Queueing Systems 40, 143–182, 2002

Diffusion Approximation in Overloaded Switching Queueing Models

VLADIMIR V. ANISIMOV vlanis@bilkent.edu.tr
Bilkent University, Bilkent 06533, Ankara, Turkey, and
Kiev University, Kiev-17, 001017, Ukraine

Received 15 June 2000; Revised 24 August 2001

Abstract. The asymptotic behavior of a queueing process in overloaded state-dependent queueing models (systems and networks) of a switching structure is investigated. A new approach to study fluid and diffusion approximation type theorems (without reflection) in transient and quasi-stationary regimes is suggested. The approach is based on functional limit theorems of averaging principle and diffusion approximation types for so-called Switching processes. Some classes of state-dependent Markov and non-Markov overloaded queueing systems and networks with different types of calls, batch arrival and service, unreliable servers, networks $(M_{SM,Q}/M_{SM,Q}/1/\infty)^r$ switched by a semi-Markov environment and state-dependent polling systems are considered.

Keywords: queueing systems, networks, Markov process, semi-Markov process, switching process, averaging principle, fluid limit, diffusion approximation

AMS subject classification: 60K25, 60J27, 60F17, 60K37, 60J60

1. Introduction

The complexity of real models of computing and information systems leads to the necessity of the creation of more complicated queueing models and the development of new approaches to the approximate (asymptotic) investigation.

A large number of papers is devoted to the analysis of queueing models in heavy traffic conditions. This usually means that the characteristics of the system depend on some parameter, say n, and as $n \to \infty$, the average load in the system tends to one with the rate $O(1/\sqrt{n})$. The study of heavy traffic limits has a long history and there are several directions oriented on different classes of queueing models. Many authors deal with the renewal input process, the independent service times and the routing processes not depending on the current size of a queue or a workload process. For this case, the convergence of a normalized queue length or a workload process to a solution of a differential equation (fluid limits) or to a reflecting Brownian motion in a corresponding domain (Brownian approximation) is proved for a single-class network [43] and for various classes of multiclass networks (see survey [48] and papers [17,20–22,28,29,44,49]). Several classes of service disciplines for multiclass networks are studied in the latest paper [18]. The methods of analysis in these papers essentially use the functional central

limit theorems for arrival, service and routing processes and the continuous mapping theorems for the corresponding reflection map (or the continuity of the solution of Skorokhod reflection problem [46] and its generalizations).

Another direction is related to the analysis of Markov state-dependent queueing models. The method of analysis here is mostly based on a martingale technique [36] and again uses the continuous mapping theorems. Basing on this technique, the convergence of a queueing process in heavy traffic conditions for a state-dependent $(M/M_Q/1/c_n)^r$ network to the diffusion process with the reflection in the rectangle is proved in [12], for $(M_Q/M_Q/1/\infty)^r$ type networks the fluid limits and the convergence to the diffusion process with the reflection in the orthant are studied in the papers [38,39], and the book [15]. Markov time-dependent models are considered in [37,38]. Some results for the state-dependent arrival process and the general service time distribution are given in [34].

The fluid limits and the diffusion approximation (without reflection) for state-dependent Markov queueing systems (networks) of the type $(M_Q/M_Q/k/\infty)^r$ are studied in the book [13] basing on the averaging principle and the diffusion approximation for so-called recurrent processes of a semi-Markov type [5,10]. Some types of Markov state-dependent models $(M_Q/M_Q/1/\infty)^r$ and non-Markov models $G_Q/M_Q/1/\infty$, $(G_Q/M_Q/1/\infty)^r$ are considered in [5,6,10] as examples of using this approach.

The aim of this paper is to extend fluid and diffusion approximation type results to more general classes of queueing models of a switching structure. That is, the corresponding queueing process can be represented in terms of so-called Switching processes (SP's). SP has the property that the character of its operation varies spontaneously (switches) at some epochs of time which can be random functionals of the previous trajectory or possibly jumps of some random environment. The environment may reflect some outer perturbations, a type of operating regime, a number of working servers, a domain of operation for queueing process, a type of priority, etc. Note that on the intervals between switches the process may have a non-Markov structure (for a general description of SP see section 2 and papers [2,5]).

The class of switching queueing models, in particular, includes open and closed Jackson's type Markov and semi-Markov systems and networks with the dependence of the arrival, service and routing processes on the current state of the queueing process and possibly some additional Markov or semi-Markov environment (for instance, a batch semi-Markov arrival process, a service rate depending on the current size of the queue and the environment, etc.). This class also includes some models with multiple calls, calls of a random size and different priorities, models with negative and impatient calls, semi-Markov models with unreliable servers, nonhomogeneous in time Markov and semi-Markov models, networks $(G_Q/M_Q/s/m)^r$ with the state-dependent non-exponential arrival process [5], some classes of state-dependent retrial models [8,9,11] and polling systems. In terms of SP's we can also describe an output process jointly with the queueing process and some other types of additive functionals on the trajectory of the queueing process such as flows of lost calls, etc.

Taking into account that the queueing processes in these models are more complicated, the reflected process in general cannot be represented as a functional of the independent primary processes (arrival, service, routing), and a martingale technique cannot be applied directly, we restrict our analysis to study overloaded models without reflection on the boundary. This means that we study the convergence on the interval $[0, T]$ such that in each component $s(t) > 0$, $t \in [0, T]$, where $s(t)$ is a fluid limit.

A new quite general approach to study limit theorems for models of these types is suggested. It is based on Averaging Principle (AP) and Diffusion Approximation (DA) type results for SP's [2–5], and also uses the representation of a queueing process in terms of SP's.

This approach gives us the possibility to extend fluid and diffusion approximation type results (without reflection) to new more general classes of queueing models, in particular, to state-dependent Markov models (networks) $(M_{Q,B}/M_{Q,B}/k/\infty)^r$ with batch arrival process and service, state-dependent Markov models (networks) $(M_{M,Q}/M_{M,Q}/k/\infty)^r$ in a Markov environment, state-dependent semi-Markov type models $(M_{SM,Q}/M_{SM,Q}/k/\infty)^r$, retrial queues and some types of non-semi-Markov models. From the other side, it also gives us a new technique to study known classes of Markov state-dependent and time-dependent models $(M_Q/M_Q/1/\infty)^r$.

In the paper, we concentrate our attention to study mostly state-dependent Markov and semi-Markov models (networks) and their modifications at the presence (or not) of the ergodic Markov or semi-Markov environment as well. We suppose that characteristics of the system depend on some parameter $n \to \infty$, and the arrival and service processes as well as the routing matrix may depend on the current value of the queueing process $Q_n(t)$ (a vector of queues or a workload process) and possibly some random environment $x_n(t)$. In specific applications the environment may appear due to some external or inner factors. In general, the environment may depend on the queueing process and be not a Markov or a semi-Markov process (case of feedback). We suppose also that a number of calls (or a value of a workload process) in the system is asymptotically large, which may be caused by a high load or by a large initial value of the queueing process.

For queueing models of these types we prove that under quite general assumptions the multidimensional queueing process $n^{-1}Q_n(nt)$ on some interval $[0, T]$ uniformly converges in probability to some function $s(t)$ which is a positive solution of an ordinary differential equation (fluid limit), and the process $n^{-1/2}(Q_n(nt) - ns(t))$ weakly converges (in the sense of a weak convergence of probability measures induced by the process on the space D_T^r and endowed by Skorokhod topology) to a diffusion process with coefficients depending in general on $s(t)$ (diffusion approximation). Here D_T^r is the Skorokhod space of r-dimensional right-continuous functions given on $[0, T]$ with finite left limits. Readers are refereed to [16,23,45] for the definition of Skorokhod space and Skorokhod topology.

The results obtained are mostly oriented to the analysis of a transient behavior of the queueing processes. They also give the possibility to study the transient behavior of the queueing process even for ergodic systems in the case, when the initial value of the

process is large, and, in addition, to get the asymptotic behaviour of the time of hitting to zero, because the weak convergence of measures implies the weak convergence of continuous functionals of the process such as hitting times. From the other side, for some types of overloaded models the queueing process asymptotically cannot reach zero (for instance, for $M/M/\infty$ model (network) when the service rate goes to 0). For models of this type we get the approximation on the entire time horizon. It is possible to study so-called quasi-stationary regimes also. These regimes appear, when the corresponding fluid limit $s(t)$ has a point of stability $s_* > 0$. In this case $n^{-1}Q_n(nt)$ is asymptotically close to s_*, as $n \to \infty$ and then $t \to \infty$. In particular, if $n^{-1}Q_n(0) \to s_*$ in probability, then the coefficients of the limiting diffusion process do not depend on time, and the queueing process is balancing near some asymptotically high level ns_* as a homogeneous diffusion process multiplied by $\sqrt{n}$.

The rest of the paper is organized as follows. A description of some important subclasses of SP's and some classes of switching queueing models is given in section 2. Section 3 deals with the asymptotic analysis (fluid limits and diffusion approximation) of some classes of overloaded state-dependent Markov queueing systems and networks in transient conditions. Some classes of non-Markov models (systems and networks in a semi-Markov environment), state-dependent systems with unreliable servers and polling systems are considered in section 4. Some theoretical results related to AP and DA for some special subclasses of SP's are given in appendix.

2. Switching models

We consider here some rather general models of switching queueing systems and networks. These models can be described in terms of switching processes (SP's). First, to illustrate our approach and to give some basic ideas on the analysis of more general switching systems, we consider rather simple Markov state-dependent system.

2.1. A system $M_Q/M_Q/1/\infty$

A system consists of one server with infinite buffer. The calls arrive one at a time and wait in the queue according to FIFO discipline. Let nonnegative functions $\{\lambda(q),\ \mu(q),\ q \geqslant 0\}$ be given. Denote the total number of calls in the system at time t by $Q(t)$, $t \geqslant 0$. The system operates as follows. If at time t $Q(t) = q$, then the local arrival rate is $\lambda(q)$, and the local service rate is $\mu(q)$. After service completion a call leaves the system.

It is well known, that in this case the process $Q(t)$ is a birth-and-death process. Let us represent it in a recurrent form. Denote by $t_1 < t_2 < \cdots$ the times of any changing in the system (arrival of a call or service completion), and put $Q_k = Q(t_k + 0)$, $k \geqslant 0$. Suppose that $t_0 = 0$, $Q(0+) = Q_0$.

First, we construct the family of jointly independent random variables $\{\tau_k(q), \xi_k(q),\ q \geqslant 0\},\ k \geqslant 0$. Here $\tau_k(q)$ has an exponential distribution with parameter $\Lambda(q) = \lambda(q) + \mu(q)\chi(q > 0)$, $\xi_k(q)$ is an independent of $\tau_k(q)$ variable and

$$\xi_k(q) = \begin{cases} +1, & \text{with probab. } \lambda(q)\Lambda(q)^{-1}, \\ -1, & \text{with probab. } \mu(q)\chi(q > 0)\Lambda(q)^{-1}, \end{cases}$$

where $\chi(A)$ is the indicator of the set A.

Let us introduce the following recurrent sequences:

$$\begin{aligned} &\widetilde{Q}_0 = Q_0, \qquad && \widetilde{Q}_{k+1} = \widetilde{Q}_k + \xi_k\big(\widetilde{Q}_k\big), \\ &\widetilde{t}_0 = 0, && \widetilde{t}_{k+1} = \widetilde{t}_k + \tau_k\big(\widetilde{Q}_k\big), \quad k \geqslant 0, \end{aligned} \tag{2.1}$$

and put

$$\widetilde{Q}(t) = \widetilde{Q}_k, \quad \text{as } \widetilde{t}_k \leqslant t < \widetilde{t}_{k+1},\ t \geqslant 0. \tag{2.2}$$

It is easy to check that the process $\widetilde{Q}(t)$ is equivalent (by finite dimensional distributions) to the queueing process $Q(t)$.

The advantage of this representation is that $\widetilde{Q}(t)$ is written as a superposition of two more simple recurrent processes in discrete time, $\widetilde{t}_k$ and $\widetilde{Q}_k, k \geqslant 0$. Processes, represented in this form, are called recurrent processes of a semi-Markov type [3,5,10]. This representation gives also an idea, how to study the limiting behavior of $Q(t)$. If we can prove, that appropriately scaled processes $\widetilde{t}_k$ and $\widetilde{Q}_k$ weakly converge to some (maybe dependent) processes $y(u)$ and $q(u)$, $u \geqslant 0$, then under some regular assumptions we can expect that the appropriately scaled process $\widetilde{Q}(t)$ weakly converges to the superposition of $y(u)$ and $q(u)$ in the form $q(y^{-1}(t))$, where $y^{-1}(t)$ is the inverse function.

The representation (2.1), (2.2) has a similar form for Markov networks and also for batch arrivals and service. In this case the variables $\xi_k(q)$ may take vector values, and variables $\tau_k(q)$ again have exponential distributions. By analogy, we can write similar representations for more general systems with non-Markov arrival process and non-exponential service. For these cases we need to choose in the appropriate way times $\widetilde{t}_k$ and construct corresponding processes, reflecting the behavior of queueing processes, on the intervals $[\widetilde{t}_k, \widetilde{t}_{k+1})$.

For further exploration we note that, actually, the exponentiality of $\tau_k(q)$ is not essential for the asymptotic analysis. That means, if we can prove quite general theorems on the convergence of the recurrent processes, constructed according to relations (2.1), (2.2), then these theorems can be used for the analysis of more general queueing models, for which the queueing processes have representations similar to (2.1), (2.2).

In this way we came to the idea to analyze switching queueing models. For these models, the queueing processes can be represented in terms of SP in the form similar to (2.1), (2.2). From the other side, for SP rather general results on averaging principle and diffusion approximation are proved in [3–5,10]. Thus, we can use the class of SP as a very convenient tool to describe wide classes of state-dependent queueing models and to study their asymptotic behavior.

Let us give now a general definition of an SP.

2.2. Switching processes

Let $\mathcal{F}_k = \{(\zeta_k(t,x,\alpha), \tau_k(x,\alpha), \beta_k(x,\alpha)), t \geqslant 0, x \in X, \alpha \in \mathcal{R}^r\}, k \geqslant 0$, be jointly independent parametric families. Here $(X, \mathcal{B}_X)$ is some measurable space, $\zeta_k(t,x,\alpha)$ at each fixed k, x, α, is a stochastic process with sample trajectories belonging to Skorokhod space $\mathcal{D}^r_\infty$ (the space of right-continuous functions given on $[0,\infty)$ with values in $\mathcal{R}^r$ and finite left limits) [45], and $\tau_k(x,\alpha)$, $\beta_k(x,\alpha)$ are possibly dependent on $\zeta_k(\cdot,x,a)$ random variables, $\tau_k(\cdot) \geqslant 0$, $\beta_k(\cdot) \in X$. Furthermore, we suppose that the vectors from R^r are column vectors and the variables introduced are measurable in the ordinary way in the pair (x,α) concerning σ-algebra $\mathcal{B}_X \times \mathcal{B}_{R^r}$. Let also (x_0, S_0) be an independent of $\mathcal{F}_k$, $k \geqslant 0$, initial value in $X \times \mathcal{R}^r$. We introduce the following recurrent sequences:

$$\begin{aligned} t_0 = 0, \qquad & t_{k+1} = t_k + \tau_k(x_k, S_k), \qquad S_{k+1} = S_k + \xi_k(x_k, S_k), \\ & x_{k+1} = \beta_k(x_k, S_k), \quad k \geqslant 0, \end{aligned} \tag{2.3}$$

where $\xi_k(x,\alpha) = \zeta_k(\tau_k(x,\alpha), x, \alpha)$, and set

$$\zeta(t) = S_k + \zeta_k(t - t_k, x_k, S_k), \qquad x(t) = x_k, \quad \text{as } t_k \leqslant t < t_{k+1},\ t \geqslant 0. \tag{2.4}$$

Then the two-component process $\{(x(t), \zeta(t));\ t \geqslant 0\}$ is called a switching process (SP). Times t_k are usually called switching times, $x(t)$ is a switching random environment. We introduce also the imbedded process

$$S(t) = S_k, \quad \text{as } t_k \leqslant t < t_{k+1}, t \geqslant 0, \tag{2.5}$$

and call it a recurrent process of a semi-Markov type (RPSM) (see [5,10]). Furthermore, we assume that SP is regular, i.e., the component $x(t)$ has with probability one a finite number of jumps on each finite interval.

The class of SP's was introduced in [1,2]. Note that this class is a natural generalization of such well-known classes of stochastic processes as Markov processes homogeneous in the second component [24], piecewise Markov aggregates [19], and Markov and semi-Markov evolutions [30,32,33,35,41,42].

Relations (2.3)–(2.5) show that we may have the dependence (feedback) between components $x(t)$ and $S(t)$. That is, the sequence x_k itself is not in general a Markov process (MP), and the process $x(t)$ also in general is not an MP or a semi-Markov process (SMP), respectively. We do not have feedback, if we consider a semi-Markov random evolution or a queueing model in some external Markov or semi-Markov environment.

Consider some particular cases. Suppose that characteristics of $\mathcal{F}_k$ do not depend on $k \geqslant 0$ (homogeneous case). Then $\{(x_k, S_k);\ k \geqslant 0\}$ is a homogeneous MP, and $\{(x(t), S(t));\ t \geqslant 0\}$ is an SMP with the sojourn time in the state (x,α), $\tau_1(x,\alpha)$, and transition probability

$$\begin{aligned} &\mathbf{P}\{x_{k+1} \in A,\ S_{k+1} \in B,\ t_{k+1} - t_k < t \mid x_k = x,\ S_k = \alpha\} \\ &\quad = \mathbf{P}\big\{\beta_1(x,\alpha) \in A,\ \xi_1(x,\alpha) \in B - \{\alpha\},\ \tau_1(x,\alpha) < t\big\}, \end{aligned}$$

where $B - \{\alpha\} = \{b\colon\ \alpha + b \in B\}$.

If, in addition, the distributions of variables $\tau_k(x,\alpha)$, $\beta_k(x,\alpha)$ do not depend on the parameter α, then $\{x(t);\ t \geqslant 0\}$ is itself an SMP. Assume also that at each $x \in X$ variables $\tau_k(x)$ are independent of $\{\zeta_k(t,x,\alpha);\ t \geqslant 0\}$. If in this case $\tau_k(x)$ has an exponential distribution, then $\{x(t);\ t \geqslant 0\}$ is an MP. If $\tau_k(x)$ has an arbitrary distribution, then the component $\zeta(t)$ can be described as a stochastic process with semi-Markov switches (or in a semi-Markov environment). In particular, if at each $(x,\alpha) \in X \times \mathcal{R}^r$, $\{\zeta_k(t,x,\alpha);\ t \geqslant 0\}$ is a process with independent increments, then $\{(x(t),\zeta(t));\ t \geqslant 0\}$ is a process with independent increments and semi-Markov switches (see [2]). If at each (x,α) $\{\zeta_k(t,x,\alpha);\ t \geqslant 0\}$ is an MP with the initial value α and $\{x(t);\ t \geqslant 0\}$ is an MP or an SMP, then $\{(x(t),\zeta(t));\ t \geqslant 0\}$ is a Markov or a semi-Markov random evolution.

Using the construction of SP we can describe the nonhomogeneous in time models also. For this purpose, in the definition of families $\mathcal{F}_k$ we add an additional parameter u. Then relations (2.3)–(2.5) have the form: $t_{k+1} = t_k + \tau_k(x_k, S_k, t_k)$, $S_{k+1} = S_k + \xi_k(x_k, S_k, t_k)$, $x_{k+1} = \beta_k(x_k, S_k, t_k)$.

Now we say, a switching queuing model is a model with the property that a queueing process $Q(t)$ can be described in terms of SP's. This means that we can construct on some probability space an auxiliary SP $(\widetilde{x}(t), \widetilde{Q}(t))$ such that the component $\widetilde{Q}(t)$ is equivalent (by finite-dimensional distributions) to $Q(t)$.

Consider some special subclasses of SP's which are useful at the analysis of queueing models.

2.3. Recurrent processes of a semi-Markov type

Let $\mathcal{F}_k = \{(\xi_k(\alpha), \tau_k(\alpha)),\ \alpha \in \mathcal{R}^r\}$, $k \geqslant 0$, be jointly independent families of random variables with values in $\mathcal{R}^r \times [0,\infty)$. Let also S_0 be an independent of $\mathcal{F}_k$, $k \geqslant 0$, initial value, $S_0 \in \mathcal{R}^r$. Denote

$$\begin{aligned} &t_0 = 0, \qquad t_{k+1} = t_k + \tau_k(S_k), \qquad S_{k+1} = S_k + \xi_k(S_k), \quad k \geqslant 0, \\ &\qquad\qquad S(t) = S_k, \quad \text{as } t_k \leqslant t < t_{k+1},\ t \geqslant 0. \end{aligned} \tag{2.6}$$

Then the process $\{S(t);\ t \geqslant 0\}$ is called a Recurrent Process of a Semi-Markov type (RPSM) [5,10]. In homogeneous case (distributions of introduced variables do not depend on k) the process $S(t)$ is a homogeneous SMP.

Suppose now that jointly independent families of random variables $\mathcal{F}_k = \{(\xi_k(x,\alpha), \tau_k(x,\alpha)), x \in X, \alpha \in \mathcal{R}^r\}$, $k \geqslant 0$, with values in $\mathcal{R}^r \times [0,\infty)$ be given. Let $\{x_l;\ l \geqslant 0\}$ be an independent of $\mathcal{F}_k$, $k \geqslant 0$, MP with values in X, (x_0, S_0) be an initial value. We put $t_0 = 0$, $t_{k+1} = t_k + \tau_k(x_k, S_k)$, $S_{k+1} = S_k + \xi_k(x_k, S_k)$, $k \geqslant 0$, and denote

$$S(t) = S_k, \qquad x(t) = x_k, \quad \text{as } t_k \leqslant t < t_{k+1},\ t \geqslant 0. \tag{2.7}$$

Then the process $\{(x(t), S(t));\ t \geqslant 0\}$ is an RPSM with Markov switches.

Consider a general case. Let $\mathcal{F}_k = \{(\xi_k(x,\alpha), \tau_k(x,\alpha), \beta_k(x,\alpha)),\ x \in X,\ \alpha \in \mathcal{R}^r\}$, $k \geqslant 0$, be jointly independent families of random variables with values in $\mathcal{R}^r \times$

$[0, \infty) \times X$, X be some measurable space, (x_0, S_0) be an initial value. We put $t_0 = 0$, $t_{k+1} = t_k + \tau_k(x_k, S_k)$, $S_{k+1} = S_k + \xi_k(x_k, S_k)$, $x_{k+1} = \beta_k(x_k, S_k)$, $k \geqslant 0$, and denote

$$S(t) = S_k, \qquad x(t) = x_k, \quad \text{as } t_k \leqslant t < t_{k+1},\ t \geqslant 0. \tag{2.8}$$

Then the pair $\{(x(t), S(t)); t \geqslant 0\}$ forms a general RPSM (case of feedback between components $x(t)$ and $S(t)$). In particular, when distributions of the variables $\beta_k(x, \alpha)$ do not depend on the parameter α, the sequence x_k is an MP and $(x(t), S(t))$ is an RPSM with Markov switches.

2.4. Processes with semi-Markov switches

Consider an operation of some stochastic process in a semi-Markov environment. Let $\mathcal{F}_k = \{\zeta_k(t, x, \alpha),\ t \geqslant 0,\ x \in X,\ \alpha \in \mathcal{R}^r\}$, $k \geqslant 0$, be jointly independent families of stochastic processes, where $\zeta_k(t, x, \alpha)$ at each fixed k, x, α is a process with trajectories in Skorokhod space $\mathcal{D}^r_\infty$. Let also $x(t)$, $t \geqslant 0$, be an independent of $\mathcal{F}_k$, $k \geqslant 0$, right-continuous SMP in X, and S_0 be an initial value. Denote by $0 = t_0 \leqslant t_1 \leqslant \cdots$ the epochs of sequential jumps of $x(\cdot)$ and set $x_k = x(t_k)$, $k \geqslant 0$. We construct a process with semi-Markov switches (or in a semi-Markov environment) as follows: put $S_{k+1} = S_k + \xi_k$, $\tau_k = t_{k+1} - t_k$, $k \geqslant 0$, where $\xi_k = \zeta_k(\tau_k,\ x_k,\ S_k)$, and denote

$$\zeta(t) = S_k + \zeta_k(t - t_k,\ x_k,\ S_k), \quad \text{as } t_k \leqslant t < t_{k+1},\ t \geqslant 0. \tag{2.9}$$

Then a two-component process $\{(x(t), \zeta(t));\ t \geqslant 0\}$ is called a process with semi-Markov switches (PSMS). We introduce also an imbedded process

$$S(t) = S_k, \quad \text{as } t_k \leqslant t < t_{k+1},\ t \geqslant 0. \tag{2.10}$$

By construction, $\{(x(t), S(t));\ t \geqslant 0\}$ is an RPSM with Markov switches. In particular, if at each (x, α) $\zeta_k(t, x, \alpha)$ is an MP with the initial value α, and $x(t)$ is either an MP or an SMP, then $\{(x(t), \zeta(t));\ t \geqslant 0\}$ is a random evolution.

2.5. Switching queueing models

In this section we consider several examples of state-dependent queueing models and a technique of the representation of queueing processes in terms of SP's.

2.5.1. State-dependent Markov network

Consider a state-dependent queueing network $(M_Q/M_Q/1/\infty)^r$ consisting of r nodes. Suppose for simplicity that there is one server at each node with infinite buffer. Denote by R^r_+ the space of vectors with non-negative components. To distinguish the cases of systems and networks, we denote by $\overline{q}$ column-vectors $(q_1, \ldots, q_r) \in \mathcal{R}^r$. Let non-negative functions $\{\lambda(\overline{q}), \mu_i(\overline{q}), i = \overline{1, r}, \overline{q} \in R^r_+\}$ be given. Let also the independent families of random vectors $\{\overline{\eta}(\overline{q}),\ \overline{q} \in R^r_+\}$ and $\{(\kappa_i(\overline{q}),\ \overline{\gamma}_i(\overline{q})),\ i = \overline{1, r}, \overline{q} \in R^r_+\}$ with values in R^r_+ and $(\mathcal{R}_+ \times \mathcal{R}^{r+1}_+)$, respectively, be given. An arrival flow is consisting of calls of a random size (a portion of work, an information package, etc.). Let $Q_i(t)$ be

the total amount of work (or information) in the buffer at node i at time t (a queue size in the case of ordinary arrivals). Put $\overline{Q}(t) = (Q_1(t), \dots, Q_r(t))$.

The network operates as follows. Suppose that at time t $\overline{Q}(t) = \overline{q}$. Then we have the following possibilities. A call of a random size $\overline{\eta}(\overline{q})$ may enter the network with the local arrival rate $\lambda(\overline{q})$ (it means that the ith component $\eta_i(\overline{q})$ of the vector $\overline{\eta}(\overline{q})$ enters node i). Correspondingly, with the local service rate $\mu_i(\overline{q})$ a random portion of work of the size $\min\{\kappa_i(\overline{q}), q_i\}$ may finish service at node i. Immediately after this, this portion is transformed to the vector $\overline{\gamma}_i(\overline{q})$ which is added to current amounts of work at the nodes. This means that jth component $\gamma_i^{(j)}(\overline{q})$ of the vector $\overline{\gamma}_i(\overline{q})$ goes to node j, $j = \overline{1, r}$, and the portion $\gamma_i^{(0)}(\overline{q})$ leaves the network. We can always assume that $\mu_i(\overline{q}) = 0$, if the ith component of $\overline{q}$ is equal to zero.

Let us describe the process $\{\overline{Q}(t);\ t \geqslant 0\}$ as an RPSM. In our case, $\overline{Q}(t)$ is a multidimensional MP. We define here switching times t_k, $k \geqslant 0$, as times of any changing in the network. Let us introduce the independent random variable $\tau(\overline{q})$ and vector $\overline{\xi}(\overline{q})$ such that $\tau(\overline{q})$ has an exponential distribution with parameter $\Lambda(\overline{q}) = \lambda(\overline{q}) + \sum_{i=1}^{r} \mu_i(\overline{q})$, and

$$\overline{\xi}(\overline{q}) = \begin{cases} \overline{\eta}(\overline{q}), & \text{with probab. } \lambda(\overline{q})\Lambda(\overline{q})^{-1}, \\ -\min\{\kappa_i(\overline{q}), q_i\}\overline{e}_i + \overline{[\gamma_i(\overline{q})]_r}, & \text{with probab. } \mu_i(\overline{q})\Lambda(\overline{q})^{-1},\ i = \overline{1, r}. \end{cases}$$

Here $\overline{e}_i$ is a vector with the ith component equals to one and the other components equal to 0, and for any vector $\overline{a} = (a_1, \dots, a_r, a_{r+1})$, $\overline{[a]}_r$ denotes the vector $(a_1, \dots, a_r)$.

Now we put $\overline{Q}_k = \overline{Q}(t_k + 0)$. It is easy to see that for any $k \geqslant 0$, $\overline{z} \in \mathcal{R}^r$, $u > 0$

$$\mathbf{P}\big(\overline{Q}_{k+1} - \overline{Q}_k \leqslant \overline{z},\ t_{k+1} - t_k \leqslant u \mid \overline{Q}_k = \overline{q}\big) = \mathbf{P}\big(\overline{\xi}(\overline{q}) \leqslant \overline{z}\big)\mathbf{P}\big(\tau(\overline{q}) \leqslant u\big).$$

This relation shows that the process $\overline{Q}(t)$ is equivalent to an RPSM which is defined by families $\{\overline{\xi}(\overline{q}), \tau(\overline{q})\}$ according to (2.6). In this way we can also represent an output process $Z(t)$. We add an additional node $r + 1$ and consider it as an accumulating node for $Z(t)$. Then the process $(\overline{Q}(t), Z(t))$ by analogy can be described as an RPSM.

If we consider the case of negative calls introduced in [26], then $\eta_i(\overline{q})$ may take negative values. In this case, according to a standard truncation procedure, we can assume that the total amount of work q_i at node i after arrival is transformed into $\min\{0, q_i + \eta_i(\overline{q})\}$ and leave the rest of notation.

Note that in these terms it is also possible to describe state-dependent Markov models with different classes of calls, impatient calls, priority models, etc.

2.5.2. Markov system in a Markov environment

Consider a system $M_M/M_M/1/\infty$ in a Markov environment. There is one server and an infinite number of waiting places. Let $\{x(t);\ t \geqslant 0\}$ be an MP with finite state space $X = \{1, \dots, d\}$ and transition rates a_{xy}, $x, y \in X$, $x \neq y$, and let nonnegative functions $\{\lambda(x, q), \mu(x, q),\ x \in X,\ q \geqslant 0\}$ be given. The calls arrive one at a time and wait in the queue according to FIFO discipline. Denote by $Q(t)$ the total number of calls in

the system at time t. If $x(t) = x$, $Q(t) = q$, then the local arrival rate is $\lambda(x, q)$ and the local service rate is $\mu(x, q)$ (for simplicity assume that $\mu(x, 0) \equiv 0$). A call being served leaves the system.

For this model we may choose switching times in a different way. For instance, let us define switching times t_k as times of any changing (either $Q(t)$ or $x(t)$) in the system. Denote $x_k = x(t_k)$, $Q_k = Q(t_k)$, $k \geqslant 0$. It is easy to see that we can represent $(x(t), Q(t))$ as a general RPSM, for which $\tau_1(x, q)$ has an exponential distribution with parameter $\Lambda(x, q) = a_{xx} + \lambda(x, q) + \mu(x, q)$, and

$$\xi_1(x, q) = \begin{cases} +1, & \text{with probab. } \lambda(x, q)\Lambda(x, q)^{-1}, \\ -1, & \text{with probab. } \mu(x, q)\Lambda(x, q)^{-1}, \\ 0, & \text{with probab. } a_{xx}\Lambda(x, q)^{-1}, \quad x = \overline{1, r}, \end{cases}$$

where $a_{xx} = \sum_{y \neq x} a_{xy}$. In this case

$$\begin{aligned} &\mathbf{P}(x_{k+1} = y \mid x_k = x, Q_k = q) = a_{xy}\Lambda(x, q)^{-1}, \quad y \neq x, \\ &\mathbf{P}(x_{k+1} = x \mid x_k = x, Q_k = q) = \big(\lambda(x, q) + \mu(x, q)\big)\Lambda(x, q)^{-1}. \end{aligned}$$

Here x_k is not, in general, an MP and we have feedback between components x_k and Q_k.

2.5.3. State-dependent semi-Markov type network

Consider a network $(M_{SM,Q}/M_{SM,Q}/1/\infty)^r$ switched by a semi-Markov environment, which in some sense is a generalization of a model considered in section 2.5.1. Suppose that there are r nodes and one server at each node with infinite buffer. Let $\{x(t);\ t \geqslant 0\}$ be an SMP with state space $X = \{1, 2, \ldots, d\}$, which stands for the external environment. Let also the families of nonnegative functions $\{\lambda(x, \overline{q}), \mu_j(x, \overline{q}),\ j = \overline{1, r}, x \in X\}$, routing matrices $P(x, \overline{q}) = \|p_{ij}(x, \overline{q})\|_{i=\overline{1,r},\ j=\overline{1,r+1}}$, $x \in X$, and the independent families of random vectors $\{\overline{\eta}(x, \overline{q}),\ x \in X\}$ with values in $\mathcal{R}^r_+$ and nonnegative random variables $\{\kappa_j(x, \overline{q}),\ x \in X,\ j = \overline{1, r}\}$ be given (here $\overline{q} \in \mathcal{R}^r_+$).

As in section 2.5.1, denote the total amount of work in the buffer at node i at time t by $Q_i(t)$ and put $\overline{Q}(t) = (Q_1(t), \ldots, Q_r(t))$. If at time t $(x(t), \overline{Q}(t)) = (x, \overline{q})$, then with the local arrival rate $\lambda(x, \overline{q})$ a call of a random size $\overline{\eta}(x, \overline{q})$ may enter the system (the ith component of the vector $\overline{\eta}(x, \overline{q})$ enters node i). Correspondingly, with the local service rate $\mu_i(x, \overline{q})$ a random portion of work of a size $\widetilde{\kappa}_i(x, \overline{q}) = \min\{\kappa_i(x, \overline{q}), q_i\}$ may leave node i. Immediately after this, either with probability $p_{ij}(x, \overline{q})$ this portion goes to node j, $j = \overline{1, r}$, or with probability $p_{i,r+1}(x, \overline{q})$ it leaves the network. Here we may assume for simplicity that $\mu_i(x, \overline{q}) \equiv 0$ if $q_i = 0$, where $\overline{q} = (q_1, \ldots, q_r)$.

To describe the process $\{(x(t), \overline{Q}(t));\ t \geqslant 0\}$ in the network as an SP, we introduce the independent families of multidimensional MP's $\{\overline{\gamma}_k(t, x, \overline{q}), t \geqslant 0, x \in X, \overline{q} \in R^r_+\}$, $k \geqslant 0$, with distributions not depending on k and with values in R^r_+ in the following way: $\overline{\gamma}_k(0, x, \overline{q}) = \overline{q}$, and if at time t $\overline{\gamma}_k(t, x, \overline{q}) = \overline{s}$, then the process $\overline{\gamma}_k(t, x, \overline{q})$ can make a jump of the size $\overline{\delta}(x, \overline{s})$ with the local rate $\Lambda(x, \overline{s}) = \lambda(x, \overline{s}) + \sum_{i=1}^r \mu_i(x, \overline{s})$,

where

$$\overline{\delta}(x,\overline{s}) = \begin{cases} \overline{\eta}(x,\overline{s}), & \text{with probab. } \lambda(x,\overline{s})\Lambda(x,\overline{s})^{-1}, \\ (-\overline{e}_i + \overline{e}_j)\widetilde{\kappa}_i(x,\overline{s}), & \text{with probab. } \mu_i(x,\overline{s})p_{ij}(x,\overline{s})\Lambda(x,\overline{s})^{-1}, \\ -\overline{e}_i\widetilde{\kappa}_i(x,\overline{s}), & \text{with probab. } \mu_i(x,\overline{s})p_{i,r+1}(x,\overline{s})\Lambda(x,\overline{s})^{-1}, \\ & i,\ j = \overline{1,r}. \end{cases} \tag{2.11}$$

Now we construct the family of processes $\{\overline{\zeta}_k(t,x,\overline{q});\ t \geqslant 0\}$ in the following way. Let at each $x \in X$, $\tau_k(x)$, $k \geqslant 0$, be a sequence of i.i.d.r.v. having the same distribution as the sojourn time $\tau(x)$ in state x. Then $\overline{\zeta}_k(t,x,\overline{q})$ is defined on the interval $[0, \tau_k(x)]$, and $\overline{\zeta}_k(t,x,\overline{q}) = \overline{\gamma}_k(t,x,\overline{q}) - \overline{q}$, $0 \leqslant t \leqslant \tau_k(x)$. We choose switching times t_k as times of sequential jumps of $x(t)$. Then the process $\{(x(t), \widetilde{\overline{Q}}(t));\ t \geqslant 0\}$, which is constructed using introduced processes $\overline{\zeta}_k(\cdot)$ and an SMP $x(\cdot)$ according to (2.9), is a process with semi-Markov switches (PSMS). It is equivalent to the process $\{(x(t), \overline{Q}(t));\ t \geqslant 0\}$ in our system.

For this case, an arrival process may be called a semi-Markov modulated Poisson process by analogy to Markov modulated arrival process [40].

If we add an additional node $r+1$ and consider it as an accumulating node for the output process $Z(t)$, then in the same way we can describe the process $(x(t), \overline{Q}(t), Z(t))$ as PSMS.

By analogy, we can consider different classes of calls, a priority service, etc.

2.5.4. Models with dependent arrival flows

Consider a system $G_Q/M_Q/1/\infty$. There is one server and an infinite number of waiting places. The function $\mu(\alpha) \geqslant 0$, $\alpha \geqslant 0$, and the independent families of nonnegative random variables $\{\tau_k(\alpha), \alpha \geqslant 0\}$, $k \geqslant 0$, with distributions not depending on index k are given. The system operates as follows: the calls enter the system one at a time and wait according to FIFO discipline. Denote the total number of calls in the system at time t by $Q(t)$. If a call enters the system at time t_k and $Q(t_k+0) = q$, then the next call enters the system at time $t_{k+1} = t_k + \tau_k(q)$, and the service rate on the interval (t_k, t_{k+1}) is $\mu(q)$.

In this case we do not have a switching component x_k. Now we choose $\tau_k(q)$ as switching intervals. Let us construct the process $\zeta_k(t,q)$ as follows: $\zeta_k(t,q) = -\min\{q, \Pi_k(t, \mu(q))\}$ as $t < \tau_k(q)$, and $\zeta_k(\tau_k(q), q) = 1 - \min\{q, \Pi_k(\tau_k(q), \mu(q))\}$, where $\Pi_k(t, \mu_k)$ are jointly independent Poisson processes with parameters μ_k. Then we can represent $Q(t)$ as an SP according to (2.3), (2.4).

In the same way we can describe models with the dependent batch arrival and service and extend this description to networks.

2.5.5. Polling systems

Consider a system with r stations and a single moving server. An arrival flow to station i is a Poisson flow with rate λ_i. Denote by $Q_i(t)$ a number of calls at station i at time t, $\overline{Q}(t) = (Q_1(t), \ldots, Q_r(t))$. Let $\kappa_k(i)$ and $\kappa_k(i,j)$ be the independent at different $k \geqslant 0$,

$i, j = 1, \ldots, r$, random variables with distributions not depending on index k. If the server comes to station i at time t and $\overline{Q}(t) = \overline{Q} = (Q_1, \ldots, Q_r)$, then it occupies the station for the time $\kappa_k(i)$, and the service rate on this period is $\mu_i(Q_i)$. All calls being served at the station during this period leave the system. After completing the time $\kappa_k(i)$, the server with probability p_{ij} goes to station j, and it takes a random time $\kappa_k(i, j)$. During this time no service is provided. When the server arrives to station j, the service immediately begins with rate $\mu_j(Q_j)$, where Q_j is the number of calls at station j at the time of arrival, and so on.

Let us represent this system as a switching system. Denote by t_k, $k \geqslant 0$, the sequential times of arrivals of the server at any station ($t_0 = 0$). We construct the process $x(t)$ in the following way: $x(t) = i$, $t_k \leqslant t < t_{k+1}$, if at time t_k the server arrives to station i. Note that $x(t)$ is an SMP. Put $x_k = x(t_k + 0)$. Let also $\widetilde{\kappa}_k(i)$ be an independent of $\kappa_k(i)$ random variable such that $\mathbf{P}(\widetilde{\kappa}_k(i) \leqslant z) = \sum_j p_{ij}\mathbf{P}(\kappa_k(i, j) \leqslant z)$. Then $\mathbf{P}(t_{k+1} - t_k \leqslant z \mid x_k = i) = \mathbf{P}(\kappa_k(i) + \widetilde{\kappa}_k(i) \leqslant z)$, $k \geqslant 0$.

Let $\{y_k(t, i, Q);\ t \geqslant 0\}$ be the independent at different $k \geqslant 0$, $i = 1, \ldots, r$, birth-and-death processes with constant rates of birth λ_i and death $\mu_i(Q)$, respectively, and the initial value Q (the rates do not depend on the current state except state 0). Let also $\{\Pi_k(t, i, \lambda_k);\ t \geqslant 0\}$ be the independent at different k, i Poisson processes with parameters λ_k. We introduce the process $\overline{\zeta}_k(t, i, \overline{Q}) = (\zeta_k^{(j)}(t, i, Q_j),\ j = \overline{1, r})$ on the interval $[0, \kappa_k(i) + \widetilde{\kappa}_k(i)]$ as follows:

$$\zeta_k^{(i)}(t, i, Q_i) = y_k(t, i, Q_i) - Q_i, \quad \text{as } 0 \leqslant t \leqslant \kappa_k(i);$$

$$\zeta_k^{(i)}(t, i, Q_i) = y_k\big(\kappa_k(i), i, Q_i\big) - Q_i + \Pi_k\big(t - \kappa_k(i), i, \lambda_i\big),$$

$$\text{as } \kappa_k(i) < t \leqslant \kappa_k(i) + \widetilde{\kappa}_k(i);$$

$$\zeta_k^{(j)}(t, i, Q_j) = \Pi_k(t, j, \lambda_j), \qquad \text{as } 0 \leqslant t \leqslant \kappa_k(i) + \widetilde{\kappa}_k(i), \quad j = \overline{1, r},\ j \neq i.$$

Then, using families $\{x_k, \overline{\zeta}_k(t, i, \overline{Q})\}$ and relations (2.9), we can construct a PSMS which is equivalent to the queueing process $\{(x(t), \overline{Q}(t));\ t \geqslant 0\}$.

We can also consider a workload process $W_i(t)$ at station i (the total time that a call arriving at time t will spend in the system). If $Q_i(t) = Q_i$, then for any fixed t, $W_i(t)$ can be represented as the hitting time to level Q_i of a Poisson type process switched by an SMP $x(t)$.

It is also possible to consider other types of service policy. For instance, under the gated policy we suppose that if the server upon arrival to station i sees Q_i calls in the queue, it spends at the station the time which is necessary to complete the service of all those Q_i calls. Other calls, arriving during this time, go to the queue and wait until the next arrival of the server.

In this case, the total time $\kappa(i) = \kappa(i, Q_i)$ spent at station i depends also on Q_i and is represented in the form: $\kappa(i, Q_i) = \sum_{1 \leqslant l \leqslant Q_i} \eta_l(Q_i)$, where $\eta_l(Q_i)$, $l \geqslant 1$, are jointly independent and exponentially distributed with parameter $\mu(Q_i)$ variables (we assume that $\sum_1^0 = 0$). The family of processes $\zeta_k(t, i, Q)$ is constructed in a similar way. Note that here $x(t)$ is not an SMP.

Remark 2.1. In terms of SP it is also possible to describe some classes of Markov and semi-Markov queueing systems and networks with unreliable servers and some classes of retrial queues [8,9,11].

3. Diffusion approximation in Markov queueing models

In overloaded switching queueing models various multidimensional characteristics (numbers of calls at different nodes, volumes of information in buffers, output flows, flows of lost calls, waiting times, etc.) can be approximated by the solutions of differential equations or by the diffusion processes. The method of the analysis is based on the asymptotic results of Averaging Principle (AP) and Diffusion Approximation (DA) types for SP's (see appendix) and uses the representation of corresponding queueing processes as SP's.

As it was mentioned in introduction, the queueing processes here are in general more complicated comparatively to known models. Therefore, we restrict our analysis to study queueing processes without reflection and consider the convergence on the interval $[0, T]$ such that in each component $s(t) > 0$, $t \in [0, T]$. The analysis of reflecting processes should be detached into a separate problem.

In this section, as an illustration of a general approach we consider some classes of overloaded state-dependent Markov queueing systems and networks.

3.1. Markov queueing systems

Consider rather general Markov system $\overline{M}_{\overline{Q},B}/\overline{M}_{\overline{Q},B}/1/\infty$, which includes state-dependent systems with batch arrivals and service, systems with different types of calls, impatient calls, etc.

Suppose that characteristics of the system depend on some scaling parameter $n \to \infty$. Let nonnegative functions $\lambda(\overline{q})$, $\mu(\overline{q})$, $\nu_i(\overline{q})$, $i = \overline{1, m}$, $\overline{q} \in \mathcal{R}_+^m$, be given. Let also $\overline{\alpha}(\overline{q})$, $\overline{\gamma}(\overline{q})$, $\overline{\beta}_i(\overline{q})$, $i = \overline{1, m}$, $\overline{q} \in \mathcal{R}_+^m$, be random variables with values in $\mathcal{R}_+^m$. There is one server and an infinite number of waiting places. Denote by $\overline{Q}_n(t)$ the number of calls in the system at time t, $\overline{Q}_n(t) \in R_+^m$. Vector values may denote the different classes of calls (or different priorities). The system operates in the following way: if $\overline{Q}_n(t) = n\overline{q}$, then with the local rate $\lambda(\overline{q})$ a batch of $\overline{\alpha}(\overline{q})$ calls may enter the system. Correspondingly, with the local service rate $\mu(\overline{q})$ a batch of $\min\{\overline{\gamma}(\overline{q}), n\overline{q}\}$ calls may finish service (in the case of vector-valued variables the minimum is taken in each component). In addition to this, each call of type i in the queue independently of others with the local rate $n^{-1}\nu_i(\overline{q})$ may be transformed into $\overline{e}_i + \overline{\beta}_i(\overline{q})$ calls, where $\overline{e}_i$ is a vector with ith component is equal to one and other components are equal to 0. Calls after service completion leave the system. If a vector $\overline{\beta}_i(\overline{q})$ may have negative components (for instance, we have impatient calls), then after transformation we get $\min\{\overline{0}, n\overline{q} + \overline{\beta}_i(\overline{q})\}$ calls in the system.

Denote $\Lambda(\overline{q}) = \lambda(\overline{q}) + \mu(\overline{q}) + \nu(\overline{q})$, where $\nu(\overline{q}) = \sum_{i=1}^{m} q_i \nu_i(\overline{q})$, $\overline{q} = (q_1, \ldots, q_m)$, and introduce the following moment functions:

$$\begin{array}{lll} \overline{m}^{(1)}(\overline{q}) = \mathbf{E}\overline{\alpha}(\overline{q}), & \overline{m}^{(2)}(\overline{q}) = \mathbf{E}\overline{\gamma}(\overline{q}), & \overline{m}_i^{(3)}(\overline{q}) = \mathbf{E}\overline{\beta}_i(\overline{q}), \\ d^{(1)}(\overline{q}) = \mathbf{E}\overline{\alpha}(\overline{q})\overline{\alpha}(\overline{q})^*, & d^{(2)}(\overline{q}) = \mathbf{E}\overline{\gamma}(\overline{q})\overline{\gamma}(\overline{q})^*, & d_i^{(3)}(\overline{q}) = \mathbf{E}\overline{\beta}_i(\overline{q})\overline{\beta}_i(\overline{q})^*, \end{array}$$

where an expectation is taken in each component, and a^* denotes the conjugate vector. Put

$$\overline{b}(\overline{q}) = \overline{m}^{(1)}(\overline{q})\lambda(\overline{q}) - \overline{m}^{(2)}(\overline{q})\mu(\overline{q}) + \sum_{i=1}^{m} \overline{m}_i^{(3)}(\overline{q}) q_i \nu_i(\overline{q}),$$

$$B^2(\overline{q}) = d^{(1)}(\overline{q})\lambda(\overline{q}) + d^{(2)}(\overline{q})\mu(\overline{q}) + \sum_{i=1}^{m} d_i^{(3)}(\overline{q}) q_i \nu_i(\overline{q}).$$

Let also $G(\overline{q})$ be the matrix of partial derivatives for $\overline{b}(\overline{q})$:

$$\lim_{h\to 0} h^{-1}\big(b(\overline{q} + h\overline{z}) - b(\overline{q})\big) = G(\overline{q})\overline{z}, \quad \overline{z} \in \mathcal{R}^m.$$

Furthermore, for any two vectors $\overline{a}$ and $\overline{b}$, the inequality $\overline{a} > \overline{b}$ means that $a_i > b_i$ for all components. Denote by $\overline{s}(t)$ a solution of the equation

$$\mathrm{d}\overline{s}(t) = \overline{b}\big(\overline{s}(t)\big)\,\mathrm{d}t, \qquad \overline{s}(0) = \overline{s}_0. \tag{3.1}$$

Theorem 3.1. (1) Suppose that in any bounded and closed domain in $int\{\mathcal{R}_+^m\}$ the variables $\overline{\alpha}(\overline{q})$, $\overline{\gamma}(\overline{q})$, $\overline{\beta}(\overline{q})$ are uniformly in $\overline{q}$ integrable, functions $\lambda(\overline{q})$, $\mu(\overline{q})$, $\nu_i(\overline{q})$, $\overline{m}_i^{(j)}(\overline{q})$ are locally Lipschitz, and $\Lambda(\overline{q}) > 0$. Let also

$$n^{-1}\overline{Q}_n(0) \xrightarrow{\mathrm{P}} \overline{s}_0, \tag{3.2}$$

where $\overline{s}_0 > \overline{0}$ is some deterministic value, there exist $T > 0$ such that $\overline{s}(t) > \overline{0}$, $t \in [0, T]$, and also $y(+\infty) > T$, where $y(t) = \int_0^t \Lambda(\overline{\eta}(u))^{-1}\,\mathrm{d}u$, and the function $\overline{\eta}(t)$ satisfies the equation

$$\overline{\eta}(0) = \overline{s}_0, \qquad \mathrm{d}\overline{\eta}(t) = \overline{b}\big(\overline{\eta}(t)\big)\Lambda\big(\overline{\eta}(t)\big)^{-1}\,\mathrm{d}t, \tag{3.3}$$

a unique solution of which exists on any interval.

Then a unique solution of (3.1) exists on $[0, T]$ and

$$\sup_{0\leqslant t\leqslant T} \big|n^{-1}\overline{Q}_n(nt) - \overline{s}(t)\big| \xrightarrow{\mathrm{P}} 0. \tag{3.4}$$

(2) Suppose, in addition, that variables $|\alpha(\overline{q})|^2$, $|\gamma(\overline{q})|^2$, $|\beta(\overline{q})|^2$ are integrable uniformly in $\overline{q}$ in any bounded and closed domain in $int\{\mathcal{R}_+^m\}$, functions $B^2(\overline{q})$ and $G(\overline{q})$ are continuous in $int\{\mathcal{R}_+^m\}$, and $n^{-1/2}(\overline{Q}_n(0) - n\overline{s}_0) \overset{\mathrm{w}}{\Rightarrow} \overline{\zeta}_0$.

Then the sequence of processes $\overline{\zeta}_n(t) = n^{-1/2}(\overline{Q}_n(nt) - n\overline{s}(t))$ weakly converges in $\mathcal{D}_T^r$ to a diffusion process $\overline{\zeta}(t)$ satisfying the following stochastic differential equation:

$$\mathrm{d}\overline{\zeta}(t) = G\big(\overline{s}(t)\big)\overline{\zeta}(t)\,\mathrm{d}t + B\big(\overline{s}(t)\big)\,\mathrm{d}\overline{w}(t), \qquad \overline{\zeta}(0) = \overline{\zeta}_0,$$

a unique solution of which exists on the interval $[0, T]$.

Here $int\{\mathcal{R}_+^m\} = \mathcal{R}_+^m \setminus \partial\mathcal{R}_+^m$ (the interior of $\mathcal{R}_+^m$), the matrix $B(\overline{q})$ satisfies the relation $B(\overline{q})B(\overline{q})^* = B(\overline{q})^2$, $\overline{w}(t)$ is a standard Wiener process in $\mathcal{R}^m$, symbols $\overset{\mathrm{P}}{\longrightarrow}$ and $\overset{\mathrm{w}}{\Rightarrow}$ mean the convergence in probability and the convergence in distribution, respectively, and the weak convergence of random processes in $\mathcal{D}_T^r$ means the weak convergence of probability measures induced by the processes on Skorokhod space D_T^r and endowed by Skorokhod topology [45].

Proof. Let us introduce jointly independent families of random variables $\{(\tau_{nk}(n\overline{q}), \overline{\xi}_{nk}(n\overline{q}))\}$, $k \geqslant 0$. Here $\tau_{nk}(n\overline{q})$ has an exponential distribution with parameter $\Lambda(\overline{q}) = \lambda(\overline{q}) + \mu(\overline{q}) + \nu(\overline{q})$, where $\nu(\overline{q}) = \sum_{i=1}^m q_i\nu_i(\overline{q})$, $\overline{q} = (q_1, \ldots, q_m)$. $\overline{\xi}_{nk}(n\overline{q})$ is independent of $\tau_{nk}(n\overline{q})$ and can be represented in the form:

$$\overline{\xi}_{n1}\big(n\overline{q}\big) = \begin{cases} \overline{\alpha}\big(\overline{q}\big), & \text{with probab. } \lambda\big(\overline{q}\big)\Lambda\big(\overline{q}\big)^{-1}, \\ -\overline{\gamma}\big(\overline{q}\big), & \text{with probab. } \mu_i\big(\overline{q}\big)\Lambda\big(\overline{q}\big)^{-1}, \\ \overline{\beta}_i\big(\overline{q}\big), & \text{with probab. } q_i\nu_i\big(\overline{q}\big)\Lambda\big(\overline{q}\big)^{-1},\ i = \overline{1, m}. \end{cases} \tag{3.5}$$

Now, to avoid the consideration of truncated random variables, we construct an auxiliary RPSM $\widetilde{Q}_n(t)$ defined in the whole space $\mathcal{R}^m$. Let $s_i(t)$ be the ith component of the function $s(t)$. Put $\delta = \min_{1\leqslant i\leqslant m}\min_{0\leqslant t\leqslant T} s_i(t)$. By the construction, $\delta > 0$. Take $\varepsilon = \delta/2$ and consider the orthant $R_+^m(\varepsilon) = \{\overline{a}\colon\ \overline{a} \in \mathcal{R}_+^m,\ a_i \geqslant \varepsilon,\ i = 1, \ldots, m\}$. Now we extend the introduced functions and random variables from the domain $R_+^m(\varepsilon)$ to the whole space $\mathcal{R}^m$ in the following way.

Let $f(\overline{q}), \overline{q} \in R_+^m(\varepsilon)$, be some given function. We define a function $\widetilde{f}(\overline{a}), \overline{a} \in \mathcal{R}^m$, according to the transformation: $\widetilde{f}(a_1, \ldots, a_m) = f(\max(a_1, \varepsilon), \ldots, \max(a_m, \varepsilon))$. By construction, in the domain $R_+^m(\varepsilon)$, $\widetilde{f}(\overline{q}) = f(\overline{q})$. If $f(\overline{q})$ is a continuous (locally Lipschitz) in $R_+^m(\varepsilon)$ function, then it is easy to check that $\widetilde{f}(\overline{a})$ is also continuous (locally Lipschitz) in R^m.

Using this transformation, we define the functions $\widetilde{\lambda}(\overline{a})$, $\widetilde{\mu}(\overline{a})$, $\widetilde{\nu}_i(\overline{a})$, $i = \overline{1, m}$, $\overline{a} \in \mathcal{R}^m$, and random variables $\widetilde{\alpha}(\overline{a})$, $\widetilde{\gamma}(\overline{a})$, $\widetilde{\beta}_i(\overline{a})$, $i = \overline{1, m}$, for any $\overline{a} \in \mathcal{R}^m$. Construct variables $\widetilde{\tau}_{nk}(n\overline{a})$ and $\widetilde{\xi}_{nk}(n\overline{a})$ as in (3.5) and above.

Now, using these variables, we can define according to (2.6) an RPSM $\widetilde{Q}_n(t)$. It may take values in $\mathcal{R}^m$, and, by construction, if on some interval $[0, T]$ $\widetilde{Q}_n(t) \geqslant n\varepsilon$, then its trajectory coincides with the trajectory of queue $\overline{Q}_n(t)$ on $[0, T]$.

Let us study the behavior of $\widetilde{Q}_n(t)$. As we can see, all conditions of theorem A.1 in appendix A are satisfied. Calculating the expectation of $\overline{\xi}_{n1}(n\overline{q})$, we get that $\widetilde{Q}_n(nt)$ satisfies relation (3.4) with the same function $\overline{s}(t)$. Consider now an interval $[0, T]$,

where $\overline{s}(t) > 0$, $t \in [0, T]$. Then, for chosen above $\varepsilon > 0$, we have $\overline{s}(t) \geqslant 2\varepsilon$, $t \in [0, T]$, and (3.4) implies that

$$\mathbf{P}\big(n^{-1}\widetilde{Q}_n(nt) \geqslant \varepsilon,\ t \in [0, T]\big) \to 1. \tag{3.6}$$

Let us construct now on the same probability space the queueing process $\overline{Q}_n(nt)$ and RPSM $\widetilde{Q}_n(nt)$ in a recurrent way as follows. We put $\widetilde{Q}_n(0) = \overline{Q}_n(0)$. Then we generate a sequence of uniformly distributed on [0, 1] random variables $\omega_1, \omega_2, \ldots$, and construct recursively on this sequence the variables $\widetilde{Q}_{nk}, \widetilde{\tau}_{nk}(\widetilde{Q}_{nk}), \widetilde{\xi}_{nk}(\widetilde{Q}_{nk}), k \geqslant 0$, according to (2.6) and using a standard simulation technique. For instance, we construct an exponential random variable by the formula $\tau_{nk}(\overline{Q}) = -\Lambda(n^{-1}\overline{Q})^{-1} \ln \omega_{3k}$, and $\overline{\xi}_{nk}(\overline{Q})$ is constructed by variables $\omega_{2k+1}, \omega_{2k+2}$ in two stages according to (3.5). Then we construct trajectories of $\overline{Q}_n(nt)$ and $\widetilde{Q}_n(nt)$, where a trajectory of $\widetilde{Q}_n(nt)$ is constructed directly according to relations (2.6) for variables with tilde. By construction, if on some interval $[0, T]$ $\widetilde{Q}_n(nt) \geqslant n\varepsilon$, then $\widetilde{Q}_n(nt) = \overline{Q}_n(nt)$, $t \in [0, T]$. Now for any measurable set A of functions from σ-algebra $\mathcal{B}_{\mathcal{D}_T^r}$ we have according to (3.6) as $n \to \infty$

$$\begin{aligned}
&\big|\mathbf{P}\big(n^{-1}\overline{Q}_n(nt) \in A,\ t \in [0, T]\big) - \mathbf{P}\big(n^{-1}\widetilde{Q}_n(nt) \in A,\ t \in [0, T]\big)\big| \\
&\quad \leqslant \big|\mathbf{P}\big(n^{-1}\overline{Q}_n(nt) \in A,\ \widetilde{Q}_n(nt) \geqslant n\varepsilon,\ t \in [0, T]\big) \\
&\qquad - \mathbf{P}\big(n^{-1}\widetilde{Q}_n(nt) \in A,\ \widetilde{Q}_n(nt) \geqslant n\varepsilon,\ t \in [0, T]\big)\big| \\
&\qquad + 2\mathbf{P}\big(\text{exists } u,\ u \in [0, T] \text{ such that } \widetilde{Q}_n(nu) < n\varepsilon\big) \\
&\quad = 2\mathbf{P}\big(\text{exists } u,\ u \in [0, T] \text{ such that } \widetilde{Q}_n(nu) < n\varepsilon\big) \to 0.
\end{aligned}$$

This relation shows that the asymptotic behavior of trajectories of the queue and auxiliary RPSM $\widetilde{Q}_n(nt)$ is the same and, finally, implies relation (3.4).

To prove the 2nd part of theorem 3.1, we first prove DA for the process $\widetilde{Q}_n(nt)$. The proof uses theorem A.2 given in appendix A. Then this result is extended using the same considerations as above to the process $\overline{Q}_n(nt)$. □

Remark 3.2. The result of theorem 3.1 is also valid if the value s_0 is a random variable, and corresponding relations involving s_0 are satisfied with probability one. These results can be also extended to the case of r servers.

Consider now as examples some special models.

3.1.1. A system $M_Q/M_Q/1/\infty$

Suppose that calls arrive and are served one at a time, there is only one type of calls, and there is no transformation of calls in the system. That is, $\alpha(q) \equiv 1$, $\nu_i(q) \equiv 0$, $q \geqslant 0$, $\gamma(q) \equiv 1, q > 0$, $\gamma(0) = 0$. Denote by $s(t)$ a solution of the equation:

$$\mathrm{d}s(t) = b\big(s(t)\big)\,\mathrm{d}t, \qquad s(0) = s_0, \tag{3.7}$$

where $b(q) = \lambda(q) - \mu(q)$.

The following result is a consequence of theorem 3.1 for the scalar case.

Corollary 3.3. (1) Suppose that (3.2) is true, $s_0 > 0$, on the open interval $(0, \infty)$ the functions $\lambda(q)$, $\mu(q)$ satisfy a local Lipschitz condition and $\lambda(q) + \mu(q) > 0$, there exists $T > 0$ such that $s(t) > 0$, as $0 < t \leqslant T$, and in addition $y(+\infty) > T$, where

$$y(t) = \int_0^t \big(\lambda\big(\eta(u)\big) + \mu\big(\eta(u)\big)\big)^{-1}\,\mathrm{d}u, \tag{3.8}$$

and the function $\eta(t)$ satisfies the equation

$$\eta(0) = s_0, \qquad \mathrm{d}\eta(t) = b\big(\eta(t)\big)\big(\lambda\big(\eta(u)\big) + \mu\big(\eta(u)\big)\big)^{-1}\,\mathrm{d}t, \tag{3.9}$$

a unique solution of which exists. Then

$$\sup_{0\leqslant t\leqslant T} \big|n^{-1}Q_n(nt) - s(t)\big| \xrightarrow{\mathrm{P}} 0. \tag{3.10}$$

(2) Suppose, in addition, that functions $\lambda(q)$, $\mu(q)$ are continuously differentiable in $(0, \infty)$, and $n^{-1/2}(Q_n(0) - ns_0) \overset{\mathrm{w}}{\Rightarrow} \zeta_0$. Then the sequence of processes $\zeta_n(t) = n^{-1/2}(Q_n(nt) - ns(t))$ weakly converges in $\mathcal{D}_T$ to the diffusion process $\zeta(t)$:

$$\mathrm{d}\zeta(t) = \big(\lambda'\big(s(t)\big) - \mu'\big(s(t)\big)\big)\zeta(t)\,\mathrm{d}t + \big(\lambda\big(s(t)\big) + \mu\big(s(t)\big)\big)^{1/2}\,\mathrm{d}w(t), \qquad \zeta(0) = \zeta_0. \tag{3.11}$$

Remark 3.4. Suppose that $s_0 = 0$, other conditions of corollary 3.3 hold and, in addition, $\lambda(q)$ is continuous in 0, there exists a limit $\mu(+0) = \lim \mu(q)$ as $q \searrow 0$, and $\lambda(0) > \mu(+0)$. Then (3.10) also holds.

Proof of remark 3.4. Suppose for simplicity that $Q_n(0) = 0$. Let there exist $T > 0$ such that $s(t) > 0$, as $0 < t \leqslant T$. As $\lambda(0) > \mu(+0)$, using the continuity of $\lambda(q)$ and $\mu(q)$ in $(0, T)$ we can find $\varepsilon > 0$ such that $\lambda_* - \mu^* = \delta > 0$, where $\lambda_* = \inf\{\lambda(q)\colon\ 0 \leqslant q \leqslant \varepsilon\}$, $\mu^* = \sup\{\mu(q)\colon\ 0 < q \leqslant \varepsilon\}$.

Let $\Pi_1(t)$ and $\Pi_2(t)$ be two independent Poisson processes with parameters λ_* and μ^*, respectively. Note that in the domain $Q_n(nt) \leqslant n\varepsilon$ the queue $Q_n(nt)$ stochastically dominates the process $\Pi_1(nt) - \Pi_2(nt)$. This implies for any $c > 0$ that

$$\mathbf{P}\big(\tau_n(\varepsilon) > c\big) \leqslant \mathbf{P}\big(\widetilde{\tau}_n(\varepsilon) > c\big),$$

where $\tau_n(\varepsilon) = \inf\{u\colon\ Q_n(nu) \geqslant n\varepsilon\}$, $\widetilde{\tau}_n(\varepsilon) = \inf\{u\colon\ \Pi_1(nt) - \Pi_2(nt) \geqslant n\varepsilon\}$. It is easy to see that as $n \to \infty$, $\widetilde{\tau}_n(\varepsilon) \xrightarrow{\mathrm{P}} \varepsilon/\delta$. Then for any $c > 0$

$$\lim_{\varepsilon\to 0} \limsup_{n\to\infty} \mathbf{P}\big(\tau_n(\varepsilon) > c\big) = 0,$$

and also for any $\varepsilon > 0$

$$\lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\big(\tau_n(\varepsilon) > c\big) = 0. \tag{3.12}$$

Now let us consider the behavior of $Q_n(nt)$ on the interval $[\tau_n(\varepsilon), T]$. As the sequence $\tau_n(\varepsilon)$ is stochastically bounded (see (3.12)), then for any sequence $n_k \to \infty$ we can

choose a subsequence n_{k_l} such that $\tau_{n_{k_l}}(\varepsilon) \stackrel{\mathrm{w}}{\Rightarrow} \tau_0(\varepsilon)$. Using Skorokhod construction of a common probability space, we can always assume without loss of generality that $\tau_{n_{k_l}}(\varepsilon) \xrightarrow{\mathrm{P}} \tau_0(\varepsilon)$.

By construction, $n^{-1}Q_n(n\tau_n(\varepsilon)) \xrightarrow{\mathrm{P}} \varepsilon > 0$. Thus, applying corollary 3.3, we get

$$\sup_{\tau_{n_{k_l}}(e)\leqslant t\leqslant T} \left|n_{k_l}^{-1}Q_{n_{k_l}}(n_{k_l}t) - s_\varepsilon(t)\right| \xrightarrow{\mathrm{P}} 0, \tag{3.13}$$

where $s_\varepsilon(t)$ is a solution of equation (3.7) on the interval $[\tau_0(\varepsilon), T]$ with initial value ε. As $\tau_0(\varepsilon) \to 0$, $\varepsilon \to 0$, using the continuity of the solution of a differential equation in the initial value, we get that $s_\varepsilon(t) \to s(t)$ as $\varepsilon \to 0$ uniformly on any interval $[\delta, T]$, $\delta > 0$. Now from (3.12), (3.13) and the relation

$$\sup_{0\leqslant t\leqslant \tau_n(e)} \left|n^{-1}Q_n(nt) - s(t)\right| \leqslant \varepsilon + \frac{1}{n} + \sup_{0\leqslant t\leqslant\tau_n(e)} s(t),$$

we, finally, get for any $\varepsilon > 0$, when $n = n_{k_l} \to \infty$,

$$\begin{aligned}
&P\lim_{n\to\infty} \sup_{0\leqslant t\leqslant T} \left|n^{-1}Q_n(nt) - s(t)\right| \\
&\quad\leqslant P\lim_{n\to\infty} \max\Big\{ \sup_{0\leqslant t\leqslant\tau_n(e)} \left|n^{-1}Q_n(nt) - s(t)\right|, \\
&\qquad\qquad \sup_{\tau_n(e)\leqslant t\leqslant T} \left\{\left|n^{-1}Q_n(nt) - s_\varepsilon(t)\right| + \left|s_\varepsilon(t) - s(t)\right|\right\}\Big\} \\
&\quad\leqslant \max\Big\{\varepsilon + \sup_{0\leqslant t\leqslant\tau_0(e)} s(t), \sup_{\tau_0(e)\leqslant t\leqslant T} \left|s_\varepsilon(t) - s(t)\right|\Big\},
\end{aligned}$$

where symbol $P\lim$ means the limit in probability, and the last term tends to 0 as $\varepsilon \to 0$.

Now we see that for any sequence n_k we can choose some subsequence n_{k_l}, for which (3.10) is true. So that (3.10) is true as $n \to \infty$. □

As we can see from remark 3.4, the result of theorem 3.1 can be extended to the case when some components of $\overline{s}_0$ may take values zero. For this case, we need to have some additional assumptions of non-ergodicity on the border. In addition, we have to prove that if the process starts in any point s on the border, then the 1st time $\tau_n(s, \varepsilon)$, when all components are greater then ε, should satisfy the property: for any $c > 0$ $\lim_{\varepsilon\to 0} \limsup_{n\to\infty} \mathbf{P}(\tau_n(s, \varepsilon) > c) = 0$.

Consider some particular applications of theorem 3.1.

Case 1. Let $\lambda(q) \equiv \lambda$, $q \geqslant 0$, $\mu(q) \equiv \mu$, $q > 0$ ($\mu(0) = 0$). Then our system is equivalent to a classical system $M/M/1/\infty$. In this case $s(t) = s_0 + (\lambda - \mu)t$ as $s_0 > 0$. Consider the relation between T and parameters of the system. Obviously, $y(+\infty) > T$ for any T (see (3.8)). If $\lambda \geqslant \mu$, then $s(t) > 0$ for any $t > 0$, and (3.4) is true for any $T > 0$. If $\lambda < \mu$, then $s(t) > 0$ for $0 < t < s_0(\mu - \lambda)^{-1}$, and (3.4) is true for any $T < s_0(\mu - \lambda)^{-1}$.

Consider the behavior of the first time when the queue becomes zero: $\psi_n(Q) = \inf\{t\colon\ t \geqslant 0,\ Q_n(t) = 0 \text{ given that } Q_n(0) = Q\}$. This time is a continuous functional concerning uniform convergence in probability to a monotone function. Therefore, if $\lambda < \mu$ and $n^{-1}Q_n(0) \xrightarrow{\mathrm{P}} s_0 > 0$, then $n^{-1}\psi_n(Q_n(0)) \xrightarrow{\mathrm{P}} s_0(\mu - \lambda)^{-1}$.

From (3.11) we easy get that $\zeta(t) = \zeta_0 + (\lambda + \mu)^{1/2} w(t)$, $0 \leqslant t \leqslant T$.

Case 2. Let $\lambda(q) \equiv \lambda,\ \mu(q) \equiv \mu q,\ q \geqslant 0$. Then our system is equivalent to a system $M/M/\infty$. In this case (3.1) has the form: $s(0) = s_0 > 0$, $\mathrm{d}s(t) = (\lambda - \mu s(t))\,\mathrm{d}t$, and $s(t) = \lambda/\mu + (s_0 - \lambda/\mu)\mathrm{e}^{-\mu t}$, $t \geqslant 0$.

Let us show that (3.10) holds for any $T > 0$. In our case (3.9) has the form

$$\mathrm{d}\eta(t) = \big(\lambda - \mu\eta(t)\big)\big(\lambda + \mu\eta(t)\big)^{-1}\,\mathrm{d}t. \tag{3.14}$$

We can see that the function $\eta(t)$ strictly monotonically increases in the domain $\eta(t) < \lambda/\mu$, and it strictly monotonically decreases in the domain $\eta(t) > \lambda/\mu$. That means $\eta(t) > 0$ for any $t > 0$, and there exists a limit $\eta_\infty = \lim_{t\to\infty}\eta(t)$. If $\eta_\infty \neq \lambda/\mu$, then (3.14) implies that there exists a limit $\eta'_\infty = \lim_{t\to\infty}\eta'(t) = (\lambda - \mu\eta_\infty)(\lambda + \mu\eta_\infty)^{-1} \neq 0$. In this way we get a contradiction, because from the one side for any $a > 0$, $\eta(t + a) - \eta(t) \to 0$, as $t \to \infty$, and from the another side $\eta(t + a) - \eta(t) = \int_t^{t+a}\eta'(u)\,\mathrm{d}u \to a\eta'_\infty \neq 0$. That is why it should be $\eta'_\infty = 0$ and $\eta_\infty = \lambda/\mu$.

This implies according to (3.8) that $y(t) = \int_0^t(\lambda + \mu\eta(u))^{-1}\,\mathrm{d}u \to \infty$ as $t \to \infty$. Finally, we get $y(\infty) > T$, and (3.10) holds for any $T > 0$.

Equation (3.11) has the form: $\mathrm{d}\zeta(t) = -\mu\zeta(t)\,\mathrm{d}t + (\lambda + \mu s(t))^{1/2}\,\mathrm{d}w(t)$. This is an Ornstein–Uhlenbeck type process, and a solution can be written in the explicit form. Note that the convergence of the process $n^{-1/2}(Q_n(nt) - n\lambda/\mu)$ to Ornstein–Uhlenbeck process for the system $M/M/\infty$ was obtained in [31].

3.1.2. An output process

Consider a system $M_Q/M_Q/1/\infty$ described above. Denote by $Z_n(t)$ the total number of calls served on the interval $[0, t]$.

Corollary 3.5. If conditions of corollary 3.3 are satisfied, then (3.10) holds and

$$\sup_{0\leqslant t\leqslant T}\big|n^{-1}Z_n(nt) - g(t)\big| \xrightarrow{\mathrm{P}} 0, \tag{3.15}$$

where

$$\mathrm{d}s(t) = \big(\lambda\big(s(t)\big) - \mu\big(s(t)\big)\big)\,\mathrm{d}t, \quad s(0) = s_0 > 0, \quad g(t) = \int_0^t \mu\big(s(u)\big)\,\mathrm{d}u. \tag{3.16}$$

Correspondingly, the sequence $(n^{-1/2}(Q_n(nt) - ns(t)), n^{-1/2}(Z_n(nt) - ng(t)))$ weakly converges in $\mathcal{D}_T$ to the diffusion process $(\zeta(t), \kappa(t))$ satisfying the system of stochastic differential equations:

$$\mathrm{d}\zeta(t) = \big(\lambda'\big(s(t)\big) - \mu'\big(s(t)\big)\big)\zeta(t)\,\mathrm{d}t + \frac{1}{\sqrt{2}}\Big(\Big[\sqrt{\lambda\big(s(t)\big)} + \sqrt{\mu\big(s(t)\big)}\Big]\,\mathrm{d}w_1(t)$$

$$+\Big[\sqrt{\lambda\big(s(t)\big)}-\sqrt{\mu\big(s(t)\big)}\Big]\,\mathrm{d}w_2(t)\Big),\qquad \zeta(0)=\zeta_0, \tag{3.17}$$

$$\mathrm{d}\kappa(t)=\mu'\big(s(t)\big)\zeta(t)\,\mathrm{d}t-\frac{1}{\sqrt{2}}\sqrt{\mu\big(s(t)\big)}\big(\mathrm{d}w_1(t)-\mathrm{d}w_2(t)\big),\qquad \kappa(0)=0,$$

where $w_1(t)$ and $w_2(t)$ are two independent standard Wiener processes.

Proof. We can represent the process $(Q_n(t), Z_n(t))$, $t \geqslant 0$, as a vector-valued RPSM. Here $t_{nk}, k \geqslant 0$, are constructed in the same way as above. If at time t_{nk} $(n^{-1}Q_n(t_{nk}), n^{-1}Z_n(t_{nk})) = (q, g)$, then distributions of variables $\tau_{nk}(nq)$ and $\overline{\xi}_n(nq) = (\xi_n^{(1)}(nq), \xi_n^{(2)}(nq))$ depend only on the first component q, $\tau_{nk}(nq)$ has an exponential distribution with parameter $\Lambda(q) = \lambda(q) + \mu(q)$, and

$$\overline{\xi}_{n1}(nq)=\begin{cases}(1,0), & \text{with probab. } \lambda(q)\Lambda(q)^{-1},\\ (-1,1), & \text{with probab. } \mu(q)\Lambda(q)^{-1}.\end{cases}$$

Now we use theorems A.1, A.2 from appendix A. Following notation of these theorems it is easy to calculate that if $\alpha = (q, g)$, $z = (z_1, z_2)$, then $m(\alpha) = \Lambda(q)^{-1}$, $b(\alpha) = (\lambda(q) - \mu(q), \mu(q))\Lambda(q)^{-1}$, $q(\alpha, z) = ((\lambda'(q) - \mu'(q))z_1, \mu'(q)z_2)$, and

$$D^2(\alpha)=\begin{pmatrix}\lambda(q)+\mu(q) & -\mu(q)\\ -\mu(q) & \mu(q)\end{pmatrix}\Lambda(q)^{-1}.$$

Now we can calculate the matrix $D(\alpha)$ using the relation $D^2(\alpha) = D(\alpha)D(\alpha)^*$, and from equations (A.7), (A.10) it is not difficult to get (3.16), (3.17). □

Note that results of this part can be extended to nonhomogeneous in time models also. Consider for the illustration the following model.

3.1.3. Time-dependent system $M_{Q,t}/M_{Q,t}/1/\infty$

Consider a queueing system described in section 3.1.1 with the additional dependence of service and arrival rates on time: if at time nt $Q_n(nt) = nq$, then the local arrival rate is $\lambda_n(q, t)$ and the service rate is $\mu_n(q, t)$.

Suppose that functions $\lambda_n(q, t)$ and $\mu_n(q, t)$ satisfy the following condition: in each bounded domain $\max\{t_1, t_2\} \leqslant N$, $\max\{q_1, q_2\} \leqslant L$, $q_1, q_2 > 0$,

$$\big|\lambda_n(q_1,t_1)-\lambda_n(q_2,t_2)\big|\leqslant C_{N,L}\big(|q_1-q_2|+|t_1-t_2|\big), \tag{3.18}$$

(the same for $\mu_n(\cdot)$), there exist constants $0 < C_0 < C_1 < \infty$ and functions $\lambda(q, t)$, $\mu(q, t)$ such that for any $t \geqslant 0, q > 0$

$$C_0\leqslant\lambda_n(q,t)+\mu_n(q,t)\leqslant C_1, \tag{3.19}$$

$$\lim_{n\to\infty}\lambda_n(q,t)=\lambda(q,t),\qquad \lim_{n\to\infty}\mu_n(q,t)=\mu(q,t). \tag{3.20}$$

Denote $\Lambda(q, t) = \lambda(q, t) + \mu(q, t)$. Let $s(t)$ be a solution of the equation

$$\mathrm{d}s(t)=\big(\lambda\big(s(t),t\big)-\mu\big(s(t),t\big)\big)\,\mathrm{d}t,\qquad s(0)=s_0. \tag{3.21}$$

Corollary 3.6. (1) Suppose that (3.2) is true with $s_0 > 0$, there exists $T > 0$ such that $s(t) > 0$, as $0 < t \leqslant T$, and $y(+\infty) > T$, where $y(t) = \int_0^t \Lambda(\eta(u), u)^{-1}\,\mathrm{d}u$, and $\eta(t)$ satisfies the equation

$$\eta(0) = s_0, \qquad \mathrm{d}\eta(t) = \big(\lambda\big(\eta(t), t\big) - \mu\big(\eta(t), t\big)\big)\Lambda\big(\eta(t)\big)^{-1}\,\mathrm{d}t,$$

a unique solution of which exists. Then relation (3.10) holds.

(2) Suppose in addition that functions $\lambda(q, t)$, $\mu(q, t)$ are continuously differentiable in q in the domain $(0, \infty) \times [0, T]$, and $n^{-1/2}(Q_n(0) - ns_0) \overset{\mathrm{w}}{\Rightarrow} \zeta_0$. Then the sequence $\zeta_n(t) = n^{-1/2}(Q_n(nt) - ns(t))$ weakly converges in $\mathcal{D}_T$ to the diffusion process $\zeta(t)$:

$$\mathrm{d}\zeta(t) = \big(\lambda'_q\big(s(t), t\big) - \mu'_q\big(s(t), t\big)\big)\zeta(t)\,\mathrm{d}t + \Lambda\big(s(t), t\big)^{1/2}\,\mathrm{d}w(t), \qquad \zeta(0) = \zeta_0.$$

Proof. The proof follows the same scheme as above. We use theorem A.1. Switching times $t_{n1} < t_{n2} < \cdots$ are chosen as times of any changing in the system. Put $S_{nk} = (Q_n(t_{nk}), t_{nk})$, $k > 0$. Then the argument α in theorem A.1 has the form $\alpha = (q, t)$. At any $q \geqslant 0$, $t \geqslant 0$, define the family of jointly independent in k variables $(\xi_{nk}(nq, nt), \tau_{nk}(nq, nt))$, $k > 0$, as follows:

$$\begin{aligned}&\mathbf{P}\big(\xi_{nk}(nq, nt) \leqslant z,\ \tau_{nk}(nq, nt) \leqslant u\big)\\ &\quad = \mathbf{P}\big(Q_n(t_{n,k+1}) - Q(t_{nk}) \leqslant z,\ t_{n,k+1} - t_{nk} \leqslant u \mid Q_n(t_{nk}) = nq,\ t_{nk} = nt\big).\end{aligned}$$

Here the variable $\xi_{nk}(nq, nt)$ takes values $+1$ or -1 with some probabilities $p_n(q, t)$ or $1 - p_n(q, t)$, respectively.

Using relations (3.18)–(3.20) it is not difficult to prove that for any $k > 0$ the variables $\xi_{nk}(nq, nt)$ and $\tau_{nk}(nq, nt)$ are asymptotically independent, $\tau_{nk}(nq, nt)$ is asymptotically close to the exponential distribution with parameter $\lambda(q, t) + \mu(q, t)$, and, as $n \to \infty$, uniformly in each bounded domain $\max\{t_1, t_2\} \leqslant N$, $c \leqslant \min\{q_1, q_2\}$, $\max\{q_1, q_2\} \leqslant L$ $(c > 0)$,

$$\mathbf{E}\tau_{nk}(nq, nt) \to \Lambda(q, t)^{-1}, \qquad \mathbf{E}\xi_{nk}(nq, nt) \to \big(\lambda(q, t) - \mu(q, t)\big)\Lambda(q, t)^{-1}.$$

These relations correspond to condition (A.4). Then we follow the same lines as in the proof of theorem 3.1 and construct an auxiliary RPSM, for which all other conditions of theorem A.1 can be checked. Finally, this implies relation (3.10) with $s(t)$ defined in (3.21). In a similar way we can prove DA. □

Note that time-dependent and state-dependent Markov queueing models in heavy traffic conditions are studied using a martingale technique in [37–39]. We consider a simple overloaded model $M_{Q,t}/M_{Q,t}/1/\infty$ just for the illustration of possibilities of a suggested approach. Using the same technique, these results can be extended to time-dependent and state-dependent Markov queueing networks, models in nonhomogeneous quasi-ergodic Markov environment (limit theorems for SP in quasi-ergodic Markov environment are considered in [3]), and also to non-Markov models considered in section 4.

3.1.4. A system with impatient calls

Consider a time-homogeneous system $M_Q/M_Q/1/\infty$ with impatient calls. Suppose that calls arrive and are served one at a time, and, as $Q_n(t) = nq$, the local arrival and service rates are $\lambda(q)$ and $\mu(q)$, respectively. In addition, each call in the queue independently of others with rate $n^{-1}\nu(q)$ may leave the system.

Then in notation of theorem 3.1 $\alpha(q) \equiv 1$, $q \geqslant 0$, $\gamma(q) = 1$, $\beta(q) = -1$, for $q > 0$, and $\gamma(0) = 0$, $\beta(0) = 0$, $\Lambda(q) = \lambda(q) + \mu(q) + q\nu(q)$, $b(q) = \lambda(q) - \mu(q) - q\nu(q)$, $B^2(q) = \lambda(q) + \mu(q) + q\nu(q)$, $G(q) = \lambda'(q) - \mu'(q) - \nu(q) - q\nu'(q)$, $q > 0$, and equations (3.1), (3.3) can be written in the general form.

Consider a particular case, when $\lambda(q) \equiv \lambda$, $q \geqslant 0$, $\mu(q) \equiv \mu$, $\nu(q) \equiv \nu$, $q > 0$. Then

$$\mathrm{d}s(t) = \big(\lambda - \mu - \nu s(t)\big)\,\mathrm{d}t, \qquad \mathrm{d}\zeta(t) = -\nu\zeta(t)\,\mathrm{d}t + \big(\lambda + \mu + \nu s(t)\big)^{1/2}\,\mathrm{d}w(t),$$

where $s(0) = s_0$, $\zeta(0) = \zeta_0$. Solving these equations we find:

$$s(t) = \nu^{-1}(\lambda - \mu) + \big(s_0 - \nu^{-1}(\lambda - \mu)\big)\mathrm{e}^{-\nu t}, \qquad \zeta(t) = \mathrm{e}^{-\nu t}\big(\zeta_0 + w\big(\psi(t)\big)\big),$$

where $\psi(t) = \nu^{-1}(\lambda - \mu)(\mathrm{e}^{2t\nu} - 1) - \nu^{-1}(\lambda - \mu - \nu s_0)(\mathrm{e}^{t\nu} - 1)$.

If $\lambda \geqslant \mu$, then in the same way, as it was done in section 3.1.1, we can show that (3.4) holds for any $T > 0$. In this case we have a quasi-stationary point $s^* = \nu^{-1}(\lambda - \mu)$, that is, as $n \to \infty$ and $t \to \infty$, $n^{-1}Q_n(nt) \xrightarrow{\mathrm{P}} s^*$.

If $\lambda < \mu$, then (3.4) holds on the interval $[0, T]$, where $T < \nu^{-1}\ln((\mu - \lambda + \nu s_0)/(\mu - \lambda))$.

3.2. Markov state-dependent networks

Consider a queueing network $(M_{Q,B}/M_{Q,B}/1/\infty)^r$ with batch state-dependent arrival process and service. It consists of r nodes with one server at each node and an infinite number of waiting places. The local characteristics of the network depend on some scaling parameter n. Denote by $Q_n(i, t)$ a number of calls at node i at time t and put $\overline{Q}_n(t) = (Q_n(i, t), i = \overline{1, r})$. Let the following values be given:

(1) nonnegative functions $\lambda_i(\overline{q})$, $\mu_i(\overline{q})$ and $\nu_i(\overline{q})$, $i = \overline{1, r}$, where $\overline{q} = (q_1, \ldots, q_r)$;

(2) families of integer random variables $\delta_i(\overline{q})$, $\gamma_i(\overline{q})$ with values in $\{0, 1, \ldots\}$ and variables $\beta_i(\overline{q})$ with values in $\{0, \pm 1, \ldots\}$, $i = \overline{1, r}$;

(3) a family of stochastic matrices $P(\overline{q}) = \|p_{ij}(\overline{q})\|_{i=\overline{1,r},\, j=\overline{1,r+1}}$;

(4) the initial vector $\overline{Q}_n(0)$.

The system operates as follows. If at time t, $\overline{Q}_n(t) = n\overline{q}$, then:

(1) with local arrival rate $\lambda_i(\overline{q})$, $\delta_i(\overline{q})$ calls may enter node i, $i = \overline{1, r}$;

(2) with local rate $\mu_i(\overline{q})$, $\min\{\gamma_i(\overline{q}), q_i\}$ calls may complete service at node i and all of them either with probability $p_{ij}(\overline{q})$ go to node j, $j = \overline{1, r}$, or with probability $p_{i,r+1}(\overline{q})$ leave the network;

(3) each call in the queue at node i independently of others with local rate $n^{-1}\nu_i(\overline{q})$, may be transformed into $\max\{\beta_i(\overline{q}), 1 - nq_i\}$ calls, $i = \overline{1,r}$.

In this case the process $\overline{Q}_n(t)$, $t \geqslant 0$, is a multidimensional MP. Suppose that there exist the 1st and 2nd moment functions of introduced variables. Denote

$$m_i(\overline{q}) = \mathbf{E}\delta_i(\overline{q}), \qquad g_i(\overline{q}) = \mathbf{E}\gamma_i(\overline{q}), \qquad e_i(\overline{q}) = \mathbf{E}\beta_i(\overline{q}) - 1,$$
$$\Lambda(\overline{q}) = \sum_{i=1}^{r} \bigl(\lambda_i(\overline{q}) + \mu_i(\overline{q}) + q_i\nu_i(\overline{q})\bigr),$$
$$a_i^2(\overline{q}) = \mathbf{E}\delta_i^2(\overline{q}), \qquad c_i^2(\overline{q}) = \mathbf{E}\gamma_i^2(\overline{q}), \qquad d_i^2(\overline{q}) = \mathbf{E}\bigl(\beta_i(\overline{q}) - 1\bigr)^2, \quad i = \overline{1,r}.$$

Let us introduce the following column vector-functions:

$$\overline{m}(\overline{q}) = \bigl(\lambda_1(\overline{q})m_1(\overline{q}), \ldots, \lambda_r(\overline{q})m_r(\overline{q})\bigr), \qquad \overline{g}(\overline{q}) = \bigl(\mu_1(\overline{q})g_1(\overline{q}), \ldots, \mu_r(\overline{q})g_r(\overline{q})\bigr),$$
$$\overline{e}(\overline{q}) = \bigl(q^{(1)}\nu_1(\overline{q}), \ldots, q^{(r)}\nu_r(\overline{q})\bigr), \qquad \overline{b}(\overline{q}) = \overline{m}(\overline{q}) - \bigl(I - P_0(\overline{q})^*\bigr)\overline{g}(\overline{q}) + \overline{e}(\overline{q}),$$

where I is the unit matrix, $P_0(\overline{q}) = ||p_{ij}(\overline{q})||_{i,j=\overline{1,r}}$, and symbol "*" denotes the operation of transposition.

Let $G(\overline{q}) = \overline{b}'(\overline{q})$ be the matrix of partial derivatives of $\overline{b}(\overline{q})$, and $B^2(\overline{q}) = \|b_{ij}(\overline{q})\|_{i,j=\overline{1,r}}$ be the matrix with the following elements:

$$b_{ij}(\overline{q}) = -\mu_i(\overline{q})p_{ij}(\overline{q})c_i^2(\overline{q}) - \mu_j(\overline{q})p_{ji}(\overline{q})c_j^2(\overline{q}), \quad i \neq j,$$
$$b_{ii}(\overline{q}) = -2\mu_i(\overline{q})p_{ii}(\overline{q})c_i^2(\overline{q}) + \lambda_i(\overline{q})a_i^2(\overline{q}) + \mu_i(\overline{q})c_i^2(\overline{q}) + \sum_{k=1}^{r} \mu_k(\overline{q})p_{ki}(\overline{q})c_k^2(\overline{q}) + q_i\nu_i(\overline{q})d_i^2(\overline{q}), \quad i = \overline{1,r}.$$

Denote by $\overline{s}(t)$ a solution of the equation

$$\overline{s}(0) = \overline{s}_0, \qquad \mathrm{d}\overline{s}(t) = \overline{b}\bigl(\overline{s}(t)\bigr)\,\mathrm{d}t. \tag{3.22}$$

Theorem 3.7. (1) Suppose that in any bounded and closed domain in $int\{\mathcal{R}_+^r\}$ the variables $\delta_i(\overline{q})$, $\gamma_i(\overline{q})$, $\beta_i(\overline{q})$, $i = \overline{1,r}$, are uniformly in $\overline{q}$ integrable, the functions $\lambda_i(\overline{q})$, $\mu_i(\overline{q})$, $\nu_i(\overline{q})$, $m_i(\overline{q})$, $g_i(\overline{q})$, $e_i(\overline{q})$, $i = \overline{1,r}$, $P(\overline{q})$ satisfy local Lipschitz condition, $\Lambda(\overline{q}) > 0$, $n^{-1}\overline{Q}_n(0) \xrightarrow{\mathrm{P}} \overline{s}_0 > \overline{0}$, the equation $\mathrm{d}\overline{\eta}(t) = \overline{b}(\overline{\eta}(t))\Lambda(\overline{\eta}(t))^{-1}\,\mathrm{d}t$, $\overline{\eta}(0) = \overline{s}_0$, has a unique solution $\overline{\eta}(t)$, and there exists $T > 0$ such that $\overline{s}(t) > \overline{0}$ in each component as $t \in [0, T]$, and $\int_0^\infty \Lambda(\overline{\eta}(t))^{-1}\,\mathrm{d}t > T$.

Then

$$\sup_{0 \leqslant t \leqslant T} \bigl|n^{-1}\overline{Q}_n(nt) - \overline{s}(t)\bigr| \xrightarrow{\mathrm{P}} 0. \tag{3.23}$$

(2) If in addition in any bounded and closed domain in $int\{\mathcal{R}_+^r\}$ random variables $\delta_i(\overline{q})^2$, $\gamma_i(\overline{q})^2$, $\beta_i(\overline{q})^2$, $i = \overline{1,r}$, are integrable uniformly in $\overline{q}$, functions $\lambda_i(\overline{q})$, $\mu_i(\overline{q})$, $\nu_i(\overline{q})$, $m_i(\overline{q})$, $g_i(\overline{q})$, $e_i(\overline{q})$, $i = \overline{1,r}$, $P(\overline{q})$ are continuously differentiable, and

$n^{-1/2}(\overline{Q}_n(0) - n\overline{s}_0) \stackrel{\mathrm{w}}{\Rightarrow} \overline{\gamma}_0$, then the sequence $\overline{\gamma}_n(t) = n^{-1/2}(\overline{Q}_n(nt) - n\overline{s}(t))$ weakly converges in $\mathcal{D}_T^r$ to the multidimensional diffusion process $\overline{\gamma}(t)$:

$$\mathrm{d}\overline{\gamma}(t) = G\big(\overline{s}(t)\big)\overline{\gamma}(t)\,\mathrm{d}t + B\big(\overline{s}(t)\big)\,\mathrm{d}\overline{w}(t), \qquad \overline{\gamma}(0) = \overline{\gamma}_0, \tag{3.24}$$

where $B(\overline{q})B(\overline{q})^* = B^2(\overline{q})$, and $\overline{w}(t)$ is a standard Wiener process in $\mathcal{R}^r$.

Proof. First, we construct an auxiliary RPSM $\widetilde{Q}_n(t)$ by analogy to theorem 3.1. It is an MP which is constructed by families of random variables $\{(\tau_{nk}(n\overline{q}), \overline{\xi}_{nk}(n\overline{q}))\}$, $k \geqslant 0$. Here $\tau_{n1}(n\overline{q})$ has an exponential distribution with parameter $\Lambda(\overline{q})$, $\xi_{n1}(n\overline{q})$ does not depend on $\tau_{n1}(n\overline{q})$, and

$$\overline{\xi}_{n1}\big(n\overline{q}\big) = \begin{cases} \delta_i\big(\overline{q}\big)\overline{e}_i, & \text{with probab. } \lambda_i\big(\overline{q}\big)\Lambda\big(\overline{q}\big)^{-1} \\ \gamma_i\big(\overline{q}\big)\big(\overline{e}_j - \overline{e}_i\big), & \text{with probab. } p_{ij}\big(\overline{q}\big)\mu_i\big(\overline{q}\big)\Lambda\big(\overline{q}\big)^{-1} \\ -\gamma_i\big(\overline{q}\big)\overline{e}_i, & \text{with probab. } p_{i,r+1}\big(\overline{q}\big)\mu_i\big(\overline{q}\big)\Lambda\big(\overline{q}\big)^{-1} \\ \big(\beta_i\big(\overline{q}\big) - 1\big)\overline{e}_i, & \text{with probab. } q_i\nu_i\big(\overline{q}\big)\Lambda\big(\overline{q}\big)^{-1},\ i, j = \overline{1, r}. \end{cases}$$

Now we follow the same lines about the equivalence of trajectories of $\widetilde{Q}_n(t)$ and $\overline{Q}_n(t)$ as in theorem 3.1. Finally, calculating moment characteristics of these variables and using theorems A.1, A.2, we get the statement of theorem 3.7. □

In particular, if $\lambda_i(\overline{q}) \equiv 0$, $p_{ir+1}(\overline{q}) \equiv 0$, $i = \overline{1, r}$, then our network is closed.

Example 3.8. Let $\lambda_i(\overline{q}) \equiv \lambda_i$, $\mu_i(\overline{q}) \equiv \mu_i q_i$, $p_{ij}(\overline{q}) \equiv p_{ij}$ for $\overline{q} \geqslant \overline{0}$, $i = \overline{1, r}$, $j = \overline{1, r+1}$. Then our network is equivalent to a classical network $(M/M/\infty)^r$. In this case

$$\mathrm{d}\overline{s}(t) = \big(\overline{\lambda} + \big(P_0^* - I\big)A\overline{s}(t)\big)\,\mathrm{d}t, \tag{3.25}$$

where $\overline{\lambda} = (\lambda_1, \ldots, \lambda_r)$, A is a diagonal matrix with elements μ_i, $i = \overline{1, r}$, and $P_0 = \|p_{ij}\|_{i,j=\overline{1,r}}$.

Suppose that the matrix $P_0^* - I$ is invertible. Then, iterating equation (3.25), we obtain a representation

$$\overline{s}(t) = \overline{q}_* + \exp\big\{\big(P_0^* - I\big)At\big\}\big(\overline{s}_0 - \overline{q}_*\big),$$

where $\overline{q}_* = A^{-1}(I - P_0^*)^{-1}\overline{\lambda}$ ($\overline{q}_*$ is the stationary point). Equation (3.24) has the form

$$\mathrm{d}\overline{\gamma}(t) = \big(P_0^* - I\big)A\overline{\gamma}(t)\,\mathrm{d}t + B\big(\overline{s}(t)\big)\,\mathrm{d}\overline{w}(t). \tag{3.26}$$

If $\overline{s}_0 = \overline{q}_*$, then for all $t > 0$ $\overline{s}(t) \equiv \overline{q}_*$, and we have a quasi-stationary regime and a stationary form for the equation (3.26) with the matrix of diffusion $B = B(\overline{q}_*)$.

The general model of our system gives us the possibility to consider networks with impatient calls and unreliable servers also. Some other examples of Markov state-dependent models are studied in the book [13]. Some Markov models with state-dependent routing (overloaded and in heavy traffic conditions) are considered in the

book [15]. In the case, when calls arrive and are served one at a time without transformation, equations (3.22), (3.24) are in agreement with results [39], where state-dependent networks $(M_\xi/M_\xi/1)^K$ in heavy traffic conditions are studied.

4. Diffusion approximation in non-Markov queueing models

We consider now a fluid and a diffusion approximation for some non-Markov models considered in section 2.5.3. The method of analysis consists of several stages. First, we need to represent a queueing process as an equivalent SP (to choose in the appropriate way switching times and construct corresponding processes on switching intervals). As in general, we have a truncation by the level zero, these processes in most cases are not so simple. Then on the next stage we construct an auxiliary SP which is asymptotically equivalent to the queueing process, and elementary processes on switching intervals are constructed without truncation. On the last stage we prove AP and DA for the auxiliary SP using limit theorems for SP (see appendix).

4.1. A network $(M_{SM,Q}/M_{SM,Q}/1/\infty)^r$

Consider a queueing network $(M_{SM,Q}/M_{SM,Q}/1/\infty)^r$ described in section 2.5.3. Suppose that characteristics of the network depend on parameter n in the following way. SMP $x(t)$ and variables introduced there do not depend on n. But if at time t, $x(t) = x$, $n^{-1}\overline{Q}_n(t) = \overline{q}$, then the local arrival and service rates and transition probabilities as well as random sizes of batches $\eta(x,\overline{q})$, $\kappa_i(x,\overline{q})$ depend on the pair $(x,\overline{q})$. We keep all notation given in section 2.5.3. Denote as before by $t_1 < t_2 < \cdots$ the times of sequential jumps of $x(t)$. Suppose that the imbedded MP $x_k = x(t_k)$, $k \geqslant 0$, is ergodic with stationary distribution π_x, $x \in X = \{1, 2, \ldots, d\}$. Let $Q_n^{(i)}(t)$ be the total amount of work (queue) at node i at time t, and $\overline{Q}_{n0}$ be the initial value. We put $\overline{Q}_n(t) = (Q_n^{(1)}(t), \ldots, Q_n^{(r)}(t))$, $t \geqslant 0$, and denote for any $x \in X$, $i = \overline{1,r}$, $\overline{q} \in \mathcal{R}^r$,

$$
\begin{aligned}
&m(x) = \mathbf{E}\tau(x), \qquad P_0\big(x,\overline{q}\big) = \big\|\, p_{ij}\big(x,\overline{q}\big)\big\|_{i,j=\overline{1,r}}, \qquad \overline{a}\big(x,\overline{q}\big) = \mathbf{E}\overline{\eta}\big(x,\overline{q}\big),\\
&g_i\big(x,\overline{q}\big) = \mathbf{E}\kappa_i\big(x,\overline{q}\big), \qquad \overline{g}\big(x,\overline{q}\big) = \big(\mu_1\big(x,\overline{q}\big)g_1\big(x,\overline{q}\big), \ldots, \mu_r\big(x,\overline{q}\big)g_r\big(x,\overline{q}\big)\big),\\
&m = \sum_{x\in X} m(x)\pi_x, \qquad \overline{c}\big(x,\overline{q}\big) = \lambda\big(x,\overline{q}\big)\overline{a}\big(x,\overline{q}\big) + \big(P_0\big(x,\overline{q}\big)^* - I\big)\overline{g}\big(x,\overline{q}\big),\\
&\overline{b}\big(\overline{q}\big) = \sum_{x\in X} m(x)\overline{c}\big(x,\overline{q}\big)\pi_x, \qquad d^2(x) = \mathbf{Var}\tau(x), \qquad d_i^2\big(x,\overline{q}\big) = \mathbf{E}\kappa_i^2\big(x,\overline{q}\big),\\
&J^2\big(x,\overline{q}\big) = \lambda\big(x,\overline{q}\big)\mathbf{E}\eta\big(x,\overline{q}\big)\eta\big(x,\overline{q}\big)^*.
\end{aligned}
$$

Let $F^2(x,\overline{q}) = \|\, f_{ij}(x,\overline{q})\|_{i,j=\overline{1,r}}$ be the matrix with the following elements:

$$
\begin{aligned}
&f_{ij}\big(x,\overline{q}\big) = -\mu_i\big(x,\overline{q}\big)p_{ij}\big(x,\overline{q}\big)d_i^2\big(x,\overline{q}\big) - \mu_j\big(x,\overline{q}\big)p_{ji}\big(x,\overline{q}\big)d_j^2\big(x,\overline{q}\big), \quad i,j = \overline{1,r},\ i \neq j;\\
&f_{ii}\big(x,\overline{q}\big) = \mu_i\big(x,\overline{q}\big)\big(1 - 2p_{ii}\big(x,\overline{q}\big)\big)d_i^2\big(x,\overline{q}\big) + \sum_{k=1}^{r} \mu_k\big(x,\overline{q}\big)p_{ki}\big(x,\overline{q}\big)d_k^2\big(x,\overline{q}\big).
\end{aligned}
$$

Denote

$$D^2(x,\overline{q}) = d^2(x)\big(\overline{c}(x,\overline{q}) - m^{-1}\overline{b}(\overline{q})\big)\big(\overline{c}(x,\overline{q}) - m^{-1}\overline{b}(\overline{q})\big)^* + m(x)\big(F^2(x,\overline{q}) + J^2(x,\overline{q})\big), \tag{4.1}$$

$$D^2(\overline{q}) = \sum_{x\in X} D^2(x,\overline{q})\pi_x, \qquad \overline{\gamma}(x,\overline{q}) = m(x)\big(\overline{c}(x,\overline{q}) - m^{-1}\overline{b}(\overline{q})\big).$$

Let the matrix $B^2(\overline{q})$ be calculated by variables $\overline{\gamma}(x,\overline{q})$ with the help of MP x_k according to (B.9), (B.10). We put $H^2(\overline{q}) = D^2(\overline{q}) + B^2(\overline{q})$. Define $H(\overline{q})$ according to the relation $H(\overline{q})H(\overline{q})^* = H^2(\overline{q})$. Let $\overline{s}(t)$ be a solution of the equation

$$\mathrm{d}\overline{s}(t) = m^{-1}\overline{b}\big(\overline{s}(t)\big)\,\mathrm{d}t, \quad \overline{s}(0) = \overline{s}_0. \tag{4.2}$$

Theorem 4.1. (1) Suppose that functions $\lambda(x,\overline{q})$, $\mu_i(x,\overline{q})$, $\overline{a}(x,\overline{q})$, $g_i(x,\overline{q})$, $p_{ij}(x,\overline{q})$ for any $x \in X$, $i = \overline{1,r}$, $j = \overline{1,r+1}$, are locally Lipschitz with respect to $\overline{q} \in int\{\mathcal{R}_+^m\}$, $\mathbf{E}\tau(x)^2 < \infty$, $x \in X$. Let also $m > 0$, for any bounded and closed domain $G \in int\{\mathcal{R}_+^m\}$

$$\mathbf{E}\kappa_i\big(x,\overline{q}\big)^2 \leqslant C_G, \qquad \mathbf{E}\big|\eta\big(x,\overline{q}\big)\big|^2 \leqslant C_G, \quad i = \overline{1,r},\ x \in X,\ \overline{q} \in G, \tag{4.3}$$

where $C_G < \infty$, the function $\overline{b}(\overline{q})$ has no more than linear growth, $n^{-1}\overline{Q}_n(0) \xrightarrow{\mathrm{P}} \overline{s}_0 > \overline{0}$, and there exists $T > 0$ such that $\overline{s}(t) > \overline{0}$, $t \in [0,T]$, in each component.

Then relation (3.23) holds with $\overline{s}(t)$ defined in (4.2).

(2) Suppose in addition that there exists a continuous matrix derivative $G(\overline{q}) = \overline{b}'(\overline{q})$, $\overline{q} \in int\{\mathcal{R}_+^m\}$, $\mathbf{E}\tau(x)^3 < \infty$, $x \in X$, and for any bounded and closed domain $G \in int\{\mathcal{R}_+^m\}$

$$\mathbf{E}\kappa_i\big(x,\overline{q}\big)^3 \leqslant C_G, \qquad \mathbf{E}\big|\eta\big(x,\overline{q}\big)\big|^3 \leqslant C_G, \quad i = \overline{1,r},\ x \in X,\ \overline{q} \in G. \tag{4.4}$$

Let also $n^{-1/2}(\overline{Q}_n(0) - n\overline{s}(0)) \overset{\mathrm{w}}{\Rightarrow} \overline{\gamma}_0$, and the function $H^2(\overline{q})$ is continuous.

Then the sequence $\overline{\gamma}_n(t) = n^{-1/2}(\overline{Q}_n(nt) - n\overline{s}(t))$ weakly converges in D_T^r to the diffusion process $\overline{\gamma}(t)$:

$$\mathrm{d}\overline{\gamma}(t) = G\big(\overline{s}(t)\big)\overline{\gamma}(t)\,\mathrm{d}t + m^{-1/2}H\big(\overline{s}(t)\big)\,\mathrm{d}\overline{w}(t), \qquad \overline{\gamma}(0) = \overline{\gamma}_0. \tag{4.5}$$

Proof. First, we consider an auxiliary queueing network $\widetilde{QN}$ switched by SMP $x(t)$. The network is described with the help of the families of functions and random variables $\lambda(x,\overline{q})$, $\mu_i(x,\overline{q})$, $\overline{\eta}(x,\overline{q})$, $\kappa_i(x,\overline{q})$, $p_{ij}(x,\overline{q})$, $x \in X$, $i = \overline{1,r}$, $j = \overline{1,r+1}$, introduced in section 2.5.3, in the following way: on each interval $[t_k, t_{k+1})$ the rates $\lambda(\cdot)$, $\mu(\cdot)$, probabilities $p_{ij}(\cdot)$ and variables $\overline{\eta}(\cdot)$, $\kappa_i(\cdot)$ depend only on the values $x(t_k) = x$, $\overline{q} = n^{-1}\overline{Q}_n(t_k)$, at the initial point t_k. That is, at given values $x(t_k) = x$, $n^{-1}\overline{Q}_n(t_k) = \overline{q}$, parameters of the network on the interval $[t_k, t_{k+1})$ do not depend on the changes of the current size of the queue. This network is a bit simpler, but we prove that asymptotically it is equivalent to the initial network.

Let us construct a corresponding PSMS. Denote by $\Pi_a(t,\xi)$ a compound Poisson process with parameter a and a size of a jump ξ (sizes of different jumps are independent

random variables). Denote by $\overline{\Pi}_a(t, \overline{\xi})$ a vector-valued compound Poisson process with a size of a jump $\overline{\xi}$. Put

$$\widetilde{\zeta}\big(t, x, \overline{q}\big) = \overline{\Pi}_{\lambda(x,\overline{q})}\big(t, \overline{\eta}\big(x, \overline{q}\big)\big) + \sum_{i,j=1}^{r} \Pi_{\mu_i(x,\overline{q})p_{ij}(x,\overline{q})}\big(t, \kappa_i\big(x, \overline{q}\big)\big)\big(\overline{e}_j - \overline{e}_i\big) - \sum_{i=1}^{r} \Pi_{\mu_i(x,\overline{q})p_{i,r+1}(x,\overline{q})}\big(t, \kappa_i\big(x, \overline{q}\big)\big)\overline{e}_i, \quad t \geqslant 0, \tag{4.6}$$

(here all Poisson processes are independent). We introduce a family of independent at different k processes $\widetilde{\zeta}_{nk}(t, x, n\overline{q})$, $k > 0$, such that their distributions coincide with the distribution of $\widetilde{\zeta}(t, x, \overline{q})$. Denote by $\{(x(t), \widetilde{Q}_n(t)); t \geqslant 0\}$ an auxiliary PSMS, which is constructed with the help of SMP $x(t)$ and processes $\widetilde{\zeta}_{nk}(t, x, n\overline{q})$ according to relations (2.9). By analogy to the proof of theorem 3.1, we can define $\widetilde{Q}_n(t)$ in the whole space $\mathcal{R}^r$. Now we prove AP for $\widetilde{Q}_n(t)$. Let us check the conditions of theorem B.1. In our notation the distribution of $\xi_n(x, n\overline{q})$ coincides with the one of $\widetilde{\zeta}(\tau(x), x, \overline{q})$. Conditions (B.2), (B.4) are automatically satisfied. Furthermore, for any random variables $\tau > 0$ and ξ with the properties $\mathbf{E}\tau^2 < \infty$, $\mathbf{E}|\xi|^2 < \infty$, we can calculate that $\mathbf{E}\Pi_a^2(\tau, \xi) \leqslant a\mathbf{E}\tau\mathbf{E}|\xi|^2 + a^2(\mathbf{E}\xi)^2\mathbf{E}\tau^2$. Using Chebyshev's inequality we get

$$n\mathbf{P}\Big(n^{-1}\sup_{t\leqslant\tau}\big|\Pi_a(t, \xi)\big| > \varepsilon\Big) \leqslant n\mathbf{P}\big(\Pi_a\big(\tau, |\xi|\big) > n\varepsilon\big) \leqslant n(n\varepsilon)^{-2}\mathbf{E}\Pi_a^2\big(\tau, |\xi|\big) \to 0,$$

for any $\varepsilon > 0$ as $n \to \infty$. This implies condition (B.3). As is easy to calculate, $\mathbf{E}\widetilde{\zeta}(\tau(x), x, \overline{q}) = \overline{c}(x, \overline{q})m(x)$. Using theorem B.1 we get that $\widetilde{Q}_n(t)$ satisfies relation (3.23) with $\overline{s}(t)$ defined in (4.2). Now, following the same lines as in the proof of theorem 3.1, we get that the multidimensional process generated by the queue in the system $\widetilde{QN}$ also satisfies relation (3.23).

Now we return to the initial network. First, introduce independent families of multidimensional MP $\{\overline{\gamma}_{nk}(t, x, n\overline{q}), t \geqslant 0, x \in X, \overline{q} \in R_+^r\}$, $k \geqslant 0$, with values in R_+^r in the same way as it was done in section 2.5.3. Put $\overline{\gamma}_{nk}(0, x, n\overline{q}) = n\overline{q}$. If $\overline{\gamma}_{nk}(t, x, n\overline{q}) = n\overline{s}$, then with the local rate $\Lambda(x, \overline{s}) = \lambda(x, \overline{s}) + \sum_{i=1}^{r}\mu_i(x, \overline{s})$ the process can make a jump of the size $\overline{\delta}(x, \overline{s})$. Here $\overline{\delta}(x, \overline{s})$ is defined in (2.11), where we take $\kappa_i(x, \overline{s})$ instead of $\widetilde{\kappa}_i(x, \overline{s})$. Denote $\overline{\zeta}_{nk}(t, x, n\overline{q}) = \overline{\gamma}_{nk}(t, x, n\overline{q}) - n\overline{q}$. Let $\widetilde{\widetilde{Q}}_n(t)$ be an auxiliary PSMS defined with the help of $x(t)$ and processes $\overline{\zeta}_{nk}(t, x, n\overline{q})$ according to relations (2.9). Again we can define it in the whole space $\mathcal{R}^r$. Note that by construction the trajectory of $\widetilde{\widetilde{Q}}_n(t)$ coincides with the trajectory of the queue $\overline{Q}_n(t)$ on any interval $[0, T]$ such that $\widetilde{\widetilde{Q}}_n(t) > 0$, $t \in [0, T]$.

Let us prove that $\widetilde{\widetilde{Q}}_n(t)$ also satisfies (3.23) with $\overline{s}(t)$ defined in (4.2). Again we need to check conditions of theorem B.1. Now we have that $\xi_n(x, n\overline{q}) = \overline{\zeta}_{n1}(\tau(x), x, n\overline{q})$. Let us follow the same steps as in [7, proof of theorem 1 and lemmas 1, 2]. Using condition (4.3) we can prove for any $\overline{q} \in \mathcal{R}^r$ that $\mathbf{E}\xi_n(x, n\overline{q}) \to \mathbf{E}\widetilde{\zeta}_1(\tau(x), x, \overline{q})$ (see (4.6)) and check other conditions of theorem B.1. This implies (3.23) for $\widetilde{\widetilde{Q}}_n(t)$. Now, by analogy to proof of theorem 3.1, we get that the asymptotic

behavior of the queueing process $\overline{Q}_n(t)$ and PSMS $\widetilde{\overline{Q}}_n(t)$ is the same, and the 1st part of theorem 4.1 is proved.

To prove DA we use theorem B.3. We get, in the same way, that it is enough to calculate the characteristics of the auxiliary PSMS $\widetilde{\overline{Q}}_n(t)$ defined above. To find the function $D_n^2(x,\alpha)$ (see (B.7)) we can use again [7, theorem 1 and lemmas 1, 2]. Using condition (4.4) we can prove that for any $\overline{q} \in \mathcal{R}^r$ as $n \to \infty$

$$\mathbf{E}\xi_n\big(x, n\overline{q}\big)\xi_n\big(x, n\overline{q}\big)^* \to \mathbf{E}\widetilde{\zeta}\big(\tau(x), x, \overline{q}\big)\widetilde{\zeta}\big(\tau(x), x, \overline{q}\big)^*.$$

Now according to (B.7) we need to calculate $D^2(x,\overline{q})$, where in notation of theorem 4.1 $\rho_{n1}(\cdot) = \widetilde{\zeta}(\tau(x), x, \overline{q}) - m(x)\overline{c}(x,\overline{q}) - m^{-1}\overline{b}(\overline{q})(\tau(x) - m(x))$ (see (4.6)). It is not difficult to calculate that

$$\begin{aligned} D^2\big(x,\overline{q}\big) = {} & m(x)\lambda\big(x,\overline{q}\big)\mathbf{E}\eta(x,q)\eta(x,q)^* \\ & + \sum_{i,j=1}^{r} \mu_i\big(x,\overline{q}\big)p_{ij}\big(x,\overline{q}\big)d_i^2\big(x,\overline{q}\big)\big(\overline{e}_j - \overline{e}_i\big)\big(\overline{e}_j - \overline{e}_i\big)^* \\ & + \sum_{i=1}^{r} \mu_i\big(x,\overline{q}\big)p_{i,r+1}d_i^2\big(x,\overline{q}\big)\overline{e}_i\overline{e}_i^*, \end{aligned}$$

and after some algebra we get the expression for $D^2(x,\overline{q})$ in the form (4.1). All other conditions of theorem B.3 are also satisfied. □

In particular, if $\lambda(x,\overline{q}) \equiv 0$, $p_{i,r+1}(x,\overline{q}) \equiv 0$ for all $i = \overline{1,r}$, $x \in X$, $\overline{q}$, then the network is closed.

Example 4.2. Consider a state-dependent system $M_{SM,Q}/M_{SM,Q}/1/\infty$ with semi-Markov switches. Let $\{x(t); t \geqslant 0\}$ be an SMP with state space $X = \{1, 2, \ldots, d\}$, $\tau(x)$ be the sojourn time in state $x \in X$. Suppose that the imbedded Markov chain is ergodic and denote by π_x, $x = 1, \ldots, r$, it is stationary distribution. Let also nonnegative functions $\{\lambda(x,q), \mu(x,q), x \in X, q \geqslant 0\}$, be given. Suppose that calls arrive and are served one at a time. Denote by $Q(t)$ the total number of calls in the system at time t. Assume that as $x(t) = x$, $n^{-1}Q(t) = q$, the arrival rate is $\lambda(x,q)$ and the service rate is $\mu(x,q)$. That is, we have a semi-Markov arrival process and a semi-Markov service. After service completion a call leaves the system. Denote

$$\begin{aligned} & m(x) = \mathbf{E}\tau(x), \qquad m = \sum_{x\in X} m(x)\pi_x, \qquad b(q) = \sum_{x\in X} m(x)\big(\lambda(x,q) - \mu(x,q)\big)\pi_x, \\ & d^2(x) = \mathbf{Var}\tau(x), \qquad \gamma(x,q) = m(x)\big(\lambda(x,q) - \mu(x,q) - m^{-1}b(q)\big), \\ & D^2(q) = \sum_{x\in X}\big[d^2(x)\big(\lambda(x,q) - \mu(x,q) - m^{-1}b(q)\big)^2 + m(x)\big(\lambda(x,q) + \mu(x,q)\big)\big]\pi_x. \end{aligned}$$

Let the function $B^2(q)$ be calculated by variables $\gamma(x,q)$ with the help of MP x_k according to relations (B.9), (B.10). We put $H^2(q) = D^2(q) + B^2(q)$. Denote by $s(t)$ a solution of the equation

$$\mathrm{d}s(t) = m^{-1}b\big(s(t)\big)\,\mathrm{d}t, \qquad s(0) = s_0. \tag{4.7}$$

Corollary 4.3. Suppose that for any $x \in X$ the functions $\lambda(x,q)$, $\mu(x,q)$ are locally Lipschitz with respect to $q > 0$, $m > 0$, $\mathbf{E}\tau(x)^2 < \infty$, $b(q)$ has no more than linear growth, $n^{-1}Q_n(0) \xrightarrow{\mathrm{P}} s_0 > 0$, and there exists $T > 0$ such that $s(t) > 0$, $t \in [0,T]$. Then relation (3.10) holds with $s(t)$ defined in (4.7).

If in addition there exists a continuous derivative $g(q) = b'(q)$, $q > 0$, $\mathbf{E}\tau(x)^3 < \infty$, $x \in X$, $n^{-1/2}(Q_n(0) - ns(0)) \overset{\mathrm{w}}{\Rightarrow} \gamma_0$, and the function $H^2(q)$, $q > 0$, is continuous, then the sequence $\gamma_n(t) = n^{-1/2}(Q_n(nt) - ns(t))$ weakly converges on $\mathcal{D}_T$ to the diffusion process $\gamma(t)$:

$$\mathrm{d}\gamma(t) = g\big(s(t)\big)\gamma(t)\,\mathrm{d}t + m^{-1/2}H\big(s(t)\big)\,\mathrm{d}w(t), \qquad \gamma(0) = \gamma_0.$$

In particular, when $\lambda(x,q) \equiv \lambda(x)$, $\mu(x,q) \equiv q\mu(x)$, our system is equivalent to a system $M_{SM}/M_{SM}/\infty$ in a semi-Markov environment. If we denote

$$\lambda = m^{-1}\sum_{x\in X} m(x)\lambda(x)\pi_x, \qquad \mu = m^{-1}\sum_{x\in X} m(x)\mu(x)\pi_x,$$

then (4.7) has the form $\mathrm{d}s(t) = (\lambda - \mu s(t))\,\mathrm{d}t$, which coincides with the equation for the system $M/M/\infty$ (see section 3.1.1, case 2).

Remark 4.4. In the same way it is possible to study systems and networks of the type $SM/M_{SM,Q}/1/\infty$ and $(SM/M_{SM,Q}/1/\infty)^r$. Here the calls arrive at times of jumps of some SMP $x(t)$, and the rate of service may depend on $x(t)$ and the current number of calls (or the amount of work) in the system. Note that the diffusion approximation of the system $GI/M/1/\infty$ was considered in [47].

4.2. A system with unreliable servers

To illustrate the wide possibilities of the approach suggested we consider a system $GI/M_Q/r/\infty$ with unreliable servers. Calls enter the system one at a time and interarrival times are i.i.d.r.v. τ_k, $k \geqslant 1$. The system consists of r identical servers subject to random failures and an infinite number of waiting places. Suppose that rates $\{\mu_i(q),\ q > 0,\ \nu_i,\ i = \overline{1,r},\ \kappa_i,\ i = \overline{0,r-1}\}$ be given. Denote by $Q_n(t)$, $t \geqslant 0$, a number of calls in the system at time t. Assume that the service rate depends on the queue size in the following way: if a call enters the system at time t_k and $Q_n(t_k+0) = Q$, then the service rate on the interval (t_k, t_{k+1}) for each operating server is $\mu(n^{-1}Q)$.

Let $y(t)$ be a number of operating (not failed) servers at time t. If $y(t) = i$, then each operating server with rate ν_i may fail. If there is a call on service, then this call goes back to the queue. Each failed server with rate κ_i may be repaired. After repair

a server immediately takes a call for service if there are waiting calls. By construction, the process $y(t)$ is a birth-and-death process with state space $\{0, 1, \dots, r\}$ and rate of birth $(r-i)\kappa_i$ and death $i\nu_i$, respectively. It does not depend on n. Assume that $\nu_i > 0$, $i = \overline{1, r}$, $\kappa_i > 0$, $i = \overline{0, r-1}$. Denote by ρ_i, $i = \overline{0, r}$, a stationary distribution of $y(t)$. Put $m = \mathbf{E}\tau_1$, $\hat{\rho} = \sum_{i=1}^{r} i\rho_i$.

Proposition 4.5. Suppose that the function $\mu(q)$, $q > 0$, is locally Lipschitz, $m > 0$, $n^{-1}Q_n(0) \xrightarrow{\mathrm{P}} s_0 > 0$, a unique solution of the equation

$$\mathrm{d}s(t) = \big(m^{-1} - \hat{\rho}\mu\big(s(t)\big)\big)\,\mathrm{d}t, \qquad s(0) = s_0, \tag{4.8}$$

exists on some interval $[0, T]$, and $s(t) > 0$, $t \in [0, T]$. Then (3.10) holds.

Proof. Denote by $\{y_i(t);\ t \geqslant 0\}$ a birth-and-death process $y(t)$ with initial state i. Let $\{\Pi_i(t, y_i(\cdot), q);\ t \geqslant 0\}$ be a Poisson process modulated by $y_i(t)$ in the following way: $\Pi_i(0, y_i(\cdot), q) = 0$, and if $y(t) = j$, then the local rate of a jump is $j\mu(q)$. Denote by $\{x(t);\ t \geqslant 0\}$ an imbedded SMP with state space $\{0, 1, \dots, r\}$ which is constructed using $y(t)$ as follows: times of jumps are chosen as arrival times of calls t_k, $k \geqslant 0$. If $y(t_k + 0) = i$, then we put $x(t_k + 0) = i$. Sojourn times in any state have the same distribution as the variable τ_1, and transition probabilities p_{ij} of the imbedded Markov chain $x_k = x(t_k + 0)$ are calculated in the following way:

$$p_{ij} = \mathbf{P}\big(y(\tau_1 + 0) = j \mid y(0) = i\big), \quad i, j = \overline{0, r}.$$

Let us introduce the family of jointly independent at different $k \geqslant 0$ processes $\zeta_{nk}(t, i, Q)$, having the same distributions as the process $1 - \Pi_i(t, y_i(\cdot), n^{-1}Q)$, $t \geqslant 0$, $i = \overline{0, r}$, $Q > 0$.

Now let $\{(x(t), \widetilde{Q}_n(t)); t \geqslant 0\}$ be an auxiliary RPSM which is constructed with the help of $x(t)$ and processes $\{\zeta_{nk}(t, i, Q);\ t \geqslant 0\}$, $k > 0$, according to (2.9). By construction a trajectory of the queue $Q_n(t)$ coincides with $\widetilde{Q}_n(t)$ in the domain $\widetilde{Q}_n(t) > 0$, $t \in [0, T]$. Then, according to previous arguments, it is enough to prove the AP for $\widetilde{Q}_n(t)$. We use theorem B.1. In our case $\xi_{n1}(i, nq) = 1 - \Pi_i(\tau_1, y_i(\cdot), q)$. It it easy to calculate, that the stationary distribution of the imbedded MP $x_k = y(t_k)$ is also ρ_i, $i = \overline{0, r}$, and for any $t > 0$,

$$\sum_{i=0}^{r} \mathbf{E}\Pi_i\big(t, y_i(\cdot), q\big)\rho_i = t\hat{\rho}\mu(q).$$

Therefore, $\sum_{i=0}^{r} \rho_i \mathbf{E}\xi_{n1}(i, nq) = 1 - m\hat{\rho}\mu(q)$. All other conditions of theorem B.1 are satisfied, and equation (B.6) has the form (4.8). □

Remark 4.6. If a service rate depends on the number of operating devices (equal to $\mu_i(q)$ when $y(t) = i$), then (3.10) also holds, but in (4.8) we have to write $\hat{\mu}(s(t)) = \sum_{i=1}^{r} \rho_i i\mu_i(s(t))$ instead of $\hat{\rho}\mu(s(t))$.

Remark 4.7. Using theorem B.3 we can prove DA for $Q_n(t)$ also.

Remark 4.8. Results of proposition 4.5 can be extended to systems $G_Q/M_Q/r/\infty$ with interarrival times and rates ν_i, κ_i depending also on the current size of the queue $n^{-1}Q_n(t)$. In this case it is not possible to construct an auxiliary SMP, which stands for the external environment, because sojourn times and transition probabilities may depend on the queue. But it is possible to use a general representation for the queue in terms of SP and use AP for so-called quasi-ergodic MP [3].

4.3. Polling systems

Consider a polling system defined in section 2.5.5. The system consists of r stations and a single moving server. Suppose that service rates depend on the normalized size of the queue in the following way: if upon arrival to station j a server sees Q_j calls waiting there, then the service rate on the time interval of the length $\kappa_k(j)$ is $\mu_j(n^{-1}Q_j)$. Denote by $Q_n(i,t)$ a number of calls at station i at time t, $\overline{Q}_n(t) = (Q_n(1,t), \dots, Q_n(r,t))$. We keep all notation of section 2.5.5. Suppose that an MP with transition probabilities p_{ij}, $i, j = \overline{1,r}$, is ergodic with stationary distribution π_i.

Proposition 4.9. Assume that for any $i = \overline{1,r}$ the values $m_i = \mathbf{E}\kappa_1(i)$, $\widetilde{m}_i = \mathbf{E}\widetilde{\kappa}_1(i)$ exist, functions $\mu_i(q)$, $q > 0$, satisfy local Lipschitz condition, $\overline{Q}_n(0) \xrightarrow{\mathrm{P}} \overline{s}_0 = (s_{01}, \dots, s_{0r})$, on some interval $[0,T]$ at each $i = \overline{1,r}$ a unique solution of the equation

$$\mathrm{d}s_i(t) = \big(\lambda_i - \hat{\rho}_i \mu_i\big(s_i(t)\big)\big)\,\mathrm{d}t, \qquad s_i(0) = s_{0i}, \tag{4.9}$$

exists, and $s_i(t) > 0$, $t \in [0,T]$, where $\hat{\rho}_i = \pi_i m_i (\sum_{j=1}^r \pi_j (m_j + \widetilde{m}_j))^{-1}$.

Then relation (3.23) holds with $\overline{s}(t) = (s_i(t),\ i = \overline{1,r})$.

Proof. First, we construct an auxiliary PSMS. Let $\Pi_k(t,i,\lambda_i)$ and $\widetilde{\Pi}_k(t,i,\mu_i)$ be independent at different k, i Poisson processes with parameters λ_i and μ_i, respectively. We introduce processes $\overline{\zeta}_{nk}(t,i,\overline{Q}) = (\zeta_{nk}^{(j)}(t,i,Q_j),\ j = \overline{1,r})$ as follows:

$$\zeta_{nk}^{(i)}(t,i,Q_i) = \Pi_k(t,i,\lambda_i) - \widetilde{\Pi}_k\big(t,i,\mu_i\big(n^{-1}Q_i\big)\big), \quad \text{as } 0 \leqslant t \leqslant \kappa_k(i);$$
$$\zeta_{nk}^{(i)}(t,i,Q_i) = \Pi_k(t,i,\lambda_i) - \widetilde{\Pi}_k\big(\kappa_k(i),i,\mu_i\big(n^{-1}Q_i\big)\big), \quad \text{as } \kappa_k(i) < t \leqslant \kappa_k(i) + \widetilde{\kappa}_k(i);$$
$$\zeta_{nk}^{(j)}(t,i,Q_j) = \Pi_k(t,j,\lambda_j), \quad \text{as } 0 \leqslant t \leqslant \kappa_k(i) + \widetilde{\kappa}_k(i),\ j = \overline{1,r},\ j \neq i.$$

Denote by $\{(x(t), \widetilde{Q}_n(t));\, t \geqslant 0\}$ an auxiliary PSMS constructed according to (2.9) by introduced processes and SMP $x(t)$, introduced in section 2.5.5. By construction, if on some interval $[0,T]$ $\widetilde{Q}_n(t) > 0$ in each component, then the trajectory of $\widetilde{Q}_n(t)$ coincides with the trajectory of the queue $\overline{Q}_n(t)$ on $[0,T]$. Thus, it is enough to prove AP for $\widetilde{Q}_n(t)$.

Let us use theorem B.1. In this case $\overline{\xi}_n(i, \overline{Q}) = \overline{\zeta}_{n1}(\kappa_1(i) + \widetilde{\kappa}_1(i), i, \overline{Q})$, $\tau_n(i) = \kappa_1(i) + \widetilde{\kappa}_1(i)$. It is not difficult to check all conditions of theorem B.1 and calculate that the function $m^{-1}b(\overline{q})$ in (B.6) has the form $(\lambda_i - \hat{\rho}_i\mu_i(q_i),\ i = \overline{1,r})$. □

In particular, if $\mu_i(q) = \alpha_i + \mu_i q$ and $\hat{\rho}_i\alpha_i < \lambda_i, i = \overline{1,r}$, then relation (3.23) holds for any $T > 0$, and equation (4.9) has a point of stability $\overline{s}^* = ((\lambda_i - \hat{\rho}_i\alpha_i)/\mu_i,\ i = \overline{1,r})$.

Remark 4.10. Using theorem B.3 we can also prove that the sequence $\overline{\gamma}_n(t) = n^{-1/2} \times (\overline{Q}_n(nt) - n\overline{s}(t))$ weakly converges in $\mathcal{D}_T$ to the diffusion process $\overline{\gamma}(t)$ satisfying the equation:

$$\mathrm{d}\overline{\gamma}(t) = G\big(\overline{s}(t)\big)\overline{\gamma}(t)\,\mathrm{d}t + m^{-1/2}H\big(\overline{s}(t)\big)\,\mathrm{d}\overline{w}(t), \qquad \overline{\gamma}(0) = \overline{\gamma}_0,$$

where $G(\overline{q})$ is a diagonal matrix with elements $-\hat{\rho}_i\mu_i'(q_i)$, and matrix $H^2(\overline{q})$ is calculated by vectors $\overline{\xi}_n(i, \overline{Q})$ and the imbedded MP $x(t_k + 0)$ according to relations (B.7), (B.9), (B.10).

Remark 4.11. Using the same approach, some other examples of queueing systems $G_Q/M_Q/1/\infty$, $SM_Q/M_Q/1/\infty$ and networks $(G_Q/M_Q/1/\infty)^r$ are considered in [5,6].

Another possible direction of applications is the class of so-called retrial queues [25]. Using suggested approach, AP- and DA-type theorems for some classes of overloaded retrial queueing models $\overline{M}/\overline{G}/\overline{1}/w.r.$, $M/M/m/w.r.$ and $M_Q/G/1/w.r.$ with state-dependent Markov arrival process, general or exponential service and asymptotically small rate of retrial calls are studied in [8,9,11].

4.4. Conclusion

A new approach to study fluid and diffusion approximation type theorems (without reflection on the boundary) in transient and quasi-stationary regimes for queueing processes in overloaded state-dependent systems and networks of a switching structure is suggested. The approach is based on functional limit theorems of averaging principle and diffusion approximation types for the so-called switching processes.

This approach gives us the possibility to provide an asymptotic analysis of wide classes of Markov and non-Markov models in a convenient standard way. The analysis of the initial model is reduced to the analysis of some auxiliary switching process, which is asymptotically equivalent to the queueing process and usually has more simple structure (corresponding processes on switching intervals are constructed without truncation by level zero). Then the coefficients of the equations in fluid and diffusion limits are calculated using the first and second moment functions of the increments of corresponding processes on switching intervals.

For Markov systems, we basically calculate an increment of a queueing process on the exponential interval between two sequential changes in the system. For more complicated systems possibly in a random environment, we calculate increments on the intervals between changes of the environment or some auxiliary switching component.

From the practical point of view, it is much simpler to calculate (or estimate) these characteristics rather than simulate the whole system on a large interval of time.

The wide possibilities of the suggested approach are illustrated for various classes of state-dependent Markov, semi-Markov and more general non-Markov open and closed queueing systems and networks.

Acknowledgements

The author is thankful to the associate editor and the referee for their valuable comments which helped to improve the presentation of the results.

Appendix

Here we study limit theorems for SP's in the case of fast switches. Consider a sequence of SP's $\{(x_n(t), \zeta_n(t));\ t \geqslant 0\}$ on the interval $[0, nT]$, $n \to \infty$. Suppose that SP depends on the scaling parameter n so that the number of switches on each interval $[na, nb]$, $0 < a < b < T$, tends in probability to infinity. Then, under some natural assumptions, a normalized trajectory of $\zeta_n(nt)$ uniformly converges in probability to some deterministic function which is a solution of some differential equation (Averaging Principle – AP), and a normalized difference between trajectory and this solution weakly converges in Skorokhod space $\mathcal{D}_T$ to some diffusion process (Diffusion Approximation – DA). As sample trajectories of a limiting process are continuous, this convergence implies weak convergence of functionals, which are continuous with respect to the uniform convergence [23,45].

Appendix A. Averaging principle and diffusion approximation for RPSM

Consider first AP and DA type theorems for simple RPSM (see section 2.3). Let for each $n = 1, 2, \ldots$, $\mathcal{F}_{nk} = \{(\xi_{nk}(\alpha)), \tau_{nk}(\alpha)), \alpha \in \mathcal{R}^r\}$, $k \geqslant 0$, be jointly independent families of random variables with values in $\mathcal{R}^r \times [0, \infty)$. Suppose that their distributions do not depend on index k. Let S_{n0} be an independent of $\mathcal{F}_{nk}$, $k \geqslant 0$, initial value in $\mathcal{R}^r$. Put

$$\begin{aligned} &t_{n0} = 0, \qquad t_{nk+1} = t_{nk} + \tau_{nk}(S_{nk}), \qquad S_{nk+1} = S_{nk} + \xi_{nk}(S_{nk}), \quad k \geqslant 0, \\ &\qquad\qquad S_n(t) = S_{nk}, \quad \text{as } t_{nk} \leqslant t < t_{nk+1},\ t \geqslant 0. \end{aligned} \tag{A.1}$$

Let there exist functions $m_n(\alpha) = \mathbf{E}\tau_{n1}(n\alpha)$, $b_n(\alpha) = \mathbf{E}\xi_{n1}(n\alpha)$, $\alpha \in \mathcal{R}^r$.

Theorem A.1 (Averaging principle). Suppose that for any $N > 0$,

$$\lim_{L\to\infty} \limsup_{n\to\infty} \sup_{|\alpha|\leqslant N} \left\{\mathbf{E}\tau_{n1}(n\alpha)\chi\left(\tau_{n1}(n\alpha) > L\right) + \mathbf{E}\left|\xi_{n1}(n\alpha)\right|\chi\left(\left|\xi_{n1}(n\alpha)\right| > L\right)\right\} = 0; \tag{A.2}$$

$$\left|m_n(\alpha_1) - m_n(\alpha_2)\right| + \left|b_n(\alpha_1) - b_n(\alpha_2)\right| \leqslant C_N|\alpha_1 - \alpha_2| + \alpha_n(N), \tag{A.3}$$

as $\max(|\alpha_1|, |\alpha_2|) \leqslant N$, where C_N are some bounded constants, $\alpha_n(N) \to 0$ uniformly in $|\alpha_1| \leqslant N$, $|\alpha_2| \leqslant N$, and there exist a deterministic value s_0 and functions $m(a) > 0$, $b(a)$ such that as $n \to \infty$, $n^{-1}S_{n0} \xrightarrow{\mathrm{P}} s_0$, and for any $\alpha \in \mathcal{R}^r$

$$m_n(\alpha) \to m(\alpha), \qquad b_n(\alpha) \to b(\alpha). \tag{A.4}$$

Let also there exist T such that $y(+\infty) > T$, where $y(t) = \int_0^t m(\eta(u))\,\mathrm{d}u$, and the function $\eta(u)$ satisfies the equation

$$\eta(0) = s_0, \qquad \mathrm{d}\eta(u) = b\big(\eta(u)\big)\,\mathrm{d}u, \tag{A.5}$$

a unique solution of which exists on each interval.

Then

$$\sup_{0\leqslant t\leqslant T} \big|n^{-1}S_n(nt) - s(t)\big| \xrightarrow{\mathrm{P}} 0, \tag{A.6}$$

where the function $s(t)$ satisfies the equation

$$s(0) = s_0, \qquad \mathrm{d}s(t) = \big(s(t)\big)^{-1} b\big(s(t)\big)\,\mathrm{d}t, \tag{A.7}$$

a unique solution of which exists.

Now we consider the process $\gamma_n(t) = n^{-1/2}(S_n(nt) - ns(t))$, $t \in [0, T]$. Denote

$$\begin{aligned}
&\widetilde{b}_n(\alpha) = m_n(\alpha)^{-1} b_n(\alpha), \qquad \widetilde{b}(\alpha) = m(\alpha)^{-1} b(\alpha),\\
&\rho_n(\alpha) = \xi_{n1}(n\alpha) - b_n(\alpha) - \widetilde{b}(\alpha)\big(\tau_{n1}(n\alpha) - m_n(\alpha)\big),\\
&q_n(\alpha, z) = \sqrt{n}\left(\widetilde{b}_n\left(\alpha + \frac{1}{\sqrt{n}}z\right) - \widetilde{b}(\alpha)\right), \qquad D_n^2(\alpha) = \mathbf{E}\rho_n(\alpha)\rho_n(\alpha)^*.
\end{aligned}$$

Theorem A.2 (Diffusion approximation). Let conditions (A.3), (A.4) be satisfied, where in (A.3) $\sqrt{n}\alpha_n(N) \to 0$, there exist continuous vector-valued function $q(\alpha, z)$ and matrix-valued function $D^2(\alpha)$ such that in any domain $|\alpha| \leqslant N$ $|q(\alpha, z)| \leqslant C_N(1 + |z|)$, and uniformly in $|\alpha| \leqslant N$ at each fixed z

$$\sqrt{n}\big(\widetilde{b}_n\big(\alpha + n^{-1/2}z\big) - \widetilde{b}(\alpha)\big) \to q(\alpha, z), \qquad D_n^2(\alpha) \to D^2(\alpha), \tag{A.8}$$

$\gamma_n(0) \xRightarrow{\mathrm{w}} \gamma_0$, and for any $N > 0$

$$\lim_{L\to\infty} \limsup_{n\to\infty} \sup_{|\alpha|<N} \big\{\mathbf{E}\tau_{n1}^2(n\alpha)\chi\big(\tau_{n1}(n\alpha) > L\big) + \mathbf{E}\big|\xi_{n1}(n\alpha)\big|^2\chi\big(\big|\xi_{n1}(n\alpha)\big| > L\big)\big\} = 0. \tag{A.9}$$

Then the sequence $\gamma_n(t)$ weakly converges in $\mathcal{D}_T$, where T is defined in theorem A.1, to the diffusion process $\gamma(t)$ satisfying the following stochastic differential equation, a unique solution of which exists:

$$\mathrm{d}\gamma(t) = q\big(s(t), \gamma(t)\big)\,\mathrm{d}t + D\big(s(t)\big)m\big(s(t)\big)^{-1/2}\,\mathrm{d}w(t), \qquad \gamma(0) = \gamma_0. \tag{A.10}$$

Here $s(\cdot)$ satisfies (A.7), $D(q)D(q)^* = D^2(q)$, and $w(t)$ is a standard Wiener process in $\mathcal{R}^r$.

Remark A.3. Let $G \in \mathcal{R}^r$ be a closed bounded domain such that $s(t) \in G$, $0 \leqslant t \leqslant T$. Denote by $G(\varepsilon)$ a closure of ε-neighborhood of G. Then theorems A.1, A.2 are true, if for some $\varepsilon > 0$ conditions (A.2)–(A.4) and (A.8), (A.9) hold uniformly in $\alpha \in G(\varepsilon)$.

Proof of theorems A.1, A.2. We give the proof in a shorten way. Details can be found in [5]. Let us introduce the sequences $\eta_{nk} = n^{-1}S_{nk}$, $y_{nk} = n^{-1}t_{nk}$, $k \geqslant 0$, and denote $\eta_n(u) = \eta_{nk}$, $y(u) = y_{nk}$, as $n^{-1}k \leqslant u < n^{-1}(k+1)$, $u \geqslant 0$. Put

$$\nu_n(t) = \min\{k\colon\ k > 0,\ t_{n,k+1} > nt\}, \qquad \mu_n(t) = \inf\bigl\{u\colon\ u > 0,\ y_n(u) > t\bigr\}.$$

As far as $S_n(nt) = S_{n\nu_n(t)}$, we have the representation $n^{-1}S_n(nt) = \eta_n(n^{-1}\nu_n(t)) = \eta_n(\mu_n(t) - 1/n)$. Thus, RPSM $n^{-1}S_n(nt)$ is constructed as a superposition of two processes: $\eta_n(t)$ and $\mu_n(t)$. First, we study the behavior of the processes $\eta_n(t)$ and $y_n(t)$, then $\mu_n(t)$ and their superposition. According to (A.1), $\eta_{nk+1} = \eta_{nk} + n^{-1}b_n(\eta_{nk}) + \varphi_{nk}$, $y_{nk+1} = y_{nk} + n^{-1}m_n(\eta_{nk}) + \psi_{nk}$, $k \geqslant 0$, where $\varphi_{nk} = n^{-1}(\xi_{nk}(n\eta_{nk}) - b_n(\eta_{nk}))$, $\psi_{nk} = n^{-1}(\tau_{nk}(n\eta_{nk}) - m_n(\eta_{nk}))$.

Sequences φ_{nk} and ψ_{nk}, $k \geqslant 0$, are martingale differences with respect to the sequence of σ-algebras generated by variables $\{\eta_{ni}, i \leqslant k\}$. Applying results of [27] we get

$$\sup_{0\leqslant u\leqslant t} \bigl|\eta_n(u) - \eta(u)\bigr| \xrightarrow{\mathrm{P}} 0, \qquad \sup_{0\leqslant u\leqslant t} \bigl|y_n(u) - y(u)\bigr| \xrightarrow{\mathrm{P}} 0 \tag{A.11}$$

(see (A.5)). As $m(a) > 0$, the process $y(t)$ strictly monotonically increases. Thus, the process $y^{-1}(t) = \mu(t)$ exists for all $t < y(+\infty)$, is continuous and

$$\sup_{0\leqslant u\leqslant t} \bigl|\mu_n(u) - \mu(u)\bigr| \xrightarrow{\mathrm{P}} 0. \tag{A.12}$$

Using the result about the uniform convergence of a superposition of random functions [16] and relations (A.11), (A.12), we obtain (A.6).

Denote $v_{nk} = \gamma_n(y_{nk})$. We introduce a stochastic process $v_n(t) = v_{nk}$, as $k/n \leqslant u < (k+1)/n$, $u \geqslant 0$. Using relations (A.1) and results [27] it is possible to prove that the sequence $v_n(u)$ weakly converges in $\mathcal{D}_T$ for any $T > 0$ to the diffusion process $v(u)$ satisfying the following stochastic differential equation: $v(0) = \gamma_0$, $\mathrm{d}v(u) = m(\eta(u))q(\eta(u), v(u))\,\mathrm{d}u + D(\eta(u))\,\mathrm{d}w(u)$. Also, we can get the relation

$$\sup_{0\leqslant t\leqslant T} \left|\gamma_n(t) - v_n\left(\mu_n(t) - \frac{1}{n}\right)\right| \xrightarrow{\mathrm{p}} 0.$$

But the sequence $v_n(\mu_n(t) - 1/n)$ weakly converges in $\mathcal{D}_T$ to the process $v(\mu(t)) = \gamma(t)$. As far as $\mu'(t) = m(s(t))^{-1}$, we calculate a stochastic differential of $\gamma(t)$ using the relation $\mathrm{d}w(\mu(t)) = \sqrt{\mu'(t)}\,\mathrm{d}w(t)$ and get (A.10). □

The result of theorem A.1 is also valid, if the value s_0 is a random variable, and corresponding relations involving s_0 are satisfied with probability one.

Consider a particular case, when $S_n(t)$ is a homogeneous MP. Suppose that $S_n(t)$ is a regular stepwise process with transition rates $q_n(\alpha, A), \alpha \in \mathcal{R}^r$, $A \in B^r_{\mathcal{R}}$, $\alpha \notin A$, where $q_n(\alpha) = q_n(\alpha, \mathcal{R}^r \backslash \{\alpha\}) < \infty$. We introduce independent families of random variables $\{\xi_{nk}(\alpha),\ \alpha \in \mathcal{R}^r\}$, $k \geqslant 0$, and $\{\tau_{nk}(\alpha),\ \alpha \in R\}$, $k \geqslant 0$, with values in $\mathcal{R}^r$ and $[0, \infty)$, respectively. Here $\tau_{nk}(n\alpha)$ has an exponential distribution with parameter $q_n(\alpha)$ and $\mathbf{P}\{\xi_{nk}(n\alpha) \in A\} = q_n(\alpha)^{-1} q_n(\alpha, A + \alpha)$, $\alpha \notin A$, where $A + \alpha = \{z\colon\ z - \alpha \in A\}$. Then RPSM defined by variables $(\zeta_{nk}(\alpha), \tau_{nk}(\alpha))$, $k \geqslant 0$, is equivalent to a MP $S_n(t)$, $t \geqslant 0$, $m_n(\alpha) = q_n(\alpha)^{-1}$, and it is easy to calculate that $D_n^2(\alpha) = \mathbf{E}\xi_{n1}(n\alpha)\xi_{n1}(n\alpha)^*$.

Appendix B. AP and DA for processes with semi-Markov switches

Consider now AP- and DA-type theorems for PSMS. Let for each $n = 1, 2 \ldots$, $\mathcal{F}_{nk} = \{\zeta_{nk}(t, x, \alpha),\ t \geqslant 0,\ x \in X,\ \alpha \in \mathcal{R}^r\}$, $k \geqslant 0$, be jointly independent families of stochastic processes in $\mathcal{D}^r_\infty$, $\{x_n(t);\ t \geqslant 0\}$ be an independent of $\mathcal{F}_{nk}$ SMP with values in some measurable space X, S_{n0} be an initial value. Let also $0 = t_{n0} < t_{n1} < \cdots$ be the times of sequential jumps of $x_n(\cdot)$, $x_{nk} = x_n(t_{nk})$, $k \geqslant 0$. We construct a PSMS according to (2.9): put $S_{nk+1} = S_{nk} + \xi_{nk}$, where $\xi_{nk} = \zeta_{nk}(\tau_{nk},\ x_{nk},\ S_{nk})$, $\tau_{nk} = t_{nk+1} - t_{nk}$, and denote

$$\zeta_n(t) = S_{nk} + \zeta_{nk}(t - t_{nk}, x_{nk}, S_{nk}), \quad \text{as } t_{nk} \leqslant t < t_{nk+1},\ t \geqslant 0. \tag{B.1}$$

Then the process $\{(x_n(t), \zeta_n(t));\ t \geqslant 0\}$ is a PSMS.

First, we study an AP for the switched component $\zeta_n(\cdot)$. Consider for simplicity a homogeneous case (distributions of $\zeta_{nk}(\cdot)$ do not depend on index $k \geqslant 0$). Let $\tau_n(x)$ be the sojourn time in state x for SMP $x_n(\cdot)$. Denote for each $x \in X$, $\alpha \in \mathcal{R}^r$,

$$\xi_n(x, \alpha) = \zeta_{n1}\big(\tau_n(x), x, \alpha\big), \qquad g_n(x, \alpha) = \sup_{t < \tau_n(x)} \big|\zeta_{n1}(t, x, n\alpha)\big|.$$

Suppose that MP x_{nk}, $k \geqslant 0$, at each $n > 0$ has a stationary measure $\pi_n(A)$, $A \in \mathcal{B}_X$, and denote

$$\begin{aligned}
&m_n(x) = \mathbf{E}\tau_n(x), && b_n(x, \alpha) = \mathbf{E}\xi_n(x, n\alpha),\\
&m_n = \int_X m_n(x)\pi_n(\mathrm{d}x), && b_n(\alpha) = \int_X b_n(x, \alpha)\pi_n(\mathrm{d}x),\\
&\alpha_n(k) = \sup_{A, B \in \mathcal{B}_X, i \geqslant 0} \big|\mathbf{P}\{x_{ni} \in A,\ x_{n,i+k} \in B\} - \mathbf{P}\{x_{ni} \in A\}\mathbf{P}\{x_{n,i+k} \in B\}\big|.
\end{aligned}$$

Theorem B.1 (Averaging principle). Suppose that $n^{-1}S_{n0} \xrightarrow{\mathrm{P}} s_0$, there exists a sequence of integers r_n such that

$$n^{-1}r_n \to 0, \qquad \sup_{k \geqslant r_n} \alpha_n(k) \to 0, \tag{B.2}$$

for any $N > 0$, $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{|\alpha|\leqslant N} \sup_x n\mathbf{P}\{n^{-1}g_n(x,\alpha) > \varepsilon\} = 0, \tag{B.3}$$

$$\lim_{L\to\infty} \limsup_{n\to\infty} \sup_{|\alpha|\leqslant N} \sup_x \{\mathbf{E}\tau_{n1}(x)\chi(\tau_{n1}(x) > L) + \mathbf{E}|\xi_{n1}(x,n\alpha)|\chi(|\xi(x,n\alpha)| > L)\} = 0, \tag{B.4}$$

for any x, $|b_n(x,\alpha_1) - b_n(x,\alpha_2)| \leqslant C_N|\alpha_1 - \alpha_2| + \alpha_n(N)$, as $\max(|\alpha_1|, |\alpha_2|) < N$, where C_N are some constants, and $\alpha_n(N) \to 0$ uniformly in $|\alpha_1| \leqslant N$, $|\alpha_2| \leqslant N$. Let also there exist a function $b(\alpha)$ and a constant $m > 0$ such that for any $\alpha \in \mathcal{R}^r$ $b_n(\alpha) \to b(\alpha)$, $m_n \to m$.

Then for any $T > 0$,

$$\sup_{0\leqslant t\leqslant T} |n^{-1}\zeta_n(nt) - s(t)| \xrightarrow{\mathrm{P}} 0, \tag{B.5}$$

where

$$s(0) = s_0, \qquad \mathrm{d}s(t) = m^{-1}b(s(t))\,\mathrm{d}t \tag{B.6}$$

(it is supposed that a unique solution of (B.6) exists on each interval).

Remark B.2. Condition (B.2) can be satisfied also in some more general cases, when the process x_{nk} is not ergodic in the limit. For instance, a state space can form an S-set (see [6]).

Consider a DA for the sequence of processes $\gamma_n(t) = n^{-1/2}(\zeta_n(nt) - ns(t))$. Introduce a uniformly strong mixing coefficient for the process x_{nk}:

$$\varphi_n(k) = \sup_{x,y,A} |\mathbf{P}\{x_{nk} \in A \mid x_{n0} = x\} - \mathbf{P}\{x_{nk} \in A \mid x_{n0} = y\}|, \quad k > 0.$$

Put

$$\begin{aligned} \widetilde{b}_n(\alpha) &= b_n(\alpha)m_n^{-1}, \qquad \widetilde{b}(\alpha) = b(\alpha)m^{-1}, \\ \rho_{nk}(x,\alpha) &= \xi_{nk}(x,n\alpha) - b_n(x,\alpha) - \widetilde{b}(\alpha)(\tau_{nk}(x) - m_n(x)), \\ \gamma_n(x,\alpha) &= b_n(x,\alpha) - b_n(\alpha) - \widetilde{b}(\alpha)(m_n(x) - m_n), \\ D_n^2(x,\alpha) &= \mathbf{E}\rho_{n1}(x,\alpha)\rho_{n1}(x,\alpha)^*. \end{aligned} \tag{B.7}$$

Theorem B.3 (Diffusion approximation). Suppose that $\gamma_n(0) \overset{\mathrm{w}}{\Rightarrow} \gamma_0$, there exist a fixed $r > 0$ and $q < 1$ such that $\varphi_n(r) \leqslant q$ for any $n > 0$, conditions of theorem B.1 hold, where $\sqrt{n}\alpha_n(N) \to 0$, and for any $N > 0$ the following conditions are satisfied:

$$\begin{aligned}
&\lim_{n\to\infty}\sup_{|\alpha|\leqslant N}\sup_{x} n\mathbf{P}\{n^{-1/2}g_n(x,\alpha)>\varepsilon\}=0,\quad \forall\varepsilon>0;\\
&\lim_{L\to\infty}\lim_{n\to\infty}\sup_{|\alpha|\leqslant N}\sup_{x}\{\mathbf{E}\tau_{n1}(x)^2\chi(\tau_{n1}(x)>L)\\
&\quad+\mathbf{E}|\xi_{n1}(x,n\alpha)|^2\chi(|\xi_{n1}(x,n\alpha)|>L)\}=0;\\
&|D_n^2(x,\alpha_1)-D_n^2(x,\alpha_2)|\leqslant C_N|\alpha_1-\alpha_2|+\alpha_n(N),\quad x\in X
\end{aligned}\tag{B.8}$$

as $\max(|\alpha_1|,|\alpha_2|)\leqslant N$, where $\alpha_n(N)\to 0$ uniformly in $|\alpha_1|\leqslant N$, $|\alpha_2|\leqslant N$; there exist a continuous vector-valued function $q(\alpha,z)$ and matrix-valued functions $D^2(\alpha)$ and $B^2(\alpha)$ such that in any domain $|\alpha|\leqslant N$, $|q(\alpha,z)|\leqslant C_N(1+|z|)$; uniformly in $|\alpha|\leqslant N$ at each fixed z

$$\sqrt{n}\big(\widetilde{b}_n\big(\alpha+n^{-1/2}z\big)-\widetilde{b}(\alpha)\big)\to q(\alpha,z);$$

for any $\alpha\in\mathcal{R}^m$

$$\begin{aligned}
D_n^2(\alpha)&=\int_X D_n^2(x,\alpha)\pi_n(\mathrm{d}x)\to D^2(\alpha),\\
\big(B_n^{(1)}(\alpha)\big)^2+\big(B_n^{(2)}(\alpha)\big)^2+\big(B_n^{(2)}(\alpha)^*\big)^2&\to B^2(\alpha),
\end{aligned}\tag{B.9}$$

where $(B_n^{(1)}(\alpha))^2=\int_X\gamma_n(x,\alpha)\gamma_n(x,\alpha)^*\pi_n(\mathrm{d}x)$, and

$$B_n^{(2)}(\alpha)^2=\sum_{k\geqslant 1}\mathbf{E}\gamma_n(x_{n0},\alpha)\gamma_n(x_{nk},\alpha)^*,\tag{B.10}$$

with $\mathbf{P}\{x_{n0}\in A\}=\pi_n(A)$, $A\in\mathcal{B}_X$.

Then for any $T>0$ the sequence $\gamma_n(t)$ weakly converges in $\mathcal{D}_T$ to the diffusion process $\gamma(t)$:

$$\mathrm{d}\gamma(t)=q\big(s(t),\gamma(t)\big)\,\mathrm{d}t+m^{-1/2}\big(D^2\big(s(t)\big)+B^2\big(s(t)\big)\big)^{1/2}\,\mathrm{d}w(t),\qquad \gamma(0)=\gamma_0,\tag{B.11}$$

where $C=A^{1/2}$ means that $CC^*=A$, $w(t)$ is a standard Wiener process in $\mathcal{R}^r$, and a unique solution of (B.11) exists.

Proof. The proof of theorems B.1, B.3 follows the same scheme as the proof of theorems A.1, A.2 and uses the results on the convergence of stochastic recurrent sequences in a Markov environment to the solutions of stochastic differential equations [14]. More details can be found in [4,5]. □

Conditions (B.3), (B.8) mean that there are no large jumps on the switching intervals.

Results of theorems B.1, B.3 show that at given condition (B.3) (or (B.8), respectively) the asymptotic behavior of PSMS $\zeta_n(\cdot)$ and simpler RPSM $S_n(\cdot)$, which is constructed only by accumulating increments on switching intervals, is the same. This is very useful in applications, because we do not need to keep track of the whole trajectory of the process on switching intervals.

These results can be extended to nonhomogeneous in time models also [5].

References

[1] V.V. Anisimov, Limit theorems for random processes and their applications to discrete summation schemes, Theory Probab. Appl. 20 (1975) 692–694.

[2] V.V. Anisimov, Switching processes, Cybernetics 13(4) (1977) 590–595.

[3] V.V. Anisimov, Averaging principle for switching processes, Theory Probab. Math. Statist. 46 (1992) 1–10.

[4] V.V. Anisimov, Limit theorems for processes with semi-Markov switching and their applications, Random Oper. Stochastic Equations 2(4) (1994) 333–352.

[5] V.V. Anisimov, Switching processes: averaging principle, diffusion approximation and applications, Acta Appl. Math. 40 (1995) 95–141.

[6] V.V. Anisimov, Asymptotic analysis of switching queueing systems in conditions of low and heavy loading, in: *Matrix-Analytic Methods in Stochastic Models*, eds. A.S. Alfa and S.R. Chakravarthy, Lecture Notes in Pure and Applied Mathematics, Vol. 183 (Marcel Dekker, New York, 1996) pp. 241–260.

[7] V.V. Anisimov, Averaging principle for near-critical branching processes with semi-Markov switching, Theory Probab. Math. Statist. 52 (1996) 13–26.

[8] V.V. Anisimov, Averaging methods for transient regimes in overloading retrial queuing systems, Math. Comput. Modelling 30(3/4) (1999) 65–78.

[9] V.V. Anisimov, Switching stochastic models and applications in retrial queues, Top 7(2) (1999) 169–186.

[10] V.V. Anisimov and A.O. Aliev, Limit theorems for recurrent processes of semi-Markov type, Theory Probab. Math. Statist. 41 (1990) 7–13.

[11] V.V. Anisimov and J.R. Artalejo, Analysis of Markov multiserver retrial queues with negative arrivals, Queueing Systems 39 (2001) 157–182.

[12] V.V. Anisimov and V.S. Chabanyuk, On applying of Skorokhod reflecting problem at diffusion approximation of queueing networks, Soviet Phys. Dokl. 35(6) (1990) 505–506.

[13] V.V. Anisimov and E.A. Lebedev, *Stochastic Queueing Networks. Markov Models* (Kiev Univ., Kiev, 1992) (in Russian).

[14] V.V. Anisimov and A.P. Yurachkovskiy, A limit theorem for stochastic difference schemes with random coefficients, Theory Probab. Math. Statist. 33 (1986) 1–9.

[15] G.P. Basharin, P.P. Bocharov and J.A. Kogan, *Analysis of Queues in Computing Networks* (Nauka, Moscow, 1989) (in Russian).

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

[17] M. Bramson, State space collapse with applications to heavy traffic limits for multiclass queueing networks, Queueing Systems 30(1/2) (1998) 89–148.

[18] M. Bramson and J.G. Dai, Heavy traffic limits for some queueing networks, Ann. Appl. Probab. 11 (2001) 49–90.

[19] N.P. Buslenko, V.V. Kalashnikov and I.N. Kovalenko, *Lectures on the Theory of Complex Systems* (Sov. Radio, Moscow, 1973) (in Russian).

[20] H. Chen and H. Zhang, Diffusion approximations for some multiclass queueing networks under FIFO disciplines, Math. Oper. Res. 25(4) (2000) 679–707.

[21] J.G. Dai, W. Dai, A heavy traffic limit theorems for a class of open queueing networks with finite buffers, Queueing Systems 32 (1999) 5–40.

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

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

[24] I.I. Ežov and A.V. Skorokhod, Markov processes which are homogeneous in the second component, Theory Probab. Appl. 14 (1969) 679–692.
[25] G.I. Falin and J.G.C. Templeton, *Retrial Queues* (Chapman and Hall, London, 1997).
[26] E. Gelenbe and G. Pujolle, *Introduction to Queueing Networks* (Wiley, Chichester, 1998).
[27] I.I. Gikhman and A.V. Skorokhod, *Theory of Random Processes III* (Springer, Berlin, 1978).
[28] J.M. Harrison, Balanced fluid models of multiclass queueing network: A heavy traffic conjecture, in: *Stochastic Networks*, eds. F. Kelly and R. Williams, IMA Volumes in Mathematics and Its Applications, Vol. 71 (Springer, New York, 1995) pp. 1–20.
[29] J.M. Harrison and R.J. Williams, A multiclass closed queueing network with unconventional heavy traffic behavior, Ann. Appl. Probab. 6(1) (1996) 1–47.
[30] R. Hersh, Random evolutions: survey of results and problems, Rocky Mountain J. Math. 4(3) (1974) 443–475.
[31] D.L. Iglehart, Limit diffusion approximation for the many server queue and the repairman problem, J. Appl. Probab. 2 (1965) 429–441.
[32] R. Kertz, Random evolutions with underlying semi-Markov processes, Publ. Res. Inst. Math. Sci. 14 (1978) 589–614.
[33] V.S. Korolyuk and A.V. Swishchuk, *Random Evolutions* (Kluwer Academic, Dordrecht, 1994).
[34] E.V. Krichagina, R.S. Liptser and A.A. Puhalsky, Diffusion approximation for the system with arrival process depending on queue and arbitrary service distribution, Theory Probab. Appl. 33 (1988) 124–135.
[35] T. Kurtz, A limit theorem for perturbed operator semigroups with applications to random evolutions, J. Funct. Anal. 12 (1973) 55–67.
[36] R.S. Liptser and A.N. Shiryaev, *Theory of Martingales* (Kluwer Academic, Dordrecht, 1989).
[37] A. Mandelbaum and W. Masey, Strong approximation for time-dependent queues, Math. Oper. Res. 20(1) (1995) 33–64.
[38] A. Mandelbaum, W. Masey and M.I. Reiman, Strong approximation for Markovian service networks, Queueing Systems 30(1/2) (1998) 149–202.
[39] A. Mandelbaum and G. Pats, State-dependent stochastic networks. Part I: Approximations and applications with continuous diffusion limits, Ann. Appl. Probab. 8(2) (1998) 569–646.
[40] M. Neuts, *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications* (Marcel Dekker, New York/Basel, 1989).
[41] G. Papanicolaou and R. Hersh, Some limit theorems for stochastic equations and applications, Indiana Univ. Math. J. 21 (1972) 815–840.
[42] M. Pinsky, *Random Evolutions*, Lecture Notes in Mathematics, Vol. 451 (Springer, Berlin, 1975) pp. 89–100.
[43] M.I. Reiman, Open queueing networks in heavy traffic, Math. Oper. Res. 9(3) (1984) 441–458.
[44] M.I. Reiman, A multiclass feedback queue in heavy traffic, Adv. in Appl. Probab. 20 (1988) 179–207.
[45] A.V. Skorokhod, Limit theorems for random processes, Theory Probab. Appl. 1 (1956) 289–319.
[46] A.V. Skorokhod, Stochastic equations for diffusion processes with boundaries. I, II, Theory Probab. Appl. 6 (1961) 287–298; 7 (1962) 5–25.
[47] W. Whitt, On the heavy-traffic limit theorem for $GI/G/\infty$ queues, Adv. in Appl. Probab. 14 (1982) 171–190.
[48] R.J. Williams, On the approximation of queueing networks in heavy traffic, in: *Stochastic Networks. Theory and Applications*, eds. F.P. Kelly, S. Zachary and I. Zieding (Oxford Univ. Press, Oxford, 1996) pp. 35–56.
[49] R.J. Williams, Diffusion approximation for open multiclass queueing networks: Sufficient conditions involving state space collapse, Queueing Systems 30(1/2) (1998) 27–88.