DIFFUSION APPROXIMATION FOR STATIONARY ANALYSIS OF QUEUES AND THEIR NETWORKS: A REVIEW

Masakiyo Miyazawa
Tokyo University of Science

Abstract   Diffusion processes have been widely used for approximations in the queueing theory. There are different types of diffusion approximations. Among them, we are interested in those obtained through limits of a sequence of models which describe queueing networks. Such a limit is typically obtained by the weak convergence of either stochastic processes or stationary distributions. We already have nice reviews and text books for them. However, this area is still actively studied, and it seems getting hard to have a comprehensive overview because mathematical results are highly technical. We try to fill this gap presenting technical background. Although those diffusion approximations have been well developed, there remains a big problem, which is difficulty to get useful information from the limiting diffusion processes. Their state spaces are multidimensional, whose dimension corresponds to the number of nodes for a single-class case and the number of customer types for a multi-class case. We now have a better view for the two dimensional case, but still know very little about the higher dimensional case. This intractability is somehow against a spirit of diffusion approximation. This motivates us to reconsider diffusion approximation from scratch. For this, we highlight the stationary distributions, and make clear a mechanism to produce diffusion approximations.

1. Introduction

The aim of this paper is to review some technical aspects on diffusion approximation arising in queues and their networks in heavy traffic. We only consider open networks which have exogenous arrivals. The diffusion approximation is meant to use a diffusion process for approximation. Here, a diffusion process is a continuous time strong Markov process which is characterized by a stochastic integral equation using a Brownian motion and continuous processes with bounded variations (e.g., see (6.4) for this integration). This diffusion process has a continuous path, and is called Itô diffusion in some literature, but we will just call it a diffusion process. A comprehensive introduction about it can be found in Harrison [30] (see also Kushner [48] for control problems). Mathematical details can be found in Section 5 of Protter [54].

Queueing models can not be described by diffusion processes in general because customers are discrete in nature, and therefore their sample paths are not continuous. However, analytical studies on queueing models are generally very hard particularly for queueing networks except for special cases. This motivates to use diffusion processes for their approximations because diffusion processes are characterized by fewer modeling parameters, that is, fewer primitive data, and have simpler state spaces. For example, the first two moments including covariances are typical primitive data for diffusion processes. Thus, diffusion approximation may be considered as a two-moment approximation. Their state spaces are

finite dimensional vector spaces, while queueing processes require supplementary states for describing their dynamics. Those two features are analytical advantages of diffusion approximation. However, one must be careful of such simplification because important information might be lost.

Another motivation for diffusion approximation comes from queueing dynamics. In general, queues and their networks are input-output systems, and inputs and outputs can be described by additive processes. For example, a counting process of arriving customers and accumulated workload brought by customers are such additive processes. If the additive process has independent or weakly dependent increments, then it is natural to use Brownian motion for approximation. In this case, queueing characteristics such as queue lengths are functionals of Brownian motion, which may not be easily analyzed particularly for the multi-dimensional case, but fewer primitive data are advantage to study feature of the original model.

Diffusion approximation is so versatile, and the literature about it is so huge. It is hard to survey its whole area. Thus, we focus on three things. The first is about a method to derive diffusion approximations. We are only interested in those approximations supportable by theory. Namely, they should be asymptotically exact under certain limiting operations. Because objects of our interest are a stochastic process and its stationary distribution, the limiting operation is convergence in distribution under scaling of state (and time if necessary). Here, we will show that scaling in time may not be necessary if we are only concerned with the stationary distribution. This is different from the standard approach of diffusion approximation, but may not be surprising because the stationary distribution is independent of time. Furthermore, Kingman [42] uses such scaling for the waiting time of a single server queue.

The next step is to derive approximations based on those limiting results. However, this step is often ignored in the literature because it may be considered obvious. For example, the limiting processes and distributions may be considered as approximations themselves. Here is a gap between mathematical results and approximations. To fill this gap, we go back to the original idea of Kingman [42]. He proposes an approximation for the stationary waiting time distribution for a single server queue by moving a scaling factor into the limiting distribution. We think this is a right way to have theoretical support. Of course, we may be motivated by intuitive arguments, and should not exclude heuristics. However, it also is important to legitimize them by theory.

The second is about a class of queueing models. Our primary interest is in queueing networks, but also consider some of single node queues because they are components of network systems. Key features of a queueing network are routing of customers and idling of servers. There are two types of routing, choosing next nodes or next types. The latter is used for a multi-class queueing network. In either case, we assume Markovian routing, which means that nodes or types are chosen with given probabilities only depending on the current nodes or types, respectively, independently of everything else. If exogenous arrivals at nodes are subject to independent renewal processes, if service times at each server are *i.i.d.* (independent identically distributed) and if service discipline is first-come first-served, FCFS for short, then this network model is called a generalized Jackson network, GJN for short. This network is a main subject of this review.

The third is about objects for approximations. As we mentioned above, a stochastic process describing a model is a primary object. However, from the application viewpoint, its stationary distribution is also very important. Other objects are moments and tail asymptotics of the stationary distribution. Note that the weak convergence of the sequence of

stochastic processes does not imply that of their stationary distributions. That is, the exchange of those limits needs proofs. If it is possible, the approximation is considered to have better support. Similarly, the convergence of the sequence of the stationary distributions does not imply the convergences of their moments and tail asymptotics. If they do so, we may say the approximation has better quality.

Theoretical support for diffusion approximation generally requires a heavy traffic condition. Under this condition, the traffic intensity at each node is close to unit, and therefore the behavior of a small queue at a fixed node is ignorable. In other words, diffusion approximation is for the behavior of a large queue. This greatly simplifies diffusion approximation for a single queue. On the other hand, in a queueing network, there are multiple queues at different nodes, and they may not simultaneously get large. For example, there may be idle servers when some queues are large. This is a crucial aspect of the queueing network, and should not be excluded in its approximation. Thus, it is interesting to see how such a behavior is implemented in diffusion approximation.

Historically, the study of diffusion approximation started from single node queues. Among them, a $GI/GI/1$ queue is known as a basic model. This model assumes that customers arrive according to a renewal process, there is a single server and service times are $i.i.d.$ Even this simple model is known to be hard to get the stationary distribution in closed form. This motivated Kingman [42, 43] to originate a two-moment approximation for the stationary waiting time distribution of the $GI/GI/1$ queue in heavy traffic. This approximation is generalized for a $GI/GI/s$ queue with homogeneous servers by Köllerström [46]. Those studies directly consider the stationary distribution, so do nothing with diffusion approximation. In particular, scaling is taken only for state, as we already mentioned. However, their stationary distributions are identical with those obtained by the standard diffusion approximations (e.g., compare (2.11) with Corollary 3.1 together with (3.34)).

Independent of those studies, Borovkov [5] studied diffusion approximation for the transient behavior of a many server queue. It refers to Provkov's 1963 paper, but we are unable to get this paper. However, Iglehart and Whitt [37] credited that those Russian work is one of the origin of diffusion approximation. Thus, the work may be the first one to study diffusion limits for a single queue by scaling state and time. The book [6] of Borovkov is a good source for those studies.

A next step was an extension from a single queue to a feedforward network. This step was firstly taken by seminal papers of Iglehart and Whitt. They obtained diffusion processes as weak limits of the sequence of suitably scaled processes, called a diffusion scaling, in [36, 37]. Their results are basically for a single queue because functional relations of input and output of a single queue are essential. For example, they only studied a marginal process of a given node, but were unable to study a joint process for multiple nodes. However, there are two things to be notable. The first is an idea for process limit, that is, a weak limit of the sequence of stochastic processes. The second is that a reflecting Brownian motion is obtained as one of the process limits. This means that the diffusion approximation is more than a stochastic process version of the central limit theorem. This approximation is referred to as a functional limit theorem under normalization similar to that of the central limit theorem. A similar approach is also studied for a single queue in [6].

A real step toward a queueing network was taken by Harrison [28], who considered a tandem queue with two nodes, and got a two-dimensional reflecting Brownian motion on the nonnegative orthant. It still uses an explicit functional relation of input and output, but it was a big step to move from one-dimensional diffusion process to the two-dimensional one. Harrison and Reiman [34] studied this reflecting process for an arbitrary dimension.

It is a multidimensional reflecting Brownian motion on the nonnegative orthant. Because it has a semi-martingale representation, it is called a semi-martingale reflecting Brownian motion, SRBM for short. Those studies were culminated by the work of Reiman [56]. It obtains SRBM as a weak limit of suitably scaled processes for GJNs, where scaling is taken for both state and time. This scaling is called diffusion scaling, and the weak limit is called a diffusion limit.

Another source for diffusion approximation is heuristic arguments. A queueing network including GJN is analytically intractable except for some special cases. Typical examples for this exceptional case are a Jackson network and its extensions, called product form networks due to Kelly [40] (see [9, 58] for further development in this line). They are analytically tractable, but require stronger assumptions. Thus, there have been many attempts for heuristic approximations for a GJN, particularly for its stationary distribution. For example, the diffusion approximation of Kobayashi [45] is one of the earliest work. Among them, Queueing Network Analyzer, QNA for short, proposed by Whitt [62, 63], is well known. This is a heuristic approximation based on the product form solution and the two-moment approximation for the queue length at each node.

The concept of QNA has been widely accepted, and similar types of approximations have been considered for more complicated network models (e.g., see [44]). However, it lacks the theoretical supports which we have discussed. A theoretical improvement of QNA was studied by Harrison and Nguyen [31]. They propose another approximation concept, called QNET, incorporating information from a diffusion limit for the sequence of GJNs in heavy traffic. This limiting process is an SRBM. We note that this QNET is different from Q'NET (alternatively spelled as Q-NET) which has been developed for business solutions by IBM.

The SRBM is much simpler than the GJN, and its primitive data can be expressed by the mean and covariances of Brownian motion and the reflecting matrix. An approximation based on it may be considered as a two-moment approximation. However, the stationary distribution of the SRBM is still hard to get except for either one-dimensional case or the skew-symmetric case. The latter corresponds to the Jackson network, and has a similar product form solution. Thus, QNET requires approximation for the stationary distribution of the SRBM. For this, it uses either the product form or numerically approximated solution for the stationary distribution. Thus, QNET may not be fully supported by theory. Nevertheless, QNA and QNET had great impacts on theoretical studies for diffusion approximation through their design concepts and practical applications using their software packages.

In present days, diffusion approximations have been studied in various models including those with customer abandonment and using a different scaling in state, namely, increasing the number of servers, which is called Halfin-Whitt regime. We will briefly discuss them.

1.1. Paper structure

This paper is made up by seven sections. In Section 2, we introduce a stochastic model for the GJN, that is, generalized Jackson network, with a single-class of customers. This section have four subsections.

In Section 2.1, we start with a sample path level of modeling, then introduce stochastic assumptions for describing a GJN in Section 2.2. We are devoted to the single node case of the GJN, that is, the $GI/G/1$ queue, in Sections 2.3 and 2.4. A main interest of these two subsections is to see how we should scale state and/or time to get a two-moments approximation for the stationary distribution. For this, we first review Kingman's [42] approximation

for the waiting time in Section 2.3, which is only scaled by state. In Section 2.4, we then consider the queue length process in continuous-time, and show that scaling is only required for state similarly to the waiting time case. They are turned out to be identical with the corresponding diffusion approximations. We conjecture that similar approximations can be obtained for the GJN with an arbitrary number of nodes.

In Section 3, we consider a diffusion approximation for a stochastic process. This section is divided into four subsections. In Section 3.1, we formally define the SRBM on a nonnegative multidimensional orthant, and derive a stationary equation to characterize the stationary distribution of SRBM, which is called a basic adjoint relationship, BAR for short (e.g., see [35]). In Section 3.2, we present basic results for deriving a diffusion process as process limit. In Section 3.3. those results will be used to derive the SRBM from the queue length processes of the GJNs under scaling state and time. We finally review the case that each node may have multiple servers in Section 3.4.

In Section 4, we consider quality of diffusion approximations for the stationary distribution of the GJN and some other characteristics. For the single node case, that is, a single queue, this approximation can be directly obtained for the stationary distribution as discussed in Section 2.4. In other words, a limiting process is not needed to get a weak limit of the stationary distribution. This also would be the case for the network case. However, in the literature, the SRBM is firstly obtained as a process limit, then the stationary distributions of the GJNs under diffusion scaling are shown to weakly converges to that of the SRBM. Surprisingly, this verification has been relatively recently done by Gamarnik and Zeevi [24] then by Budhiraja and Lee [8] under weaker assumptions. We also review about diffusion approximations for moments and tail asymptotics of the stationary distribution.

In Section 5, we consider a multi-class queueing network. That is, types are attributed to customers, and nodes are exclusively attributed to types. Thus, the type of a customer uniquely specifies its service node. Each type at each node has its own primitive data, inter-arrival times (if it specifies exogenous arrivals) and service times. Customers may change their types when their services are completed. We assume that this change of types is Markovian. This network can not be handled by single-class networks in general because there may be more than one types of customers at a single node. In the nineties, multi-class queueing networks were highlighted because such models are found useful in production lines and unconventional stability conditions have been reported (e.g., see [33, 47]). These facts in addition to analytical difficulty stimulated people to study diffusion approximation for a multi-class queueing network. Harrison and Nguyen [31, 32] proposed a grand design for such diffusion approximation with some conjectures. Right after that, Dai and Wang [21] negatively answered to those conjectures.

For a multi-class queueing network, a problem is under what conditions diffusion approximation exists in what sense. Harrison [29] suggested to consider non-conventional diffusion approximation, but there seems no good alternative at least for open queueing networks. Thus, we here only consider the conventional diffusion approximation by diffusion scaling and SRBM. As for the conditions, state-space collapse and the completely-$\mathcal{S}$ condition for reflection matrix R have been recognized as key features. We review diffusion approximation for a multi-class queueing network in this line.

In Section 6, we review another type of diffusion approximations. Consider a $GI/G/s$ queue. For each fixed s, we can employ the standard diffusion approximation for this model. However, it ignores information on the number of busy servers, equivalently, the number of idle servers, so this approximation may be too rough for large s. This motivated Halfin and Whitt [27] to increase both of the arrival rate and the number s of servers for a $GI/M/s$

queue while keeping the service time distribution, which is exponential in this model. This approximation also assumes heavy traffic condition, and called Halfin-Whitt regime. Its limiting process is also a diffusion process but state-dependent. This approximation and its variants have been largely studied because they can be used for call center application.

We finally give some concluding remarks in Section 7.

2. Single-class queueing networks and heavy traffic approximation

We first introduce a stochastic process for a single-class queueing model. Here, a single-class means that there is only one customer type, in other words, customers have no type. We assume that the model is open, and has exogenous arrivals. The system is composed of a finite number of nodes which have single waiting lines with possibly multiple servers, where waiting capacity is assumed to be infinite. Service times are attached at each servers. We assume that customers are served in the first-come first-served manner. If there are multiple nodes, customers may be routed to other nodes.

2.1. Primitive data and two characteristics

In this section, we assume that each node has a single server for simplicity. We describe this model starting at time 0 by the following notations. Let $J \equiv \{1, 2, \ldots, d\}$ be the set of nodes, where d is a positive integer. For node $i \in J$, the sequence of service times $\{T_{si}(\ell); \ell \in \mathbb{N}\}$ is attached, where $\mathbb{N}$ is the set of natural numbers. Denote the set of nodes which have exogenous arrivals by J_e, and, for $i \in J_e$, let $T_{ei}(\ell)$ be the inter-arrival time between the $(\ell - 1)$-th and ℓ-th exogenous arrivals at node i for $\ell \in \mathbb{N}$. Note that $T_{ei}(\ell)$ may be 0, which means a batch arrival. Let $\mathbb{Z}_+$ is the set of all nonnegative integers. For $n \in \mathbb{Z}_+$, let $\Phi_{ij}(n)$ be the number of customers who are routed to node j among the first n departures from node i, where $\Phi_{ij}(0) = 0$. This j may be 0, which means that customers leave the network. Thus,

$$\{T_{ei}(\ell); \ell \in \mathbb{N}\}, \quad i \in J_e, \quad \{T_{si}(\ell); \ell \in \mathbb{N}\}, \quad \{\Phi_{ij}(n); n \in \mathbb{Z}_+, j \in J \cup \{0\}\}, \quad i \in J, \quad (2.1)$$

specify this queueing network. We call them primitive data. To exclude trivial case, we assume that $J_e \neq \emptyset$.

For $i \in J_e$ and $i' \in J$, let

$$V_{ei}(n) = \sum_{\ell=1}^{n} T_{ei}(\ell), \qquad V_{si'}(n) = \sum_{\ell=1}^{n} T_{si'}(\ell), \qquad n \in \mathbb{N},$$

then the first two of the primitive data in (2.1) can be replaced by

$$\{V_{ei}(n); n \in \mathbb{N}\}, \quad i \in J_e, \qquad \{V_{si}(n); n \in \mathbb{N}\}, \quad i \in J. \qquad (2.2)$$

We denote the counting processes for arrivals and service completions at node i by

$$N_{ei}(t) \equiv \sup\{n; V_{ei}(n) \leq t\}, \qquad N_{si}(t) \equiv \sup\{n; V_{si}(n) \leq t\}, \qquad (2.3)$$

where $N_{ei}(t) \equiv 0$ for $i \in J \setminus J_e$. Then, (2.3) is equivalent to (2.2), so can be used for the primitive data as well. The advantage of (2.3) is that the nodes which have no arrivals are not necessarily distinguished. Namely, we have just put $N_{ei}(t) \equiv 0$ for such nodes. It should be noted that $N_{ei}(t)$ counts exogenous arrivals at node i but $N_{si}(t)$ may not count departures from node i. The latter does so if the server is always busy up to time t.

So far, we have assumed nothing about probability laws on those primitives, but we can define queue length and workload in sample path sense. Let $L_i(t)$ be the number of customers at node i at time t, and $U_i(t)$ be the total busy time of the server at node i up to time t. Then, it is easy to see that, for $t \geq 0$ and $i \in J$,

$$L_i(t) = L_i(0) + N_{ei}(t) + \sum_{j \in J} \Phi_{ji}(N_{sj}(U_j(t))) - N_{si}(U_i(t)), \tag{2.4}$$

$$U_i(t) = \int_0^t 1(L_i(u) \geq 1)du, \tag{2.5}$$

because $N_{si}(U_i(t))$ is the number of departing customers from node i in the time interval $[0, t]$, where $1(\cdot)$ is the indicator function of the statement "$\cdot$", which takes 1 if the statement is true, and vanishes otherwise. Conversely, it is not hard to see that $\{L_i(t); t \geq 0\}$ and $\{U_i(t); t \geq 0\}$ for $i \in J$ are uniquely determined by (2.4) and (2.5). Hence, we can define them by those equations.

Since $N_{ei}(t) + \sum_{j \in J} \Phi_{ji}(N_{sj}(U_j(t)))$ is the total number of arriving customers at node i by time t, the workload $W_i(t)$ at node i can be defined as

$$W_i(t) = W_i(0) + V_{si}\Big(N_{ei}(t) + \sum_{j \in J} \Phi_{ji}(N_{sj}(U_j(t)))\Big) - U_i(t). \tag{2.6}$$

Let $\boldsymbol{L}(t)$ and $\boldsymbol{W}(t)$ be the vectors whose i-th entries are $L_i(t)$ and $W_i(t)$, respectively. Thus, we have obtained the mappings from the primitive data and the initial data $\boldsymbol{L}(0)$ and $\boldsymbol{W}(0)$ to $\boldsymbol{L}(\cdot) \equiv \{\boldsymbol{L}(t); t \geq 0\}$ and $\boldsymbol{W}(\cdot) \equiv \{\boldsymbol{W}(t); t \geq 0\}$. These mappings are called reflection mappings. In what follows we mainly consider the queue length process $\boldsymbol{L}(\cdot)$ because similar arguments can be applied for $\boldsymbol{W}(\cdot)$ with help of $\boldsymbol{L}(\cdot)$.

2.2. Stochastic assumptions for GJN

We now introduce the following stochastic assumptions on the primitive data (2.1).

(2a) $\{T_{ei}(\ell); \ell \in \mathbb{N}\}$ with $i \in J_e$, $\{T_{si}(\ell); \ell \in \mathbb{N}\}$ and $\{\Phi_{ij}(n); n \in \mathbb{Z}_+, j \in J \cup \{0\}\}$ with $i \in J$ are independent.

(2b) $\{T_{ei}(\ell); \ell \in \mathbb{N}\}$ for $i \in J_e$ and $\{T_{si}(\ell); \ell \in \mathbb{N}\}$ for $i \in J$ are sequences of i.i.d. random variables taking values in $\mathbb{R}_+$, and, for each i, $T_{ei}(\ell)$ and $T_{si}(\ell)$ have finite and positive first moments, which are denoted by m_{ei} and m_{si}, respectively. Let $\lambda_{ei} = m_{ei}^{-1}$ for $e \in J_e$ and $\lambda_{ei} = 0$ for $i \in J \setminus J_e$. Thus, λ_{ei} is the exogenous arrival rate at node $i \in J$. Similarly, the service rate at node i is denoted by $\lambda_{si} \equiv m_{si}^{-1}$.

(2c) $\{\Phi_{ij}(n); n \in \mathbb{Z}_+, j \in J \cup \{0\}\}$ is a discrete-time $(d+1)$-dimensional counting process with i.i.d. increments for each $i \in J$, and the increment $\Delta\Phi_{ij}(n) \equiv \Phi_{ij}(n) - \Phi_{ij}(n-1)$ takes values 0 or 1, and has the distribution given by

$$\mathbb{P}(\Delta\Phi_{ij}(n) = 1) = p_{ij} \geq 0, \qquad j \in J.$$

Let P be the d-dimensional square matrix whose (i, j) entry is p_{ij} for $i, j \in J$, and let $\overline{P}$ be the $(d+1)$-dimensional square matrix whose (i, j) entry is p_{ij} for $i, j \in J \cup \{0\}$, where $p_{0j} = \lambda_{ej}(\sum_{i=1} \lambda_{ei})^{-1}$ for $j \in J$ and $p_{00} = 0$. We assume that $\overline{P}$ is stochastic and irreducible, which means that customers will leave the network in finite times with probability one.

The d-node queueing network satisfying those three sets of the conditions is called a generalized Jackson network, GJN for short.

Under the assumptions (2a)–(2c), we can construct a Markov process for $\boldsymbol{L}(\cdot)$ supplemented by auxiliary characteristics. To this end, let $R_{ei}(t)$ and $R_{si}(t)$ be the residual times for the next exogenous arrival and service completion, respectively, at node i at time t, where $R_{ei}(t) \equiv 0$ for $i \in J \setminus J_e$ and $R_{si}(t) = 0$ if the server at node i is idle at time t. They are formally defined as

$$R_{ei}(t) = V_{ei}(N_{ei}(t) + 1) - t, \quad R_{si}(t) = (V_{si}(N_{si}(U_i(t)) + 1) - U_i(t))\mathbf{1}(L_i(t) \geq 1).$$

Denote the vectors whose i-th entries are $R_{ei}(t)$ and $R_{ei}(t)$ by $\boldsymbol{R}_e(t)$ and $\boldsymbol{R}_s(t)$. Then, it is not hard to see that $\{(\boldsymbol{L}(t), \boldsymbol{R}_e(t), \boldsymbol{R}_{si}(t)); t \geq 0\}$ is a Markov process. Furthermore, it is a strong Markov process (see, e.g., [16]). We refer to this process as a supplemented Markov process for the GJN.

We next consider the stability of this Markov process, that is, the condition under which the stationary distribution exists. Recall that λ_{ei} is the exogenous arrival rate at node $i \in J$. Consider the set of the following traffic equations for variables λ_{ai} with $i \in J$.

$$\lambda_{aj} = \lambda_{ej} + \sum_{i \in J} \lambda_{ai} p_{ij}, \qquad j \in J. \tag{2.7}$$

Denote the d-dimensional vectors whose entries are λ_{ei} and λ_{ai} by $\boldsymbol{\lambda}_e$ and $\boldsymbol{\lambda}_a$, respectively. Since $I - P$ is invertible by (2c), the solution $\boldsymbol{\lambda}_a$ for the equations (2.7) is given by

$$\boldsymbol{\lambda}_a = (I - P^{\mathrm{T}})^{-1} \boldsymbol{\lambda}_e,$$

where P^{T} is the transposition of P. Because $J_e \neq \emptyset$ and the irreducibility of $\overline{P}$ in (2c), $\boldsymbol{\lambda}_a$ is positive. Let $\lambda_{si} = m_{si}^{-1}$, and denote the d-dimensional vector whose i-th entry is λ_{si} by $\boldsymbol{\lambda}_s$. Then, it is well known that the supplemented Markov process is stable if and only if

$$(I - P^{\mathrm{T}})^{-1} \boldsymbol{\lambda}_e < \boldsymbol{\lambda}_s, \tag{2.8}$$

where the inequality of vectors represents entry-wise inequalities (see, e.g. [59]).

2.3. Kingman's two-moment approximation in heavy traffic

In this and next sections, we consider the $GI/G/1$ queue, as a pilot example for diffusion approximation. As we mentioned in Section 1, Kingman [42] is the first one to get the approximation which corresponds with diffusion approximation. He directly considered the stationary distribution of the waiting times in the $GI/G/1$ queue, and derived a limiting distribution under a certain scaling, using the first two moments of the interarrival and service times. In his formulation, no continuous-time parameter is involved, which is different from those in the previous two sections. Nevertheless, his approximation agrees with that obtained from diffusion approximation.

Here two questions arise. Is a similar approximation possible for the stationary distribution of the queue length process $L(t)$? If so, how it is related to diffusion approximation ? To address those questions for the $GI/G/1$ queue, which is the single node case of GJN, we first summarize Kingman's [42] approximation for the stationary waiting time distribution, and make clear how approximation is produced from the limiting distribution. We then answer the questions in the next section.

We use the same notations of the GJN introduced in the previous two sections, but drop the index specifying a node for simplicity. For example, $L(t), R_e(t), R_s(t)$ and $W(t)$ stand for $L_1(t), R_{e1}(t), R_{s1}(t)$ and $W_1(t)$. We further drop the time index t for those characteristics in the steady state like L, R_e, R_s and W.

Let $D(n)$ be the waiting time, also called delay, of the n-th arriving customer. Obviously, we have

$$D(n) = W((T_e(1) + T_e(2) + \ldots + T_e(n))-), \qquad n = 1, 2, \ldots.$$

However, this expression is not so useful because continuous-time process $W(t)$ is involved. Instead of it, Kingman directly considers the stationary distribution, using so called Lindley's equation. Assume that $\rho \equiv \lambda_e/\lambda_s < 1$, and let D be a random variable subject to the stationary distribution of $D(n)$, then the distributional invariance of one step transition in the steady state yields Lindley's equation:

$$D \stackrel{w}{=} \max(0, D + T_s - T_e), \tag{2.9}$$

where D, T_s, T_e are independent, and $\stackrel{w}{=}$ stands for the equality in distribution. In this formulation, no continuous-time process is needed.

Kingman's [42] idea is to scale the waiting time D in the steady state by

$$\alpha = \frac{2(\lambda_s - \lambda_e)}{\lambda_e^2(\sigma_e^2 + \sigma_s^2)},$$

and to prove that αD weakly converges to the exponential distribution with unit mean, which is the stationary distribution of a one-dimensional SRBM, as $\alpha \downarrow 0$, under the finiteness conditions on some moments of T_e and T_s. See Theorem 1 of Köllerström [46] for a comprehensive proof. Based on this convergence, Kingman [41] suggests the following approximation for the stationary waiting time D.

$$\mathbb{P}(D \le x) \approx 1 - \exp\left(-\frac{2(\lambda_s - \lambda_e)}{\lambda_e^2(\sigma_e^2 + \sigma_s^2)}x\right), \qquad x \ge 0. \tag{2.10}$$

Here, it is notable to distinguish the following two steps.
(S1) Prove a weak convergence, that is, convergence in distribution, under appropriate changes of primitive data.
(S2) Derive an approximation from (S1).
(S1) is a purely mathematical issue, while (S2) is not more than proposal. Nevertheless, we have a clear rule for (S2) as we observed for the above D. Namely, the scaling factor is moved into the stationary distribution for approximation. When we are discussing about diffusion approximation, we often only consider (S1). This is because (S2) is automatically done according to the rule.

Kingman's as well as Köllerström's arguments are very clear, but lack the idea of a sequence of models for the convergence. Such a sequence makes clear how the primitive data are changed. For this, we here introduce a sequence of the $GI/G/1$ queues indexed by n, where the primitive data of the n-th system are denoted by $\{T_e^{(n)}(\ell)\}$ and $\{T_s^{(n)}(\ell)\}$. Denote the stationary waiting time of the n-th queue by $D^{(n)}$. Define $\alpha^{(n)}$ as

$$\alpha^{(n)} = \frac{2(\lambda_s^{(n)} - \lambda_e^{(n)})}{(\lambda_s^{(n)}\sigma_s^{(n)})^2 + (\lambda_e^{(n)}\sigma_e^{(n)})^2}, \qquad n = 1, 2, \ldots.$$

Then, Theorem 1 of Köllerström [46] reads as

$$\lim_{n\to\infty} \mathbb{P}(\alpha^{(n)}D^{(n)} \le x) = 1 - e^{-x}, \qquad x \ge 0, \tag{2.11}$$

if the following conditions are satisfied.

$$\lim_{n\to\infty} \alpha^{(n)} = 0, \qquad \alpha^{(n)} > 0, \quad n \geq 1, \tag{2.12}$$

$$\inf_{n\geq 1}((\sigma_s^{(n)})^2 + (\sigma_e^{(n)})^2) > 0, \tag{2.13}$$

$$\sup_{n\geq 1} \mathbb{E}((T_e^{(n)})^{2+\delta} + (T_s^{(n)})^{2+\delta})) < \infty, \qquad \text{for some } \delta > 0, \tag{2.14}$$

Because of (2.14), $(\sigma_s^{(n)})^2 + (\sigma_e^{(n)})^2$ are uniformly bounded in n, and therefore (2.12) is equivalent to

$$\lim_{n\to\infty}(\lambda_s^{(n)} - \lambda_e^{(n)}) = 0, \tag{2.15}$$

which is further equivalent to $\rho^{(n)} \equiv \lambda_e^{(n)}/\lambda_s^{(n)} \to 1$ if $\lambda_s^{(n)}$ is uniformly away from 0. Thus, we may divide (S1) into further two steps.

(S1-1) Set up a sequence of models indexed by positive integer n.

(S1-2) Scale the state space of the n-th system, and derive the limiting distribution.

In particular, if we choose $\lambda_u^{(n)}$ and $\sigma_u^{(n)}$ for $u = e, s$ in such a way that

$$\lim_{n\to\infty} \lambda_u^{(n)} = \lambda_u > 0, \qquad \lim_{n\to\infty} \sigma_u^{(n)} = \sigma_u > 0, \qquad u = e, s, \tag{2.16}$$

where λ_u and σ_u are positive constants, then (2.11) becomes

$$\lim_{n\to\infty} \mathbb{P}((1 - \rho^{(n)})D^{(n)} \leq x) = \lim_{n\to\infty} \mathbb{P}\left(\alpha^{(n)}D^{(n)} \leq \frac{\alpha^{(n)}}{1 - \rho^{(n)}}x\right)$$

$$= 1 - \exp\left(-\frac{2}{\lambda_e(\sigma_e^2 + \sigma_s^2)}x\right), \qquad x \geq 0, \tag{2.17}$$

since $\lambda_e = \lambda_s$. This shows that, in (S1-2), $D^{(n)}$ can be scaled by $1 - \rho^{(n)}$.

2.4. Extension of Kingman's approximation to the queue length

We next consider diffusion approximation for the stationary distribution of $L(t)$, which is the queue length including customers being served, that is, the number of customers in system. In this case, continuous-time is involved, and one may think that it would be easier to consider scaling a process rather than scaling a stationary distribution. This may be the reason why diffusion approximation on the queue length has been much less studied through the stationary equation, while process limits have been well studied as we shall see in Section 3.3.

We here challenge diffusion approximation for $L(t)$ through the stationary equation, and like to answer whether or not time scaling is really needed. Intuitively, no time scaling is needed because the stationary distribution is independent of time. Kingman's approximation supports this. However, this has not yet been confirmed, while process limit mandatorily requires state and time scaling to get a diffusion process.

To attack this problem, we use slightly stronger assumptions on the tails of the inter-arrival time T_e and the service time T_s. We here recall the simplified notation system introduced in the previous section, which drops the index for a node. Let $\widetilde{F}_e$ and $\widetilde{F}_s$ be the moment generating functions of T_e and T_s, respectively. That is,

$$\widetilde{F}_e(v) = \mathbb{E}(e^{vT_e}), \qquad \widetilde{F}_s(v) = \mathbb{E}(e^{vT_s}), \qquad v \in \mathbb{R}.$$

Assume the following light tail conditions.

(2d) There exist $\delta_e, \delta_s > 0$ such that

$$\widetilde{F}_u(v) < \infty, \quad v < \delta_u, \qquad \lim_{v \uparrow \delta_u} \widetilde{F}_u(v) = \infty, \qquad u = e, s. \tag{2.18}$$

These conditions are a bit stranger than what we actually need, but typically assumed to study large deviations (e.g., see [26]).

For $\theta, \xi \leq 0$, $\eta < \delta_e$ and some function $g(\theta)$ which will be determined later, let

$$X(t) = \begin{cases} g(\theta)e^{\eta R_e(t)}, & L(t) = 0, \\ e^{\theta L(t) + \eta R_e(t) + \xi R_s(t)}, & L(t) \geq 1, \end{cases}$$

and let N_d be the counting process for departing customers. Since $X(t)$ has jumps at instants when $N_e(t)$ or $N_d(t)$ is increased, it is obvious to see that

$$X(t) = X(0) + \int_0^t X'(u)du + \int_0^t \Delta X(u)(N_e(du) + N_d(du)), \tag{2.19}$$

where $\Delta X(u) = X(u+) - X(u-)$. Let $\mathbb{E}_e, \mathbb{E}_d$ be the expectations under the Palm distributions with respect to N_e, N_d. Namely, for $u = e, d$ and a continuous function f on $\mathbb{R}$,

$$\mathbb{E}_u(f(X(0))) = \lambda_u^{-1}\mathbb{E}\left(\int_0^1 f(X(u))N_u(ds) \right), \tag{2.20}$$

as long as the expectations are well defined, where $\lambda_d = \lambda_e$.

Assume that $(L(0), R_e(0), R_s(0))$ is subject to the stationary distribution, which is also assumed to exist. Then, taking the expectation of both sides of (2.19) yields

$$0 = \mathbb{E}(X'(0)) + \lambda_e\mathbb{E}_e(\Delta X(0)) + \lambda_d\mathbb{E}_d(\Delta X(0)),$$

where the expectations are all finite for $\theta, \xi \leq 0$, $\eta < \delta_e$. This formula is called a rate conservation law in the literature (e.g., see [51]). Hence, letting

$$\varphi(\theta, \eta, \xi) = \mathbb{E}(e^{\theta L(0) + \eta R_e(0) + \xi R_s(0)}), \quad \varphi_0(\eta) = \mathbb{E}(e^{\eta R_e(0)}1(L(0) = 0)),$$
$$\varphi_e(\theta, \xi) = \mathbb{E}_e(e^{\theta L(0-) + \xi R_s(0-)}), \qquad \varphi_{e0-}(0) = \mathbb{P}_e(L(0-) = 0),$$
$$\varphi_d(\theta, \eta) = \mathbb{E}_d(e^{\theta L(0-) + \eta R_e(0-)}), \qquad \varphi_{d0+}(\eta) = \mathbb{E}_d(e^{\eta R_e(0)}1(L(0+) = 0)),$$

we have the stationary equation for $\theta, \xi \leq 0$, $\eta < \delta_e$:

$$-(\eta + \xi)\varphi(\theta, \eta, \xi) + (\eta(1 - g(\theta)) + \xi)\varphi_0(\eta)$$
$$+ \lambda_e(e^\theta \widetilde{F}_e(\eta) - 1)\varphi_e(\theta, \xi) + \lambda_e(e^\theta \widetilde{F}_e(\eta)(\widetilde{F}_s(\xi) - 1) + 1 - g(\theta))\varphi_{e0-}(0)$$
$$+ \lambda_d(e^{-\theta} \widetilde{F}_s(\xi) - 1)\varphi_d(\theta, \eta) + \lambda_d e^{-\theta}(g(\theta) - \widetilde{F}_s(\xi))\varphi_{d0+}(\eta) = 0. \tag{2.21}$$

Choose η, ξ and $g(\theta)$ for $\theta \in \mathbb{R}$ in such a way that

$$e^\theta \widetilde{F}_e(\eta) = 1, \qquad e^{-\theta}\widetilde{F}_s(\xi) = 1, \qquad g(\theta) = e^\theta. \tag{2.22}$$

Then, they do exist by the assumption (2.18), and (2.21) boils down to

$$(\eta + \xi)\varphi(\theta, \eta, \xi) = (\eta(1 - e^\theta) + \xi)\varphi_0(\eta), \qquad \theta, \xi \leq 0, \ \eta < \delta_e, \tag{2.23}$$

where $\varphi(\theta, \eta, \xi)$ and $\varphi_0(\eta)$ are finite because $\theta, \xi \leq 0$ and

$$\mathbb{E}(e^{\eta R_e}) = \lambda_e \mathbb{E}\Big(\int_0^{T_e} e^{\eta(T_e - x)} dx \Big) = \frac{\lambda_e}{\eta} \big(\widetilde{F}_e(\eta) - 1 \big) < \infty, \qquad \eta < \delta_e,$$

by the cycle formula, where the singularity of the right-hand term at the origin can be removed as it can be seen from the middle term.

We next note that η and ξ have nice representations as functions of θ. For this, we define logarithmic moment generating functions for counting processes N_e and N_s as

$$\gamma_u(\theta) = \lim_{t \to \infty} \frac{1}{t} \log \mathbb{E}(e^{\theta N_u(t)}), \qquad u = e, s. \tag{2.24}$$

These functions play important roles in the theory of large deviations (e.g., see [22]). We here note the following fact, which is immediate from Theorem 1 of [26] (see also Lemma 4.1 of [53] in a more general context).

Lemma 2.1. *Under the assumption (2.18), let $\eta(\theta)$ and $\xi(\theta)$ be the solutions η and ξ of (2.22), then*

$$\eta(\theta) = -\gamma_e(\theta), \qquad \xi(\theta) = -\gamma_s(-\theta), \qquad \theta \in \mathbb{R}. \tag{2.25}$$

We here sketch a proof for the first equation of (2.25) to see the reason why it is obtained (see [26] for a complete proof). The second equation can be proved similarly. Define function h for a given constant α as

$$h(t) = \mathbb{E}\big(e^{\theta N_e(t) + \alpha t}\big), \qquad t \geq 0.$$

It is easy to see that h satisfies the following renewal equation.

$$h(t) = e^{\alpha t} \mathbb{P}(T_e > t) + \int_0^t h(t - s) e^{\theta + \alpha s} d\mathbb{P}(T_e \leq s).$$

If we choose α such that $\widetilde{F}_e(\alpha) e^\theta = 1$, then this renewal equation has a proper interval distribution, and therefore we can see that $h(t)$ converges to a finite constant by the key renewal theorem as long as $\widetilde{F}_e(t)$ is finite around a neighborhood of α. This is always the case under the assumption (2.18). On the other hand, from the definition of h,

$$\log h(t) = \log \mathbb{E}\big(e^{\theta N_e(t)}\big) + \alpha t.$$

Dividing both sides of this equation by t and letting t to infinity, we have (2.25).

For scaling L, we expand $\gamma_e(\theta)$ and $\gamma_s(-\theta)$ as polynomials of θ around the origin.

Lemma 2.2. *Under the same assumptions of Lemma 2.1, we have, as $\theta \to 0$,*

$$\gamma_e(\theta) = \lambda_e \theta + \frac{1}{2} \lambda_e^3 \sigma_e^2 \theta^2 + o(\theta^2), \tag{2.26}$$

$$\gamma_s(-\theta) = -\lambda_s \theta + \frac{1}{2} \lambda_s^3 \sigma_s^2 \theta^2 + o(\theta^2). \tag{2.27}$$

This lemma is easily obtained from (2.22) and (2.25). For convenience of the reader, we prove it in Appendix A.

We are now ready to consider a scaled limit for a sequence of the $GI/G/1$ queues, indexed by n. Assume that those queues are all stable, that is, $\rho^{(n)} < 1$ for all $n \geq 1$. As we did,

characteristics of the n-th queue are indexed by superscript (n). By Lemma 2.1, (2.23) for the n-th system can be written as

$$-(\gamma_e^{(n)}(\theta) + \gamma_s^{(n)}(-\theta))\varphi^{(n)}(\theta, -\gamma_e^{(n)}(\theta), -\gamma_s^{(n)}(-\theta))$$
$$= (\gamma_e^{(n)}(\theta)(e^\theta - 1) - \gamma_s^{(n)}(-\theta))\varphi_0^{(n)}(-\gamma_e^{(n)}(\theta)).$$

Divide both sides of this equation by θ and applying Lemma 2.2, we have

$$\left((\lambda_s^{(n)} - \lambda_e^{(n)}) - \frac{1}{2}((\lambda_e^{(n)})^3(\sigma_e^{(n)})^2 + (\lambda_s^{(n)})^3(\sigma_s^{(n)})^2)\theta + o(\theta)\right)$$
$$\times \varphi^{(n)}(\theta, -\lambda_e^{(n)}\theta + o(\theta), \lambda_s^{(n)}\theta + o(\theta))$$
$$= \left(\lambda_e^{(n)}(e^\theta - 1) + \lambda_s^{(n)} - \frac{1}{2}(\lambda_s^{(n)})^3(\sigma_s^{(n)})^2\theta + o(\theta)\right)$$
$$\times (1 - \rho^{(n)})\mathbb{E}(e^{(-\lambda_e^{(n)}\theta + o(\theta))R_e^{(n)}(0)}|L(0) = 0), \qquad (2.28)$$

where we have used the fact that $\mathbb{P}(L^{(n)} = 0) = 1 - \rho^{(n)}$.

Let us consider what is an appropriate scaling factor for $L^{(n)}$. Denote it by $\beta^{(n)} > 0$. That is, we scale $L^{(n)}$ as $\beta^{(n)}L^{(n)}$. This can be done by replacing θ by $\beta^{(n)}\theta$ in (2.28). We first assume that $\beta^{(n)} \to 0$ as $n \to \infty$ because we hardly expect to have a reasonable limiting distribution otherwise. We then examine (2.28), and easily see that we have a meaningful limiting distribution only when $\beta^{(n)}$ is proportional to $\lambda_s^{(n)} - \lambda_e^{(n)} > 0$. Under this limiting operation, it is preferable for the terms on the residual times to vanish. Namely, we like to verify the following conditions.

$$\lim_{n\to\infty} \mathbb{E}(e^{-\beta^{(n)}\theta R_e^{(n)}}) = 1, \qquad \text{for some } \theta < 0, \qquad (2.29)$$

$$\lim_{n\to\infty} \beta^{(n)}R_s^{(n)} = 0 \text{ almost surely.} \qquad (2.30)$$

Among these conditions, (2.30) is easily verified if the previous conditions (2.14) and (2.16) are satisfied because (2.14) implies that $\mathbb{E}(R_s^{(n)})$ is uniformly bounded in n. However, they may not be sufficient for (2.29). So far, we here use the following simple assumption.

$$\text{The distribution of } T_e^{(n)} \text{ is independent of } n. \qquad (2.31)$$

We now arrive at the following theorem, which corresponds to Kingman's approximation for the queue length in the steady state.

Theorem 2.1. *For a sequence of the stable $GI/G/1$ queues satisfying (2.18) for $\widetilde{F}_s = \widetilde{F}_s^{(n)}$, if the conditions (2.14), (2.15), (2.16) and (2.31) hold and if $\rho^{(n)} < 1$ for $n \geq 1$, then*

$$\lim_{n\to\infty} \mathbb{P}((1 - \rho^{(n)})L^{(n)} \leq x) = 1 - \exp\left(-\frac{2}{\lambda_e^2(\sigma_e^2 + \sigma_s^2)}x\right), \qquad x \geq 0. \qquad (2.32)$$

We already explained major ideas to have this theorem, but its full proof requires careful arguments. Since they are technical, we defer them into Appendix B. Note that the exponential distribution (2.32) is identical with the stationary distribution of the SRBM, which will given in Corollary 3.1.

Thus, Theorem 2.1 concludes that time scaling is not needed to get the limiting distribution corresponding to that of SRBM. We here only considered a single node queue, but can expect a similar result for a general GJN although we have to work on the stationary

equation instead of the stationary distribution because the latter is unknown for $d \geq 2$. Thus, we may have a completely different story about diffusion approximation for the stationary distribution of a queueing network in heavy traffic. As we will see in Section 4, this approach is closely related to a study on the tail decay rates of the stationary distributions.

In what follows, we will review another story, which requires scaling in both state and time.

3. Diffusion approximation for a single class GJN

We review diffusion approximation for a single class GJN introduced in Sections 2.1 and 2.2 by an SRBM, that is, semi-martingale reflecting Brownian motion on a nonnegative orthant. This is now standard, and can be found in a text book (e.g., see [12]). Nevertheless, it is good to see how the diffusion approximation is derived because it is a prototype for diffusion approximations for stochastic processes.

3.1. From a single class GJN to SRBM

We have observed that the supplemented Markov process may be useful to consider a limit of the stationary distributions. However, this is not the case for a limit of processes because an evolution in time needs to match scaling in state. Instead of it, the sample path characterizations (2.4) and (2.5) are known to be useful. To see this, let $R = I - P^{\mathrm{T}}$, and we rewrite (2.4) as

$$\boldsymbol{L}(t) = \boldsymbol{L}(0) + \boldsymbol{X}(t) + R\boldsymbol{Y}(t), \qquad t \geq 0, \tag{3.1}$$

where the i-th entry of $\boldsymbol{X}(t)$ is given by

$$X_i(t) = N_{ei}(t) - \lambda_{ei}t + \sum_{j \in J} \big(N_{sj}(U_j(t)) - \lambda_{sj}U_j(t)\big)p_{ji} - \big(N_{si}(U_i(t)) - \lambda_{si}U_i(t)\big)$$
$$+ \sum_{j \in J} \big(\Phi_{ji}(N_{sj}(U_j(t))) - N_{sj}(U_j(t))p_{ji}\big) + \Big(\lambda_{ei} + \sum_{j \in J} \lambda_{sj}p_{ji} - \lambda_{si}\Big)t, \tag{3.2}$$

and the i-th entry of $\boldsymbol{Y}(t)$ is given by

$$Y_i(t) = \lambda_{si}(t - U_i(t)). \tag{3.3}$$

Clearly, $t - U_i(t)$ is the total idle time of the server at node i by time t, and therefore $Y_i(t)$ is non-decreasing in t. Furthermore, $Y_i(t)$ increases only when $L_i(t) = 0$. These are precisely what are meant by (2.5), and we have

$$\int_0^t L_i(u)dY_i(u) = 0, \qquad i \in J. \tag{3.4}$$

Thus, (2.4) and (2.5) imply (3.1) and (3.4). Conversely, for a given $\boldsymbol{X}(\cdot)$, $(\boldsymbol{L}(\cdot), \boldsymbol{Y}(\cdot))$ is uniquely determined by (3.1) and (3.4) as will be given in Theorem 3.1 below. Thus, we have the reflection mapping from $\boldsymbol{X}(\cdot)$ to $(\boldsymbol{L}(\cdot), \boldsymbol{Y}(\cdot))$, which is denoted by $(\phi(\boldsymbol{X}), \Psi(\boldsymbol{X}))$. This mapping will be used to study process limit for diffusion approximation.

A basic idea for diffusion approximation is to replace $\boldsymbol{X}(t)$ in (3.1) by a d-dimensional Brownian motion $\boldsymbol{B}(t)$ with a drift vector $\boldsymbol{\mu}$ satisfying $\boldsymbol{B}(0) = \boldsymbol{0}$. Namely,

$$\boldsymbol{X}(t) = \boldsymbol{B}(t) + \boldsymbol{\mu}t. \tag{3.5}$$

Roughly speaking, $\{\boldsymbol{B}(t)\}$ is an independent incremental process with continuous sample path. The reader who is not familiar with Brownian motion may refer to Section 5.3 of [12] for a short summary. Thorough introductions are found in [2, 38, 60].

For this $\boldsymbol{X}(t)$, we rewrite (3.1) and (3.4) for a nonnegative d-dimensional process $\boldsymbol{Z}(t)$ and a nondecreasing d-dimensional process $\boldsymbol{Y}(t)$ as

$$\boldsymbol{Z}(t) = \boldsymbol{Z}(0) + \boldsymbol{X}(t) + R\boldsymbol{Y}(t), \tag{3.6}$$

$$\int_0^t Z_i(u)dY_i(u) = 0, \qquad i \in J, \qquad t \geq 0. \tag{3.7}$$

Let $D(\mathbb{R}^d)$ be the set of functions from $[0,\infty)$ to $\mathbb{R}^d$ which are right continuous and have left-hand limit. In general, finding a solution $(\boldsymbol{Z}(\cdot), \boldsymbol{Y}(\cdot))$ in $D(\mathbb{R}^{2d})$ satisfying (3.6) and (3.7) for a given $\boldsymbol{Z}(0)$ and $\boldsymbol{X}(\cdot)$ with $\boldsymbol{X}(0) = \boldsymbol{0}$ is called a Skorohod problem. The following theorem answers this problem for $R = I - P^{\mathrm{T}}$.

Theorem 3.1 (Theorem 1 of [24], Theorem 7.2 of [12])**.** *Let $R = I - P^{\mathrm{T}}$. For each function $\boldsymbol{x}(\cdot) \in D(\mathbb{R}^d)$, there exists a unique pair of nonnegative function $\boldsymbol{z}(\cdot)$ and a nondecreasing function $\boldsymbol{y}(\cdot)$ with $\boldsymbol{y}(0) = \boldsymbol{0}$ which satisfy*

$$\boldsymbol{z}(t) = \boldsymbol{x}(t) + R\boldsymbol{y}(t), \tag{3.8}$$

$$\int_0^t z_i(u)dy_i(u) = 0, \qquad i \in J, \qquad t \geq 0. \tag{3.9}$$

Furthermore, the mapping ϕ and Ψ from $\boldsymbol{x}(\cdot)$ to $\boldsymbol{z}(\cdot) = \phi(\boldsymbol{x})$ and $\boldsymbol{y}(\cdot) = \Psi(\boldsymbol{x})$ are Lipschitz continuous. Namely, there exist constants C_1 and C_2 for each t such that, for $H_1 = \phi$ and $H_2 = \Psi$,

$$\sup_{0 \leq u \leq t} \|H_k(\boldsymbol{x}_1)(u) - H_k(\boldsymbol{x}_2)(u)\| \leq C_k \sup_{0 \leq u \leq t} \|\boldsymbol{x}_1(u) - \boldsymbol{x}_2(u))\|, \qquad k = 1, 2,$$

for $\boldsymbol{x}_1(\cdot), \boldsymbol{x}_2(\cdot) \in D(\mathbb{R}^d)$, where $\|\cdot\|$ is the Euclidean metric in the vector space $\mathbb{R}^d$.

If we choose $\boldsymbol{X}(\cdot)$ of (3.5) and take $\{\boldsymbol{Z}(0) + \boldsymbol{X}(t)\}$ for $\boldsymbol{x}(\cdot)$ in this theorem, we have a unique solution $(\boldsymbol{Z}(\cdot), \boldsymbol{Y}(\cdot))$ of (3.6) and (3.7). Here, it is important for R to be $I - P^{\mathrm{T}}$, which is always the case for GJN. However, some other models may require more general matrix R, which is indeed the case for a multi-class queueing network (see Section 5). Thus, we first define $(\boldsymbol{Z}(\cdot), \boldsymbol{Y}(\cdot))$ for a general square matrix R.

Definition 3.1. *Continuous d-dimensional processes $\{\boldsymbol{Z}(t); t \geq 0\}$ and $\{\boldsymbol{Y}(t); t \geq 0\}$ are called a semimartingale reflecting Brownian motion, SRBM for short, and a regulator, respectively, if there exists a family of probability measures $\{\mathbb{P}_{\boldsymbol{x}}; \boldsymbol{x} \in \mathbb{R}_+^d\}$ on some filtration $\{\mathcal{F}_t\}$ and the following conditions are satisfied for $\boldsymbol{Z}(0) = \boldsymbol{x}$ under $\mathbb{P}_{\boldsymbol{x}}$-almost surely for every $\boldsymbol{x} \in \mathbb{R}_+^d$.*

(i) $\{\boldsymbol{X}(t)\}$ is a d-dimensional Brownian motion with drift vector $\boldsymbol{\mu}$, that is, given by (3.5), and $\{\boldsymbol{B}(t)\}$ is $\{\mathcal{F}_t\}$-martingale.

(ii) $\{\boldsymbol{Y}(t)\}$ is a $\{\mathcal{F}_t\}$-adapted nondecreasing process with $\boldsymbol{Y}(0) = \boldsymbol{0}$.

(iii) $\{\boldsymbol{Z}(t)\}$ is a $\{\mathcal{F}_t\}$-adapted process.

(iv) (3.6) and (3.7) are satisfied.

To use this definition, it is important to see when $\boldsymbol{Z}(\cdot)$ and $\boldsymbol{Y}(\cdot)$ uniquely exist. From Theorem 3.1, we can see that, if $R = I - P^{\mathrm{T}}$, then they uniquely exist, which is independent of a choice of $\mathbb{P}_{\boldsymbol{x}}$. This existence problem was completely solved by Reiman and Williams [57] for necessity and by Taylor and Williams [61] for sufficiency. The answer is:

Theorem 3.2. *The SRBM $\boldsymbol{Z}(\cdot)$ uniquely exists if and only if R is a completely-$\mathcal{S}$ matrix, which is a square matrix whose all principal submatrices are $\mathcal{S}$-matrices, where an ℓ-dimensional square matrix A is called an $\mathcal{S}$-matrix if there is an ℓ-dimensional vector $\boldsymbol{v} > 0$ such that $A\boldsymbol{v} > \boldsymbol{0}$.*

It is easy to see that $I - P^{\mathrm{T}}$ is a special case of a completely-$\mathcal{S}$ matrix because we can choose $(I - P^{\mathrm{T}})^{-1}\mathbf{1}$ for the $\boldsymbol{v}$, where $\mathbf{1}$ is the d-dimensional vector whose entries are all units.

In the remaining part of this section, we assume that R is a completely $\mathcal{S}$-matrix, so the SRBM $\{\boldsymbol{Z}(t)\}$ exists. We denote the covariance matrix of the Brownian motion $\boldsymbol{B}(t)$ by $\Sigma \equiv \{\sigma_{ij}; i, j \in J\}$, and assume that Σ is non-singular. Thus, $(\Sigma, \boldsymbol{\mu}, R)$ is the primitive data. Clearly, this SRBM is a Markov process with respect to filtration $\{\mathcal{F}_t\}$, and its dynamics is specified by Itô integration formula. Namely, let $C_b^2(\mathbb{R}^d)$ be the set of twice continuously differentiable and bounded functions from $[0, \infty)$ to $\mathbb{R}^d$, then, for $f \in C_b^2(\mathbb{R}^d)$, computing the difference $f(\boldsymbol{Z}(u)) - f(\boldsymbol{Z}(0))$ by the integration with respect to the integrator $d(\boldsymbol{X}(u) + R\boldsymbol{Y}(u))$ yields the following formula almost surely.

$$
f(\boldsymbol{Z}(u)) - f(\boldsymbol{Z}(0)) = \int_0^t (\nabla f(\boldsymbol{Z}(u)))^{\mathrm{T}}(\boldsymbol{\mu}\, du + d\boldsymbol{B}(u))
$$
$$
+ \int_0^t \mathcal{L}f(\boldsymbol{Z}(u))\, du + \int_0^t (\nabla f(\boldsymbol{Z}(u))^{\mathrm{T}} R\, d\boldsymbol{Y}(u), \qquad t \geq 0, \quad (3.10)
$$

where

$$
\nabla f(\boldsymbol{x}) = \Big(\frac{\partial}{\partial x_1} f(\boldsymbol{x}), \dots, \frac{\partial}{\partial x_d} f(\boldsymbol{x}) \Big)^{\mathrm{T}}, \qquad \mathcal{L}f(\boldsymbol{x}) = \frac{1}{2} \sum_{i,j \in J} \sigma_{ij} \frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x}).
$$

If one is not familiar with Itô integration, he may have a trouble with the second integration term in (3.10). This term is so called an integration by quadratic variations, which arise from unbounded variation of the sample path of a Brownian motion. On the other hand, the integration with respect to $d\boldsymbol{B}(u)$ in the first integration is just a Riemann integral, whose expectation vanishes. We now have nice text books for Itô integration. For example, Chung and Williams [15] is recommended for first reading.

We next consider a stable SRBM, that is, an SRBM which has a stationary distribution. The stability condition has not yet been completely obtained in terms of the primitive data for $d \geq 4$ (see [7]). However, if R is an $\mathcal{M}$-matrix, which is a matrix of the form $I - G$ for some nonnegative matrix G such that $I - G$ is invertible, then the SRBM is stable if and only if

$$
R^{-1}\boldsymbol{\mu} < \boldsymbol{0}. \tag{3.11}
$$

In particular, this is the stability condition for $R = I - P^{\mathrm{T}}$ since this R is an $\mathcal{M}$-matrix.

We return to a general R which is completely $\mathcal{S}$, and assume that the SRBM is stable. We denote its stationary distribution by π, and define measure ν_i on $\mathbb{R}_+^d$ as

$$
\nu_i(A) = \mathbb{E}_\pi\Big(\int_0^1 1(\boldsymbol{Z}(u) \in A) dY_i(u) \Big), \qquad A \in \mathcal{B}(\mathbb{R}_+^d),\ i \in J, \tag{3.12}
$$

where $\mathcal{B}(\mathbb{R}_+^d)$ is the Borel σ-field on $\mathbb{R}_+^d$, and $\mathbb{E}_\pi$ stands for the expectation under π, that is, under the condition that $\boldsymbol{Z}(0)$ is subject to π. Obviously, ν_i is a finite measure because $\mathbb{E}(X_i(t) - X_i(0))$ is bounded for $t \in [0, 1]$. In the literature (e.g., see [38]), the measure ν_i is

called a Palm measure with respect to the random measure generated by $Y_i(\cdot)$. Since $Y_i(t)$ does not increases for $Z_i(t) > 0$, this measure concentrates on the face F_i given by

$$F_i = \{\boldsymbol{x} \in \mathbb{R}^d_+; x_i = 0\}.$$

Let us take the expectation of (3.10) under π, then it is immediate to have the part (a) of the following theorem.

Theorem 3.3. *Let R be a completely $\mathcal{S}$-matrix. (a) If the SRBM has the stationary distribution π, then, for $f \in C^2_b(\mathbb{R}^d)$,*

$$\int_{\mathbb{R}^d_+} \mathcal{L}f(\boldsymbol{x})\pi(d\boldsymbol{x}) + \int_{\mathbb{R}^d_+} \langle \nabla f(\boldsymbol{x}), \boldsymbol{\mu} \rangle \pi(d\boldsymbol{x}) + \sum_{i \in J} \int_{\mathbb{R}^d_+} \langle \nabla f(\boldsymbol{x}), R^{[i]} \rangle \nu_i(d\boldsymbol{x}) = 0, \qquad (3.13)$$

where $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ is the inner product $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}$ of vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, and $R^{[i]}$ is the i-th column of R. (b) Conversely, if there are a probability distribution π on $(\mathbb{R}^d_+, \mathcal{B}(\mathbb{R}^d_+))$ and finite measures ν_i on $(F_i, \mathcal{B}(F_i))$ for $i \in J$, then π is the stationary distribution of the SRBM, and (3.12) holds.

Remark 3.1. *(3.13) for all $f \in C^2_b(\mathbb{R}^d)$ is equivalent to*

$$\left(\frac{1}{2} \langle \boldsymbol{\theta}, \Sigma \boldsymbol{\theta} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle \right) \tilde{\varphi}(\boldsymbol{\theta}) + \sum_{i \in J} \langle R^{[i]}, \boldsymbol{\theta} \rangle \tilde{\varphi}_i(\boldsymbol{\theta}) = 0, \qquad \boldsymbol{\theta} \le \boldsymbol{0}, \qquad (3.14)$$

where $\tilde{\varphi}(\boldsymbol{\theta})$ and $\tilde{\varphi}_i(\boldsymbol{\theta})$ are the moment generating functions of the stationary distribution π and the Palm measure ν_i, respectively. (3.14) is immediate from (3.13) with $f(\boldsymbol{x}) = e^{\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle}$ for $\boldsymbol{x} \in \mathbb{R}^d_+$. The converse is also easy to see at least intuitively because Laplace transform uniquely determines a distribution and measure on the nonnegative orthant. A detailed proof for the converse is given in Appendix D of the arXiv version [19].

This equation (3.13) is called a basic adjoint relationship, BAR for short (e.g., see [35]). Since (3.13) is nothing but a stationary equation, (b) must be true intuitively. However, its proof is technically quite involving (see [17, 39]).

3.2. Basic tips for process limit

For diffusion approximation, we will consider a sequence of stochastic processes which converges to some stochastic process. Recall that this limit is called a process limit. To formally define this convergence, we need topology or metric on the set of functions. We summarize some basic tips for them. The reader may skip this subsection if it is familiar with them. The materials here can be found in standard text books [2, 12, 38, 64]. The book chapter [25] is a good reference for OR readers.

Let S be a complete and separable normed space with norm $\|\cdot\|_S$, and let $C(E, S)$ $(D(E, S))$ be the set of all functions from an interval $E \subset \mathbb{R}_+$ to S which are continuous (are right-continuous and have left-hand limits, respectively). For $E = \mathbb{R}_+$, we omit E, so $C(\mathbb{R}_+, S)$ and $D(\mathbb{R}_+, S)$ are simply denoted by $C(S)$ or $D(S)$, which we already used for $C(\mathbb{R}^d)$ or $D(\mathbb{R}^d)$ in Section 3.1.

Stochastic processes which we consider are functions from the sample space Ω to either $C(S)$ or $D(S)$. In particular, a sample path of a Brownian motion belongs to $C(S)$ with $S = \mathbb{R}^d$, while a sample path for the GJN belongs to $D(S)$. Thus, we need topology or metric on both $C(S)$ and $D(S)$ to consider diffusion approximation. They can be defined through metrics on $C([0, t], S)$ and $D([0, t], S)$ for each $t > 0$. Let

$$\|x - y\|_t = \sup_{u \in [0, t]} \|x(u) - y(u)\|_S, \qquad x, y \in C([0, t], S) \text{ or } D([0, t], S),$$

which is called a uniform norm. $C([0,t],S)$ typically uses this uniform norm as its metric. On the other hand, $D([0,t],S)$ uses J_1-metric, which is defined as

$$d_{J_1}(x,y,t) = \inf_{\lambda \in \Lambda(t)} \max(\|x \circ \lambda - y\|_t, \|\lambda - \lambda_0\|_t), \qquad x, y \in D([0,t], S),$$

where $\Lambda(t)$ is the set of strictly increasing and continuous function from $[0,t]$ to $[0,t]$, $x \circ \lambda(u) = x(\lambda(u))$ and $\lambda_0(u) = u$ for $u \in [0,t]$. As one can see from its definition, J_1-metric is to minimize discontinuity of functions by a suitable time scaling.

Let $X_n(\cdot) \equiv \{X_n(t); t \geq 0\}$ for $n = 1, 2, \ldots$ and $X(\cdot) \equiv \{X(t); t \geq 0\}$ be stochastic processes whose sample paths are in $D(S)$. Then, $X_n(\cdot)$ is said to converge to $X(\cdot)$ in distribution, which is denoted by $X_n(\cdot) \xrightarrow{w} X(\cdot)$, if, for any bounded continuous function f from $D(S)$ to $\mathbb{R}$,

$$\lim_{n \to \infty} \mathbb{E}(f(X_n(\cdot))) = \mathbb{E}(f(X(\cdot))). \tag{3.15}$$

We also refer to this convergence as weak convergence.

The following theorem is an immediate consequence of these definitions.

Theorem 3.4 (Continuous mapping theorem)**.** *Let g be a continuous mapping from $D(S)$ to $D(S)$, and assume that $X_n(\cdot) \xrightarrow{w} X(\cdot)$ for the sequence of stochastic processes $X_n(\cdot)$ and that of $X(\cdot)$ whose sample paths in $D(S)$, then we have*

$$g(X_n)(\cdot) \xrightarrow{w} g(X)(\cdot).$$

Another useful tool is to replace convergence in distribution by almost surely convergence. For this, sample paths for the sequence of stochastic processes must be suitably chosen, but it is quite useful to prove the weak convergence of functionals of those stochastic processes. A random variable version of this theorem is not hard to prove (e.g., see [3]), but its proof is quite technical for stochastic processes.

Theorem 3.5 (Skorohod representation theorem, Theorem 6.7 of [2])**.** *For the sequence of stochastic processes $X_n(\cdot)$ and that of $X(\cdot)$ whose sample paths are in $D(S)$, if $X_n(\cdot) \xrightarrow{w} X(\cdot)$, then there exist stochastic processes $\tilde{X}_n(\cdot)$ and $\tilde{X}(\cdot)$ such that they have the same distributions as those of $X_n(\cdot)$ and $X(\cdot)$, respectively, and $\tilde{X}_n(\cdot)$ almost surely converges to $\tilde{X}(\cdot)$ as $n \to \infty$.*

In general, verifying the convergence condition (3.15) is not easy. A typical situation is to use a finite dimensional convergence of $X_n(\cdot)$ in distribution. Namely, for each $k \geq 1$, each sequence $t_1, t_2, \ldots, t_k \in \mathbb{R}_+$ and each bounded continuous function f_k from S^k to $\mathbb{R}$,

$$\lim_{n \to \infty} \mathbb{E}(f_k(X_n(t_1), X_n(t_2), \ldots, X_n(t_k))) = \mathbb{E}(f_k(X(t_1), X(t_2), \ldots, X(t_k))).$$

Clearly, $X_n(\cdot) \xrightarrow{w} X(\cdot)$ implies this finite dimensional convergence in distribution, but its converse is generally not true although the limiting distribution of $X_n(\cdot)$ is uniquely identified if it exists (e.g., see [25]). For the converse to be true, an extra condition is required. A typical condition for this is tightness. The sequence of $X_n(\cdot)$ is said to be tight if, for any $\epsilon > 0$, there exists a compact set K of $D(S)$ such that

$$\liminf_{n \to \infty} \mathbb{P}(X_n(\cdot) \in K) > 1 - \epsilon.$$

This tightness also may not be easy to verify. So, various sufficient conditions have been studied in the literature (e.g., see [2, 38]). Those are quite technical, and we will not get into their details.

Another useful technique is a random change of time.

Lemma 3.1 (Lemma on page 151 of [2]). *Let $\{X^{(n)}(\cdot); n = 1, 2, \ldots\}$ be the sequence of processes whose sample paths are in $D(S)$, and let $A^{(n)}(\cdot)$ be the sequence of non-decreasing non-negative valued processes. If there are $X^{(n)}(\cdot) \in D(S)$ and nondecreasing $A^{(n)}(\cdot)$ such that $(X^{(n)}(\cdot), A^{(n)}(\cdot)) \xrightarrow{w} (X(\cdot), A(\cdot))$ and if $\mathbb{P}(X(\cdot) \in C(S)) = 1$, then $X^{(n)}(\cdot) \circ A^{(n)}(\cdot) \xrightarrow{w} X(\cdot) \circ A(\cdot)$, where the composition $f \circ g$ is defined as $f \circ g(x) = f(g(x))$.*

We finally refer to process versions of the central limit theorems (called Lindeberg-Feller Theorems, e.g., see Theorem 7.2.1 of [14]). They are all sources of diffusion approximation.

Theorem 3.6 (An extended version of Donsker's theorem (e.g., see Problem 8.4 of [2])). *For each $n \geq 1$, let $\tau_i^{(n)}$, $\tau_2^{(n)}$, $\ldots$ be independent and identically distributed random variables with finite mean $m^{(n)}$ and finite variance $(\sigma^{(n)})^2$. Assume that*

$$\lim_{n \to \infty} \sigma^{(n)} = \sigma > 0, \tag{3.16}$$

$$\sup_{n \geq 1} \mathbb{E}((\tau^{(n)})^{2+\delta}) < \infty, \qquad \exists \delta > 0, \tag{3.17}$$

and define $\widehat{X}^{(n)}(t)$ as

$$\widehat{X}^{(n)}(t) = \frac{1}{\sqrt{n}\sigma^{(n)}} \sum_{\ell=1}^{[nt]} (\tau_\ell^{(n)} - m^{(n)}),$$

then $\widehat{X}^{(n)}(\cdot) \xrightarrow{w} B(\cdot)$, where $[a]$ is the largest integer not greater than real number a, and $B(\cdot)$ is the standard Brownian motion.

Remark 3.2. *The conditions (3.16) and (3.17) guarantees for Lindeberg condition to hold (see Theorem 27.3 of [3]).*

A similar scaling limit can be obtained for the counting process $N^{(n)}$, which is defined as

$$N^{(n)}(t) = \sup \left\{ \ell \geq 0; \sum_{k=1}^{\ell} \tau_k^{(n)} \leq t \right\}.$$

The next theorem is obtained from Theorem 3.6 using Lemma 3.1 (e.g., see [25] for details).

Theorem 3.7 (Counting process version of Donsker's theorem, Theorem 14.6 of [2]). *Assume that there exist positive and finite m and σ such that $\widehat{X}^{(n)}(\cdot) \xrightarrow{w} B(\cdot)$ holds under the assumptions of Theorem 3.6. Let $\lambda^{(n)} = 1/m^{(n)}$. Assume that*

$$\lim_{n \to \infty} \lambda^{(n)} = \lambda,$$

and define

$$\widehat{Z}^{(n)}(t) = \frac{1}{\sqrt{n}\sigma^{(n)}(\lambda^{(n)})^{\frac{3}{2}}} \left(N^{(n)}(nt) - nt\lambda^{(n)} \right), \qquad t \geq 0,$$

then $\widehat{Z}^{(n)}(\cdot) \xrightarrow{w} B(\cdot)$.

From Lemma 2.2, we can see that the variance $V(N^{(n)}(t)) \sim (\lambda^{(n)})^3(\sigma^{(n)})^2 t$ as $t \to \infty$. This is the reason why $\sigma^{(n)}(\lambda^{(n)})^{\frac{3}{2}}$ appears in the normalizing factor above. For convenience of the reader, we directly verify this fact in Appendix C.

3.3. Process limit for diffusion approximation

We now consider a process limit of the joint queue length process $\{\boldsymbol{L}(t)\}$ of the GJN under suitable scaling in time and state. A key step for obtaining this limit is to change $X_i(t)$ of (3.2) to the following $\underline{X}_i(t)$ by replacing accumulated busy times $U_i(t)$ and $U_j(t)$ by deterministic time t.

$$\underline{X}_i(t) = N_{ei}(t) - \lambda_{ei}t + \sum_{j \in J} \left(N_{sj}(t) - \lambda_{sj}t\right)(p_{ji} - \delta_{ij})$$
$$+ \sum_{j \in J} \left(\Phi_{ji}(N_{sj}(t)) - N_{sj}(t)p_{ji}\right) + \left(\lambda_{ei} + \sum_{j \in J} \lambda_{sj}p_{ji} - \lambda_{si}\right)t, \qquad (3.18)$$

where $\delta_{ij} = 1(i = j)$. We denote the vector whose i-th coordinate is $\underline{X}_i(t)$ by $\underline{\boldsymbol{X}}(t)$.

One can guess that the process $\{\underline{\boldsymbol{X}}(t)\}$ converges to a Brownian motion under normalization. For this, we require additional conditions on the finiteness of variances. Denote generic random variables which have the same distributions as $T_{ei}(\ell)$ and $T_{si}(\ell)$ respectively by T_{ei} and T_{si}. In what follow, we assume:

(3a) T_{ei} and T_{si} have finite variances, which are denoted respectively by σ_{ei}^2 and σ_{si}^2, where we let $\sigma_{ei} = 0$ for $i \in J \setminus J_e$, and

$$\sum_{i \in J} (\sigma_{ei}^2 + \sigma_{si}^2) > 0. \qquad (3.19)$$

The condition (3.19) is just to exclude a trivial case in a simple way, but can be weakened as we will see later.

We next introduce the sequence of the GJN's for diffusion approximation. We index those GJN's by positive integer n. For the n-th GJN, all random elements as well as constant parameters are indexed by n such as $T_{ei}^{(n)}(\ell)$, $T_{si}^{(n)}(\ell)$, $\lambda_{ei}^{(n)}$, $\boldsymbol{L}^{(n)}(t)$, $\boldsymbol{X}^{(n)}(t)$ and so on, while keeping the non-indexed notations for the original GJN satisfying the assumptions (2a)–(2c) and (3a). In particular, we will not change the routing matrix P, so R is unchanged. We define the n-th GJN by

$$T_{ei}^{(n)}(\ell) = T_{ei}(\ell), \qquad i \in J_e \qquad (3.20)$$

$$T_{si}^{(n)}(\ell) = T_{si}(\ell) - m_{si} + \left(1 - \frac{1}{\sqrt{n}}\right)\lambda_{ai}^{-1}, \qquad i \in J, \qquad (3.21)$$

keeping the other primitive data, where we recall that $\lambda_{ai}(\equiv \lambda_{ai}^{(n)})$ is the total arrival rate at node i, which is obtained by (2.7). Then, it is easy to see that

$$\sqrt{n}\left(\lambda_{ai}^{(n)} - \lambda_{si}^{(n)}\right) = -\lambda_{si} < 0, \qquad i \in J. \qquad (3.22)$$

On the other hand, the variances of $T_{ei}^{(n)}$ and $T_{si}^{(n)}$ are unchanged, that is,

$$\sigma_{ei}^{(n)} = \sigma_{ei}, \qquad \sigma_{si}^{(n)} = \sigma_{si}, \quad i \in J.$$

Instead of those assumptions, one can choose $\lambda_{ei}^{(n)}, \lambda_{si}^{(n)}, \sigma_{ei}^{(n)}, \sigma_{si}^{(n)}$ as long as the following two sets of conditions are satisfied (e.g., see [56]).

(3b) $\lim_{n \to \infty} \lambda_{ei}^{(n)} = \lambda_{ei}$, $\lim_{n \to \infty} \lambda_{si}^{(n)} = \lambda_{si}$, $\lim_{n \to \infty} \sigma_{ei}^{(n)} = \sigma_{ei}$, $\lim_{n \to \infty} \sigma_{si}^{(n)} = \sigma_{si}$, and $\lim_{n \to \infty} \sqrt{n}\left(\lambda_{ai}^{(n)} - \lambda_{si}^{(n)}\right) = c_i$ for some $c_i \in \mathbb{R}$.

(3c) $\sup_{n \geq 1} \mathbb{E}((T_{ei}^{(n)})^{2+\delta} 1(i \in J_e) + (T_{si}^{(n)})^{2+\delta}) < \infty$ for some $\delta > 0$.

In what follows we simply assume (3.20) and (3.21). Obviously, (3b) is satisfied for this case, and (3c) is not needed because Lindeberg condition for the central limit theorem is automatically satisfied in this case. For the n-th GJN, we define diffusion scaling processes by

$$\widehat{\boldsymbol{L}}^{(n)}(t) = \frac{1}{\sqrt{n}} \boldsymbol{L}^{(n)}(nt), \qquad \widehat{\boldsymbol{Y}}^{(n)}(t) = \frac{1}{\sqrt{n}} \boldsymbol{Y}^{(n)}(nt),$$

$$\widehat{\boldsymbol{X}}^{(n)}(t) = \frac{1}{\sqrt{n}} \boldsymbol{X}^{(n)}(nt), \qquad \underline{\widehat{\boldsymbol{X}}}^{(n)}(t) = \frac{1}{\sqrt{n}} \underline{\boldsymbol{X}}^{(n)}(nt).$$

We would like to show that $\{\widehat{\boldsymbol{X}}^{(n)}(t)\}$ has the same limiting process as that of $\{\underline{\widehat{\boldsymbol{X}}}^{(n)}(t)\}$. For this, we use the following fact.

Lemma 3.2 (Proposition 4 of [56]). $\overline{U}_i^{(n)}(t) \equiv n^{-1} U_i^{(n)}(nt)$ *converges to* t *in probability as* $n \to \infty$.

By this lemma and Lemma 3.1, we have

$$\widehat{X}_i^{(n)}(t) = \frac{1}{\sqrt{n}} \underline{X}_i^{(n)}(U_i^{(n)}(nt)) = \frac{1}{\sqrt{n}} \underline{X}_i^{(n)}(n\overline{U}_i^{(n)}(t)) \simeq \underline{\widehat{X}}_i^{(n)}(t), \quad n \to \infty, \qquad (3.23)$$

where $\simeq$ stands to have the same distribution asymptotically as $n \to \infty$.

Thus, for $\widehat{\boldsymbol{L}}^{(n)}(t)$ to weakly converges to an SRBM, it remains to show that

$$\underline{\widehat{X}}_i^{(n)}(t) = \frac{1}{\sqrt{n}} \big(N_{ei}^{(n)}(nt) - \lambda_{ei}^{(n)} nt \big) + \sum_{j \in J} \frac{1}{\sqrt{n}} \big(N_{sj}^{(n)}(nt) - \lambda_{sj}^{(n)} nt \big) (p_{ji} - \delta_{ji})$$

$$+ \sum_{j \in J} \frac{1}{\sqrt{n}} \big(\Phi_{ji}(N_{sj}^{(n)}(nt)) - N_{sj}^{(n)}(nt) p_{ji} \big) + \sqrt{n} \Big(\lambda_{ei}^{(n)} + \sum_{j \in J} \lambda_{sj}^{(n)} p_{ji} - \lambda_{si}^{(n)} \Big) t$$

weakly converges to a d-dimensional Brownian motion with a constant drift term, where $\lambda_{ei}^{(n)} = \lambda_{ei}$ by (3.20), but we keep the notation $\lambda_{ei}^{(n)}$ for arguments to be parallel for service times. Letting

$$\widehat{N}_{ei}^{(n)}(t) \equiv \frac{1}{\sqrt{n}} \big(N_{ei}^{(n)}(nt) - \lambda_{ei}^{(n)} nt \big),$$

it follows from Theorem 3.7 that

$$\widehat{N}_{ei}^{(n)}(\cdot) \xrightarrow{w} \sqrt{\lambda_{ei}^3 \sigma_{ei}^2} B_{ei}^*(\cdot), \qquad (3.24)$$

where $\{B_{ei}^*(t)\}$ is the standard Brownian motion. We here choose $\{B_{ei}^*(t)\}$ to be independent for $i \in J_e$.

Let $\widehat{\boldsymbol{N}}_e^{(n)}(t)$ is the random vector whose i-th coordinate is $\widehat{N}_{ei}^{(n)}(t)$. We similarly define $\widehat{\boldsymbol{N}}_s^{(n)}(t)$. It is immediate from (3.24) that

$$\widehat{\boldsymbol{N}}_e^{(n)}(\cdot) \xrightarrow{w} \mathrm{diag}(\{\lambda_{ei}^{\frac{3}{2}} \sigma_{ei}\}) \boldsymbol{B}_e^*(\cdot), \qquad (3.25)$$

where $\mathrm{diag}(\{a_i\})$ is the diagonal matrix whose i-th diagonal entry is a_i, and $\boldsymbol{B}_e^*(\cdot)$ is the d-dimensional standard Brownian motion, which means that its covariance matrix is the

identity matrix. Note that the right-hand side of (3.25) is $|J_e|$-dimensional Brownian motion because $\lambda_{ei} = 0$ for $i \in J \setminus J_e$. Similarly, we have

$$\widehat{\boldsymbol{N}}_s^{(n)}(\cdot) \xrightarrow{w} \operatorname{diag}(\{\lambda_{si}^{\frac{3}{2}}\sigma_{si}\})\boldsymbol{B}_s^*(\cdot), \tag{3.26}$$

where $\boldsymbol{B}_s^*(\cdot)$ is the d-dimensional standard Brownian motion. By the modeling assumptions, we can choose $\boldsymbol{B}_s^*(\cdot)$ to be independent of $\boldsymbol{B}_e^*(\cdot)$.

Let $\boldsymbol{\chi}^{(n)}(t)$ be the random vector whose i-th component is given by

$$\chi_i^{(n)}(t) \equiv \sum_{j \in J} \frac{1}{\sqrt{n}} \big(N_{sj}^{(n)}(nt) - \lambda_{sj}^{(n)}nt\big)(p_{ji} - \delta_{ji}),$$

then, it follows from (3.26) that

$$\boldsymbol{\chi}^{(n)}(\cdot) \xrightarrow{w} \Lambda_\chi \boldsymbol{B}_s^*(\cdot), \tag{3.27}$$

where $\Lambda_\chi \Lambda_\chi^{\mathrm{T}}$ is the covariance matrix whose ik-entry is given by

$$\sum_{j \in J} \lambda_{sj}^3 \sigma_{sj}^2 (p_{ji} - \delta_{ji})(p_{jk} - \delta_{jk}).$$

We define $\widehat{\boldsymbol{\Phi}}^{(n)}(t)$ as the random vector whose i-th component is given by

$$\widehat{\Phi}_i^{(n)}(t) \equiv \frac{1}{\sqrt{n}} \sum_{j \in J} \big(\Phi_{ji}([nt]) - [nt]p_{ji}\big),$$

and define the fluid scaling of $N_{sj}^{(n)}(t)$ as

$$\overline{N}_{sj}^{(n)}(t) = \frac{1}{n} N_{sj}^{(n)}(nt).$$

Let us compute the covariance matrix of $\widehat{\boldsymbol{\Phi}}^{(n)}(1)$, whose $k\ell$-entry is given by

$$\frac{1}{n}\mathbb{E}\Big(\sum_{i,j \in J} \big(\Phi_{ik}(n) - np_{ik}\big)\big(\Phi_{j\ell}(n) - np_{j\ell}\big)\Big) = \sum_{i \in J} p_{ik}(\delta_{k\ell} - p_{i\ell}), \tag{3.28}$$

where recall that $\delta_{ij} = 1(i = j)$. Let

$$\widehat{\boldsymbol{\Phi}^{(n)} \circ N_{sj}^{(n)}}(t) = \frac{1}{\sqrt{n}} \sum_{j \in J} \big(\Phi_{ji}(N_{sj}^{(n)}(nt)) - N_{sj}^{(n)}(nt)p_{ji}\big).$$

Since $\overline{N}_{sj}^{(n)}(t) \to \lambda_{sj}t$ almost surely as $n \to \infty$ and

$$\frac{1}{\sqrt{n}} \sum_{j \in J} \big(\Phi_{ji}(N_{sj}^{(n)}(nt)) - N_{sj}^{(n)}(nt)p_{ji}\big) = \frac{1}{\sqrt{n}} \sum_{j \in J} \big(\Phi_{ji}(n\overline{N}_{sj}^{(n)}(t)) - n\overline{N}_{sj}^{(n)}(t)p_{ji}\big)$$

$$\simeq \frac{1}{\sqrt{n}} \sum_{j \in J} \big(\Phi_{ji}([n\lambda_{sj}t]) - [n\lambda_{sj}t]p_{ji}\big),$$

applying Lemma 3.1 yields

$$\widehat{\boldsymbol{\Phi}^{(n)} \circ N_{sj}^{(n)}}(\cdot) \xrightarrow{w} \Lambda_{\Phi} \boldsymbol{B}_{\Phi}^*(\cdot), \tag{3.29}$$

where $\boldsymbol{B}_{\Phi}^*(\cdot)$ is the d-dimensional standard Brownian motion which is independent of $\boldsymbol{B}_e^*(\cdot)$ and $\boldsymbol{B}_s^*(\cdot)$, and $\Lambda_{\Phi}\Lambda_{\Phi}^{\mathrm{T}}$ is the d-dimensional covariance matrix whose ik-entry is given by

$$\sum_{j \in J} \lambda_{sj} p_{ji}(\delta_{ik} - p_{jk}),$$

which comes from (3.28).

Finally, the last term in the formula for $\underline{\widehat{X}}_i^{(n)}(s)$ is

$$\sqrt{n}\Big(\lambda_{ei}^{(n)} + \sum_{j \in J} \lambda_{sj}^{(n)} p_{ji} - \lambda_{si}^{(n)}\Big) = \sqrt{n}\Big(\lambda_{ai}^{(n)} - \sum_{j \in J} \lambda_{aj}^{(n)} p_{ji} + \sum_{j \in J} \lambda_{sj}^{(n)} p_{ji} - \lambda_{si}^{(n)}\Big)$$

$$= \sqrt{n} \sum_{j \in J} (\lambda_{aj}^{(n)} - \lambda_{sj}^{(n)})(\delta_{ji} - p_{ji})$$

$$= -\left[R\boldsymbol{\lambda}_s\right]_i,$$

where the last equality comes from (3.22).

Summarizing all the computations, we have that $\underline{\widehat{\boldsymbol{X}}}^{(n)}(\cdot)$ weakly converges to the Brownian motion with the covariance matrix Σ whose ik-entry Σ_{ik} is

$$\Sigma_{ik} = \lambda_{ei}^3 \sigma_{ei}^2 \delta_{ik} + \sum_{j \in J} \big(\lambda_{sj}^3 \sigma_{sj}^2 (p_{ji} - \delta_{ji})(p_{jk} - \delta_{jk}) + \lambda_{sj} p_{ji}(\delta_{ik} - p_{jk})\big), \quad i, k \in J, \tag{3.30}$$

and drift vector $\boldsymbol{\mu}$ whose i-entry μ_i is given by

$$\boldsymbol{\mu} = -R\boldsymbol{\lambda}_s. \tag{3.31}$$

Combing this with (3.23) and Theorem 3.1 and applying Theorem 3.4, we arrive at the following theorem.

Theorem 3.8 (A version of Theorem 1 of Reiman [56])**.** Assume (2a)–(2c), and define the n-th GJN by (3.20) and (3.21) and the routing matrix P, then the diffusion scaled process $\{\frac{1}{\sqrt{n}}\boldsymbol{L}^{(n)}(nt); t \geq 0\}$ converges in distribution to the SRBM $\{\boldsymbol{Z}(t)\}$ with the primitive data $(\Sigma, -R\boldsymbol{\lambda}_s, R)$, where Σ by (3.30). Thus, this SRBM is always stable.

Remark 3.3. *If we assume (3b) and (3c) instead of (3.20) and (3.21), then the limiting SRBM has primitive data $(\Sigma, -R\boldsymbol{c}, R)$, which is stable if and only if $\boldsymbol{c} > 0$.*

The following corollary is immediate from Theorems 3.3 and 3.8 and (3.30).

Corollary 3.1. *Under the assumptions of Theorem 3.8, if $d = 1$, then the limiting SRBM has the stationary distribution given by*

$$\mathbb{P}(Z_1(\infty) \leq x) = 1 - \exp\Big(-\frac{2}{\lambda_{e1}^2(\sigma_{e1}^2 + \sigma_{s1}^2)}x\Big), \qquad x \geq 0, \tag{3.32}$$

where $Z_1(\infty)$ is a random variable subject to the stationary distribution.

Let $\rho_i^{(n)} = \lambda_{ei}^{(n)}/\lambda_{si}^{(n)}$. Since (3.22) implies that

$$\frac{1}{\sqrt{n}} = 1 - \rho_i^{(n)}, \qquad i \in J,$$

we have the following approximation for the stationary queue length process $\boldsymbol{L}(\cdot)$.

$$\boldsymbol{L}(\cdot) \approx \operatorname{diag}(\{1 - \rho_i\})^{-1} \boldsymbol{Z}(\cdot), \tag{3.33}$$

where $\boldsymbol{Z}(\cdot)$ is the SRBM with the primitive data $(\Sigma, -R\boldsymbol{\lambda}_s, R)$.

We have not discussed about the workload process $\boldsymbol{W}(t)$. Let $\widehat{\boldsymbol{W}}^{(n)}(t) = \frac{1}{\sqrt{n}} \boldsymbol{W}^{(n)}(nt)$, then one may guess that

$$\widehat{\boldsymbol{W}}^{(n)}(\cdot) \xrightarrow{w} \operatorname{diag}(\{\lambda_{si}\})^{-1} \boldsymbol{Z}(\cdot). \tag{3.34}$$

and therefore, as $\lambda_{si} - \lambda_{ei} \downarrow 0$ for all $i \in J$,

$$\boldsymbol{W}(\cdot) \approx \operatorname{diag}(\{\lambda_{si} - \lambda_{ei}\})^{-1} \boldsymbol{Z}(\cdot). \tag{3.35}$$

We check that (3.34) indeed holds true. For simplicity, we assume that $\boldsymbol{W}^{(n)}(0) = \boldsymbol{0}$ and $\boldsymbol{L}^{(n)}(0) = \boldsymbol{0}$. Then, using (2.4) for the n-th GJN and the fact that

$$U_i^{(n)}(nt) = V_{si}^{(n)}(N_{si}^{(n)}(U_i^{(n)}(nt))) + A_{si}^{(n)}(nt),$$

where $A_{si}^{(n)}(t)$ is the attained service time of a customer being served at node i if any while $A_{si}^{(n)}(t) = 0$ if no customer in service there, the workload representation (2.6) for the n-th GJN is rewritten as

$$\widehat{W}_i^{(n)}(t) = \frac{1}{\sqrt{n}} V_{si}^{(n)}\big(L_i^{(n)}(nt) + N_{si}^{(n)}(U_i^{(n)}(nt))\big) - \frac{1}{\sqrt{n}}\big(V_{si}^{(n)}(N_{si}^{(n)}(U_i^{(n)}(nt))) + A_{si}^{(n)}(nt)\big)$$

$$= \frac{L_i^{(n)}(nt)}{\sqrt{n}} \frac{1}{L_i^{(n)}(nt)} \sum_{\ell = N_{si}^{(n)}(U_i^{(n)}(nt))+1}^{N_{si}^{(n)}(U_i^{(n)}(nt))+L_i^{(n)}(nt)} T_{si}^{(n)} - \frac{1}{\sqrt{n}} A_{si}^{(n)}(nt).$$

Since $L_i^{(n)}(nt) \xrightarrow{a.s.} \infty$, which can be checked by fluid approximation, we have

$$\frac{1}{L_i^{(n)}(nt)} \sum_{\ell = N_{si}^{(n)}(U_i^{(n)}(nt))+1}^{N_{si}^{(n)}(U_i^{(n)}(nt))+L_i^{(n)}(nt)} T_{si}^{(n)} \xrightarrow{a.s.} \frac{1}{\lambda_{si}}, \qquad (n \to \infty).$$

Hence, $\frac{1}{\sqrt{n}} A_{si}^{(n)}(nt) \xrightarrow{a.s.} 0$ and Theorem 3.8 yields (3.34).

We have derived Theorem 3.8 along the line due to Reiman [56]. Some years later, Chen and Mandelbaum [10] simplified a part of the proof using Skorohod representation of Theorem 3.5. This does not have so much benefit for proving Theorem 3.8, but it does for extensions or modifications of the GJN. A closed GJN in [10] is such an example.

3.4. Extension to many server nodes

Theorem 3.8 can be extended for the case where each node may have multiple servers and their service times may have different distributions, while keeping $i.i.d$ service times at each server. This extension is already mentioned in [56], but was firstly proved in [11]. The basic idea for this extension is called a sandwich method (see also [13] for a single node queue). We present this extension following Chen and Shanthikumar [11].

Assume that node i has c_i servers and the j-th server there has service times $\{T_{sij}(\ell)\}$ with finite mean service time μ_{sij}^{-1} and variance σ_{sij}^2. Define the counting process of departures from this server by $N_{sij}(t)$ as

$$N_{sij}(t) = \sup \left\{ n \geq 0; \sum_{\ell=1}^{n} T_{sij}(\ell) \leq t \right\},$$

and let $U_{ij}(t)$ be the total time when the j-th server at node i is busy in the time interval $[0, t]$. Then, for the GJN with multiple servers at nodes, $X_i(t)$ of (3.2) are changed to

$$
\begin{aligned}
X_i(t) = & N_{ei}(t) - \lambda_{ei}t + \sum_{j \in J} \sum_{k=1}^{c_j} \left(N_{sjk}(U_{jk}(t)) - \lambda_{sjk}U_{jk}(t) \right) p_{ji} \\
& + \sum_{j \in J} \sum_{k=1}^{c_j} \left(\Phi_{ji}(N_{sjk}(U_{jk}(t))) - N_{sjk}(U_{jk}(t)) p_{ji} \right) \\
& - \sum_{k=1}^{c_i} (N_{sik}(U_{ik}(t)) - \lambda_{sik}U_{ik}(t)) + \left(\lambda_{ei} + \sum_{j \in J} \sum_{k=1}^{c_j} \lambda_{sjk}p_{ji} - \sum_{k=1}^{c_i} \lambda_{sik} \right) t, \qquad (3.36)
\end{aligned}
$$

but we can keep $\boldsymbol{L}(t)$ of (3.1) with $\boldsymbol{X}(t)$ whose i-th component is this $X_i(t)$. However, we can not uniquely determine $\boldsymbol{L}(t)$ and $\boldsymbol{Y}(t)$ by (3.1) and (3.4) obviously because the information for $U_{ik}(t)$ is not sufficient, where

$$Y_i(t) = \sum_{k=1}^{c_i} \lambda_{sik}(t - U_{ik}(t)).$$

We modify $\boldsymbol{X}(t)$ similar to $\underline{\boldsymbol{X}}(t)$ of (3.18) as

$$
\begin{aligned}
\underline{X}_i(t) = & N_{ei}(t) - \lambda_{ei}t + \sum_{j \in J} \sum_{k=1}^{c_j} \left(N_{sjk}(t) - \lambda_{sjk}t \right)(p_{ji} - \delta_{ij}) \\
& + \sum_{j \in J} \sum_{k=1}^{c_j} \left(\Phi_{ji}(N_{sjk}(t)) - N_{sjk}(t)p_{ji} \right) + \left(\lambda_{ei} + \sum_{j \in J} \sum_{k=1}^{c_j} \lambda_{sjk}p_{ji} - \sum_{k=1}^{c_i} \lambda_{sik} \right) t. \quad (3.37)
\end{aligned}
$$

We next let $(\phi(\boldsymbol{X}), \Psi(\boldsymbol{X}))$ be the reflection map for (3.1) and (3.4) with $\boldsymbol{X}(\cdot)$ of (3.37). Since $\Psi(\boldsymbol{X})$ is the minimal solution, we have

$$\Psi(\boldsymbol{X})(t) \leq \boldsymbol{Y}(t). \qquad (3.38)$$

On the other hand, for the d-dimensional vector $\boldsymbol{c}$ whose i-th component is the number of servers c_i at node i, it can be shown that $\Psi(\boldsymbol{X} - \boldsymbol{c})$ is the maximal element in the set:

$$\left\{ \boldsymbol{y} \in \mathcal{T}; (x_i(t) + [R\boldsymbol{y}(t)]_i)dy_i(t) \leq c_i, i \in J, \forall t \geq 0 \right\},$$

where where $\mathcal{T}$ is the set of non-decreasing functions, and therefore

$$\boldsymbol{Y}(t) \leq \Psi(\boldsymbol{X} - \boldsymbol{c})(t). \qquad (3.39)$$

Thus, we have

$$\Psi(\boldsymbol{X})(t) \leq \boldsymbol{Y}(t) \leq \Psi(\boldsymbol{X} - \boldsymbol{c})(t).$$

We are now ready to introduce the sequence of the GJN with multiple servers at each node. For the n-th GJN, we use the following notation at each node i. The mean arrival rate and the variance of the inter-arrival time are denoted by $\lambda_{ei}^{(n)}$ and $(\sigma_{ei}^{(n)})^2$, which are the same as those for the single server case. For the j-th server of node i, the mean service rate and the variance of the service time are denoted by $\lambda_{sij}^{(n)}$ and $(\sigma_{sij}^{(n)})^2$. We assume that there exist constants λ_{ei}, λ_{sik}, h_{ei} and h_{sik} such that

$$\lim_{n\to\infty} \sqrt{n}(\lambda_{ei}^{(n)} - \lambda_{ei}) = h_{ei}, \qquad \lim_{n\to\infty} \sqrt{n}(\lambda_{sij}^{(n)} - \lambda_{sij}) = h_{sij}, \tag{3.40}$$

$$\lambda_{ei} + \sum_{j\in J}\sum_{k=1}^{c_j} \lambda_{sjk}p_{ji} - \sum_{k=1}^{c_i} \lambda_{sik} = 0, \qquad \lim_{n\to\infty} \sigma_{ei}^{(n)} = \sigma_{ei}, \qquad \lim_{n\to\infty} \sigma_{sij}^{(n)} = \sigma_{sij}, \tag{3.41}$$

$$\sup_{n\geq 1} \mathbb{E}\Big((T_{ei}^{(n)})^{2+\delta} + \sum_{j=1}^{c_j}(T_{sij}^{(n)})^{2+\delta}\Big) < \infty, \qquad \exists\delta > 0. \tag{3.42}$$

Then, (3.38), (3.39) and (3.37) for the n-th GJN concludes the following theorem.

Theorem 3.9 (Theorem 5.2 of [11]). *Under the assumptions (3.40) (3.41) and (3.42), the diffusion scaled process $\widehat{L}^{(n)}(\cdot) \equiv \{\frac{1}{\sqrt{n}}L^{(n)}(nt); t \geq 0\}$ converges in distribution to the SRBM $Z(\cdot)$ with the primitive data (Σ, μ, R), where Σ and μ are given by*

$$\Sigma_{ik} = \lambda_{ei}^3 \sigma_{ei}^2 \delta_{ik} + \sum_{j\in J}\sum_{\ell=1}^{c_j} \big(\lambda_{sj\ell}^3 \sigma_{sj\ell}^2 (p_{ji} - \delta_{ji})(p_{jk} - \delta_{jk}) + \lambda_{sj\ell}p_{ji}(\delta_{ik} - p_{jk})\big),$$

$$\mu_i = h_{ei} + \sum_{j\in J}\sum_{k=1}^{c_j} h_{sjk}p_{ji} - \sum_{k=1}^{c_i} h_{sik}.$$

This theorem tells us that the diffusion approximation for the GJN with heterogeneous many servers at all nodes is identical with that for the GNJ with single servers at all nodes with the mean service rate and the variance of service times at node j, respectively, given by $\sum_{\ell=1}^{c_j} \lambda_{sj\ell}$ and

$$\Big(\sum_{\ell=1}^{c_j} \lambda_{sj\ell}\Big)^{-3} \sum_{\ell=1}^{c_j} \lambda_{sj\ell}^3 \sigma_{sj\ell}^2.$$

In particular, if servers are identical at each node, that is, $\lambda_{sj\ell} = \lambda_{sj}$ and $\sigma_{sj\ell} = \sigma_{sj}$, then the above variance becomes σ_{sj}^2/c_j^2, and therefore the result is compatible with Kingman's conjecture for the stationary waiting time in the $GI/GI/k$ queue for $k = c_j$, which is proved by Köllerström [46].

4. Quality of the diffusion approximation

We have discussed about the diffusion approximation through a process limit. This may not imply the weak convergence for the stationary distribution. Let $\widehat{L}^{(n)}(\infty)$ and $Z(\infty)$ be random vectors subject to the stationary distributions of the diffusion scaled process $\widehat{L}^{(n)}(\cdot)$ and the SRBM $Z(\cdot)$, respectively. Then, the following implication, which is referred to as continuity of the stationary distribution, is important for application because the stationary distribution is widely used for performance evaluation.

$$\widehat{L}^{(n)}(\cdot) \xrightarrow{w} Z(\cdot) \quad \text{implies} \quad \widehat{L}^{(n)}(\infty) \xrightarrow{w} Z(\infty). \tag{4.1}$$

However, this continuity does not automatically hold in general, and it requires a proof. Similar continuity also may be considered for characteristics of the stationary distribution such as moments and tail asymptotics. Those continuity can be considered as quality support for the diffusion approximation, and therefore their verifications are important for application. In this section, we review the current status for those continuity.

4.1. Continuity of the stationary distribution and its moments

As already mentioned in Section 1, this continuity has been verified by Gamarnik and Zeevi [24] and Budhiraja and Lee [8]. They use the fact that the tightness for the sequence of $\widehat{\boldsymbol{L}}^{(n)}(\infty)$, that is, the sequence of the stationary distributions, is sufficient to prove (4.1). This is because we can construct a stationary process for the n-th GJN taking its stationary distribution as the initial distribution at time 0 and choose the subsequence of those stationary distributions which weakly converges to a distribution by the tightness.

In [24], uniformly light tails over n (see (1) and (2) of [24] for their precise definitions) are assumed for sequences of the inter-arrival time $T_{ei}^{(n)}$ and the service time $T_{si}^{(n)}$, where we dropped the index (ℓ) from $T_{ei}^{(n)}(\ell)$ and $T_{si}^{(n)}(\ell)$. This convention for indexed random variables will be used in what follows if their distributions are only concerned. Then, it is proved that there are positive constants a_1 and a_2 for a positive vector $\boldsymbol{w} \in \mathbb{R}^d$ such that

$$\mathbb{P}(\langle \boldsymbol{w}, \widehat{\boldsymbol{L}}^{(n)}(\infty) \rangle > x) \le a_1 e^{-a_2 x}, \qquad \forall x > 0.$$

See Theorem 7 of [24]. This is a stronger result more than what is needed for the tightness, and gives an exponential upper bound for the tails of the stationary distribution of the corresponding SRBM, which implies that the moment with any order has the continuity. Furthermore, a finer upper bound is obtained in [24]. However, those bounds may not be so interesting because a stable SRBM is known to have the light tailed stationary distribution if reflection matrix R is an $\mathcal{M}$-matrix.

In [8], (4.1) is proved under the standard uniformly integrable assumptions for $(T_{ei}^{(n)})^2$ and $(T_{si}^{(n)})^2$. Furthermore, they proved that, if $\mathbb{E}([T_{ei}^{(n)}]^p)$ for $i \in J_e$ and $\mathbb{E}([T_{si}^{(n)}]^p)$ for $i \in J$ are uniformly finite over n and if certain asymptotic condition on $N_{ei}^{(n)}(\cdot)$ and $N_{si}^{(n)}(\cdot)$ are satisfied (see (i) in (A8.p) of [8]), then the moment with order less than p of the stationary distribution also has the continuity (see Theorem 3.1 of [8]). However, it is known from the finite moment condition for the $GI/GI/1$ queue that $\mathbb{E}([L_i]^{p+1}) = \infty$ if $\mathbb{E}([T_{si}]^p) = \infty$ at least for $i \in J_e$ (e.g., see Theorem 2.1 of Section X of [1]). Thus, we can not have the continuity for the $(p+1)$-th moment of the stationary distribution in this case because the limiting distribution, which is the stationary distribution of the SRBM, always has light tail and therefore has finite moments for all order (e.g., see [4]). We may say that the diffusion approximation by SRBM is not good if service time distributions have heavy tails although the continuity holds for the stationary distribution as long as the uniformly integrability holds.

As we demonstrated in Section 2.4, the continuity of the stationary distribution has been directly verified for the single node case of the GJN. As suggested there, the same verification may be possible for a general GJN. Since no time scaling is required for this limiting operation, this continuity has a stronger aspect than the obtained in [8] and [24]. Furthermore, more information on the convergence may be obtained. In fact, we will see such an example for the tail decay rate of the stationary distribution in the next section. However, a large part of this approach is still open for future study.

4.2. Continuity of the tail decay rate of the stationary distribution

The continuity of the tail decay rate of the stationary distribution is studied for a feedforward queueing network with deterministic service times by Majewski [49]. This network is assumed to have multi-class customers and FCFS discipline. See Section 5 for a multi-class queueing network. It seems that the feedforward assumption and deterministic service times greatly simplifies arguments but its application is limited. It is claimed that the continuity holds for the tail decay rate, precisely, for the large deviations rate function (see Theorem 9 of [49]).

For a general routing but single-class and two node GJN, the author [52] recently obtained the tail decay rate of the marginal stationary distribution in an arbitrary direction, assuming a Markov modulated Poisson arrivals and phase type service time distributions, which includes the special case that an arrival process is renewal with a phase type inter-arrival distribution. For simplicity, we assume this renewal arrival.

In the rest of this subsection, we will discuss the continuity on those decay rates. Assume that $d = 2$. Recall that $N_{ei}(t)$ is the number of exogenous arrivals at node i during the time period $[0, t]$, and define the logarithmic generating functions similar to (2.24) as

$$\gamma_{ei}(\theta) = \lim_{t \to \infty} \frac{1}{t} \log \mathbb{E}(e^{\theta N_{ei}(t)}), \qquad i = 1, 2. \tag{4.2}$$

Because T_{ei} has a phase-type distribution, the assumption (2d) is satisfied for T_{ei}, and therefore these functions are well defined for all $\theta \in \mathbb{R}$. We similarly define $\gamma_{si}(\theta)$ for the counting process $N_{si}(t)$ for service completions, that is, departures, from node i during a busy period.

Similarly in Section 2.4, we denote the moment generating functions of T_{ui} is denoted by

$$\widetilde{F}_{ui}(\theta) = \mathbb{E}(e^{\theta T_{ui}}), \qquad u = e, s, \ i = 1, 2,$$

then we have the following facts similar to Lemma 2.1.

Lemma 4.1. $\widetilde{F}_{ui}$ *has a unique inverse* $\widetilde{F}_{ui}^{-1}$ *whose domain is* $\mathbb{R}$*, and, for* $u = e, s$ *and* $i = 1, 2$*,*

$$\gamma_{ui}(\theta) = -\widetilde{F}_{ui}^{-1}(e^{-\theta}), \qquad \theta \in \mathbb{R}. \tag{4.3}$$

We next consider a logarithmic moment generating function for customers' movement at their departure. Recall that departures from node i are routed to node j according to the Markovian routing function $\Phi_{ij}(\cdot)$. Hence, if a server at node i is busy for all time up to t, then we define the logarithmic moment generating function for departing and routing customers for busy node i as

$$\gamma_{di}(\boldsymbol{\theta}) = \lim_{t \to \infty} \frac{1}{t} \log \mathbb{E}(e^{-\theta_i N_{si}(t) + \theta_{3-i} \Phi_{i(3-i)}(N_{si}(t))}), \qquad \boldsymbol{\theta} \equiv (\theta_1, \theta_2) \in \mathbb{R}^2.$$

Let $t_i(\boldsymbol{\theta}) = e^{-\theta_i}(p_{i0} + p_{ij}e^{\theta_j})$ for $j = 3 - i$, then it is not hard to see that

$$\gamma_{di}(\boldsymbol{\theta}) = \lim_{t \to \infty} \frac{1}{t} \log \mathbb{E}(t_i(\boldsymbol{\theta})^{N_{si}(t)})$$

$$= \lim_{t \to \infty} \frac{1}{t} \log \mathbb{E}(e^{N_{si}(t) \log t_i(\boldsymbol{\theta})}) = \gamma_{si}(\log t_i(\boldsymbol{\theta})). \tag{4.4}$$

We are now ready to present main results. For this, we need some notations. Let

$$\gamma_+(\boldsymbol{\theta}) = \gamma_{e1}(\theta_1) + \gamma_{e2}(\theta_2) + \gamma_{d1}(\boldsymbol{\theta}) + \gamma_{d2}(\boldsymbol{\theta}),$$

which stands for the logarithmic generating function for the total net flow. The following notations are geometric objects.

$$\Gamma = \{\boldsymbol{\theta} \in \mathbb{R}^2; \gamma_+(\boldsymbol{\theta}) < 0\}, \qquad \partial\Gamma = \{\boldsymbol{\theta} \in \mathbb{R}^2; \gamma_+(\boldsymbol{\theta}) = 0\},,$$
$$\Gamma_i = \{\boldsymbol{\theta} \in \Gamma; t_{3-i}(\boldsymbol{\theta}) > 1\}, \qquad \partial\Gamma_i = \{\boldsymbol{\theta} \in \partial\Gamma; t_{3-i}(\boldsymbol{\theta}) > 1\}, \qquad i = 1, 2,$$
$$\Gamma_{\max} = \{\boldsymbol{\theta} \in \mathbb{R}^2; \exists \boldsymbol{\theta}' \in \partial\Gamma, \boldsymbol{\theta} < \boldsymbol{\theta}'\}.$$

Let $\varphi(\boldsymbol{\theta}) \equiv \mathbb{E}(e^{\langle \boldsymbol{\theta}, \boldsymbol{L}(\infty) \rangle})$ be the moment generating function of $\boldsymbol{L}(\infty)$ subject to the marginal stationary distribution with respect to the numbers of customers at nodes 1 and 2, and define $\mathcal{D}$ as

$$\mathcal{D} = \text{the interior of } \{\boldsymbol{\theta} \in \mathbb{R}^2; \varphi(\boldsymbol{\theta}) < \infty\},$$

which is referred to as a convergence domain of φ. Denote the two extreme points of Γ_i by

$$\boldsymbol{\theta}^{(i,\Gamma)} = \arg_{\boldsymbol{\theta}\in\mathbb{R}^2} \sup\{\theta_i \geq 0; \boldsymbol{\theta} \in \partial\Gamma_i\}, \quad i = 1, 2.$$

Using these points, we define the vector $\boldsymbol{\tau}$ by

$$\tau_1 = \sup\{\theta_1 \in \mathbb{R}; \boldsymbol{\theta} \in \partial\Gamma_1; \theta_2 < \theta_2^{(2,\Gamma)}\},$$
$$\tau_2 = \sup\{\theta_2 \in \mathbb{R}; \boldsymbol{\theta} \in \partial\Gamma_2; \theta_1 < \theta_1^{(1,\Gamma)}\}.$$

We refer to the following results (see [52] for their proofs).
Theorem 4.1 (Theorems 3.1 and 3.3 of Miyazawa [52]). *For the two node generalized Jackson network under phase type setting, if it is stable, then $\mathcal{D} = \{\boldsymbol{\theta} \in \Gamma_{\max}; \boldsymbol{\theta} < \boldsymbol{\tau}\}$, and, for any non-zero vector $\boldsymbol{c} \geq \boldsymbol{0}$,*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}(\langle \boldsymbol{c}, \boldsymbol{L}(\infty) \rangle > x) = -\sup\{u > 0; u\boldsymbol{c} \in \mathcal{D}\}. \tag{4.5}$$

Note that this theorem shows that the decay rates are determined by (4.2) and (4.4). Hence, we can expect that Theorem 4.1 hold beyond the phase-type setting and more general routing mechanism.

It is also notable that these decay rates are obtained in the exactly same way as those of the stationary distribution of the two-dimensional SRBM (see Theorems 2.1, 2.2 and 2.3 of [18]). To prove the continuity of the decay rates, we here briefly present results in [18].

Let $\{(\boldsymbol{Z}(t), \boldsymbol{Y}(t))\}$ be the two dimensional SRBM with the primitive data $(\Sigma, \boldsymbol{\mu}, R)$, and assume that R is a completely-$\mathcal{S}$ matrix and the stability condition $R^{-1}\boldsymbol{\mu} < \boldsymbol{0}$ is satisfied. Define the following functions and sets.

$$\widetilde{\gamma}_+(\boldsymbol{\theta}) = \frac{1}{2}\langle \boldsymbol{\theta}, \Sigma\boldsymbol{\theta} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle, \qquad \widetilde{\gamma}_i(\boldsymbol{\theta}) = \langle R^{[i]}, \boldsymbol{\theta} \rangle, \qquad i = 1, 2,$$
$$\widetilde{\Gamma} = \{\boldsymbol{\theta} \in \mathbb{R}^2; \widetilde{\gamma}_+(\boldsymbol{\theta}) < 0\}, \qquad \widetilde{\Gamma}_i = \{\boldsymbol{\theta} \in \widetilde{\Gamma}; \widetilde{\gamma}_{3-i}(\boldsymbol{\theta}) < 0\},$$
$$\widetilde{\Gamma}_{\max} = \{\boldsymbol{\theta} \in \mathbb{R}^2; \exists \boldsymbol{\theta}' \in \widetilde{\Gamma}, \boldsymbol{\theta} < \boldsymbol{\theta}'\},$$

where we recall that $R^{[i]}$ is the i-th column of R. Let $\boldsymbol{Z}(\infty)$ be a random vector subject to the stationary distribution of $\{\boldsymbol{Z}(t)\}$. Note that this stationary distribution is uniquely

determined by $\widetilde{\gamma}_+$ and $\widetilde{\gamma}_i$ (see Remark 3.1 for its reason). Recall that $\tilde{\varphi}(\boldsymbol{\theta}) = \mathbb{E}(e^{\langle \boldsymbol{\theta}, \boldsymbol{Z}(\infty) \rangle})$. Define the convergence domain of $\tilde{\varphi}(\boldsymbol{\theta})$ as

$$\widetilde{\mathcal{D}} = \text{the interior of } \{\boldsymbol{\theta} \in \mathbb{R}^2; \tilde{\varphi}(\boldsymbol{\theta}) < \infty\}.$$

We also define $\widetilde{\boldsymbol{\theta}}^{(i,\Gamma)}$ and $\widetilde{\boldsymbol{\tau}}$ in the exactly same way as $\boldsymbol{\theta}^{(i,\Gamma)}$ and $\boldsymbol{\tau}$. Then, similar to Theorem 4.1, we have the following results from Theorems 2.1, 2.2 and 2.3 in [18], in which finer tail asymptotics are obtained, but we do not use them here.

Theorem 4.2 (Dai and Miyazawa [18]). *For the two node generalized Jackson network under phase type setting, if it is stable, then* $\widetilde{\mathcal{D}} = \{\boldsymbol{\theta} \in \widetilde{\Gamma}_{\max}; \boldsymbol{\theta} < \widetilde{\boldsymbol{\tau}}\}$, *and, for any non-zero vector* $\boldsymbol{c} \geq \boldsymbol{0}$,

$$\lim_{x \to \infty} \frac{1}{x} \log \mathbb{P}(\langle \boldsymbol{c}, \boldsymbol{Z}(\infty) \rangle > x) = -\sup\{u > 0; u\boldsymbol{c} \in \widetilde{\mathcal{D}}\}. \tag{4.6}$$

We now consider the sequence of the GJNs whose joint queue length processes under diffusion scaling weakly converge to a stable SRBM. For this, we use the same notations and assumptions in Section 2. For example, recall that $\lambda_{ei}^{(n)}, \lambda_{si}^{(n)}$ and $(\sigma_{ei}^{(n)})^2, (\sigma_{si}^{(n)})^2$ are the arrival and service rates and the variances of the inter-arrival and service times, respectively. We assume the heavy traffic condition (3b) with $c_i = \lambda_{ai} - \lambda_{si}$ and regularity condition (3c). Then, the covariance matrix Σ and the drift vector $\boldsymbol{\mu}$ of this SRBM are given by (3.30) and (3.31), respectively, and $R = I - P^{\mathrm{T}}$. The logarithmic generating function of the n-th system under diffusion scaling is defined as

$$\widehat{\gamma}_{ui}^{(n)}(\theta) = \lim_{t \to \infty} \frac{1}{t} \log \mathbb{E}(e^{\theta \frac{1}{\sqrt{n}} N_{ui}^{(n)}(nt)}) = n\gamma_{ui}^{(n)}\left(\frac{1}{\sqrt{n}}\theta\right), \qquad u = e, s. \tag{4.7}$$

Hence, by Lemmas 2.2 and 4.1, for $u = e, s$,

$$\widehat{\gamma}_{ui}^{(n)}(\theta) = -n(\widetilde{F}_{ui}^{(n)})^{-1}\left(e^{-\frac{\theta}{\sqrt{n}}}\right) = \sqrt{n}\lambda_{ui}^{(n)}\theta + \frac{1}{2}(\lambda_{ui}^{(n)})^3(\sigma_{ui}^{(n)})^2\theta^2 + o(1),$$

and therefore it follows from (4.4) that, for $i = 1, 2$ and $j = 3 - i$,

$$\begin{aligned}
\widehat{\gamma}_{di}^{(n)}(\boldsymbol{\theta}) &= -n(\widetilde{F}_{si}^{(n)})^{-1}\left(e^{-\log t_i\left(\frac{1}{\sqrt{n}}\boldsymbol{\theta}\right)}\right) \\
&= n\left(\lambda_{si}^{(n)} \log t_i\left(\frac{1}{\sqrt{n}}\boldsymbol{\theta}\right) + \frac{1}{2}(\lambda_{si}^{(n)})^3(\sigma_{si}^{(n)})^2 \log^2 t_i\left(\frac{1}{\sqrt{n}}\boldsymbol{\theta}\right)\right) + o(1) \\
&= \sqrt{n}\lambda_{si}^{(n)}(-\theta_i + p_{ij}\theta_j) + \frac{1}{2}\lambda_{si}^{(n)} p_{ij}(1 - p_{ij})\theta_j^2 + \frac{1}{2}(\lambda_{si}^{(n)})^3(\sigma_{si}^{(n)})^2(\theta_i - \theta_j)^2 + o(1).
\end{aligned}$$

From these facts, we can see that

$$\lim_{n \to \infty} \widehat{\gamma}_+^{(n)}(\boldsymbol{\theta}) = \widetilde{\gamma}_+(\boldsymbol{\theta}), \qquad \lim_{n \to \infty}(-(\lambda_{si}^{(n)}\sqrt{n})^{-1})\widehat{\gamma}_{di}^{(n)}(\boldsymbol{\theta}) = \widetilde{\gamma}_i(\boldsymbol{\theta}), \quad i = 1, 2. \tag{4.8}$$

Thus, in the view of Theorems 4.1 and 4.2, we are very close to have the continuity of the decay rates. However, the second equation of (4.8) is a bit short to prove the following theorem. We give a complete proof in Appendix D.

Theorem 4.3. *For a sequence of two node GJNs with phase type* $T_e^{(n)} \equiv T_e$ *and* $T_s^{(n)}$, *if they are stable and satisfy the condition* (3.21), *then the convergence domain* $\mathcal{D}^{(n)}$ *of the moment generating function of the stationary distribution of the n-th GJN converges to the domain* $\widetilde{\mathcal{D}}$ *for the limiting SRBM, and therefore the decay rate of* $\langle \boldsymbol{c}, \boldsymbol{L}^{(n)}(\infty) \rangle$ *for any non-zero* $\boldsymbol{c} \geq \boldsymbol{0}$ *converges to that of the limiting SRBM.*

5. Multi-class queueing network

We discuss how diffusion approximation is obtained for a multi-class queueing network with FCFS service discipline, following Williams [66]. We modify the GJN for this. We denote the set of customer types by $K \equiv \{1, 2, \ldots, c\}$. Type k customers are served only at node $\beta(k)$, and changed to type k' customer after service completion with probability $p_{kk'}$ independent of everything else. Similarly, T_{ei} and T_{si} are changed to T_{ek} and T_{sk}, respectively, for $k \in K$.

Let $N_{ek}(t), N_{ak}(t), N_{sk}(t), N_{dk}(t)$ be the counting processes for exogenous arrivals, total arrivals, potential service completions, departures of type k customers. Let $L_k(t), U_k(t)$ be the number of customers and the accumulated service time for type k customers. Let $W_i(t), Y_i(t)$ be the workload and the accumulated idle time at node i at time t, respectively. For simplicity, we assume that $\boldsymbol{L}(0) = \boldsymbol{0}$ and $\boldsymbol{W}(0) = \boldsymbol{0}$. Then, they satisfy the following set of equations for $i \in J$ and $k \in K$. Recall that the composition $f \circ g$ is defined as $f \circ g(x) = f(g(x))$ for appropriate functions f and g.

$$N_{ak}(t) = N_{ek}(t) + \sum_{k' \in K} (\Phi_{k'k} \circ N_{dk'})(t), \tag{5.1}$$

$$L_k(t) = N_{ak}(t) - N_{dk}(t), \tag{5.2}$$

$$N_{dk}(t) = (N_{sk} \circ U_k)(t), \tag{5.3}$$

$$W_i(t) = \sum_{k \in K} c_{ik}(V_{sk} \circ N_{ak})(t) - t + Y_i(t), \tag{5.4}$$

$$Y_i(t) = t - \sum_{k \in K} c_{ik}U_k(t), \tag{5.5}$$

$$\int_0^t W_i(u)dY_i(u) = 0, \tag{5.6}$$

where $c_{ik} = 1(\beta(k) = i)$. One can see that (5.2) with (5.1), (5.4) with (5.5), and (5.6) are parallel to (2.4), (2.6) and (3.7) of a single-class queueing network. An essential feature here is that the workload $\boldsymbol{W}(t)$ plays a key role while the queue length $\boldsymbol{L}(t)$ did so in the single-class case. It also indicates the lack of information because $\boldsymbol{L}(t)$ is higher dimensional than $\boldsymbol{W}(t)$. To fill this gap, we need to specify service discipline among customer types at each node.

Once the service discipline is specified, we can fully characterize $\boldsymbol{W}(t)$ and $\boldsymbol{L}(t)$ by (5.1)–(5.6). For example, if the service discipline is FCFS independent of types, then the following conditions are sufficient.

$$N_{ak}(t) = N_{dk}(t + W_{\beta(k)}(t)), \tag{5.7}$$

which represents that the type k customers arriving up to time t will depart up to time $t + W_{\beta(k)}(t)$. However, this is said in principle, and it is almost impossible to compute them. On the other hand, one may guess that the number of each type customers in node i would be proportional to their arrival rate in heavy traffic because the queue there would be very long and service does not discriminate types. This is called a state-space collapse, which will be formally defined below. Thus, if the state-space collapse occurs, then we may compute $\boldsymbol{L}(t)$ from the lower dimensional $\boldsymbol{W}(t)$. This is exactly a key idea to get diffusion approximation for a multi-class queueing network.

We now introduce the sequence of the multi-class GJN's. They are indexed by natural numbers similar to those of the single-class GJN's. Thus, the n-th system have the following notations. $T_{ek}^{(n)}(\ell)$ and $T_{sk}^{(n)}(\ell)$ are the inter-arrival (from the outside) and service times of

type k customers. It is assume that they are *i.i.d.* for each fixed n and $k \in K$. Counting processes of type k customers and their mean rates are denoted by $N_{uk}^{(n)}(t)$ and $\lambda_{uk}^{(n)}$ for $u = e, a, s, d$. Routing function, the number of type k customers, the total workload brought by ℓ customers of type k are denoted by $\Phi_{kk'}^{(n)}(\ell)$, $L_k^{(n)}(t)$, $V_{sk}^{(n)}(\ell)$ for $k \in K$. The workload and total idle time at node i are denoted by $W_i^{(n)}(t)$, $Y_i^{(n)}(t)$ for $i \in J$. Vectors and matrix whose entries are given by them are denoted by $\boldsymbol{N}_u^{(n)}(t)$ and $\boldsymbol{\lambda}_u^{(n)}$ for $u = e, a, s, d$, $\boldsymbol{\Phi}^{(n)}(\ell)$, $\boldsymbol{L}^{(n)}(t)$, $\boldsymbol{V}_s^{(n)}(\ell)$, $\boldsymbol{W}^{(n)}(t)$ and $\boldsymbol{Y}^{(n)}(t)$, respectively. We also introduce $|K| \times |K|$ matrices $P \equiv \{p_{kk'}; k, k' \in K\}$ and $M^{(n)} \equiv \mathrm{diag}(\{m_{sk}^{(n)}; k \in K\})$, $d \times K$ matrix $C \equiv \{1(\beta(k) = i); i \in J, k \in K\}$, where $m_{sk}^{(n)} = (\lambda_{sk}^{(n)})^{-1}$. We define the traffic intensity $\rho_k^{(n)} = \lambda_{ak}^{(n)} m_{sk}^{(n)}$, where the total arrival rate of type k customer λ_{ak} is given by

$$\boldsymbol{\lambda}_a^{(n)} = (I - P^{\mathrm{T}})^{-1} \boldsymbol{\lambda}_e^{(n)}.$$

We define fluid scaling versions of the above processes as

$$\overline{N}_{uk}^{(n)}(t) = \frac{1}{n} N_{uk}^{(n)}(nt), \qquad \overline{\Phi}_{kk'}^{(n)}(t) = \frac{1}{n} \Phi_{kk'}^{(n)}([nt]),$$

$$\overline{L}_k^{(n)}(t) = \frac{1}{n} L_k^{(n)}(nt), \qquad \overline{V}_{sk}^{(n)}(t) = \frac{1}{n} V_{sk}^{(n)}([nt]),$$

$$\overline{W}_i^{(n)}(t) = \frac{1}{n} W_i^{(n)}(nt), \qquad \overline{Y}_i^{(n)}(t) = \frac{1}{n} Y_i^{(n)}(nt).$$

Similarly, diffusion scaling versions are defined as

$$\widehat{N}_{uk}^{(n)}(t) = \frac{1}{\sqrt{n}} \big(N_{uk}^{(n)}(nt) - \lambda_{uk}^{(n)} nt\big), \qquad \widehat{\Phi}_{kk'}^{(n)}(t) = \frac{1}{\sqrt{n}} \big(\Phi_{kk'}^{(n)}([nt]) - p_{kk'}[nt]\big)$$

$$\widehat{L}_k^{(n)}(t) = \frac{1}{\sqrt{n}} L_k^{(n)}(nt), \qquad \overline{V}_{sk}^{(n)}(t) = \frac{1}{\sqrt{n}} \big(V_{sk}^{(n)}([nt]) - m_{sk}^{(n)}[nt]\big),$$

$$\widehat{W}_i^{(n)}(t) = \frac{1}{\sqrt{n}} W_i^{(n)}(nt), \qquad \widehat{Y}_i^{(n)}(t) = \frac{1}{\sqrt{n}} Y_i^{(n)}(nt),$$

$$\widehat{V_{sk}^{(n)} \circ N_{ak}^{(n)}}(t) = \frac{1}{\sqrt{n}} \big(V_{sk}^{(n)} \circ N_{ak}^{(n)}(nt) - \lambda_{ak}^{(n)} m_{sk}^{(n)} nt\big),$$

$$\widehat{\Phi_{kk'}^{(n)} \circ N_{dk}^{(n)}}(t) = \frac{1}{\sqrt{n}} \big(\Phi_{kk'}^{(n)} \circ N_{dk}^{(n)}(nt) - \lambda_{dk}^{(n)} p_{kk'}^{(n)} nt\big).$$

From those definitions, we have

$$\widehat{\Phi}_{k'k}^{(n)}(\overline{N}_{dk'}^{(n)}(t)) = \frac{1}{\sqrt{n}} \big(\Phi_{k'k}^{(n)}(N_{dk'}^{(n)}(nt)) - p_{k'k} N_{dk'}^{(n)}(nt)\big)$$

$$= \widehat{\Phi_{k'k}^{(n)} \circ N_{dk'}^{(n)}}(t) - p_{k'k} \widehat{N}_{dk'}^{(n)}(t). \tag{5.8}$$

Similarly,

$$\widehat{V}_{sk}^{(n)}(\overline{N}_{ak}^{(n)}(t)) = \widehat{V_{sk}^{(n)} \circ N_{ak}^{(n)}}(t) - m_{sk}^{(n)} \widehat{N}_{ak}^{(n)}(t). \tag{5.9}$$

Substituting (5.8) into the n-th diffusion scaling version of (5.1), we have

$$\widehat{N}_{ak}^{(n)}(t) = \widehat{N}_{ek}^{(n)}(t) + \sum_{k' \in K} \widehat{\Phi_{k'k}^{(n)} \circ N_{dk'}^{(n)}}(t)$$

$$= \widehat{N}_{ek}^{(n)}(t) + \sum_{k' \in K} \big(\widehat{\Phi}_{k'k}^{(n)}(\overline{N}_{dk'}^{(n)}(t)) + p_{k'k} \widehat{N}_{dk'}^{(n)}(t)\big).$$

Substituting the *n*-th diffusion scaling version of (5.2) into this equation, we have

$$\widehat{N}_{ak}^{(n)}(t) = \widehat{N}_{ek}^{(n)}(t) + \sum_{k' \in K} \left(\widehat{\Phi}_{k'k}^{(n)}(\overline{N}_{dk'}^{(n)}(t)) + p_{k'k}(\widehat{N}_{ak'}^{(n)}(t) - \widehat{L}_{k'}^{(n)}(t)) \right),$$

which is expressed as, in vector-matrix notation,

$$\widehat{\boldsymbol{N}}_a^{(n)}(t) = \widehat{\boldsymbol{N}}_e^{(n)}(t) + \sum_{k' \in K} \widehat{\boldsymbol{\Phi}}_{k'}^{(n)}(\overline{N}_{dk'}^{(n)}(t)) + P^{\mathrm{T}}(\widehat{\boldsymbol{N}}_a^{(n)}(t) - \widehat{\boldsymbol{L}}^{(n)}(t)).$$

Hence, multiplying $Q \equiv (I - P^{\mathrm{T}})^{-1}$ from the left, we have

$$\widehat{\boldsymbol{N}}_a^{(n)}(t) = Q\widehat{\boldsymbol{N}}_e^{(n)}(t) + Q\sum_{k' \in K} \widehat{\boldsymbol{\Phi}}_{k'}^{(n)}(\overline{N}_{dk'}^{(n)}(t)) - QP^{\mathrm{T}}\widehat{\boldsymbol{L}}^{(n)}(t). \tag{5.10}$$

We further substitute (5.10) into the *n*-th diffusion version of (5.4), then, using (5.9) and again (5.10), arrive at

$$\begin{aligned}
\widehat{\boldsymbol{W}}^{(n)}(t) &= C\widehat{\boldsymbol{V}_s^{(n)} \circ \boldsymbol{N}_a^{(n)}}(t) + \sqrt{n}t(\boldsymbol{\rho}^{(n)} - \mathbf{1}) + \widehat{\boldsymbol{Y}}^{(n)}(t) \\
&= C\left(\widehat{\boldsymbol{V}}_s^{(n)}(\overline{\boldsymbol{N}}_a^{(n)}(t)) + M^{(n)}\widehat{\boldsymbol{N}}_a^{(n)}(t) \right) + \sqrt{n}t(\boldsymbol{\rho}^{(n)} - \mathbf{1}) + \widehat{\boldsymbol{Y}}^{(n)}(t) \\
&= C\left(\widehat{\boldsymbol{V}}_s^{(n)}(\overline{\boldsymbol{N}}_a^{(n)}(t)) + M^{(n)}Q(\widehat{\boldsymbol{N}}_e^{(n)}(t) + \sum_{k' \in K} \widehat{\boldsymbol{\Phi}}_{k'}^{(n)}(\overline{N}_{dk'}^{(n)}(t))) \right) \\
&\qquad - CM^{(n)}QP^{\mathrm{T}}\widehat{\boldsymbol{L}}^{(n)}(t) + \sqrt{n}t(\boldsymbol{\rho}^{(n)} - \mathbf{1}) + \widehat{\boldsymbol{Y}}^{(n)}(t). \tag{5.11}
\end{aligned}$$

We now formally define the state space collapse according to [65].

Definition 5.1. *State space collapse is said to hold for $K \times J$ nonnegative matrix $\Delta^{(n)}$, which is called a lifting matrix, if ki-entry of $\Delta^{(n)}$ vanishes if $\beta(k) \neq i$ and, for each $t \geq 0$,*

$$\|\widehat{\boldsymbol{L}}^{(n)}(\cdot) - \Delta^{(n)}\widehat{\boldsymbol{W}}^{(n)}(\cdot)\|_t \xrightarrow{P} 0, \qquad n \to \infty, \tag{5.12}$$

where $\xrightarrow{P}$ stands for convergence in probability.

Assume that state space collapse holds for the sequence of lifting matrices $\Delta^{(n)}$, and $\Delta^{(n)}$ converges to $\Delta \neq 0$. In particular, we assume FCFS service discipline. Then, as we expected, it can be proved that the condition (5.12) holds if

$$\Delta_{ki}^{(n)} = \frac{\lambda_{ak}^{(n)}}{\sum_{\beta(k)=i} \lambda_{ak}^{(n)} m_{sk}^{(n)}}, \tag{5.13}$$

and therefore

$$CM^{(n)}\Delta^{(n)} = I. \tag{5.14}$$

We write the condition (5.12) as

$$\widehat{\boldsymbol{L}}^{(n)}(t) = \Delta^{(n)}\widehat{\boldsymbol{W}}^{(n)}(t) + \widehat{\boldsymbol{\epsilon}}^{(n)}(t) \quad \text{with} \quad \|\widehat{\boldsymbol{\epsilon}}^{(n)}(t)\|_t \xrightarrow{P} 0, \quad n \to \infty,$$

and substitute $\widehat{\boldsymbol{L}}^{(n)}(t)$ into (5.11), then we have

$$(I + CM^{(n)}QP^{\mathrm{T}}\Delta^{(n)})\widehat{\boldsymbol{W}}^{(n)}(t) = C\big(\widehat{\boldsymbol{V}}_s^{(n)}(\overline{\boldsymbol{N}}_a^{(n)}(t)) + M^{(n)}Q(\widehat{\boldsymbol{N}}_e^{(n)}(t) + \sum_{k' \in K} \widehat{\boldsymbol{\Phi}}_{k'}^{(n)}(\overline{N}_{dk'}^{(n)}(t)))$$

$$- CM^{(n)}QP^{\mathrm{T}}\widehat{\boldsymbol{\epsilon}}^{(n)}(t) + \sqrt{n}t(\boldsymbol{\rho}^{(n)} - \mathbf{1}) + \widehat{\boldsymbol{Y}}^{(n)}(t). \quad (5.15)$$

Define $G^{(n)}$ as

$$G^{(n)} = CM^{(n)}QP^{\mathrm{T}}\Delta^{(n)}.$$

Because of (5.14), we have

$$I + G^{(n)} = CM^{(n)}(I + QP^{\mathrm{T}})\Delta^{(n)} = CM^{(n)}Q\Delta^{(n)}. \quad (5.16)$$

We now assume the heavy traffic conditions: For some constant $c_k \in \mathbb{R}$,

$$\lambda_{ek}^{(n)} \to \lambda_{ek}, \qquad m_{sk}^{(n)} \to m_{sk}, \qquad \sqrt{n}(\rho_k^{(n)} - 1) \to c_k, \quad (5.17)$$

where $\rho_k = \lambda_{ak}m_{sk}$ and λ_{ak} is given by

$$\boldsymbol{\lambda}_a = Q\boldsymbol{\lambda}_e.$$

For diffusion approximation, we assume the finiteness of variances. Let

$$(\sigma_{ek}^{(n)})^2 = \mathbb{E}((T_{ek}^{(n)}(1))^2), \qquad (\sigma_{sk}^{(n)})^2 = \mathbb{E}((T_{sk}^{(n)}(1))^2), \qquad k \in K,$$

where we use the convention that $(\sigma_{ek}^{(n)})^2 = 0$ if $\lambda_{ek}^{(n)} = 0$, and assume that

$$\sigma_{ek}^{(n)} \to \sigma_{ek}, \qquad \sigma_{sk}^{(n)} \to \sigma_{sk}. \quad (5.18)$$

Because we did not choose specific sequences for the inter-arrival times $T_{ek}^{(n)}(\ell)$ and the service times $T_{sk}^{(n)}(\ell)$ as we did for the single-class GJN, we here require the following Lindeberg type conditions.

$$\max_{k \in K} \sup_{n \in \mathbb{N}} \mathbb{E}\big((T_{ek}^{(n)}(1))^2; T_{ek}^{(n)}(1) > n\big) \to 0, \quad (5.19)$$

$$\max_{k \in K} \sup_{n \in \mathbb{N}} \mathbb{E}\big((T_{sk}^{(n)}(1))^2; T_{sk}^{(n)}(1) > n\big) \to 0. \quad (5.20)$$

Under these assumptions, it can be proved the following convergence in distribution.

$$\widehat{\boldsymbol{V}}_s^{(n)}(\overline{\boldsymbol{N}}_a^{(n)}(t)) \xrightarrow{w} \mathrm{diag}(\{\lambda_{sk}^{\frac{1}{2}}\sigma_{sk}\})\boldsymbol{B}_s^*(\cdot) \quad (5.21)$$

$$\widehat{\boldsymbol{N}}_e^{(n)}(t) \xrightarrow{w} \mathrm{diag}(\{\lambda_{ek}^{\frac{3}{2}}\sigma_{ek}\})\boldsymbol{B}_e^*(\cdot) \quad (5.22)$$

$$\widehat{\boldsymbol{\Phi}}_{k'}^{(n)}(\overline{N}_{dk'}^{(n)}(t)) \xrightarrow{w} \lambda_{ak'}^{\frac{1}{2}}\Gamma_{k'}\boldsymbol{B}_{k'}^*(\cdot), \quad (5.23)$$

where $\boldsymbol{B}_s^*(\cdot)$, $\boldsymbol{B}_e^*(\cdot)$ and $\boldsymbol{B}_k^*(\cdot)$ are independent $|K|$, $|K_e|$ and $|K|$ dimensional standard Brownian motions, and $\Gamma_{k'}\Gamma_{k'}^{\mathrm{T}}$ is the $K \times K$ covariance matrix defined by

$$[\Gamma_{k'}\Gamma_{k'}^{\mathrm{T}}]_{\ell\ell'} = \begin{cases} p_{k\ell}(1 - p_{k\ell}), & \ell' = \ell, \\ -p_{k\ell}p_{k\ell'}, & \ell' \neq \ell. \end{cases}$$

Let $G = \lim_{n\to\infty} G^{(n)}$, and assume that $I + G$ is invertible. Let $R \equiv (I + G)^{-1}$. Taking (5.15) and (5.21)–(5.23) into account, define $\boldsymbol{X}(t)$ as

$$\boldsymbol{X}(t) = RC\Big(\Lambda^{\frac{1}{2}}\operatorname{diag}(\{\sigma_{sk}\})\boldsymbol{B}_s^*(t) + MQ\big(\operatorname{diag}(\{\lambda_{ek}^{\frac{3}{2}}\sigma_{ek}\})\boldsymbol{B}_e^*(t)$$
$$+ \sum_{k'\in K}\lambda_{ak'}^{\frac{1}{2}}\Gamma_{k'}\boldsymbol{B}_{k'}^*(t)\big)\Big) + tR\boldsymbol{c}.$$

Since $R^{(n)} \equiv (I + G^{(n)})^{-1}$ is well defined for sufficiently large n, and for such n, (5.15) is can be written as

$$\widehat{\boldsymbol{W}}^{(n)}(t) = \boldsymbol{X}^{(n)}(t) + R^{(n)}\widehat{\boldsymbol{Y}}^{(n)}(t), \tag{5.24}$$

where

$$\boldsymbol{X}^{(n)}(t) = R^{(n)}C\big(\widehat{\boldsymbol{V}}_s^{(n)}(\overline{\boldsymbol{N}}_a^{(n)}(t)) + M^{(n)}Q(\widehat{\boldsymbol{N}}_e^{(n)}(t) + \sum_{k'\in K}\widehat{\boldsymbol{\Phi}}_{k'}^{(n)}(\overline{N}_{dk'}^{(n)}(t)))$$
$$- R^{(n)}CM^{(n)}QP^{\mathrm{T}}\widehat{\boldsymbol{\epsilon}}^{(n)}(t) + \sqrt{n}tR^{(n)}(\boldsymbol{\rho}^{(n)} - \boldsymbol{1}).$$

It can be shown that $\boldsymbol{X}^{(n)}(\cdot) \xrightarrow{w} \boldsymbol{X}(\cdot)$. Furthermore, $\widehat{Y}_k^{(n)}(t)$ increases only when $\widehat{W}_k^{(n)}(t) = 0$. Similar to the reflection mapping, if R is a completely $\mathcal{S}$-matrix, this suggests that (5.24) implies that there are unique nonnegative $\boldsymbol{W}(\cdot)$ and $\boldsymbol{Y}(\cdot)$ such that

$$\boldsymbol{W}(t) = \boldsymbol{X}(t) + R\boldsymbol{Y}(t), \tag{5.25}$$

This implication is actually proved in [67]. One also needs to check the tightness for $\boldsymbol{W}^{(n)}(\cdot) \xrightarrow{w} \boldsymbol{W}(\cdot)$ and $\boldsymbol{Y}^{(n)}(\cdot) \xrightarrow{w} \boldsymbol{Y}(\cdot)$.

Theorem 5.1 (Theorem 7.1 of [66])**.** Under the assumptions (5.17) and (5.18)–(5.23), if the service discipline is FCFS and if $I + G$ is invertible and its inverse is a completely $\mathcal{S}$-matrix, then we have

$$(\boldsymbol{X}^{(n)}(\cdot), \boldsymbol{W}^{(n)}(\cdot), \boldsymbol{Y}^{(n)}(\cdot), \boldsymbol{L}^{(n)}(\cdot)) \xrightarrow{w} (\boldsymbol{X}(\cdot), \boldsymbol{W}(\cdot), \boldsymbol{Y}(\cdot), \boldsymbol{L}(\cdot)).$$

Furthermore, this convergence still holds for any head of line service discipline as long as the state collapse condition holds for some sequence $\Delta^{(n)}$ converging Δ.

For the multi-class GJN, there is no results about the continuity of the stationary distribution. The difficulty for this case is remarked in Section 5 of [24].

6. Halfin-Whitt regime

As we have observed in Section 3.4, the diffusion approximation by SRBM is still possible for the GJN with many server nodes. However, it ignores the influence of the number of idle servers which varies from 0 to the total number of servers at one node. So, the approximation may be too coarse. On the other hand, if there are infinitely many servers at a node, one may apply the central limit theorem to get asymptotics for approximation. In this case, there is no queue, and this is another extreme. Furthermore, the limiting process is Gaussian, and therefore not a diffusion process. So, we will not discuss about them. Comprehensive discussions about such a limiting process is found in Section 9 of [25].

One may wonder what approximation is possible between those two extremes. This motivated Halfin and Whitt [27] to increase the number n of servers as well as the arrival rate

of customers in a single queue with many servers while keeping service time distributions. In this case, a network model may be too complicate because different nodes may have different numbers of customers. So, this type of limits has been studied mainly for a single node, that is, a single queue. We here refer to results in [27] for such a model.

Assume that the n-th system has n servers, and let $L^{(n)}(t)$ be the number of customers in system at time t. Customers arrives according to a renewal process, and their service times are *i.i.d.*. An arriving customer has to wait if all servers are busy. An important assumption here is that the service time distribution is exponential. This is the crucial assumption which makes analysis possible and results tractable.

We will use the following notations. $\lambda^{(n)}$ is the arrival rate of customers, μ is the service rate, $\rho^{(n)} = \lambda^{(n)}/(n\mu)$. We assume the following heavy traffic condition. There is a constant β such that

$$\lim_{n\to\infty} \sqrt{n}(1 - \rho^{(n)}) = \beta, \qquad \beta > 0. \tag{6.1}$$

We further assume the following condition on the moments of the inter-arrival time $T^{(n)}(\ell)$ for the n-th system.

$$\lim_{n\to\infty} (\lambda^{(n)})^2 Var(T^{(n)}(1)) = c^2, \tag{6.2}$$

$$\sup_{n\geq 1}(\lambda^{(n)})^3 \mathbb{E}((T^{(n)}(1))^3) < \infty. \tag{6.3}$$

Define the following diffusion scaling for the number of customers in system.

$$\widetilde{L}^{(n)}(t) = \frac{1}{\sqrt{n}}(L^{(n)}(t) - n).$$

The following theorems are obtained by Halfin and Whitt [27].

Theorem 6.1 (Theorem 3 of [27]). *Assume (6.1)–(6.3). If $\widetilde{L}^{(n)}(0) \xrightarrow{w} Z(0)$, then $\widetilde{L}^{(n)}(\cdot) \xrightarrow{w} Z(\cdot)$, where $Z(\cdot) \equiv \{Z(t); t \geq 0\}$ is a diffusion process characterized by the following stochastic integration.*

$$Z(t) = Z(0) + \int_0^t \sigma(Z(u))dB(u) + \int_0^t m(Z(u))du, \qquad t \geq 0, \tag{6.4}$$

where $B(\cdot)$ is the standard Brownian motion, and $\sigma^2(x)$ and $m(x)$ are given by

$$m(x) = -\mu(\beta + \mu x 1(x < 0)), \qquad \sigma^2(x) = \mu(1 + c^2).$$

Theorem 6.2 (Theorem 4 of [27]). *Assume (6.1)–(6.3), and let $\widetilde{L}^{(n)}(\infty)$ and $Z(\infty)$ be random variables subject to the stationary distributions of $\widetilde{L}^{(n)}(\cdot)$ and $Z(\cdot)$, respectively, of Theorem 6.1, then $\widetilde{L}^{(n)}(\infty) \xrightarrow{w} Z(\infty)$. Furthermore, let $\beta^* = 2\beta/(1 + c^2)$ and let $\Phi_N(x)$ be the standard normal distribution function, then*

$$\lim_{n\to\infty} \mathbb{P}(L^{(n)} \geq n) = \mathbb{P}(Z(\infty) \geq 0) = (1 + \sqrt{2\pi}\beta^*\Phi_N(\beta^*)exp((\beta^*)^2/2))^{-1},$$

and

$$\mathbb{P}(Z(\infty) > x | Z(\infty) \geq 0) = e^{-x\beta^*}, \qquad x > 0,$$

$$\mathbb{P}(Z(\infty) \leq x | Z(\infty) \leq 0) = \frac{\Phi_N(\beta^* + x)}{\Phi_N(\beta^*)}, \qquad x \leq 0.$$

This approximation is popular for the performance evaluation of call centers, and there have been several attempts for its generalization concerning the service time distribution (e.g., see [23, 55]). In particular, Reed [55] identifies the limiting process $Z(\cdot)$ for $\widetilde{L}^{(n)}(\cdot)$ as the solution of a certain integral equation. However, it is hard to see the property of this solution. So, Gamarnik and Goldberg [23, 55]) derive bounds for the tail probability of the stationary distribution.

7. Concluding Remarks

From the diffusion approximations in Sections 2, 3 and 5, we can see the following facts.

(7a) For both single-class and multi-class queueing networks, SRBM is obtained as a process limit of the sequence of diffusion scaling processes.

(7b) The SRBM for a single-class GJN is determined by the asymptotic means and variances of the counting processes $N_{ei}(\cdot)$ and $N_{si}(\cdot)$ and random routing function $\Phi_{ij}(\cdot)$.

(7c) The SRBM for a multi-class GJN requires the asymptotic mean and variance of the accumulated workload $V_{si}(\cdot)$ in addition to those for a single-class queueing network.

For (7b), we can get those data of $N_{ei}(\cdot)$ and $N_{si}(\cdot)$ from their logarithmic moment generating functions as defined in (4.2). Those logarithmic generating functions play important roles in the stationary equation of the GJN, which is the reason why they arise in the tail decay rate of the stationary distribution in Section 4.2.

Although SRBM is very important for diffusion approximation, very little is known about its stationary distribution for the more than two dimensional case. Thus, it is often to use a product form stationary distribution, which is obtained only when the so called skew symmetric condition holds. However, a choice of the primitive data is very limited under this condition. Thus, it would be very nice if we can find some features of the stationary distribution.

The author and his colleagues [20] recently studied to decompose the stationary distribution of SRBM into two marginals. They show that, if the stationary distribution is decomposable, then the marginal distributions are obtained as lower dimensional SRBM, and their primitive data are identified. This suggests that we may have more flexible choice of the primitive data than the skew symmetric case, so we may have a better approximation.

There are two interesting observations in Section 2.4 and Appendix D. One is the fact that time scaling is not essential to get the limit of a sequence of the stationary distributions for diffusion compatible approximation. This may be useful for simulation to get the stationary distribution of the SRBM. Another is that large deviations rate functions are closely related to diffusion approximation through logarithmic moment generating function. Both observations may not be surprising, but it is notable that they are mathematically confirmed.

Appendix

A. Proof of Lemma 2.2

We only prove the first formula of (2.26) because (2.27) can be obtained similarly. By the assumption (2.18), $\widetilde{F}_e(w)$ with complex variable w is analytic in a neighborhood of the origin in the complex plain $\mathbb{C}$. Since

$$\frac{\partial}{\partial w}\big(e^z \widetilde{F}_e(w) - 1\big)\big|_{(z,w)=(0,0)} = \widetilde{F}_e'(0) = \mathbb{E}(T_e) \neq 0,$$

$e^z \widetilde{F}_e(w) - 1 = 0$ has a unique root $w = w(z)$ which is analytic in z in a neighborhood of the origin by the implicit function theorem (e.g., see Theorem 3.11 of the Volume II of [50]). Hence, $\gamma_e(z)$ is analytic in a neighborhood of the origin, and therefore, it follows from Taylor expansion that

$$\gamma_e(\theta) = -\eta(\theta) = -\Big(\eta(0) + \eta'(0)\theta + \frac{1}{2}\eta''(0)\theta^2 + o(\theta^2)\Big), \qquad \theta \to 0.$$

Obviously, $\eta(0) = 0$. By differentiating both sides of $e^\theta \widetilde{F}_e(\eta) = 1$, we have

$$\widetilde{F}_e'(\eta)\frac{d\eta}{d\theta} + \widetilde{F}_e(\eta) = 0, \tag{A.1}$$

where η is a function θ, namely, $\eta(\theta)$, but we drop θ for simplicity. This implies that $\eta'(0) = -\lambda_e$. Differentiating (A.1), we have

$$\widetilde{F}_e'(\eta)\frac{d^2\eta}{d\theta^2} + \widetilde{F}_e''(\eta)\Big(\frac{d\eta}{d\theta}\Big)^2 + \widetilde{F}_e'(\eta)\frac{d\eta}{d\theta} = 0, \tag{A.2}$$

which yields that $\eta''(0) = -\lambda_e^3 \sigma_e^2$. Thus, we obtain (2.26).

B. Proof of Theorem 2.1

Let $\beta^{(n)} = 1 - \rho^{(n)}$. Since (2.29) and (2.30) hold, (2.28) yields, for $\theta \leq 0$,

$$\lim_{n\to\infty} \varphi^{(n)}(\beta^{(n)}\theta, -\lambda_e^{(n)}\beta^{(n)}\theta + o(\beta^{(n)}\theta), \lambda_s^{(n)}\beta^{(n)}\theta + o(\beta^{(n)}\theta)) = \frac{2}{2 - \lambda_e^2(\sigma_e^2 + \sigma_s^2)\theta}, \tag{B.1}$$

where the condition (2.14) is tacitly used to the small order terms $o(\beta^{(n)}\theta)$ as $n \to \infty$. To complete the proof, we need to show that

$$\lim_{n\to\infty} \varphi^{(n)}(\beta^{(n)}\theta, 0, 0) = \frac{2}{2 - \lambda_e^2(\sigma_e^2 + \sigma_s^2)\theta}, \qquad \theta \leq 0. \tag{B.2}$$

For this, denote the right-hand side of (B.2) by $f(\theta)$. Then, by a triangular inequalities, we have

$$\begin{aligned}
&|\varphi^{(n)}(\beta^{(n)}\theta, 0, 0) - f(\theta)| \\
&\quad \leq |\varphi^{(n)}(\beta^{(n)}\theta, -\lambda_e^{(n)}\beta^{(n)}\theta + o(\beta^{(n)}\theta), \lambda_s^{(n)}\beta^{(n)}\theta + o(\beta^{(n)}\theta)) - f(\theta)| \\
&\quad\quad + |\varphi^{(n)}(\beta^{(n)}\theta, 0, 0) - \varphi^{(n)}(\beta^{(n)}\theta, -\lambda_e^{(n)}\beta^{(n)}\theta + o(\beta^{(n)}\theta), \lambda_s^{(n)}\beta^{(n)}\theta + o(\beta^{(n)}\theta))|.
\end{aligned}$$

From (B.1), the first absolute term in the right-hand side of this inequality vanishes as $n \to \infty$. Thus, we only need to show that its second absolute term vanishes as $n \to \infty$. For this, we write down it in terms of expectations for $\theta \le 0$.

$$\text{The second absolute term} = \left| \mathbb{E}\left(e^{\beta^{(n)}\theta L} \left(1 - e^{(-\lambda_e^{(n)}\beta^{(n)}\theta + o(\beta^{(n)}\theta))R_e + (\lambda_s^{(n)}\beta^{(n)}\theta + o(\beta^{(n)}\theta))R_s^{(n)}} \right) \right) \right|$$

$$\le \left| 1 - e^{(-\lambda_e^{(n)}\beta^{(n)}\theta + o(\beta^{(n)}\theta))R_e} \right|$$

$$+ \left| e^{(-\lambda_e^{(n)}\beta^{(n)}\theta + o(\beta^{(n)}\theta))R_e} \left(1 - e^{(\lambda_s^{(n)}\beta^{(n)}\theta + o(\beta^{(n)}\theta))R_s^{(n)}} \right) \right|.$$

The right-hand side converges to 0 by the bounded convergence theorem because $\theta \le 0$ and $\beta^{(n)} > 0$ converges to 0. This completes the proof of Theorem 2.1 because the convergence of Laplace transfers implies the convergence of the corresponding distributions.

C. Normalizing factor in Theorem 3.7

We check the normalizing factor to be correct in Theorem 3.7. For this, we derive a limiting distribution for the one-dimensional marginal. It follows from the definition of $N^{(n)}$ that

$$N^{(n)}(nt) > [kt] \quad \text{if and only if} \quad \sum_{\ell=1}^{[kt]} \tau_\ell^{(n)} < nt.$$

Hence,

$$\frac{1}{\sqrt{nt}}(N^{(n)}(nt) - \lambda^{(n)}nt) > \frac{1}{\sqrt{nt}}([kt] - \lambda^{(n)}nt)$$

$$\text{if and only if} \quad \frac{1}{\sqrt{[kt]}\sigma^{(n)}} \sum_{\ell=1}^{[kt]} (\tau_\ell^{(n)} - m^{(n)}) < \frac{1}{\sqrt{[kt]}\sigma^{(n)}}(nt - m^{(n)}[kt]).$$

We let $n, k \to \infty$ so that $\frac{1}{\sqrt{[kt]}\sigma^{(n)}}(nt - m^{(n)}[kt])$ converges to an arbitrarily given x. Then,

$$\lambda^{(n)}nt - [kt] \sim \lambda^{(n)}\sqrt{[kt]}\sigma^{(n)}x,$$

where we recall that $\lambda^{(n)} = 1/m^{(n)}$, and therefore

$$\frac{1}{\sqrt{nt}}([kt] - \lambda^{(n)}nt) \sim -\frac{\lambda^{(n)}\sqrt{[kt]}\sigma^{(n)}x}{\sqrt{(\sqrt{[kt]}\sigma^{(n)}x + m^{(n)}[kt])t}} \sim -\frac{\lambda^{(n)}\sigma^{(n)}x}{\sqrt{m^{(n)}}} = -\sqrt{(\lambda^{(n)})^3}\sigma^{(n)}x,$$

$$[kt] \sim \lambda^{(n)}nt \sim \lambda nt$$

Thus, we have

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{1}{\sqrt{nt}}(N^{(n)}(nt) - \lambda^{(n)}nt) > -\sqrt{(\lambda^{(n)})^3}\sigma^{(n)}x \right)$$

$$= \lim_{n \to \infty} \mathbb{P}\left(\frac{1}{\sqrt{[\lambda nt]}\sigma^{(n)}} \sum_{\ell=1}^{[\lambda nt]} (\tau_\ell^{(n)} - m^{(n)}) < x \right).$$

Since the right-hand side is the standard normal distribution function, $\sigma\lambda^{\frac{3}{2}}$ correctly appears in the theorem.

D. Proof of Theorem 4.3

We only need to prove that (4.8) implies the convergence of $\mathcal{D}^{(n)}$ to $\widetilde{\mathcal{D}}$. A major problem is to show that $1/\sqrt{n}$ in the second formula of (4.8) is a right scaling factor. For this, we first derive the stationary equation similar to (2.21) for a general GJN. The reason why we consider general d rather than $d = 2$ is that there are not so much difference between them and the general case may have own interest. For $i, j \in J$ and $\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi} \in \mathbb{R}^d$, let

$$\varphi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) = \mathbb{E}(e^{\langle \boldsymbol{\theta}, \boldsymbol{L}(0)\rangle + \langle \boldsymbol{\eta}, \boldsymbol{R}_e(0)\rangle + \langle \boldsymbol{\xi}, \boldsymbol{R}_s(0)\rangle}),$$

$$\varphi_{i0}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) = \mathbb{E}(e^{\langle \boldsymbol{\theta}, \boldsymbol{L}(0)\rangle + \langle \boldsymbol{\eta}, \boldsymbol{R}_e(0)\rangle + \langle \boldsymbol{\xi}, \boldsymbol{R}_s(0)\rangle} 1(L_i(0) = 0)),$$

$$\varphi_{ei}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) = \mathbb{E}_{ei}(e^{\langle \boldsymbol{\theta}, \boldsymbol{L}(0-)\rangle + \langle \boldsymbol{\eta}, \boldsymbol{R}_e(0-)\rangle + \langle \boldsymbol{\xi}, \boldsymbol{R}_s(0-)\rangle}),$$

$$\varphi_{ei0-}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) = \mathbb{E}_{ei}(e^{\langle \boldsymbol{\theta}, \boldsymbol{L}(0-)\rangle + \langle \boldsymbol{\eta}, \boldsymbol{R}_e(0-)\rangle + \langle \boldsymbol{\xi}, \boldsymbol{R}_s(0-)\rangle} 1(L_i(0-) = 0)),$$

$$\varphi_{si}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) = \mathbb{E}_{si}(e^{\langle \boldsymbol{\theta}, \boldsymbol{L}(0-)\rangle + \langle \boldsymbol{\eta}, \boldsymbol{R}_e(0-)\rangle + \langle \boldsymbol{\xi}, \boldsymbol{R}_s(0-)\rangle}),$$

$$\varphi_{si0+}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) = \mathbb{E}_{si}(e^{\langle \boldsymbol{\theta}, \boldsymbol{L}(0+)\rangle + \langle \boldsymbol{\eta}, \boldsymbol{R}_e(0+)\rangle + \langle \boldsymbol{\xi}, \boldsymbol{R}_s(0+)\rangle} 1(L_i(0+) = 0)),$$

$$\varphi_{sij0}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) = \mathbb{E}_{si}(e^{\langle \boldsymbol{\theta}, \boldsymbol{L}(0+)\rangle + \langle \boldsymbol{\eta}, \boldsymbol{R}_e(0+)\rangle + \langle \boldsymbol{\xi}, \boldsymbol{R}_s(0+)\rangle} 1(L_i(0+) = 0, L_j(0-) = 0)),$$

where $\mathbb{E}_{ui}$ represents the expectation with respect to the Palm measure concerning N_{ui} for $u = e, s$, which is defined similarly to (2.20). We then apply the rate conservation law similarly to (2.19) for the following $X(t)$.

$$X(t) = \begin{cases} g_i(\boldsymbol{\theta})e^{\langle \boldsymbol{\theta}, \boldsymbol{L}(t)\rangle + \langle \boldsymbol{\eta}, \boldsymbol{R}_e(t)\rangle + \langle \boldsymbol{\xi}, \boldsymbol{R}_s(t)\rangle}, & L_i(t) = 0, \\ e^{\langle \boldsymbol{\theta}, \boldsymbol{L}(t)\rangle + \langle \boldsymbol{\eta}, \boldsymbol{R}_e(t)\rangle + \langle \boldsymbol{\xi}, \boldsymbol{R}_s(t)\rangle}, & L_i(t) \geq 1. \end{cases}$$

This yields the following formula for $\boldsymbol{\theta} \leq \boldsymbol{0}$, where $\theta_0 = 0$.

$$
\begin{aligned}
&-\sum_{i \in J}(\eta_i + \xi_i)\varphi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) + \sum_{i \in J}(\eta_i(1 - g_i(\boldsymbol{\theta})) + \xi_i)\varphi_{i0}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) \\
&+ \sum_{i \in J}\lambda_{ei}(e^{\theta_i}\widetilde{F}_{ei}(\eta_i) - 1)\varphi_{ei}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) \\
&+ \sum_{i \in J}\lambda_{ei}\big(e^{\theta_i}\widetilde{F}_{ei}(\eta_i)(\widetilde{F}_{si}(\xi_i) - 1) + 1 - g_i(\boldsymbol{\theta})\big)\varphi_{ei0-}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) \\
&+ \sum_{i \in J}\lambda_{si}\Big(\sum_{j \in J\cup\{0\}} e^{-\theta_i+\theta_j}p_{ij}\widetilde{F}_{si}(\xi_i) - 1\Big)\varphi_{si}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) \\
&+ \sum_{i \in J}\sum_{j \in J\cup\{0\}}\lambda_{si}e^{-\theta_i+\theta_j}p_{ij}(g_i(\boldsymbol{\theta}) - \widetilde{F}_{si}(\xi_i))\varphi_{si0+}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) \\
&+ \sum_{i \in J}\sum_{j \in J}\lambda_{si}e^{-\theta_i+\theta_j}p_{ij}g_i(\boldsymbol{\theta})(\widetilde{F}_{sj}(\xi_j) - 1)\varphi_{sij0}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\xi}) = 0. \quad (\text{D.1})
\end{aligned}
$$

We next choose η_i, ξ_i and $g_i(\boldsymbol{\theta})$ in such a way that

$$e^{\theta_i}\widetilde{F}_{ei}(\eta_i) = 1, \qquad \widetilde{F}_{si}(\xi_i)\sum_{j \in J\cup\{0\}} e^{-\theta_i+\theta_j}p_{ij} = 1, \qquad g_i(\boldsymbol{\theta}) = \widetilde{F}_{si}(\xi_i), \qquad i \in J.$$

Denote these η_i and ξ_i by $\eta_i(\theta_i)$ and $\xi_i(\boldsymbol{\theta})$, and let $\boldsymbol{\eta}(\boldsymbol{\theta})$ and $\boldsymbol{\xi}(\boldsymbol{\theta})$ be their vectors, then we

arrive at

$$-\sum_{i\in J}(\eta_i(\theta_i)+\xi_i(\boldsymbol{\theta}))\varphi(\boldsymbol{\theta},\boldsymbol{\eta}(\boldsymbol{\theta}),\boldsymbol{\xi}(\boldsymbol{\theta}))$$

$$+\sum_{i\in J}(\eta_i(\theta_i)(1-\widetilde{F}_{si}(\xi_i(\boldsymbol{\theta})))+\xi_i(\boldsymbol{\theta}))\varphi_{i0}(\boldsymbol{\theta},\boldsymbol{\eta}(\boldsymbol{\theta}),\boldsymbol{\xi}(\boldsymbol{\theta}))$$

$$+\sum_{i,j\in J,i\neq j}\lambda_{si}e^{-\theta_i+\theta_j}p_{ij}\widetilde{F}_{si}(\xi_i(\boldsymbol{\theta}))(\widetilde{F}_{sj}(\xi_j(\boldsymbol{\theta}))-1)\varphi_{sij0}(\boldsymbol{\theta},\boldsymbol{\eta}(\boldsymbol{\theta}),\boldsymbol{\xi}(\boldsymbol{\theta}))=0. \qquad (\text{D.2})$$

For $d=2$, it follows from (4.3) and (4.4) that

$$\eta_i(\theta_i)=-\gamma_{ei}(\theta_i),\qquad \xi_i(\boldsymbol{\theta})=-\gamma_{di}(\boldsymbol{\theta}).$$

We now consider the n-th GJN with $d=2$, and its characteristics are indexed by superscript "(n)". Hence, recalling the definition of γ_+, (D.2) becomes, for $\boldsymbol{\theta}\leq\boldsymbol{0}$,

$$\gamma_+^{(n)}(\boldsymbol{\theta})\varphi^{(n)}(\boldsymbol{\theta},-\boldsymbol{\gamma}_e^{(n)}(\boldsymbol{\theta}),-\boldsymbol{\gamma}_d^{(n)}(\boldsymbol{\theta}))-\sum_{i=1,2}\gamma_{di}^{(n)}(\boldsymbol{\theta})\varphi_{i0}^{(n)}(\boldsymbol{\theta},-\boldsymbol{\gamma}_e^{(n)}(\boldsymbol{\theta}),-\boldsymbol{\gamma}_d^{(n)}(\boldsymbol{\theta}))$$

$$-\sum_{i=1,2}\gamma_{ei}^{(n)}(\theta_i)(1-t_i^{-1}(\boldsymbol{\theta}))\varphi_{i0}^{(n)}(\boldsymbol{\theta},-\boldsymbol{\gamma}_e^{(n)}(\boldsymbol{\theta}),-\boldsymbol{\gamma}_d^{(n)}(\boldsymbol{\theta}))$$

$$+\sum_{(i,j)=(1,2),(2,1)}\lambda_{si}^{(n)}e^{-\theta_i+\theta_j}p_{ij}t_i^{-1}(\boldsymbol{\theta})$$

$$\times(t_j^{-1}(\boldsymbol{\theta})-1)\varphi_{sij0}^{(n)}(\boldsymbol{0},-\boldsymbol{\gamma}_e^{(n)}(\boldsymbol{\theta}),-\boldsymbol{\gamma}_d^{(n)}(\boldsymbol{\theta}))=0, \qquad (\text{D.3})$$

where we have used the fact that

$$\varphi_{sij0}^{(n)}(\boldsymbol{\theta},-\boldsymbol{\gamma}_e^{(n)}(\boldsymbol{\theta}),-\boldsymbol{\gamma}_d^{(n)}(\boldsymbol{\theta}))=\varphi_{sij0}^{(n)}(\boldsymbol{0},-\boldsymbol{\gamma}_e^{(n)}(\boldsymbol{\theta}),-\boldsymbol{\gamma}_d^{(n)}(\boldsymbol{\theta})),$$

because $\boldsymbol{L}(0)=\boldsymbol{0}$ implies that $\langle\boldsymbol{\theta},\boldsymbol{L}(0)\rangle=0=\langle\boldsymbol{0},\boldsymbol{L}(0)\rangle$. Thus, using the diffusion scaling $\widehat{\gamma}_{ui}^{(n)}(\boldsymbol{\theta})$ of (4.7) and substituting $\frac{1}{\sqrt{n}}\boldsymbol{\theta}$ into the first variable $\boldsymbol{\theta}$ of $\varphi^{(n)},\varphi_{i0}^{(n)},\varphi_{sij0}^{(n)}$ in (D.3), we have, under diffusion scaling,

$$\widehat{\gamma}_+^{(n)}(\boldsymbol{\theta})\varphi^{(n)}\Big(\frac{1}{\sqrt{n}}\boldsymbol{\theta},-\frac{1}{n}\widehat{\boldsymbol{\gamma}}_e^{(n)}(\boldsymbol{\theta}),-\frac{1}{n}\widehat{\boldsymbol{\gamma}}_d^{(n)}(\boldsymbol{\theta})\Big)$$

$$-\sum_{i=1,2}\widehat{\gamma}_{di}^{(n)}(\boldsymbol{\theta})\varphi_{i0}^{(n)}\Big(\frac{1}{\sqrt{n}}\boldsymbol{\theta},-\frac{1}{n}\widehat{\boldsymbol{\gamma}}_e^{(n)}(\boldsymbol{\theta}),-\frac{1}{n}\widehat{\boldsymbol{\gamma}}_d^{(n)}(\boldsymbol{\theta})\Big)$$

$$-\sum_{i=1,2}\widehat{\gamma}_{ei}^{(n)}(\theta_i)\Big(1-t_i^{-1}\Big(\frac{1}{\sqrt{n}}\boldsymbol{\theta}\Big)\Big)\varphi_{i0}^{(n)}\Big(\frac{1}{\sqrt{n}}\boldsymbol{\theta},-\frac{1}{n}\widehat{\boldsymbol{\gamma}}_e^{(n)}(\boldsymbol{\theta}),-\frac{1}{n}\widehat{\boldsymbol{\gamma}}_d^{(n)}(\boldsymbol{\theta})\Big)$$

$$+\sum_{(i,j)=(1,2),(2,1)}n\lambda_{si}^{(n)}e^{-\theta_i+\theta_j}p_{ij}t_i^{-1}\Big(\frac{1}{\sqrt{n}}\boldsymbol{\theta}\Big)$$

$$\times\Big(t_j^{-1}\Big(\frac{1}{\sqrt{n}}\boldsymbol{\theta}\Big)-1\Big)\varphi_{sij0}^{(n)}\Big(\boldsymbol{0},-\frac{1}{n}\widehat{\boldsymbol{\gamma}}_e^{(n)}(\boldsymbol{\theta}),-\frac{1}{n}\widehat{\boldsymbol{\gamma}}_d^{(n)}(\boldsymbol{\theta})\Big)=0. \qquad (\text{D.4})$$

Since $\frac{1}{n}\widehat{\gamma}_{ui}^{(n)}(\boldsymbol{\theta})\to 0$ as $n\to\infty$ for $u=e,d$ by (4.8), we have, using similar arguments in Appendix B,

$$\lim_{n\to\infty}\varphi^{(n)}\Big(\frac{1}{\sqrt{n}}\boldsymbol{\theta},-\frac{1}{n}\widehat{\boldsymbol{\gamma}}_e^{(n)}(\boldsymbol{\theta}),-\frac{1}{n}\widehat{\boldsymbol{\gamma}}_d^{(n)}(\boldsymbol{\theta})\Big)=\tilde{\varphi}(\boldsymbol{\theta}).$$

Since $t_j^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{\theta}\right) - 1$ has the order $1/\sqrt{n}$ as $n \to \infty$, this and the fact that (D.4) converges to the BAR (3.14) of the limiting SRBM concludes

$$\lim_{n\to\infty} \lambda_{si}^{(n)} \sqrt{n} \varphi_{i0}^{(n)}\left(\frac{1}{\sqrt{n}}\boldsymbol{\theta}, -\frac{1}{n}\widehat{\boldsymbol{\gamma}}_e^{(n)}(\boldsymbol{\theta}), -\frac{1}{n}\widehat{\boldsymbol{\gamma}}_d^{(n)}(\boldsymbol{\theta})\right) = \tilde{\varphi}_i(\boldsymbol{\theta}),$$

and therefore $-(\lambda_{si}^{(n)}\sqrt{n})^{-1}\widehat{\gamma}_{di}^{(n)}(\boldsymbol{\theta})$ is a right term to converge to $\tilde{\gamma}_i(\boldsymbol{\theta})$ so that the continuity of the stationary distribution holds. Since $-\widehat{\gamma}_{di}^{(n)}(\boldsymbol{\theta})$ and $-\gamma_{di}^{(n)}(\boldsymbol{\theta})$ have the same sign and

$$-\gamma_{di}^{(n)}(\boldsymbol{\theta}) = (\tilde{F}_{si}^{(n)})^{-1}(t_i^{-1}(\boldsymbol{\theta})),$$

$-\gamma_{di}^{(n)}(\boldsymbol{\theta}) < 0$ if and only if $t_i(\boldsymbol{\theta}) > 1$, which is the condition to determine Γ_{3-i} for Theorem 4.1. This completes the proof of Theorem 4.3.

References

[1] S. Asmussen: *Applied probability and queues,* 2nd edition (Springer, New York, 2003).

[2] P. Billingsley: *Convergence of probability measures*, 2nd edition (Wiley, New York, 1999).

[3] P. Billingsley: *Probability and measure*, 2nd edition (Wiley, New York, 2000).

[4] J. Blanchet: Private communication (2012).

[5] A.A. Borovkov: Some limit theorems in the theory of mass service, II. *Theory of Probability and its Applications*, **10** (1965), 375–400.

[6] A.A. Borovkov: *Stochastic processes in queueing theory* (Springer-Verlag, New York, 1976).

[7] M. Bramson, J.G. Dai, and J.M. Harrison: Positive recurrence of reflecting Brownian motion in three dimensions. *Annals of Applied Probability*, **20** (2010), 753–783.

[8] A. Budhiraja and C. Lee: Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Mathematics of Operations Research*, **34** (2009), 45–56.

[9] X. Chao, M. Miyazawa, and M. Pinedo: *Queueing networks: customers, signals and product form solutions* (Wiley, Chichester, 1999).

[10] H. Chen and A. Mandelbaum: Stochastic discrete flow networks: Diffusion approximation and bottlenecks. *Annals of Probability*, **19** (1991), 1463–1519.

[11] H. Chen and J.G. Shanthikumar: Fluid limits and diffusion approximations for networks of multi-server queues in heavy traffic. *Discrete Event Dynamic Systems: Theory and Applications*, **4** (1994), 269–291.

[12] H. Chen and D. Yao: *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization* (Springer-Verlag, New York, 2001).

[13] H. Chen and H.Q. Ye: *Methods in Diffusion Approximation for Multi-Server System: Sandwich, Uniform Attraction and State-Space Collapse*, In R.J. Boucherie and N.M. van Dijk (eds.): *Queueing Networks, A Fundamental Approach* (Springer, New York, 2011), Chapter **11**, 489–530.

[14] K.L. Chung: *A Course in Probability Theory*, 3rd edition (Academic Press (An Imprint of Elsevier), San Diego, 2001).

[15] K.L. Chung and R.J. Williams: *Introduction to Stochastic Integration*, 2nd edition (Birkhäuser, Boston, 1990).

[16] J.G. Dai: On positive harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Annals of Applied Probability*, **5** (1995), 49–77.

[17] J.G. Dai and T.G. Kurtz: Characterization of the stationary distribution for a semi-martingale reflecting Brownian motion in a convex polyhedron (Preprint, 1994).

[18] J.G. Dai and M. Miyazawa: Reflecting Brownian motion in two dimensions: Exact asymptotics for the stationary distribution. *Stochastic Systems*, **1** (2011), 146–208.

[19] J.G. Dai, M. Miyazawa, and J. Wu: A multi-dimensional SRBM: geometric views of its product form stationary distribution. arXiv version (http://arxiv.org/abs/1312.1758), 2014. *Queueing Systems* **78** (2014), 313–335.

[20] J.G. Dai, M. Miyazawa, and J. Wu: Decomposable stationary distribution of a multi-dimensional SRBM. *Stochastic Processes and their Applications* (to appear, 2015).

[21] J.G. Dai and Y. Wang: Nonexistence of Brownian models for certain multiclass queueing networks. *Queueing Systems*, **13** (1993), 41–46.

[22] A. Dembo and O. Zeitouni: *Large Deviations Techniques and Applications*, volume 38 of *Applications of Mathematics* (Springer, New York, 1998).

[23] D. Gamarnik and D. Goldberg: Steady-state $GI/GI/N$ queue in the Halfin-Whitt regime. *Annals of Applied Probability*, **23** (2013), 2382–2429.

[24] D. Gamarnik and A. Zeevi: Validity of heavy traffic steady-state approximation in generalized Jackson networks. *Annals of Applied Probability*, **16** (2006), 56–90.

[25] P.W. Glynn: Diffusion approximations. In D. P. Heyman and M. J. Sobel (eds.): *Stochastic Models*, volume 2 of *Handbooks in Operations Research and Management Science* (North Holland, Amsterdam, 1990), 145–198.

[26] P.W. Glynn and W. Whitt: Large deviations behavior of counting processes and their inverses. *Queueing Systems*, **17** (1994), 107–128.

[27] S. Halfin and W. Whitt: Heavy-traffic limits for queues with many exponential servers. *Operations Research*, **29** (1981), 567–588.

[28] J.M. Harrison: The diffusion approximation for tandem queues in heavy traffic. *Advances in Applied Probability*, **10** (1987), 886–905.

[29] J.M. Harrison: Balanced fluid models of multiclass queueing networks: a heavy traffic conjecture. In F. Kelly and R. J. Williams (eds.): *Stochastic Networks*, volume 71 of *The IMA volumes in mathematics and its applications* (Springer, New York, 1995), 1–20.

[30] J.M. Harrison: *Brownian Models of Performance and Control* (Cambridge University Press, New York, 2013).

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

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

[33] J.M. Harrison and V. Nguyen: Some badly behaved closed queueing networks. In F. P. Kelly and R. J. Williams (eds.) *Stochastic Networks*, volume 71 of *The IMA volumes in mathematics and its applications* (Springer, New York, 1995), 117–124.

[34] J.M. Harrison and M.I. Reiman: Reflected Brownian motion on an orthant. *Annals of Probability*, **9** (1981), 302–308.

[35] J.M. Harrison and R.J. Williams: Brownian models of open queueing networks with homogeneous customer populations. *Stochastics*, **22** (1987), 77–115.

[36] D.L. Iglehart and W. Whitt: Multiple channel queues in heavy traffic I. *Advances in Applied Probability*, **2** (1970), 150–177.

[37] D.L. Iglehart and W. Whitt: Multiple channel queues in heavy traffic II: sequences, networks, and batches. *Advances in Applied Probability*, **2** (1970), 355–369.

[38] O. Kallenberg: *Foundations of Modern Probability* (Springer, New York, 1997).

[39] W. Kang and K. Ramanan: Characterization of stationary distributions of reflected diffusions. *Annals of Applied Probability*, **24** (2014), 1329–1374.

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

[41] J.F.C. Kingman: The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society*, **57** (1961), 902–904.

[42] J.F.C. Kingman: On queues in heavy traffic. *Journal of the Royal Statistical Society. Series B. Methodological*, **24** (1962), 383–392.

[43] J.F.C. Kingman: The heavy traffic approximation in the theory of queues. In W. L. Smith and *et al.* (eds.): *Proc. Symp. on Congestion Theory* (University of North Carolina Press, 1965), 137–159.

[44] I. Kino: Two-layer queueing networks. *Journal of the Operations Research Society of Japan*, **40** (1997), 163–185.

[45] H. Kobayashi: Application of the diffusion approximation to queueing networks i: Equilibrium queue distributions. *Journal of the ACM*, **21** (1974), 316–328.

[46] J. Köllerström: Heavy traffic theory for queues with several servers. I. *Journal of Applied Probability*, **11** (1974), 544–552.

[47] P.K. Kumar: Scheduling queueing networks: stability, performance analysis and design. In F. Kelly and R. J. Williams (eds.): *Stochastic Networks*, volume 71 of *The IMA volumes in mathematics and its applications* (Springer-Verlag, New York, 1995), 21–70.

[48] H.H. Kushner: *Heavy Traffic Analysis of Controlled Queueing and Communication Networks* (Springer, New York, 2001).

[49] K. Majewski: Heavy traffic approximations of large deviations of feedforward queueing networks. *Queueing Systems*, **28** (1998), 125–155.

[50] A.I. Markushevich: *Theory of functions of a complex variable. Vol. I, II, III*, English edition, Translated and edited by Richard A. Silverman (Chelsea Publishing Co., New York, 1977).

[51] M. Miyazawa: Rate conservation laws: a survey. *Queueing Systems*, **15** (1994), 1–58.

[52] M. Miyazawa: Superharmonic vector for a nonnegative matrix with QBD block structure and its application to a Markov modulated two dimensional reflecting process. Submitted (2014).

[53] P. Ney and E. Nummelin: Markov additive processes I. eigenvalue properties and limit theorems. *Annals of Probability*, **15** (1987), 561–592.

[54] P.E. Protter: *Stochastic integration and differential equations*, 2nd edition, Version 2.1, Corrected third printing (Springer-Verlag, Berlin, 2005).

[55] J. Reed: The $G/GI/N$ queue in the Halfin-Whitt regime. *Annals of Applied Probability*, **19** (2009), 2211–2269.

[56] M.I. Reiman: Open queueing networks in heavy traffic. *Mathematics of Operations Research*, **9** (1984), 441–458.

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

[58] R. Serfozo: *Introduction to stochastic networks* (Springer, New York, 1999).

[59] K. Sigman: The stability of open queueing networks. *Stochastic Processes and their Applications*, **35** (1990), 11–25.

[60] D.W. Stroock and S.R.S. Varadhan: *Multidimensional Diffusion Processes* (Springer, New York, 1979).

[61] L.M. Taylor and R.J. Williams: Existence and uniqueness of Semimartingale reflecting Brownian motions in an orthant. *Probability Theory and Related Fields*, **96** (1993), 283–317.

[62] W. Whitt: The queueing network analyzer. *Bell System Technical Journal*, **62** (1983), 2779–2815.

[63] W. Whitt: Performance of the queueing network analyzer. *Bell System Technical Journal*, **62** (1983), 2817–2843.

[64] W. Whitt: *Stochastic-process limits* (Springer, New York, 2002).

[65] R.J. Willams: Some recent developments for queueing networks. In L. Accardi and C. C. Heyde (eds.): *Probability Towards 2000* (Springer, New York, 1998), 340–456.

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

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

Masakiyo Miyazawa
Department of Information Sciences
Tokyo University of Science
Yamazaki 2641, Noda
Chiba 278-8510, Japan
E-mail: `miyazawa@rs.tus.ac.jp`