl(Z^*(s_j)\bigr)\right].\end{aligned}$$

Similarly, we can show that

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{\sqrt{n_k}}\widehat{U}_i^{n_k}\bigl(A_i^{n_k}\bigl(n_k(t+s)\bigr)\bigr)\right)\prod_{j=1}^r f_j\bigl(\widetilde{X}^{n_k}(s_j)\bigr)g_j\bigl(\widetilde{Y}^{n_k}(s_j)\bigr)h_j\bigl(\widetilde{Y}^{n_k}(s_j)\bigr)\right]\\ &\quad\to \mathbb{E}^*\left[-\frac{E_i^*\bigl(\alpha_i(t+s)\bigr)}{\alpha_i}\prod_{j=1}^r f_j\bigl(X^*(s_j)\bigr)g_j\bigl(Y^*(s_j)\bigr)h_j\bigl(Z^*(s_j)\bigr)\right].\end{aligned}$$

Therefore, we have

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{n_k}}\widehat{U}_i^{n_k}\bigl(A_i^{n_k}\bigl(n_k(t+s)\bigr)\bigr) - \frac{1}{\sqrt{n_k}}\widehat{U}_i^{n_k}\bigl(A_i^{n_k}(n_k s)\bigr)\right)\right.$$
$$\left.\times\prod_{j=1}^{r} f_j\bigl(\widetilde{X}^{n_k}(s_j)\bigr)g_j\bigl(\widetilde{Y}^{n_k}(s_j)\bigr)h_j\bigl(\widetilde{Y}^{n_k}(s_j)\bigr)\right]$$
$$\to -\frac{1}{\alpha_i}\mathbb{E}^*\left[\bigl(E_i^*(\alpha_i(t+s) - E_i^*(\alpha_i s))\bigr)\prod_{j=1}^{r} f_j\bigl(X^*(s_j)\bigr)g_j\bigl(Y^*(s_j)\bigr)h_j\bigl(Z^*(s_j)\bigr)\right].$$

From (8.18), we have

$$\mathbb{E}^*\left[\bigl(E_i^*\bigl(\alpha_i(t+s)\bigr) - E_i^*(\alpha_i s)\bigr)\prod_{j=1}^{r} f_j\bigl(X^*(s_j)\bigr)g_j\bigl(Y^*(s_j)\bigr)h_j\bigl(Z^*(s_j)\bigr)\right] = 0.$$

By the exact same proof, we have

$$\mathbb{E}^*\left[\bigl(S_i^*\bigl(\mu_i(t+s)\bigr) - S_i^*(\mu_i s)\bigr)\prod_{j=1}^{r} f_j\bigl(X^*(s_j)\bigr)g_j\bigl(Y^*(s_j)\bigr)h_j\bigl(Z^*(s_j)\bigr)\right] = 0.$$

Therefore, (8.16) follows from the fact that

$$X_i^*(t+s) - X_i^*(s) - \theta_i t = E_i^*\bigl(\alpha_i(t+s)\bigr) - E_i^*(\alpha_i s) - \bigl(S_i^*\bigl(\mu_i(t+s)\bigr) - S_i^*(\mu_i s)\bigr). \ \square$$

9. Extensions

Consider the queueing network described in section 2, except that probabilistic routing is allowed. Assume that a customer leaving station $i \in \boldsymbol{I}$ goes to station $j \in \boldsymbol{I}$ with probability P_{ij} or exits the network with probability $1 - \sum_{j\in\boldsymbol{I}} P_{ij}$, independent of all previous history. Assume the network is feedforward, i.e., the stations can be numbered so that $P_{ij} = 0$ for $j \leqslant i$. Furthermore, we assume that each station has at most one predecessor. That is,

$$\sigma(i) \cap \sigma(j) = \emptyset \quad \text{for any } i \neq j,$$

where $\sigma(i) = \{j \in \boldsymbol{I}\colon\ P_{ij} > 0\}$.

For this network, using the techniques developed in this paper, we can show that the heavy traffic limit theorem in theorem 3.1 holds with Γ replaced by the formula

$$\Gamma = \operatorname{diag}\bigl(\alpha_1 c_1^a, \ldots, \alpha_d c_d^a\bigr) + \bigl(I - P'\bigr)\operatorname{diag}\bigl(\mu_1 c_1^s, \ldots, \mu_d c_d^s\bigr)(I - P) + \sum_{j\in\boldsymbol{I}} \mu_j \Gamma^j, \tag{9.1}$$

where

$$\Gamma_{lk}^j = \begin{cases} P_{jl}(1 - P_{jl}) & \text{if } l = k,\\ -P_{jl}P_{jk} & \text{if } l \neq k.\end{cases}$$

See [19, sections 2.2, 4.3] for more discussion on this network.

Consider another modification to the network in section 2, where general probabilistic routing is allowed, but a customer arriving at a full buffer is lost. Therefore, the network is a generalized Jackson network [40] except that a customer arriving at a full buffer station is lost. It can be shown that the heavy traffic limit theorem in theorem 3.1 holds with

$$R = \big(I - P', -I\big),$$

and Γ given in (9.1).

Acknowledgements

We thank Tom Kurtz for providing the proof of lemma 8.4. We thank John Hasenbein, Takis Konstantopoulos, Ruth Williams and two referees for suggesting various improvements to the paper.

References

[1] I. Bardhan and S. Mithal, Heavy-traffic limits for an open network of finite-buffer overflow queues: The single class case, preprint (1993).

[2] A. Bernard and A. El Kharroubi, Régulations déterministes et stochastiques dans le premier “orthant” de $\mathbb{R}^n$, Stochastics Stochastics Rep. 34 (1991) 149–167.

[3] D. Bertsekas and R. Gallagher, *Data Networks* (Prentice-Hall, Englewood Cliffs, NJ, 1992).

[4] P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).

[5] M. Bramson, Convergence to equilibria for fluid models of FIFO queueing networks, Queueing Systems 22 (1996) 5–45.

[6] M. Bramson, Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks, Queueing Systems 23 (1997) 1–26.

[7] M. Bramson, State space collapse with application to heavy traffic limits for multiclass queueing networks, Queueing Systems 30 (1998) 89–148.

[8] A. Brondstred, *An Introduction to Convex Polytopes* (Springer, New York, 1983).

[9] J.A. Buzacott, Automatic transfer lines with buffer stocks, Internat. J. Prod. Res. 5 (1967) 183–200.

[10] H. Chen and H. Zhang, Stability of multiclass queueing networks under FIFO service discipline, Math. Oper. Res. 22 (1997) 691–725.

[11] H. Chen and H. Zhang, Diffusion approximations for multiclass FIFO queueing networks, preprint.

[12] D.W. Cheng, Second order properties in a tandem queue with general blocking, Oper. Res. Lett. 12 (1992) 139–144.

[13] D. Cheng and D.D. Yao, Tandem queues with general blocking: A unified model and comparison results, Discrete Event Dyn. Systems 2 (1993) 207–234.

[14] K.L. Chung and J.R. Williams, *Introduction to Stochastic Integration* (Birkhäuser, Boston, 1983).

[15] J.G. Dai and J.M. Harrison, Steady-state analysis of RBM in a rectangle: Numerical methods and a queueing application, Ann. Appl. Probab. 1 (1991) 16–35.

[16] J.G. Dai and T.G. Kurtz, A multiclass station with Markovian feedback in heavy traffic, Math. Oper. Res. 20 (1995) 721–742.

[17] J.G. Dai, G. Wang and Y. Wang, Nonuniqueness of the Skorohod problem arising from FIFO Kelly type network, private communication (1992).

[18] J.G. Dai and R.J. Williams, Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons, Theory Probab. Appl. 40 (1995) 3–53.

[19] W. Dai, Brownian approximations for queueing networks with finite buffers: modeling, heavy traffic analysis and numerical implementations, Ph.D. thesis, School of Mathematics, Georgia Institute of Technology (1996).

[20] A.I. Elwalid and D. Mitra, Analysis and design of rate-based congestion control of high speed networks, I: Stochastic fluid models, access regulation, Queueing Systems 9 (1991) 29–64.

[21] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence* (Wiley, New York, 1986).

[22] L.M. Graves, *The Theory of Functions of Real Variables* (McGraw-Hill, New York, 1956).

[23] A. Gut, *Stopped Random Walks: Limit Theorems and Applications* (Springer, Berlin, 1988).

[24] J.M. Harrison, Brownian models of queueing networks with heterogeneous customer populations, in: *Proc. of the IMA Workshop on Stochastic Differential Systems* (Springer, Berlin, 1988).

[25] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic I, Adv. in Appl. Probab. 2 (1970) 150–177.

[26] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic II, Adv. in Appl. Probab. 2 (1970) 355–364.

[27] D.P. Johnson, Diffusion approximations for optimal filtering of jump processes and for queueing networks, Ph.D. thesis, University of Wisconsin (1983).

[28] T. Konstantopoulos and J. Walrand, On the ergodicity of networks of $\cdot/GI/1/N$ queues, Adv. in Appl. Probab. 22 (1990) 263–267.

[29] H. Kroner, M. Eberspacher, T.H. Theimer, P.J. Kuhn and U. Briem, Approximate analysis of the end to end delay in ATM networks, in: *Proc. of the IEEE INFOCOM '92*, Florence, Italy (1992) pp. 978–986.

[30] G. Last and A. Brandt, *Marked Point Processes on the Real Line: The Dynamic Approach* (Springer, New York, 1995).

[31] D. Mitra and I. Mitrani, Analysis of a Kanban discipline for cell coordination in production lines: I, Managm. Sci. 36 (1990) 1458–1566.

[32] H. Perros and T. Altiok, Queueing networks with blocking: A bibliography, Performance Evaluation Rev.: ACM Sigmetrics 12 (1984) 8–12.

[33] W.P. Peterson, A heavy traffic limit theorem for networks of queues with multiple customer types, Math. Oper. Res. 16 (1991) 90–118.

[34] M.I. Reiman, Open queueing networks in heavy traffic, Math. Oper. Res. 9 (1984) 441–458.

[35] M.I. Reiman, A multiclass feedback queue in heavy traffic, Adv. in Appl. Probab. 20 (1988) 179–207.

[36] M.I. Reiman and R.J. Williams, A boundary property of semimartingale reflecting Brownian motions, Probab. Theory Related Fields 77 (1988) 87–97 and 80 (1989) 633.

[37] L.M. Taylor and R.J. Williams, Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant, Probab. Theory Related Fields 96 (1993) 283–317.

[38] R.J. Williams, An invariance principle for semimartingale reflecting Brownian motions in an orthant, Queueing Systems 30 (1998) 5–25.

[39] R.J. Williams, Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse, Queueing Systems 30 (1998) 27–88.

[40] D.D. Yao, *Probability Models in Manufacturing Systems*, Springer Series in Operations Research (Springer, Berlin, 1994).

Queueing Systems 42, 33–62, 2002

The Finite Element Method for Computing the Stationary Distribution of an SRBM in a Hypercube with Applications to Finite Buffer Queueing Networks

XINYANG SHEN[*] and HONG CHEN[**]
Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, Canada

J.G. DAI[***]
School of Industrial and Systems Engineering and School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA

WANYANG DAI[****]
Department of Mathematics, Nanjing University, Nanjing, China

Received 23 August 2000; Revised 31 January 2002

Abstract. This paper proposes an algorithm, referred to as BNAfm (Brownian network analyzer with finite element method), for computing the stationary distribution of a semimartingale reflecting Brownian motion (SRBM) in a hypercube. The SRBM serves as an approximate model of queueing networks with finite buffers. Our BNAfm algorithm is based on the finite element method and an extension of a generic algorithm developed by Dai and Harrison [14]. It uses piecewise polynomials to form an approximate subspace of an infinite-dimensional functional space. The BNAfm algorithm is shown to produce good estimates for stationary probabilities, in addition to stationary moments. This is in contrast to the BNAsm algorithm (Brownian network analyzer with spectral method) of Dai and Harrison [14], which uses global polynomials to form the approximate subspace and which sometimes fails to produce meaningful estimates of these stationary probabilities. Extensive computational experiences from our implementation are reported, which may be useful for future numerical research on SRBMs. A three-station tandem network with finite buffers is presented to illustrate the effectiveness of the Brownian approximation model and our BNAfm algorithm.

Keywords: Brownian approximation, heavy traffic, semimartingale reflecting Brownian motion, stationary distribution, performance analysis, queueing networks, finite buffer, finite element method, BNAfm, BNAsm, basic adjoint relationship, numerical algorithm

[*] Supported in part by an NSERC (Canada) research grant and an NSERC (Canada) Postgraduate Scholarship.

[**] Supported in part by research grants from NSERC (Canada), RGC (Hong Kong) and NSFC (China).

[***] Supported in part by NSF grants DMI-9457336 and DMI-9813345, by TLI-AP, a partnership between National University of Singapore and Georgia Institute of Technology, and by a grant from Chinese National Science Foundation.

[****] Supported in part by research grants from Nanjing University and Ministry of Education of China.

1. Introduction

This paper proposes a numerical algorithm for computing the stationary distribution of a semimartingale reflecting Brownian motion (SRBM). The SRBM is a certain diffusion process that lives in a hypercube state space. Such an SRBM often serves as an approximate model for finite buffer queueing networks.

Queueing networks have long been used to model manufacturing systems and communication networks, and have provided a very useful tool for the design and the operations management of these systems. (See, for example, [6,29,31,42].) In modeling and analyzing these systems, one of the fundamental issues is the performance analysis of queueing networks. Despite much effort, exact analysis of queueing networks has been largely limited to exponential networks with infinite buffers. (See, for example, [30,38,42].) Almost all real-world systems modeled by queueing networks have finite buffer capacity. In many applications, buffer constraints are not essential (or are not hard constraints); in this case, analytically simpler queueing networks with infinite buffers have been used. But in some other applications, buffer constraints have an important impact on the performance of the systems and may not be ignored. (See examples in [5,6,42].)

For certain queueing networks with finite buffers, Brownian models can be formulated for approximate analysis of these networks. See, for example, [14,19]. A Brownian model of a three-station tandem network is given in section 6 of this paper. In the Brownian model of a queueing network with finite buffers, an SRBM in a hypercube is used to approximate the queue length process. The data specifying the SRBM can be computed explicitly from certain parameters of the queueing network. The parameters involved are the first and second moments of the interarrival time and service time distributions, and the routing probabilities.

The theoretic foundation for our SRBM is the work of Dai and Williams [18], which provides a necessary and sufficient condition for the existence of an SRBM in a convex polyhedron. For a given SRBM, one would like to compute its certain characteristics. Motivated by queueing network applications, one often focuses on the stationary distribution of an SRBM. Computed quantities from the stationary distribution are used to estimate certain performance measures of the corresponding queueing network. Only in some special cases (see [28]) does the SRBM have an explicit formula for stationary distributions.

In this paper, we propose an algorithm for computing the stationary distribution of an SRBM in a hypercube. In general, we shall use a Brownian network analyzer (BNA) to refer to an algorithm for computing the stationary distribution of an SRBM. (This is motivated by Whitt [40], who uses a queueing network analyzer (QNA) to refer to an algorithm for computing the stationary distribution of a queueing network.) Our algorithm is closely related to a numeric algorithm developed by Dai and Harrison [14] for computing the stationary distribution in a two-dimensional rectangle. Their algorithm consists of two parts: the first part requires a finite dimensional approximation of an infinite-dimensional functional space, and the second part uses a specific sequence of global polynomials to form the approximation subspace. For convenience, we will refer to the first part of their algorithm as a generic algorithm, and the second part as a BNAsm algorithm (a BNA algorithm with a spectral method). (The latter follows a convention in numerical literature [7].) The specific BNA algorithm we propose is based on an extension of the generic algorithm of Dai and Harrison [14], and uses a finite element method or piecewise polynomials to form the approximation subspace. We shall refer to it as a BNAfm algorithm.

The BNAsm algorithm has been shown to often produce accurate estimates of the stationary mean of an SRBM. However, it sometimes fails to produce good estimates for stationary probabilities. Stationary probabilities and tail probabilities are important quantities of an SRBM that can be used to answer some important questions regarding quality of service for the system modeled by the corresponding queueing network. Even in computing the stationary mean of an SRBM, there have been cases where BNAsm fails to provide a meaningful estimate. See case A.1 in table 2 of [17], although we point out that the case is for an SRBM living in a high-dimensional orthant, not a hypercube. Our BNAfm algorithm is shown to produce accurate estimates of the stationary mean as well as the stationary probabilities. (See section 4.4 for more comparisons between the two algorithms.)

Implementing the BNAfm algorithm in arbitrary dimensions has been a long, difficult project. An exploratory implementation was done in [15] by Dai for SRBMs in one and two dimensions. W. Dai [19] implemented a version in his thesis for SRBMs in two and three dimensions with uniform mesh. Finally, Shen [36] implemented a version for SRBMs in arbitrary dimensions with general lattice mesh. His general implementation, in C++ programming language, supersedes all the previous implementations. The numerical results and experiments reported in this paper are from his implementation. In addition to developing the BNAfm algorithm and reporting its successful implementation, we also summarize our numerical experiences from our extensive computations using the implementation. It is hoped that these experiences can guide further numerical research on SRBMs.

Once an approximating subspace is chosen, there is still a choice of which basis to use to represent the subspace. With a fixed subspace, choice of a basis can affect the computational accuracy significantly due to round-off errors in numerical computation. We should point out that the sometimes poor performance of BNAsm in [14] may not be intrinsic to the algorithm. It may be due to the poor choice of basis for global polynomials.

Both the generic and BNAsm algorithms were generalized to an SRBM living in a high-dimensional orthant and simplex in [14,15]. In a companion paper, Chen and Shen [10] extended the BNAfm algorithm to compute the stationary distribution of an SRBM in an orthant. Schwerer [35] proposed to use a linear program to compute stationary moments of an SRBM.

Brownian approximation, a version of diffusion approximation or the functional central limit theorem, has long been used for approximating the queueing network. The SRBMs arise as the limits of certain performance processes of queuing networks with appropriate scaling in time and space under a heavy traffic condition. Most of these limit theorems, known as functional central limit theorems or heavy traffic limit theorems, have been focused on the queueing networks with infinite buffers, where the corresponding SRBMs are defined in a nonnegative orthant. For a survey in this area, readers are referred to [9,11,23,27,32,39,41]. Relatively much less effort has been made on the Brownian approximation for the network with finite buffers. Bardhan and Mithal [3] first attempted to establish such a theorem. Dai and W. Dai [13] established a limit theorem for certain feedforward finite buffer networks that identifies the SRBM in a hypercube as its limit.

As will be discussed in section 4.2, our BNAfm algorithm, like the BNAsm of Dai and Harrison, has the "curse of dimensionality". The complexity of the algorithm grows exponentially in the dimension of the state space. In most Brownian approximation of a queueing network, the dimension corresponds to the number of stations of the queueing network. For a queueing network with a large number of stations, we admit that it may be more efficient to simulate the queueing network itself than to use the Brownian model. On the other hand, for a multiclass queueing network, the network can get "large" by having a large number of job classes but a small number of stations. In such a case, performance analysis based on formulating the Brownian model and solving the stationary density is an attractive alternative to brute force simulation of the queueing network.

The rest of the paper is organized as follows. In the next section, we define the semimartingale reflecting Brownian motion (SRBM) in a hypercube. We also present the basic adjoint relationship that characterizes the stationary density of the SRBM. In section 3, we start with recapitulating the generic algorithm of Dai and Harrison [14] with an extension to multi-dimensional hypercube, and then propose our BNAfm algorithm. In section 4, we report several important issues emerging from our implementation of the algorithm. Some numerical experiments are presented in section 5 to show the accuracy of the BNAfm algorithm. In section 6, we present a three-station tandem network with finite buffers, and show how SRBMs, armed with the BNAfm algorithm, can effectively be used for its performance analysis. We conclude the paper with section 7.

Finally, we introduce some notation to be used in this paper. Let $\Re^k$ denote the k-dimensional Euclidean space, and $\Re^k_+$ denote the nonnegative k-dimensional orthant. For a subset S of $\Re^k$, let $C^2_b(S)$ be the functional space of twice differentiable functions whose first and second order partial derivatives are continuous and bounded on S, and let $\mathcal{B}(\mathcal{S})$ be the set of functions which are Borel measurable.

2. SRBM in a hypercube

Let $K \geqslant 1$ be a fixed integer. A K-dimensional hypercube S is defined as

$$S \equiv \{x \in \Re^K \colon 0 \leqslant x \leqslant b\}, \tag{1}$$

where b is a K-dimensional strictly positive vector. In this section, we define a semimartingale reflecting Brownian motion (SRBM) that lives in the state space S. We then state the basic adjoint relationship that characterizes the stationary distribution of the SRBM (theorem 2.5). The characterization is the starting point for computing the stationary distribution which is the primary quantity that we wish to compute in this paper.

Given a K-dimensional vector θ, a $K \times K$ symmetric and strictly positive definite matrix Γ, and a $K \times 2K$ matrix R, we now define an SRBM associated with the data (θ, Γ, R) on the hypercube state space S. Readers who choose to work with the analytical problem associated with data (S, θ, Γ, R) without going through SRBMs may go directly to theorem 2.5 at the end of this section.

Definition 2.1. For $x \in S$, an (S, θ, Γ, R)-SRBM that starts from x is an $\{\mathcal{F}_t\}$-adapted, K-dimensional process Z defined on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \boldsymbol{P}_x)$ such that

$$Z = X + RY, \tag{2}$$

where

1. Z has continuous paths in S, $\boldsymbol{P}_x$-a.s.,
2. under $\boldsymbol{P}_x$, X is a K-dimensional Brownian motion with drift θ and covariance matrix Γ such that $\{X(t) - \theta t, \mathcal{F}_t, t \geqslant 0\}$ is a martingale, and $X(0) = x$, $\boldsymbol{P}_x$-a.s.,
3. Y is an $\{\mathcal{F}_t\}$-adapted, $2K$-dimensional process such that $\boldsymbol{P}_x$-a.s.,
   (a) $Y(0) = 0$,
   (b) Y is continuous and nondecreasing,
   (c) for $i = 1, \ldots, 2K$, Y_i can increase only when Z is on the face F_i,

and $F_i \equiv \{x \in S\colon x_i = 0\}$ and $F_{K+i} \equiv \{x \in S\colon x_i = b_i\}$ are the ith lower and upper boundary face of the hypercube S, respectively.

In (3c), we mean that, for each $t > 0$, $Z(t) \notin F_i$ implies $Y_i(t - \delta) = Y_i(t + \delta)$ for some $\delta > 0$. This is equivalent to $\int_0^\infty \mathbb{1}_{\{Z(s) \in F_i\}}\, \mathrm{d}Y_i(s) = 0$ for all i. Loosely speaking, an SRBM behaves like a Brownian motion with drift vector θ and covariance matrix Γ in the interior of the hypercube S, with the processes being confined to the hypercube by instantaneous "reflection" (or "pushing") at the boundary, where the direction of "reflection" on the ith face F_i is given by the ith column of R. The parameters θ, Γ, and R are called the *drift vector*, *covariance matrix*, and *reflection matrix* of the SRBM, respectively.

The existence of an SRBM depends on the properties of the reflection matrix R. Dai and Williams [18] provided a sufficient condition on R for the existence of an SRBM in a general polyhedron state space. For convenience, we partition R as $R = (R_1, R_2)$, where both R_1 and R_2 are $K \times K$ matrices formed by the first and the last K columns of R, respectively. To specialize their condition into our case, we introduce the notion of reflection matrix associated with a vertex. Note that our hypercube has 2^K vertexes, and each vertex is given by $\bigcap_{i \in \alpha} F_i \bigcap_{i \in \beta} F_{K+i}$ for a (unique) index set $\alpha \subset \{1, \ldots, K\}$

with $\beta = \{1, \dots, K\} \setminus \alpha$. For each vertex α, the reflection matrix R^α associated with the vertex is the $K \times K$ matrix, given by

$$R^\alpha = (I_\alpha - I_\beta)[R_1 I_\alpha + R_2 I_\beta],$$

where I_α is a $K \times K$ diagonal matrix whose ith component equals one if $i \in \alpha$ and equals zero otherwise, and I_β is similarly defined.

Definition 2.2. A square matrix A is said to be an $\mathcal{S}$ matrix if there is a vector $x \geqslant 0$ such that $Ax > 0$. The matrix A is said to be completely-$\mathcal{S}$ if each principal submatrix of A is an $\mathcal{S}$-matrix.

Definition 2.3. The $K \times 2K$ reflection matrix R is said to satisfy the completely-$\mathcal{S}$ condition if for each vertex α, R^α is a completely-$\mathcal{S}$ matrix.

It follows from propositions 1.1 and 1.2 of [18] that a necessary condition for the existence of the SRBM Z associated with (S, θ, Γ, R), for each initial $x \in S$, is that the reflection matrix R satisfy the completely-$\mathcal{S}$ condition. When R satisfies the completely-$\mathcal{S}$ condition, it follows from theorem 1.3 of [18] that there exist processes (Z, X, Y) defined on a common filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\})$, on which a family of probability measures $\{\boldsymbol{P}_x\}$ is defined, such that for each $x \in S$, under $\boldsymbol{P}_x$, Z is an (S, θ, Γ, R)-SRBM starting from x. Furthermore, Z is a strong Markov process that is Feller continuous.

Definition 2.4. A probability measure π_0 on S is called a stationary distribution for Z if for each bounded Borel measurable function f on S

$$\int_S \boldsymbol{E}_x\big[f\big(Z(t)\big)\big]\,\mathrm{d}\pi_0(x) = \int_S f(x)\,\mathrm{d}\pi_0(x) \quad \text{for all } t \geqslant 0. \tag{3}$$

Here, $\boldsymbol{E}_x$ denotes the expectation under $\boldsymbol{P}_x$.

Using the same argument as in section 7 of [28], one can show that the stationary distribution π_0 is unique and has a density p_0 with respect to Lebesgue measure $\mathrm{d}x$ on S. As stated in the introduction, the primary purpose of this paper is to compute the stationary density p_0. We now provide an analytical characterization for p_0. To this end, for each $k = 1, \dots 2K$, define the measure π_k on boundary face F_k via

$$\pi_k(\cdot) = 2\boldsymbol{E}_{\pi_0}\left[\int_0^1 \mathbb{1}_{\{Z(s)\in\cdot\}}\,\mathrm{d}Y_k(s)\right], \tag{4}$$

where $\boldsymbol{E}_{\pi_0}$ denotes the expectation under probability measure $\boldsymbol{P}_{\pi_0}(\cdot) = \int_S \boldsymbol{P}_x(\cdot)\,\pi_0(\mathrm{d}x)$. It then follows again from the arguments in [28] that π_k has a density p_k with respect to the surface Lebesgue measure $\mathrm{d}\sigma_k$ on F_k. Furthermore, $p_0, p_1, \ldots, p_{2K}$ satisfy the following basic adjoint relationship (BAR):

$$\int_S \big(\mathcal{L} f(x)\, p_0(x)\big)\,\mathrm{d}x + \sum_{k=1}^{2K} \int_{F_k} \big(\mathcal{D}_k f(x)\, p_k(x)\big)\,\mathrm{d}\sigma_k = 0, \quad \forall f \in C_b^2(S), \tag{5}$$

where

$$\mathcal{L} f(x) = \frac{1}{2} \sum_{j,k=1}^{K} \Gamma_{j,k} \frac{\partial^2 f(x)}{\partial x_j \partial x_k} + \sum_{j=1}^{K} \theta_j \frac{\partial f(x)}{\partial x_j}, \tag{6}$$

$$\mathcal{D}_k f(x) = v_k' \nabla f(x), \tag{7}$$

v_k is the kth column of the reflection matrix R, and ∇f is the gradient of f.

The following theorem is a special case of [16], where general polyhedron state space was considered. As before, θ is a K-dimensional vector, Γ is a $K \times K$ symmetric and strictly positive definite matrix, and R is a $K \times 2K$ matrix.

Theorem 2.5. Assume that R satisfies the completely-$\mathcal{S}$ condition in definition 2.3. There exists a unique nonnegative function $p = (p_0, p_1, \ldots, p_{2K})$ with $\int_S p_0(x)\,\mathrm{d}x = 1$ and $\int_{F_k} p_k(x)\,\mathrm{d}\sigma_k < \infty$ for $k = 1, \ldots, 2K$ that satisfies the basic adjoint relationship (5). Furthermore, $\pi_0(\cdot) = \int_{\cdot} p_0(x)\,\mathrm{d}x$ is the stationary distribution of the SRBM Z associated with data (S, θ, Γ, R), and $\pi_k(\cdot) = \int_{\cdot} p_k(x)\,\mathrm{d}\sigma_k$ is the measure on F_k defined in (4).

Theorem 2.5 provides an analytical characterization of the stationary density of an SRBM. One would hope to find an analytical solution from the characterization. This has been possible only for some very special cases. Harrison et al. [24] derived an analytical expression for a two-dimensional driftless SRBM. Harrison and Williams [28] identified a certain skew symmetry condition for an SRBM to have a product-form stationary distribution. In general, a numerical algorithm is needed to compute the stationary distribution. As we will see in the next section, the characterization provides a starting point for a generic algorithm for computing the stationary density p.

We now define some quantities related to the stationary distribution of an SRBM. For $i = 1, \ldots, K$ and $k = 1, \ldots, 2K$, define

$$q_i = \int_S x_i\, p_0(x)\,\mathrm{d}x, \tag{8}$$

$$\delta_k = \int_{F_k} p_k(x)\,\mathrm{d}\sigma_k. \tag{9}$$

The vector $q = (q_1, \ldots, q_K)$ is called the stationary mean. It is also the long-run average value of Z. The quantity δ_k represents the long-run average amount of pushing per unit of time needed on boundary F_k in order to keep the SRBM Z inside the state space S. These quantities, along with stationary probabilities, are of interest in the queueing network applications.

3. The BNAfm algorithm

In this section, we develop the BNAfm algorithm for computing the stationary density p of an SRBM. Dai and Harrison [14] developed a BNAsm algorithm for computing the stationary distribution of an SRBM in a two-dimensional rectangle. Both their BNAsm and our BNAfm algorithms are specialized versions of a generic algorithm, which involves a finite dimensional approximation of an infinite-dimensional functional space. It is in the schemes of approximations that BNAsm and BNAfm differ. In BNAsm, global polynomials are used to form approximating subspaces, whereas in our BNAfm algorithm, piecewise polynomials are used. A piecewise polynomial is defined through a partition of state space; within each subdomain of the partition it is a polynomial. A global polynomial is one defined on the entire state space. The spectral algorithm achieves its accuracy by increasing the maximum degree of polynomials, whereas the BNAfm algorithm achieves its accuracy by refining the partition of the state space.

Pros and cons of both the spectral method and the finite element method in many problem domains, notably in fluid dynamics, are well documented; see, for example, [4,7]. As it was discussed in the introduction, the BNAsm of Dai and Harrison [14] generally produces a good estimate of the stationary mean of an SRBM. However, it sometimes produces poor estimates of stationary probabilities. As will be shown in section 5, our BNAfm algorithm produces accurate estimates for stationary probabilities as well.

In the remainder of this section, we first recapitulate the generic algorithm of Dai and Harrison [14] with an extension to a multi-dimensional hypercube. We also extend their framework by allowing approximating functions not necessarily C^2 smooth. Such extension is essential when we propose our BNAfm algorithm in section 3.2.

3.1. The generic algorithm

3.1.1. Functional space $L^2(S)$

To facilitate the description of the generic algorithm, we adopt some new notation to present the basic adjoint relationship (5) in a compact form.

First we define a linear space of functions:

$$L^2(S) \equiv \left\{ g = (g_0, g_1, \dots, g_{2K}) \in \mathcal{B}(S) \times \mathcal{B}(F_1) \times \cdots \times \mathcal{B}(F_{2K}): \int_S |g_0|^2 \, \mathrm{d}x + \sum_{k=1}^{2K} \int_{F_k} |g_k|^2 \, \mathrm{d}\sigma_k < \infty \right\}.$$

The space $L^2(S)$ is a tensor product of the L^2 space in the interior and the L^2 spaces on boundaries. For $u, v \in \mathcal{B}(S) \times \mathcal{B}(F_1) \times \cdots \times \mathcal{B}(F_{2K})$, define

$$(u, v) \equiv \int_S u_0 v_0 \, \mathrm{d}x + \sum_{k=1}^{2K} \int_{F_k} u_k v_k \, \mathrm{d}\sigma_k$$

whenever the right side is well defined. When $u, v \in L^2(S)$, (u, v) defines a proper inner product on $L^2(S)$. The norm of a function $u \in L^2(S)$ is defined as a nonnegative real number $\|u\|$ given by

$$\|u\| = \sqrt{(u, u)}. \tag{10}$$

For two functions $u, v \in L^2(S)$, we say that u and v are orthogonal in $L^2(S)$ if $(u, v) = 0$. With new notation, basic adjoint relationship (5) can be rewritten as

$$(\mathcal{A}f, p) = 0 \quad \text{for all } f \in C_b^2(S), \tag{11}$$

where $p = (p_0, p_1, \dots, p_{2K})$ and $\mathcal{A}f = (\mathcal{L}f, \mathcal{D}_1 f, \dots, \mathcal{D}_{2K} f)$.

3.1.2. The least square problem

Since the hypercube S is compact, it is easy to verify that $\mathcal{A}f \in L^2(S)$ for each $f \in C_b^2(S)$. Thus, we can define

$$H = \text{closure of}\{\mathcal{A}f\colon\ f \in C_b^2(S)\},$$

where the closure is taken with respect to norm (10) in $L^2(S)$. If we assume that the unknown density p is in $L^2(S)$, then (11) implies that p is orthogonal to $\mathcal{A}f$ for all $f \in C_b^2(S)$, and thus for all $f \in H$. In other words, if we assume that $p \in L^2(S)$, then that p satisfies the basic adjoint relationship (5) is equivalent to $p \in H^\perp$, where $H^\perp$ denotes the orthogonal space of H in $L^2(S)$.

Let us assume for the moment that the unknown density function p is in $L^2(S)$. For any $h^0 \notin H$, $h^0 - \bar{h}^0 \in H^\perp$, where $\bar{h}^0$ is the projection of h^0 onto H or

$$\bar{h}^0 = \arg\min_{h \in H} \|h^0 - h\|^2.$$

Thus, $h^0 - \bar{h}^0$, in place of p, satisfies the basic adjoint relationship (11). If the function $h^0 - \bar{h}^0$ does not change sign, it follows from theorem 2.5 that

$$p = \kappa\bigl(h^0 - \bar{h}^0\bigr), \tag{12}$$

where κ is a constant such that the integral of p_0 on S equals one. The question of whether function $h^0 - \bar{h}^0$ changes sign remains an open research problem. It was conjectured by Dai and Harrison [14] that the function does not change sign. We state their conjecture, adapted to the high-dimensional hypercube, in the following.

Conjecture 3.1. Suppose that p_0 is an integrable Borel function in S and p_k, $k = 1, \dots, 2K$ are finite Borel measures on $F_1, \dots, F_{2K}$, respectively. If they jointly satisfy the basic adjoint relationship (5), then p_0 does not change sign in S.

Supporting the numerical experiences of Dai and Harrison [14], we found that the function $h^0 - \bar{h}^0$ does not change sign in all our numerical experiments.

For all numerical examples shown in this paper, we choose $h^0 = (1, 0, \dots, 0) \in L^2(S)$. If we assume that p is in L^2, then

$$\int_S \big(h^0 \cdot p\big)\, \mathrm{d}\lambda = 1;$$

this immediately implies $h^0 \notin H$.

At several points in this section, we have made the assumption that $p \in L^2(S)$. Unfortunately, this assumption does not hold for some (S, θ, Γ, R)-SRBMs. See, for example, [24]. When $p \notin L^2(S)$, the key relation (12) fails to hold. However, the algorithm to estimate p proposed later in this section remains valid. (Also see the example in section 5.1.)

3.1.3. Galerkin approximations

Let us again assume that $p \in L^2(S)$ and fix $h^0 = (1, 0, \dots, 0)$. To find p using equation (12), one needs to compute $\bar{h}^0$, i.e., the projection of h^0 onto H. The space H is linear and infinite-dimensional. (By infinite dimensionality of H, we mean that it is necessary to have infinite many functions to form a basis for the space.) Solving the least square problem exactly in an infinite-dimensional space is in general impossible. Instead we seek an approximate solution to (11) by using a finite-dimensional subspace H_n to approximate the space H. This is known as Galerkin approximation in numerical analysis (cf. [4]).

Suppose that we have a sequence of finite-dimensional subspaces $\{H_n\}$ that satisfies $H_n \to H$ in $L^2(S)$ as $n \to \infty$. (By $H_n \to H$ we mean that, for any $h \in H$, there exists a sequence $\{h_n\}$ with $h_n \in H_n$ such that $\|h_n - h\| \to 0$ as $n \to \infty$.) Let

$$h^n = \arg\min_{h \in H_n} \big\|h^0 - h\big\|^2.$$

Since $H_n \to H$, we have $\|h^n - \bar{h}^0\| \to 0$, as $n \to \infty$. Let

$$w^n(x) = \kappa^n\big[h^0(x) - h^n(x)\big], \tag{13}$$

where κ^n is a normalizing constant that makes the integral of w_0^n on S equal one. Dai and Harrison [14] proposed to use w^n to approximate the stationary density p. Indeed, when $p \in L^2(S)$, it was proved that

$$\big\|w^n - p\big\| \to 0 \quad \text{as } n \to \infty, \tag{14}$$

assuming that conjecture 3.1 holds. When $p \notin L^2(S)$, convergence (14) in L^2 cannot be expected. However, w^n in (13) is still well defined. Dai and Harrison [14] conjectured that w^n converges to p in a certain weaker sense.

As in [14], our choice of finite-dimensional subspace H_n will be of the form

$$H_n = \{\mathcal{A} f\colon\ f \in C_n\} \tag{15}$$

for some finite-dimensional space C_n. However, there is an important difference. In [14], C_n was chosen as a subspace of $C_b^2(S)$, whereas in the current exposition we do not make such restriction. For a function f that is not in C^2, the operator $\mathcal{A}f$ is undefined in the conventional sense because the second order derivatives of f do not exist at some point. In such cases, $\mathcal{A}f$ in (15) will be interpreted through general derivatives as described in [34].

To introduce the general derivatives, let us define the norm $\|\cdot\|_{H^2}$ via

$$\|f\|_{H^2}^2 = \max_{1\leqslant i,j\leqslant K}\int_S\left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)^2 \mathrm{d}x + \max_{1\leqslant i\leqslant K}\int_S\left(\frac{\partial f}{\partial x_i}\right)^2 \mathrm{d}x + \int_S f^2\,\mathrm{d}x$$
$$+ \max_{1\leqslant i,j\leqslant k}\int_{F_i}\left(\frac{\partial f}{\partial x_i}\right)^2 \mathrm{d}\sigma_j + \max_{1\leqslant i\leqslant K}\int_{F_i} f^2\,\mathrm{d}\sigma_i$$

for $f \in C_b^2(S)$. One can check that there exists a constant $\kappa_1 > 0$ such that

$$\|\mathcal{A}f\| \leqslant \kappa_1 \|f\|_{H^2} \tag{16}$$

for any $f \in C_b^2(S)$. We use $\overline{C}_b^2(S)$ to denote the closure of $C_b^2(S)$ under the norm $\|\cdot\|_{H^2}$. A standard procedure can be used to define the first-order and second-order derivatives for each $f \in \overline{C}_b^2(S)$. Thus, the operator $\mathcal{A}f$ can be extended to $f \in \overline{C}_b^2(S)$. The inequality (16) can be extended for any $f \in \overline{C}_b^2(S)$.

Suppose that one is given a sequence of finite-dimensional subspaces $\{C_n\}$ of $\overline{C}_b^2(S)$ with $C_n \to \overline{C}_b^2(S)$ in the sense that for every $f \in \overline{C}_b^2(S)$, one can find a sequence $\{f_n\}$ with $f_n \in C_n$ such that $\|f - f_n\|_{H^2} \to 0$ as $n \to \infty$. One can then verify that $H_n \to H$ via (16).

To numerically compute h^n, let f_i^n, $i = 1, \ldots, N_n$, be a finite set of linearly independent basis functions of C_n, where N_n is the dimension of subspace C_n. Then, we can express h^n as

$$h^n = \sum_{i=1}^{N_n} u_i \mathcal{A} f_i^n \tag{17}$$

for some scalars $\{u_i\}$. To find the coefficients $\{u_i\}$, observing that $(h^0 - h^n, \mathcal{A}f_i^n) = 0$ for $i = 1, \ldots, N_n$, we obtain the following linear equations:

$$Au = y, \tag{18}$$

where

$$A_{i,j} = \big(\mathcal{A}f_i^n, \mathcal{A}f_j^n\big), \qquad u = (u_1, \ldots, u_{N_n})', \qquad y = \big(\big(h^0, \mathcal{A}f_1^n\big), \ldots, \big(h^0, \mathcal{A}f_{N_n}^n\big)\big)'. \tag{19}$$

The matrix A is symmetric and semi-positive definite. By deleting some redundant basis functions if necessary, we can and will assume that the matrix A is positive definite. Thus there exists a unique solution to the linear system of equations (18).

To summarize the generic algorithm, let C_n be a given finite-dimensional subspace of $\overline{C}_b^2(S)$. First solve u from the system of linear equations (18) with coefficients computed via formulas in (19). Then form projection h_n using u via (17). Finally, construct function w^n via (13). The resulting function w_n is proposed to be an estimate of the unknown density p.

Each choice of C_n, and consequently H_n, yields an approximation w^n of p. Even for a fixed C_n, different choice of a basis produces a different set of coefficients A and y in (19). Because of numerical round-off error, the resulting h^n, and hence w^n, depends on the choice of basis. In the next section, we propose to use the finite element method to generate the approximate sequence $\{C_n\}$ for which a natural choice of basis exists.

3.2. The BNAfm algorithm

In this section, we construct a sequence of functional subspaces C_n using the finite element method (FEM). The resulting algorithm to compute w^n is called the BNAfm algorithm. The BNAfm algorithm differs significantly from BNAsm used in [14]. As evidenced in section 5, our BNAfm algorithm is able to produce accurate approximations of stationary probabilities.

A mesh is a partition of the state space into a finite number of subdomains called finite elements. Since the domain S is a hypercube, it is natural to use lattice mesh to divide the domain S, where each finite element is again a hypercube. The lattices are allowed to be non-uniform so that we can choose the sizes of lattices freely. Each corner of a finite element is called a node. Figure 1 shows, for example, the domain of a two-dimensional hypercube (rectangle) that is partitioned into 8 × 6 elements with 9 × 7 nodes.

Let $x = (x_1, \dots, x_K)$ denote a free variable in S. For every dimension $j = 1, \dots, K$, we divide interval $[0, b_j]$ into n_j subintervals. Let $y_j^0 = 0 < y_j^1 < \cdots < y_j^{n_j} = b_j$ be the partition points in dimension j. We have $\prod_{j=1}^K (n_j + 1)$ nodes with $\prod_{j=1}^K n_j$ finite elements. The corresponding mesh is denoted as $n_1 \times n_2 \times \cdots \times n_K$. We use Δ to denote a generic mesh. Also, we label nodes in such a way that node

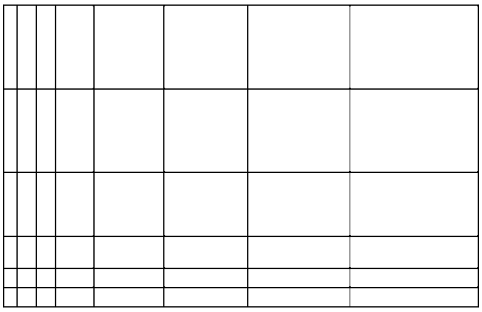

Figure 1. A finite element mesh of a two-dimensional hypercube state space.

$(i_1, \ldots, i_K)$ corresponds to spatial coordinate $(y_1^{i_1}, \ldots, y_K^{i_K})$. For future reference, we define

$$h_j^k = y_j^{k+1} - y_j^k, \quad k = 0, \ldots, n_j - 1 \text{ and } j = 1, \ldots, K,$$

and $\|\Delta\| = \max_{k,j} h_j^k$.

For each mesh Δ, we now construct the finite dimensional space C_Δ. The corresponding H_Δ is constructed via (15). Each function f in C_Δ is a polynomial when it is restricted in each finite element. It is C^2 in the interior of each element and is C^1 globally. With these requirements, we use third order Hermit functions to construct the basis for the subspace C_Δ. See [8] for some basic properties of Hermit functions and other possibilities for constructing bases.

The one-dimensional Hermit basis functions over interval $[-1, 1]$ are

$$\begin{aligned} \phi(x) &= \big(|x| - 1\big)^2\big(2|x| + 1\big), && \text{for } -1 \leqslant x \leqslant 1, \\ \psi(x) &= x\big(|x| - 1\big)^2, && \text{for } -1 \leqslant x \leqslant 1. \end{aligned}$$

For an interval $[y_j^{k-1}, y_j^{k+1}]$ in the jth dimension, define

$$\phi_k(x_j) = \begin{cases} \phi\left(\dfrac{x_j - y_j^k}{h_j^{k-1}}\right) & \text{if } x_j \in \left[y_j^{k-1}, y_j^k\right] \text{ and } k > 0, \\ \phi\left(\dfrac{x_j - y_j^k}{h_j^k}\right) & \text{if } x_j \in \left[y_j^k, y_j^{k+1}\right] \text{ and } k < n_j, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\psi_k(x_j) = \begin{cases} h_j^{k-1}\psi\left(\dfrac{x_j - y_j^k}{h_j^{k-1}}\right) & \text{if } x_j \in \left[y_j^{k-1}, y_j^k\right] \text{ and } k > 0, \\ h_j^k\psi\left(\dfrac{x_j - y_j^k}{h_j^k}\right) & \text{if } x_j \in \left[y_j^k, y_j^{k+1}\right] \text{ and } k < n_j, \\ 0 & \text{otherwise.} \end{cases}$$

Now by using tensor product, we are able to construct tensor-product Hermit basis functions for each node in high dimensions. At node $(i_1, \ldots, i_K)$, the basis functions are of the form

$$f_{i_1,\ldots,i_K,r_1,\ldots,r_K}(x_1, \ldots, x_K) = \prod_{j=1}^{K} g_{i_j,r_j}(x_j),$$

where r_j is 0 or 1, and

$$g_{i_j,r_j}(x_j) = \begin{cases} \phi_{i_j}(x_j), & \text{if } r_j = 0, \\ \psi_{i_j}(x_j), & \text{if } r_j = 1. \end{cases} \tag{20}$$

Each node has 2^K tensor-product basis functions. Hence, we have a total of $n = 2^K \prod_{i=1}^{K}(n_i + 1)$ basis functions. Furthermore, for ease of programming, we re-index these basis functions as

$$f_i(x_1, \ldots, x_K) = f_{i_1,\ldots,i_K,r_1,\ldots,r_K}(x_1, \ldots, x_K), \tag{21}$$

where

$$i = 2^K \sum_{k=1}^{K} i_k \prod_{i=1}^{k-1}(n_k + 1) + \sum_{k=1}^{K} 2^{k-1} r_k. \tag{22}$$

Now we have completed the construction of finite-dimensional subspaces C_Δ. One can check that $C_\Delta \subset \overline{C}_b^2(S)$. The following theorem is needed to justify the use of the BNAfm algorithm.

Theorem 3.2. As $\|\Delta\| \to 0$,

$$C_\Delta \to \overline{C}_b^2(S)$$

in the $\|\cdot\|_{H^2}$ norm.

Proof. Let $f \in \overline{C}_b^2(S)$ be fixed. For any $\varepsilon > 0$, we would like to show the following assertion: there exists $\delta > 0$ such that for any partition Δ with $\|\Delta\| < \delta$,

$$\|g - f\|_{H^2} < \varepsilon \tag{23}$$

for some $g \in C_\Delta$.

By the definition of the Sobolev space $\overline{C}_b^2(S)$, it is enough to prove the assertion for $f \in C_b^2(S)$. It follows from proposition 7.1 in the appendix of [22] that for any $\varepsilon > 0$ there exists a polynomial f_1 such that $\|f_1 - f\|_{H^2} < \varepsilon$. Thus, it is enough to prove the assertion for a polynomial f.

For each partition Δ, let g be the finite element interpolation of f. Since any polynomial function is C^4 smooth, the theorem follows from the following interpolation error estimate in theorem 6.6 of [34]:

$$\|f - g\|_{H^2} \leqslant \kappa \max_{x\in S} \max_{0\leqslant|\alpha|\leqslant 4} \left| \frac{\partial^\alpha f(x)}{\partial x_1^{\alpha_1} \ldots \partial x_K^{\alpha_K}} \right| \|\Delta\|^2,$$

where κ is a constant independent of Δ and f, $\alpha = (\alpha_1, \ldots, \alpha_K)$, and $|\alpha| = \sum_k \alpha_k$. □

The implementation of the BNAfm algorithm requires us to solve the system of linear equations (18) with matrix A and vector y constructed as in (19). The computation of A_{ij} and y_i can be quite tedious. Explicit formulas for their computation were given in (4.11), (4.12), and section 5.4.2 of [19] when the mesh is uniform. Extension to non-uniform mesh is provided in [36].

4. Computational issues of the BNAfm algorithm

We have implemented the BNAfm algorithm in a software package using the C++ programming language. The software runs in both Linux and Sun Solaris operating systems. Although the algorithm itself is easy to understand, it is a big challenge to program the algorithm because of the complexity of BNAfm implementation. In this section, we discuss several important issues emerging from our implementation. They are very critical to the success of applying our BNAfm algorithm to solve practical problems. Some of challenges such as the curse of dimensionality apply to other algorithms as well.

4.1. Solving linear systems of equations

Recall that the BNAfm algorithm uses the subspace C_{Δ} constructed in section 3.2 for a given mesh Δ. The total number of basis functions is

$$n = 2^K \prod_{j=1}^{K} (n_j + 1), \tag{24}$$

where K is the dimension of the state space S and n_j is the number of partition points in the jth dimension. To obtain a numerical estimate of the density function p, we must solve the system of linear equations (18), $Au = y$, where the $n \times n$ matrix A and the n-vector y are given in (19). The most computationally expensive part of the BNAfm algorithm is to solve the linear system of equations (18).

In general, there are two types of methods to solve a system of linear equations: direct methods and iterative methods. A direct method would yield an exact solution in a finite number of steps if all calculations were exact (without round-off error). An iterative computation ends when a solution with a prescribed precision is found. There is no prior knowledge of the number of steps needed in an iterative method. Because of the round-off error, there is no guarantee that the iterative method will converge at all. There has been a huge literature in studying the pros and cons of both methods. Whether one method dominates the other is often problem-specific, and depends on fine tuning such as pivoting and preconditioning that is performed.

In the software, we have implemented both the iterative methods and direct methods. Users can experiment with both methods and choose a better one depending on a specific problem when they run the software. As mentioned above, both of these methods have their own advantages and disadvantages. Interested readers are referred to [36] for more details.

4.2. Computational complexity

The size of matrix A is $n \times n$. Because of the sparseness of matrix A, it may take $\mathrm{O}(n)$ calculations to generate matrix A. But the number of arithmetic operations needed to solve the linear system is $\mathrm{O}(n^3)$ via either LU factorization or Gaussian elimination. For example, if we set $n_i = 5$ for $i = 1, \ldots, K$, then $n = \mathrm{O}(12^K)$. Thus, the computational complexity increases exponentially with the number of dimensions K. In other words, the computing time needed may increase exponentially as the dimension of the problem increases. For example, to solve a 3-dimensional problem with a $4 \times 4 \times 4$ mesh, it takes our software about 9 seconds to obtain an estimate on a computer. But for a 4-dimensional problem with a $4 \times 4 \times 4 \times 4$ mesh, it takes our software more than 24 minutes to obtain an estimate with similar accuracy on the same computer.

4.3. Mesh selection

As motivation, consider a special case of one-dimensional (S, θ, σ^2, R)-SRBM, where $S = [0, b]$ and $R = (1, -1)$. Such an SRBM is also called a two-sided regulated Brownian motion by Harrison [25]. It is known that the stationary density is given by

$$p(x) = \frac{(2\theta/\sigma^2)\mathrm{e}^{2\theta x/\sigma^2}}{\mathrm{e}^{2\theta b/\sigma^2} - 1} \quad \text{for } x \in [0, b],$$

if $\theta \neq 0$, and $p(x) = 1/b$ for $x \in [0, b]$ if $\theta = 0$. (See, for example, [25].) When $\theta < 0$, it is clear that the peak of the derivative of the density is at $x = 0$. In this case, intuitively the numerical algorithm would do better by selecting mesh with smaller subintervals near the origin and (relatively) larger subintervals near the upper bound b. Similarly, when $\theta > 0$, the smaller subintervals would be preferred near the upper bound b for mesh selection. For the boundary case $\theta = 0$, a uniform mesh would be the best. This is indeed the case with the actual implementation of our numerical algorithm.

Unfortunately, determining where the density makes the quick changes itself is a difficult problem. For a driftless ($\theta = 0$) SRBM in a two-dimensional rectangle, Harrison et al. [24] have a conformal mapping representation of the stationary density. In particular, they were able to explicitly identify which corner has a singular pole. Prior information on the location of singularities can be used to build a more refined mesh.

4.4. Ill-conditioned system matrix

In using our computed w^n to approximate the stationary density p, there are two sources of error. The first source is due to the fact that C_Δ is an approximate of $\overline{C}_b^2(S)$. Such an error is called the approximation error. Even when the computation of w^n can be carried out using infinite precision, this error exists. It decreases when the mesh gets finer. The other source is from the numerical round-off error in computing w^n once an approximate subspace C_Δ is given. Round-off error occurs because only finite precision arithmetic is carried on a computer.

Our numerical computation of w^n consists primarily of two parts: the system generation, i.e., calculating coefficient matrix A, and system solution, i.e., solving linear equations. There can be some round-off errors in the calculation of A. But significantly more round-off errors occur in computing the solution to the large linear system (18), $Au = y$. The accuracy of u depends on the property of A. If A is nearly singular, the solution u is extremely sensitive to small changes in the coefficient matrix A and the right-hand side y. In this case, A or the system is said to be ill-conditioned.

The degree of ill-conditioning of linear systems is measured by the *condition number* of matrix A. The larger the condition number is, the worse-conditioned the system is. The condition number can be determined using the extreme eigenvalues of A. The formal mathematical definition of the condition number is

$$\mathrm{Cond}(A) = \|A\| \cdot \left\|A^{-1}\right\|,$$

where $\|\cdot\|$ is the usual matrix norm. Estimating the condition number of A is not an easy task since it involves obtaining the inverse of A, which takes much more effort than solving the linear system directly.

As the mesh is refined, the size of the system increases, and so does the condition number of the system as we have observed in our numerical experiments. From some experiments we performed, we note that as mesh is refined, the system becomes progressively more ill-conditioned, and the round-off error increases. At some point, the round-off error can completely dominate the approximation error. In such cases, further refining the mesh actually decreases the quality of approximation w^n. We note that in running the current implementation of the BNAsm algorithm of [15], we sometimes observe that their algorithm fails to produce positive numbers when the maximum degree of polynomials used is as small as 8. In such cases, we believe that the round-off error dominates the approximation error even when a moderate accuracy of the final estimate is attempted. In all of our cases, the final estimate degrades only after it reaches a high level degree of accuracy.

There are several other factors that affect the conditioning number in our BNAfm algorithm. The uniform or non-uniform mesh has an effect, as does the basis function chosen. We have used third order Hermite functions. Other orders or hybrid polynomials are possible. See, for example, [8].

Currently, the entry $A_{ij} = (\mathcal{A}f_i, \mathcal{A}f_j)$, where $\mathcal{A}$ involves second order derivatives. Such construction of A follows naturally from the current form of the basic adjoint relationship (5) that characterizes the stationary density. The condition number for such A is several orders of magnitude larger than the one for a matrix formed by (f_i, f_j). See p. 197 of [8] for a similar observation. If one can find an alternative characterization of the stationary density, for example, by carrying out integration by parts once in the basic adjoint relationship (5), one may be able to formulate a system matrix that has a much smaller condition number. Such an investigation is a possible future research direction.

4.5. Scaling

For an (S, θ, Γ, R)-SRBM, our computational experiences show that the proper scale of the data (S, θ, Γ, R) has a significant effect on the accuracy and efficiency of our numerical approximation of the stationary density. The fact that the data can be scaled is based on the following proposition whose proof readily follows from (3), definition 2.1, and theorem 1.3 of [18]. Dai and Harrison [15] have a similar proposition for SRBM in an orthant.

Proposition 4.1. Suppose that Z is a K-dimensional SRBM with data (S, θ, Γ, R), and that Z has a stationary distribution π with mean vector m. Let S be the hypercube as defined in (1), D be a positive diagonal matrix, and α be a positive scalar. The new process Z^* defined by

$$Z^*(t) = DZ(\alpha t), \tag{25}$$

is also an SRBM with data $(S^*, \theta^*, \Gamma^*, R^*)$, where

$$\theta^* = \alpha D\theta, \qquad \Gamma^* = \alpha D\Gamma D, \qquad R^* = DR, \tag{26}$$

and

$$S^* \equiv \left\{x \in \Re^K \colon 0 \leqslant x \leqslant Db\right\}.$$

Moreover, Z^* has a stationary distribution π^* with a finite mean vector m^*; they are related to π and m via

$$\pi^*(x) = \pi\left(D^{-1}x\right), \tag{27}$$

$$m^* = Dm. \tag{28}$$

To illustrate the scaling effects, we consider a two-dimensional SRBM example which has a product form stationary density function. The data associated with this SRBM are

$$R = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \end{pmatrix},$$

$\theta' = (10, -10)$, $\Gamma = I$, and $S = [0, 1] \times [0, 1]$. As in chapter 2 of [12], one can check that the data satisfy a skew symmetry condition of [28]. Thus, this SRBM has a product form stationary density function and the mean vector of the stationary distribution can be computed to be $(0.5, 0.95)$. Following from proposition 28, if we scale Z by $D = I$ and a scalar α, we will get a different SRBM Z^* but with the same stationary distribution as SRBM Z. We list our numerical approximation of the means of stationary distribution of Z^* in table 1 for several different α by using the uniform 10×10 mesh. From this table, it can be observed that the smaller α is, the more accurate estimates the results are. However, this does not mean that the BNAfm algorithm gives poor estimates for this problem when α is large. Instead, it indicates that a mesh denser than 10×10 should be used in order to produce good approximations when α is large. In this table, we also show the number of iterations needed for the iterative method. Loosely speaking, more iterations means that the system matrix A is more ill-conditioned. Thus, we can conclude partially that smaller Γ and θ would give better approximations in our algorithm. In practice, if some elements of θ or some diagonal elements of matrix Γ are large, we should scale them properly before carrying out the numerical computation.

Table 1
Comparisons of different scaling.

α	q_1	q_2	Iterations
50.00	0.399164	0.783404	530
10.00	0.521410	0.955050	330
1.00	0.513462	0.95336	116
0.10	0.500231	0.950072	88
0.01	0.499998	0.950001	407
Exact	0.50000	0.950000	

5. Numerical examples

In this section, we present two SRBMs whose stationary distributions can be obtained through methods other than the algorithm proposed in this paper. We compare the accuracy of our algorithm with those known methods. In the first case, we show that the BNAfm algorithm produces estimates as good as the BNAsm algorithm. In the second case, we show that the BNAfm algorithm produces good estimates of stationary probabilities. We also present empirical evidence of the complexity of our algorithm.

5.1. Comparison with SC solution

In this subsection we apply our BNAfm algorithm to a two-dimensional SRBM that was studied by Dai and Harrison [14]. The data of this SRBM are $\theta = 0$, $\Gamma = 2I$, $S = [0, a] \times [0, 1]$, and

$$R = \begin{pmatrix} 1 & 0 & -1 & 1 \\ -1 & 1 & 0 & -1 \end{pmatrix}.$$

As discussed in section 2.5 of [14], the density function $p \notin L^2(S)$. However the BNAfm algorithm still gives a very accurate approximation that is consistent with the results obtained by Dai and Harrison [14] using the BNAsm algorithm.

As in [14], we fix the height of the rectangle and vary the length a of the rectangle. For the various values of the length parameter a, table 2 compares three different estimates of q_1 and q_2. The BNAfm is obtained by our algorithm with a 9×9 uniform mesh. The SC estimates were obtained by Trefethen and Williams [37] using an explicit expression of the stationary density. The expression was obtained by Harrison et al. [24] for general two-dimensional driftless SRBMs, and is based on the Schwarz–Christoffel (SC) transformation in complex variables. BNAsm and SC estimates are taken from [14].

It is clear from the table that the accuracy of our BNAfm algorithm is at least as good as BNAsm in [14]. It takes less than 1 second CPU time and 800 Kilobyte of memory for both iterative and direct methods to obtain BNAfm estimates for every value of length parameter a.

A very coarse estimate of the condition number of matrix A is 4.7×10^{11}, which is very large. Because of the ill-conditioning, we have observed that the number of iterations performed in order to get 6-decimal precision is very close to the size of the

Table 2
Estimates of stationary means from different algorithms for a special two-dimensional SRBM.

a	Method	q_1	q_2
0.5	BNAsm	0.258229	0.380822
	BNAfm	0.258548	0.380244
	SC	0.258585	0.380018
1.0	BNAsm	0.551325	0.448675
	BNAfm	0.551511	0.448571
	SC	0.551506	0.448494
1.5	BNAsm	0.878800	0.471640
	BNAfm	0.879476	0.471676
	SC	0.879534	0.471624
2.0	BNAsm	1.238442	0.483103
	BNAfm	1.239767	0.482937
	SC	1.239964	0.482830

linear system. For example, the number of iterations for $a = 1$ is 374 while the size of the linear system is 400 when using a 9×9 mesh.

5.2. A 3-dimensional SRBM with product form solution

One of the main reasons that we develop the BNAfm algorithm is to approximate the stationary distribution function, not just its mean values. To see how effective this algorithm is, we introduce a special 3-dimensional SRBM whose stationary density has an explicit product form solution. Then we compare numerical results from our BNAfm algorithm with analytical solutions.

The data of the SRBM are given as

$$R = \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix},$$

$\theta' = (1, -1, -0.5)$, $\Gamma = I$, and $S = [0, 1] \times [0, 1] \times [0, 1]$. Since the data satisfies the skew symmetry condition in [28], the stationary density function p_0 is of exponential form,

$$p_0(x) = \frac{2\exp(-2x_2 - x_3)}{(1 - \mathrm{e}^{-2})(1 - \mathrm{e}^{-1})}, \quad \text{for } x \in S. \tag{29}$$

Table 3 compares the exact means of the stationary distribution with the approximate results obtained by our BNAfm algorithm. In this numerical example, we use uniform mesh. The index i in the table denotes the total number of partitions at each dimension, and the index n denotes the size of the linear system for each different mesh. In this table, we also show the computing time and memory usage for both iterative and direct methods. The computing time is measured by second and the memory is measured

Table 3
Comparisons for a 3-dimensional SRBM with product form stationary density.

	Means			LU method		Iterative method		n
	q_1	q_2	q_3	Time	Memory	Time	Memory	
$i = 4$	0.500043	0.344660	0.418012	9	4.1	15	3.1	1,000
$i = 6$	0.500033	0.343893	0.418021	33	13.6	62	8.4	2,744
$i = 8$	0.500021	0.343677	0.418023	86	35.9	180	17.6	5,832
$i = 10$	0.500013	0.343592	0.418023	200	76.7	420	33.2	10,648
$i = 12$	0.500009	0.343551	0.418023	441	149.0	930	57.4	17,576
Exact	0.500000	0.343482	0.418023					

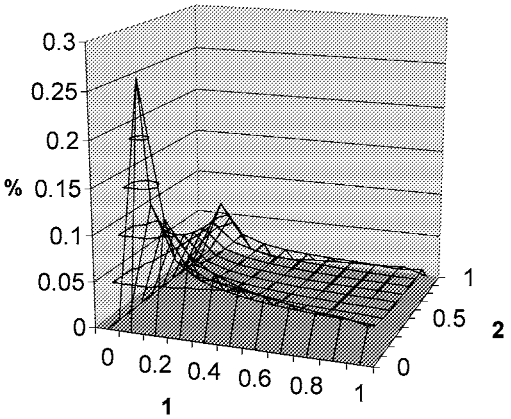

Figure 2. Percentage errors of approximate marginal stationary distribution $\boldsymbol{P}_1$.

by Megabytes. The approximate results obtained by both direct and iterative methods are very close, so we only list the results obtained by the direct method. A very coarse estimate of the condition number of matrix A is 6.7×10^{14} for $i = 10$ and 9.6×10^{14} for $i = 12$, which is much larger than the case for the previous two-dimensional example (section 5.1).

Table 3 shows that if we require 1% accuracy (which is usually good enough in queueing network applications), the convergence is very fast ($i = 4$ is good enough). It also shows that when the mesh is refined, the accuracy of approximate means increases slowly, while the required computing time and memory increase exponentially. Compared with the direct method, the iterative method takes almost twice as much computing time but only takes about half as much memory as the direct method. Using less memory will definitely help us to solve large-scale practical problems although it will take longer computing time.

Figures 2–4 are plots regarding the computation of three two-dimensional marginal stationary distributions $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, and $\boldsymbol{P}_3$, where $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, and $\boldsymbol{P}_3$ are defined as

$$\boldsymbol{P}_1(x_1, x_2) = \int_{0 \leqslant s_1 \leqslant x_1, 0 \leqslant s_2 \leqslant x_2} p_0(s)\, \mathrm{d}s,$$

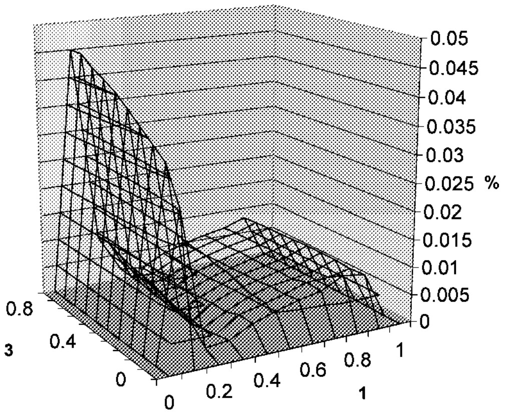

Figure 3. Percentage errors of approximate marginal stationary distribution $\boldsymbol{P}_2$.

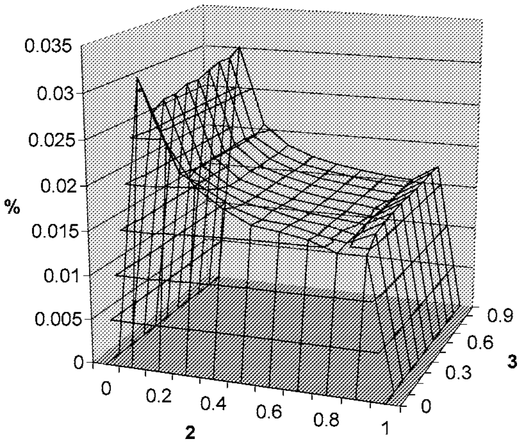

Figure 4. Percentage errors of approximate marginal stationary distribution $\boldsymbol{P}_3$.

$$\boldsymbol{P}_2(x_1, x_3) = \int_{0\leqslant s_1\leqslant x_1, 0\leqslant s_3\leqslant x_3} p_0(s)\,\mathrm{d}s,$$

$$\boldsymbol{P}_3(x_2, x_3) = \int_{0\leqslant s_2\leqslant x_2, 0\leqslant s_3\leqslant x_3} p_0(s)\,\mathrm{d}s.$$

The vertical axes represent the percentage errors of our computation results compared against the exact results. As can be seen from these three figures, the BNAfm algorithm provides very accurate estimations for the stationary distribution.

6. A queueing network application

In this section, we show how our BNAfm algorithm, proposed for solving the stationary distribution of SRBMs, can be used to predict the performance of a 3-station finite-buffer queueing network.

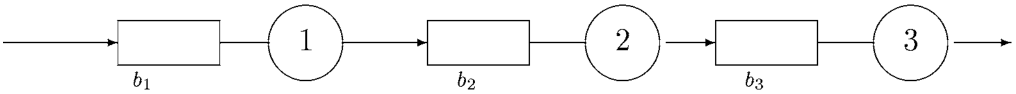

Figure 5. Finite queues in tandem.

Pictured in figure 5 is a queueing network of 3 stations in series. Each station has a single server with a first-in-first-out service discipline. The buffer size at each station is assumed to be finite. We use b_i to denote the buffer size (the number of waiting rooms plus 1) at station i, $i = 1, 2, 3$. Jobs arrive at station 1 according to a Poisson process with rate $\lambda = 1$. After completing services at station 1, they go to station 2, and after completing services there, they proceed to station 3. They exit the system after completing services at station 3. To deal with the finiteness of buffers, we make the convention that a job entering a full buffer is simply discarded (or lost). Such a network is referred to as a loss network, which is commonly used to model computer networks.

The service times at each station are assumed to be i.i.d. positive random variables, and service times at different stations are assumed to be independent. The service time distribution at station 1 is taken to be Erlang of order 4. Thus, the squared coefficient of variation (variance divided by the mean squared) of the service time distribution is $c_1^2 = 1/4 = 0.25$. The service time distribution at station 2 is taken to be exponential, and thus $c_2^2 = 1$. The service time distribution at station 3 is taken to be a Gamma distribution with $c_3^2 = 2$. The service rate μ_i and the buffer size b_i at each station i, $i = 1, 2, 3$, are shown in table 4.

Let $Z_i(t)$ be the queue length, including possibly the one being served, at station i at time t, $i = 1, 2, 3$. Following the approach in [26], W. Dai [19] proposed an SRBM in the 3-dimensional box to approximate the queue length process $Z = \{Z(t),\ t \geqslant 0\}$, where $Z(t) = (Z_1(t), Z_2(t), Z_3(t))$. The SRBM has the following data: $S = \{z \in \Re^3_+ : 0 \leqslant z_i \leqslant b_i,\ i = 1, 2, 3\}$, $\theta = (1 - \mu_1, \mu_1 - \mu_2, \mu_2 - \mu_3)'$,

$$R = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & -1 & 0 \\ 0 & -1 & 1 & 0 & 0 & -1 \end{pmatrix},$$

and

$$\Gamma = \begin{pmatrix} 1 + \dfrac{\mu_1}{4} & -\dfrac{\mu_1}{4} & 0 \\ -\dfrac{\mu_1}{4} & \dfrac{\mu_1}{4} + \mu_2 & -\mu_2 \\ 0 & -\mu_2 & \mu_2 + 2\mu_3 \end{pmatrix}. \tag{30}$$

Using the BNAfm algorithm proposed in section 3, one can compute the stationary distribution and the stationary mean of the SRBM. The stationary mean is then used to estimate the long-run average queue lengths of the loss network. The SRBM rows in table 5 lists the estimates of average queue lengths, q_1, q_2, and q_3, for different cases.

For comparison, we have simulated the loss network in each case. The corresponding estimates are given in the simulation rows. In each case, the simulation estimates are

Table 4
The parameters of the queueing network.

System	b_1	b_2	b_3	μ_1	μ_2	μ_3
1	10	10	10	1/0.9	1/0.9	1/0.9
2	20	25	25	0.9	0.9	0.9
3	10	15	15	0.9	0.9	0.9
4	3	5	5	1/0.9	1/0.9	1/0.9

Table 5
The average queue lengths of the queueing network.

System	Approximate method	q_1	q_2	q_3
1	BNAfm	3.619 (1.4%)	3.651 (3.2%)	4.172 (12.5%)
	Simulation	3.669 (0.7%)	3.539 (1.1%)	3.709 (0.6%)
2	BNAfm	14.669 (1.6%)	12.137 (1.7%)	11.565(2.5%)
	Simulation	14.912 (0.5%)	12.344 (1.5%)	11.286 (0.9%)
3	BNAfm	6.304 (2.2%)	6.731 (2.2%)	6.780 (4.1%)
	Simulation	6.443 (0.4%)	6.883 (1.1%)	6.515 (1.2%)
4	BNAfm	1.370 (0.4%)	1.795 (10.2%)	2.086 (22.6%)
	Simulation	1.364 (0.2%)	1.629 (0.5%)	1.701 (0.5%)

based on 10 batches of 200,000 units of time, with the simulation in the first 10,000 units of time truncated. The numbers in parentheses after the simulation figures show 95% confidence intervals as the percentage of the simulation figures. The numbers in parentheses following all other figures are percentage errors (in absolute values) as compared to simulation results.

For this loss network, there are other performance measures that are important in practice. For example, one might be interested in the throughput at each station. (The throughput γ_i at station i is the long-run average number of jobs leaving station i per unit of time.) The throughput γ_i at station i is related to the utilization rate ρ_i at the station via

$$\gamma_i = \mu_i \rho_i, \quad i = 1, 2, 3.$$

Let $m_i = 1/\mu_i$. Note the definition of δ_k ($k = 1, \ldots, 6$) in (9). Then the Brownian estimate of ρ_k is given by

$$\rho_i = 1 - m_i \delta_i, \quad i = 1, 2, 3. \tag{31}$$

Also, the long-run fraction of jobs lost at station i can be estimated via δ_{3+i}, $i = 1, 2, 3$. Tables 6 and 7 list the simulation results and SRBM estimates of average throughput rates and job loss rates for different cases.

In obtaining the Brownian model with the covariance matrix given in (30), we implicitly assumed that the actual utilization rate ρ_i can be replaced by 1. This assumption

Table 6
The average throughput rates of the queueing network.

System	Approximate method	γ_1	γ_2	γ_3
1	BNAfm	0.976 (0.3%)	0.947 (0.4%)	0.859 (3.9%)
	Simulation	0.973 (0.2%)	0.951 (0.2%)	0.894 (0.3%)
2	BNAfm	0.896 (0.0%)	0.875 (0.1%)	0.836 (0.3%)
	Simulation	0.896 (0.0%)	0.876 (0.1%)	0.839 (0.2%)
3	BNAfm	0.876 (0.3%)	0.848 (0.8%)	0.788 (1.9%)
	Simulation	0.879 (0.2%)	0.855 (0.3%)	0.803 (0.2%)
4	BNAfm	0.838 (0.1%)	0.784 (3.2%)	0.611 (17.2%)
	Simulation	0.837 (0.0%)	0.810 (0.2%)	0.738 (0.3%)

Table 7
The average job loss rates of the queueing network.

System	Approximate method	δ_4	δ_5	δ_6
1	BNAfm	0.024 (4.0%)	0.030 (25.0%)	0.088 (63.0%)
	Simulation	0.025 (3.1%)	0.024 (3.1%)	0.054 (2.3%)
2	BNAfm	0.104 (0.0%)	0.021 (0.0%)	0.039 (11.4%)
	Simulation	0.104 (2.1%)	0.021 (4.4%)	0.035 (3.1%)
3	BNAfm	0.124 (2.5%)	0.028 (3.7%)	0.061 (27.1%)
	Simulation	0.121 (1.1%)	0.027 (3.3%)	0.048 (3.0%)
4	BNAfm	0.162 (1.3%)	0.054 (80.0%)	0.173 (162.1%)
	Simulation	0.160 (0.5%)	0.030 (1.6%)	0.066 (1.2%)

requires that each station is heavily loaded and each buffer size is large. As discussed in [14], one can refine the Brownian model by replacing the covariance matrix by

$$\Gamma = \begin{pmatrix} 1 + \dfrac{\rho_1\mu_1}{4} & -\dfrac{\rho_1\mu_1}{4} & 0 \\ -\dfrac{\rho_1\mu_1}{4} & \dfrac{\rho_1\mu_1}{4} + \rho_2\mu_2 & -\rho_2\mu_2 \\ 0 & -\rho_2\mu_2 & \rho_2\mu_2 + 2\rho_3\mu_3 \end{pmatrix}. \tag{32}$$

Since the utilization rate $\rho = (\rho_1, \rho_2, \rho_3)$ itself is unknown, we denote the covariance in (32) by $\Gamma(\rho)$. We now use an iterative procedure to find ρ and other performance measures simultaneously. We initialize $\rho(0) = (1, 1, 1)$. Assume that $\rho(n-1)$ is known. We use the BNAfm algorithm to find the stationary density corresponding to covariance matrix $\Gamma(\rho(n-1))$. The associated $\delta(n-1)$ can be obtained at the same time using formula (9). Then we use (31) to get an update for $\rho(n)$. The iterations, along with the refined Brownian estimates, are given in tables 8–10. The case $n = 1$ corresponds to the original Brownian model whose results have been shown in tables 5–7.

By observing the numerical results in tables 8–10, we can see that the above iterative procedure provides a slightly better Brownian model for performance evaluation

Table 8
The iterations of the SRBM approximation for average queue lengths.

System	n	$q_1(n)$	$q_2(n)$	$q_3(n)$
1	1	3.619 (1.4%)	3.651 (3.2%)	4.172 (12.5%)
	2	3.585 (2.3%)	3.507 (0.9%)	4.022 (8.4%)
	3	3.585 (2.3%)	3.515 (0.7%)	4.046 (9.1%)
	Simulation	3.669 (0.7%)	3.539 (1.1%)	3.709 (0.6%)
2	1	14.669 (1.6%)	12.137 (1.7%)	11.565 (2.5%)
	2	14.671 (1.6%)	12.170 (1.4%)	11.534 (2.2%)
	3	14.671 (1.6%)	12.141 (1.6%)	11.532 (2.2%)
	Simulation	14.912 (0.5%)	12.344 (1.5%)	11.286 (0.9%)
3	1	6.304 (2.2%)	6.731 (2.2%)	6.780 (4.1%)
	2	6.310 (2.1%)	6.700 (2.7%)	6.730 (3.3%)
	3	6.310 (2.1%)	6.702 (2.7%)	6.733 (3.3%)
	Simulation	6.443 (0.4%)	6.883 (1.1%)	6.515 (1.2%)
4	1	1.370 (0.4%)	1.795 (10.2%)	2.086 (22.6%)
	2	1.363 (0.0%)	1.634 (0.3%)	1.896 (11.5%)
	3	1.363 (0.0%)	1.656 (1.7%)	1.967 (15.6%)
	Simulation	1.364 (0.2%)	1.629 (0.5%)	1.701 (0.5%)

Table 9
The iterations of the SRBM approximation for average throughput rates.

System	n	$\gamma_1(n)$	$\gamma_2(n)$	$\gamma_3(n)$
1	1	0.976 (0.3%)	0.947 (0.4%)	0.859 (3.9%)
	2	0.978 (0.5%)	0.955 (0.4%)	0.892 (0.2%)
	3	0.978 (0.5%)	0.954 (0.3%)	0.889 (0.6%)
	Simulation	0.973 (0.2%)	0.951 (0.2%)	0.894 (0.3%)
2	1	0.896 (0.0%)	0.875 (0.1%)	0.836 (0.3%
	2	0.897 (0.1%)	0.876 (0.0%)	0.839 (0.0%)
	3	0.896 (0.0%)	0.876 (0.0%)	0.839 (0.0%)
	Simulation	0.896 ().0%)	0.876 (0.1%)	0.839 (0.2%)
3	1	0.876 (0.3%)	0.848 (0.8%)	0.788 (1.9%)
	2	0.876 (0.3%)	0.850 (2.6%)	0.797 (0.7%)
	3	0.876 (0.3%)	0.850 (2.6%)	0.797 (0.7%)
	Simulation	0.879 (0.2%)	0.855 (0.3%)	0.803 (0.2%)
4	1	0.838 (0.1%)	0.784 (3.2%)	0.611 (17.2%)
	2	0.849 (1.4%)	0.819 (1.1%)	0.744 (0.8%)
	3	0.848 (1.3%)	0.816 (0.7%)	0.717 (2.8%)
	Simulation	0.837 (0.0%)	0.810 (0.2%)	0.738 (0.3%)

compared to the original Brownian model, especially for system No. 4. By comparing numerical results to simulation results, the SRBM model gives fairly good approximations. Performance approximation to station 3 is not as good as that to stations 1 and 2. This may be due to the large variation of service time at station 3.

Table 10
The iterations of the SRBM approximation for average job loss rates.

System	n	$\delta_4(n)$	$\delta_5(n)$	$\delta_6(n)$
1	1	0.024 (4.0%)	0.030 (25.0%)	0.088 (63.0%)
	2	0.022 (12.0%)	0.023 (4.3%)	0.062 (14.8%)
	3	0.022 (12.0%)	0.0233 (4.3%)	0.065 (20.4%)
Simulation		0.025 (3.1%)	0.024 (3.1%)	0.054 (2.3%)
2	1	0.104 (0.0%)	0.021 (0.0%)	0.039 (11.4%)
	2	0.104 (0.0%)	0.020 (2.0%)	0.037 (5.7%)
	3	0.104 (0.0%)	0.020 (2.0%)	0.037 (5.7%)
Simulation		0.104 (2.1%)	0.021 (4.4%)	0.035 (3.1%)
3	1	0.124 (2.5%)	0.028 (3.7%)	0.061 (27.1%)
	2	0.124 (2.5%)	0.026 (3.7%)	0.053 (10.4%)
	3	0.124 (2.5%)	0.026 (3.7%)	0.053 (10.4%)
Simulation		0.121 (1.1%)	0.027 (3.3%)	0.048 (3.0%)
4	1	0.162 (1.3%)	0.054 (80.0%)	0.173 (162.1%)
	2	0.151 (0.6%)	0.030 (0.0%)	0.075 (13.6%)
	3	0.152 (5.0%)	0.032 (6.7%)	0.099 (50.0%)
Simulation		0.160 (0.5%)	0.030 (1.6%)	0.066 (1.2%)

Table 11
The comparison of tail probabilities of system No. 1.

k	$1 - \boldsymbol{P}_1(k)$		$1 - \boldsymbol{P}_2(k)$		$1 - \boldsymbol{P}_3(k)$	
	BNAfm	Simul.	BNAfm	Simul.	BNAfm	Simul.
1	0.8035	0.7232	0.7850	0.6846	0.8421	0.6675
2	0.6391	0.5802	0.6190	0.5446	0.6988	0.5547
3	0.5016	0.4566	0.4846	0.4290	0.5702	0.4561
4	0.3866	0.3519	0.3742	0.3324	0.4553	0.3689
5	0.2905	0.2633	0.2827	0.2519	0.3532	0.2920
6	0.2100	0.1877	0.2063	0.1839	0.2627	0.2237
7	0.1427	0.1238	0.1421	0.1260	0.1830	0.1628
8	0.0864	0.0702	0.0878	0.0764	0.1130	0.1092
9	0.0394	0.0248	0.0412	0.0345	0.0521	0.0614
10	0	0	0	0	0	0

Another performance measure related to the queueing network is the (tail) probability that the total number of jobs in the system is at least k for a positive integer k. Such performance measures are needed to assess the *quality of service* for a queueing network. In tables 11 and 12, we use systems No. 1 and 2 to compare (tail) probabilities for each station calculated from both SRBM and simulation. From these two tables, we find that the SRBM estimates are not very close to simulation results, but they are reasonably good enough in practice when high precision is not required. We note that there are two possible errors here: one is from the BNAfm algorithm itself; the other results from using SRBMs to approximate original queueing networks. We strongly believe that the main error here is the SRBM approximation error.

Table 12
The comparison of tail probabilities of system No. 2.

k	$1 - \boldsymbol{P}_1(k)$		$1 - \boldsymbol{P}_2(k)$		$1 - \boldsymbol{P}_3(k)$	
	BNAfm	Simul.	BNAfm	Simul.	BNAfm	Simul.
1	0.9930	0.9898	0.9566	0.9336	0.9525	0.8818
2	0.9847	0.9817	0.9135	0.8935	0.9054	0.8353
4	0.9635	0.9600	0.8283	0.8128	0.8123	0.7477
6	0.9340	0.9298	0.7441	0.7307	0.7211	0.6638
8	0.8932	0.8868	0.6611	0.6484	0.6322	0.5831
10	0.8367	0.8263	0.5791	0.5667	0.5459	0.5048
12	0.7583	0.7421	0.4983	0.4864	0.4623	0.4298
14	0.6495	0.6258	0.4186	0.4084	0.3817	0.3575
16	0.4988	0.4658	0.3401	0.3310	0.3042	0.2881
18	0.2898	0.2430	0.2627	0.2553	0.2300	0.2219
19	0.1567	0.1005	0.2238	0.2178	0.1942	0.1897
20	0	0	0.1864	0.1804	0.1593	0.1583
22	0	0	0.1113	0.1054	0.0924	0.0979
24	0	0	0.0370	0.0328	0.0296	0.0406
25	0	0	0	0	0	0

7. Concluding remarks

In this paper, we have proposed the finite element method algorithm to compute the stationary distribution of a semimartingale reflecting Brownian motion in a hypercube. This algorithm extends and complements previous algorithms. In particular, we find this algorithm accurate, stable, and capable of computing the stationary density function (in addition to the mean of the stationary distribution). Computing the density function would allow us to predict some important performance measures in real applications such as the service level in a production or communication network. We have applied the algorithm to a finite buffer queueing network, and our numerical results indicate that the algorithm in general provides very good approximations.

References

[1] C. Ashcraft, Ordering sparse matrices and transforming front trees, Working paper (1999).

[2] C. Ashcraft and R. Grimes, SPOOLES: An object-oriented sparse matrix library, in: *Proc. of the 1999 SIAM Conf. on Parallel Processing for Scientific Computing*, 22–27 March 1999.

[3] I. Bardhan and S. Mithal, Heavy traffic limits for an open network of finite buffer overflow queues: The single class case, Preprint (1993).

[4] E.B. Becker, G.F. Carey and J. Tinsley Oden, *Finite Elements: An Introduction* (Prentice-Hall, Englewood Cliffs, NJ, 1981).

[5] D. Bertsekas and R. Gallager, *Data Networks* (Prentice-Hall, Englewood Cliffs, NJ, 1992).

[6] J. Buzacott and J.G. Shanthikumar, Design of manufacturing systems using queueing models, Queueing Systems 12 (1992) 135–213.

[7] C. Canuto, M.Y. Hussaini, A. Quarteroni and T.A. Zang, *Spectral Methods in Fluid Dynamics* (Springer, Berlin, 1988).

[8] G.F. Carey and J. Tinsley Oden, *Finite Elements: A Second Course* (Prentice-Hall, Englewood Cliffs, NJ, 1981).

[9] H. Chen and A. Mandelbaum, Hierarchical modelling of stochastic networks, Part II: Strong approximations, in: *Stochastic Modeling and Analysis of Manufacturing Systems*, ed. D.D. Yao (Springer, Berlin, 1994) pp. 107–131.

[10] H. Chen and X. Shen, Computing the stationary distribution of SRBM in an orthant, Preprint (2000).

[11] H. Chen and D.D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization* (Springer, Berlin, 2001).

[12] J.G. Dai, Steady-state analysis of reflected Brownian motions: characterization, numerical methods and queueing applications, Ph.D. thesis, Stanford University (1990).

[13] J.G. Dai and W. Dai, A heavy traffic limit theorem for a class of open queueing networks with finite buffers, Queueing Systems 32 (1998) 5–40.

[14] J.G. Dai and J.M. Harrison, Steady-state analysis of RBM in a rectangle: Numerical methods and a queueing application, Ann. Appl. Probab. 1 (1991) 16–35.

[15] J.G. Dai and J.M. Harrison, Reflected Brownian motion in an orthant: Numerical methods for steady-state analysis, Ann. Appl. Probab. 2 (1992) 65–86.

[16] J.G. Dai and T.G. Kurtz, Characterization of the stationary distribution for a semimartingale reflecting Brownian motion in a convex polyhedron, Preprint (1997).

[17] J.G. Dai, V. Nguyen and M.I. Reiman, Sequential bottleneck decompositions: An approximation method for generalized Jackson networks, Oper. Res. 42 (1994) 119–136.

[18] J.G. Dai and R. Williams, Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons, Theory Probab. Appl. 40 (1995) 1–40.

[19] W. Dai, Brownian approximations for queueing networks with finite buffers: Modeling, heavy traffic analysis and numerical implementations, Ph.D. dissertation, Georgia Institute of Technology (1996).

[20] J. Dongarra, A. Lumsdaine, R. Pozo and K. Remington, A sparse matrix library in C++ for high performance architectures, in: *Proc. of the 2nd Object Oriented Numerics Conf.*, 1994, pp. 214–218.

[21] J. Dongarra, A set of level 3 basic linear algebra subprograms, ACM Trans. Math. Software 16 (1990) 1–17.

[22] S.N. Ethier and T.G. Kurtz, *Markov Process: Characterization and Convergence* (Wiley, New York, 1986).

[23] P.W. Glynn, Diffusion approximations, in: *Handbooks in Operations Research and Management Science, II: Stochastic Models*, eds. D.P. Heyman and M.J. Sobel (North-Holland, Amsterdam, 1990) pp. 145–198.

[24] J.M. Harrison, H. Landau and L.A. Shepp, The stationary distribution of reflected Brownian motion in a planar region, Ann. Probab. 13 (1985) 744–757.

[25] J.M. Harrison, *Brownian Motion and Stochastic Flow Systems* (Wiley, New York, 1985).

[26] J.M. Harrison and V. Nguyen, The QNET method for two-moment analysis of open queueing networks, Queueing Systems 6 (1990) 1–32.

[27] J.M. Harrison and V. Nguyen, Brownian models of multiclass queueing networks: Current status and open problems, Queueing Systems 13 (1993) 5–40.

[28] J.M. Harrison and R.J. Williams, Multidimensional reflected Brownian motions having exponential stationary distributions, Ann. Probab. 15 (1987) 115–137.

[29] J.R. Jackson, Job shop-like queueing systems, Managm. Sci. 10 (1963) 131–142.

[30] F.P. Kelly, *Reversibility and Stochastic Networks* (Wiley, New York, 1979).

[31] L. Kleinrock, *Queueing Systems II: Computer Applications* (Wiley, New York, 1976).

[32] A.J. Lemoine, Network of queues – a survey of weak convergence results, Managm. Sci. 24 (1978) 1175–1193.

[33] M.J. Maron, *Numerical Analysis: A Practical Approach*, 2nd ed. (Macmillan, New York, 1987).

[34] J.T. Oden and J.N. Reddy, *An Introduction to the Mathematical Theory of Finite Elements* (Wiley-Interscience, New York, 1976).

[35] E. Schwerer, A linear programming approach to the steady-state analysis of Markov process, Ph.D. dissertation, Stanford University (1997).

[36] X. Shen, Performance evaluation of multiclass queueing networks via Brownian motions, Ph.D. dissertation, Faculty of Commerce and Business Administration, University of British Columbia (2001).

[37] L. Trefethen and R.J. Williams, Conformal mapping solution of Laplace's equation on a polygon with oblique derivative boundary conditions, J. Comput. Appl. Math. 14 (1985) 227–249.

[38] J. Walrand, *An Introduction to Queueing Networks* (Prentice-Hall, Englewood Cliffs, NJ, 1988).

[39] W. Whitt, Heavy traffic theorems for queues: A survey, in: *Mathematical Methods in Queueing Theory*, ed. A.B. Clarke (Springer, Berlin, 1974) pp. 307–350.

[40] W. Whitt, The queueing network analyzer, Bell System Tech. J. 62(9) (1983) 2779–2815.

[41] R.J. Williams, On the approximation of queueing networks in heavy traffic, in: *Stochastic Networks: Theory and Applications*, eds. F.P. Kelly, S. Zachary and I. Ziedins (Oxford Univ. Press, Oxford, 1996) pp. 35–56.

[42] D.D. Yao, ed., *Stochastic Modeling and Analysis of Manufacturing Systems* (Springer, New York, 1994).