APPLIED STOCHASTIC METHODS SERIES

Switching Processes in Queueing Models

Vladimir Anisimov

ISTE

WILEY

Switching Processes in Queueing Models

Vladimir V. Anisimov

Series Editor
Nikolaos Limnios

iSTE

WILEY

First published in Great Britain and the United States in 2008 by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd	John Wiley & Sons, Inc.
6 Fitzroy Square	111 River Street
London W1T 5DX	Hoboken, NJ 07030
UK	USA
www.iste.co.uk	www.wiley.com

Library of Congress Cataloging-in-Publication Data

Anisimov, V. V. (Vladimir Vladislavovich)
Switching processes in queueing models / Vladimir V. Anisimov.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-84821-045-5
1. Telecommunication--Switching systems--Mathematical models. 2. Telecommunication--Traffic--Mathematical models. 3. Queuing theory. I. Title.
TK5102.985.A55 2008
519.8'2--dc22

2008008995

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN: 978-1-84821-045-5

Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire.

Contents

To my wife, Zoya

Preface

Contemporary communication systems and computer networks usually have a rather complex structure and therefore require creating more complicated mathematical models of queues and developing new approaches for modeling and asymptotic investigation. The main features of these systems are the stochasticity of the processes describing the behavior in time, influence of various internal and external events which may change (switch) the behavior of the system, the presence of different time scales for different subsystems (very fast internal computer time and user interaction time, etc), and the hierarchic structure. Wide classes of such systems can be adequately described with the help of so-called "switching" stochastic processes.

Switching processes (SP) have been developed by the author for describing the operation of stochastic systems with the property that their development in time varies spontaneously (switches) at some random points of time which may depend on the previous system trajectory. According to Kolmogorov, these processes can be called random processes with discrete interference of chance or with discrete components. Processes of this type often appear in the theory of queueing and communication systems and networks, branching, population and migration processes, in the analysis of stochastic dynamical systems with random perturbations, random movements and various other applications.

SP can be represented as a two-component process $(x(t), \zeta(t))$, $t \geq 0$, with the property that there exists a sequence of Markov points of time $t_1 < t_2 < \cdots$ such that in each interval $[t_k, t_{k+1})$, $x(t) = x(t_k)$, and the behavior of the process $\zeta(t)$ in this interval depends only on the value $(x(t_k), \zeta(t_k))$. $x(t)$ is a discrete switching component and the points of time $\{t_k\}$ are called switching times. SP can be described in terms of constructive characteristics and is very suitable in analyzing and asymptotic investigating of stochastic systems with "rare" and "fast" switching.

The class of SPs is the natural generalization of well-known classes of random processes such as Markov processes that are homogenous in the 2nd component, processes with independent increments and Markov or semi-Markov switches, piecewise

Markov aggregates, and Markov processes with Markov and semi-Markov switching (random evolutions). Wide classes of queueing models can be described in terms of SPs. The class of switching queueing models includes, as examples, various types of state-dependent queueing systems and networks in a Markov or semi-Markov environment, queueing models under the influence of flows of external events or internal perturbations, unreliable systems, retrial queues, hierarchic queueing systems, etc. Therefore, the asymptotic theory of SPs can be effectively applied to the investigation of wide classes of queueing systems and networks.

In the book several large directions of asymptotic results for SP are investigated and successfully applied to various classes of switching queueing models.

The first direction is devoted to the limit theorems of averaging principle (AP) and diffusion approximation (DA) type in the case of fast switching. Theorems on the convergence of the trajectory of an SP to a solution of a differential equation (AP) and the convergence of the normalized difference to a diffusion process (DA) are proved for different subclasses of SP: recurrent processes of a semi-Markov type (RPSMs), processes with semi-Markov switching and general SP with feedback between both components. The results are based on the investigation of the asymptotic properties of a special subclass of SP – RPSMs theorems on the convergence of recurrent sequences with Markov switching to the solutions of stochastic differential equations and the convergence of superpositions of random functions.

This class of theorems is the basis of a new approach to the investigation of transient phenomena for service processes in overloading queueing systems and Markov and semi-Markov type networks, retrial queues, etc. Numerous examples for the illustration AP and DA for queueing models are considered.

The second direction is devoted to the limit theorems for SP with slow switching. Models of this type appear at the investigation of hierarchic systems in different scales of time (slow and fast). The conditions, when an SP of a rather complicated structure can be approximated by an SP of a simpler structure, in particular, by a Markov or semi-Markov process, are established and various applications to processes with Markov and semi-Markov switching are considered. The method of investigation uses the results on the convergence of the accumulating type processes constructed on the trajectory of Markov or semi-Markov process satisfying some form of the asymptotic mixing condition in triangular scheme to processes with independent increments (homogenous or non-homogenous in time). A special class of non-homogenous in time Markov processes with transition probabilities slowly varying in the expanding time scale is introduced. These processes have quasi-ergodic properties and are called quasi-ergodic Markov processes. Under rather general conditions it is proved a Poisson approximation of the flows of rare events governed by a Markov process satisfying an asymptotic mixing condition, in particular with the state space forming a so-called S-set (asymptotically connected set), and the exponential approximation of the exit time from S-set. Special attention is paid to the analysis of the flow of rare events defined on stochastic systems satisfying asymptotic mixing conditions, in particular, with state space forming an S-set. These models naturally appear at study queueing models with asymptotically "fast" service (or low traffic). Applications of a method of S-sets are considered for different classes of queueing systems.

Using these results and the results on the convergence of SP with slow switching, the models of the asymptotic aggregation of the state space of Markov and semi-Markov processes (homogenous and non-homogenous in time) are investigated. These results create the basis for a theory of the asymptotic decreasing dimension and aggregation (consolidation) of the state space of stochastic systems. Special attention is paid to the hierarchic Markov and semi-Markov systems operating in different time scales. These systems under rather general conditions can be approximated by a simpler Markov system with averaged transition characteristics. The applications to the asymptotic aggregation of a state space and approximation by Markov models with averaged characteristics are considered for different classes of Markov and non-Markov queueing models in a random environment.

The asymptotic aggregation of SP in different time scales is the next natural level of development. The conditions of the convergence of SP to solutions of differential and stochastic differential equations with coefficients depending on a limiting aggregated Markov or semi-Markov process are obtained. Various applications to the asymptotic aggregation of overloaded queueing systems and networks under the influence of hierarchic random environment in different time scales are considered.

The results of the book were obtained while the author was working at Kiev University as Head and Professor of Applied Statistics Department at the Faculty of Cybernetics (1978–2002) and also as Visiting Professor at Bilkent University, Ankara (1997–2002). Some results were reflected in different courses on stochastic processes and queueing models that the author taught at Kiev University and Bilkent University for graduate and post-graduate students.

The book contains many practical examples of asymptotic results for queueing models and is directed to applied mathematicians and researchers, post-graduate students and engineers working in the field of stochastic systems, queueing models and applications to computer sciences, biology, ecology, physical and social sciences. Some theoretical results are illustrated by examples of simulation in R.

The author is sincerely grateful to professors Vladimir Korolyuk, Anatoli Skorokhod, Igor Kovalenko and Nikolaos Limnios for a fruitful long-term collaboration and useful discussions.

Vladimir V. Anisimov
March 2008

Definitions

Throughout the book we use the following abbreviations:

$\mathbf{P}(A)$	probability of event A
$\mathbf{E}\xi$	expectation of a random variable ξ
$\mathbf{Var}\xi$	variance of a random variable ξ
MP	Markov process
SMP	semi-Markov process
RPSM	recurrent process of semi-Markov type
PSMS	process with semi-Markov switching
SP	switching process
$\stackrel{\text{w}}{\Longrightarrow}$	the weak convergence of random variables in the sense of weak convergence of probability distributions
$\stackrel{\text{P}}{\longrightarrow}$	the convergence in probability
$\mathcal{D}_T$	the space of right-continuous functions defined on $[0, T]$ with finite left limits (Skorokhod space)
$\mathcal{C}_T$	the space of continuous functions defined on $[0, T]$
$[a]$	usually means the integer part of a
$\bar{a}$	vector in space $\mathcal{R}^r$
$\bar{a}'$	transposed vector

We say that the sequence of random processes $\zeta_n(\cdot)$ J-converges to the process $\zeta_0(\cdot)$ in an interval $[0, T]$ as $n \to \infty$, if the sequence of measures generated by $\zeta_n(\cdot)$ in Skorokhod space $\mathcal{D}_T$ weakly converges to the corresponding measure generated by $\zeta_0(\cdot)$.

J-convergence in $[0, T]$ means a weak convergence of finite dimensional distributions in all points in $[0, T]$ except possibly a countable set and a relative compactness of corresponding measures in $[0, T]$ (see [BIL 68]).

Readers are refered to [SKO 56, BIL 68, ETH 86] for the definition of Skorokhod space and J-convergence.

Bibliography

[BIL 68] BILLINGSLEY P., *Convergence of Probability Measures*, Wiley, New York, 1968.

[ETH 86] ETHIER S. and KURTZ T., *Markov Processes, Characterization and Convergence*, John Wiley & Sons, New York, 1986.

[SKO 56] SKOROKHOD A., "Limit theorems for random processes", *Theory Prob. Appl.*, vol. 1, p. 289–319, 1956.

Chapter 1

Switching Stochastic Models

This chapter is devoted to the description of the main types of switching stochastic models.

1.1. Random processes with discrete component

The operation of a wide range of stochastic systems can be described in terms of random processes such that the character of their operation over time varies spontaneously (switches) at random times.

Models of this type appear at the investigation of systems operating in a random environment, systems with different stochastic perturbations and also systems of a complex hierarchic structure. Switching may be determined by external factors (for example, when a system operates in an environment and does not influence the environment itself), and also by internal and interconnected factors (switching is determined by some events depending on the trajectory of the system). The latter situation corresponds to the case of feedback. Thus, in the general case, the switching times may be random functionals of the previous trajectory of the system.

Switching models appear in the theory of queueing systems and networks, branching processes, in the analysis of stochastic dynamic systems with random perturbations, in recurrent algorithms with random calculation times and other various applications.

According to Kolmogorov, these processes are called random processes with discrete interference of chance or with a discrete component. A wide range of different modifications of processes with discrete interference of chance have been studied: Markov processes that are homogenous in the second component [EZO 69], processes with independent increments and semi-Markov switching [ANI 73], piecewise Markov aggregates [BUS 73], Markov processes with semi-Markov interference of chance [GIK 75] and Markov and semi-Markov evolutions [KAC 74, GRI 69, HER 74, HER 03, KER 78a, KER 78b, KUR 73, PAP 72, PIN 75, KOR 94, KOR 05]. SPs were introduced by the author [ANI 75, ANI 77, ANI 78, ANI 88b].

Different classes of asymptotic results for processes with a discrete component were investigated by many authors. The Law of Large Numbers and the Central Limit Theorem for random evolutions have been proved by different authors [GRI 69, PAP 72, KUR 73, KER 78a, KER 78b, PIN 75, ANI 73, ANI 88b, KOR 93, KOR 94, WAT 84, KOR 90]. The next natural stage is the investigation of the asymptotic behavior of random evolution in transient conditions.

There are two major directions of asymptotic results for processes with a discrete component. The first direction is devoted to the limit theorems of averaging principle (AP) and diffusion approximation (DA) in the case of fast switches. Various AP and DA results for dynamic systems and stochastic differential equations with fast Markov switching were proved by [KHA 68, SKO 89, TSA 93a, TSA 93b].

The author investigated AP and DA for special subclasses of SP [ANI 90, ANI 92a, ANI 93, ANI 94, ANI 95] and these results were applied to branching processes [ANI 96], stochastic differential equations [ANI 86, ANI 89] and to different classes of queueing systems [ANI 99a, ANI 99c, ANI 02b, ANI 92b, ANI 99b, ANI 01, ANI 04]. Another approach based on the asymptotic results for semi-groups of perturbed operators associated with corresponding Markov processes was developed by V.S. Korolyuk *et al.* and various applications to dynamic systems and queueing models [KOR 94, KOR 99, KOR 04] are studied.

The second direction is related to the aggregation (merging) of the states of stochastic systems. These situations appear when we consider two-level (hierarchic) systems with “fast” transitions on the first level and “slow” transitions on the second level. For example, a stochastic system with fast transitions operates in an external environment with slow transitions. The approach based on the asymptotic results for semi-groups of perturbed operators was also successfully applied to this class of problems [KOR 93, KOR 99, KOR 00, KOR 04, KOR 05].

The author developed a different approach based on the limit theorems for SP in the case of slow switches [ANI 73, ANI 78, ANI 88b, ANI 88a] and investigated the conditions when an SP of a rather complicated structure can be approximated by an SP of a simpler structure, in particular, by a Markov or a semi-Markov process. These results created the basis for a new approach to problems of asymptotic decreasing dimension and aggregation of state space of stochastic systems. Different applications to reliability and queueing models and dynamic systems are considered in [ANI 88b, ANI 00a, ANI 02a, ANI 04, ANI 87, ANI 99b, ANI 00b].

In the following sections we consider some special models of processes with a discrete component such as Markov or semi-Markov processes, processes with independent increments and Markov or semi-Markov switching.

1.1.1. *Markov and semi-Markov processes*

First let us consider a well-known constructive definition of homogenous Markov process (MP) in continuous time. Let $\{\tau^{(k)}(x),\ x \in X\}$, $k \geq 0$, be jointly independent at different k families of random variables such that the variable $\tau^{(k)}(x)$ at each x, k has an exponential distribution with parameter $\lambda(x)$ where the values $\lambda(x)$, $x \in X$, are given. Here X is a metric space with Borel σ-algebra $\mathcal{B}_X$ (we call it a measurable space $(X, \mathcal{B}_X)$). The family of transition probabilities $\{p(x, A),\ x \in X,\ A \in \mathcal{B}_X\}$ is also given. Suppose that x_0 is the initial value. We construct a Markov sequence x_k, $k \geq 0$, using transition probabilities in the following recurrent way: $t_0 = 0$, $t_{k+1} = t_k + \tau^{(k)}(x_k)$, and

$$P\big\{x_{k+1} \in A \mid x_k = x\big\} = p(x, A), \quad x \in X,\ A \in B_X,\ k \geq 0.$$

Then the process

$$x(t) = x_k \quad \text{as } t_k \leq t < t_{k+1},\ t \geq 0,$$

is a homogenous MP in continuous time. The transition rates from state x to the region A ($x \notin A$) are calculated using the formula: $\lambda(x, A) = p(x, A)\lambda(x)$.

If the variables $\{\tau^{(k)}(x),\ x \in X\}$, $k \geq 0$, have arbitrary distributions, then the process $x(t)$ is a semi-Markov process (SMP). In both cases the times t_k can be considered as switching times for the trajectory of the process.

Note that the process $x(t)$ is correctly defined only in the interval $[0, t_\infty)$, where $t_\infty = \lim_{k\to\infty} t_k$. If $t_k \xrightarrow{\text{P}} \infty$ as $k \to \infty$, then this process is defined in the interval $[0, \infty)$ and we call it regular. Here a symbol $\xrightarrow{\text{P}}$ means the convergence in probability.

In the following we will say that the process $x(\cdot)$ is regular if in any finite interval with probability one it has a finite number of jumps.

1.1.2. *Processes with independent increments and Markov switching*

Let $x(t)$, $t \geq 0$, be a regular homogenous MP with values in a measurable space $(X, \mathcal{B}_X)$ and let the following families that are independent of $x(\cdot)$ be given: the family of jointly independent at different k homogenous processes with independent increments $\{\xi^{(k)}(t, x),\ t \geq 0,\ x \in X\}$, $k \geq 0$, (Lévy processes) and the family of jointly independent at different k random variables $\{\gamma^{(k)}(x),\ x \in X\}$, $k \geq 1$, with values in $\mathcal{R}$. Suppose also that the distributions of processes $\xi^{(k)}(t, x)$ and variables $\gamma^{(k)}(x)$

do not depend on index k. It is known that a characteristic function of the Lévy process can be represented in the form

$$\mathbf{E}\exp\left\{i\theta\xi^{(1)}(t,x)\right\}=\exp\left\{t\psi(\theta,x)\right\},$$

where the cumulant function $\psi(\theta,x)$ has a well-known Lévy-Khintchine representation. Denote

$$\mathbf{E}\exp\left\{i\theta\gamma^{(1)}(x)\right\}=\varphi(\theta,x),\quad x\in X,\ \theta\in(-\infty,\infty).$$

Suppose that the functions $\psi(\theta,x)$, $\varphi(\theta,x)$ at each fixed θ as functions in x are measurable with respect to σ-algebra $\mathcal{B}_X$.

We define a two-component MP $(x(t),\zeta(t))$, $t\geq 0$, with values in (X,R) using the families introduced above. Let (x_0,ζ_0) be an initial value, $x(t)$ a right-continuous homogenous MP and $x(0)=x_0$. Denote by $0=t_0<t_1<t_2<\cdots$ the times of successive jumps of $x(t)$ and let $x_k=x(t_k)$, $k\geq 0$, be the values of the embedded MP. We put $\zeta(0)=\zeta_0$ and set

$$\begin{aligned}\zeta(t)&=\zeta\bigl(t_k\bigr)+\xi^{(k)}\bigl(t,x_k\bigr)-\xi^{(k)}\bigl(t_k,x_k\bigr),\quad \text{as } t\in\bigl[t_k,t_{k+1}\bigr),\\ \zeta\bigl(t_k\bigr)&=\zeta\bigl(t_k-0\bigr)+\gamma^{(k)}\bigl(x_k\bigr),\quad k>0.\end{aligned}\tag{1.1}$$

This means that in the interval $[t_k,t_{k+1})$ the process $\zeta(t)$ takes the increments of the process $\xi^{(k)}(t,x_k)$ and at time t_k it has a jump of the size $\gamma^{(k)}(x_k)$. Thus, by the construction, at the fixed trajectory of $x(t)$ the process $\zeta(t)$ develops as a non-homogenous process with independent increments and momentary value of the cumulant function $\psi(\theta,x(t))$ and it takes additional jumps at the times t_k, $k>0$, of the size $\gamma^{(k)}(x_k)$. Let, for simplicity, $\zeta_0=0$. Thus, the characteristic function of the one-dimensional distribution of the process $\zeta(t)$ is given by the expression:

$$\mathbf{E}\exp\left\{i\theta\zeta(t)\right\}=\mathbf{E}\exp\left\{\int_0^t\psi\bigl(\theta,x(u)\bigr)\mathrm{d}u+\sum_{0<t_k\leq t}\ln\varphi\bigl(\theta,x_k\bigr)\right\}.\tag{1.2}$$

Here the expectation is taken with respect to all possible trajectories of $x(u)$, $0\leq u\leq t$. If the intervals $[a_j,b_j]$, $j=\overline{1,r}$, are not overlapping, then

$$\begin{aligned}&\mathbf{E}\exp\left\{i\sum_{j=1}^r\theta_j\bigl(\psi\bigl(b_j\bigr)-\psi\bigl(a_j\bigr)\bigr)\right\}\\ &\quad=\mathbf{E}\prod_{j=1}^r\exp\left\{\int_{a_j}^{b_j}\psi\bigl(\theta_j,x(u)\bigr)du+\sum_{a_j<t_k\leq b_j}\ln\varphi\bigl(\theta_j,x_k\bigr)\right\}.\end{aligned}$$

This construction can be easily extended to the case where the processes $\xi^{(k)}(t,x)$ are non-homogenous in time and the distributions of variables $\gamma^{(k)}(x,t)$ also dependent on t. In this case the process $(x(t),\zeta(t))$, $t \geq 0$, is non-homogenous in time. Note that as $x(\cdot)$ is regular, the process $(x(t),\zeta(t))$, $t \geq 0$, is defined in the interval $[0,\infty)$.

By definition the two-component process $(x(t),\zeta(t))$, $t \geq 0$ is an MP which is homogenous in the 2nd component. This means that

$$\begin{aligned}
&\mathbf{P}\{x(t+s) \in A,\ \zeta(t+s) < z \mid x(t) = x,\ \zeta(t) = y\} \\
&\quad = \mathbf{P}\{x(t+s) \in A,\ \zeta(t+s) < z - y \mid x(t) = x,\ \zeta(t) = 0\}, \\
&\qquad t \geq 0,\ s \geq 0,\ x \in X,\ y,z \in (-\infty,\infty).
\end{aligned}$$

Note that Markov processes that were homogenous in the 2nd component were defined in [EZO 69].

1.1.3. *Processes with independent increments and semi-Markov switching*

Let $x(t)$, $t \geq 0$, be an SMP with the state space X and let the family of homogenous processes with independent increments $\xi^{(k)}(t,x)$, $t \geq 0$, $x \in X$, and the family of random variables $\gamma^{(k)}(x)$, $x \in X$, satisfying the conditions above, section 1.1.2, be given. Relations (1.1) thus define the two-component process $(x(t),\zeta(t))$, $t \geq 0$, which is the process with independent increments and semi-Markov switching (PII SMS) [ANI 73].

Let us consider a three-component process $(x(t),\gamma(t),\zeta(t))$, $t \geq 0$, where $\gamma(t) = t - \max_{t_k \leq t} t_k$. This process is a homogenous MP that is homogenous in the second component (see previous section 1.1.2). The one-dimensional distributions for $\zeta(t)$ are also defined by formula (1.2).

Now let us consider as an example a Poisson process with a random rate. Let $\lambda(t)$, $t \geq 0$, be the non-negative random process which represents the instantaneous rate for a Poisson process $\Pi_{\lambda(t)}(t)$. This means that at given trajectory of $\lambda(t,\omega)$ the process $\Pi_{\lambda(t,\omega)}(t)$ is developing as a non-homogenous Poisson process with the instantaneous rate $\lambda(t,\omega)$. A Poisson process of this type is called a double stochastic Poisson process or Cox process [COX 80].

The joint distributions in non-overlapping intervals $[a_j,b_j]$, $j = \overline{1,r}$, are calculated as follows

$$\mathbf{P}\{\Pi_{\lambda(b_i)}(b_i) - \Pi_{\lambda(a_i)}(a_i) = n_i,\ i = \overline{1,r}\} = \mathbf{E}\prod_{i=1}^{r} \exp\{-\Lambda(a_i,b_i)\}\frac{\Lambda(a_i,b_i)^{n_i}}{n_i!},$$

where $\Lambda(a,b)$ is a random cumulative rate in the interval $[a,b]$: $\Lambda(a,b) = \int_a^b \lambda(u)\mathrm{d}u$.

Consider a special case. Let the family of non-negative functions $a(x)$, $x \in X$, and the random process $x(t)$ with values in X be given. Suppose that the random rate $\lambda(t)$ has the form $\lambda(t) = a(x(t))$, $t \geq 0$. In this case the process $\Pi_{\lambda(t)}(t)$ is a Poisson process with a rate which is switching by the process $x(t)$. In particular, if $x(t)$ is a homogenous MP, then the two-component process $(x(t), \Pi_{\lambda(t)}(t))$ is a homogenous PII MS (see section 1.1.2). If the process $x(t)$ is an SMP, then the process $(x(t), \Pi_{\lambda(t)}(t))$ is a homogenous PII SMS.

1.2. Switching processes

1.2.1. *Definition of switching processes*

In this section we introduce a general definition of the class of switching processes (SPs). Let

$$\mathcal{F}_k = \big\{\big(\zeta_k(t,x,\alpha), \tau_k(x,\alpha), \beta_k(x,\alpha)\big),\ t \geq 0,\ x \in X,\ \alpha \in R^r\big\}, \quad k \geq 0,$$

be jointly independent in index k parametric families, where $(X, \mathcal{B}_X)$ is a measurable space, $\zeta_k(t,x,\alpha)$ for each fixed k, x, α is a random process with trajectories belonging to the Skorokhod space $D_\infty{}^r$ (the space of right-continuous functions having left-side limits which are also called cadlag functions), and $\tau_k(x,\alpha), \beta_k(x,\alpha)$, are possibly dependent on $\zeta_k(\cdot, x, a)$ random variables, $\tau_k(\cdot) \geq 0$, $\beta_k(\cdot) \in X$. We assume that the vectors from R^r are column vectors and the variables introduced are measurable in the ordinary way in the pair (x, a) concerning σ-algebra $\mathcal{B}_X \times \mathcal{B}_{R^r}$. Let (x_0, S_0) also be the initial vector in $X \times R^r$ independent of $\mathcal{F}_k$, $k \geq 0$. We introduce the following recurrent sequences:

$$\begin{gathered} t_0 = 0, \quad t_{k+1} = t_k + \tau_k\big(x_k, S_k\big), \\ S_{k+1} = S_k + \xi_k\big(x_k, S_k\big), \quad x_{k+1} = \beta_k\big(x_k, S_k\big), \quad k \geq 0, \end{gathered} \tag{1.3}$$

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

$$\begin{gathered} \zeta(t) = S_k + \zeta_k\big(t - t_k, x_k, S_k\big), \\ x(t) = x_k, \quad \text{as } t_k \leq t < t_{k+1},\ t \geq 0. \end{gathered} \tag{1.4}$$

A two-component process $(x(t), \zeta(t))$, $t \geq 0$, is thus called an SP [ANI 75, ANI 77]. We also introduce a process

$$S(t) = S_k \quad \text{as } t_k \leq t < t_{k+1},\ t \geq 0, \tag{1.5}$$

and call it a semi-Markov recurrent process (RPSM) [ANI 90]. Note that by definition RPSM $S(t)$ is an embedded process for $\zeta(t)$.

It is worth noting that the general definition of an SP allows feedback between the discrete component $x(\cdot)$ and the switched component $\zeta(\cdot)$ (case of feedback). Different classes of random evolutions studied earlier by other authors allow only the dependence of the component $\zeta(\cdot)$ on the process $x(\cdot)$. This corresponds to the case when we have an external environment corresponding to the process $x(\cdot)$ and its operation does not depend on the process $\zeta(\cdot)$, but operation of $\zeta(\cdot)$ may depend on the state of $x(\cdot)$. The case of feedback was not considered by other authors.

Consider a particular case when there is no switching component $x(\cdot)$ and the process $\zeta(\cdot)$ is switching at some random times t_k. Let

$$\mathcal{F}_k = \big\{\big(\zeta_k(t,\alpha), \tau_k(\alpha)\big),\ t \geq 0,\ \alpha \in R^r\big\}, \quad k \geq 0,$$

be jointly independent in index k families of random processes $\zeta_k(t,\alpha)$ with trajectories belonging to the Skorokhod space $\mathcal{D}^r_\infty$ and random variables $\tau_k(\alpha)$ measurable in the natural way. We introduce the following recurrent sequences:

$$t_0 = 0, \quad t_{k+1} = t_k + \tau_k\big(S_k\big), \quad S_{k+1} = S_k + \xi_k\big(S_k\big), \quad k \geq 0, \tag{1.6}$$

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

$$\zeta(t) = S_k + \zeta_k\big(t - t_k, S_k\big) \quad \text{as } t_k \leq t < t_{k+1},\ t \geq 0. \tag{1.7}$$

Process $\zeta(t)$, $t \geq 0$, is thus a special case of an SP which is switched by its own values at the switching times t_k constructed recurrently using the values of $\zeta(\cdot)$ at the switching times. This definition will be used later to describe some classes of queueing models.

In what follows we assume that the SP is regular, i.e. the component $x(\cdot)$ with probability one has a finite number of jumps in each finite interval. Now as an illustration we consider the special subclasses of SPs.

Assume that the characteristics of the families $\mathcal{F}_k$, $k \geq 0$, do not depend on parameters a and k. Then x_k, $k \geq 0$, is a homogenous Markov process (MP) and $x(t)$, $t \geq 0$, is a semi-Markov process (SMP). Assume also that the variables $\tau_k(x)$ at each $x \in X$ are independent of the processes $\zeta_k(t,x)$, $t \geq 0$. In that case, if the variables $\tau_k(x)$ have the exponential distribution, then $x(t)$, $t \geq 0$, is an MP, and if in addition $\zeta_k(t,x)$, $t \geq 0$, is at each $x \in X$, the process with independent increments, thus the two-component process $(x(t), \zeta(t))$, $t \geq 0$, forms an MP which is homogenous in the second component [EZO 69].

If the variables $\tau_k(x)$ have arbitrary distributions, process $\zeta(t)$ is a process with independent increments and semi-Markov switching introduced in [ANI 73], see section 1.1.3.

Suppose now that the process $\zeta_k(t,x)$ at each $x \in X$ is an MP. Then the process $\zeta(t)$, $t \geq 0$, in [BUS 73] is called a piecewise Markov aggregate and in the book by Gikhman and Skorokhod it is called an MP with semi-Markov interference of chance [GIK 75]. If the processes $\zeta_k(t,x)$ are realized in a Banach space and described by a semigroup of operators, then the process $\zeta(t)$ in [GRI 69, HER 74] is called a random evolution.

Now we consider separately a class of recurrent processes of semi-Markov type (RPSM) which is a special subclass of SP introduced above. It is defined as the embedded process for an SP (see equation (1.5)). This process is a stepwise process and has a simpler structure than a general SP. Below we consider some special RPSM models which will be used later in the description of stochastic queues and the problems of asymptotic analysis.

1.2.2. *Recurrent processes of semi-Markov type (simple case)*

Let

$$\mathcal{F}_k = \big\{ \big(\xi_k(\alpha), \tau_k(\alpha)\big),\ a \in R^r \big\}, \quad k \geq 0,$$

be jointly independent families of random variables with values in the space $R^r \times [0,\infty)$, and S_0 be an independent of $\mathcal{F}_k$, $k \geq 0$ random variable in R^r. Note that the variables $\xi_k(\alpha)$ and $\tau_k(\alpha)$ can be dependent. We introduce the following recurrent sequences:

$$t_0 = 0, \quad t_{k+1} = t_k + \tau_k\big(S_k\big), \quad S_{k+1} = S_k + \xi_k\big(S_k\big), \quad k \geq 0 \tag{1.8}$$

and set

$$S(t) = S_k \quad \text{as } t_k \leq t < t_{k+1},\ t \geq 0. \tag{1.9}$$

Process $S(t)$ thus forms an RPSM (in this case a discrete switching component $x(t)$ is absent).

If the distributions of the families $\mathcal{F}_k$ do not depend on parameter k, process $S(t)$ is a homogenous SMP. If the distributions of the families $\mathcal{F}_k$ do not depend on both parameters α and k, then the times $t_0 \leq t_1 \leq \cdots \leq t_k \cdots$, form a recurrent flow and $S(t)$ can be interpreted as a reward renewal process.

1.2.3. *RPSM with Markov switching*

Let

$$\mathcal{F}_k = \big\{ \big(\xi_k(x,\alpha), \tau_k(x,\alpha)\big),\ x \in X,\ \alpha \in R^r \big\}, \quad k \geq 0$$

be jointly independent families of random variables taking values in $R^r \times [0,\infty)$, and let x_l, $l \geq 0$, be an MP independent of $\mathcal{F}_k$, $k \geq 0$, with values in a space X and (x_0, S_0) be the initial value. We put

$$t_0 = 0, \quad t_{k+1} = t_k + \tau_k\bigl(x_k, S_k\bigr), \quad S_{k+1} = S_k + \xi_k\bigl(x_k, S_k\bigr), \quad k \geq 0, \tag{1.10}$$

and set

$$S(t) = S_k, \quad x(t) = x_k, \quad t_k \leq t < t_{k+1},\ t \geq 0. \tag{1.11}$$

The two-component process $(x(t), S(t))$ then forms an RPSM with additional Markov switching. If the distributions of variables $\tau_k(x,\alpha)$ do not depend on the parameters α and k, then $x(t)$ is an SMP.

1.2.4. *General case of RPSM*

Let

$$\mathcal{F}_k = \bigl\{\bigl(\xi_k(x,\alpha), \tau_k(x,\alpha), \beta_k(x,\alpha)\bigr),\ x \in X,\ \alpha \in R^r\bigr\}, \quad k \geq 0,$$

be jointly independent families of random variables with values in the space $R^r \times [0,\infty) \times X$, (x_0, S_0) be the initial value. We put

$$\begin{gathered} t_0 = 0, \quad t_{k+1} = t_k + \tau_k\bigl(x_k, S_k\bigr), \\ S_{k+1} = S_k + \xi_k\bigl(x_k, S_k\bigr), \quad x_{k+1} = \beta_k\bigl(x_k, S_k\bigr), \quad k \geq 0, \end{gathered} \tag{1.12}$$

and set

$$S(t) = S_k, \quad x(t) = x_k \quad \text{as } t_k \leq t < t_{k+1},\ t \geq 0. \tag{1.13}$$

The two-component process $(x(t), S(t))$, $t \geq 0$ then forms a general RPSM with feedback between both components. In particular, when the distributions of the variables $\beta_k(x,\alpha)$ do not depend on parameter α, sequence x_k is an MP with values in X and this case corresponds to the previous case (RPSM with additional Markov switching). In the general case, we have feedback between components $x(t)$ and $S(t)$ at the switching times (case of feedback).

1.2.5. *Processes with Markov or semi-Markov switching*

Consider the case when a random process operates in the external Markov or semi-Markov environment. Let

$$\mathcal{F}_k = \bigl\{\zeta_k(t,x,\alpha),\ t \geq 0,\ x \in X,\ \alpha \in R^r\bigr\}, \quad k \geq 0,$$

be the jointly independent parametric families of random processes with trajectories in Skorokhod space ${D_\infty}^r$, where $(X, \mathcal{B}_X)$ is a measurable space. Let $x(t)$, $t \geq 0$, also be the independent of $\mathcal{F}_k$, $k \geq 0$, right-continuous SMP with values in X and S_0 be the initial value. We suppose that the variables introduced are measurable in the ordinary way in the pair (x, a) with respect to σ-algebra $\mathcal{B}_X \times \mathcal{B}_{R^r}$.

Denote by $0 = t_0 < t_1 < \cdots$ the times of consecutive jumps of $x(\cdot)$ and put $x_k = x(t_k)$, $k \geq 0$. We construct a process with Markov (or semi-Markov) switching in the following way. Put

$$S_{k+1} = S_k + \xi_k,$$

where

$$\xi_k = \zeta_k\big(\tau_k, x_k, S_k\big), \quad \tau_k = t_{k+1} - t_k,$$

and set

$$\zeta(t) = S_k + \zeta_k\big(t - t_k, x_k, S_k\big) \quad \text{as } t_k \leq t < t_{k+1},\ t \geq 0. \tag{1.14}$$

Then the two-component process $(x(t), \zeta(t))$, $t \geq 0$, is a process with Markov switching (PMS) if $x(t)$ is a Markov process, or a process with semi-Markov switching (PSMS), if $x(t)$ is a semi-Markov process.

The component $x(t)$ itself is a switching environment. Let us also introduce the embedded process

$$S(t) = S_k \quad \text{as } t_k \leq t < t_{k+1}. \tag{1.15}$$

We can easily see that the process $(x(t), S(t))$ is an RPSM with independent Markov switching (section 1.2.3).

Consider a special case where $\{\zeta(t, x),\ t \geq 0\}$ is the family of Markov processes and denote by $\zeta(t, x, \alpha)$ the process $\zeta(t, x)$ with the initial value α. Process $(x(t), \zeta(t))$ thus forms a Markov random evolution (when $x(t)$ is a Markov process), or a semi-Markov random evolution (when $x(t)$ is a semi-Markov process).

1.3. Switching stochastic models

In this section we consider different examples of switching stochastic models which can be described in terms of an SP.

1.3.1. *Sums of random variables*

Let $\{\xi_k(x),\ x \in X\}$, $k \geq 0$, be the jointly independent families of random variables in $\mathcal{R}^r$, and let x_k, $k \geq 0$, be an MP in X that is independent of these families. Denote

$$\zeta_m = \sum_{k=0}^{m} \xi_k(x_k), \quad m \geq 0. \tag{1.16}$$

The two-component process (x_m, ζ_m), $m \geq 0$, is a process with independent increments and Markov switching in discrete time (see section 1.1.2).

Furthermore, let $p(x, A, \alpha)$, $x \in X$, $A \in B_x$, be at each $a \in \mathcal{R}^r$ the family of transition probabilities and the natural conditions of measurability are satisfied. Let x_0 be some initial state and we define the sequence x_k, $k \geq 1$, in the following way:

$$P\big\{x_{k+1} \in A \mid x_i, \zeta_i,\ i \leq k\big\} = p\big(x_k, A, \zeta_k\big), \quad k \geq 0,$$

where ζ_k is given by expression (1.16). Then the two-component process (x_m, ζ_m), $m \geq 0$, is an SP with feedback between both components. In this case (x_m, ζ_m) is an MP, but component x_m itself in general is not an MP.

1.3.2. *Random movements*

In this section we describe a random movement switched by a Markov or semi-Markov environment. This means that the speed vector depends on the current state of the environment.

Let $v_1, v_2, \ldots, v_m$ be deterministic vectors in $\mathcal{R}^r$, $x(t)$, $t \geq 0$, be an MP or SMP with a finite number of states $\{1, 2, \ldots, m\}$. We put $\zeta_k(t, i) = tv_i$, $t \geq 0$, $i = \overline{1, m}$, (the movement when the state of the environment is i) and denote by $0 = t_0 < t_1 < \cdots$ the times of successive jumps of $x(t)$. Then an SP constructed by the family of processes $\zeta_k(t, i)$, $t \geq 0$, $i = \overline{1, m}$, by switching component $x(t)$ and times t_k generates a random movement in $\mathcal{R}^r$ with Markov or correspondingly semi-Markov switching.

Let us represent the position of movement $\zeta(t)$ at any time t. Denote

$$\nu(t) = \max\big\{k : k \geq 0,\ t_k < t\big\}.$$

If ζ_0 is the initial position, then setting $\sum_0^{-1} = 0$ we obtain:

$$\zeta(t) = \zeta_0 + \sum_{k=0}^{\nu(t)-1} \big(t_{k+1} - t_k\big) v_{x_k} + \big(t - t_{\nu(t)}\big) v_{x_{\nu(t)}}. \tag{1.17}$$

1.3.3. *Dynamic systems in a random environment*

Let $\{f(x,\alpha),\ x \in X,\ \alpha \in \mathcal{R}^r\}$ be the family of functions with values in the space $\mathcal{R}^r$, $\{\gamma_k(x,\alpha),\ x \in X,\ \alpha \in \mathcal{R}^r\}$, $k \geq 0$, be the jointly independent families of random variables with values in $\mathcal{R}^r$ and $x(t)$, $t \geq 0$, be a stepwise random process in X which is independent of introduced families. In addition, let $t_1 < t_2 < \cdots$ be the sequential times of jumps of process $x(t)$. We introduce process $\zeta(t)$ in the following way: at each interval (t_k, t_{k+1}),

$$\zeta(0) = \zeta_0, \quad \mathrm{d}\zeta(t) = f\big(x_k, \zeta(t)\big)\mathrm{d}t, \tag{1.18}$$

where $x_k = x(t_k + 0)$, and at times t_k,

$$\zeta\big(t_{k+1} + 0\big) = \zeta\big(t_{k+1} - 0\big) + \gamma_k\big(x_k, \zeta\big(t_{k+1} - 0\big)\big), \quad k > 0, \tag{1.19}$$

which means process $\zeta(t)$ has a jump of size $\gamma_k(x_k, \zeta(t_{k+1} - 0))$ at time t_{k+1} depending on the value of the switching component and the value of the trajectory of $\zeta(t)$ at this time. Note that when $x(t)$ is an MP or SMP, $\zeta(t)$ is a dynamic system with Markov or semi-Markov switching.

To represent process $\zeta(t)$ as an SP we introduce the family of functions $v(t, x, a)$ in the following way:

$$v(0, x, \alpha) = \alpha, \quad \mathrm{d}v(t, x, \alpha) = f\big(x, v(t, x, \alpha)\big)\mathrm{d}t, \tag{1.20}$$

and suppose that a unique solution of (1.20) exists on each interval $[0, t]$ for any x, α. For each $k \geq 0$, we put $\tau_k = t_{k+1} - t_k$, $x_k = x(t_k + 0)$, and define the family of processes $\{\zeta_k(t, x, \alpha)\}$ in the following way:

$$\zeta_k(t, x, \alpha) = v(t, x, \alpha) \quad \text{as } t < \tau_k,\ x_k = x,$$

$$\zeta_k\big(\tau_k, x, \alpha\big) = v\big(\tau_k, x, \alpha\big) + \gamma_k\big(x, v\big(\tau_k, x, \alpha\big)\big).$$

Then the family of processes $\zeta_k(t, x, \alpha)$, $t \leq \tau_k$, together with sequences t_k and x_k, define SP $(x(t), \zeta(t))$, $t \geq 0$ according to formulae (1.3), (1.4).

1.3.4. *Stochastic differential equations in a random environment*

Let $\{c(x,\alpha), b(x,\alpha),\ x \in X,\ \alpha \in R\}$ be the family of vector and matrix-valued functions of the dimension r and $r \times r$, respectively, and let $w(t)$, $t \geq 0$, be a standard Wiener process in $\mathcal{R}^r$ and $x(t)$, $t \geq 0$, be an MP or SMP, which is independent of $w(t)$. We introduce process $\zeta(t)$ as a solution to the following stochastic differential equation:

$$\zeta(0) = \zeta_0, \quad d\zeta(t) = c\big(x(t), \zeta(t)\big)dt + b\big(x(t), \zeta(t)\big)\mathrm{d}w(t), \tag{1.21}$$

where at each $x \in X$ coefficients $c(x,a)$ and $b(x,a)$ satisfy the conditions of the theorem of the existence and uniqueness of the solution. The process $\zeta(t)$ is a diffusion process with Markov of semi-Markov switching and then the pair $(x(t), \zeta(t))$, $t \geq 0$, is an SP which is a Markov process with Markov (or semi-Markov) switching. In this case the distributions of processes $\zeta_k(t,x,\alpha)$ do not depend on index k and for each x, α, process $\zeta_1(t,x,\alpha)$ is defined as follows: $\zeta_1(0,x,\alpha) = \alpha$ and $\mathrm{d}\zeta_1(t,x,\alpha) = c(x, \zeta_1(t,x,\alpha))\mathrm{d}t + b(x, \zeta_1(t,x,\alpha))\mathrm{d}w(t)$.

In this example the behavior of process $x(t)$ does not depend on the values of $\zeta(\cdot)$ and the main process $\zeta(\cdot)$ is switched by process $x(\cdot)$ (no feedback between both components). Now we describe the case of feedback. Let $\{c(x,\alpha), b(x,\alpha), x \in X, \alpha \in \mathcal{R}^r\}$ be the family of vector and matrix-valued functions of the dimension r and $r \times r$, respectively, and let the parametric family of functions $\{q(x,A,z),\ x \in X,\ A \in B_X,\ \alpha \in \mathcal{R}^r\}$, measurable in the ordinary way, be given such that at each fixed α, $q(x,A,\alpha)$ are the transition rates of some non-breaking stepwise MP in X.

We construct an SP $(x(t), \zeta(t)), t \geq 0$, in the following way: $\zeta(t)$ satisfies equation (1.21), where $x(t)$ is constructed jointly with $\zeta(t)$ as a stepwise process in X such that for all $x \notin A$, $\alpha \in \mathcal{R}^r$,

$$\mathbf{P}\big\{x(t+h) \in A \mid x(t) = x,\ \zeta(t) = \alpha\big\} = q(x,A,\alpha)h + o(h).$$

In this case $(x(t), \zeta(t))$ is an MP and we have an effect of mutual influence between both components (case of feedback).

1.3.5. *Branching processes*

Let us consider the state-dependent branching process in continuous time which is defined in the following way. The non-negative functions $\mu(m)$, $\nu(m)$, $m = 0, 1, \ldots$ and the independent families of random variables $\{\eta(m),\ m = 0, 1, \ldots\}$ and $\{\gamma(m), m = 0, 1, \ldots\}$, with values in $\{0, 1, 2, \ldots\}$ and $\{0, \pm 1, \pm 2, \ldots\}$, respectively, are given. Let $Q(t)$ be the total number of particles at time t. If $Q(t) = m$, then in a small interval $[t, t+h]$, each particle, independently of others, with probability $\mu(m)h + o(h)$, can be transformed into $\eta(m)$ particles, and also with probability $\nu(m)h + o(h)$, the whole population of m particles can be transformed into $[m + \gamma(m)]_+$ particles, where $[k]_+ = \max(0,k)$.

In this case process $Q(t)$ can be described as a simple RPSM. Let $\{\tau_k(m), m = 0, 1, \ldots\}$, $k \geq 0$, and $\{\xi_k(m),\ m = 0, 1, \ldots\}$, $k \geq 0$, be the independent families of random variables, where $\tau_k(m)$ has an exponential distribution with parameter $\lambda(m) = m\mu(m) + \nu(m)$ and

$$\begin{aligned}\mathbf{P}\big\{\xi_k(m) \in A\big\} &= m\mu(m)\lambda(m)^{-1}\mathbf{P}\big\{\eta(m) - 1 \in A\big\} \\ &\qquad + \nu(m)\lambda(m)^{-1}\mathbf{P}\big\{\big[m + \gamma(m)\big]_+ - m \in A\big\}.\end{aligned}$$

Thus, process $Q(t)$ is equivalent to a simple RPSM which is defined by the family $\{(\xi_k(m), \tau_k(m)),\ m = 0, 1, \ldots\}$, $k \geq 0$, according to formulae (1.8), (1.9).

By analogy we can construct a branching process with Markov or semi-Markov switching and with different types of interacting particles.

1.3.6. *State-dependent flows*

Consider a flow of random events which is constructed in the following way. Let $\{\tau_k(\alpha),\ \alpha \geq 0\}$, $k \geq 0$, be families of non-negative variables, which are jointly independent of index k. We introduce the following recurrent sequence:

$$t_0 = 0, \quad t_{k+1} = t_k + \tau_k\big(t_k\big), \quad k \geq 0.$$

Then sequence $\{t_k,\ k \geq 0\}$ represents a state-dependent flow of random events. By definition,

$$\mathbf{P}\big\{t_{k+1} - t_k < z \mid t_k = \alpha,\ t_i = \alpha_i,\ i = \overline{1, k-1}\big\} = \mathbf{P}\big\{\tau_k(\alpha) < z\big\}.$$

Note that by definition, a sequence t_k, $k \geq 0$, can be represented as an RPSM. If the distributions of variables $\tau_k(\alpha)$ do not depend on index k and parameter α, then sequence $\{t_k,\ k \geq 0\}$ represents a well-known recurrent flow.

By analogy we can construct a state-dependent flow in a random environment. Let $\{(\tau_k(x, \alpha),\ \beta_k(x, \alpha)),\ \alpha \geq 0,\ x \in X,\ k \geq 0\}$ be parametric families of random variables which are jointly independent of index k. Suppose that the initial value x_0 is given. Let us define the following recurrent sequences:

$$t_0 = 0, \quad t_{k+1} = t_k + \tau_k\big(x_k, t_k\big), \quad x_{k+1} = \beta_k\big(x_k, t_k\big), \quad k \geq 0.$$

The pair (x_k, t_k) represents a state-dependent flow in a random environment.

We can also construct a corresponding SP. Denote

$$\nu(t) = \max\big\{k : k \geq 0,\ t_k < t\big\}, \quad x(t) = x_{\nu(t)}, \quad S(t) = t_{\nu(t)}.$$

Then a two-component process $(x(t), S(t))$, $t \geq 0$, is an SP.

1.3.7. *Two-level Markov systems with feedback*

Consider n interacting Markov systems which are operating in the following way. Each system has a finite state space $\{1, 2, \ldots, r\}$. Let the family of non-negative functions $\lambda_{ij}(\overline{q})$, $i, j = \overline{1, r}$, $i \neq j$, $\overline{q} = (q_1, \ldots, q_r)$, $q_i \geq 0$, $\sum_{i=1}^{r} q_i = 1$, be given.

Denote by $\nu_n(i,t)$ the total number of systems in state i at time t and put

$$\overline{\nu}_n(t) = \left(\frac{1}{n}\nu_n(i,t),\ i = \overline{1,r}\right).$$

If at time t, $\overline{\nu}_n(t) = \overline{q}$, and a system is in state i, then this system in a small interval $[t, t+h]$ independently of other systems can jump to state j with probability

$$n^{-1}\lambda_{ij}(\overline{q})h + o(h), \quad j \neq i,$$

or, correspondingly, with probability

$$1 - n^{-1}\sum_{j\neq i}\lambda_{ij}(\overline{q})h + o(h)$$

the system stays in state i. Then process $\overline{\nu}_n(t)$ represents a two-level Markov system.

Denote by $x_k(t)$ the state of the kth system at time t. Then the process

$$\big(\overline{\nu}_n(t), \big(x_1(t), \ldots, x_n(t)\big)\big), \quad t \geq 0,$$

can be represented as an SP and, by definition, at each interval $[s,t]$ such that $\overline{\nu}_n(u) = \overline{q}$, $s \leq u \leq t$, the process $x_k(t)$ is an homogenous MP with transition rates $\lambda_{ij}(\overline{q})$, $i, j = \overline{1,r}$, $i \neq j$. In this case we have feedback between components $\overline{\nu}_n(t)$ and $(x_1(t), \ldots, x_n(t))$.

1.4. Bibliography

[ANI 73] ANISIMOV V., "Asymptotic consolidation of the states of random processes", *Cybernetics*, vol. 9, no. 3, p. 494–504, 1973.

[ANI 75] ANISIMOV V., "Limit theorems for random processes and their applications to discrete summation schemes", *Theor. Probab. Appl.*, vol. 20, 1975.

[ANI 77] ANISIMOV V., "Switching processes", *Cybernetics*, vol. 13, no. 4, p. 590–595, 1977.

[ANI 78] ANISIMOV V., "Limit theorems for switching processes and their applications", *Cybernetics*, vol. 14, no. 6, p. 917–929, 1978.

[ANI 86] ANISIMOV V. and YURACHKOVSKIY A., "A limit theorem for stochastic difference schemes with random coefficients", *Theor. Prob. and Math. Stat.*, vol. 33, p. 1–9, 1986.

[ANI 87] ANISIMOV V., ZAKUSILO O. and DONTCHENKO V., *The Elements of Queueing Theory and Asymptotic Analysis of Systems*, Visca Scola (Russian), Kiev, Ukraine, 1987.

[ANI 88a] ANISIMOV V., "Limit theorems for switching processes", *Theor. Probab. and Math. Stat.*, vol. 37, p. 1–5, 1988.

[ANI 88b] ANISIMOV V., *Random Processes with Discrete Component. Limit Theorems*, Kiev University (Russian), Kiev, Ukraine, 1988.

[ANI 89] ANISIMOV V. and YURACHKOVSKIY A., "Averaging principle for stochastic difference equations", *Ukrainian Math. J.*, vol. 41, p. 1022–1028, 1989.

[ANI 90] ANISIMOV V. and ALIEV A., "Limit theorems for recurrent processes of semi-Markov type", *Theor. Prob. and Math. Stat.*, vol. 41, p. 7–13, 1990.

[ANI 92a] ANISIMOV V., "Averaging principle for switching processes", *Theor. Probab. and Math. Stat.*, vol. 46, p. 1–10, 1992.

[ANI 92b] ANISIMOV V. and LEBEDEV E., *Stochastic Queueing Networks. Markov Models*, Kiev University (Russian), Kiev, Ukraine, 1992.

[ANI 93] ANISIMOV V., "Averaging principle for the processes with fast switching", *Random Oper. and Stoch. Eqv.*, vol. 1, no. 2, p. 151–160, 1993.

[ANI 94] ANISIMOV V., "Limit theorems for processes with semi-Markov switching and their applications", *Random Oper. and Stoch. Eqv.*, vol. 2, no. 4, p. 333–352, 1994.

[ANI 95] ANISIMOV V., "Switching processes: averaging principle, diffusion approximation and applications", *Acta Applicandae Mathematicae*, vol. 40, p. 95–141, 1995.

[ANI 96] ANISIMOV V., "Averaging principle for near-critical branching processes with semi-Markov switching", *Theor. Probab. and Math. Stat.*, vol. 52, p. 13–26, 1996.

[ANI 99a] ANISIMOV V., "Averaging methods for transient regimes in overloading retrial queuing systems", *Mathematical and Computing Modelling*, vol. 30, no. 3/4, p. 65–78, 1999.

[ANI 99b] ANISIMOV V., "Diffusion approximation for processes with semi-Markov switches and applications in queueing models", in JANSSEN J. and LIMNIOS N., Eds., *Semi-Markov Models and Applications*, p. 77–101, Kluwer Acad. Publ., Dordrecht, 1999.

[ANI 99c] ANISIMOV V., "Switching stochastic models and applications in retrial queues", *Top*, vol. 7, no. 2, p. 169–186, 1999.

[ANI 00a] ANISIMOV V., "Asymptotic analysis of reliability for switching systems in light and heavy traffic conditions", in LIMNIOS N. and NIKULIN M., Eds., *Recent Advances in Reliability Theory: Methodology, Practice, and Inference*, p. 119–133, Birkhäuser Boston, Massachusetts, 2000.

[ANI 00b] ANISIMOV V., "J-convergence for switching processes with rare perturbations to diffusion processes with Poisson type jumps", in KOROLYUK V., PORTENKO N. and SYTA H., Eds., *Skorokhod's Ideas in Probability Theory*, p. 81–98, Inst. of Math. Nat. Acad. Sci. of Ukraine, Kiev, 2000.

[ANI 01] ANISIMOV V. and ARTALEJO J., "Analysis of Markov multiserver retrial queues with negative arrivals", *Queueing Systems*, vol. 39, no. 2/3, p. 157–182, 2001.

[ANI 02a] ANISIMOV V., "Averaging in Markov models with fast Markov switches and applications to queueing models", *Annals of Operations Research*, vol. 112, no. 1, p. 63–82, 2002.

[ANI 02b] ANISIMOV V., "Diffusion approximation in overloaded switching queueing models", *Queueing Systems*, vol. 40, no. 2, p. 141–180, 2002.

[ANI 04] ANISIMOV V., "Averaging in Markov models with fast semi-Markov switches and applications", *Communications in Statistics - Theory and Methods*, vol. 33, no. 3, p. 517–531, 2004.

[BUS 73] BUSLENKO N., KALASHNIKOV V. and KOVALENKO I., *Lectures on the Theory of Complex Systems (Russian)*, Sov. Radio, Moscow, 1973.

[COX 80] COX D. and ISHAM V., *Point Processes*, Chapman & Hall, London, 1980.

[EZO 69] EZOV I. and SKOROKHOD A., "Markov processes which are homogenous in the second component", *Theor. Probab. Appl.*, vol. 14, p. 679–692, 1969.

[GIK 75] GIKHMAN I. and SKOROKHOD A., *Theory of Random Processes, II*, Springer-Verlag, New York, 1975.

[GRI 69] GRIEGO R. and HERSH R., "Random evolutions, Markov chains, and systems of partial differential equations", *Proc. Nat. Acad. Sci. USA*, vol. 62, p. 305–308, 1969.

[HER 74] HERSH R., "Random evolutions: survey of results and problems", *Rocky Mount. J. Math.*, vol. 4, no. 3, p. 443–477, 1974.

[HER 03] HERSH R., "The birth of random evolutions", *Mathematical Intelligence*, vol. 25, no. 1, p. 53–60, 2003.

[KAC 74] KAC M., "A stochastic model related to the telegrapher's equation", *Rocky Mount. J. Math.*, vol. 4, p. 497–509, 1974.

[KER 78a] KERTZ R., "Limit theorems for semigroups with perturbed generators, with applications to multiscaled random evolutions", *J. Funct. Anal.*, vol. 27, no. 2, p. 215–233, 1978.

[KER 78b] KERTZ R., "Random evolutions with underlying semi-Markov processes", *Publ. Res. Inst. Math. Sci.*, vol. 14, p. 589–614, 1978.

[KHA 68] KHAS'MINSKII R., "About the averaging principle for ITO stochastic differential equations", *Kybernetika*, vol. 4, no. 3, p. 260–279, 1968.

[KOR 90] KOROLYUK V., "Central limit theorem for semi-Markov random evolutions", *Comp. Math. Appl.*, vol. 19, no. 1, p. 83–88, 1990.

[KOR 93] KOROLYUK V. and TURBIN A., *Mathematical Foundation of the State Lumping of Large Systems*, Kluwer, Dordrecht, 1993.

[KOR 94] KOROLYUK V. and SWISHCHUK A., *Random Evolutions*, Kluwer, Dordrecht, 1994.

[KOR 99] KOROLYUK V. and KOROLYUK V., *Stochastic Models of Systems*, Kluwer, Dordrecht, 1999.

[KOR 00] KOROLYUK V. and LIMNIOS N., "Evolutionary systems in an asymptotic split phase space", in LIMNIOS N. and NIKULIN M., Eds., *Recent Advances in Reliability Theory: Methodology, Practice and Inference*, p. 145–161, Birkhäuser Boston, Massachusetts, 2000.

[KOR 04] KOROLYUK V. and LIMNIOS N., "Average and diffusion approximation for evolutionary systems in an asymptotic split phase state", *Ann. Appl. Prob.*, vol. 14, no. 1, p. 489–516, 2004.

[KOR 05] KOROLYUK V. and LIMNIOS N., *Stochastic Systems in Merging Phase Space*, World Scientific, Singapore, 2005.

[KUR 73] KURTZ T., "A limit theorem for perturbed operator semigroups with applications to random evolutions", *J. Funct. Anal.*, vol. 12, p. 55–67, 1973.

[PAP 72] PAPANICOLAOU G. and HERSH R., "Some limit theorems for stochastic equations and applications", *Indiana Univ. Math. J.*, vol. 21, p. 815–840, 1972.

[PIN 75] PINSKY M., "Random evolutions", in *Probabilistic Methods in Differential Equations*, vol. 451 of *Lecture Notes in Math.*, p. 89–99, Springer, Berlin, 1975.

[SKO 89] SKOROKHOD A., *Asymptotic Methods in the Theory of Stochastic Differential Equations*, Amer. Math. Soc., Rhode Island, 1989.

[TSA 93a] TSARKOV J., "Averaging and stability of impulse systems with rapid Markov switchings", *Proc. Latv. Prob. Sem.*, vol. 2, p. 49–63, 1993.

[TSA 93b] TSARKOV J., "Limit theorems for impulse systems with rapid Markov switchings", *Proc. Latv. Prob. Sem.*, vol. 2, p. 74–96, 1993.

[WAT 84] WATKINS J., "A central limit problem in random evolutions", *Ann. Prob.*, vol. 12, no. 2, p. 480–513, 1984.

Chapter 2

Switching Queueing Models

2.1. Introduction

A class of switching processes is a convenient tool for the description and asymptotic investigation of wide classes of queueing systems. In this chapter we provide various examples of switching queueing systems. Traditionally a queueing system is defined by the input flow of calls (or customers), by the queue discipline (in which order incoming calls are taken for service) and by the character of service. In what follows we use a general classification of queueing systems introduced by Kendall. According to this classification a standard queueing system is defined by a sequence of 4 basic symbols. The first symbol defines the input flow, the second characterizes a service time, the third defines the number of servers, the fourth defines the number of waiting places.

For example, the letters M, D, E, GI for the first symbol mean that the durations of the intervals between times when the calls (customers) enter the system are independent, identically distributed, random variables and have exponential, degenerative, Erlang or general distribution, respectively. The letters M, D, E, GI for the second symbol in a similar way define the type of distribution of the service time.

Additional sub-indexes can mean different specifications of the system, e.g. index B at the first symbol means that calls arrive in batches, e.g. M_B, means that calls arrive at the times of events of a Poisson process in batches. Index B at the second symbol, e.g. M_B, means that the service if provided in batches. Index Q means the additional dependence of input flow (or service) on the value of the current queue (state-dependent system). For example, notation M_Q at the first place means that the rate of a Poisson input process depends on the current value of the queue. Index M or SM means the dependence of input flow (or service time) on the state of an external

Markov or semi-Markov process. This case can be interpreted as a queueing system in a Markov or semi-Markov environment, or a system governed by an external Markov or semi-Markov process. If this is not specified, it is usually assumed that the calls are taken for service in the order of their arrival at the system (the so-called FIFO discipline, first-in-first-out).

For example, a traditional system $M/M/m/s$ means that the calls arrive one at a time according to a Poisson process (exponential interarrival times), there are m identical servers with exponential service times and s waiting places. The call upon arrival at the system either takes one of available servers, if any, or joins the queue (occupies one of the waiting places), if there are free places. If there are no available servers or waiting places, this call leaves the system. If the call takes one of the available servers, then after the exponential service time this call leaves the system. Immediately after service completion the server is ready to take the next call from the queue, if any. Otherwise, the server waits for the next call to arrive. If the call joins the queue, then it will take the first available server only when all calls standing in the queue before him have been taken for service.

By switching queueing model we mean that the queueing process can be represented as an SP. A large class of switching queueing models is the queues in a random environment. Usually the environment is considered as an external process which is not influenced by the behavior of the internal characteristics, for example, queueing processes. The systematic investigation of queues in a Markov environment was started by Neuts [NEU 81, NEU 89] who also introduced the notion of the Markov arrival process also known as the compound Poisson arrival process with Markov switching. Different classes of queueing models in a Markov environment and later in a semi-Markov environment have been investigated by many authors. See for example [PUR 74, O'C 86, SZT 87, FAL 88a, FAL 88b, GEL 90, ANI 92b, FIS 93, LUC 94b, LUC 94a, HE 96, BOC 03, KIM 07b]. A good survey of papers on queueing models in a Markov environment is provided in [KIM 07a].

In the cases when there is feedback between the environment and queueing processes in the system, the environment itself is not a Markov or semi-Markov process and in this case it is more natural to describe these models as switching models. In the next sections we consider some classes of queueing systems and networks which can be represented using switching processes.

2.2. Queueing systems

2.2.1. *Markov queueing models*

Let us illustrate the basic ideas of how to represent the queueing process in the system as an SP with the example of a rather simple, state-dependent, Markov queueing model.

2.2.1.1. *A state-dependent system* $M_Q/M_Q/1/\infty$

A system consists of one server with infinite buffer (an infinite number of waiting places). The calls arrive one at a time and are served according to the FIFO discipline. Let non-negative functions $\{\lambda(q), \mu(q),\ q \geq 0\}$ be given. Denote by $Q(t)$ the total number of calls in the system at time t. The system operates as follows. If at time t, $Q(t) = q$, then the local arrival rate is $\lambda(q)$. This means that the probability that a new call arrives in a small interval $[t, t+h]$ is $\lambda(q)h + o(h)$. Correspondingly, the local service rate is $\mu(q)$ (the probability that a call in service completes service in a small interval $[t, t+h]$ is $\mu(q)h + o(h)$). After service completion the call leaves the system.

It is well known that in this case process $Q(t)$, $t \geq 0$, is a Birth-and-Death process.

Let us represent this in a recurrent form. Denote by $t_1 < t_2 < \cdots$ the times of any change in the system (arrival of a call or service completion), and put $Q_k = Q(t_k + 0)$, $k \geq 0$. Suppose that $t_0 = 0$, $Q(0+) = Q_0$.

First, let us define the family of jointly independent random variables $\{\tau_k(q), \xi_k(q),\ q \geq 0\}$, $k \geq 0$, where $\tau_k(q)$ has an exponential distribution with parameter $\Lambda(q) = \lambda(q) + \mu(q)\chi(q > 0)$, $\xi_k(q)$ is a variable that is independent of $\tau_k(q)$ such that

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

and $\chi(A)$ is the indicator of set A. Define the following recurrent sequences:

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

and put

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

As we can see, process $\widetilde{Q}(t)$ is a simple RPSM (see section 1.2.2) and by definition the finite dimensional distributions of process $\widetilde{Q}(t)$ coincide with corresponding distributions of queueing process $Q(t)$.

The advantage of the representation of the queueing process as an RPSM is that $\widetilde{Q}(t)$ is a superposition of the two more simple recurrent processes in discrete time, $\widetilde{t}_k$ and $\widetilde{Q}_k$, $k \geq 0$.

This representation also provides an idea for how to study the limiting behavior of $Q(t)$. If we can prove that the appropriately scaled two component process $(\widetilde{t}_k, \widetilde{Q}_k)$

weakly converges to process $(y(u), q(u))$, $u \geq 0$, where components $y(u)$ and $q(u)$ are possibly dependent, 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.

Similar representation is true for a system $M_Q/M_Q/r/\infty$. In this case, given that $Q(t) = q$, we assume that the local rate of incoming calls is $\lambda(q)$ and the service rate for each busy server is $\mu(q)$. $Q(t)$ is thus a Birth-and-Death process with birth and death rates $\lambda(q)$ and $\min(q,r)\mu(q)$, respectively, and in the expressions above, $\Lambda(q) = \lambda(q) + \min(q,r)\mu(q)$,

$$\xi_k(q) = \begin{cases} +1, & \text{with prob. } \lambda(q)\Lambda(q)^{-1}, \\ -1, & \text{with prob. } \min(q,r)\mu(q)\Lambda(q)^{-1}. \end{cases}$$

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

For further exploration note that in fact the exponentiality of $\tau_k(q)$ is not essential for the asymptotic analysis. This 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 can analyze rather general 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, rather general results on AP and DA for an SP are proved in [ANI 90, ANI 92a, ANI 94, ANI 95] (see Chapter 4). These results provide us with the new methodology for asymptotic investigation of switching queueing models.

2.2.1.2. *Queueing system* $M_{M,Q}/M_{M,Q}/1/m$

Consider, as a more complicated example, a state-dependent queueing system in a Markov environment. Let $z(t)$, $t \geq 0$, be an homogenous MP with finite state space $I = \{1, \ldots, d\}$ and transition rates $a(i,l)$, $i,l = \overline{1,d}$, $i \neq l$. $z(t)$ stands for the external Markov environment. In addition, let the family of non-negative functions $\lambda(i,j)$, $\mu(i,j)$, $i = \overline{1,d}$, $j = \overline{0,m+1}$, be given. The system consists of one server with m waiting places. The calls enter the system one at a time. Denote by $Q(t)$ the number of calls in the system at time t, $0 \leq Q(t) \leq m+1$. If $z(t) = i$ and $Q(t) = j$, then the local input rate for incoming calls is $\lambda(i,j)$ and the local service rate is $\mu(i,j)$ ($\mu(i,0) \equiv 0$). The call, upon arrival at the empty system, is immediately taken for service. When the server is busy, the call joins the queue. After completion of service the call leaves the system and the next call from the queue, if any, is immediately taken for service. If a call enters the system and at that time $Q(t) = m+1$, then this call is lost.

To describe this system as a switching system consider a two-component MP $x(t) = (z(t), Q(t))$, $t \geq 0$, with state space $I \times \{0, \ldots, m+1\}$ and transition rates $a((i,j),(l,q))$, $i,l = \overline{1,d}$, $j,q = \overline{0,m+1}$, where

$$a\big((i,j),(l,j)\big) = a(i,l), \quad i,l = \overline{1,d},\ j = \overline{0,m+1};$$

$$a\big((i,j),(i,j+1)\big) = \lambda(i,j), \quad i = \overline{1,d},\ j = \overline{0,m};$$

$$a\big((i,j),(i,j-1)\big) = \mu(i,j), \quad i = \overline{1,d},\ j = \overline{1,m+1},$$

(other rates are zeros). The two-component process $x(t)$ stands for the switching component (or environment). Note that process $x(t)$ belongs to the class of so-called quasi-Birth-and-Death processes introduced by [NEU 89].

In addition, let $\zeta_k(t,(i,m+1))$ be a Poisson process with parameter $\lambda(i,m+1)$, $i = \overline{1,d}$, and $\zeta_k(t,(i,j)) \equiv 0$ as $j < m+1$.

We construct an SP $(x(t), \zeta(t))$, $t \geq 0$, using the Markov component $x(t)$ and processes $\zeta_k(\cdot)$ according to formula (1.14) (or (1.1)), where $x(t) = (z(t), Q(t))$ and $S_0 = 0$. This process is then a process with independent increments and Markov switching, see section 1.1.2, component $Q(t)$ is the value of the queue and $\zeta(t)$ is the number of calls lost in interval $[0,t]$.

Observing process $(x(t), \zeta(t))$ we can also calculate other characteristics of the system. Let $\nu^+(t)$ $(\nu^-(t))$ be the number of jumps up $(+1)$ (and down (-1) correspondingly) of the process $Q(t)$ in interval $[0,t]$. $\nu^+(t)$ is thus the number of calls which entered the system in interval $[0,t]$ and $\nu^-(t)$ is the number of calls served in this time interval.

Note that if the rate of input process $\lambda(i,j) \equiv \lambda(i)$ (depends only on the state of Markov environment), then the input process is usually called a Markov modulated input process [NEU 89] and in fact is a Poisson process with random rate $\lambda(z(t))$ or a doubly-stochastic Poisson process [COX 80].

2.2.1.3. *System* $\overline{M}_{\overline{Q},B}/\overline{M}_{\overline{Q},B}/1/\infty$

Consider a rather general Markov system which includes state-dependent systems with batch arrivals and service, systems with different types of calls, impatient calls, etc.

Let the non-negative functions $\lambda(\overline{q})$, $\mu(\overline{q})$, $\nu_i(\overline{q})$, $i = \overline{1,m}$, $\overline{q} \in \mathcal{R}_+^m$, be given. In addition, let $\overline{\alpha}(\overline{q})$, $\overline{\gamma}(\overline{q})$, $\overline{\beta}_i(\overline{q})$, $i = \overline{1,m}$, $\overline{q} \in \mathcal{R}_+^m$, be the random variables with values in $\mathcal{R}_+^m$. There is one server and an infinite number of waiting places. Assume that there are different types of calls or calls of different priorities. Denote by $\overline{Q}(t)$ the vector describing the number of calls of different types in the system at time t, $\overline{Q}(t) \in R_+^m$.

The system operates in the following way: if $\overline{Q}(t) = \overline{Q}$, then the local arrival rate is $\lambda(\overline{Q})$ and a batch of $\overline{\alpha}(\overline{Q})$ calls may enter the system. Correspondingly, the local service rate is $\mu(\overline{Q})$ and a batch of $\min\{\overline{\gamma}(\overline{Q}), \overline{Q}\}$ calls may complete service (in the case of vector-valued variables the minimum is taken in each component). In addition, each call of type i in the queue independently of the others may be transformed into $\overline{e}_i + \overline{\beta}_i(\overline{Q})$ calls with the local rate $\nu_i(\overline{Q})$, where $\overline{e}_i$ is a vector with the ith component which is equal to one and other components are equal to 0. Calls after service completion leave the system. If a vector $\overline{\beta}_i(\overline{Q})$ can have negative components (for instance, there are impatient calls), then after transformation we obtain $\max\{\overline{0}, \overline{Q} + \overline{\beta}_i(\overline{Q})\}$ calls in the system.

2.2.2. *Non-Markov systems*

2.2.2.1. *Semi-Markov system* $SM/M_{SM,Q}/1$

Consider a queueing system of the type $SM/M_{SM,Q}/1$ which is described in the following way. Let $x(t)$, $t \geq 0$, be a right continuous SMP with values in a measurable space (X, B_X) and let the functions $\mu(x,m)$, $x \in X$, $m = 0, 1, 2, \dots$ be given ($\mu(x,m)$ are measurable relatively σ-algebra B_X and stand for the local transition rates). Let $t_1 < t_2 < \cdots$ be a sequence of the times of jumps of $x(t)$. We say that the input flow is semi-Markov if the calls enter the system one at a time at the times t_k. The system has one server and the service rate at time t is $\mu(x(t), Q(t))$, where $Q(t)$ is the number of calls in the system at time t. After completion of service the calls leave the system.

To describe process $(x(t), Q(t))$ as an SP, we introduce the jointly independent families of stepwise decreasing MPs $\{\eta_k(t,x,m),\ t \geq 0,\ x \in X,\ m = 1, 2, \dots\}$, $k \geq 0$, with values in $\{0, 1, 2, \dots\}$ such that $\eta_k(0,x,m) = m$ for any x, m, k, the transition from state j is possible only to state $j-1$, and at small h,

$$\mathbf{P}\big\{\eta_k(t+h,x,m) = j-1 \mid \eta_k(t,x,m) = j\big\} = \mu(x,j)h + o(h),$$

where we assume that $\mu(x,0) \equiv 0$. In addition, let $\{(\tau_k(x), \beta_k(x)),\ x \in X\}$, $k \geq 0$, be the jointly independent families of random variables, which are independent of the introduced processes, defining the transitions of an SMP $x(t)$ in the following way:

$\tau_k(x) > 0$, $\beta_k(x) \in X$, and

$$\begin{aligned} P(t,x,A) &= \mathbf{P}\{\tau_k(x) < t,\ \beta_k(x) \in A\} \\ &= \mathbf{P}\{t_{k+1} - t_k < t,\ x(t_{k+1}) \in A \mid x(t_k) = x\}, \\ &\qquad\qquad t > 0,\ x \in X,\ A \in B_X,\ k \geq 0, \end{aligned} \tag{2.3}$$

where $P(t,x,A)$ is a transition probability for SMP $x(t)$. Then we introduce the families of processes $\zeta_k(t,x,m)$ in the following way:

$$\zeta_k(t,x,m) = \eta(t,x,m), \quad t < \tau_k(x),$$
$$\zeta_k(\tau_k(x),x,m) = \eta_k(\tau_k(x),x,m) + 1.$$

By definition process $(x(t), Q(t))$ is an SP which is defined by the families

$$\{(\zeta_k(t,x,m), \tau_k(x), \beta_k(x)),\ t \geq 0,\ x \in X,\ m = 0,1,\dots\}, \quad k \geq 0,$$

according to relations (1.3), (1.4). This process belongs to the class of Markov processes with semi-Markov switching.

By analogy we can describe a more complicated queueing system of the type $SM_Q/M_{SM,Q}/1/\infty$ where the input process depends on the values of the queue in the system and is constructed using independent families of random variables $\{\tau_k(x,m),\ x \in X,\ m \geq 0\}$, $k \geq 0$, and an MP x_k, $k \geq 0$, with values in X as follows. The calls enter the system one at a time. If a call enters the system at time t_k and the total number of calls in the system becomes Q, then the next call enters at time

$$t_{k+1} = t_k + \tau_k(x_k, Q).$$

The service in the system is provided in the same way as in the system $SM/M_{SM,Q}/1/\infty$.

In this case the two-component process $(x(t), Q(t))$ is an SP, but the component $x(t)$ itself is not a semi-Markov process and there is feedback between the input flow and the values of the queue.

Now we consider a queueing system where the input flow is a Poisson process modulated by an external semi-Markov environment and the values of the queue in the system.

2.2.2.2. *System* $M_{SM,Q}/M_{SM,Q}/1/\infty$

Let $x(t)$, $t \geq 0$, be an SMP with values in a measurable space X and functions $\lambda(x,m)$, $\mu(x,m)$, $x \in X$, $m = 0,1,2,\dots$ be given, where $\lambda(\cdot)$ and $\mu(\cdot)$ stand for the input and service rates, respectively. The input flow is a Poisson flow switched by process $x(t)$ and the value of the queue and is constructed in the following way. The calls enter the system one at a time and in the interval where $x(t) = x$ and $Q(t) = m$, the input rate is $\lambda(x, m)$ and the service rate is $\mu(x, m)$. The call, after service completion, leaves the system.

To describe process $(x(t), Q(t))$ as an SP we introduce the families of Birth-and-Death processes $\{\eta_k(t, x, m),\ t \geq 0,\ x \in X,\ m = 0, 1, 2, \ldots\}$, $k \geq 0$, such that $\eta_k(0, x, m) = m$ and in state j, the birth and death rates are $\lambda(x, j)$ and $\mu(x, j)$, respectively. Denote $\zeta_k(t, x, m) = \eta_k(t, x, m) - m$, $t \geq 0$, where we assume that $\mu(x, 0) \equiv 0$.

Then the processes $\zeta_k(t, x, m)$ together with a process $x(t)$ according to relations (1.14) determine a process with semi-Markov switching which is equivalent to process $(x(t), Q(t))$.

2.2.2.3. *System* $M_{SM,Q}/M_{SM,Q}/1/V$

Consider a similar system as above with the input calls and service portions of a random size depending on the current state of the system (queue size or total volume of the information in the system, etc.). The system has a buffer of a restricted capacity. Let $x(t)$, $t \geq 0$, be an SMP with state space X given by the embedded Markov process x_k, $k \geq 0$, and the family of sojourn times $\{\tau(x),\ x \in X\}$. Also let the jointly independent parametric families of non-negative random variables $\{\eta(x, \alpha),\ x \in X,\ \alpha \geq 0\}$ and $\{\kappa(x, \alpha),\ x \in X,\ \alpha \geq 0\}$, and the families of non-negative functions $\{\lambda(x, \alpha), \mu(x, \alpha), r(x, \alpha),\ x \in X,\ \alpha \geq 0\}$ be given where $0 \leq r(x, \alpha) \leq 1$.

An input flow is a Poisson flow modulated by process $x(t)$ and the current state of the buffer. The calls have a random size which can be interpreted as a volume of information, required amount of work, etc. The input flow is defined in the following way. Denote by $Q(t)$ the total volume of the information in the buffer at time t. If $x(t) = x$ and $Q(t) = q$, then the local input rate is $\lambda(x, q)$ and the size of incoming call is $\eta(x, q)$. At the fixed state of the system the sizes of different incoming calls are jointly independent. If a call of size $\eta(x, q)$ enters the system, then either with probability $1 - r(x, q)$ the total volume of information becomes $\min(q + \eta(x, q), V)$ (this size is added to the total volume of the information in the system up to threshold V) or with probability $r(x, q)$ the call is lost.

The system has one server with local service rate $\mu(x, q)$. Each time the server completes the service, it takes a random portion of the information of size $\min(Q(t), \kappa(x(t), Q(t)))$ and after completion of service this portion leaves the system. If the buffer is empty, the servers waits until some portion of information arrives at the buffer.

To describe process $(x(t), Q(t))$, $t \geq 0$, as a process with semi-Markov switching we introduce jointly independent families of stepwise MPs $\{\gamma_k(t, x, \alpha), t \geq 0, x \in X, \alpha > 0\}$, $k \geq 0$, with values in $[0, \infty)$ and distributions not dependent on index k. Process $\gamma_k(t, x, \alpha)$ is constructed as follows. For each x, α, k, $\gamma_k(0, x, \alpha) = \alpha$, and if at time t, $\gamma_k(t, x, \alpha) = s$, then with the local rate $\lambda(x, s) + \mu(x, s)$ the process can get a jump of the size $\beta(x, s)$ where

$$\beta(x,s) = \begin{cases} \min\{\eta(x,s), V - s\} & \text{with probability } \lambda(x,s)\big(\lambda(x,s) + \mu(x,s)\big)^{-1}, \\ -\min\{\kappa(x,s), s\} & \text{with probability } \mu(x,s)\big(\lambda(x,s) + \mu(x,s)\big)^{-1}. \end{cases}$$

We introduce the families of processes $\zeta_k(t, x, \alpha)$: $\zeta_k(t, x, \alpha) = \gamma_k(t, x, \alpha) - \alpha$, $t < \tau_k(x)$. Then process $(x(t), \widetilde{Q}(t))$ defined as an SP according to relations (1.3), (1.4) using the introduced families $\zeta_k(\cdot)$ and SMP $x(\cdot)$ is equivalent to process $(x(t), Q(t))$. This process is an MP with semi-Markov switching.

2.2.3. *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 non-negative function $\mu(\alpha) \geq 0$, $\alpha \geq 0$, and the independent families of non-negative random variables $\{\tau_k(\alpha), \alpha \geq 0\}$, $k \geq 0$, with distributions not depending on index k are given. The system operates as follows: the calls enter the system one at a time and are served according to the FIFO discipline. Denote by $Q(t)$ the total number of calls in the system at time 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 in the interval (t_k, t_{k+1}) is $\mu(q)$.

In this case we do not have a switching component x_k. Let us choose $\tau_k(q)$ as switching intervals and construct processes $\zeta_k(t, q)$ in interval $[0, \tau_k(q)]$ as follows:

$$\zeta_k(t, q) = -\min\big\{q, \Pi_k\big(t, \mu(q)\big)\big\} \quad \text{as } t < \tau_k(q),$$
$$\zeta_k\big(\tau_k(q), q\big) = 1 - \min\big\{q, \Pi_k\big(\tau_k(q), \mu(q)\big)\big\},$$

where $\Pi_k(t, \mu)$ are jointly independent Poisson processes with parameter μ. Then we can represent $Q(t)$ as an SP where formulae (1.3), (1.4) are modified as follows:

$$t_0 = 0,\ t_{k+1} = t_k + \tau_k\big(Q_k\big),\ Q_{k+1} = Q_k + \zeta_k\big(t_{k+1} - t_k, Q_k\big),\ k \geq 0,$$
$$Q(t) = Q_k + \zeta_k\big(t - t_k, Q_k\big) \text{ as } t_k \leq t < t_{k+1},\ t \geq 0. \tag{2.4}$$

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

2.2.4. *Polling systems*

Consider a system with r stations and a single moving server. An arrival flow at 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, $\overline{Q}(t) = (Q_1(t), \dots, Q_r(t))$. Let $\kappa_k(i)$ and $\kappa_k(i,j)$ be the random variables that are independent at different $k \geq 0, i, j = 1, \dots, r$, with distributions not depending on index k. If the server arrives at station i at time t_k and $\overline{Q}(t) = \overline{Q} = (Q_1, \dots, Q_r)$, then it occupies the station for time $\kappa_k(i)$ and the service rate in this period is $\mu_i(Q_i)$. All calls being served at the station during this period leave the system. After completing time $\kappa_k(i)$, the server with probability p_{ij} chooses station j, and it takes a random time $\kappa_k(i,j)$ to arrive at this station. During the travel time no service is provided. When the server arrives at station j, the service immediately starts 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, $k \geq 0$, the sequential times of arrivals of the server at any station ($t_0 = 0$). We construct process $x(t)$ in the following way: $x(t) = i$, $t_k \leq t < t_{k+1}$, if at time t_k the server arrived at station i. Then $x(t)$ is an SMP with the embedded Markov chain $x_k = x(t_k + 0)$ and occupation times constructed in the following way. Denote by $\widetilde{\kappa}_k(i)$ the random variables that are independent of $\kappa_k(i)$ such that

$$\mathbf{P}\big(\widetilde{\kappa}_k(i) \leq z\big) = \sum_j p_{ij} \mathbf{P}\big(\kappa_k(i,j) \leq z\big).$$

Thus $\mathbf{P}(t_{k+1} - t_k \leq z \mid x_k = i) = \mathbf{P}(\kappa_k(i) + \widetilde{\kappa}_k(i) \leq z)$, $k \geq 0$. Let $y_k(t,i,Q)$, $t \geq 0$, be the Birth-and-Death processes that are independent at different $k \geq 0$, $i = 1, \dots, r$, with initial value Q and constant birth rate λ_i and death rate $\mu_i(Q)$, respectively. In addition, let $\Pi_k(t,i,\lambda)$, $t \geq 0$, be the Poisson processes that are independent at different k, i, with parameter λ. We introduce process $\overline{\zeta}_k(t,i,\overline{Q}) = (\zeta_k^{(j)}(t,i,Q_j),\ j = \overline{1,r})$ in interval $[0, \kappa_k(i) + \widetilde{\kappa}_k(i)]$ as follows:

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

$$\zeta_k^{(i)}\big(t,i,Q_i\big) = y_k\big(\kappa_k(i),i,Q_i\big) - Q_i + \Pi_k\big(t - \kappa_k(i), i, \lambda_i\big)$$
$$\text{as } \kappa_k(i) < t \leq \kappa_k(i) + \widetilde{\kappa}_k(i);$$

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

Using the families $\{\overline{\zeta}_k(t,i,\overline{Q})\}$ and an SMP $x(t)$ we can construct an SP $(x(t), \overline{Q}(t))$, $t \geq 0$, according to relations (1.14). This process belongs to the class of Markov processes with semi-Markov switching and relations (1.14) in our case have the form:

$$\begin{aligned} &t_0 = 0, \quad \overline{Q}_{k+1} = \overline{Q}_k + \overline{\zeta}_k\big(t_{k+1} - t_k, x_k, \overline{Q}_k\big), \quad k \geq 0, \\ &\overline{Q}(t) = \overline{Q}_k + \overline{\zeta}_k\big(t - t_k, x_k, \overline{Q}_k\big) \quad \text{as } t_k \leq t < t_{k+1},\ t \geq 0. \end{aligned} \tag{2.5}$$

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 at 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 by the server at station i depends also on Q_i and is represented in the form: $\kappa(i, Q_i) = \sum_{1 \leq l \leq Q_i} \eta_l(Q_i)$, where $\eta_l(Q_i)$, $l \geq 1$, 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.

Note that in terms of SPs it is also possible to describe different classes of Markov and semi-Markov queueing systems and networks with unreliable servers and also some classes of retrial queues [ANI 99a, ANI 99b, ANI 01].

2.2.5. *Retrial queueing systems*

In retrial systems the calls finding the server busy may join a special retrial queue and repeat their attempts for service after a random time.

Consider as an example a system $M_Q/G/1/w.r$ with one server and an infinite number of waiting places. Calls enter the system one at a time. If the server is free, it immediately takes the call for service. If the server is busy, the call joins a special retrial queue (orbit) and repeats its attempts for service according to a random procedure described below.

Let $\lambda(q), \nu(q), q \geq 0$, be given non-negative functions. Denote by $Q(t)$ the number of calls in the orbit at time t. Let $t_1 < t_2 < \cdots$ be the sequential times of service completion, $t_0 = 0$. Denote $Q_k = Q(t_k + 0)$. We assume that in the interval $[t_k, t_{k+1})$ the input flow is a Poisson flow with parameter $\lambda(Q_k)$ and each call in the queue independently of other calls with local rate $\nu(Q_k)$ may re-apply for service. If the server is free, it takes the call for service. If the server is busy, the call remains in the queue and repeats its attempts for service in the same way. A service time κ does not depend on the type of call (if the call appears from the input flow or from the orbit) and has a general distribution function $B(x) = \mathbf{P}(\kappa \leq x)$. We denote this system as $M_Q/G/1/w.r$. This is a state-dependent retrial system. In particular, if $\lambda(q) \equiv \lambda$, $\nu(q) \equiv \nu$, we obtain a classical retrial system $M/G/1/w.r$ [FAL 90, FAL 97].

Let us represent the process $Q(t)$ as an SP. We choose times t_k as the switching times. Given that $Q_k = q$, denote $\tau_k(q) = t_{k+1} - t_k$. By definition

$$\mathbf{P}\big(\tau_k(q) \le x\big) = \mathbf{P}\big(\eta\big(\Lambda(q)\big) + \kappa \le x\big),$$

where $\eta(a)$ is an exponentially distributed random variable that is independent of κ with parameter a and $\Lambda(q) = \lambda(q) + q\nu(q)$. Further, denote $\xi_k(Q_k) = Q_{k+1} - Q_k$. We can then write a representation:

$$\begin{aligned}
&\mathbf{P}\big(\xi_k(Q_k) \le x \mid Q_k = q,\ \kappa = z\big) \\
&\quad = \big(\lambda(q) + q\nu(q)\big)^{-1}\big(q\nu(q)\mathbf{P}\big(\Pi_{\lambda(q)}(z) - 1 \le x\big) + \lambda(q)\mathbf{P}\big(\Pi_{\lambda(q)}(z) \le x\big)\big) \\
&\quad = \mathbf{P}\big(\Pi_{\lambda(q)}(z) - \chi(q) \le x\big),
\end{aligned}$$

where $\Pi_b(t)$ is a Poisson process with parameter b, and $\chi(q)$ is a Bernoulli random variable, $\mathbf{P}(\chi(q) = 1) = 1 - \mathbf{P}(\chi(q) = 0) = q\nu(q)/\Lambda(q)$.

Let us define the family of processes $\zeta_k(t,q)$, $t \ge 0$, $q \in \{0,1,2,\ldots\}$, $k = 0,1,2,\ldots$, as follows: $\zeta_k(t,q)$ are jointly independent at different k processes with distributions not depending on index k, and

$$\zeta_1(t,q) = \begin{cases} 0, & t < \eta(q) \\ \Pi_{\lambda(q)}\big(t - \eta(q)\big) & \eta(q) \le t < \eta(q) + \kappa, \\ \Pi_{\lambda(q)}(\kappa) - \chi(q) & t = \eta(q) + \kappa, \end{cases} \tag{2.6}$$

Then this family of processes jointly with switching times t_k define an SP $\widetilde{Q}(t)$ according to relations (1.3), (1.4), where component x_k and variables $\beta_k(\cdot)$ are absent, and the jump value at switching time is $\xi_1(x,q) = \Pi_{\lambda(q)}(\kappa) + \chi(q)$. The process $\widetilde{Q}(t)$ is equivalent to retrial queue $Q(t)$.

Note that for the system above the input and service rates in each switching interval depend on the value of $Q(t_k + 0)$ just after the switching time. In a similar way we can represent a retrial queueing model where the input and service rates at any time depend on the current value of queue $Q(t)$.

2.3. Queueing networks

A stochastic queueing network in a usual understanding consists of several nodes each of which works as a queueing system. Calls arrive at different nodes according to a multivariate process. Nodes might be of several types: the nodes taking the incoming calls, the nodes responsible for service and the nodes directing the calls which are completed service out of the network.

Let the network consist of r nodes. Usually it is assumed that the call after service completion at node i either with probability p_{ij} passes to node j, $j = 1, \ldots, r$, or with probability $p_{i,0}$ leaves the network. Matrix $P = \|p_{ij}\|_{i=1,\ldots,r;\ j=0,\ldots,r}$ is called a routing matrix. Note that P is a stochastic matrix as in each node i, $\sum_{j=0}^{r} p_{ij} = 1$.

The network is defined if each node is described as a queueing system and the routing matrix is given. For example, if each node is working as $\cdot/M/\overline{m}/s$ system and the input process is a multidimensional Poisson process, then according to Kendall's classification we denote this network as $(M/M/\overline{m}/s)^r$. This means that the input at node i is a Poisson input with rate λ_i, there are m_i servers and s_i waiting places, and after service completion the call either with probability p_{ij} is immediately sent to node j, $j = 1, \ldots, r$, or with probability p_{i0} leaves the network. The calls entering node i (from the input process or from other nodes) are served according to the FIFO discipline. If a new call enters node i and there are already $m_i + s_i$ calls (all servers and waiting places are busy), then this call is lost. The networks of type $(GI/G/\overline{m}/s)^r$ can be defined in a similar way.

A network is called closed if for all i, $p_{i0} = 0$ and there are no input calls. Otherwise, the network is called open. A standard Jackson network means that input flows in all nodes are Poisson flows, calls are identical and service times are exponential.

In the following sections we consider a representation of queueing processes for different classes of state-dependent Markov and non-Markov networks in terms of SPs. This representation plays a key role in the forthcoming sections at the investigation of limit theorems of averaging principle and diffusion approximation type for normalized queueing processes.

2.3.1. *Markov state-dependent networks*

In this section we consider different sub-classes of Markov state-dependent networks.

2.3.1.1. *Markov network* $(M_Q/M_Q/\overline{m}/\infty)^r$

Consider a queueing network $(M_Q/M_Q/\overline{m}/\infty)^r$ which consists of r nodes. There are m_i servers in each node and an infinite number of waiting places. Denote a number of calls in the ith node at time t by $Q(i,t)$ and let $\overline{Q}(t) = (Q(i,t),\ i = \overline{1,r})$ be a column vector. Let the family of functions $\{\lambda_i(\overline{q}), \mu_i(\overline{q})\}$, the family of stochastic matrices $\|p_{ij}(\overline{q})\|_{i=\overline{1,r},\ j=\overline{0,r}}$, $\overline{q} \in [0,\infty)^r$ and the initial vector $\overline{Q}(0)$ also be given.

The network operates as follows: if $\overline{Q}(t) = \overline{q}$, then at time t a local input rate to the ith node is $\lambda_i(\overline{q})$ and the local service rate at each busy server is $\mu_i(\overline{q})$. After service completion in node i a call either with probability $p_{ij}(\overline{q})$ enters jth node, $j = \overline{1,r}$, or with probability $p_{i0}(\overline{q})$ leaves the network.

In this case the process $\overline{Q}(t)$, $t \geq 0$, is a multi-dimensional Markov process and it can be described as a quasi-Birth-and-Death process where level j corresponds to the total number of calls in the network [NEU 89].

2.3.1.2. *Markov networks* $(M_{Q,B}/M_{Q,B}/\overline{m}/\infty)^r$ *with batches*

Consider a queueing network $(M_{Q,B}/M_{Q,B}/\overline{m}/\infty)^r$ with batch state-dependent Markov arrival process and service. It consists of r nodes with m_i servers at node i and an infinite number of waiting places. Denote by $Q(i,t)$ a number of calls in node i at time t and put $\overline{Q}(t) = (Q(i,t),\ i = \overline{1,r})$. Let there be given:

1) a family of non-negative functions $\{\lambda_i(\overline{q}), \mu_i(\overline{q}), \nu_i(\overline{q})\}$, $i = \overline{1,r}$, where $\overline{q} = (q_1, \ldots, q_r)$;

2) the families of integer random variables $\{\delta_i(\overline{q}), \gamma_i(\overline{q})\}$ with values in $\{0, 1, \ldots\}$ and the 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 $\overline{Q}(0)$.

The system operates as follows. If at time t, $\overline{Q}(t) = \overline{q}$, 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 $q_i\mu_i(\overline{q})$, $\min\{\gamma_i(\overline{q}), q_i\}$ calls may complete service at node i and all of them either with probability $p_{ij}(\overline{q})$ are passed 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 $\nu_i(\overline{q})$, may be transformed into $\max\{\beta_i(\overline{q}), 1 - q_i\}$ calls, $i = \overline{1,r}$.

In this case process $\overline{Q}(t)$, $t \geq 0$, is also a multidimensional MP. However if the size of batches is more than one with positive probability, then this process is not a quasi-Birth-and-Death process.

Markov models of the type $(M_{M,Q}/M_{M,Q}/m_i/s_i)^r$ with local input and service rates depending on the current state of the queueing processes in the nodes and on the state of some external Markov environment can also be described as multidimensional Markov processes that are a particular case of an MP with Markov switching.

In the next section we consider more general state-dependent queueing networks with input and service batches of a random size switched by some semi-Markov environment.

2.3.2. *Non-Markov networks*

2.3.2.1. *State-dependent semi-Markov networks*

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 the models considered in section 2.3.1. Suppose that there are r nodes and one server at each node with infinite buffer.

Let $x(t)$, $t \geq 0$, be an SMP with state space $X = \{1, 2, \dots, d\}$ which stands for the external environment. Let the families of non-negative 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 non-negative random variables $\{\kappa_j(x, \overline{q}),\ x \in X,\ j = \overline{1, r}\}$ also be given (here $\overline{q} \in \mathcal{R}_+^r$).

Denote by $Q_i(t)$ the total amount of work in the buffer at node i at time t and put $\overline{Q}(t) = (Q_1(t), \dots, Q_r(t))$. 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 vector $\overline{\eta}(x, \overline{q})$ enters node i). Correspondingly, with local service rate $\mu_i(x, \overline{q})$ a random portion of work of 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 $p_{i,r+1}(x, \overline{q})$ leaves the network. Here we may assume for simplicity that $\mu_i(x, \overline{q}) \equiv 0$ if $q_i = 0$, where $\overline{q} = (q_1, \dots, q_r)$.

To describe process $(x(t), \overline{Q}(t))$, $t \geq 0$, in the network as an SP, we introduce the independent families of multi-dimensional MPs $\{\overline{\gamma}_k(t, x, \overline{q}),\ t \geq 0,\ x \in X, \overline{q} \in R_+^r\}$, $k \geq 0$, 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

$$\overline{\delta}(x,\overline{s}) = \begin{cases} \overline{\eta}(x,\overline{s}), & \text{with pr. } \lambda(x,\overline{s})\Lambda(x,\overline{s})^{-1}, \\ \big(-\overline{e}_i + \overline{e}_j\big)\widetilde{\kappa}_i(x,\overline{s}), & \text{with pr. } \mu_i(x,\overline{s})p_{ij}(x,\overline{s})\Lambda(x,\overline{s})^{-1}, \qquad i,j = \overline{1,r}. \\ -\overline{e}_i\widetilde{\kappa}_i(x,\overline{s}), & \text{with pr. } \mu_i(x,\overline{s})p_{i,r+1}(x,\overline{s})\Lambda(x,\overline{s})^{-1}, \end{cases} \tag{2.7}$$

Now we construct the family of processes $\overline{\zeta}_k(t, x, \overline{q})$, $t \geq 0$, in the following way. Let at each $x \in X$, $\tau_k(x)$, $k \geq 0$, be a sequence of iidrv 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 $\overline{\zeta}_k(t, x, \overline{q}) = \overline{\gamma}_k(t, x, \overline{q}) - \overline{q}$, $0 \leq t \leq \tau_k(x)$. We choose switching times t_k as times of sequential jumps of $x(t)$. Then process $(x(t), \widetilde{\overline{Q}}(t))$, $t \geq 0$, which is constructed using introduced processes $\overline{\zeta}_k(\cdot)$ and an SMP $x(\cdot)$ according to section 1.2.5, is a process with semi-Markov switching. This process is equivalent to process $(x(t), \overline{Q}(t))$, $t \geq 0$. For this case, an arrival process may be called a semi-Markov modulated Poisson process by analogy to a Markov modulated arrival process [NEU 89].

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.

2.3.2.2. *Semi-Markov networks with random batches*

Now consider a more complicated non-Markov network of the type $(M_{SM,Q}/M_{SM,Q}/1/\infty)^r$ which in some sense is a generalization of the system described in section 2.3.1.2. The network consists of r nodes with one server in each node. Suppose for simplicity that each node has an infinite capacity.

Let SMP $x(t)$ with values in a space X, parametric families of non-negative functions $\{\lambda(x,\bar{q}), \mu_i(x,\bar{q}), p_{ij}(x,\bar{q}),\ i=\overline{1,r},\ j=\overline{0,r},\ x\in X,\ \bar{q}\in R_+^r\}$, independent parametric families of random vectors $\{\bar{\gamma}(x,\bar{q}),\ x\in X,\ \bar{q}\in R_+^r\}$ and random variables $\{\kappa_i(x,\bar{q}),\ i=\overline{1,r},\ x\in X,\ \bar{q}\in R_+^r\}$ with values in R_+^r and R_+ respectively be given. Here $\sum_{j=0}^r p_{ij}(x,\bar{q}) = 1$ for any $i,x,\bar{q}$. Denote by $Q_i(t)$ the total volume of the information or total amount of work in the ith node at time t and put $\bar{Q}(t) = (Q_1(t),\ldots,Q_r(t))$.

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 $\overline{\gamma}(x,\overline{q})$ may enter the system (the ith component of the vector $\overline{\gamma}(x,\overline{q})$ 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 $(x(t),\bar{Q}(t))$ in the network as an SP, we introduce the independent families of multi-dimensional Markov processes $\{\bar{\gamma}_k(t,x,\bar{q}),\ t\geq 0,\ x\in X,\ \bar{q}\in R_+^r\}$, $k\geq 0$ with values in R_+^r and distributions not depending on k in the following way: $\bar{\gamma}_k(0,x,\bar{q}) = \bar{q}$ and if at time t, $\bar{\gamma}_k(t,x,\bar{q}) = \bar{s}$, then the process $\bar{\gamma}_k(t,x,\bar{q})$ can make a jump of the size $\bar{\beta}(x,\bar{s})$ with the local rate $\Lambda(x,\bar{s}) = \lambda(x,\bar{s}) + \sum_{i=1}^r \mu^{(i)}(x,\bar{s})$, where:

$$\bar{\beta}(x,\bar{s})$$
$$= \begin{cases} \bar{\gamma}(x,\bar{s}) & \text{with probability } \lambda(x,\bar{s})\Lambda(x,\bar{s})^{-1}, \\ \big(-e_i+e_j\big)\tilde{\kappa}_i(x,\bar{s}) & \text{with probability } \mu^{(i)}(x,\bar{s})p_{ij}(x,\bar{s})\Lambda(x,\bar{s})^{-1},\ i=\overline{1,r}, \\ -e_i\tilde{\kappa}_i(x,\bar{s}) & \text{with probability } \mu^{(i)}(x,\bar{s})p_{i0}(x,\bar{s})\Lambda(x,\bar{s})^{-1}, \end{cases}$$

where we denote by e_i a column-vector in R^r with the ith component equal to one and the other components equal to zero.

Let us introduce now the family of processes $\{\overline{\zeta}_k(t,x,\overline{q}),\ t\geq 0\}$ in the following way. Let at each $x\in X$, $\tau_k(x)$, $k\geq 0$, be a sequence of iidrv having the same distribution as the sojourn time $\tau(x)$ in state x. Then $\overline{\zeta}_k(t,x,\overline{q})$ is defined on interval $[0,\tau_k(x)]$, and $\overline{\zeta}_k(t,x,\overline{q}) = \overline{\gamma}_k(t,x,\overline{q}) - \overline{q}$, $0\leq t\leq \tau_k(x)$. We choose switching times t_k as times of sequential jumps of $x(t)$. Then process $\{(x(t),\overline{\widetilde{Q}}(t));\ t\geq 0\}$, which is constructed using introduced processes $\overline{\zeta}_k(\cdot)$ and an SMP $x(\cdot)$ in section 1.2.5 according to (1.14), is a process with semi-Markov switching (PSMS). This process is equivalent to process $\{(x(t), \overline{Q}(t));\ t \geq 0\}$ in the network.

For this case, an arrival process may be called a semi-Markov modulated Poisson process by analogy with the Markov modulated arrival process [NEU 89].

If we add an additional node $r+1$ and consider it as an accumulating node for the output process $Z(t)$ (total number of calls that left the network in the interval $[0,t]$), then in the same way the process $(x(t), \overline{Q}(t), Z(t))$ can be described as a PSMS.

By analogy we can also describe the network $(SM/M_Q/1/\infty)^r$ with semi-Markov input (calls enter the system at times of jumps of some SMP), consider different classes of calls, priority service, etc.

These examples show that various classes of stochastic queueing networks operating in a random environment can be described in terms of SP.

2.3.2.3. *Networks with state-dependent input*

Consider a network $(G_Q/M_Q/1/\infty)^r$ where the input flow may depend on the current state of the network. Let there be r nodes with one server in each node and an infinite number of waiting places. In addition, let $Q(i,t)$ be the number of calls in the ith node at time t and $\overline{Q}(t) = (Q(i,t),\ i = \overline{1,r})$. Assume that functions $\mu_i(\overline{q})$, probabilities $q_i(\overline{q})$ and $p_{ij}(\overline{q})$, $i = \overline{1,r}$, $j = \overline{0,r}$, $\overline{q} \in R$, and the jointly independent family of non-negative variables $\{\tau_k(\overline{q})\}$, $k \geq 0$, with distributions not dependent on index k are given, where for any $i = \overline{1,r}, \overline{q} \in R_+^r$, $\sum_{j=1}^r p_{ij}(\overline{q}) = \sum_{j=1}^r q_j(\overline{q}) = 1$. The network operates in the following way: if a call enters the system at time t_k and $\overline{Q}(t_k - 0) = \overline{Q}$, then this call with probability $q_j(\overline{Q})$ enters the jth node, the next call enters the system at time $t_{k+1} = t_k + \tau_k(\overline{Q})$, and the service rate in this period in the ith node is $\mu_i(\overline{Q})$, $i = \overline{1,r}$. In addition, if a call in this interval completes its service in the ith node, then this call with probability $p_{ij}(\overline{Q})$ goes to the jth node, $j = \overline{1,r}$, and with probability $p_{i0}(\overline{Q})$ leaves the network.

To describe process $\overline{Q}(t)$ as an SP, note that in this case there is no switching component x_k. We consider $\tau_k(\overline{Q})$ as switching intervals. Denote for any fixed vector $\overline{\mu} = (\mu_1, \ldots, \mu_r)$ and stochastic matrix $P = \|p_{ij}\|_{i=1,\ldots,r,\ j=0,\ldots,r}$ by $(\cdot/M_\mu/1/\infty/P)^r$ a queueing network without input flow, with service rate in ith node μ_i and routing matrix P. This means that a call after completion of service in the ith node with probability p_{ij} goes to the jth node, and with probability p_{i0} leaves the network. Note that for any fixed $\overline{\mu}$ and P, a network $(\cdot/M_\mu/1/\infty/P)^r$ is a Markov network.

Assume that the initial number of calls in this network is a vector $\overline{q}$ and denote by $\xi_k(t, \overline{q}, \overline{\mu}, P)$ the number of calls in the network in node k at time t. Put $\overline{\xi}(t, \overline{q}, \overline{\mu}, P) = (\xi_1(t, \overline{q}, \overline{\mu}, P), \ldots, \xi_r(t, \overline{q}, \overline{\mu}, P))$, this is a multidimensional MP.

Let us define the family of processes $\overline{\zeta}_k(t,q)$ as follows: in interval $0 \leq t < \tau_k(\overline{q})$, $\overline{\zeta}(t,\overline{q}) = \overline{\xi}(t,\overline{q},\overline{\mu},P) - \overline{q}$, where $\overline{\mu} = (\mu_1(\overline{q}),\ldots,\mu_r(\overline{q}))$ and $P = P(\overline{q})$. When $t = \tau_k(\overline{q})$, then $\overline{\zeta}(\tau_k(\overline{q})) = \overline{\zeta}(\tau_k(\overline{q})-) + \overline{\delta}(\overline{\zeta}_-(\overline{q})) - \overline{q}$, where $\overline{\delta}(\overline{z})$ is a vector with jth component equal to one with probability $q_j(\overline{z})$ and other components are equal to zeros, and $\overline{\zeta}_-(\overline{q}) = \overline{\zeta}(\tau_k(\overline{q})-)$. We can thus represent $\overline{Q}(t)$ as a special case of an SP according to relations (1.6), (1.7).

2.4. Bibliography

[ANI 90] ANISIMOV V. and ALIEV A., "Limit theorems for recurrent processes of semi-Markov type", *Theor. Prob. and Math. Stat.*, vol. 41, p. 7–13, 1990.

[ANI 92a] ANISIMOV V., "Averaging principle for switching processes", *Theor. Probab. and Math. Stat.*, vol. 46, p. 1–10, 1992.

[ANI 92b] ANISIMOV V. and LEBEDEV E., *Stochastic Queueing Networks. Markov Models*, Kiev University (Russian), Kiev, Ukraine, 1992.

[ANI 94] ANISIMOV V., "Limit theorems for processes with semi-Markov switching and their applications", *Random Oper. and Stoch. Eqv.*, vol. 2, no. 4, p. 333–352, 1994.

[ANI 95] ANISIMOV V., "Switching processes: averaging principle, diffusion approximation and applications", *Acta Applicandae Mathematicae*, vol. 40, p. 95–141, 1995.

[ANI 99a] ANISIMOV V., "Averaging methods for transient regimes in overloading retrial queuing systems", *Mathematical and Computing Modelling*, vol. 30, no. 3/4, p. 65–78, 1999.

[ANI 99b] ANISIMOV V., "Switching stochastic models and applications in retrial queues", *Top*, vol. 7, no. 2, p. 169–186, 1999.

[ANI 01] ANISIMOV V. and ARTALEJO J., "Analysis of Markov multiserver retrial queues with negative arrivals", *Queueing Systems*, vol. 39, no. 2/3, p. 157–182, 2001.

[BOC 03] BOCHAROV P., D'APICE C., PECHINKIN A. and SALERNO S., "The stationary characteristics of the $G/MSP/1/r$ queueing system", *Automation and Remote Control*, vol. 64, no. 2, p. 288–301, 2003.

[COX 80] COX D. and ISHAM V., *Point Processes*, Chapman & Hall, London, 1980.

[FAL 88a] FALIN G., "Analysis of buffered floating-threshold hybrid switching system", *Problems of Information Transmission*, vol. 24, no. 4, p. 318–328, 1988.

[FAL 88b] FALIN G., "Single server queueing system with randomly varying service rate", *Engineering Cybernetics*, vol. 26, no. 1, p. 134–137, 1988.

[FAL 90] FALIN G., "A survey of retrial queues", *Queueing systems*, vol. 7, p. 127–168, 1990.

[FAL 97] FALIN G. and TEMPLETON J., *Retrial Queues*, Chapman & Hall, London, 1997.

[FIS 93] FISHER W. and MEIER-HELLSTERN K., "The Markov-modulated Poisson process (MMPP) cookbook", *Performance Evaluation*, vol. 18, p. 149–171, 1993.

[GEL 90] GELENBE E. and ROSENBERG C., "Queues with slowly varying arrival and service processes", *Management Science*, vol. 36, no. 8, p. 928–937, 1990.

[HE 96] HE Q., "Queues with marked customers", *Advances in Applied Probability*, vol. 28, p. 567–587, 1996.

[KIM 07a] KIM C., DUDIN A., KLIMENOK V. and KHRAMOVA V., "Erlang loss queueing system with batch arrivals operating in a random environment", *Computers & Operations Research*, forthcoming, 2007.

[KIM 07b] KIM C., KLIMENOK V., SANG C. and DUDIN A., "The BMAP/PH/1 retrial queueing system operating in random environment", *Journal of Statistical Planning and Inference*, vol. 137, p. 3904–3916, 2007.

[LUC 94a] LUCANTONI D., CHOUDHURY G. and WHITT W., "The transient BMAP$/G/1$ queue", *Comm. Statist. Stochastic Models*, vol. 10, no. 1, p. 145–182, 1994.

[LUC 94b] LUCANTONI D. and NEUTS M., "Some steady-state distributions for the $MAP/SM/1$ queue", *Comm. Statist. Stochastic Models*, vol. 10, no. 3, p. 575–598, 1994.

[NEU 81] NEUTS M., *Matrix-Geometric Solutions in Stochastic Models*, John Hopkins University Press, Baltimore, 1981.

[NEU 89] NEUTS M., *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York & Basel, 1989.

[O'C 86] O'CINNEIDE C. and PURDUE P., "The $M/M/\infty$ queue in a random environment", *Journal of Applied Probability*, vol. 23, no. 1, p. 175–184, 1986.

[PUR 74] PURDUE P., "The $M/M/1$ queue in a Markovian environment", *Operations Research*, vol. 22, p. 562–569, 1974.

[SZT 87] SZTRIK J., "On the heterogeneous M/G/n blocking system in a random environment", *Journal of Operational Research Society*, vol. 38, no. 1, p. 57–63, 1987.

Chapter 3

Processes of Sums of Weakly-dependent Variables

In this chapter we consider limit theorems for stepwise random processes constructed by the sums of independent random variables defined on a random sequence satisfying some form of mixing condition in a triangular scheme. Special attention is given to the case when a switching sequence is a Markov process (homogenous or inhomogenous in time). A special notion of quasi-ergodic non-homogenous in time Markov processes is introduced [ANI 83, ANI 88]. The processes have the property that their one-step transition probabilities (or transition rates) vary slowly in time, and it is proved that the sums of random variables switched by a quasi-ergodic process converge to processes with independent increments with the local characteristics averaged by local stationary distribution (quasi-ergodic distribution) of a quasi-ergodic process.

The results of this chapter are used in the following chapters in the study of asymptotic behavior of a special type of switching processes and applications to switching queueing systems.

3.1. Limit theorems for processes of sums of conditionally independent random variables

In this section we study in a triangular scheme J-convergence of stepwise processes constructed by the sums of independent random variables on a random sequence to the processes with independent increments assuming that the sequence satisfies a uniform mixing condition. The exposition partially follows [ANI 83, ANI 88].

Let x_{nk}, $k \geq 0$, be a random sequence with values in a measurable space (X, B_X) and let

$$F_{nk} = \{\xi_{nk}(x),\ x \in X\}, \quad k \geq 0,$$

be the families of jointly independent and independent of $\{x_{nl},\ l \geq 0\}$ random variables with values in R^r. We suppose that at each k function $\xi_{nk}(x, \omega)$, $\omega \in \Omega$ ((Ω, B_Ω) is the original probability space) is measurable in the pair (x, ω) with respect to $B_X \times B_\Omega$ (i.e., $\xi_{nk}(x_{nk})$ is a random variable). Consider the sequence of sums

$$S_n(m) = \sum_{k=0}^{m} \xi_{nk}\big(x_{nk}\big), \quad m \geq 0,$$

and the sequence of stepwise random processes

$$\zeta_n(t) = S_n\big([nt]\big), \quad t \geq 0,$$

where $[a]$ denotes the integer part of value a.

Denote by $\mathcal{F}_m^k$ a σ-algebra generated by the random variables $\{x_{nl},\ m \leq l \leq k\}$ and let

$$\varphi_n(k, j) = \sup_{A \in F_o^k,\ B \in F_j^\infty} \big|\mathbf{P}_n(B \mid A) - \mathbf{P}_n(B)\big|, \quad k < j, \tag{3.1}$$

be the uniformly strong mixing coefficient ($\mathbf{P}_n(\cdot)$ is the measure in (Ω, B_Ω) induced by $x_n(\cdot)$). Denote

$$a_n(\lambda, k, x) = \ln \mathbf{E} \exp\big\{i\big(\lambda, \xi_n(k, x)\big)\big\}, \quad x \in X,\ \lambda \in R^r,\ k \geq 0,$$

$$\eta_n(\lambda, k) = a_n\big(\lambda, k, x_{nk}\big), \quad \alpha_n(\lambda, k) = \mathbf{E}\eta_n(\lambda, k),$$

$$\beta_n^2(\lambda, k) = \mathbf{E}\big|\eta_n(\lambda, k) - \alpha_n(\lambda, k)\big|^2, \quad k \geq 0,$$

$$A_n(\lambda, t) = \sum_{k=0}^{[nt]} \alpha_n(\lambda, k), \quad 0 \leq t \leq T,$$

where it is assumed that the corresponding expressions are defined for all $\lambda \in R^r$. In addition, let $\tilde{\xi}_{nk}$, $k \geq 0$ be the sequence on jointly independent random variables such that

$$\ln \mathbf{E} \exp\big\{i\big(\lambda, \tilde{\xi}_{nk}\big)\big\} = \alpha_n(\lambda, k), \quad \lambda \in R^r,\ k \geq 0.$$

We introduce the following stepwise process with independent increments

$$\tilde{\zeta}_n(t) = \sum_{k=0}^{[nt]} \tilde{\xi}_{nk}, \quad 0 \leq t \leq T,$$

that in some sense is the process approximating the behavior of $\zeta_n(t)$, and put

$$q_n(\lambda) = \sum_{k=0}^{[nT]} \beta_{nk}^2(\lambda, k) + 8 \sum_{0 \leq k < j \leq [nT]} \sqrt{\varphi_n(k, j)}\beta_n(\lambda, k)\beta_n(\lambda, j).$$

THEOREM 3.1. *If for any $\lambda \in R^r$, as $n \to \infty$,*

$$q_n(\lambda) \longrightarrow 0, \tag{3.2}$$

then for any integer $s > 0$, $\lambda_j \in R^r$, $j = \overline{1, s}$, as $n \to \infty$,

$$\sup_{0 \le t_j \le T,\ j=\overline{1,s}} \left| \mathbf{E} \exp\left\{ i \sum_{j=1}^{s} \big(\lambda_j, \zeta_n(t_j)\big) \right\} - \mathbf{E} \exp\left\{ i \sum_{j=1}^{s} \big(\lambda_j, \tilde{\zeta}_n(t_j)\big) \right\} \right| \longrightarrow 0. \tag{3.3}$$

(finite dimensional distributions of processes $\zeta_n(t)$ and $\tilde{\zeta}_n(t)$ are equivalent).

If in addition there is a sequence N_n such that, as $n \to \infty$:

$$\begin{aligned} &N_n \longrightarrow \infty, \quad n^{-1} N_n \longrightarrow 0, \\ &\max_{k \le nT} \varphi_n\big(k, k + N_n\big) \longrightarrow 0, \end{aligned} \tag{3.4}$$

at each $\lambda \in R^r$,

$$\max_{l \le nT} \sup \left\{ \left| \sum_{k=1}^{N_n} a_n(\lambda, k + l, y_k) \right| : y_k \in X,\ k = \overline{1, N_n} \right\} \longrightarrow 0, \tag{3.5}$$

$$\lim_{h \to +0} \limsup_{n \to \infty} \max_{k \le nT} \max_{u \le h} \left| \sum_{l=0}^{[nu]} \alpha_n(\lambda, k + l) \right| = 0, \tag{3.6}$$

and for any $t \in [0, T]$,

$$\lim_{\lambda \to 0} \limsup_{n \to \infty} \big| A_n(\lambda, t) \big| = 0, \tag{3.7}$$

then for each bounded functional $f(\cdot)$, continuous with respect to the convergence in Skorokhod J-topology in the space $\mathcal{D}^r_{[0,T]}$,

$$\mathbf{E} f\big(\zeta_n(\cdot)\big) - \mathbf{E} f\big(\tilde{\zeta}_n(\cdot)\big) \longrightarrow 0. \tag{3.8}$$

Proof. Since variables $\xi_{nk}(x_{nk})$ are conditionally independent at the fixed sample trajectory x_{nk}, the following representation holds:

$$\begin{aligned} \mathbf{E} \exp\big\{ i\big(\lambda, S_n(m)\big) \big\} &= \mathbf{E} \prod_{k=0}^{m} \mathbf{E}\big[e^{i(\lambda, \xi_{nk}(x_{nk}))} \mid x_{nk} \big] \\ &= \mathbf{E} \prod_{k=0}^{m} e^{a_n(\lambda, k, x_{nk})} = \mathbf{E} \exp\big\{ \Lambda_n(\lambda, m) \big\}, \end{aligned} \tag{3.9}$$

where $\Lambda_n(\lambda, m) = \sum_{k=0}^{m} \eta_n(\lambda, k)$, and also $\operatorname{Re} \eta_n(\lambda, k) \leq 0$ for each $\lambda \in R^r$, $k \geq 0$, with probability one, where $\operatorname{Re} a$ is the real part of complex number a.

Now we use the well-known inequality ([BIL 68], Lemma 1 in section 20): if the real variables ξ and η are F_0^k and F_j^∞-measurable, respectively, $k < j$, then

$$|\mathbf{E}\xi\eta - \mathbf{E}\xi\mathbf{E}\eta| \leq 2\sqrt{\varphi_n(k,j)}\sqrt{\mathbf{E}|\xi|^2}\sqrt{\mathbf{E}|\eta|^2}.$$

If variables ξ and η take complex values, then from inequality $\sqrt{\mathbf{E}(\operatorname{Re} z)^2} + \sqrt{\mathbf{E}(\operatorname{Im} z)^2} \leq \sqrt{2}\sqrt{\mathbf{E}|z|^2}$, that is true for any complex z, we obtain

$$|\mathbf{E}\xi\eta - \mathbf{E}\xi\mathbf{E}\eta| \leq 4\sqrt{\varphi_n(k,j)}\sqrt{\mathbf{E}|\xi|^2}\sqrt{\mathbf{E}|\eta|^2}.$$

Taking into account this inequality it is not hard to show that as $m \leq nT$,

$$\begin{aligned}
\mathbf{E}\big|\Lambda_n(\lambda,m) - A_n(\lambda,m)\big|^2 &= \mathbf{E}\left|\sum_{k=0}^{m}\big(\eta_n(\lambda,k) - \mathbf{E}\eta_n(\lambda,k)\big)\right|^2 \\
&\leq \sum_{k=0}^{m} \mathbf{E}\big|\eta_n(\lambda,k) - \mathbf{E}\eta_n(\lambda,k)\big|^2 \\
&\quad + 2\sum_{0=k<j\leq m} \mathbf{E}\big|\eta_n(\lambda,k) - \mathbf{E}\eta_n(\lambda,k)\big|\big|\eta_n(\lambda,j) - \mathbf{E}\eta_n(\lambda,j)\big| \leq q_n(\lambda).
\end{aligned}$$

Furthermore, from the relation $\int_0^1 \exp\{x + t(y-x)\}\mathrm{d}t = (y-x)^{-1}(e^y - e^x)$, which is true for any complex x, y, it follows that as $\operatorname{Re} x \leq 0$, $\operatorname{Re} y \leq 0$,

$$\big|e^y - e^x\big| \leq |x - y|. \tag{3.10}$$

Since $\operatorname{Re} a_n(\lambda,k,x) \leq 0$, and $\operatorname{Re} \eta_n(\lambda,k) \leq 0$ with probability one, it follows that for any $t \geq 0$,

$$\begin{aligned}
&\big|\mathbf{E}\exp\big\{i\big(\lambda, \zeta_n(t)\big)\big\} - \mathbf{E}\exp\big\{i\big(\lambda, \tilde{\zeta}_n(t)\big)\big\}\big| \\
&\quad \leq \mathbf{E}\big|\exp\big\{\Lambda_n\big(\lambda,[nt]\big)\big\} - \exp\big\{A_n\big(\lambda,[nt]\big)\big\}\big| \\
&\quad \leq \mathbf{E}\big|\Lambda_n\big(\lambda,[nt]\big) - A_n\big(\lambda,[nt]\big)\big| \\
&\quad \leq \Big(\mathbf{E}\big|\Lambda_n\big(\lambda,[nt]\big) - A_n\big(\lambda,[nt]\big)\big|^2\Big)^{1/2} \leq \sqrt{q_n(\lambda)} \longrightarrow 0.
\end{aligned} \tag{3.11}$$

This relation implies the convergence of one-dimensional distributions. The convergence of the multi-dimensional distributions follows from the formula

$$\mathbf{E}\exp\left\{i\sum_{j=1}^{m}\big(\lambda_j, \zeta_n(t_j)\big)\right\} = \mathbf{E}\exp\left\{\sum_{k=0}^{[nt]}\hat{\eta}_n(k)\right\},$$

where

$$\hat{\eta}_n(k) = \begin{cases} \eta_n\left(\sum_{s=1}^{m} \lambda_s, k\right) & \text{as } 0 \le k \le [nt_1], \\ \eta_n\left(\sum_{s=j+1}^{m} \lambda_s, k\right) & \text{as } [nt_j] < k \le [nt_{j+1}],\ 1 \le j < r. \end{cases}$$

This relation proves the first part of Theorem 3.1.

Now we prove that under the conditions of this theorem the measures generated by the sequence of processes $\zeta_n(\cdot)$ are weakly compact in space $\mathcal{D}^m_{[0,T]}$. According to [GRI 73] we need to prove that

$$\lim_{h\to+0} \limsup_{n\to\infty} \sup \Big\{ \big|\mathbf{E}\big[\exp\big\{i\big(\lambda, \zeta_n(t+s) - \zeta_n(t)\big)\big\} A\big] - 1\big| : \\ 0 \le t \le T,\ s < h,\ A \in F_0^{[nt]} \Big\} = 0. \tag{3.12}$$

Taking into account inequality $|e^{\alpha+\beta} - e^{\alpha+\gamma}| \le |e^{\beta} - e^{\gamma}|$ which is true for any complex α, β, γ such that $\operatorname{Re}\alpha \le 0$, and adding and subtracting several terms, we obtain the following representation:

$$\begin{aligned}
&\big|\mathbf{E}\big[\exp\big\{i\big(\lambda, \zeta_n(t+s) - \zeta_n(t)\big)\big\} \mid A\big] - 1\big| \\
&= \Bigg|\mathbf{E}\Bigg[\exp\Bigg\{\sum_{k=[nt]+1}^{[n(t+s)]} \eta_n(\lambda,k)\Bigg\} \mid A\Bigg] \pm \mathbf{E}\Bigg[\exp\Bigg\{\sum_{k=[nt]+N_n+1}^{[n(t+s)]} \eta_n(\lambda,k)\Bigg\} \mid A\Bigg] \\
&\quad \pm \mathbf{E}\exp\Bigg\{\sum_{k=[nt]+N_n+1}^{[n(t+s)]} \eta_n(\lambda,k)\Bigg\} \pm \mathbf{E}\exp\Bigg\{\sum_{k=[nt]+N_n+1}^{[n(t+s)]} \alpha_n(\lambda,k)\Bigg\} - 1\Bigg| \\
&\le \mathbf{E}\Bigg[\Bigg|\exp\Bigg\{\sum_{k=[nt]+1}^{[nt]+N_n} \eta_n(\lambda,k)\Bigg\} - 1\Bigg| \mid A\Bigg] \\
&\quad + \Bigg|\mathbf{E}\Bigg[\exp\{\sum_{k=[nt]+N_n+1}^{[n(t+s)]} \eta_n(\lambda,k)\Bigg\} \mid A\Bigg] - \mathbf{E}\exp\Bigg\{\sum_{k=[nt]+N_n+1}^{[n(t+s)]} \eta_n(\lambda,k)\Bigg\}\Bigg| \\
&\quad + \mathbf{E}\Bigg|\exp\Bigg\{\sum_{k=[nt]+N_n+1}^{[n(t+s)]} \big(\eta_n(\lambda,k) - \alpha_n(\lambda,k)\big)\Bigg\} - 1\Bigg| \\
&\quad + \Bigg|\exp\Bigg\{\sum_{k=[nt]+N_n+1}^{[n(t+s)]} \alpha_n(\lambda,k)\Bigg\} - 1\Bigg| = \delta_1 + \delta_2 + \delta_3 + \delta_4.
\end{aligned}$$

Furthermore,

$$\delta_1 \leq \mathbf{E}\left[\left|\sum_{k=[nt]+1}^{[nt]+N_n} \eta_n(\lambda,k)\right| \mid A\right]$$
$$\leq \sup\left\{\left|\sum_{k=1}^{N_n} a_n(\lambda,k+l,x_k)\right| : l \leq nT,\ x_k \in X,\ k=\overline{1,N_n}\right\} \longrightarrow 0$$

as $n \to \infty$ according to relation (3.5). Then, using the inequality,

$$\left|\int_X f(x)P(\mathrm{d}x) - \int_X f(x)Q(\mathrm{d}x)\right| \leq 2\sqrt{2}\sup_x \left|f(x)\right| \sup_A \left|P(A)-Q(A)\right| \quad (3.13)$$

which holds for any complex bounded function $f(x)$ and any probabilistic measures $P(\cdot), Q(\cdot)$ (this inequality follows from the analogous inequality for real functions [BIL 68]) and taking into consideration that $A \in \mathcal{F}_0^{[nt]}$, we obtain

$$\delta_2 \leq 2\sqrt{2}\varphi_n\big([nt],[nt]+N_n+1\big) \leq 2\sqrt{2}\max_{k\leq nT}\varphi_n\big(k,k+N_n+1\big) \longrightarrow 0,$$

according to relation (3.4). Variable δ_3 tends to 0 according to inequality (3.10), and $\delta_4 \to 0$ due to condition (3.6).

Now it remains to prove that the weak compactness of the measures generated by $\zeta_n(\cdot)$ in the space $\mathcal{D}^m_{[0,T]}$ together with condition (3.6) implies (3.8). Using the known inequality

$$\mathbf{P}\left\{|\xi| > \frac{2}{u}\right\} \leq \frac{1}{u}\int_{-u}^{+u}\left|1-\mathbf{E}\exp\left\{i(\lambda,\xi)\right\}\right|\mathrm{d}\lambda,$$

and condition (3.7) we find that for any $0 \leq t \leq T$,

$$\lim_{L\to\infty}\limsup_{n\to\infty}\mathbf{P}\left\{\left|\zeta_n(t)\right| > L\right\} = 0. \quad (3.14)$$

Note that the measures generated by $\tilde{\zeta}_n(\cdot)$ in the space $\mathcal{D}^m_{[0,T]}$ are also weakly compact as using condition (3.6) it is easy to check the condition of the (3.13) type. Now we choose from n an arbitrary subsequence $n_k \to \infty$. Consider the sequence of processes $\zeta_{n_k}(\cdot)$. According to the weak compactness it is possible to choose a subsequence n_{k_l} such that the measures generated by sequences $\zeta_{n_{k_l}}(\cdot)$ and $\tilde{\zeta}_{n_{k_l}}(\cdot)$ weakly converge in $\mathcal{D}^m_{[0,T]}$ to a measure generated by a proper (due to relation (1.13)) random process $\zeta_0(\cdot)$. Thus

$$\mathbf{E}f\big(\zeta_{n_{k_l}}(\cdot)\big) - \mathbf{E}f\big(\tilde{\zeta}_{n_{k_l}}(\cdot)\big) \longrightarrow \mathbf{E}f\big(\zeta_0(\cdot)\big) - \mathbf{E}f\big(\zeta_0(\cdot)\big) = 0.$$

This relation according to the arbitrary choice of the sequence n_k implies (3.8). □

NOTE 3.1. Condition (3.5) can be replaced by a weaker condition: for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \max_{k\le nT} \sup_{A\in F_0^k} \mathbf{P}\left\{ \left| \sum_{l=1}^{N_n} a_n\big(\lambda, k+l, x_{n,k+l}\big) \right| > \varepsilon \mid A \right\} = 0. \tag{3.15}$$

Theorem 3.1 is the approximative type theorem. As a consequence let us consider the convergence to a process with independent increments.

THEOREM 3.2. *Let*

$$\lim_{n\to\infty} A_n(\lambda, t) = A(\lambda, t), \quad t > 0, \tag{3.16}$$

and the conditions of Theorem 3.1 hold where condition (3.7) is replaced by

$$A(\pm 0, t) = 0, \quad t > 0. \tag{3.17}$$

Then the sequence of processes $\zeta_n(\cdot)$ J-converges in the interval $[0, T]$ to the process with independent increments $\zeta_0(\cdot)$ such that

$$\mathbf{E} \exp\big\{ i\big(\lambda, \zeta_0(t)\big) \big\} = \exp\big\{ A(\lambda, t) \big\}. \tag{3.18}$$

Proof. Condition (3.16) implies that the multi-dimensional distributions of process $\tilde{\zeta}_n(\cdot)$ weakly converge to corresponding distributions of process $\zeta_0(\cdot)$ as $\tilde{\zeta}_n(\cdot)$ is the process with independent increments. Then from relation (3.3) it follows that the multi-dimensional distributions of process $\zeta_n(\cdot)$ also weakly converge to the distributions of process $\zeta_0(\cdot)$.

Now under conditions of the 2nd part of Theorem 3.1 the sequence of measures generated by the sequence of processes $\zeta_n(\cdot)$ is weakly compact in space $\mathcal{D}^r_{[0,T]}$. This implies our result. □

Theorems 3.1 and 3.2 state the fact that the distributions of the initial process $\zeta_n(\cdot)$ are well enough approximated by the distributions of the process with independent increments $\tilde{\zeta}_n(\cdot)$ and $\zeta_0(\cdot)$. Local characteristics of these processes can be obtained by averaging of transition probabilities by the switching sequence x_{nk}.

The natural question appears: in which cases it is possible to calculate these averaged characteristics in terms of some stationary characteristics of process x_{nk}? Now we consider the sufficient conditions when relation (3.2) holds.

COROLLARY 3.1. *Let there exist a sequence N_n such that conditions (3.4) hold and for any $\lambda \in R^r$,*

$$\limsup_{n\to\infty} \sum_{k=0}^{[nT]} \beta_n(\lambda, k) \le C_\lambda < \infty; \tag{3.19}$$

$$\beta_n(\lambda) = \max_{k \le nT} \sum_{l=0}^{N_n} \beta_n(\lambda, k+l) \longrightarrow 0. \tag{3.20}$$

Then for any $\lambda \in R^r$, $q_n(\lambda) \to 0$.

Proof. The proof follows from the relation

$$\begin{aligned}
\limsup_{n\to\infty} q_n(\lambda) &\le \limsup_{n\to\infty} \Bigg(\beta_n(\lambda) \sum_{k=0}^{[nT]} \beta_n(\lambda, k) \\
&\quad + 8 \sum_{k=0}^{[nT]} \beta_n(\lambda, k) \max_{l \le [nT]} \sum_{i=1}^{[nT]-l} \beta_n(\lambda, l+i) \sqrt{\varphi_n(l, l+i)} \Bigg) \\
&\le \limsup_{n\to\infty} \Bigg(C_\lambda \beta_n(\lambda) \\
&\quad + 8 C_\lambda \Big(\beta_n(\lambda) + C_\lambda \max_{l \le nT} \max_{N_n \le i \le nT} \sqrt{\varphi_n(l, l+i)} \Big) \Bigg) = 0,
\end{aligned}$$

which follows from conditions (3.4), (3.19) and (3.20). □

Similar results can be obtained for the accumulative processes defined on the processes in continuous time of the form:

$$\zeta_n(t) = \int_0^{nt} \mathrm{d}\xi_n\big(u, x_n(u)\big). \tag{3.21}$$

Process $\zeta_n(\cdot)$ in this case can be formally defined in the following way. Let $x_n(t)$, $t \ge 0$, be a random process with values in a measurable space (X, B_X) and $\{a_n(\lambda, t, x),\ \lambda \in R^r,\ t \ge 0,\ x \in X\}$ be a family of B_X-measurable by variable x functions such that for any fixed t, x, the function $\exp\{a_n(\lambda, t, x)\}$ is a characteristic function of an infinitely divisible law. We thus define a process $\zeta_n(t)$ by its finite-dimensional distributions in the following way:

$$\begin{aligned}
&\mathbf{E} \exp\big\{ i\big(\lambda, \zeta_n(t)\big)\big\} = \mathbf{E} \exp\bigg\{ \int_0^{nt} a_n\big(\lambda, u, x_n(u)\big) \mathrm{d}u \bigg\}, \\
&\mathbf{E} \exp\bigg\{ i \sum_{j=1}^m \big(\lambda_j, \zeta_n(t_j)\big)\bigg\} = \mathbf{E} \exp\bigg\{ \int_0^{nt_1} a_n\bigg(\sum_{k=1}^m \lambda_k, u, x_n(u) \bigg) \mathrm{d}u \\
&\qquad\qquad + \sum_{j=1}^{m-1} \int_{nt_j}^{nt_{j+1}} a_n\bigg(\sum_{k=j+1}^m \lambda_k, u, x_n(u) \bigg) \mathrm{d}u \bigg\}
\end{aligned} \tag{3.22}$$

for any $0 \le t_1 < t_2 < \cdots < t_m$, $\lambda_j \in R^r$, $m \ge 1$.

Thus, at the given trajectory $x_n(\cdot)$, process $\zeta_n(t)$ is the process with conditionally independent increments and it can be characterized as the inhomogenous in time process with independent increments switched by process $x_n(\cdot)$. The family of processes with independent increments is defined by the family of cumulant functions $a_n(\lambda, t, x)$. The results of Theorems 3.1 and 3.2 hold for process $\zeta_n(t)$ as well, where in the corresponding conditions the sums should be replaced by integrals.

3.2. Limit theorems for sums with Markov switching

Now we consider as the consequences of Theorems 3.1 and 3.2 the case where the switching sequence is a homogenous or non-homogenous Markov process. The exposition partially follows [ANI 83].

Let x_{nk}, $k \geq 0$, be a non-homogenous MP with values in a measurable space (X, B_X) given by the family of transition probabilities

$$p_n(k, x, j, A) = \mathbf{P}\big(x_{nj} \in A \mid x_{nk} = x\big), \quad x \in X,\ A \in B_X,\ k < j, \tag{3.23}$$

and an initial state x_{n0}. Let us define the uniformly strong mixing coefficient by the formula

$$\varphi_n(k, j) = \sup_{x_1, x_2, A} \big|p_n\big(k, x_1, j, A\big) - p_n\big(k, x_2, j, A\big)\big|, \quad k < j, \tag{3.24}$$

and define a weaker coefficient – strong mixing coefficient:

$$\begin{aligned}\tilde{\psi}_n(k, j) = \sup \big\{\big|&\mathbf{P}\big(x_{nk} \in A,\ x_{nj} \in B\big) \\ &- \mathbf{P}\big(x_{nk} \in A\big)\mathbf{P}\big(x_{nj} \in B\big)\big| : A, B \in B_X\big\}.\end{aligned} \tag{3.25}$$

Now we provide the sufficient conditions when Theorems 3.1 and 3.2 hold. Let us keep all the notations given above.

COROLLARY 3.2. *Let at each $\lambda \in R^r$,*

$$\varlimsup_{n\to\infty} n \max_{k \leq nT} \sup_{x \in X} \big|a_n(\lambda, k, x)\big| \leq C_\lambda < \infty, \tag{3.26}$$

there is a sequence N_n such that, as $n \to \infty$,

$$N_n \longrightarrow \infty, \quad n^{-1} N_n \longrightarrow 0, \tag{3.27}$$

and

$$\max_{k \leq nT} \max_{m \geq N_n} \tilde{\psi}_n(k, k+m) \longrightarrow 0. \tag{3.28}$$

Then relation (3.3) is true.

If in addition conditions (3.16), (3.17) hold, then the sequence of processes $\zeta_n(\cdot)$ J*-converges in interval* $[0,T]$ *to the process with independent increments* $\zeta_0(\cdot)$ *(see equation (3.18)).*

Indeed, it is easy to check that under our assumptions the conditions of Theorem 3.1 are valid.

NOTE 3.2. Condition (3.28) holds if there is an integer sequence r_n and a q, $0 \le q < 1$, such that $n^{-1} r_n \to 0$ and

$$\sup_{k \ge 0} \varphi_n\big(k, k + r_n\big) \le q. \tag{3.29}$$

Indeed, by definition $\psi_n(k,j) \le \varphi_n(k,j)$. Moreover, according to [DOO 53], (3.29) implies

$$\varphi_n(k, k+j) \le q^{[j/r_n]-1}, \quad k \ge 0,\ j \ge 1. \tag{3.30}$$

Thus, we can choose a sequence N_n in the form: $N_n = [\sqrt{n r_n}]$. This sequence satisfies conditions (3.27) and according to (3.30) relation (3.28) is also valid.

Now consider a homogenous case. Let x_{nk}, $k \ge 0$, be a homogenous MP with values in a measurable space (X, B_X) given by the family of transition probabilities

$$p_n(x,k,A) = \mathbf{P}\big(x_{nk} \in A \mid x_n(0) = x\big), \quad x \in X,\ A \in B_X,\ k \ge 1, \tag{3.31}$$

and an initial state $x_n(0)$. Also let $F_{nk} = \{\xi_{nk}(x),\ x \in X\}$, $k \ge 0$, be the families of jointly independent random variables with values in R^r that are independent of $\{x_{nl},\ l \ge 0\}$ and whose distributions do not depend on index k. Denote

$$a_n(\lambda, x) = \ln \mathbf{E} \exp\big\{i\big(\lambda, \xi_n(1,x)\big)\big\}, \quad x \in X,\ \lambda \in R^r.$$

Suppose that there is a sequence of probability measures $\pi_n(A)$, $A \in B_X$, such that process x_{nk}, $k \ge 0$, is uniformly ergodic in the following sense: there exists a sequence $N_n \to \infty$ such that $n^{-1} N_n \to \infty$ and

$$\lim_{n\to\infty} \sup_{k > N_n} \sup_{x \in X,\ A \in B_X} \big|p_n(x,k,A) - \pi_n(A)\big| = 0. \tag{3.32}$$

We put

$$a_n(\lambda) = \int_X a_n(\lambda, x) \pi_n(\mathrm{d}x). \tag{3.33}$$

COROLLARY 3.3. *Let condition (3.32) hold and at each* $\lambda \in R^r$,

$$\limsup_{n\to\infty} n \sup_x \big|a_n(\lambda, x)\big| \le C_\lambda < \infty, \tag{3.34}$$

$$\lim_{n\to\infty} na_n(\lambda) = A(\lambda), \tag{3.35}$$

where $A(\pm 0) = 0$.

Then the sequence of processes $\zeta_n(\cdot)$ J*-converges on interval* $[0, T]$ *to the homogenous process with independent increments* $\zeta_0(\cdot)$ *such that*

$$\mathbf{E}\exp\big\{i\big(\lambda, \zeta_0(t)\big)\big\} = \exp\big\{A(\lambda)t\big\}. \tag{3.36}$$

The proof follows straightforwardly from Corollary 3.2.

Note that condition (3.29) implies condition (3.32). In particular applications condition (3.29) can be checked. For example, if x_{nk} is an MP with finite state space X and matrix of one-step transition probabilities P_n and $P_n \to P_0$ as $n \to \infty$, where an MP with matrix P_0 is ergodic, then condition (3.29) is true where instead of r_n we can take a large enough value L.

3.2.1. *Flows of rare events*

Now consider the applications of Theorems 3.1 and 3.2 and the corollaries above to the case where we consider the sums of rare indicators with Markov switching.

3.2.1.1. *Discrete time*

Let x_{nk}, $k \ge 0$, at any $n = 1, 2, \dots$ be a non-homogenous MP with values in a measurable space (X, B_X) given by the family of transition probabilities (3.23) and $\varphi_n(k, j)$ is the uniformly strong mixing coefficient defined by formula (3.24). Let $\{\chi_{nk}(x),\ x \in X\}$, $k \ge 0$, be the families of jointly independent and independent of $\{x_{nl},\ l \ge 0\}$ random indicators, where $\mathbf{P}(\chi_{nk}(x) = 1) = 1 - \mathbf{P}(\chi_{nk}(x) = 0) = p_{nk}(x)$. Consider the sequence of stepwise random processes

$$\zeta_n(t) = \sum_{k=0}^{[nt]} \chi_{nk}\big(x_{nk}\big), \quad t \ge 0.$$

Denote

$$p_{nk} = \mathbf{E}p_{nk}\big(x_{nk}\big), \quad b_{nk} = \sup_{x\in X} p_{nk}(x), \quad k \ge 0, \tag{3.37}$$

and put

$$\Lambda_n(t) = \sum_{k=0}^{[nt]} p_{nk}.$$

Let $\Pi_n(t)$ be a Poisson process with cumulative rate $\Lambda_n(t)$ in interval $[0, t]$. The following results are a consequence of Corollary 3.1.

STATEMENT 3.1. *Let there be a sequence N_n such that conditions (3.4) hold,*

$$\limsup_{n\to\infty} \sum_{k=0}^{[nT]} b_{nk} \le C_T < \infty; \tag{3.38}$$

$$\beta_n = \max_{k\le nT} \sum_{l=0}^{N_n} b_{n,k+l} \longrightarrow 0. \tag{3.39}$$

Then the finite dimensional distributions of processes $\zeta_n(t)$ and $\Pi_n(t)$ are asymptotically equivalent.

In particular this means that

$$\sup_{t\le T}\sup_{k\ge 0} \left| \mathbf{P}\big(\zeta_n(t) = k\big) - \exp\big\{ -\Lambda_n(t)\big\} \frac{\Lambda_n(t)^k}{k!} \right| \longrightarrow 0.$$

STATEMENT 3.2. *If in addition for any $t \le T$,*

$$\Lambda_n(t) \longrightarrow \Lambda_0(t), \tag{3.40}$$

where $\Lambda_0(t)$ is a continuous function, then $\zeta_n(t)$ J-converges to the Poisson processwith cumulative rate $\Pi_0(t)$.

Consider a homogenous in time MP x_{nk}, $k \ge 0$. Assume that the distributions of variables $\{\chi_{nk}(x),\ x \in X\}$, $k \ge 0$, do not depend on index k. Let the process x_{nk} at each n be ergodic with stationary measure $\pi_n(A)$. Denote

$$\widetilde{p}_n = \int_X p_{n1}(x)\pi_n(\mathrm{d}x), \quad b_n = \sup_{x\in X} p_{n1}(x). \tag{3.41}$$

Statements 3.1 and 3.2 lead to 3.3.

STATEMENT 3.3. *Let condition (3.29) be true and*

$$\limsup_{n\to\infty} n\widetilde{p}_n < \infty, \quad r_n b_n \longrightarrow 0. \tag{3.42}$$

Then the finite dimensional distributions of process $\zeta_n(t)$ are asymptotically equivalent to the distributions of a homogenous Poisson process $\Pi_n(t)$ with rate $n\widetilde{p}_n$.

If in addition, $n\widetilde{p}_n \to \lambda$, then $\zeta_n(t)$ J-converges to a Poisson process with rate λ.

3.2.1.2. *Continuous time*

Similar results can be proved for MPs in continuous time. Let $x_n(t)$, $t \geq 0$, at any $n = 1, 2, \ldots$ be a non-homogenous MP with values in a measurable space $(X, \mathcal{B}_X)$ and let $\{q_n(t,x),\ t \geq 0,\ x \in X\}$, be non-negative functions measurable in a natural way with regard to σ-algebra $\mathcal{B}_X$. Let $\zeta_n(t)$ be a doubly stochastic Poisson process [COX 80] switched by $x_n(t)$. This means that the local rate of jump at time t is $q_n(t, x_n(t))$. Denote

$$\widetilde{q}_n(t) = \mathbf{E}q_n\big(t, x_n(t)\big), \quad b_n(t) = \sup_{x \in X} q_n(t,x), \quad t \geq 0, \tag{3.43}$$

and put

$$\Lambda_n(t) = \int_0^{nt} \widetilde{q}_n(u)\mathrm{d}u.$$

The following results can be proved in a similar way as in Theorem 3.1 and Corollary 3.1.

STATEMENT 3.4. *Let there be sequence N_n such that conditions (3.4) hold,*

$$\limsup_{n\to\infty} \int_0^{nT} b_n(u)\mathrm{d}u \leq C_T < \infty; \tag{3.44}$$

$$\beta_n = \sup_{t \leq T} \int_t^{t+N_n} b_n(u)\mathrm{d}u \longrightarrow 0. \tag{3.45}$$

Then the finite dimensional distributions of processes $\zeta_n(t)$ and $\Pi_n(t)$ are asymptotically equivalent.

Consider a homogenous in time MP $x_n(t)$, $t \geq 0$, and assume that the family of local rates $\{q_n(x)\}$ is given. Let $\zeta_n(t)$ be a doubly stochastic Poisson process switched by $x_n(t)$ with the local rate of jump at time t, $q_n(x_n(t))$. Assume that the process x_{nk} at each n is ergodic with stationary measure $\pi_n(A)$. Denote

$$\widetilde{q}_n = \int_X q_n(x)\pi_n(\mathrm{d}x), \quad b_n = \sup_{x \in X} q_n(x). \tag{3.46}$$

STATEMENT 3.5. *Let condition (3.29) be true and*

$$\limsup_{n\to\infty} n\widetilde{q}_n < \infty, \quad r_n b_n \longrightarrow 0. \tag{3.47}$$

Then the finite dimensional distributions of the process $\zeta_n(t)$ are asymptotically equivalent to the distributions of the homogenous Poisson process $\Pi_n(t)$ with rate $n\widetilde{q}_n$.

If in addition, $n\widetilde{q}_n \to \lambda$, then $\zeta_n(t)$ J-converges to a Poisson process with rate λ.

At the study of flows of rare events the time of the first event can be interpreted as the time of loss of a call in the queueing system or the time of the first failure. As the results of section 3.2.1 state that the flows of rare events under rather general conditions J-converge to Poisson processes, this means that the time of the first event weakly converges to the time of the first jump of a limiting Poisson process. For homogenous in time models this means the exponential approximation of the time of the first jump.

Some consequences of these results for asymptotically connected MPs are considered later in section 7.2 and in applications to the analysis of the time of the first loss of a call in queueing models.

3.3. Quasi-ergodic Markov processes

When we study the behavior of stepwise processes switched by a non-homogenous sequence, the main problem in particular applications is to evaluate the behavior of the cumulative function $A_n(\lambda, t)$ (or $\Lambda_n(t)$ in the flow of rare events), where these functions are calculated using the distribution of the switching component at time t. When the characteristics of the process slowly vary in a corresponding scale of time, we can expect that this distribution can be approximated by a corresponding quasi-ergodic distribution of the process at time t.

In this section we consider a special subclass of non-homogenous in time MPs, so-called quasi-ergodic MPs, with the rates (or probabilities) which vary slowly in time. These processes satisfy a strong mixing condition and in some sense have the ergodic properties in a corresponding scale of time. Quasi-ergodic MPs were first introduced in [ANI 83, ANI 88].

DEFINITION 3.1. *An MP x_{nk}, $k \geq 0$, with values in $(X, \mathcal{B}_X)$ is said to be quasi-ergodic if there is a family of probability measures $\pi(t, A)$, $A \in \mathcal{B}_X$, $t \geq 0$, such that for any $T > 0$,*

$$\lim_{j\to\infty} \limsup_{n\to\infty} \sup_{A\in\mathcal{B}_X,\, t\leq T} \left|\mathbf{P}\left(x_{n,[nt]} \in A \mid x_{n,[nt]-j} = x\right) - \pi(t, A)\right| = 0. \quad (3.48)$$

Let us consider the conditions in terms of local transition characteristics when an MP is quasi-ergodic. Assume that the MP x_{nk} is given by the family of one-step transition probabilities at step k, $p_n(x, A, k)$, $x \in X$, $A \in \mathcal{B}_X$, $k \geq 1$.

LEMMA 3.1. *The process x_{nk} is quasi-ergodic, if the following conditions hold:*

1) for any $j \geq 0$,

$$\lim_{n\to\infty} \sup_{x\in X,\, A\in\mathcal{B}_X,\, t\leq T} \left|p_n\left(x, A, [nt] - j\right) - p_t(x, A)\right| = 0,$$

where at each $t \geq 0$, $p_t(x, A)$ *is the transition probability of an homogenous in time ergodic MP* $x_k^{(t)}$, $k \geq 0$, *with values in* X;

2) for any $t \geq 0$, *there is a stationary measure* $\pi(t, A)$, $A \in \mathcal{B}_X$, *such that*

$$\lim_{m\to\infty} \sup_{x\in X,\ A\in\mathcal{B}_X,\ t\leq T} \left|p_t^{(m)}(x, A) - \pi(t, A)\right| = 0,$$

where $p_t^{(m)}(x, A)$ *is the* m*-step transition probability for process* $x_k^{(t)}$.

Proof. For any $j \geq 0$,

$$\begin{aligned}&\left|\mathbf{P}\big(x_{n,[nt]} \in A \mid x_{n,[nt]-j} = x\big) - \pi(t, A)\right| \\ &\quad \leq \left|\mathbf{P}\big(x_{n,[nt]} \in A \mid x_{n,[nt]-j} = x\big) - p_t^{(j)}(x, A)\right| + \left|p_t^{(j)}(x, A) - \pi(t, A)\right|.\end{aligned}$$

Let us prove that for any fixed $j \geq 0$ the first term in the right-hand side of the inequality tends to zero as $n \to \infty$ uniformly in $x \in X$, $A \in \mathcal{B}_X$, $t \leq T$. Then the statement of Lemma 3.1 follows from condition 2).

The proof is carried out by induction on j. For $j = 0$ the assertion follows from condition 1). Let us write the following relation for $j \geq 1$:

$$\begin{aligned}&\left|\mathbf{P}\big(x_{n,[nt]} \in A \mid x_{n,[nt]-j-1} = x\big) - p_t^{(j+1)}(x, A)\right| \\ &\quad \leq \int_X \left|\mathbf{P}\big(x_{n,[nt]} \in A \mid x_{n,[nt]-j} = y\big) - p_t^{(j)}(y, A)\right| p_n\big(x, \mathrm{d}y, [nt] - j\big) \\ &\qquad + \int_X \left|p_t^{(j)}(y, A)\big(p_n\big(x, \mathrm{d}y, [nt] - j\big) - \pi_t(x, \mathrm{d}y)\big)\right|.\end{aligned}$$

The first term in the right-hand side tends to zero by the induction hypothesis. Using the known inequality

$$\left|\int_X f(x)P(\mathrm{d}x) - \int_X f(x)Q(\mathrm{d}x)\right| \leq 2\sup_x \left|f(x)\right| \sup_{A\in\mathcal{B}_X} \left|P(A) - Q(A)\right|$$

which is valid for any non-negative measures $P(\cdot)$ and $Q(\cdot)$, we see that the second term does not exceed

$$2 \sup_{x\in X,\ A\in\mathcal{B}_X,\ t\leq T} \left|p_n\big(x, A, [nt] - j\big) - p_t(x, A)\right|.$$

This implies the statement of Lemma 3.1. $\square$

Note that if x_{nk} is quasi-ergodic, then by definition for any $t > 0$,

$$\lim_{n\to\infty} \sup_{A\in\mathcal{B}_X} \left|\mathbf{P}\big(x_{n,[nt]} \in A\big) - \pi(t, A)\right| \longrightarrow 0, \tag{3.49}$$

and

$$\lim_{L\to\infty} \limsup_{n\to\infty} \sup_{k\geq 0} \varphi_n(k, k+L) \longrightarrow 0, \tag{3.50}$$

where $\varphi_n(k, j)$ is the uniformly strong mixing coefficient defined by equation (3.24). Therefore, (3.29) is satisfied where instead of r_n we can take a large enough value L. In this case the conditions of convergence become more transparent. For example, in Statement 3.1 in relation (3.37) we can replace the value $p_{nk} = \mathbf{E}p_{nk}(x_{nk})$ by the expression

$$p_{nk} = \int_X p_{nk}(x)\pi(k/n, \mathrm{d}x),$$

and Statement 3.1 holds when equation (3.38) holds and condition (3.39) is replaced by

$$\max_{k\leq nT} b_{nk} \longrightarrow 0.$$

This result can also be used to provide more transparent conditions in Corollary 3.2. If x_{nk} is quasi-ergodic and equation (3.26) holds, then the values $\alpha_n(\lambda, k)$ and $\beta_n^2(\lambda, k)$ in conditions of Theorem 3.1 can be replaced by the values:

$$\widetilde{\alpha}_n(\lambda, k) = \int_X a_n(\lambda, k, x)\pi(k/n, \mathrm{d}x),$$

$$\beta_n^2(\lambda, k) = \int_X \big(a_n(\lambda, k, x) - \alpha_n(\lambda, k)\big)^2 \pi(k/n, \mathrm{d}x).$$

Similar results are also valid for a non-homogenous MP $x_n(t)$, $t \geq 0$, in continuous time. In this case in Definition 3.1 of a quasi-ergodic process we just need to change $x_{n,[nt]}$ to $x_n([nt])$ and $x_{n,[nt]-j}$ to $x_n([nt] - j)$, respectively.

The conditions of quasi-ergodicity in terms of local transition rates can be formulated by analogy to Lemma 3.1. Consider for simplicity a finite space state. Let $x_n(t)$, $t \geq 0$, be a non-homogenous in time MP with state space $X = \{1, 2, \ldots, d\}$ given by the family of instantaneous rates $\{a_n(i, j, t),\ i, j \in X,\ i \neq j,\ t \geq 0\}$. This means that at time t the rate of jump from j to j is $a_n(i, j, t)$. Suppose that there is a family of functions $\{a_0(i, j, v),\ i, j \in X,\ i \neq j,\ v \geq 0\}$ that are continuous in v and a sequence $k_n \to \infty$ such that for any $i, j \in X$, $i \neq j$, and any fixed $T \geq 0$,

$$\lim_{n\to\infty} \sup_{v\leq T} \left|a_n\big(i, j, k_n v\big) - a_0(i, j, v)\right| = 0. \tag{3.51}$$

For each fixed $v \geq 0$ denote by $x_0^{(v)}(\cdot)$ an auxiliary homogenous MP with state space X given by the family of rates $\{a_0(i,j,v),\ i,j \in X,\ i \neq j\}$. We assume that for any $T > 0$, MP $x_0^{(v)}(\cdot)$ is ergodic uniformly in $u \leq T$. Formally this means the following. Let $\varphi^{(v)}(\cdot)$ be its uniformly strong mixing coefficient:

$$\begin{aligned}\varphi^{(v)}(u) = \sup_{t\geq 0} \max_{i,j\in X, A\subset X} \Big| &\mathbf{P}\big\{x_0^{(v)}(t+u) \in A \mid x_0^{(v)}(t) = i\big\} \\ &- \mathbf{P}\big\{x_0^{(v)}(t+u) \in A \mid x_0^{(v)}(t) = j\big\}\Big|.\end{aligned} \tag{3.52}$$

Suppose that there exists q, $0 \leq q < 1$, and for any $T > 0$ there exists a constant $r(T) > 0$ such that for any $v \leq T$,

$$\varphi^{(v)}\big(r(T)\big) \leq q. \tag{3.53}$$

Note that condition (3.53) holds if there exists a sequence of states $(i_1, i_2, \ldots, i_m)$ containing all states from X, where $i_m = i_1$, and constants $0 < c_T < C_T < \infty$ such that for any $v \leq T$,

$$\begin{aligned}c_T \leq \min_{k=1,\ldots,m-1} \inf_{v\leq T} a_0\big(i_k, i_{k+1}, v\big), \\ \max_{k=1,\ldots,m-1} \sup_{v\leq T} a_0\big(i_k, i_{k+1}, v\big) \leq C_T.\end{aligned} \tag{3.54}$$

LEMMA 3.2. *Let conditions (3.51), (3.53) hold. Then for any $v > 0$ as $n \to \infty$,*

$$\mathbf{P}\big\{x_n\big(k_n v\big) = j \mid x_n(0) = i\big\} \longrightarrow \pi^{(v)}(j), \quad i \in X,\ j \in X, \tag{3.55}$$

where $\pi^{(v)}(j)$, $j \in X$, is the stationary distribution of MP $x_0^{(v)}(\cdot)$ which exists under assumption (3.53).

If $x_n(t)$ is a quasi-ergodic process with $k_n = n$, then the conditions of Statements 3.4 and 3.5 can be expressed in terms of quasi-stationary distribution. This means that instead of the value $\widetilde{q}_n(t)$ defined in (3.43) we can use the expression

$$\widetilde{q}_n(t) = \sum_{i\in X} q_n(t,i)\pi^{(t/n)}(i).$$

3.4. Limit theorems for non-homogenous Markov processes

In this section we consider the conditions of weak convergence in Skorokhod space $D_{[0,1]}$ for stepwise processes of centered sums of non-random functions defined on a non-homogenous Markov chain in a triangular scheme to a Gaussian process $\xi_0(\cdot)$ with independent increments. Weak convergence of sums of random variables to a composition of a process $\xi_0(\cdot)$ and the process which is independent of it with independent increments is also considered. The exposition in this section follows [ANI 88, ANI 89].

3.4.1. *Convergence to Gaussian processes*

Many papers and books are devoted to the central limit theorem for non-homogenous Markov chains and closely related problems, see for example [BIL 68, IBR 71] and many others. The difference between these results and the results in this section is that the theorems on the weak convergence are considered in a triangular scheme for Markov processes with arbitrary state space, and the mixing conditions are weakened (asymptotic mixing in a certain time scale is sufficient). The state spaces satisfying conditions of this type (asymptotically connected or so-called S-sets) were introduced in the author's papers [ANI 70, ANI 74]. See the description of S-sets in sections 6.2 and 7.3. The method of proof uses the mixing properties and is common for discrete and continuous time.

For any $n > 0$, let x_{ni}, $i = 0, \ldots, n$, be an MP with values in the measurable space $(X_n, \mathcal{B}_{X_n})$ and transition probabilities

$$p_n(k, x, m, A) = \mathbf{P}\big(x_{nm} \in A \mid x_{nk} = x\big), \quad x \in X_n,\ A \in \mathcal{B}_{x_n},\ k \le m.$$

Let

$$\varphi_n(k, m) = \sup\big\{\big|p_n(k, x_1, mA) - p_n(k, x_2, m, A)\big| : x_1, x_2 \in X_n,\ A \in \mathcal{B}_{x_n}\big\},$$

$k \le m$, be the uniformly strong mixing coefficient. In addition, let $f_{nk}(x)$, $x \in X_n$, $k = 0, \ldots, n$, be a sequence of non-random real-valued measurable functions. We define the process

$$S_n(t) = \sum_{k=0}^{[nt]} f_{nk}\big(x_{nk}\big) \tag{3.56}$$

and denote

$$m_n(t) = \mathbf{E}S_n(t), \quad B_n^2(t) = \mathbf{Var}S_n(t),$$
$$\xi_n(t) = \big(S_n(t) - m_n(t)\big)B_n^{-1}, \quad t \in [0, 1].$$

Let

$$m_{nk} = \mathbf{E}f_{nk}\big(x_{nk}\big), \quad a_{nk} = \sup_x \big|f_{nk}(x) - m_{nk}\big|, \quad \sigma_{nk}^2 = \mathbf{Var}f_{nk}\big(x_{nk}\big),$$

$$D_n(t) = \sum_{k=0}^{[nt]} \sigma_{nk}^2 + 2 \sum_{0 \le i < j \le nt} \sqrt{\varphi_n(i, j)}\sigma_{ni}\sigma_{nj},$$

$$Q_n = \max_{k<n} \left(\sum_{i=k}^{n} a_{ni}^2 \varphi_n(k, i) + 4 \left(\sum_{i=k}^{n} a_{ni}\varphi_n(k, i) \right)^2 + 12 \sum_{k \le i < j \le n} a_{ni}a_{nj} \sqrt{\varphi_n(k, i)\varphi_n(i, j)} \right). \tag{3.57}$$

THEOREM 3.3. *Assume that as $n \to \infty$, there is a sequence B_n such that the following conditions are satisfied:*

1) $B_n^{-2} B_n^2(t) \to B^2(t)$, $0 \le t \le 1$;

2) $B_n^{-2} Q_n \to 0$;

3) $\lim_{n\to\infty} \sup_{0\le t\le 1-h} B_n^{-2}(D_n(t+h) - D_n(t)) \to 0$ as $h \to +0$;

4) for any $t \in [0,1]$, the sequence of variables $\xi_n^2(t)$ is uniformly integrable.

Then the sequence of processes $\xi_n(t)$ weakly converges in $\mathcal{D}_{[0,1]}$ to Gaussian process $\xi_0(t)$ with independent increments where $\mathbf{E}\xi_0(t) = 0$ and $\mathbf{Var}\xi_0(t) = B^2(t)$, $t \ge 0$.

Note that a similar result is valid for processes in continuous time, we just need to replace the sums by the corresponding integrals.

Proof. Extending the proof of Theorem 19.2 in [BIL 68] to the non-homogenous case, we establish that it suffices to verify the following conditions:

a) for any $\varepsilon > 0$ as $h \to +0$,

$$\limsup_{n\to\infty} \mathbf{P}\left(\sup_{|t_1 - t_2| < h} \left|\xi_n(t_1) - \xi_n(t_2)\right| > \varepsilon \right) \longrightarrow 0,$$

b) every weak limit of $\xi_n(\cdot)$ is a process with independent increments.

Then our results follow in view of conditions 1) and 4).

To verify condition a) we introduce continuous processes $\tilde{\xi}_n(t)$, $t \in [0,1]$, constructed by polygonal curves joining points $(k/n, \xi_n(k/n))$, $k = 0, \ldots, n$. Since

$$\sup_{0\le t\le 1} |\tilde{\xi}_n(t) - \xi_n(t)| \le \max_{0\le k\le n} B_n^{-1} |a_{nk} - m_{nk}| \longrightarrow 0$$

by condition 2), it suffices to verify that functions $\xi_n(\cdot)$ form a weakly compact set in $\mathcal{D}_{[0,1]}$. This will imply that $\tilde{\xi}_n(\cdot)$ form a weakly compact set in $\mathcal{D}_{[0,1]}$. Since a continuous process cannot converge in the metric of $\mathcal{D}_{[0.1]}$ to a discontinuous process, any weak limit of $\xi_n(\cdot)$ is a continuous process. According to [GRI 73], for weak convergence of $\xi_n(\cdot)$ in $\mathcal{D}_{[0,1]}$ it suffices that for any $\varepsilon > 0$,

$$\lim_{h\to+0} \overline{\lim_{n\to\infty}} \sup_{u\le h} \sup_{t,x} \mathbf{P}\left(\left|\xi_n(t+u) - \xi_n(t)\right| > \varepsilon \mid x_{n[nt]} = x\right) = 0. \tag{3.58}$$

Denote by $E_{n,k,x}$ the expectation under the condition $x_{nk} = x$. According to the Chebyshev inequality, the relation

$$\lim_{h\to+0} \overline{\lim_{n\to\infty}} \sup_{u\le h} \sup_{t,x} E_{n[nt],x} \left(\xi_n(t+u) - \xi_n(t)\right)^2 = 0 \tag{3.59}$$

is sufficient for (3.58). Let

$$A_n(k,m,x) = E_{n,k,x}\left(\sum_{i=k}^{m}\big(f_{ni}(x_{ni}) - m_{ni}\big)\right)^2.$$

Let us use the inequalities

$$\big|\mathbf{E}f_{ni}(x_{ni})f_{nj}(x_{nj}) - m_{ni}m_{nj}\big| \le 2\big(\varphi_n(i,j)\mathbf{E}f_{nj}^2(x_{ni})\mathbf{E}f_{nj}^2(x_{nj})\big)^{1/2}, \quad (3.60)$$

$$\left|\int f(x)P(\mathrm{d}x) - \int f(x)Q(\mathrm{d}x)\right| \le 2\sup_x\big|f(x)\big|\sup_A\big|P(A)-Q(A)\big|, \quad (3.61)$$

valid for any real function $f(x)$ and probability measures $P(\cdot)$ and $Q(\cdot)$ (if $f(x)\ge 0$, then factor 2 in the right-hand side can be omitted). Thus,

$$\begin{aligned} A_n(k,m,x) &= \sum_{i=k}^{m} E_{n,k,x}\big(f_{ni}(x_{ni}) - m_{ni}\big)^2 \\ &\quad + 2\sum_{k\le i<j\le m} E_{n,k,x}\big(f_{ni}(x_{ni}) - m_{ni}\big)\big(f_{nj}(x_{nj}) - m_{nj}\big). \end{aligned}$$

Furthermore, $|E_{n,k,x}(f_{ni}(x_{ni}) - m_{ni})^2 - \sigma_{ni}^2| \le a_{ni}^2\varphi_n(k,i)$, and since $|E_{n,k,x}f_{ni}(x_{ni}) - m_{ni}| \le 2a_{ni}\varphi_n(k,i)$, it follows that

$$\begin{aligned} &\big|E_{n,k,x}\big(f_{ni}(x_{ni}) - m_{ni}\big)\big(f_{nj}(x_{nj}) - m_{nj}\big)\big| \\ &\quad \le 4a_{ni}a_{nj}\varphi_n(k,j) + 2\big(\varphi_n(i,j)\big(\sigma_{ni}^2 + a_{ni}^2\varphi_n(k,i)\big)\big(\sigma_{nj}^2 + a_{nj}^2\varphi_n(k,j)\big)\big)^{1/2}. \end{aligned}$$

Using the fact that $\sqrt{a^2+b^2} \le |a| + |b|$, we get after some algebra that

$$A_n\big([nt],\big[n(t+u)\big],x\big) \le D_n(t+h) - D_n(t) + Q_n$$

as $0 \le u \le h$. Thus, (3.59) follows from conditions 2) and 3).

Let us verify condition b). It suffices to show that

$$E_{n,[nt],x}\big(\xi_n(t+s) - \xi_n(t)\big) \longrightarrow 0, \quad (3.62)$$

$$E_{n,[nt],x}\big(\xi_n(t+s) - \xi_n(t)\big)^2 \longrightarrow B^2(t+s) - B^2(t) \quad (3.63)$$

uniformly with respect to x for any $t, s > 0$.

Indeed, if $\xi_0(\cdot)$ is a weak limit of $\xi_n(\cdot)$, then in view of condition 4) and the preceding arguments it is a continuous process and

$$\mathbf{E}\big[\xi_0(t+s) - \xi_0(t) \mid \mathcal{F}_t\big] = 0,$$

$$\mathbf{E}\big[\big(\xi_0(t+s) - \xi_0(t)\big)^2 \mid \mathcal{F}_t\big] = B^2(t+s) = B^2(t),$$

where $\mathcal{F}_t = \sigma\{\xi_0(u),\ u \leq t\}$ and $B^2(t)$ is continuous (the continuity follows from condition 3). Thus $\xi_0(t)$ is a continuous martingale with non-random quadratic characteristics and therefore is a Gaussian process with independent increments according to [GIK 72]. Using condition 2) and inequalities (3.60) and (3.61) we now obtain (3.62) and (3.63). □

COROLLARY 3.4. *Suppose that for some fixed $q \in [0, 1)$ there exists an integer sequence r_n such that for any $k \geq 0$,*

$$\varphi_n\big(k, k + r_n\big) \leq q. \tag{3.64}$$

Assume that $a_{nk} \leq C$, $B_n^2 = nr_n$, $k = 0, \dots, n$, condition 1) in Theorem 3.3 holds, and

$$n^{-1} r_n \longrightarrow 0. \tag{3.65}$$

Then the statement of Theorem 3.3 is valid.

Proof. Let us verify the conditions of the theorem. Relation (3.64) implies

$$\varphi_n(k, k + m) \leq q^{[mr_n^{-1}]}, \quad m \geq 0. \tag{3.66}$$

Let C_i denote some constants independent of n. Using equation (3.66), we can prove that $Q_n \leq C_1 r_n^2$ and $D_n(t) \leq C_2 ntr_n$. In view of equation (3.65) this involves conditions 2) and 3) of Theorem 3.3. Carrying out the calculations in detail ([BIL 68], 20, Lemma 4) and using (3.66), we can obtain the estimate

$$\mathbf{E} S_n(t)^4 \leq C_3 t^2 n^2 r_n^2.$$

This implies that $\mathbf{E}\xi_n^4(t) \leq C_4$ for every $t > 0$, which implies condition 4). □

COROLLARY 3.5. *Suppose that x_{nk}, $k \geq 0$, is an homogenous Markov process, equations (3.64) and (3.65) hold, function $f_n(x)$ is uniformly bounded,*

$$\begin{aligned} B^2 = \lim_{n\to\infty} r_n^{-1} \bigg(& E_\pi\big(f_n(x_{n0}) - m_{n0}\big)^2 \\ & + 2\sum_{i=1}^{n} E_\pi\Big(\big(f_n(x_{n0}) - m_{n0}\big)\big(f_n(x_{ni}) - m_{n0}\big)\Big)\bigg), \end{aligned} \tag{3.67}$$

and $B^2 > 0$, where $m_{n0} = E_\pi f_n(x_{n0})$, and the expectation is calculated with respect to the stationary initial distribution of X_{n0}, which exist in view of (3.64).

Then for any initial distribution the sequence of processes

$$\left(nr_n\right)^{-1/2}\left(\sum_{k=0}^{[nt]} f_n\left(x_{nk}\right) - ntm_{n0}\right)$$

converges weakly in $\mathcal{D}_{[0,1]}$ *to process* $BW(t)$*, where* $W(t)$ *is a standard Wiener process.*

3.4.2. *Convergence to processes with independent increments*

Now we consider the conditions of the convergence of the sums of random variables defined on a Markov chain to processes with independent increments. Let for each $n > 0$, $\{\eta_{nk}(x),\ x \in X_n\}$, $k \geq 0$, be the families of jointly independent random variables with values in $\mathcal{R}^1$ that are also independent of x_{ni}, $i \geq 0$, and their characteristic functions are B_{X_n}-measurable with respect to x. The existence of $f_{nk}(x) = \mathbf{E}\eta_{nk}(x)$, $x \in X_n$, $k \geq 0$, is assumed. Let the following condition hold:

$$\begin{gathered}\mathbf{E}\exp\left\{i\lambda B_n^{-1}\left(\eta_{nk}(x) - f_{nk}(x)\right)\right\} = 1 + b_{nk}(\lambda, x) + o_{nk}(\lambda, x),\\ \lambda \in R^1,\ x \in X_n,\ k \geq 0,\end{gathered} \tag{3.68}$$

where B_n is a normalizing factor, and for any λ as $n \to \infty$,

$$\sum_{k=0}^{n} \sup_x \left|o_{nk}(\lambda, x)\right| \longrightarrow 0. \tag{3.69}$$

Let us keep the notation of Theorem 3.3, section 3.4.1, where $m_{nk} = \mathbf{E}f_{nk}(x_{nk})$, $k \geq 0$, and the value $S_n(t)$ is defined by equation (3.56). Denote

$$\begin{aligned}
\alpha_{nk}(\lambda) &= \mathbf{E}b_{nk}\left(\lambda, x_{nk}\right), \quad A_n(\lambda, t) = \sum_{k=0}^{[nt]} \alpha_{nk}(\lambda),\\
\beta_{nk}^2(\lambda) &= \mathbf{E}\left|b_{nk}\left(\lambda, x_{nk}\right) - \alpha_{nk}(\lambda)\right|^2,\\
q_n(\lambda) &= \sum_{k=0}^{n} \beta_{nk}^2(\lambda) + \sum_{0 \leq k < j \leq n} \sqrt{\varphi_n(k,j)}\beta_{nk}(\lambda)\beta_{nj}(\lambda),\\
g_{nk}(\lambda) &= \sup_x \left|b_{nk}(\lambda, x) - \alpha_{nk}(\lambda)\right|,\\
G_n(k,l) &= \sum_{i=k}^{l} g_{ni}^2(\lambda)\varphi_n(k,i) + 4\left(\sum_{i=k}^{l} g_{ni}(\lambda)\varphi_n(k,i)\right)^2\\
&\quad + 12 \sum_{k \leq i < j \leq l} \sqrt{\varphi_n(i,j)\varphi_n(k,i)}g_{ni}(\lambda)g_{nj}(\lambda).
\end{aligned}$$

THEOREM 3.4. *Assume that variables $f_{nk}(x)$ and MP x_{nk} satisfy the conditions of Theorem 3.3, conditions (3.68) and (3.69) hold, and for any $\lambda \in R^1$,*

$$q_n(\lambda) \to 0, \tag{3.70}$$

$$\lim_{n\to\infty} A_n(\lambda, t) = A(\lambda, t), \quad t \in [0, 1], \tag{3.71}$$

where $A(0, t) = 0$, $t \in [0, 1]$.

Then the finite-dimensional distributions of the process

$$\eta_n(t) = B_n^{-1} \sum_{k=0}^{[nt]} \big(\eta_{nk}(x_{nk}) - m_{nk}\big)$$

weakly converge in $[0, 1]$ to the distributions of the process with independent increments $\eta_0(t)$ such that

$$\mathbf{E} \exp\big\{i\lambda\eta_0(t)\big\} = \exp\bigg\{ -\frac{1}{2}\lambda^2 B^2(t) + A(\lambda, t)\bigg\}.$$

If in addition for any λ, (3.71) holds uniformly with respect to t and

$$\lim_{h\to+0} \overline{\lim_{n\to\infty}} \max_{k\le n} G_n\big(k, k + [nh]\big) = 0, \tag{3.72}$$

then the sequence $\eta_n(\cdot)$ weakly converges to $\eta_0(\cdot)$ in $\mathcal{D}_{[0,1]}$.

COROLLARY 3.6. *Assume that $B_n = \sqrt{nr_n}$, $\sup_{k\le n} a_{nk} < C_1$, (see (3.57)), condition 1) of Theorem 3.3 holds along with conditions (3.64), (3.65), (3.68), (3.69) and*

$$n \max_k \sup_x \big|b_{nk}(\lambda, x)\big| < C_2. \tag{3.73}$$

Then the statement of Theorem 3.4 holds.

COROLLARY 3.7. *Under the conditions of Corollary 3.6 suppose that distributions of variables $\eta_{nk}(x)$ do not depend on k, conditions (3.68), (3.69), and (3.73) hold, and $a_{n1} < C_1$. If $n\alpha_n(\lambda) \to A(\lambda)$, where $\alpha_n(\lambda) = E_\pi b_{n1}(\lambda, x_{n0})$ (expectation is taken over the stationary distribution of x_{nk} which exists under conditions (3.64), (3.65)) and $A(0) = 0$, then the statement of Theorem 3.4 is valid with*

$$\mathbf{E} \exp\big\{i\lambda\eta_0(t)\big\} = \exp\bigg\{ -\frac{1}{2}\lambda^2 t^2 B^2 + A(\lambda)t\bigg\}.$$

Proof. Denote $a_{nk}(\lambda, x) = \ln \mathbf{E} \exp\{i\lambda B_n^{-1}(\eta_{nk}(x) - m_{nk})\}$. By definition of process $\eta_n(t)$ the following representation holds:

$$\begin{aligned}\mathbf{E} \exp\big\{i\lambda\eta_n(t)\big\} &= \mathbf{E} \exp\left\{\sum_{k=0}^{[nt]} a_{nk}\big(\lambda, x_{nk}\big)\right\} \\ &= \mathbf{E} \exp\big\{i\lambda\xi_n(t) + A_n(\lambda, t) + \rho_n(\lambda, t) + \delta_n(\lambda, t)\big\},\end{aligned} \tag{3.74}$$

where

$$\rho_n(\lambda, t) = \sum_{k=0}^{[nt]} \big(b_{nk}\big(\lambda, x_{nk}\big) - \alpha_{nk}(\lambda)\big), \quad \delta_n(\lambda, t) = \sum_{k=0}^{[nt]} o_{nk}\big(\lambda, x_{nk}\big).$$

Condition (3.69) implies $\delta_n(\lambda, t) \xrightarrow{\text{P}} 0$, correspondingly, equation (3.70) implies $\rho_n(\lambda, t) \to 0$, and according to Theorem 3.3, $\xi_n(t)$ weakly converges to $\xi_0(t)$ in the interval $[0, 1]$. Since $\operatorname{Re} a_{nk}(\lambda, x) \le 0$ and, from the conditions of Theorem 3.4, process $\sum_{k=0}^{[nt]} a_{nk}(\lambda, x_{nk})$ weakly converges to $i\lambda\xi_0(t) + A(\lambda, t)$, then Helly's theorem implies that

$$\mathbf{E} \exp\big\{i\lambda\eta_n(t)\big\} \longrightarrow \mathbf{E} \exp\big\{i\lambda\xi_0(t) + A(\lambda, t)\big\},$$

which proves the convergence of one-dimensional distributions, and, similarly, the convergence of finite-dimensional distributions. To verify that sequence $\eta_n(\cdot)$ forms a weakly compact set in $\mathcal{D}_{[0,1]}$ it suffices according to [GRI 73] to establish that

$$\lim_{h\to+0} \overline{\lim_{n\to\infty}} \sup_{t,x} \sup_{u<h} \big|E_{n,[nt],x} \exp\big\{i\lambda\big(\eta_n(t+u) - \eta_n(t)\big)\big\} - 1\big| = 0. \tag{3.75}$$

By virtue of equation (3.74) and inequality $|e^a - e^b| \le |a - b|$, which is true for $\operatorname{Re} a, \operatorname{Re} b \le 0$, it suffices to check that

$$\begin{aligned}\lim_{h\to+0} \overline{\lim_{n\to\infty}} \sup_{t,x} \sup_{u<h} \big(&E_{n,[nt],x}\big|\xi_n(t+u) - \xi_n(t)\big|\big|A_n(\lambda, t+u) - A_n(\lambda, t)\big| \\ &+ E_{n,[nt],x}\big|\rho_n(\lambda, t+u) - \rho_n(\lambda, t)\big|\big) = 0.\end{aligned} \tag{3.76}$$

Note that

$$E_{n,[nt],x}\big|\xi_n(t+u) - \xi_n(t)\big| \le \sqrt{A_n\big([nt], \big[n(t+u)\big], x\big)},$$

and the right-hand part of Theorem 3.3 tends to zero. As in Theorem 3.3, conditions (3.70) and (3.72) imply that $E_{n,[nt],x}|\rho_n(\lambda, t+u) - \rho_n(\lambda, t)|^2$ is small. In view of the uniform convergence in equation (3.71) this implies equation (3.76), which proves Theorem 3.4.

The corollaries are proved by direct verification of the conditions of Theorem 3.4. □

The fundamental complexity in real problems is verification of conditions related to mixing properties. Let us consider the following example, which arises in the models of asymptotic aggregation of state space (see Chapter 8 and [ANI 73, ANI 78, ANI 88]).

EXAMPLE 3.1. Let x_{nk}, $k \geq 0$, be a homogenous MP with values in X and one-step transition probabilities $p_n(x, A)$, $x \in X$, $A \in \mathcal{B}_X$. Assume that

$$p_n(x, A) = p_0(x, A) + n^{-\alpha} b(x, A) + o\big(n^{-a}\big),$$

where $\alpha > 0$, $n^{\alpha} o(n^{-\alpha}) \to 0$ uniformly with respect to x and A, and $\sup_{x,A} |b(x, A)| < C$. The state space of the Markov chain x_k with one-step transition probabilities $p_0(x, A)$ can be subdivided into several essential classes X_y, $y \in Y$. Uniformly with respect to $y \in Y$, each class X_y for the chain x_k is uniformly ergodic with stationary measure $\pi^{(y)}(A)$, $A \in \mathcal{B}_{X_y}$, and

$$\widehat{p}(y, B) = \int_{X_y} b\big(x, \cup_{u \in B} X_u\big) \pi^{(y)}(\mathrm{d}x), \quad y \in Y,\ B \in \mathcal{B}_Y,\ y \notin B.$$

Assume that the MP with transition probabilities $\widehat{p}(y, B)$ is uniformly ergodic.

For this case it can be proved that the original MP x_{nk} satisfies condition (3.64), where $r_n = C_q n^{\alpha}$, and hence Theorem 3.3 is applicable to x_{nk} for $\alpha < 1$.

3.5. Bibliography

[ANI 70] ANISIMOV V., "Limit distributions of functionals of a semi-Markov process given on a fixed set of states up to the time of first exit", *Soviet Math. Dokl.*, vol. 11, no. 4, p. 1002–1004, 1970.

[ANI 73] ANISIMOV V., "Asymptotic consolidation of the states of random processes", *Cybernetics*, vol. 9, no. 3, p. 494–504, 1973.

[ANI 74] ANISIMOV V., "Limit theorems for sums of random variables in an array of sequences defined on a subset of states of a Markov chain up to the exit time", *Theor. Probab. and Math. Stat.*, , no. 4, p. 1–12, 1974.

[ANI 78] ANISIMOV V., "Limit theorems for switching processes and their applications", *Cybernetics*, vol. 14, no. 6, p. 917–929, 1978.

[ANI 83] ANISIMOV V., "Limit theorems for non-homogenous weakly dependent summation schemes", *Theor. Probab. and Math. Stat.*, vol. 27, p. 9–21, 1983.

[ANI 88] ANISIMOV V., *Random Processes with Discrete Component. Limit Theorems*, Kiev University (Russian), Kiev, Ukraine, 1988.

[ANI 89] ANISIMOV V., "Functional limit theorems for switching processes for nonhomogeneuos Markov chains", *Theor. Probab. and Math. Stat.*, vol. 39, p. 5–12, 1989.

[BIL 68] BILLINGSLEY P., *Convergence of Probability Measures*, Wiley, New York, 1968.

[COX 80] COX D. and ISHAM V., *Point Processes*, Chapman & Hall, London, 1980.

[DOO 53] DOOB J. L., *Stochastic Processes*, Wiley, New York, 1953.

[GIK 72] GIKHMAN I. and SKOROKHOD A., *Stochastic Differential Equations and their Applications*, Springer-Verlag, New York, 1972.

[GRI 73] GRIGELIONIS B., "The relative compactness of sets of probability measures in $D_{(0,\infty)}(X)$", *Math. Trans. Acad. Sci. Lithuanian SSR*, vol. 13, 1973.

[IBR 71] IBRAGIMOV I. A. and LINNIK Y. V., *Independent and Stationary Sequences of Random Variables*, Wolters-Noordhoff Publishing, Groningen, 1971.

Chapter 4

Averaging Principle and Diffusion Approximation for Switching Processes

4.1. Introduction

In this chapter we study the limit theorems for recurrent sequences and SPs in the case of "fast" switching. Consider a sequence of SPs $(x_n(t), \zeta_n(t))$, $t \geq 0$, depending on a scaling parameter n on the expanding interval $[0, nT]$, where $n \to \infty$. Suppose that SP depends on n in such a way that the number of switches on each interval $[na, nb]$, $0 < a < b < T$, tends in probability to infinity. In this case we can expect that under some natural assumptions a normalized trajectory of $\zeta_n(nt)$ uniformly converges in probability to a deterministic function which is a solution of a differential equation (AP), and a normalized difference between the trajectory of $\zeta_n(nt)$ and this solution weakly converges in Skorokhod space $\mathcal{D}_T$ to a diffusion process (DA). As sample trajectories of a limiting process are continuous, this convergence implies a weak convergence of functionals, which are continuous with respect to the uniform convergence [ETH 86, SKO 56]. Note that after time re-scaling we can consider the process in interval $[0, T]$ in the scale of time nt and in this case the number of switches in each interval $[a, b]$ tends to infinity (switches occur rapidly).

A new approach based on the investigation of the asymptotic properties of a special subclass of SP – recurrent process of semi-Markov type (RPSM), theorems about the convergence of recurrent sequences to the solutions of stochastic differential equations and the convergence of superposition of random functions are developed.

SPs (see section 1.2) are described in terms of constructive characteristics [ANI 75, ANI 77] and are represented in the recurrent form. This representation, as we see from examples in the previous chapters, is convenient for describing wide classes of stochastic systems. It also plays the basic role in the analysis of asymptotic properties of stochastic systems with "rare" and "fast" switches [ANI 78, ANI 88, ANI 92a].

For the special classes of random evolutions (processes with independent increments and Markov and semi-Markov switches), the Law of Large Numbers and the Central Limit Theorem have been proved by many authors [GRI 69, PAP 72, KUR 73, KER 78a, KER 78b, PIN 75, ANI 73, ANI 88, KOR 93, WAT 84, KOR 94, ANI 95]. The averaging principle for stochastic differential equations in the case of independent Markov switches was investigated in [KHA 68] and some Markov models in the case of feedback were studied in [SKO 89]. Models of Markov evolutions in the scheme of asymptotic phase consolidation were investigated in [ANI 73, ANI 78, ANI 88, ANI 99b, ANI 00a, ANI 00b, ANI 02a, ANI 04, ANI 87, KOR 93, KOR 94, KOR 99, KOR 00, KOR 04, KOR 05].

AP for RPSMs was considered in [ANI 90, ANI 92a]. Nonhomogenous in time models and additional semi-Markov switches were studied in [ANI 93, ANI 94a, ANI 95]. The applications of these results to the dynamic systems with fast semi-Markov switches [ANI 95], stochastic differential equations [ANI 86, ANI 89] and branching processes in a fast semi-Markov environment [ANI 96] have been obtained.

A wide area of applications is switching queueing models and networks. For Markov queueing systems and networks, various results of AP and DA type are investigated in [ANI 92a, ANI 92b, ANI 95, ANI 97, ANI 99a, ANI 99c, ANI 02b, ANI 91b, ANI 94b] for semi-Markov and non-Markov models, different results are obtained in [ANI 92a, ANI 99b, ANI 02b].

This chapter consists of five sections. In section 4.2 the averaging principle for stochastic recurrent sequences is considered. In section 4.3 the averaging principle and diffusion approximation for RPSMs are proved. In sections 4.4 and 4.5 these results are extended to RPSMs with additional semi-Markov switching and with feedback. Section 4.6 is devoted to AP and DA for general SPs.

4.2. Averaging principle for switching recurrent sequences

In this section we study the convergence of stepwise processes generated by trajectories of recurrent sequences to solutions of ordinary differential equations. Let for any $n > 0$, a sequence ξ_{nk}, $k \geq 0$, of random variables with values in $\mathcal{R}^r$ and a monotone flow of σ-algebras $\mathcal{F}_{nk}$ ($\mathcal{F}_{nk} \subset \mathcal{F}_{nk+1}$) be given such that variables ξ_{nk} are $\mathcal{F}_{nk}$-measurable and satisfy the relation

$$\xi_{nk+1} = \xi_{nk} + a_{nk}\big(\xi_{nk}\big)\frac{1}{n} + \beta_{nk}, \quad k \geq 0, \tag{4.1}$$

where ξ_{n0} is given, $a_{nk}(y)$, $y \in \mathcal{R}^m$ are some, possibly random, $\mathcal{F}_{nk}$-measurable functions and β_{nk} are $\mathcal{F}_{nk+1}$-measurable random variables.

THEOREM 4.1. *Assume that functions $a_{nk}(y)$ for some given $T > 0$ satisfy the following conditions: for any n, $k \le nT$, $y_1, y_2 \in \mathcal{R}^m$,*

$$\big|a_{nk}(y_1) - a_{nk}(y_2)\big| \le C\big|y_1 - y_2\big| + c_{nk}, \tag{4.2}$$

where C is constant, $c_{nk} \ge 0$, and as $n \to \infty$,

$$\frac{1}{n}\sum_{k=0}^{[nT]} c_{nk} \xrightarrow{\text{P}} 0; \tag{4.3}$$

$$\max_{k \le nT}\left|\sum_{i=0}^{k} \beta_{ni}\right| \xrightarrow{\text{P}} 0. \tag{4.4}$$

Then

$$\max_{k \le nT}\big|\xi_{nk} - y_{nk}\big| \xrightarrow{\text{P}} 0, \tag{4.5}$$

where variables y_{nk}, $k \ge 0$, satisfy the relation:

$$y_{n0} = \xi_{n0}, \quad y_{nk+1} = y_{nk} + a_{nk}\big(y_{nk}\big)\frac{1}{n}, \ k \ge 0. \tag{4.6}$$

Proof. Let us introduce the variables $z_{nk} = \xi_{nk} - y_{nk}$ and $v_{nk} = \max_{k \le m}|z_{nk}|$, $k \ge 0$. Relations (4.1), (4.6) imply that

$$z_{nk+1} = z_{nk} + \big(a_{nk}\big(\xi_{nk}\big) - a_{nk}\big(y_{nk}\big)\big)\frac{1}{n} + \beta_{nk}, \quad k \ge 0.$$

Summing both parts in k from 0 till m we obtain

$$z_{nm+1} = \frac{1}{n}\sum_{k=0}^{m}\big(a_{nk}\big(\xi_{nk}\big) - a_{nk}\big(y_{nk}\big)\big) + \sum_{k=0}^{m}\beta_{nk}.$$

Using this relation and relation (4.2) we obtain the inequality

$$v_{nm+1} \le \frac{C}{n}\sum_{k=1}^{m} v_{nk} + \gamma_{nm}, \quad m \ge 0, \tag{4.7}$$

where

$$\gamma_{nm} = \max_{k \le m}\left|\sum_{i=0}^{k}\beta_{ni}\right| + \frac{1}{n}\sum_{k=0}^{m} c_{nk}.$$

Now let us use the following lemma:

LEMMA 4.1. *Assume that a real-valued sequence v_k, $k \geq 0$, satisfies the relation:*

$$v_0 \leq \gamma_0, \quad v_{k+1} \leq \sum_{i=0}^{k} b_i v_i + \gamma_{k+1},\ k \geq 0,$$

where $b_i \geq 0$, $\gamma_i \geq 0$, $i \geq 0$.

Then

$$v_{k+1} \leq \hat{\gamma}_{k+1} \exp\left\{ \sum_{i=0}^{k} b_i \right\},$$

where $\hat{\gamma}_k = \max_{i \leq k} \gamma_k$.

Proof. Using inequality $1 + b \leq e^b$, we obtain recurrently

$$\begin{aligned} v_1 &\leq b_0 v_0 + \gamma_1 \leq b_0\gamma_0 + \gamma_1 \leq \hat{\gamma}_1 e^{b_0}, \\ v_2 &\leq b_1\hat{\gamma}_1 e^{b_0} + b_0\gamma_0 + \gamma_2 \leq b_1\hat{\gamma}_1 e^{b_0} + \hat{\gamma}_2 e^{b_0} \leq \hat{\gamma}_2 e^{b_0+b_1}, \\ v_3 &\leq b_2\hat{\gamma}_2 e^{b_0+b_1} + b_1\hat{\gamma}_1 e^{b_0} + b_0\gamma_0 + \gamma_3 \leq b_2\hat{\gamma}_2 e^{b_0+b_1} + \hat{\gamma}_3 e^{b_0+b_1} \\ &\leq \hat{\gamma}_3 e^{b_0+b_1+b_2}, \end{aligned}$$

and so on. Then by induction we obtain

$$\begin{aligned} v_{k+1} &\leq b_k v_k + \sum_{i=0}^{k-1} b_i v_i + \gamma_{k+1} \\ &\leq b_k \hat{\gamma}_k \exp\left\{ \sum_{i=0}^{k-1} b_i \right\} + \hat{\gamma}_{k+1} \exp\left\{ \sum_{i=0}^{k-1} b_i \right\} \\ &\leq \hat{\gamma}_{k+1} \exp\left\{ \sum_{i=0}^{k} b_i \right\}. \end{aligned}$$

This relation proves the statement of Lemma 4.1. □

Furthermore, according to equation (4.7), $v_{nm+1} \leq \gamma_{nm} \exp\{\frac{m}{n} C\}$. Finally this relation according to equations (4.3) and (4.4) implies equation (4.5) and proves Theorem 4.1. □

Theorem 4.1 is an approximating theorem. Let us now consider the conditions of the convergence of a random process constructed by sequence y_{nk} to a solution of an ordinary differential equation of the form

$$\mathrm{d}y(t) = a\big(t, y(t)\big)\mathrm{d}t, \quad y(0) = y_0. \tag{4.8}$$

Consider a simple but useful case for applications when coefficients $a_{nk}(y)$ have the form:

$$a_{nk}(y) = a\left(\frac{k}{n}, y\right), \quad k \geq 0, \tag{4.9}$$

where $a(t, y)$ is a given function.

THEOREM 4.2. *Assume that relations (4.1) and (4.9) are true, function $a(t, y)$ uniformly in y in each bounded region $|y| \leq L$, $0 \leq t \leq T$, is continuous with respect to t, and for any $t \leq T$, $y_1, y_2 \in \mathcal{R}^m$,*

$$\left|a(t, y_1) - a(t, y_2)\right| \leq C\left|y_1 - y_2\right|, \tag{4.10}$$

condition (4.4) is satisfied and $\xi_{n0} \xrightarrow{\mathrm{P}} y_0$.

Then

$$\max_{k \leq nT} \left|\xi_{nk} - y\left(\frac{k}{n}\right)\right| \xrightarrow{\mathrm{P}} 0, \tag{4.11}$$

where $y(t)$ is a solution of equation (4.8).

Proof. Condition (4.10) implies that for any $t \leq T$, $y \in \mathcal{R}^m$, $|a(t, y)| \leq C_1(1 + |y|)$, where C_1 is a constant. This relation together with relation (4.10) implies the existence of a unique solution to equation (4.8). Now let us define a sequence y_{nk}, $k \geq 0$, according to the relations:

$$y_{n0} = y_0, \quad y_{nk+1} = y_{nk} + a\left(\frac{k}{n}, y_{nk}\right)\frac{1}{n}, \ k \geq 0.$$

Using (4.8) we obtain

$$\begin{aligned} y\left(\frac{k+1}{n}\right) &= y\left(\frac{k}{n}\right) + a\left(\frac{k}{n}, y_{nk}\right)\frac{1}{n} \\ &\quad + \int_{\frac{k}{n}}^{\frac{k+1}{n}} \left(a\big(t, y(t)\big) - a\left(\frac{k}{n}, y\left(\frac{k}{n}\right)\right)\right)\mathrm{d}t. \end{aligned}$$

As function $a(t, y)$ is continuous, denote $G = \sup_{t\in[0,T]} |a(t, y(t))| < \infty$. Relation (4.8) implies that $|y(t + s) - y(t)| \leq G|s|$ as $t \vee (t + s) \leq T$. Thus,

$$\begin{aligned} &\left|\int_{k/n}^{(k+1)/n} \big(a\big(t, y(t)\big) - a\big(k/n, y(k/n)\big)\big)\mathrm{d}t\right| \\ &\quad \leq \sup_{t\leq T} \sup_{|u|\leq 1/n} \left|a\big(t + u, y(t)\big) - a\big(t, y(t)\big)\right|\frac{1}{n} + \frac{G}{2n^2}. \end{aligned}$$

Using uniform continuity of function $a(t, y)$ and Theorem 4.1 we obtain

$$\max_{k \le nT} \bigl|y_{nk} - y(k/n)\bigr| \longrightarrow 0, \tag{4.12}$$

as $n \to \infty$, and finally Theorem 4.1 implies (4.11). □

Condition (4.10) can be replaced by a weaker local Lipschitz condition: for any $t \le T$, $y \in \mathcal{R}^m$,

$$\bigl|a(t, y)\bigr| \le C\bigl(1 + |y|\bigr),$$

and for any $L > 0$ as $|y_1| \vee |y_2| \le L$, $t \le T$,

$$\bigl|a\bigl(t, y_1\bigr) - a\bigl(t, y_2\bigr)\bigr| \le C_L\bigl|y_1 - y_2\bigr|.$$

4.3. Averaging principle and diffusion approximation for RPSMs

In this section we study the limit theorems for RPSMs in the triangular scheme in the case of fast switching and prove that under natural assumptions, the normalized trajectory of an RPSM uniformly converges in probability to some function which is the solution of an ordinary differential equation (AP) and the normalized difference between the trajectory and this solution weakly converges in Skorokhod space D to some diffusion process (DA) (see [ANI 90]).

Let at each $n = 1, 2, \ldots$, $\mathcal{F}_{nk} = \{(\xi_{nk}(z), \tau_{nk}(z)),\ z \in R^r\}$, $k \ge 0$, be jointly independent at different k families of random variables with values in $\mathcal{R}^r \times [0, \infty)$ and distributions not depending on k, and let S_{n0} be the initial value in $\mathcal{R}^r$ which is independent of $\mathcal{F}_{nk}$ $k \ge 0$. According to section 1.1.2 let us introduce recurrent sequences

$$t_{n0} = 0,\ \ t_{nk+1} = t_{nk} + \tau_{nk}\bigl(S_{nk}\bigr),\ \ S_{nk+1} = S_{nk} + \xi_{nk}\bigl(S_{nk}\bigr),\ \ k \ge 0, \tag{4.13}$$

and define RPSM as follows:

$$S_n(t) = S_{nk} \quad \text{as } t_{nk} \le t < t_{nk+1},\ t > 0. \tag{4.14}$$

As under natural assumptions the normalized trajectory of an RPSM after n switches is of the order n, we will consider the dependence of the argument in recurrent equations on the re-scaled trajectory S_{nk}/n with the purpose of obtaining a state-dependent property in the limiting equations. Note that we can define the new variables $\widetilde{\xi}_{nk}(\alpha) = \xi_{nk}(n\alpha), \widetilde{\tau}_{nk}(\alpha) = \tau_{nk}(n\alpha)$, re-write the relations above in the form

$$t_{n0} = 0, \quad t_{nk+1} = t_{nk} + \widetilde{\tau}_{nk}\bigl(S_{nk}/n\bigr), \quad S_{nk+1} = S_{nk} + \widetilde{\xi}_{nk}\bigl(S_{nk}/n\bigr), \quad k \ge 0,$$

where the newly introduced variables now depend on the re-scaled values, and formulate the conditions of theorems in terms of variables $\widetilde{\xi}_{nk}(\alpha), \widetilde{\tau}_{nk}(\alpha)$. However, for simplicity we use the original notation.

Assume that there exist the functions $m_n(\alpha) = \mathbf{E}\tau_{n1}(n\alpha)$, $b_n(\alpha) = \mathbf{E}\xi_{n1}(n\alpha)$.

THEOREM 4.3 (AP). *Suppose that for any $N > 0$,*

$$\lim_{L\to\infty} \limsup_{n\to\infty} \sup_{|\alpha|\le N} \big\{\mathbf{E}\tau_{n1}(n\alpha)\chi\big(\tau_{n1}(n\alpha) > L\big) + \mathbf{E}\big|\xi_{n1}(n\alpha)\big|\chi\big(\big|\xi_{n1}(n\alpha)\big| > L\big)\big\} = 0, \tag{4.15}$$

and as $\max(|\alpha_1|, |\alpha_2|) \le N$,

$$\big|m_n(\alpha_1) - m_n(\alpha_2)\big| + \big|b_n(\alpha_1) - b_n(\alpha_2)\big| \le C_N\big|\alpha_1 - \alpha_2\big| + \alpha_n(N), \tag{4.16}$$

where C_N are some constants, $\alpha_n(N) \to 0$ uniformly in $|\alpha_1| \le N$, $|\alpha_2| \le N$, there exist functions $m(a) > 0$ and $b(a)$ such that for any $\alpha \in \mathcal{R}^r$ as $n \to \infty$,

$$m_n(\alpha) \longrightarrow m(\alpha), \quad b_n(\alpha) \longrightarrow b(\alpha), \tag{4.17}$$

and

$$n^{-1}S_{n0} \xrightarrow{\text{P}} s_0. \tag{4.18}$$

Then

$$\sup_{0\le t\le T} \big|n^{-1}S_n(nt) - s(t)\big| \xrightarrow{\text{P}} 0, \tag{4.19}$$

where function $s(t)$ satisfies the following ordinary differential equation

$$\mathrm{d}s(t) = m\big(s(t)\big)^{-1} b\big(s(t)\big)\mathrm{d}t, \tag{4.20}$$

and T is any positive number such that $y(+\infty) > T$ with probability one, where

$$y(t) = \int_0^t m\big(\eta(u)\big)\mathrm{d}u, \tag{4.21}$$

and $\eta(t)$ is a solution of the differential equation

$$\mathrm{d}\eta(u) = b\big(\eta(u)\big)\mathrm{d}u, \quad \eta(0) = S_0, \tag{4.22}$$

(it is assumed that a unique solution to equation (4.22) exists in each interval).

Proof. Let us introduce the sequences $\eta_{nk} = n^{-1}S_{nk}$, $y_{nk} = n^{-1}t_{nk}$, $k \geq 0$, and construct stepwise processes $\eta_n(\cdot)$ and $y_n(\cdot)$ as follows:

$$\eta_n(u) = \eta_{nk}, \quad y_n(u) = y_{nk} \quad \text{as } n^{-1}k \leq u < n^{-1}(k+1),\ u \geq 0.$$

Put

$$\nu_n(t) = \min\big\{k : k > 0,\ t_{nk+1} > nt\big\},$$
$$\mu_n(t) = \inf\big\{u : u > 0,\ y_n(u) > t\big\}.$$

By definition, $y_n(n^{-1}\nu_n(t)) \leq t < y_n(n^{-1}\nu_n(t)+1)$ and $\mu_n(t) = n^{-1}(\nu_n(t)+1)$. As $S_n(nt) = S_{n\nu_n(t)}$, the following representation is true:

$$n^{-1}S_n(nt) = \eta_n\big(n^{-1}\nu_n(t)\big) = \eta_n\big(\mu_n(t) - 1/n\big). \tag{4.23}$$

Thus, RPSM $n^{-1}S_n(nt)$ is constructed as a superposition of two processes: $\eta_n(t)$ and $\mu_n(t)$. Therefore, we study first the behavior of processes $\eta_n(t)$ and $y_n(t)$, then the behavior of $\mu_n(t)$ and their superposition. Using (4.13) we can write the relations

$$\eta_{nk+1} = \eta_{nk} + n^{-1}b_n\big(\eta_{nk}\big) + \varphi_{nk}, \quad k \geq 0, \tag{4.24}$$
$$y_{nk+1} = y_{nk} + n^{-1}m_n\big(\eta_{nk}\big) + \psi_{nk}, \quad k \geq 0, \tag{4.25}$$

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}, $k \geq 0$, are martingale differences with respect to the sequence of σ-algebra σ_{nk} generated by variables $\{\eta_{ni},\ i \leq k\}$. Assume first that condition (4.15) holds uniformly on $\alpha \in \mathcal{R}^r$. It follows from the paper [GRI 73] that for any $t > 0$,

$$\max_{m \leq nt} \left| \sum_{k=o}^{m} \varphi_{nk} \right| \xrightarrow{\text{P}} 0. \tag{4.26}$$

Furthermore, applying the results of book [GIK 78] and using relations (4.17), (4.26), we obtain

$$\sup_{u \leq t} \big|\eta_n(u) - \eta(u)\big| \xrightarrow{\text{P}} 0. \tag{4.27}$$

By analogy, we can prove that

$$\sup_{u \leq t} \big|y_n(u) - y(u)\big| \xrightarrow{\text{P}} 0. \tag{4.28}$$

As $m(a) > 0$, process $y(t)$ increases strictly monotonically. Thus, the process $y^{-1}(t) = \mu(t)$ exists for any t such that $y(+\infty) > t$ with probability one, is continuous and

$$\sup_{u\le t} \big|\mu_n(u) - \mu(u)\big| \xrightarrow{\text{P}} 0. \tag{4.29}$$

Using the results on the U-convergence of the superposition of random functions [BIL 68] and relations (4.27) and (4.28), we obtain (4.19). Finally, for any t such that $y(+\infty) > t$ with probability one,

$$\mathbf{P}\Big\{\sup_{u\le t} \big|s(t)\big| > N\Big\} \longrightarrow 0$$

as $N \to \infty$. This means that it is sufficient to check all conditions in each bounded region $|\alpha| \le N$. Thus, Theorem 4.3 is proved. □

Now we prove the convergence of the process

$$\gamma_n(t) = \frac{1}{\sqrt{n}}\big(S_n(nt) - ns(t)\big), \quad t \in [0, T],$$

to a diffusion process. Denote

$$\begin{aligned} \tilde{b}_n(\alpha) &= m_n(\alpha)^{-1} b_n(\alpha), \quad \tilde{b}(\alpha) = m(\alpha)^{-1} b(\alpha), \\ \rho_n(\alpha) &= \xi_{n1}(n\alpha) - b_n(\alpha) - \tilde{b}(\alpha)\big(\tau_{n1}(n\alpha) - m_n(\alpha)\big), \\ D_n^2(\alpha) &= \mathbf{E}\rho_n(\alpha)\rho_n(\alpha)^* \end{aligned} \tag{4.30}$$

(here and in what follows we denote the conjugate vector by symbol $*$), and put

$$q_n(\alpha, z) = \sqrt{n}\bigg(\widetilde{b}_n\Big(\alpha + \frac{1}{\sqrt{n}} z\Big) - \widetilde{b}(\alpha)\bigg). \tag{4.31}$$

THEOREM 4.4 (DA). *Let conditions (4.16)-(4.18) hold where in equation (4.16) a condition $\alpha_n(N) \to 0$ is replaced by $\sqrt{n}\alpha_n(N) \to 0$, there exist continuous matrix-valued functions $D^2(\alpha)$ and $Q(\alpha)$ and a vector-valued function $g(a)$ such that as $n \to \infty$, uniformly in each bounded region,*

$$D_n^2(\alpha) \longrightarrow D^2(\alpha), \tag{4.32}$$

$$q_n(\alpha, z) \longrightarrow Q(\alpha)z + g(a), \tag{4.33}$$

for any $z \in \mathcal{R}^r$,

$$\gamma_n(0) \overset{\text{w}}{\Longrightarrow} \gamma_0, \tag{4.34}$$

where γ_0 is a proper random variable, and for any $N > 0$,

$$\lim_{L\to\infty} \limsup_{n\to\infty} \sup_{|\alpha|<N} \Big\{ \mathbf{E}\tau_{n1}^2(n\alpha)\chi\big(\tau_{n1}(n\alpha) > L\big) \\ + \mathbf{E}\big|\xi_{n1}(n\alpha)\big|^2 \chi\big(\big|\xi_{n1}(n\alpha)\big| > L\big)\Big\} = 0. \tag{4.35}$$

Then for any T such that $y(+\infty) > T$, the sequence of processes $\gamma_n(t)$ J-converges in the interval $[0, T]$ to the diffusion process $\gamma(t)$ satisfying the following stochastic differential equation:

$$\mathrm{d}\gamma(t) = \big(Q\big(s(t)\big)\gamma(t) + g\big(s(t)\big)\big)\mathrm{d}t + D\big(s(t)\big)m\big(s(t)\big)^{-1/2}\mathrm{d}w(t), \\ \gamma(0) = \gamma_0, \tag{4.36}$$

where $s(\cdot)$ satisfies equation (4.20).

Proof. Condition (4.16) and relation $m(a) > 0$ imply that the limiting functions $b(a), m(a), \tilde{b}(a)$ also satisfy the local Lipschitz condition, and the function $b(a)$ and the function $q_n(\alpha, z)$ with respect to z satisfy condition (4.16). Let us keep the notation of Theorem 4.3, denote $v_{nk} = \gamma_n(y_{nk})$, $\tilde{s}_{nk} = s(y_{nk})$, $k \geq 0$, and suppose first for simplicity that s_0 is a non-random variable. As relation (4.19) holds, the trajectory $\eta_{nk}, k = \overline{0, nT}$, belongs to a bounded region with probability close to one. Thus, it is sufficient to check all conditions in each bounded region. By definition,

$$v_{nk+1} = v_{nk} + \frac{1}{\sqrt{n}}\big(\xi_{nk}\big(n\eta_{nk}\big) - n\big(\tilde{s}_{nk+1} - \tilde{s}_{nk}\big)\big).$$

Using the Lagrange formula and relation (4.14), we find that

$$\tilde{s}_{nk+1} - \tilde{s}_{nk} = \frac{1}{n}\tilde{b}\big(\tilde{s}_{nk}\big)\tau_{nk} + \delta_{nk}^{(1)} = \frac{1}{n}\tilde{b}\big(\tilde{s}_{nk}\big)m_n\big(\eta_{nk}\big) + \delta_{nk}^{(1)} + \delta_{nk}^{(2)},$$

where $\tau_{nk} = \tau_{nk}(n\eta_{nk})$, $|\delta_{nk}^{(1)}| \leq C\frac{1}{n^2}\tau_{nk}^2$ and $\mathbf{E}|\delta_{nk}^{(2)}|^2 \leq Cn^{-2}$. After some algebra we obtain:

$$v_{nk+1} = v_{nk} + \frac{1}{n}m_n\big(\eta_{nk}\big)q_n\big(\tilde{s}_{nk}, v_{nk}\big) + \frac{1}{\sqrt{n}}\alpha_{nk} + \delta_{nk}^{(3)}, \tag{4.37}$$

where $\alpha_{nk} = \xi_{nk}(\eta_{nk}) - b_n(\eta_{nk}) - \tilde{b}(\tilde{s}_{nk})(\tau_{nk} - m_n(\eta_{nk}))$, and $\mathbf{E}|\delta_{nk}^{(3)}|^2 \leq Cn^{-3/2}$. We can also prove by analogy to Theorem 4.3 that

$$\max_{k\leq nT}\left|\sum_{t=o}^{k}\delta_{ni}^{(3)}\right| \xrightarrow{\mathrm{P}} 0. \tag{4.38}$$

Following the lines of the proof of Theorem 4.3, we see that for any $u > 0$,

$$\eta_{n,[nu]} \xrightarrow{\mathrm{P}} \eta(u), \quad \tilde{s}_{n,[nu]} \xrightarrow{\mathrm{P}} s\big(y(u)\big) = \eta\big(\mu\big(y(u)\big)\big) = \eta(u).$$

Thus, as $n \to \infty$, $k/n \to t$ and $v_{nk} = z$, the coefficient at $1/n$ in the right-hand side of (4.37) tends in probability to the value $m(\eta(t))(Q(\eta(t))z + g(\eta(t)))$. Furthermore, $\mathbf{E}[\alpha_{nk} \mid \eta_{nk}] = 0$, and as $n \to \infty$ for any α, $\mathbf{E}[\alpha_{nk}\alpha^*_{nk} \mid \eta_{nk} = \alpha] \to D(\alpha)^2$, and relation (4.35) implies that the variables $|\alpha_{nk}|^2$ are uniformly integrable in each bounded region.

Let us introduce the random process $v_n(u) = v_{nk}$ as $k/n \le u < (k+1)/n$, $u \ge 0$. Then (4.37) and the results [GIK 75] imply that the sequence of processes $v_n(u)$ J-converges in the interval $[0, T]$ to the diffusion process $v(u)$ satisfying the following stochastic differential equation:

$$\begin{aligned} &\mathrm{d}v(u) = m\big(\eta(u)\big)\big(Q\big(\eta(u)\big)v(u) + g\big(\eta(u)\big)\big)\mathrm{d}u + D\big(\eta(u)\big)\mathrm{d}w(u), \\ &v(0) = \gamma_0. \end{aligned} \tag{4.39}$$

Note that

$$\left| s(t) - s\left(\frac{1}{n}t_{nk}\right) \right| \le \frac{1}{n}\tau_{nk} \sup_{\frac{1}{n}t_{nk} \le u \le \frac{1}{n}t_{nk+1}} \big|\tilde{b}\big(s(u)\big)\big|,$$

as $\frac{1}{n}t_{nk} \le t < \frac{1}{n}t_{nk+1}$. Thus, as $\mu_n(T) < \mu(T) + \varepsilon$,

$$\sup_{0 \le t \le T} \left| \gamma_n(t) - v_n\left(\mu_n(t) - \frac{1}{n}\right) \right| \le \frac{1}{\sqrt{n}} C_T \max_{k \le n(\mu(T)+\varepsilon)} \tau_{nk}, \tag{4.40}$$

where $C_T = \sup_{u \le \mu(T)+\varepsilon} |\tilde{b}(s(u))|$. Let us prove that for any $L > 0$,

$$\max_{k \le nL} \frac{1}{\sqrt{n}}\tau_{nk} \xrightarrow{\mathrm{p}} 0. \tag{4.41}$$

Denote $\widetilde{\tau}_{nk} = \tau_{nk}/\sqrt{n}$, $k \ge 0$, and let A_k be the sequence of the following events: $A_k = \{\widetilde{\tau}_{ni} < \varepsilon,\ i < k\}$, $k \ge 1$, $A_0 = \Omega$. Then for any $m > 0$,

$$\begin{aligned} \ln \mathbf{P}\Big(\max_{k \le m} \widetilde{\tau}_{nk} < \varepsilon\Big) &= \ln \prod_{k=0}^{m} \mathbf{P}\big(\widetilde{\tau}_{nk} < \varepsilon \mid A_k\big) \\ &= \sum_{k=0}^{m} \ln\big(1 - \mathbf{P}\big(\widetilde{\tau}_{nk} > \varepsilon \mid A_k\big)\big). \end{aligned} \tag{4.42}$$

Using the inequality $|\ln(1-z)+z| \le |z^2|$ as $|z| \le 1/2$, we find that

$$\bigl|\ln(1-z)\bigr| \le |z|\bigl(1+|z|\bigr) \quad \text{as } |z| \le 1/2. \tag{4.43}$$

Relations (4.42) and (4.43) imply that

$$\begin{aligned} &\Bigl|\ln \mathbf{P}\Bigl(\max_{k\le m} \widetilde{\tau}_{nk} < \varepsilon\Bigr)\Bigr| \\ &\qquad \le \Bigl(1 + \max_{k\le m} \mathbf{P}\bigl(\widetilde{\tau}_{nk} > \varepsilon \mid A_k\bigr)\Bigr) \sum_{k=0}^{m} \mathbf{P}\bigl(\widetilde{\tau}_{nk} > \varepsilon \mid A_k\bigr). \end{aligned} \tag{4.44}$$

It follows from condition (4.35) that for any fixed N and $\varepsilon > 0$ as $n \to \infty$,

$$\sup_{|\alpha|\le N} \mathbf{E}\tau_{n1}^2(n\alpha)\chi\bigl(\tau_{n1}(n\alpha) > \sqrt{n}\varepsilon\bigr) \longrightarrow 0. \tag{4.45}$$

In the region $\max_{k\le nC} |\eta_{nk}| \le N$ at $m \le nC$,

$$m \max_{k\le m} \mathbf{P}\bigl(\widetilde{\tau}_{nk} > \varepsilon \mid A_k\bigr) \le nC \sup_{|\alpha|\le N} \mathbf{P}\bigl\{\tau_{n1}(n\alpha) > \sqrt{n}\varepsilon\bigr\}.$$

According to (4.45),

$$\begin{aligned} n \sup_{|\alpha|\le N} \mathbf{P}\bigl\{\tau_{n1}(n\alpha) > \sqrt{n}\varepsilon\bigr\} &\le n \sup_{|\alpha|\le N} \mathbf{E}\left(\frac{\tau_{n1}(n\alpha)}{\sqrt{n}\varepsilon}\right)^2 \chi\bigl(\tau_{n1}(n\alpha) > \sqrt{n}\varepsilon\bigr) \\ &\le \sup_{|\alpha|\le N} \mathbf{E}\tau_{n1}^2(n\alpha)\chi\bigl(\tau_{n1}(n\alpha) > \sqrt{n}\varepsilon\bigr)/\varepsilon^2 \longrightarrow 0. \end{aligned}$$

Therefore, in this region as $m \le nC$,

$$\sum_{k=0}^{m} \mathbf{P}\bigl\{\widetilde{\tau}_{nk} > \varepsilon \mid A_k\bigr\} \le nC \max_{k} \mathbf{P}\bigl\{\widetilde{\tau}_{nk} > \varepsilon \mid A_k\bigr\} \longrightarrow 0,$$

and relations (4.44) and (4.45) imply that as $m \le nC$, $\mathbf{P}\{\max_{k\le m} \widetilde{\tau}_{nk} > \varepsilon\} \to 0$, and relation (4.41) is proved. Now using (4.40) we obtain

$$\sup_{0\le t\le T} \bigl|\gamma_n(t) - v_n\bigl(\mu_n(t) - 1/n\bigr)\bigr| \xrightarrow{\text{P}} 0.$$

The sequence of processes $v_n(\mu_n(t) - 1/n)$ J-converges to the process $v(\mu(t)) = \gamma(t)$. As far as $\mu'(t) = m(s(t))^{-1}$, we can calculate the stochastic differential for process $\gamma(t)$ using the formula $\mathrm{d}w(\mu(t)) \sim \sqrt{\mu'(t)}\mathrm{d}w(t)$, and obtain equation (4.36). $\square$

In conclusion to this section let us consider an important case when the process $S_n(t)$ is an homogenous MP. Suppose that $S_n(t)$ is a regular stepwise process and there exist transition rates $q_n(\alpha, A)$, $\alpha \in \mathcal{R}^r$, $A \in \mathcal{B}^r_{\mathcal{R}}$, $\alpha \neq A$ such that $q_n(\alpha) = q_n(\alpha, \mathcal{R}^r \backslash \{\alpha\}) < \infty$ for any $\alpha \in \mathcal{R}^r$. Let us define the independent families of random variables $\{\xi_{nk}(\alpha), \alpha \in \mathcal{R}^r\}$, $k \geq 0$, and $\{\tau_{nk}(\alpha), \alpha \in R\}$, $k \geq 0$, with values in $\mathcal{R}^r$ and $[0, \infty)$, respectively, where $\tau_{nk}(n\alpha)$ has an exponential distribution with parameter $q_n(\alpha)$ and

$$\mathbf{P}\big(\xi_{nk}(n\alpha) \in A\big) = q_n(\alpha)^{-1} q_n(\alpha, A + \alpha), \quad \alpha \neq A,$$

where $A + \alpha = \{z : z - \alpha \in A\}$. By definition RPSM which is defined by the families $\{(\zeta_{nk}(\cdot), \tau_{nk}(\cdot))\}$ is equivalent to an MP $S_n(t)$. Denote $m_n(\alpha) = q_n(\alpha)^{-1}$. We can easily verify that in the case where $\tau_n(\cdot)$ has an exponential distribution, $D^2_n(\alpha) = \mathbf{E}\xi_{n1}(n\alpha)\xi_{n1}(n\alpha)^*$.

COROLLARY 4.1. *If the conditions of Theorems 4.3 and 4.4 hold, then relation (4.19) takes place and the sequence of processes $\gamma_n(t)$ weakly converges to the diffusion process $\gamma(t)$ satisfying stochastic differential equation (4.36).*

Note that, as $\tau_{n1}(\cdot)$ has an exponential distribution, conditions (4.15) and (4.35) are automatically satisfied.

4.4. Averaging principle and diffusion approximation for recurrent processes of semi-Markov type (Markov case)

In this section we investigate the next level of complexity when an RPSM is switched by some external Markov process and consider the case of fast switching. Let at each $n \geq 0$,

$$F_{nk} = \big\{(\xi_{nk}(x, z), \tau_{nk}(x, z)),\ x \in X,\ z \in \mathcal{R}^r\big\}, \quad k \geq 0, \tag{4.46}$$

be jointly independent families of random variables with values in the space $R^r \times [0, \infty)$ and distributions not depending on $k \geq 0$, and let x_{ni}, $i \geq 0$, be an homogenous MP which is independent of F_{nk}, $k \geq 0$, with values in a space X and S_{n0} be the initial value. Note that the variables $\xi_{nk}(x, z)$ and $\tau_{nk}(x, z)$ can be dependent. We construct RPSM $(x_n(t), S_n(t))$, $t \geq 0$, according to section 1.2.3. Put $t_{n0} = 0$ and denote

$$S_{nk+1} = S_{nk} + \xi_{nk}\big(x_{nk}, S_{nk}\big),\ \ t_{nk+1} = t_{nk} + \tau_{nk}\big(x_{nk}, S_{nk}\big),\ \ k \geq 0. \tag{4.47}$$

Let

$$S_n(t) = S_{nk}, \quad x_n(t) = x_{nk} \quad \text{as } t_{nk} \leq t < t_{nk+1}. \tag{4.48}$$

The process $x_n(\cdot)$ stands for some external environment and in general it is not an MP or even an SMP as it depends on the values of a switching component S_{nk}.

Suppose that MP x_{nk}, $k \geq 0$, at each $n > 0$ has a stationary measure $\pi_n(A)$, $A \in \mathcal{B}_X$. Assume that the corresponding integrals exist and denote

$$\begin{aligned} m_n(x,\alpha) &= \mathbf{E}\tau_{n1}(x,n\alpha), \quad b_n(x,\alpha) = \mathbf{E}\xi_{n1}(x,n\alpha), \\ m_n(\alpha) &= \int_X m_n(x,\alpha)\pi_n(\mathrm{d}x), \quad b_n(\alpha) = \int_X b_n(x,\alpha)\pi_n(\mathrm{d}x). \end{aligned} \tag{4.49}$$

Let us introduce a strong mixing coefficient

$$\begin{aligned} \alpha_n(k) = \sup\big\{|P\{x_{ni} \in A,\ x_{n,i+k} \in B\} \\ - P\{x_{ni} \in A\}P\{x_{n,i+k} \in B\}\big| : A, B \in \mathcal{B}_X,\ i \geq 0\big\}. \end{aligned} \tag{4.50}$$

THEOREM 4.5 (Averaging principle). *Suppose that there exist a sequence of integers r_n such that*

$$n^{-1}r_n \longrightarrow 0, \quad \sup_{k\geq r_n} \alpha_n(k) \longrightarrow 0, \tag{4.51}$$

for any $N > 0$,

$$\begin{aligned} \lim_{L\to\infty} \limsup_{n\to\infty} \sup_{|\alpha|\leq N} \sup_{x} \big\{\mathbf{E}\tau_{n1}(x,n\alpha)\chi\big(\tau_{n1}(x,n\alpha) > L\big) \\ + \mathbf{E}\big|\xi_{n1}(x,n\alpha)\big|\chi\big(\big|\xi_{n1}(x,n\alpha)\big| > L\big)\big\} = 0, \end{aligned} \tag{4.52}$$

and for any x as $\max(|\alpha_1|, |\alpha_2|) \leq N$,

$$\begin{aligned} &\big|m_n\big(x,\alpha_1\big) - m_n\big(x,\alpha_2\big)\big| + \big|b_n\big(x,\alpha_1\big) - b_n\big(x,\alpha_2\big)\big| \\ &\leq C_N\big|\alpha_1 - \alpha_2\big| + \alpha_n(N), \end{aligned} \tag{4.53}$$

where C_N are bounded constants, $\alpha_n(N) \to 0$ uniformly in $|\alpha_1| \leq N$, $|\alpha_2| \leq N$, there exist functions $m(\alpha) > 0$ and $b(\alpha)$ such that for any $\alpha \in \mathcal{R}^r$,

$$m_n(\alpha) \to m(\alpha), \quad b_n(\alpha) \to b(\alpha), \tag{4.54}$$

and

$$n^{-1}S_{n0} \xrightarrow{\mathrm{P}} s_0, \tag{4.55}$$

where s_0 is some (possibly random) value. Then

$$\sup_{0\leq t\leq T} \big|n^{-1}S_n(nt) - s(t)\big| \xrightarrow{\mathrm{P}} 0, \tag{4.56}$$

where a function $s(t)$ is a solution of an ordinary differential equation

$$\mathrm{d}s(t) = m\big(s(t)\big)^{-1}b\big(s(t)\big)\mathrm{d}t, \quad s(0) = s_0, \tag{4.57}$$

and T satisfies the relation $y(+\infty) > T$ with probability one, where

$$y(t) = \int_0^t m\big(\eta(u)\big)\mathrm{d}u,$$

and $\eta(t)$ is a solution of an ordinary differential equation

$$\mathrm{d}\eta(u) = b\big(\eta(u)\big)\mathrm{d}u, \quad \eta(0) = s_0 \tag{4.58}$$

(it is assumed that a unique solution of equation (4.58) exists in each interval).

Proof. We follow the proof of Theorem 4.3. Let us introduce the sequences $\eta_{nk} = S_{nk}/n$, $y_{nk} = t_{nk}/n$, $k \geq 0$, and random processes $\eta_n(u) = \eta_{nk}$, $y_n(u) = y_{nk}$ as $k/n \leq u < (k+1)/n$, $u \geq 0$. Put $\mu_n(t) = \inf\{u : u > 0,\ y_n(u) > t\}$. The following representation is true:

$$n^{-1}S_n(nt) = \eta_n\big(n^{-1}\nu_n(t)\big) = \eta_n\big(\mu_n(t) - 1/n\big).$$

In this way RPSM $n^{-1}S_n(nt)$ is represented as a superposition of two processes: $\eta_n(t)$ and $\mu_n(t)$. First, we study the behavior of processes $\eta_n(t)$ and $y_n(t)$, then $\mu_n(t)$ and their superposition. According to equation (4.47) we can write the relations

$$\eta_{nk+1} = \eta_{nk} + \frac{1}{n} b_n\big(x_{nk}, \eta_{nk}\big) + \varphi_{nk}, \tag{4.59}$$

$$y_{nk+1} = y_{nk} + \frac{1}{n} m_n\big(x_{nk}, \eta_{nk}\big) + \psi_{nk}, \quad k \geq 0, \tag{4.60}$$

where

$$\varphi_{nk} = \frac{1}{n}\big(\xi_{nk}\big(x_{nk}, n\eta_{nk}\big) - b_n\big(x_{nk}, \eta_{nk}\big)\big),$$

$$\psi_{nk} = \frac{1}{n}\big(\tau_{nk}\big(x_{nk}, n\eta_{nk}\big) - m_n\big(x_{nk}, \eta_{nk}\big)\big).$$

Sequences φ_{nk} and ψ_{nk} are martingale-differences with respect to the flow of σ-algebras generated by the variables $\{\eta_{ni}, x_{ni},\ i \leq k\}$. Assume that condition (4.52) holds uniformly in $\alpha \in \mathcal{R}^r$. Using the results [GRI 73] we can prove that for any $t > 0$,

$$\max_{m \leq nt} \left| \sum_{k=1}^{m} \varphi_{nk} \right| \xrightarrow{\mathrm{P}} 0. \tag{4.61}$$

According to the results of Chapter 3, condition (4.51) implies that for any $\alpha \in \mathcal{R}^r$ and $t > 0$,

$$\frac{1}{n} \sum_{k=0}^{nt} b_n\big(x_{nk}, \alpha\big) - t b_n(\alpha) \xrightarrow{\mathrm{P}} 0. \tag{4.62}$$

Applying the results [ANI 86] on the convergence of stochastic difference schemes with random coefficients to solutions of ordinary differential equations and using relations (4.61) and (4.62) we find that for any $t > 0$,

$$\sup_{u\le t}\left|\eta_n(u) - \eta(u)\right| \xrightarrow{\mathrm{P}} 0. \tag{4.63}$$

By analogy we can prove that

$$\sup_{u\le t}\left|y_n(u) - y(u)\right| \xrightarrow{\mathrm{P}} 0. \tag{4.64}$$

As $m(\alpha) > 0$, process $y(t)$ increases strictly monotonically. Thus, process $y^{-1}(t) = \mu(t)$ exists for any t such that $y(+\infty) > t$ with probability one, is continuous and

$$\sup_{u\le t}\left|\mu_n(u) - \mu(u)\right| \xrightarrow{\mathrm{P}} 0. \tag{4.65}$$

Using the results [BIL 68] on U-convergence of the superposition of random functions and relations (4.63) and (4.65), we prove (4.56) where $s(t) = \eta(y^{-1}(t))$. Calculating the differential of $s(t)$ we obtain equation (4.57). In conclusion note that $\mathbf{P}\{\sup_{u\le t}|s(u)| > N\} \xrightarrow{\mathrm{P}} 0$ as $N \to \infty$. Thus, it is sufficient to check all conditions in each bounded region $|\alpha| \le N$. Finally Theorem 4.5 is proved. □

Now we study the conditions of the convergence of the sequence of processes

$$\kappa_n(t) = \frac{1}{\sqrt{n}}\big(S_n(nt) - ns(t)\big)$$

to some diffusion process. Let us introduce the uniformly strong mixing coefficient for process x_{nk}:

$$\varphi_n(r) = \sup_{x,y,A}\Big|P\big\{x_{nr}\in A \mid x_{n0} = x\big\} - P\big\{x_{nr}\in A \mid x_{n0} = y\big\}\Big|.$$

Put

$$\begin{gathered}
\tilde{b}_n(\alpha) = b_n(\alpha)m_n(\alpha)^{-1}, \quad \tilde{b}(\alpha) = b(\alpha)m(\alpha)^{-1},\\
\rho_{nk}(x,\alpha) = \xi_{nk}(x,n\alpha) - b_n(x,\alpha) - \tilde{b}(\alpha)\big(\tau_{nk}(x,n\alpha) - m_n(x,\alpha)\big),\\
D_n^2(x,\alpha) = \mathbf{E}\rho_{n1}(x,\alpha)\rho_{n1}(x,\alpha)^*,\\
\gamma_n(x,\alpha) = b_n(x,\alpha) - b_n(\alpha) - \tilde{b}(\alpha)\big(m_n(x,\alpha) - m_n(\alpha)\big).
\end{gathered} \tag{4.66}$$

THEOREM 4.6 (Diffusion approximation). *Suppose that for a fixed* $r > 0$ *and* $q \in [0, 1)$,

$$\varphi_n(r) \le q, \quad n > 0, \tag{4.67}$$

condition (4.53) with relation $\sqrt{n}\alpha_n(N) \to 0$ *holds, conditions (4.54), (4.55) are true, and for any* $N > 0$ *the following conditions are satisfied:*

1)

$$\begin{aligned} \lim_{L\to\infty} \lim_{n\to\infty} \sup_{|\alpha|\le N} \sup_x \big\{ \mathbf{E}\tau_{n1}(x, n\alpha)^2 \chi\big(\tau_{n1}(x, n\alpha) > L\big) \\ + \mathbf{E}\big|\xi_{n1}(x, n\alpha)\big|^2 \chi\big(\big|\xi_{n1}(x, n\alpha)\big| > L\big)\big\} = 0; \end{aligned} \tag{4.68}$$

2) as $\max(|\alpha_1|, |\alpha_2|) \le N$,

$$\big|D_n\big(x, \alpha_1\big)^2 - D_n\big(x, \alpha_2\big)^2\big| \le C_N\big|\alpha_1 - \alpha_2\big| + \alpha_n(N), \tag{4.69}$$

where $\alpha_n(N) \to 0$ *uniformly in* $|\alpha_1| \le N$, $|\alpha_2| \le N$;

3) there exists a function $q(\alpha, z)$ *such that for any* N *in the region* $|\alpha| \le N$,

$$\big|q(\alpha, z)\big| \le C_N\big(1 + |z|\big)$$

uniformly in $|\alpha| \le N$ *at each fixed* z,

$$\sqrt{n}\left(\tilde{b}_n\left(\alpha + \frac{1}{\sqrt{n}}z\right) - \tilde{b}(\alpha)\right) \longrightarrow q(\alpha, z), \tag{4.70}$$

and there exist functions $D(\alpha)$ *and* $B(\alpha)$ *such that for any* $\alpha \in \mathcal{R}^m$,

$$D_n^2(\alpha) = \int_X D_n^2(x, \alpha)\pi_n(\mathrm{d}x) \longrightarrow D^2(\alpha), \tag{4.71}$$

$$B_n^{(1)}(\alpha)^2 + B_n^{(2)}(\alpha)^2 \longrightarrow B(\alpha)^2, \tag{4.72}$$

where

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

and

$$B_n^{(2)}(\alpha)^2 = \sum_{k\ge 1} E\gamma_n\big(x_{n0}, \alpha\big)\gamma_n\big(x_{nk}, \alpha\big)^*,$$

with $P\{x_{n0} \in A\} = \pi_n(A)$, $A \in B_X$, *and also*

$$\kappa_n(0) \overset{\mathrm{w}}{\Longrightarrow} \kappa_0, \tag{4.73}$$

where κ_0 *is a proper random variable.*

Then for any $T > 0$ *satisfying the conditions of Theorem 4.5 the sequence of processes* $\kappa_n(t)$ *J-converges in the space* D_T^r *to the diffusion process* $\kappa(t)$ *satisfying the following stochastic differential equation:* $\kappa(0) = \kappa_0$,

$$\mathrm{d}\kappa(t) = q\big(s(t), \kappa(t)\big)\mathrm{d}t + m\big(s(t)\big)^{-\frac{1}{2}}\Big(D\big(s(t)\big)^2 + B\big(s(t)\big)^2\Big)^{\frac{1}{2}}\mathrm{d}w(t), \tag{4.74}$$

where $w(t)$ *is the standard Wiener process in* R^r, *and a solution of equation (4.74) exists and is unique.*

Proof. We keep the notation of Theorem 4.5 and denote

$$\tilde{\eta}_{nk} = s\big(y_{nk}\big), \quad \gamma_{nk} = \sqrt{n}\big(\eta_{nk} - \tilde{\eta}_{nk}\big), \quad \gamma_n(t) = \gamma_{n[nt]},$$
$$\tau_{nk} = \tau_{nk}\big(x_{nk}, n\eta_{nk}\big), \quad \xi_{nk} = \xi_{nk}\big(x_{nk}, n\eta_{nk}\big).$$

Conditions (4.53), (4.54) and relation $m(\alpha) > 0$ imply that function $\tilde{b}(\alpha)$ also satisfies the local Lipschitz condition. Therefore, using equation (4.57) after calculations we obtain the relation:

$$\gamma_{nk+1} = \gamma_{nk} + \alpha_{nk}\big(\gamma_{nk}\big)\frac{1}{n} + \beta_{nk}, \quad k \geq 0, \tag{4.75}$$

where

$$\begin{aligned}
\alpha_{nk}(z) &= m_{nk}\big(\eta_{nk}\big)\sqrt{n}\bigg(\tilde{b}_{nk}\big(\eta_{nk}\big) - \tilde{b}\bigg(\eta_{nk} - \frac{1}{n}z\bigg)\bigg),\\
\beta_{nk} &= \varphi_{nk} + \psi_{nk} + \delta_{nk},\\
\varphi_{nk} &= \frac{1}{\sqrt{n}}\Big(b_n\big(x_{nk}, \eta_{nk}\big) - b_{nk}\big(\eta_{nk}\big)\\
&\quad - \tilde{b}\big(\tilde{\eta}_{nk}\big)\big(m_n\big(x_{nk}, \eta_{nk}\big) - m_{nk}\big(\eta_{nk}\big)\big)\Big),\\
\psi_{nk} &= \frac{1}{\sqrt{n}}\Big(\xi_{nk} - b\big(x_{nk}, \eta_{nk}\big) - \tilde{b}\big(\tilde{\eta}_{nk}\big)\big(\tau_{nk} - m_n\big(x_{nk}, \eta_{nk}\big)\big)\Big),\\
b_{nk}(\alpha) &= \mathbf{E}b_n\big(x_{nk}, \alpha\big), \quad m_{nk}(\alpha) = \mathbf{E}m_n\big(x_{nk}, \alpha\big),
\end{aligned}$$

and $|\delta_{nk}| \leq \frac{C}{n\sqrt{n}}\tau_{nk}$.

It is known that for uniformly ergodic MP, condition (4.67) implies the relation

$$\varphi_n(k) \leq q^{k/r-1}, \quad k \geq 0. \tag{4.76}$$

Therefore, in each bounded region $|\alpha| \leq N$,

$$\big|b_n(\alpha) - b_{nk}(\alpha)\big| + \big|m_n(\alpha) - m_{nk}(\alpha)\big| \leq C_N q^{k/r}.$$

Thus, as the functions $\alpha_{nk}(z)$ also satisfy condition (4.53), following the lines of proof of Theorem 4.5 we find that for any $t > 0$, $z \in R^m$,

$$\frac{1}{n}\sum_{k=0}^{[nt]}\alpha_{nk}(z) \xrightarrow{\mathrm{P}} \int_0^t m\big(\eta(u)\big)q\big(\eta(u), z\big)\mathrm{d}u. \tag{4.77}$$

Let the sequence of variables γ_{nk} with values in R^r satisfy the relation:

$$\gamma_{nk+1} = \gamma_{nk} + a_{nk}\big(\gamma_{nk}\big)\frac{1}{n} + \beta_{nk},\ k \geq 0,$$

and the flow of σ-algebras σ_{nk} be given such that γ_{n0} is a σ_{n0}-measurable variable, functions $\alpha_{nk}(z)$ are σ_{nk}-measurable variables and β_{nk} are σ_{nk+1}-measurable. Define a stepwise process

$$\gamma_n(t) = \gamma_{nk} \quad \text{as } k/n \leq t < (k+1)/n,\ t \geq 0.$$

Let us formulate the following auxiliary lemma:

LEMMA 4.2. *Assume that*

1. for any $N > 0$ as $\max(|\alpha_1|, |\alpha_2|) \leq N$,

$$\big|a_{nk}\big(\alpha_1\big) - a_{nk}\big(\alpha_2\big)\big| \leq C_N\big|\alpha_1 - \alpha_2\big| + c_{nk}, \tag{4.78}$$

where for any $t > 0$, $n^{-1}\sum_{k=0}^{[nt]} c_{nk} \xrightarrow{\mathrm{P}} 0$;

2. for any $k > j \geq 0$,

$$\mathbf{E}\left|\beta_n\left(\frac{k}{n}\right) - \beta_n\left(\frac{j}{n}\right)\right|^2 \leq C\frac{k-j}{n}, \tag{4.79}$$

where $\beta_n(t) = \sum_{k=0}^{[nt]}\beta_{nk}$;

3. there exists a non-random function $g(u, z)$ such that for any $u \leq T$, $|g(u, z)| \leq C_T(1 + |z|)$;

4. for any $t > 0$, $z \in \mathcal{R}^r$,

$$\frac{1}{n}\sum_{k=0}^{[nt]}\alpha_{nk}(z) \xrightarrow{\mathrm{P}} \int_0^t g(u, z)\mathrm{d}u; \tag{4.80}$$

5. there also exists a proper random variable γ_0 and a process $\beta(t)$ given on the same probabilistic space such that for any $T > 0$, the sequence $(\gamma_{n0}, \beta_n(t))$ weakly converges in the space D_T^r to the pair $(\gamma_0, \beta(t))$.

Then the sequence of processes $\gamma_n(t)$ J-converges in the space D_T^r to the process $\gamma(t)$ satisfying the following stochastic differential equation:

$$\mathrm{d}\gamma(t) = g\big(t, \gamma(t)\big)\mathrm{d}t + \mathrm{d}\beta(t), \quad \gamma(0) = \gamma_0, \tag{4.81}$$

a solution to which exists and is unique.

Proof. The proof follows the same lines as in [ANI 89]. First it is easy to prove that equation (4.79) implies the inequality $\mathbf{E}|\beta(t) - \beta(s)|^2 \leq C|t-s|$, and it follows from equation (4.78) that the function $g(u,z)$ satisfies a local Lipschitz condition with respect to z. Therefore, the solution of equation (4.81) exists and is unique. Let us construct a random process $\alpha_n(t) = \alpha_{nk}$ as $k/n \leq t < (k+1)/n$, where

$$\alpha_{nk+1} = \alpha_{nk} + \int_{\frac{k}{n}}^{\frac{k+1}{n}} g\big(u, \alpha_{nk}\big)du + \beta_{nk}.$$

Following the standard arguments [GIK 78] we can prove that the measures generated by the processes $\gamma_n(\cdot)$ satisfy the condition of weak compactness in space D_T^r and according to the conditions of Lemma 4.2, $\sup_{t\leq T}|\gamma_n(t) - \alpha_n(t)| \xrightarrow{\mathrm{P}} 0$, and in addition the finite-dimensional distributions of process $\alpha_n(t)$ weakly converge to the distributions of process $\gamma(t)$. This completes the proof of Lemma 4.2. □

Let us now study the behavior of the processes

$$\varphi_n(t) = \sum_{k=0}^{[nt]} \varphi_{nk}, \quad \psi_n(t) = \sum_{k=o}^{[nt]} \psi_{nk}, \quad t \geq 0.$$

Process $\psi_n(t)$ is a martingale. First we prove that

$$\sum_{k=0}^{[nt]} \mathbf{E}\big[\psi_{nk}\psi_{nk}^* \mid \sigma_{nk}\big] \xrightarrow{\mathrm{P}} \int_0^t D\big(\eta(u)\big)^2 \mathrm{d}u. \tag{4.82}$$

Then this relation according to condition (4.68) and the results of [LIP 89], implies that process $\psi_n(t)$ J-converges in D_T^r to martingale $\psi(t)$ which can be represented in the form $\int_0^t D(\eta(u))\mathrm{d}w(u)$.

Indeed, Theorem 4.5 implies that if $n \to \infty$ in such a way that $k/n \to t$, then $\tilde{\eta}_{nk} \to s(y(t)) = \eta(y^{-1}(y(t))) = \eta(t)$. Using the continuity of function $\tilde{b}(\alpha)$ and conditions (4.69) and (4.71), it is not difficult to prove that equation (4.82) is true.

Now consider process $\varphi_n(t)$. Let us introduce the process

$$\tilde{\varphi}_n(t) = \frac{1}{\sqrt{n}} \sum_{k=o}^{[nt]} \gamma_n\big(x_{nk}, \eta_{nk}\big), \quad t \geq 0.$$

Note that processes ψ_n and $\tilde{\varphi}_n(t)$ are asymptotically independent. Using condition (4.72), a local Lipschitz condition of the type (4.53) for variables $\gamma_n(x,\alpha)$ in equation (4.66), a strong mixing condition (4.76) and the results of Chapter 4 (see also [ANI 88]) we prove that process $\tilde{\varphi}_n(t)$ J-converges in $\mathcal{D}_T^r$ to martingale $\varphi(t)$ which can be represented in the form

$$\varphi(t) = \int_0^t B\big(\eta(u)\big)\mathrm{d}w(u).$$

Furthermore, let us prove that

$$\sup_{t \le T} \big|\varphi_n(t) - \tilde{\varphi}_n(t)\big| \xrightarrow{\mathrm{P}} 0. \tag{4.83}$$

Consider for $\varepsilon > 0$ the event

$$A_\varepsilon(t) = \Big\{ \sup_{k \le nt} \big|\eta_{nk} - \eta(k/n)\big| \le \varepsilon, \sup_{k \le nt} \big|\tilde{\eta}_{nk} - \eta(k/n)\big| \le \varepsilon \Big\},$$

and let $\chi_\varepsilon(t)$ be the indicator of this event. The results above imply that as $n \to \infty$, $\mathbf{P}\{A_\varepsilon(t)\} \to 1$. Now let us choose $\varepsilon_n \to 0$ is such a way that $\mathbf{P}\{A_{\varepsilon_n}(t)\} \to 1$. Note that according to the definition of processes $\eta_n(t)$ and $y_n(t)$ we may choose ε_n in the form $\varepsilon_n = L_n/\sqrt{n}$, where $L_n \to \infty$. Then

$$\begin{aligned} &\mathbf{P}\Big\{ \sup_{t \le T} \big|\varphi_n(t) - \tilde{\varphi}_n(t)\big| > \delta \Big\} \\ &\quad \le \mathbf{P}\Big\{ \sup_{t \le T} \big|\varphi_n(t) - \tilde{\varphi}_n(t)\big| > \delta, A_{\varepsilon_n}(T) \Big\} + P\big\{\overline{A_{\varepsilon_n}}(T)\big\}. \end{aligned}$$

Consider process $\delta_n(t) = (\varphi_n(t) - \tilde{\varphi}_n(t))\chi_{\varepsilon_n}(T)$. Assume without loss of generality that $\mathbf{P}\{x_{n0} \in A\} = \pi_n(A)$. Let $\tilde{z}_k$ and z_k, $k \ge 0$, be some non-random sequences in $\mathcal{R}^r$. Denote

$$\varphi_{nk}\big(z_k, \tilde{z}_k\big) = \frac{1}{\sqrt{n}}\Big(b_n\big(x_{nk}, z_k\big) - b_n\big(z_k\big) - \tilde{b}\big(\tilde{z}_k\big)\big(m_n\big(x_{nk}, z_k\big) - m_n\big(z_k\big)\big)\Big),$$

$$\varphi_n\big(t, z(\cdot)\big) = \sum_{k=0}^{[nt]} \varphi_{nk}\big(z_k, \tilde{z}_k\big) - \tilde{\varphi}_n(t).$$

Note that for any function $\varphi(u, v) > 0$, random vector (ξ, η) and region Q,

$$\mathbf{E}\varphi(\xi, \eta)\chi\big((\xi, \eta) \in Q\big) \le \sup_{u \in \tilde{Q}} \mathbf{E}\varphi(u, \eta)\chi\big((\xi, \eta) \in Q\big) \le \sup_{u \in \tilde{Q}} \mathbf{E}\varphi(u, \eta),$$

where $\tilde{Q} = \{u : (u, v) \in Q\}$. Using this inequality we find that

$$\begin{aligned} \mathbf{E}\delta_n(t)\delta_n(t)^* \le \sup \Big\{ &\mathbf{E}\varphi_n\big(t, z(\cdot)\big)\varphi_n\big(t, z(\cdot)\big)^* : \\ &\big|z_k - \eta(k/n)\big| \le \varepsilon_n,\ \big|\tilde{z}_k - \eta(k/n)\big| \le \varepsilon_n,\ k \le nt \Big\}. \end{aligned}$$

Note that by definition $\mathbf{E}\varphi_{nk}(z_k, \tilde{z}_k) = 0$. Then, using the local Lipschitz condition we can prove that in the region $|z_k - \eta(k/n)| \le \varepsilon_n$, $|\tilde{z}_k - \eta(k/n)| \le \varepsilon_n$, the following relation is true

$$E\left|\left(\varphi_{nk}\big(z_k, \tilde{z}_k\big) - \frac{1}{\sqrt{n}}\gamma_n\big(x_{nk}, \eta(k/n)\big)\left(\varphi_{nk}\big(z_k, \tilde{z}_k\big) - \frac{1}{\sqrt{n}}\gamma_n\big(x_{nk}, \eta(k/n)\big)\right)\right)^*\right|$$
$$\le \frac{1}{n}C\varepsilon_n.$$

According to the results of Chapter 4 (see also [ANI 88]) this relation implies that

$$\mathbf{E}\delta_n(t) \longrightarrow 0, \quad \mathbf{E}\delta_n(t)\delta_n(t)^* \longrightarrow 0. \tag{4.84}$$

As the measures generated by the processes $\delta_n(t)$ satisfy the condition of the weak compactness in the space $\mathcal{D}_T^r$, relations (4.84) then imply $\sup_{t\le T}|\delta_n(t)| \to 0$, and relation (4.83) is true.

Finally, we prove that the sequence of processes $\varphi_n(t)+\psi_n(t)$ J-converges in D_T^r to a process which can be represented in the form

$$\int_0^t D\big(\eta(u)\big)\mathrm{d}w_1(u) + \int_0^t B\big(\eta(u)\big)\mathrm{d}w_2(u),$$

where $w_1(\cdot)$ and $w_2(\cdot)$ are two independent Wiener processes, or in the equivalent form

$$\int_0^t \Big(D\big(\eta(u)\big)^2 + B\big(\eta(u)\big)^2\Big)^{\frac{1}{2}}\mathrm{d}w(u).$$

Therefore, all conditions of Lemma 4.2 are satisfied and the sequence of processes $\gamma_n(t)$ J-converges in $\mathcal{D}_T^r$ to process $\gamma(t)$, where $\gamma(0) = \kappa_0$ and

$$\mathrm{d}\gamma(t) = m\big(\eta(t)\big)q\big(\eta(t), \gamma(t)\big)\mathrm{d}t + \Big(D\big(\eta(t)\big)^2 + B\big(\eta(t)\big)^2\Big)^{\frac{1}{2}}\mathrm{d}w(t).$$

Furthermore, as $\frac{1}{n}t_{nk} \le t < \frac{1}{n}t_{nk+1}$,

$$\big|\kappa_n(t) - \gamma_n\big(\mu_n(t) - 1/n\big)\big| \le \frac{1}{\sqrt{n}}\tau_{nk} \sup_{\frac{1}{n}t_{nk}\le u<\frac{1}{n}t_{nk+1}} \big|\tilde{b}(u)\big|.$$

It follows from Theorem 4.5 that $\mu_n(T) \to \mu(T)$. Then, as $\mu_n(T) < \mu(T) + \delta$,

$$\sup_{t\le T}\big|\kappa_n(t) - \gamma_n\big(\mu_n(t) - 1/n\big)\big| \le C_T\frac{1}{\sqrt{n}}\max_{k/n\le\mu(T)+\delta}\tau_{nk}.$$

Condition (4.68) implies that as $n \to \infty$

$$n \sup_{|\alpha|\le N} P\big\{\tau_{n1}(x,n\alpha) > \sqrt{n}\varepsilon\big\} \\ \le \varepsilon^{-2} \sup_{|\alpha|\le N} E\tau_{n1}(x,n\alpha)^2 \chi\big(\tau_{n1}(x,n\alpha) > \sqrt{n}\varepsilon\big) \longrightarrow 0. \tag{4.85}$$

Therefore, (4.85) implies that for any N in the region $|\alpha| \le N$,

$$\mathbf{P}\bigg\{\frac{1}{\sqrt{n}} \max_{k/n\le L} \tau_{nk} > \varepsilon\bigg\} \le \sum_{k=0}^{[nL]} P\big\{\tau_{nk} > \sqrt{n}\varepsilon\big\} \\ \le nL \max_{k\le nL} P\big\{\tau_{nk} > \sqrt{n}\varepsilon\big\} \longrightarrow 0. \tag{4.86}$$

Finally, using the results on the convergence of the superposition of random functions [BIL 68] we prove that the sequence of processes $\kappa_n(t)$ weakly converges to the process $\kappa(t) = \gamma(\mu(t))$. Calculating the differential of $\kappa(t)$ according to the relation $\mathrm{d}w(\mu(t)) \sim \sqrt{\mu'(t)}\mathrm{d}w(t)$ we obtain equation (4.74) and Theorem 4.6 is proved. □

4.4.1. *Averaging principle and diffusion approximation for SMP*

Now consider the case when distributions of the variables $(\xi_{nk}(x,z), \tau_{nk}(x,z))$ in equation (4.46) do not depend on the argument z. In this case, we introduce the family of random variables $F_{nk} = \{(\xi_{nk}(x), \tau_{nk}(x)),\ x \in X\}$, $k \ge 0$, with values in $\mathcal{R}^r \times [0,\infty)$ and distributions not depending on index k. Process $x_n(t)$ defined by equation (4.48) is an SMP given by the embedded MP x_{nk} and the sojourn time in state x, $\tau_{nk}(x)$. Correspondingly, process $S_n(t)$, defined by equation (4.48), represents the sum of random variables $\xi_{nk}(x)$ defined on the trajectory of an SMP $x_n(\cdot)$ in the interval $[0,t]$.

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

$$m_n(x) = \mathbf{E}\tau_{n1}(x), \quad b_n(x) = \mathbf{E}\xi_{n1}(x), \\ m_n = \int_X m_n(x)\pi_n(\mathrm{d}x), \quad b_n = \int_X b_n(x)\pi_n(\mathrm{d}x). \tag{4.87}$$

As a consequence of Theorems 4.5 and 4.6 we obtain the following results.

COROLLARY 4.2. *Suppose that conditions (4.51) and (4.55) are true,*

$$\lim_{L\to\infty} \limsup_{n\to\infty} \sup_x \big\{\mathbf{E}\tau_{n1}(x)\chi\big(\tau_{n1}(x) > L\big) \\ + \mathbf{E}\big|\xi_{n1}(x)\big|\chi\big(\big|\xi_{n1}(x)\big| > L\big)\big\} = 0, \tag{4.88}$$

and $m_n \to m > 0$, $b_n \to b$.

Then relation (4.56) holds for any $T > 0$, where $s(t) = s_0 + tb/m$.

Now we study the conditions of the convergence of the process

$$\kappa_n(t) = \frac{1}{\sqrt{n}}\big(S_n(nt) - ns(t)\big)$$

with $s(t) = s_0 + tb/m$ to a diffusion process. Denote

$$\begin{aligned} \rho_{nk}(x) &= \xi_{nk}(x) - b_n(x) - \big(\tau_{nk}(x) - m_n(x)\big)b/m, \\ D_n^2(x) &= \mathbf{E}\rho_{n1}(x)\rho_{n1}(x)^*, \\ \gamma_n(x) &= b_n(x) - b_n - \big(m_n(x) - m_n\big)b/m. \end{aligned} \tag{4.89}$$

COROLLARY 4.3. *Suppose that the assumptions of Corollary 4.2 are satisfied where condition (4.51) is replaced by condition (4.67),*

$$\begin{aligned} \lim_{L\to\infty} \limsup_{n\to\infty} \sup_x \Big\{ &\mathbf{E}\tau_{n1}^2(x)\chi\big(\tau_{n1}(x) > L\big) \\ &+ \mathbf{E}\big|\xi_{n1}(x)\big|^2\chi\big(\big|\xi_{n1}(x)\big| > L\big)\Big\} = 0; \end{aligned} \tag{4.90}$$

and there exist matrices D^2 and B^2 such that

$$D_n^2 = \int_X D_n^2(x)\pi_n(\mathrm{d}x) \longrightarrow D^2, \quad B_{n1}^2 + B_{n2}^2 \longrightarrow B^2, \tag{4.91}$$

where $B_{n1}^2 = \int_X \gamma_n(x)\gamma_n(x)^*\pi_n(\mathrm{d}x)$, and $B_{n2}^2 = \sum_{k\geq 1} E\gamma_n(x_{n0})\gamma_n(x_{nk})^*$, with $P\{x_{n0} \in A\} = \pi_n(A)$, $A \in B_X$, and also $\kappa_n(0) \overset{\mathrm{w}}{\Longrightarrow} \kappa_0$, where κ_0 is a proper random variable.

Then for any $T > 0$ the sequence of processes $\kappa_n(t)$ J-converges in the space D_T^r to the diffusion process

$$\kappa(t) = m^{-\frac{1}{2}}\big(D^2 + B^2\big)^{\frac{1}{2}} w(t).$$

This means that $\kappa(t)$ is a Wiener process in $\mathcal{R}^r$ with mean zero and covariance matrix $(D^2 + B^2)/m$.

4.5. Averaging principle for RPSM with feedback

Let us prove AP for a general RPSM. For any $n > 0$, let $F_{nk} = \{(\xi_{nk}(x,\alpha), \tau_{nk}(x,\alpha), \beta_{nk}(x,\alpha)),\ x \in X,\ \alpha \in R^r\}$, $k \geq 0$, be jointly independent families of random vectors with values in the space $\mathcal{R}^r \times [0,\infty) \times X$, where X is a measurable space. In addition, let (x_{n0}, S_{n0}) be the initial value which is independent of F_{nk}, $k \geq 0$. We put

$$\begin{gathered} t_{n0} = 0, \quad t_{nk+1} = t_{nk} + \tau_{nk}\big(x_{nk}, S_{nk}\big), \\ S_{nk+1} = S_{nk} + \xi_{nk}\big(x_{nk}, S_{nk}\big), \quad x_{nk+1} = \beta_{nk}\big(x_{nk}, S_{nk}\big), \quad k \geq 0, \end{gathered} \tag{4.92}$$

and define

$$S_n(t) = S_{nk}, \quad x_n(t) = x_{nk}, \quad \text{as } t_{nk} \leq t < t_{nk+1},\ t \geq 0. \tag{4.93}$$

The pair $(x_n(t), S_n(t))$, $t \geq 0$, according to the definition in section 1.2.4 and relations (1.12), (1.13), is a general RPSM with feedback between both components. Suppose for simplicity that the distributions of families F_{nk} do not depend on index $k \geq 0$, and let there be moment functions

$$m_n(x, \alpha) = \mathbf{E}\tau_{n1}(x, n\alpha), \quad b_n(x, \alpha) = \mathbf{E}\xi_{n1}(x, n\alpha).$$

Denote $p_n(x, A, \alpha) = \mathbf{P}\{\beta_{n1}(x, \alpha) \in A\}$, $x \in X$, $A \in B_X$, $\alpha \in R^r$, and let for any fixed α, $\tilde{x}_{nk}(\alpha)$, $k \geq 0$, be an auxiliary MP in X with transition probabilities

$$\mathbf{P}\big\{\tilde{x}_{nk+1}(\alpha) \in A \mid \tilde{x}_{nk}(\alpha) = x\big\} = p_n(x, A, \alpha).$$

Suppose that there exists a family of transition probabilities $q(x, A, \alpha)$, $x \in X$, $A \in B_X$, $\alpha \in R^r$, where the function $q(x, A, \alpha)$ is uniformly continuous in α in each bounded region $|\alpha| \leq L$ uniformly in $x \in X$, $A \in B_X$, and let for any $L > 0$,

$$\sup_{x, A, |\alpha| \leq L} \big|p_n(x, A, \alpha) - q(x, A, \alpha)\big| \longrightarrow 0. \tag{4.94}$$

Furthermore, suppose that an MP $\tilde{x}_{nk}(\alpha)$, $k \geq 0$, is uniformly ergodic with stationary measure $\pi(A, \alpha)$ uniformly in α in each bounded region and in $n > 0$. Denote

$$m_n(\alpha) = \int_X m_n(x, \alpha)\pi_n(\mathrm{d}x, \alpha), \quad b_n(\alpha) = \int_X b_n(x, \alpha)\pi_n(\mathrm{d}x, \alpha).$$

THEOREM 4.7. *Suppose that equation (4.94) holds and:*

1) for any fixed $N > 0$*,*

$$\begin{aligned} \lim_{L\to\infty} \limsup_{n\to\infty} \sup_{|\alpha|<N} \sup_{x} \Big\{ & \mathbf{E}\tau_{n1}(x, n\alpha)\chi\big(\tau_{n1}(x, n\alpha) > L\big) \\ & + \mathbf{E}\big|\xi_{n1}(x, n\alpha)\big|\chi\big(\big|\xi_{n1}(x, n\alpha)\big| > L\big)\Big\} = 0; \end{aligned}$$

2) for any x *as* $\max(|\alpha_1|, |\alpha_2|) \leq N$*,*

$$\big|m_n\big(x, \alpha_1\big) - m_n\big(x, \alpha_2\big)\big| + \big|b_n\big(x, \alpha_1\big) - b_n\big(x, \alpha_2\big)\big| \leq C_N\big|\alpha_1 - \alpha_2\big| + \alpha_n(N),$$

where C_N *are some constants, and* $\alpha_n(N) \to 0$ *uniformly in* $|\alpha_1| \leq N$*,* $|\alpha_2| \leq N$*;*

3) there exist functions $b(\alpha)$ *and* $m(\alpha) > 0$*, and a variable* s_0 *(possibly random) such that as* $n \to \infty$*,*

$$b_n(\alpha) \longrightarrow b(\alpha), \quad m_n(\alpha) \longrightarrow m(\alpha), \quad \alpha \in R^r,$$

and $n^{-1} S_{n0} \xrightarrow{P} s_0$*.*

Then

$$\sup_{0 \le t \le T} \left| n^{-1} S_n(nt) - s(t) \right| \xrightarrow{P} 0,$$

where

$$s(0) = s_0, \quad \mathrm{d}s(t) = m\big(s(t)\big)^{-1} b\big(s(t)\big)\,\mathrm{d}t, \tag{4.95}$$

and T *is any positive number such that* $y(+\infty) > T$ *with probability one where*

$$y(t) = \int_0^t m\big(\eta(u)\big)\mathrm{d}u, \tag{4.96}$$

and $\eta(0) = s_0$, $\mathrm{d}\eta(u) = b(\eta(u))\mathrm{d}u$*.*

The proof of this result follows from the averaging principle for general switching recurrent sequences and switching processes with feedback [ANI 91a, ANI 92a].

4.6. Averaging principle and diffusion approximation for switching processes

In this section we study the AP and DA for SPs in the series scheme and in the case of fast switches. Let at each $n > 0$,

$$F_{nk} = \big\{ \big(\zeta_{nk}(t,x,\alpha), \tau_{nk}(x,\alpha), \beta_{nk}(x,\alpha)\big),\ t > 0,\ x \in X,\ \alpha \in \mathcal{R}^r \big\}, \quad k \ge 0,$$

be the families of random processes $\zeta_{nk}(\cdot)$ which are jointly independent in k with values in $\mathcal{R}^r$ and trajectories belonging to Skorokhod space and random variables $(\tau_{nk}(\cdot), \beta_{nk}(\cdot))$ with values in $[0,\infty) \times \mathcal{R}^r$, and let (x_{n0}, S_{n0}) be the initial value. These families determine the sequence of SPs $(x_n(t), \zeta_n(t))$, $t \ge 0$ and the sequence of RPSMs $S_n(t)$ according to relations (1.3)–(1.5), section 1.2.1.

First let us establish the results on the asymptotic closeness of trajectories of processes $\zeta_n(\cdot)$ and $S_n(\cdot)$. Denote

$$\nu_n(t) = \min\big\{k : k > 0,\ t_{nk+1} > nt\big\}, \tag{4.97}$$

$$g_{nk}(x,\alpha) = \sup_{t < \tau_{nk}(x,\alpha)} \big|\zeta_{nk}(t,x,\alpha)\big|. \tag{4.98}$$

THEOREM 4.8. *Suppose that there exist non-random sequences $v_n \to \infty$ and c_n such that for some $T > 0$,*

$$\lim_{L\to\infty} \limsup_{n\to\infty} \mathbf{P}\big(v_n^{-1}\nu_n(T) > L\big) = 0, \tag{4.99}$$

and for any $\varepsilon > 0$, $L > 0$,

$$\lim_{n\to\infty} \sum_{k=0}^{[v_n L]} \sup_{x,\alpha} \mathbf{P}\big(c_n g_{nk}(x,\alpha) > \varepsilon\big) = 0. \tag{4.100}$$

Then

$$\lim_{n\to\infty} \mathbf{P}\left(c_n \sup_{t\le T} \big|\zeta_n(nt) - S_n(nt)\big| > \varepsilon \right) = 0. \tag{4.101}$$

Proof. By definition of an SP we obtain for any fixed $L > 0$,

$$\begin{aligned} &\mathbf{P}\left(c_n \sup_{t\le T} \big|\zeta_n(nt) - S_n(nt)\big| > \varepsilon \right) \\ &\quad\le \mathbf{P}\left(c_n \max_{k\le \nu_n(T)+1} g_{nk}\big(x_{nk}, S_{nk}\big) > \varepsilon \right) \\ &\quad\le \mathbf{P}\left(c_n \max_{k\le [v_n L]} g_{nk}\big(x_{nk}, S_{nk}\big) > \varepsilon \right) + \mathbf{P}\big(\nu_n(T) > \big[v_n L\big]\big). \end{aligned} \tag{4.102}$$

Denote $g_{nk} = c_n g_{nk}(x_{nk}, S_{nk})$ and introduce the events:

$$A_{nk} = \big\{g_{ni} \le \varepsilon,\ i < k\big\}, \quad k > 0, \qquad A_{n0} = \Omega \text{ – certain event.}$$

Using inequalities (4.42)–(4.44) and taking into account condition (4.100) we get that as $n \to \infty$,

$$\begin{aligned} &\left| \ln \mathbf{P}\left(\max_{k\le v_n L} g_{nk} \le \varepsilon \right) \right| \\ &\quad\le \left(1 + \max_{k\le v_n L} \sup_{x,\alpha} \mathbf{P}\big(c_{nk} g_{nk}(x,\alpha) > \varepsilon\big) \right) \sum_{k=0}^{[\nu_n T]} \sup_{x,\alpha} \mathbf{P}\big(c_n g_{nk}(x,\alpha) > \varepsilon\big) \longrightarrow 0. \end{aligned}$$

These relations according to equations (4.99) and (4.102) imply the statement of Theorem 4.8. □

Note that it is sufficient if condition (4.100) is satisfied for any fixed N in the region $|\alpha| \le nN$.

Theorem 4.8 shows that under simple natural assumptions the trajectories of an SP and a simple RPSM in the case of fast switching are asymptotically close. Therefore, it is sufficient to analyze the behavior of RPSM $S_n(t)$ which generally has a more simple structure.

Now we will combine the results of Theorem 4.8 and Theorems 4.3, 4.4, 4.5 and 4.6 and prove the AP and DA for SPs.

Consider two cases: the case when there is no additional Markov switching (the process is switched according to its operation) and the case when there is an external Markov environment.

Let at each $n > 0$, $F_{nk} = \{(\zeta_{nk}(t,\alpha), \tau_{nk}(\alpha)),\ t \geq 0,\ \alpha \in \mathcal{R}^m\}$, $k \geq 0$, be the families of random processes which are jointly independent in k with values in $\mathcal{R}^r$ and trajectories belonging to Skorokhod space and random variables with values in $[0,\infty)$, S_{n0} be the initial value. Denote $\xi_{nk}(\alpha) = \zeta_{nk}(\tau_{nk}(\alpha),\alpha)$ and construct RPSM $S_n(t)$ according to formulae (4.13), (4.14) and also an SP $\zeta_n(t)$:

$$\zeta_n(t) = S_{nk} + \zeta_n\big(t - t_{nk}, S_{nk}\big) \quad \text{as } t_{nk} \leq t < t_{nk+1},\ t \geq 0.$$

Suppose for simplicity that the distributions of families F_{nk} do not depend on index k. Denote $g_n(\alpha) = \sup_{t<\tau_{n1}(\alpha)} |\zeta_{n1}(t,\alpha)|$.

THEOREM 4.9 (AP). *Suppose that the conditions of Theorem 4.3 hold and for any $\varepsilon > 0$, $N > 0$,*

$$\lim_{n\to\infty} n \sup_{|\alpha|\leq Nn} \mathbf{P}\big(c_n g_n(\alpha) > \varepsilon\big) = 0, \tag{4.103}$$

where $c_n = 1/n$. Then

$$\sup_{0\leq t\leq T} \big|n^{-1}\zeta_n(nt) - s(t)\big| \xrightarrow{\text{P}} 0. \tag{4.104}$$

THEOREM 4.10 (DA). *Suppose that the conditions of Theorem 4.4 hold and equation (4.103) takes place where $c_n = 1/\sqrt{n}$. Then the sequence of processes*

$$\tilde{\gamma}_n(t) = \frac{1}{\sqrt{n}}\big(\zeta_n(nt) - ns(t)\big)$$

J-converges to the diffusion process $\gamma(t)$ satisfying equation (4.36).

Proof. If the conditions of Theorem 4.3 hold, then as follows from relation (4.29) $n^{-1}\nu_n(T) \xrightarrow{\text{P}} \mu(T)$. This relation implies condition (4.99) with $v_n = n$. Now according to Theorem 4.8 relation (4.101) with $c_n = 1/n$ is true and relation (4.104) follows from relations (4.101) and (4.19). By analogy Theorem 4.10 is proved. □

Consider a special case which is usually seen in queueing models when the trajectory of process $\zeta_{nk}(t,\alpha)$ is monotonic in the interval $[0,\tau_{nk}(\alpha))$. In this case $g_n(\alpha) = |\xi_{n1}(\alpha)|$. The following statement is true.

STATEMENT 4.1. *If the trajectory of $\zeta_{nk}(t,\alpha)$ for any α is monotone in interval $[0,\tau_{nk}(\alpha))$ with probability one, then relation (4.103) in both theorems is automatically satisfied. Thus, the conditions of Theorem 4.3 imply relation (4.104) and correspondingly the conditions of Theorem 4.4 imply J-convergence of the process $\widetilde{\gamma}(t)$ to $\gamma(t)$.*

Proof. Indeed, in this case

$$n^{-1}\sup_{t\le T}\big|\zeta_n(nt) - S_n(nt)\big| \le n^{-1}\max_{k\le\nu_n(t)}\big|\xi_{nk}\big(S_{nk}\big)\big|. \tag{4.105}$$

However, the uniform convergence in equation (4.19) automatically implies that the value of the maximum jump of the process $S_n(nt)/n$ in the interval $[0,T]$ tends in probability to zero. This implies that the right-hand side in equation (4.105) tends in probability to zero which is equivalent to relation (4.103). The same conclusion is valid for Theorem 4.4 as J-convergence to continuous process $\gamma(t)$ automatically implies U-convergence. □

Now consider the case of additional Markov switching. Let at each $n \ge 0$, $F_{nk} = \{(\zeta_{nk}(t,x,\alpha),\tau_{nk}(x,\alpha)),\ x \in X,\ \alpha \in \mathcal{R}^r\}$, $k \ge 0$, be the jointly independent in k families of random processes $\zeta_{nk}(\cdot)$ with values in $\mathcal{R}^r$ and trajectories belonging to Skorokhod space and random variables $\tau_{nk}(\cdot)$ with values in $[0,\infty)$ with distributions not depending on index k. Also let x_{ni}, $i \ge 0$, be the independent of F_{nk}, $k \ge 0$ homogenous MP with values in X and S_{n0} be the initial value. These families determine an SP $(x_n(t), S_n(t))$ in the following way. Put

$$t_{n0} = 0,\quad t_{nk+1} = t_{nk} + \tau_{nk}\big(x_{nk}, S_{nk}\big),$$
$$S_{nk+1} = S_{nk} + \xi_{nk}\big(x_{nk}, S_{nk}\big),\quad k \ge 0,$$

where $\xi_{nk}(x,\alpha) = \zeta_{nk}(\tau_{nk}(x,\alpha),x,\alpha)$, and set

$$\zeta_n(t) = S_{nk} + \zeta_{nk}\big(t - t_{nk}, x_{nk}, S_{nk}\big),$$
$$x_n(t) = x_{nk}\quad \text{as } t_{nk} \le t < t_{nk+1},\ t \ge 0.$$

THEOREM 4.11. *Let the conditions of Theorem 4.5 hold and for any $\varepsilon > 0$, $N > 0$,*

$$\lim_{n\to\infty}\sup_{|\alpha|\le Nn}\sup_{x\in X} P\big\{c_n g_{n1}(x,\alpha) > \varepsilon\big\} = 0, \tag{4.106}$$

where $g_{nk}(x,\alpha)$ is defined in (4.98) and $c_n = 1/n$.

Then relation (4.104) is true where $s(\cdot)$ and T are defined according to relations (4.57) and (4.58).

THEOREM 4.12. *Suppose that the conditions of Theorem 4.6 hold and equation (4.106) holds with $c_n = 1/\sqrt{n}$. Then the sequence of processes $\tilde{\gamma}_n(t) = n^{-1/2}(\zeta_n(nt) - ns(t))$ J-converges to diffusion process $\kappa(t)$ satisfying equation (4.74).*

The proof of these theorems follows from Theorems 4.8, 4.5 and 4.6.

4.6.1. *Averaging principle and diffusion approximation for processes with semi-Markov switching*

Consider a particular case when an SP is a process with semi-Markov switching (PSMS). Let for each $n = 1, 2, \ldots$, $\mathcal{F}_{nk} = \{\zeta_{nk}(t, x, \alpha),\ t \geq 0,\ x \in X,\ \alpha \in \mathcal{R}^r\}$, $k \geq 0$, be the jointly independent families of stochastic processes in $\mathcal{D}^r_\infty$, $x_n(t)$, $t \geq 0$, be an independent of $\mathcal{F}_{nk}$ SMP with values in some measurable space X, and S_{n0} be an initial value. Denote by $0 = t_{n0} < t_{n1} < \cdots$ the times of sequential jumps of $x_n(\cdot)$, and introduce the embedded MP $x_{nk} = x_n(t_{nk})$, $k \geq 0$. We construct a PSMS according to equation (1.14): 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

$$\zeta_n(t) = S_{nk} + \zeta_{nk}\big(t - t_{nk}, x_{nk}, S_{nk}\big), \quad \text{as } t_{nk} \leq t < t_{nk+1},\ t \geq 0. \tag{4.107}$$

Then process $(x_n(t), \zeta_n(t))$, $t \geq 0$, is a PSMS.

We consider for simplicity a homogenous case (distributions of $\zeta_{nk}(\cdot)$ do not depend on index $k \geq 0$). Let $\tau_n(x)$ be the sojourn time in state x for *SMP* $x_n(\cdot)$. Denote for each $x \in X$, $\alpha \in \mathcal{R}^r$,

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

To apply the results of Theorems 4.5 and 4.6 note that, as we consider the process with semi-Markov switching, variables $\tau_{nk}(x, n\alpha)$ in notation (4.49) of Theorem 4.5 do not depend on α as these variables are the sojourn times in the states of SMP $x_n(t)$. Therefore, we assume that MP x_{nk}, $k \geq 0$, at each $n > 0$ has a stationary measure $\pi_n(A)$, $A \in \mathcal{B}_X$, and denote

$$m_n(x) = \mathbf{E}\tau_n(x), \quad m_n = \int_X m_n(x)\pi_n(\mathrm{d}x). \tag{4.108}$$

COROLLARY 4.4. *Let the conditions of Theorem 4.5 be satisfied where $\tau_{n1}(x, n\alpha) = \tau_n(x)$, $m_n(x, \alpha) = m_n(x)$ and $m_n(\alpha) = m_n$. Also let for any $N > 0$, $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sup_{|\alpha| \leq N} \sup_x n\mathbf{P}\big\{n^{-1} g_n(x, \alpha) > \epsilon\big\} = 0. \tag{4.109}$$

Then relation (4.104) holds where $s(t)$ satisfies equation (4.57) with $m(s) = m$.

COROLLARY 4.5. *Let the conditions of Theorem 4.6 be satisfied where in notation (4.66),* $\tau_{nk}(x, n\alpha) = \tau_n(x), m_n(x, \alpha) = m_n(x)$ *and* $m_n(\alpha) = m_n$*. Also let for any* $N > 0$*,* $\varepsilon > 0$*,*

$$\lim_{n\to\infty} \sup_{|\alpha|\leq N} \sup_{x} n\mathbf{P}\{n^{-1/2} g_n(x,\alpha) > \epsilon\} = 0. \quad (4.110)$$

Then the statement of Theorem 4.6 is true, where in equation (4.74) $m(s) = m$*.*

A statement similar to Statement 4.1 is also valid.

STATEMENT 4.2. *If the trajectory of* $\zeta_{nk}(t, x, \alpha)$ *for any* α, x *is monotonic in the interval* $[0, \tau_{nk}(x, \alpha))$ *with probability one, then relation (4.106) in both theorems is automatically satisfied. Thus, the conditions of Theorem 4.5 imply relation (4.104) and correspondingly the conditions of Theorem 4.6 imply* J*-convergence of process* $\widetilde{\gamma}(t)$ *to* $\gamma(t)$*.*

Note that the averaging principle for the general SP (case of feedback) was proved in [ANI 92a].

4.7. Bibliography

[ANI 73] ANISIMOV V., "Asymptotic consolidation of the states of random processes", *Cybernetics*, vol. 9, no. 3, p. 494–504, 1973.

[ANI 75] ANISIMOV V., "Limit theorems for random processes and their applications to discrete summation schemes", *Theor. Probab. Appl.*, vol. 20, 1975.

[ANI 77] ANISIMOV V., "Switching processes", *Cybernetics*, vol. 13, no. 4, p. 590–595, 1977.

[ANI 78] ANISIMOV V., "Limit theorems for switching processes and their applications", *Cybernetics*, vol. 14, no. 6, p. 917–929, 1978.

[ANI 86] ANISIMOV V. and YURACHKOVSKIY A., "A limit theorem for stochastic difference schemes with random coefficients", *Theor. Prob. and Math. Stat.*, vol. 33, p. 1–9, 1986.

[ANI 87] ANISIMOV V., ZAKUSILO O. and DONTCHENKO V., *The Elements of Queueing Theory and Asymptotic Analysis of Systems*, Visca Scola (Russian), Kiev, Ukraine, 1987.

[ANI 88] ANISIMOV V., *Random Processes with Discrete Component. Limit Theorems*, Kiev University (Russian), Kiev, Ukraine, 1988.

[ANI 89] ANISIMOV V. and YURACHKOVSKIY A., "Averaging principle for stochastic difference equations", *Ukrainian Math. J.*, vol. 41, p. 1022–1028, 1989.

[ANI 90] ANISIMOV V. and ALIEV A., "Limit theorems for recurrent processes of semi-Markov type", *Theor. Prob. and Math. Stat.*, vol. 41, p. 7–13, 1990.

[ANI 91a] ANISIMOV V., "Averaging principle for switching recurrent sequences", *Theor. Probab. and Math. Stat.*, vol. 45, p. 1–8, 1991.

[ANI 91b] ANISIMOV V. and ATADZHANOV H., "Diffusion approximation of systems with repeated calls", *Theor. Prob. and Math. Stat.*, vol. 44, p. 3–8, 1991.

[ANI 92a] ANISIMOV V., "Averaging principle for switching processes", *Theor. Probab. and Math. Stat.*, vol. 46, p. 1–10, 1992.

[ANI 92b] ANISIMOV V. and LEBEDEV E., *Stochastic Queueing Networks. Markov Models*, Kiev University (Russian), Kiev, Ukraine, 1992.

[ANI 93] ANISIMOV V., "Averaging principle for the processes with fast switching", *Random Oper. and Stoch. Eqv.*, vol. 1, no. 2, p. 151–160, 1993.

[ANI 94a] ANISIMOV V., "Limit theorems for processes with semi-Markov switching and their applications", *Random Oper. and Stoch. Eqv.*, vol. 2, no. 4, p. 333–352, 1994.

[ANI 94b] ANISIMOV V. and ATADZHANOV H., "Diffusion approximation of systems with repeated calls and unreliable server", *J. of Math. Sci.*, vol. 72, no. 2, p. 3032–3034, 1994.

[ANI 95] ANISIMOV V., "Switching processes: averaging principle, diffusion approximation and applications", *Acta Applicandae Mathematicae*, vol. 40, p. 95–141, 1995.

[ANI 96] ANISIMOV V., "Averaging principle for near-critical branching processes with semi-Markov switching", *Theor. Probab. and Math. Stat.*, vol. 52, p. 13–26, 1996.

[ANI 97] ANISIMOV V., "Asymptotic analysis of switching queueing systems in conditions of low and heavy loading", in CHAKRAVARTHY S. and ALFA A., Eds., *Matrix-Analytic Methods in Stochastic Models*, vol. 183 of *Lecture Notes in Pure and Appl. Math.*, p. 241–260, Dekker, New York, 1997.

[ANI 99a] ANISIMOV V., "Averaging methods for transient regimes in overloading retrial queuing systems", *Mathematical and Computing Modelling*, vol. 30, no. 3/4, p. 65–78, 1999.

[ANI 99b] ANISIMOV V., "Diffusion approximation for processes with semi-Markov switches and applications in queueing models", in JANSSEN J. and LIMNIOS N., Eds., *Semi-Markov Models and Applications*, p. 77–101, Kluwer Acad. Publ., Dordrecht, 1999.

[ANI 99c] ANISIMOV V., "Switching stochastic models and applications in retrial queues", *Top*, vol. 7, no. 2, p. 169–186, 1999.

[ANI 00a] ANISIMOV V., "Asymptotic analysis of reliability for switching systems in light and heavy traffic conditions", in LIMNIOS N. and NIKULIN M., Eds., *Recent Advances in Reliability Theory: Methodology, Practice, and Inference*, p. 119–133, Birkhäuser Boston, Massachusetts, 2000.

[ANI 00b] ANISIMOV V., "J-convergence for switching processes with rare perturbations to diffusion processes with Poisson type jumps", in KOROLYUK V., PORTENKO N. and SYTA H., Eds., *Skorokhod's Ideas in Probability Theory*, p. 81–98, Inst. of Math. Nat. Acad. Sci. of Ukraine, Kiev, 2000.

[ANI 02a] ANISIMOV V., "Averaging in Markov models with fast Markov switches and applications to queueing models", *Annals of Operations Research*, vol. 112, no. 1, p. 63–82, 2002.

[ANI 02b] ANISIMOV V., "Diffusion approximation in overloaded switching queueing models", *Queueing Systems*, vol. 40, no. 2, p. 141–180, 2002.

[ANI 04] ANISIMOV V., "Averaging in Markov models with fast semi-Markov switches and applications", *Communications in Statistics - Theory and Methods*, vol. 33, no. 3, p. 517–531, 2004.

[BIL 68] BILLINGSLEY P., *Convergence of Probability Measures*, Wiley, New York, 1968.

[ETH 86] ETHIER S. and KURTZ T., *Markov Processes, Characterization and Convergence*, John Wiley & Sons, New York, 1986.

[GIK 75] GIKHMAN I. and SKOROKHOD A., *Theory of Random Processes, II*, Springer-Verlag, New York, 1975.

[GIK 78] GIKHMAN I. and SKOROKHOD A., *Theory of Random Processes, III*, Springer-Verlag, New York, 1978.

[GRI 69] GRIEGO R. and HERSH R., "Random evolutions, Markov chains, and systems of partial differential equations", *Proc. Nat. Acad. Sci. USA*, vol. 62, p. 305–308, 1969.

[GRI 73] GRIGELIONIS B., "The relative compactness of sets of probability measures in $D_{(0,\infty)}(X)$", *Math. Trans. Acad. Sci. Lithuanian SSR*, vol. 13, 1973.

[KER 78a] KERTZ R., "Limit theorems for semigroups with perturbed generators, with applications to multiscaled random evolutions", *J. Funct. Anal.*, vol. 27, no. 2, p. 215–233, 1978.

[KER 78b] KERTZ R., "Random evolutions with underlying semi-Markov processes", *Publ. Res. Inst. Math. Sci.*, vol. 14, p. 589–614, 1978.

[KHA 68] KHAS'MINSKII R., "About the averaging principle for ITO stochastic differential equations", *Kybernetika*, vol. 4, no. 3, p. 260–279, 1968.

[KOR 93] KOROLYUK V. and TURBIN A., *Mathematical Foundation of the State Lumping of Large Systems*, Kluwer, Dordrecht, 1993.

[KOR 94] KOROLYUK V. and SWISHCHUK A., *Random Evolutions*, Kluwer, Dordrecht, 1994.

[KOR 99] KOROLYUK V. and KOROLYUK V., *Stochastic Models of Systems*, Kluwer, Dordrecht, 1999.

[KOR 00] KOROLYUK V. and LIMNIOS N., "Evolutionary systems in an asymptotic split phase space", in LIMNIOS N. and NIKULIN M., Eds., *Recent Advances in Reliability Theory: Methodology, Practice and Inference*, p. 145–161, Birkhäuser Boston, Massachusetts, 2000.

[KOR 04] KOROLYUK V. and LIMNIOS N., "Average and diffusion approximation for evolutionary systems in an asymptotic split phase state", *Ann. Appl. Prob.*, vol. 14, no. 1, p. 489–516, 2004.

[KOR 05] KOROLYUK V. and LIMNIOS N., *Stochastic Systems in Merging Phase Space*, World Scientific, Singapore, 2005.

[KUR 73] KURTZ T., "A limit theorem for perturbed operator semigroups with applications to random evolutions", *J. Funct. Anal.*, vol. 12, p. 55–67, 1973.

[LIP 89] LIPTSER R. and SHIRYAEV A., *Theory of Martingales*, Kluwer, Dordrecht, 1989.

[PAP 72] PAPANICOLAOU G. and HERSH R., "Some limit theorems for stochastic equations and applications", *Indiana Univ. Math. J.*, vol. 21, p. 815–840, 1972.

[PIN 75] PINSKY M., "Random evolutions", in *Probabilistic Methods in Differential Equations*, vol. 451 of *Lecture Notes in Math.*, p. 89–99, Springer, Berlin, 1975.

[SKO 56] SKOROKHOD A., "Limit theorems for random processes", *Theory Prob. Appl.*, vol. 1, p. 289–319, 1956.

[SKO 89] SKOROKHOD A., *Asymptotic Methods in the Theory of Stochastic Differential Equations*, Amer. Math. Soc., Rhode Island, 1989.

[WAT 84] WATKINS J., "A central limit problem in random evolutions", *Ann. Prob.*, vol. 12, no. 2, p. 480–513, 1984.

Chapter 5

Averaging and Diffusion Approximation in Overloaded Switching Queueing Systems and Networks

5.1. Introduction

The complexity of real models of computing and information systems leads to the necessity of the development of more complicated queueing models and new approaches for modeling, analysis and asymptotic investigation.

A large number of papers are devoted to the analysis of queueing models in heavy traffic conditions. This usually means that the characteristics of the system depend on a parameter, say n, and as $n \to \infty$, the average load in the system tends towards one with the rate $O(1/\sqrt{n})$ (or even greater than one). The study of heavy traffic limits has a long history and there are several directions oriented to 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 (Reiman [REI 84]) and for various classes of multiclass networks (see survey of Williams [WIL 96] and papers by Reiman [REI 88], Dai and Kurtz [DAI 95], Harrison [HAR 95], Harrison and Williams [HAR 96], Bramson [BRA 98], Williams [WIL 98], Dai [DAI 99], Chen and Zhang [CHE 00]). Several classes of service disciplines for multiclass networks are studied in the latest paper by Bramson and Dai [BRA 01]. 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 Skorokhod reflection problem solution [SKO 62] and its generalizations).

Another direction is related to the analysis of Markov state-dependent queueing models. The method of analysis here is mainly based on a martingale technique [LIP 89] and again uses the continuous mapping theorems. Based on this technique, the convergence of a queueing process for a state-dependent $(M/M_Q/1/c_n)^r$ network in heavy traffic conditions to the diffusion process with the reflection in the rectangle is proved in [ANI 90b], 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 papers by Mandelbaum and Pats [MAN 98b], Mandelbaum, Massey and Reiman [MAN 98a], and the book by Basharin, Bocharov and Kogan [BAS 89]. Markov time-dependent models are considered in Mandelbaum, Massey [MAN 95] and [MAN 98a]. Some results for the state-dependent arrival process and the general service time distribution are given in Krichagina, Liptser and Puhalsky [KRI 88].

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 [ANI 92b] based on AP and DA for so-called RPSMs [ANI 90a, ANI 95]. 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 [ANI 90a, ANI 95, ANI 97] as examples of using this approach.

In this chapter the fluid and diffusion approximation type results are extended to more general classes of queueing models of a switching structure. This means, the corresponding queueing process can be represented in terms of switching processes. The SP has the property that the character of its operation varies spontaneously (switches) at some points of time which can be random functionals of the previous trajectory or possibly jumps of a random environment. The environment may reflect some outer perturbations, a type of operating regime, a number of working servers, a domain of operation for a queueing process, a type of priority, etc. In the intervals between switches the process may have a non-Markov structure. A general description of an SP is given in section 1.2.1 (see also [ANI 77, ANI 78, ANI 95]).

The class of switching queueing models in particular includes open and closed Jackson 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 an 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 different models with multiple calls, calls of a random size and different priorities, models with negative and impatient calls, semi-Markov models with unreliable servers, non-homogenous in time Markov and semi-Markov models, networks $(G_Q/M_Q/s/m)^r$ with the state-dependent non-exponential arrival process [ANI 95], some classes of state-dependent retrial models [ANI 99a, ANI 99c, ANI 01] and polling systems. In terms of SPs 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.

The queueing processes in switching 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. Therefore, we restrict our analysis to studying overloaded models without reflection on the boundary. This means that we study the convergence on interval $[0, T]$ such that in each component $s(t) > 0$, $t \in [0, T]$, where $s(t)$ is a fluid limit (a limit for a normalized queueing process).

A new approach for the investigation of the asymptotic behavior of switching queueing models is developed. It is based on AP and DA type results for SPs (see Chapter 4 and [ANI 77, ANI 92a, ANI 94a, ANI 95]), and uses the representation of queueing processes in terms of SPs. This approach allows us to extend fluid and diffusion approximation type results (without reflection) to new more general classes of queueing models, in particular, to state-dependent Markov queueing systems and networks $(M_{Q,B}/M_{Q,B}/k/\infty)^r$ with batch arrival process and service, state-dependent Markov models $(M_{M,Q}/M_{M,Q}/k/\infty)^r$ 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 for studying known classes of Markov state-dependent and time-dependent models such as $(M_Q/M_Q/1/\infty)^r$.

In this chapter we concentrate our attention on the study of state-dependent Markov and semi-Markov queueing models and their modifications in the presence of the ergodic Markov or semi-Markov environment as well. We assume that characteristics of the system depend on a parameter n, $n \to \infty$, 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 a random environment $x_n(t)$. In specific applications the environment may appear due to some external or internal factors. In general, the environment may depend on the queueing process itself and in this case it will not be a Markov or a semi-Markov process (case of feedback). We also assume 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 this type we prove that under rather general assumptions a multidimensional queueing process $n^{-1}Q_n(nt)$ in an interval $[0, T]$ uniformly converges in probability to a function $s(t)$ which is a positive solution of an ordinary differential equation (fluid limit), we also call it the averaging principle, and the normalized queueing process $n^{-1/2}(Q_n(nt) - ns(t))$ J-converges (in the sense of a weak convergence of probability measures induced by the process in 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 refered to Skorokhod [SKO 56], Billingsley [BIL 68] and Ethier and Kurtz [ETH 86] for the definition of Skorokhod space and Skorokhod topology.

These results are mainly oriented to the analysis of a transient behavior of the queueing processes. They also allow us 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 obtain the asymptotic behavior of the hitting time to zero, as 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 cannot asymptotically reach zero (for example, for the $M/M/\infty$ model (network) when the service rate goes to 0). For models of this type we obtain the approximation on the entire time horizon. It is possible to study so-called quasi-stationary regimes as well. These regimes appear when the corresponding fluid limit $s(t)$ has a point of stability $s_* > 0$. In this case, as $n \to \infty$ and then $t \to \infty$, the value $n^{-1}Q_n(nt)$ is asymptotically close to s_*. In particular, if $n^{-1}Q_n(0) \xrightarrow{\text{P}} s_*$, then the coefficients of the limiting diffusion process do not depend on time, and the queueing process balances near some asymptotically high level ns_* as a homogenous diffusion process multiplied by $\sqrt{n}$.

The results of this chapter are partially published in [ANI 02]. Section 5.2 is devoted to the asymptotic analysis (fluid limits and diffusion approximation) of some classes of overloaded state-dependent Markov queueing systems and networks in transient conditions. Non-Markov models and some special models such as polling systems are considered in section 5.3. Section 5.4 deals with a special class of queueing models – retrial queueing systems.

5.2. Markov queueing models

In overloaded switching queueing models various multidimensional characteristics (numbers of calls at different nodes, volume of information in buffers, output flows, flows of lost calls, waiting times, etc.) can be approximated by the solutions to differential equations or by the diffusion processes. The method of analysis is based on the AP and DA type results for SPs (see Chapter 4) and uses the representation of corresponding queueing processes as SPs. We restrict our analysis to studying queueing processes without reflection and consider the convergence on interval $[0, T]$ such that in each component $s(t) > 0, t \in [0, T]$. The analysis of reflecting processes should be moved into a separate problem.

In the next section in order to illustrate a general approach we consider different classes of overloaded state-dependent Markov queueing systems and networks.

5.2.1. *System* $\overline{M}_{\overline{Q},B}/\overline{M}_{\overline{Q},B}/1/\infty$

Now consider the rather general Markov system $\overline{M}_{\overline{Q},B}/\overline{M}_{\overline{Q},B}/1/\infty$ considered in section 2.2.1.3. This system 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 a scaling parameter $n \to \infty$. Let non-negative functions $\lambda(\overline{q}), \mu(\overline{q}), \nu_i(\overline{q})$, $i = \overline{1,m}$, $\overline{q} \in \mathcal{R}_+^m$, be given. Also let $\overline{\alpha}(\overline{q}), \overline{\gamma}(\overline{q}), \overline{\beta}_i(\overline{q})$, $i = \overline{1,m}$, $\overline{q} \in \mathcal{R}_+^m$, 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 arrival 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 complete 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 the ith component equal to one, and other components equal to 0 leave the system after service completion. If vector $\overline{\beta}_i(\overline{q})$ can have negative components (for instance, there are impatient calls), then after transformation we obtain $\min\{\overline{0}, n\overline{q} + \overline{\beta}_i(\overline{q})\}$ 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:

$$\overline{m}^{(1)}(\overline{q}) = \mathbf{E}\overline{\alpha}(\overline{q}), \quad \overline{m}^{(2)}(\overline{q}) = \mathbf{E}\overline{\gamma}(\overline{q}), \quad \overline{m}_i^{(3)}(\overline{q}) = \mathbf{E}\overline{\beta}_i(\overline{q}),$$

$$d^{(1)}(\overline{q}) = \mathbf{E}\overline{\alpha}(\overline{q})\overline{\alpha}(\overline{q})^*, \quad d^{(2)}(\overline{q}) = \mathbf{E}\overline{\gamma}(\overline{q})\overline{\gamma}(\overline{q})^*, \quad d_i^{(3)}(\overline{q}) = \mathbf{E}\overline{\beta}_i(\overline{q})\overline{\beta}_i(\overline{q})^*,$$

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

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

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

In addition, let $G(\overline{q})$ be the matrix of partial derivatives for $\overline{b}(\overline{q})$:

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

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 differential equation

$$\mathrm{d}\overline{s}(t) = \overline{b}(\overline{s}(t))\mathrm{d}t, \quad \overline{s}(0) = \overline{s}_0. \tag{5.1}$$

Let us prove AP and DA for the queueing process.

THEOREM 5.1. *1) Suppose that in any bounded and closed domain in* $\mathrm{int}\{\mathcal{R}_+^m\}$ *variables* $\overline{\alpha}(\overline{q}), \overline{\gamma}(\overline{q}), \overline{\beta}(\overline{q})$ *are uniformly* $\overline{q}$ *integrable, functions* $\lambda(\overline{q}), \mu(\overline{q}), \nu_i(\overline{q}), \overline{m}_i^{(j)}(\overline{q})$ *are locally Lipschitz, and* $\Lambda(\overline{q}) > 0$. *In addition, let*

$$n^{-1}\overline{Q}_n(0) \xrightarrow{\mathrm{P}} \overline{s}_0, \tag{5.2}$$

where $\overline{s}_0 > \overline{0}$ *is a 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*

$$\overline{\eta}(0) = \overline{s}_0,\ \mathrm{d}\overline{\eta}(t) = \overline{b}(\overline{\eta}(t))\Lambda(\overline{\eta}(t))^{-1}\mathrm{d}t, \tag{5.3}$$

a unique solution of which exists in any interval.

Then a unique solution of equation (5.1) exists in interval $[0, T]$ *and*

$$\sup_{0 \le t \le T} \left| n^{-1}\overline{Q}_n(nt) - \overline{s}(t) \right| \xrightarrow{\mathrm{P}} 0. \tag{5.4}$$

2) Suppose in addition that variables $|\alpha(\overline{q})|^2$, $|\gamma(\overline{q})|^2$ *and* $|\beta(\overline{q})|^2$ *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* $\mathrm{int}\{\mathcal{R}_+^m\}$, *and* $n^{-1/2}(\overline{Q}_n(0) - n\overline{s}_0) \xrightarrow{\mathrm{W}} \overline{\zeta}_0$.

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

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

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

Note that $\mathrm{int}\{\mathcal{R}_+^m\} = \mathcal{R}_+^m \backslash \partial\mathcal{R}_+^m$ (the interior of $\mathcal{R}_+^m$), matrix $B(\overline{q})$ satisfies relation $B(\overline{q})B(\overline{q})^* = B(\overline{q})^2$, $\overline{w}(t)$ is a standard Wiener process in $\mathcal{R}^m$, and J-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 [SKO 56].

Proof. Let us introduce the jointly independent families of random variables $\{(\tau_{nk}(n\overline{q}), \overline{\xi}_{nk}(n\overline{q}))\}$, $k \geq 0$. 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:

$$\overline{\xi}_{n1}(n\overline{q}) = \begin{cases} \overline{\alpha}(\overline{q}), & \text{with probab. } \lambda(\overline{q})\Lambda(\overline{q})^{-1}, \\ -\overline{\gamma}(\overline{q}), & \text{with probab. } \mu_i(\overline{q})\Lambda(\overline{q})^{-1}, \qquad i = \overline{1, m}. \\ \overline{\beta}_i(\overline{q}), & \text{with probab. } q_i\nu_i(\overline{q})\Lambda(\overline{q})^{-1}, \end{cases} \tag{5.5}$$

Now, in order 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 function $s(t)$. Put $\delta = \min_{1\leq i\leq m} \min_{0\leq t\leq T} s_i(t)$. By construction, $\delta > 0$. Take $\varepsilon = \delta/2$ and consider the orthant $R_+^m(\varepsilon) = \{\overline{a} : \overline{a} \in \mathcal{R}_+^m, a_i \geq \varepsilon,\ i = 1, \ldots, m\}$. 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 a 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 function that is continuous (locally Lipschitz) in $R_+^m(\varepsilon)$, then it is easy to check that $\widetilde{f}(\overline{a})$ is also continuous (locally Lipschitz) in R^m.

Using this transformation, we define 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 equation (5.5) and above. Using these variables, we can define according to relations (4.13), (4.14) an RPSM $\widetilde{Q}_n(t)$. It can take values in $\mathcal{R}^m$, and, by construction, if in an interval $[0,T]$ $\widetilde{Q}_n(t) \geq n\varepsilon$, then its trajectory coincides with the trajectory of queueing process $\overline{Q}_n(t)$ in $[0,T]$.

Let us study the behavior of $\widetilde{Q}_n(t)$. As we can see, all conditions of Theorem 4.3 are satisfied. We can calculate the expectation of $\overline{\xi}_{n1}(n\overline{q})$ and see that $\widetilde{Q}_n(nt)$ satisfies relation (5.4) with the same function $\overline{s}(t)$. Now consider an interval $[0,T]$, where $\overline{s}(t) > 0$, $t \in [0,T]$. Then, for $\varepsilon > 0$ chosen above, we have $\overline{s}(t) \geq 2\varepsilon$, $t \in [0,T]$, and equation (5.4) implies that

$$\mathbf{P}\big(n^{-1}\widetilde{Q}_n(nt) \geq \varepsilon,\ t \in [0,T]\big) \longrightarrow 1. \tag{5.6}$$

Let us now construct on the same probability space the queueing process $\overline{Q}_n(nt)$ and RPSM $\widetilde{Q}_n(nt)$ in a recurrent way as follows. Put $\widetilde{Q}_n(0) = \overline{Q}_n(0)$. Then we generate a sequence of random variables $\omega_1, \omega_2, \ldots,$ that are uniformly distributed in $[0,1]$ and construct recursively using this sequence variables $\widetilde{Q}_{nk}$, $\widetilde{\tau}_{nk}(\widetilde{Q}_{nk})$, $\widetilde{\xi}_{nk}(\widetilde{Q}_{nk})$, $k \geq 0$, according to relations (4.13), (4.14) and using a standard simulation technique. For example, we construct an exponential random variable using the relation $\tau_{nk}(\overline{Q}) = -\Lambda(n^{-1}\overline{Q})^{-1} \ln \omega_{3k}$, and $\overline{\xi}_{nk}(\overline{Q})$ is constructed by variables $\omega_{2k+1}, \omega_{2k+2}$ in two stages according to equation (5.5). Then we construct trajectories of $\overline{Q}_n(nt)$ and $\widetilde{Q}_n(nt)$, where a trajectory of $\widetilde{Q}_n(nt)$ is constructed according to relations (4.13), (4.14) for variables with a tilde. By construction, if in an interval $[0,T]$, $\widetilde{Q}_n(nt) \geq n\varepsilon$, then $\widetilde{Q}_n(nt) = \overline{Q}_n(nt)$, $t \in [0,T]$. Now for any measurable set A of functions from σ-algebra $\mathcal{B}_{\mathcal{D}_T^r}$ relation (5.6) as $n \to \infty$ implies

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

This relation proves that the asymptotic behavior of trajectories of the queue and auxiliary RPSM $\widetilde{Q}_n(nt)$ is the same, and finally implies relation (5.4).

To prove the second part of Theorem 5.1, we first prove DA for the process $\widetilde{Q}_n(nt)$. The proof is based on the results of Theorem 4.4. This result is then extended using the same considerations as above to the process $\overline{Q}_n(nt)$. □

NOTE 5.1. The result of Theorem 5.1 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 also be extended to the case of r servers.

Let us now consider as examples some special classes of Markov state-dependent models.

5.2.2. *System* $M_Q/M_Q/1/\infty$

Consider the system described in section 2.2.1.1. There is one server with an infinite number of waiting places. We study AP and DA for the queueing process in transient conditions and assume that the initial number of calls is of the order n. In this case we assume that the input and service rates depend on the normalized number of the calls in the system in the following way. If at time t there are Q calls in the system, then the input rate is $\lambda(Q/n)$ and the service rate is $\mu(Q/n)$ where $\lambda(q)$ and $\mu(q)$ are given functions. Denote by $Q_n(t)$ the number of calls in the system at time t. Suppose that as $n \to \infty$,

$$Q_n(0)/n \xrightarrow{\text{P}} s_0. \tag{5.7}$$

Denote by $s(t)$ a solution of the equation:

$$\mathrm{d}s(t) = b\big(s(t)\big)\mathrm{d}t, \quad s(0) = s_0, \tag{5.8}$$

where $b(q) = \lambda(q) - \mu(q)$. The following result follows from Theorems 4.3, 4.4 on AP and DA for simple RPSM and in fact this is a consequence of Theorem 5.1.

THEOREM 5.2. *1) Suppose that equation (5.7) is true,* $s_0 > 0$*, functions* $\lambda(q), \mu(q)$ *satisfy the local Lipschitz condition,* $\lambda(q) + \mu(q) > 0$ *as* $q \in (0, \infty)$*, and for some fixed* $T > 0$ *there exists an interval* $[0, A]$ *such that the equation*

$$\mathrm{d}\eta(t) = b\big(\eta(t)\big)\big(\lambda\big(\eta(u)\big) + \mu\big(\eta(u)\big)\big)^{-1}\mathrm{d}t, \quad \eta(0) = s_0, \tag{5.9}$$

has a unique solution $\eta(t) > 0$*,* $t \in (0, A)$*, and in addition* $y(A) > T$*, where*

$$y(t) = \int_0^t \big(\lambda\big(\eta(u)\big) + \mu\big(\eta(u)\big)\big)^{-1}\mathrm{d}u. \tag{5.10}$$

Then

$$\sup_{0 \le t \le T} \big|n^{-1}Q_n(nt) - s(t)\big| \xrightarrow{\mathrm{P}} 0, \tag{5.11}$$

where $s(t)$ *is a unique solution to equation (5.8).*

2) Suppose in addition that functions $\lambda(q), \mu(q)$ *are continuously differentiable in* $(0, \infty)$ *and*

$$n^{-1/2}\big(Q_n(0) - ns_0\big) \overset{\mathrm{w}}{\Longrightarrow} \zeta_0, \tag{5.12}$$

where ζ_0 *is a proper random variable.*

Then the sequence of processes $\zeta_n(t) = n^{-1/2}(Q_n(nt) - ns(t))$ *J-converges in* $\mathcal{D}_T$ *to the diffusion process* $\zeta(t)$ *satisfying the following stochastic differential equation:* $\zeta(0) = \zeta_0$*,*

$$\mathrm{d}\zeta(t) = \big(\lambda'\big(s(t)\big) - \mu'\big(s(t)\big)\big)\zeta(t)\mathrm{d}t + \big(\lambda\big(s(t)\big) + \mu\big(s(t)\big)\big)^{1/2}\mathrm{d}w(t), \tag{5.13}$$

Proof. We use the approach described in Theorem 5.1 and first construct an auxiliary process $\widetilde{Q}_n(t)$ which coincides with the queueing process in the interval $[0, T]$ where $Q_n(\cdot) > 0$. According to section 2.2.1.1, process $\widetilde{Q}_n(t)$ is represented as an RPSM according to relations (1.8), (1.9) (see also (4.13), (4.14)). In this case the variable $\tau_{n1}(nq)$ has an exponential distribution with parameter $a(q) = \lambda(a) + \mu(a)$, the variable $\xi_{n1}(nq)$ does not depend on $\tau_{n1}(nq)$ and

$$\xi_{n1}(nq) = \begin{cases} +1 & \text{with probability } \lambda(q)/a(q), \\ -1 & \text{with probability } \mu(q)/a(q), \end{cases}$$

where the distributions of variables $(\xi_{nk}(\cdot), \tau_{nk}(\cdot))$, $k > 0$, do not depend on index k. Calculating the characteristics of these variables and using Theorems 4.3, 4.4 we prove relation (5.11) for $\widetilde{Q}_n(t)$. Furthermore, as in the interval $[0, T]$, $s(t) > 0$, then following the lines of proof of Theorem 5.1 we see that the trajectories of processes $Q_n(nt)/n$ and $\widetilde{Q}_n(nt)/n$ in this interval asymptotically coincide. This implies the result of Theorem 5.2. □

NOTE 5.2. Suppose that $s_0 = 0$, other conditions of Theorem 5.2 hold, and in addition $\lambda(q)$ is continuous at point 0, there exists a limit $\mu(+0) = \lim \mu(q)$ as $q \searrow 0$, and $\lambda(0) > \mu(+0)$. Thus, equation (5.11) also holds.

Proof. Assume for simplicity that $Q_n(0) = 0$. Let there exist $T > 0$ such that $s(t) > 0$, as $0 < t \leq T$. 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 $\lambda_* = \inf\{\lambda(q) : 0 \leq q \leq \varepsilon\}$, $\mu^* = \sup\{\mu(q) : 0 < q \leq \varepsilon\}$.

Let $\Pi_1(t)$ and $\Pi_2(t)$ be two independent Poisson processes with parameters λ_* and μ^* respectively. Note that in domain $Q_n(nt) \leq n\varepsilon$ queue $Q_n(nt)$ stochastically dominates process $\Pi_1(nt) - \Pi_2(nt)$. This implies for any $c > 0$ that $\mathbf{P}(\tau_n(\varepsilon) > c) \leq \mathbf{P}(\widetilde{\tau}_n(\varepsilon) > c)$, where

$$\tau_n(\varepsilon) = \inf\big\{u : Q_n(nu) \geq n\varepsilon\big\},$$

$$\widetilde{\tau}_n(\varepsilon) = \inf\big\{u : \Pi_1(nt) - \Pi_2(nt) \geq n\varepsilon\big\}.$$

It is easy to see that as $n \to \infty$, $\widetilde{\tau}_n(\varepsilon) \xrightarrow{\mathrm{P}} \varepsilon/\delta$. Then for any $c > 0$, $\lim_{\varepsilon \to 0} \limsup_{n \to \infty} \mathbf{P}(\tau_n(\varepsilon) > c) = 0$, and also for any $\varepsilon > 0$,

$$\lim_{c \to \infty} \limsup_{n \to \infty} \mathbf{P}\big(\tau_n(\varepsilon) > c\big) = 0. \tag{5.14}$$

Now let us consider the behavior of $Q_n(nt)$ in interval $[\tau_n(\varepsilon), T]$. As sequence $\tau_n(\varepsilon)$ is stochastically bounded (see equation (5.14)), then for any sequence $n_k \to \infty$ we can choose a subsequence n_{k_l} such that $\tau_{n_{k_l}}(\varepsilon) \xrightarrow{\mathrm{W}} \tau_0(\varepsilon)$. Using Skorokhod construction of a common probability space, we can always assume without loss of generality that $\tau_{n_{k_l}}(\varepsilon) \xrightarrow{\mathrm{P}} \tau_0(\varepsilon)$. Now by definition, $n^{-1}Q_n(n\tau_n(\varepsilon)) \xrightarrow{\mathrm{P}} \varepsilon > 0$. Thus, applying Theorem 5.2, we obtain

$$\sup_{\tau_{n_{k_l}}(e) \leq t \leq T} \big|n_{k_l}^{-1} Q_{n_{k_l}}\big(n_{k_l} t\big) - s_\varepsilon(t)\big| \xrightarrow{\mathrm{P}} 0, \tag{5.15}$$

where $s_\varepsilon(t)$ is a solution to equation (5.8) in interval $[\tau_0(\varepsilon), T]$ with the initial value ε. As $\tau_0(\varepsilon) \to 0$ as $\varepsilon \to 0$, using the continuity of the solution of a differential equation in the initial value we obtain that $s_\varepsilon(t) \to s(t)$ as $\varepsilon \to 0$ uniformly on any fixed interval $[\delta, T]$, $\delta > 0$. Now using equations (5.14), (5.15) and the relation

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

we finally prove that for any $\varepsilon > 0$ as $n = n_{k_l} \to \infty$,

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

where the last term tends to 0 as $\varepsilon \to 0$. As for any sequence n_k we can choose a subsequence n_{k_l} for which equation (5.11) is true, thus equation (5.11) is true as $n \to \infty$. □

As we can see from Note 5.2, the result of Theorem 5.2 can be extended to the case when some components of $\overline{s}_0$ may take a value of 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 a point s on the border, then the first 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 5.2.

CASE 1 (System $M/M/1/\infty$). Let $\lambda(q) \equiv \lambda$, $q \ge 0$, $\mu(q) \equiv \mu$, $q > 0$ ($\mu(0) = 0$). Our system is then 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 equation (5.10)). If $\lambda \ge \mu$, then $s(t) > 0$ for any $t > 0$, and equation (5.11) is true for any $T > 0$. If $\lambda < \mu$, then $s(t) > 0$ for $0 < t < s_0(\mu - \lambda)^{-1}$, and equation (5.10) is true for any $T < s_0(\mu - \lambda)^{-1}$.

Consider the behavior of the first time the queue becomes zero:

$$\psi_n(Q) = \inf\left\{t : t \ge 0,\ Q_n(t) = 0 \text{ given that } Q_n(0) = Q\right\}.$$

This time is a continuous functional concerning the uniform convergence in probability to a monotone function. Therefore, if $\lambda < \mu$ and $n^{-1}Q_n(0) \xrightarrow{\text{P}} s_0 > 0$, then $n^{-1}\psi_n(Q_n(0)) \xrightarrow{\text{P}} s_0(\mu - \lambda)^{-1}$.

Solving equation (5.13) we can find that $\zeta(t) = \zeta_0 + (\lambda + \mu)^{1/2} w(t)$ as $0 \le t \le T$.

CASE 2 (System $M/M/\infty$). Let $\lambda(q) \equiv \lambda$, $\mu(q) \equiv \mu q$, $q \geq 0$. Then our system is equivalent to a system $M/M/\infty$. In this case equation (5.8) has the form:

$$\mathrm{d}s(t) = \big(\lambda - \mu s(t)\big)\mathrm{d}t, \quad s(0) = s_0 > 0, \tag{5.16}$$

and $s(t) = \lambda/\mu + (s_0 - \lambda/\mu)e^{-\mu t}$, $t \geq 0$.

Let us show that equation (5.11) holds for any $T > 0$. In this case equation (5.9) has the form

$$\mathrm{d}\eta(t) = \big(\lambda - \mu\eta(t)\big)\big(\lambda + \mu\eta(t)\big)^{-1}\mathrm{d}t. \tag{5.17}$$

We can see that the function $\eta(t)$ strictly monotonically increases in the domain $\eta(t) < \lambda/\mu$, and 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 (5.17) 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 obtain 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, $\eta(t+a) - \eta(t) = \int_t^{t+a}\eta'(u)\mathrm{d}u \to a\eta'_\infty \neq 0$. Thus, it should be $\eta'_\infty = 0$ and $\eta_\infty = \lambda/\mu$. This implies according to equation (5.10) that $y(t) \to \infty$ as $t \to \infty$ and therefore equation (5.11) holds for any $T > 0$.

Equation (5.13) has the form:

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

This is a linear stochastic differential equation which is an Ornstein-Uhlenbeck type process, and a solution can be written in the closed form. To find the representation for $\zeta(t)$ we use the formula [GIK 72]: if process $\xi(t)$ satisfies the equation

$$\mathrm{d}\xi(t) = \big(\alpha(t) + \beta(t)\xi(t)\big)\mathrm{d}t + \gamma(t)\mathrm{d}w(t),\ t > 0, \quad \xi(0) = \xi_0,$$

then

$$\xi(t) = \exp\bigg\{\int_0^s \beta(s)\mathrm{d}s\bigg\}\bigg(\xi_0 + \int_0^t \exp\bigg\{-\int_0^s \beta(u)\mathrm{d}u\bigg\}\alpha(s)\mathrm{d}s \\ + \int_0^t \exp\bigg\{-\int_0^s \beta(u)\mathrm{d}u\bigg\}\gamma(s)\mathrm{d}w(s)\bigg).$$

In our case $\alpha(t) \equiv 0$, $\beta(t) \equiv -\mu, \gamma(t) = (\lambda + \mu s(t))^{1/2}$, and therefore

$$\zeta(t) = e^{-\mu t}\bigg(\zeta_0 + \int_0^t e^{\mu u}\big(\lambda + \mu s(u)\big)^{1/2}\mathrm{d}w(u)\bigg). \tag{5.18}$$

In the asymptotically stationary case $s_0 = \lambda/\mu$, $s(t) \equiv \lambda/\mu$ and $\mathrm{d}\zeta(t) = -\mu\zeta(t)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}w(t)$. In this case $\zeta(t)$ is an Ornstein-Uhlenbeck process. Note that the convergence of the process $n^{-1/2}(Q_n(nt) - n\lambda/\mu)$ to an Ornstein-Uhlenbeck process for the system $M/M/\infty$ was obtained in [IGL 65].

Representation (5.18) can be written in another form. Note that if a non-deterministic function $f(u) \geq 0$, then process $\int_0^t f(u)\mathrm{d}w(u)$ is equivalent to process $w(\int_0^t f^2(u)\mathrm{d}u)$ in the sense that the finite-dimensional distributions of both processes coincide. In fact, both processes are Gaussian with independent increments and have the same distributions as their increments in interval $[t, s]$ have a Gaussian distribution $N(0, \int_t^s f(u)^2 du)$. Therefore, process $\zeta(t)$ is equivalent to process $\zeta(t) = e^{-\mu t}(\zeta_0 + w(\phi(t)))$, where $\phi(t) = \mu^{-1}\lambda(e^{2t\mu} - 1) - \mu^{-1}(\lambda - \mu s_0)(e^{t\mu} - 1)$.

Let us consider the stationary case where $s_0 = \lambda/\mu$ separately. Then $s(t) \equiv \lambda/\mu$ and for any $t > 0$, the variable $w(\phi(t))$ has the same distribution as the variable $\phi(t)^{1/2}N(0,1)$. As $e^{-\mu t}\zeta_0 \xrightarrow{\mathrm{P}} 0$ when $t \to \infty$, then at large t,

$$\zeta(t) \sim e^{-\mu t}\left(e^{2t\mu} - 1\right)^{1/2}\sqrt{\lambda/\mu}N(0,1) \stackrel{\mathrm{w}}{\Longrightarrow} \sqrt{\lambda/\mu}N(0.1).$$

For arbitrary s_0 it is easy to calculate that $s(t) \to \lambda/\mu$, and $\zeta(t) \stackrel{\mathrm{w}}{\Longrightarrow} \sqrt{\lambda/\mu}N(0,1)$ as $t \to \infty$. This means that as $n \to \infty$,

$$Q_n(nt) \sim n\lambda/\mu + \sqrt{n}\zeta(t).$$

As at large t, $\zeta(t) \sim \sqrt{\lambda/\mu}N(0,1)$, then at large n and t,

$$Q_n(nt) \sim n\lambda/\mu + \sqrt{n}\sqrt{\lambda/\mu}N(0,1),$$

and we can say that the system is in a quasi-stationary regime.

5.2.3. *Analysis of the waiting time*

If a call enters the system at time t, denote by $W(t)$ the time spent by this call in the queue until the beginning of service. Let us consider the asymptotic behavior of waiting time $W(t)$ for system $M_Q/M_Q/1/\infty$ considered in section 5.2.2.

Denote by $\Pi_{\mu(\cdot)}(t)$ a non-homogenous Poisson process with the instantaneous rate $\mu(t)$ at time t. If $Q_n(\cdot) > 0$ in interval $[t, s]$, then the output process in this interval is generated by process $\Pi_{\mu(Q_n(\cdot)/n)}(v)$. Thus,

$$W(t) = \inf\left\{s : s > 0 : \Pi_{\mu(Q_n(\cdot)/n)}(t+s) - \Pi_{\mu(Q_n(\cdot)/n)}(t) = Q_n(t)\right\} \quad (5.19)$$

First let us prove the following auxiliary statement:

STATEMENT 5.1. *Assume that the conditions of Theorem 5.2 hold. Then the sequence of processes* $n^{-1}\Pi_{\mu(Q_n(\cdot)/n)}(nt)$ *J-converges in any interval* $[0,T]$ *to a deterministic process* $\int_0^t \mu(s(v))dv$.

In fact, process $n^{-1}\Pi_{\mu(n^{-1}Q(\cdot))}(nt)$ is monotonically non-decreasing in argument t with probability one and

$$\begin{aligned} &\mathbf{E}\exp\left\{i\theta n^{-1}\Pi_{\mu(Q_n(\cdot)/n)}(nt)\right\} \\ &\quad = \mathbf{E}\exp\left\{\left(e^{i\theta/n}-1\right)\int_0^{nt}\mu\big(Q_n(v)/n\big)\mathrm{d}v\right\}. \end{aligned} \tag{5.20}$$

As $n\to\infty$, relation (5.11) implies

$$\frac{1}{n}\int_0^{nt}\mu\big(Q_n(v)/n\big)\mathrm{d}v=\int_0^t\mu\big(Q_n(nu)/n\big)\mathrm{d}u \xrightarrow{\mathrm{P}} \int_0^t\mu\big(s(u)\big)\mathrm{d}u. \tag{5.21}$$

Note that

$$n\big(e^{i\theta/n}-1\big)=-n\big(1-\cos(\theta/n)\big)+in\sin(\theta/n)\longrightarrow i\theta. \tag{5.22}$$

Let us consider a function $f_n(z) = \exp\{n(e^{i\theta/n}-1)z\}$, $z \geq 0$. The relation $\mathrm{Re}(e^{i\theta/n}-1) \leq 0$ implies that $|f_n(z)| \leq 1$. Let us use the inequality $|e^y - e^x| \leq |y-x|$, valid for any complex values x and y such that $\mathrm{Re}\, x \leq 0$, $\mathrm{Re}\, y \leq 0$. Then, using the known inequalities $|\sin\alpha| \leq |\alpha|$, $|1-\cos\alpha| \leq \alpha^2/2$, $\alpha\in\mathcal{R}$, we obtain in the region $z_1\geq 0$, $z_2\geq 0$,

$$\begin{aligned} \big|f_n(z_1)-f_n(z_2)\big| &\leq \big|n\big(e^{i\theta/n}-1\big)\big|\big|z_1-z_2\big| \\ &\leq \left(\frac{1}{2}\theta^2/n+\theta\right)\big|z_1-z_2\big|. \end{aligned} \tag{5.23}$$

Inequality (5.23) implies the uniform continuity of function $f_n(z)$ with respect to n. Therefore, relations (5.21), (5.22) imply the convergence of the right-hand side in relation (5.20) to the expression

$$\exp\left\{i\theta\int_0^t\mu\big(s(u)\big)\mathrm{d}u\right\}. \tag{5.24}$$

Statement 5.1 is proved.

Using the monotonicity of the process $n^{-1}\Pi_{\mu(n^{-1}Q(\cdot))}(nt)$, relation (5.19) and Statement 5.1 we can prove the following.

STATEMENT 5.2. *If the conditions of the first part of Theorem 5.2 hold, then the sequence of processes* $n^{-1}W(nt)$ *U-converges in any interval* $[0,T]$ *such that*

$$\int_T^{\infty} \mu\big(s(v)\big)\mathrm{d}v > \sup_{u\le T} s(u), \tag{5.25}$$

to a deterministic process

$$\widetilde{W}(t) = \inf\left\{s : s > 0,\ \int_t^{t+s} \mu\big(s(v)\big)\mathrm{d}v = s(t)\right\}.$$

Note that if relation (5.25) holds, then $\widetilde{W}(t) < \infty$ for any $t \le T$.

Representation (5.19) also makes it possible to prove the diffusion approximation of the process $(W(nt) - n\widetilde{W}(t))/\sqrt{n}$.

5.2.4. *An output process*

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

COROLLARY 5.1. *If the conditions of Theorem 5.2 hold, then equation (5.11) is true and*

$$\sup_{0\le t\le T} \left|n^{-1}Z_n(nt) - g(t)\right| \xrightarrow{\mathrm{P}} 0, \tag{5.26}$$

where $g(t) = \int_0^t \mu(s(u))\mathrm{d}u$, *and*

$$\mathrm{d}s(t) = \big(\lambda\big(s(t)\big) - \mu\big(s(t)\big)\big)\mathrm{d}t, \quad s(0) = s_0 > 0. \tag{5.27}$$

Correspondingly, the sequence of processes

$$\big(n^{-1/2}\big(Q_n(nt) - ns(t)\big), n^{-1/2}\big(Z_n(nt) - ng(t)\big)\big)$$

weakly converges in $\mathcal{D}_T$ *to a two-dimensional diffusion process* $(\zeta(t), \kappa(t))$ *satisfying the system of stochastic differential equations:*

$$\begin{aligned}
\mathrm{d}\zeta(t) = {} & \big(\lambda'\big(s(t)\big) - \mu'\big(s(t)\big)\big)\zeta(t)\mathrm{d}t \\
& + \frac{1}{\sqrt{2}}\bigg(\Big[\sqrt{\lambda\big(s(t)\big)} + \sqrt{\mu\big(s(t)\big)}\Big]\mathrm{d}w_1(t) \\
& \qquad + \Big[\sqrt{\lambda\big(s(t)\big)} - \sqrt{\mu\big(s(t)\big)}\Big]\mathrm{d}w_2(t)\bigg), \quad \zeta(0) = \zeta_0,
\end{aligned} \tag{5.28}$$

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

$$\kappa(0) = 0,$$

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

Proof. We can represent process $(Q_n(t), Z_n(t))$, $t \geq 0$, as a vector-valued RPSM. As in section 5.2.2, the switching times $0 = t_{n0} < t_{n1} < \cdots$ are the sequential times of jumps of $Q_n(t)$. Let us define the variables $\tau_{nk}(nq)$ and $\overline{\xi}_n(nq) = (\xi_n^{(1)}(nq), \xi_n^{(2)}(nq))$ as follows. If at time t_{nk}, $(n^{-1}Q_n(t_{nk}), n^{-1}Z_n(t_{nk})) = (q, g)$, then the 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 rate $\Lambda(q) = \lambda(q) + \mu(q)$, 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 4.3, 4.4. Let us follow the notation used in equations (4.30) and (4.31). If $\alpha = (q, g)$ $z = (z_1, z_2)$, then $m(\alpha) = \Lambda(q)^{-1}$, $b(\alpha) = (\lambda(q) - \mu(q), \mu(q))\Lambda(q)^{-1}$. Correspondingly, $q_n(\alpha, z) \to q(\alpha, z) = ((\lambda'(q) - \mu'(q))z_1, \mu'(q)z_2)$, and

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

Calculating matrix $D(\alpha)$ using the relation $D^2(\alpha) = D(\alpha)D(\alpha)^*$, we get from equation (4.36) relation (5.28). □

Note that results of this section can also be extended to models that are non-homogenous in time. Consider the following model for illustration.

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

Consider a queueing system described in the section 5.2.2 where the input and service rates depend on time in the following way: 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 function $\lambda_n(q, t)$ satisfies the following condition:

$$\big|\lambda_n(q_1, t_1) - \lambda_n(q_2, t_2)\big| \leq C_{N,L}\big(|q_1 - q_2| + |t_1 - t_2|\big), \tag{5.29}$$

in each bounded domain $\{\max\{t_1, t_2\} \leq N,\ \max\{q_1, q_2\} \leq L,\ q_1, q_2 > 0\}$, and the same condition holds for $\mu_n(\cdot)$.

Let there exist constants $0 < C_0 < C_1 < \infty$ and functions $\lambda(q, t), \mu(q, t)$ such that for any $t \geq 0$, $q > 0$,

$$C_0 \leq \lambda_n(q, t) + \mu_n(q, t) \leq C_1, \tag{5.30}$$

$$\lim_{n\to\infty} \lambda_n(q, t) = \lambda(q, t), \quad \lim_{n\to\infty} \mu_n(q, t) = \mu(q, t). \tag{5.31}$$

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

$$\mathrm{d}s(t) = \big(\lambda\big(s(t),t\big) - \mu\big(s(t),t\big)\big)\mathrm{d}t, \quad s(0) = s_0. \tag{5.32}$$

COROLLARY 5.2. *1) Suppose that equation (5.7) is true with $s_0 > 0$, there exists $T > 0$ such that $s(t) > 0$ as $0 < t \le T$, and $y(+\infty) > T$, where $y(t) = \int_0^t \Lambda(\eta(u),u)^{-1}\mathrm{d}u$, and function $\eta(t)$ satisfies the equation*

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

a unique solution of which exists. Thus relation (5.11) holds.

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

$$\mathrm{d}\zeta(t) = \big(\lambda_q'\big(s(t),t\big) - \mu_q'\big(s(t),t\big)\big)\zeta(t)\mathrm{d}t + \Lambda\big(s(t),t\big)^{1/2}\mathrm{d}w(t), \quad \zeta(0) = \zeta_0.$$

Proof. The proof follows the same scheme as above. We use Theorems 4.3, 4.4. Switching times $t_{n1} < t_{n2} < \cdots$ are chosen as times of any changing of a queueing process $Q_n(t)$. Denote $S_{nk} = (Q_n(t_{nk}), t_{nk})$, $k > 0$. Then the argument α in Theorem 4.3 has the form $\alpha = (q,t)$. For any $q \ge 0$, $t \ge 0$, let us define the family of jointly independent in k variables $(\xi_{nk}(nq,nt), \tau_{nk}(nq,nt))$, $k > 0$, as follows:

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

where variable $\xi_{nk}(nq,nt)$ takes values $+1$ or -1 with probabilities $p_n(q,t)$ or $1 - p_n(q,t)$, respectively. Using relations (5.29)–(5.31) 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, the distribution of $\tau_{nk}(nq,nt)$ 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 $\max\{t_1,t_2\} \le N$, $c \le \min\{q_1,q_2\}$, $\max\{q_1,q_2\} \le L$, with $c > 0$,

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

These relations correspond to condition (4.17). Then we follow the same lines as in the proof of Theorem 5.2 and construct an auxiliary RPSM which satisfies all other conditions of Theorem 5.2. Finally this implies relation (5.11) with $s(t)$ defined in relation (5.32). 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 [MAN 95, MAN 98b, MAN 98a]. 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 a non-homogenous quasi-ergodic Markov environment. Note that limit theorems for SP in a quasi-ergodic Markov environment are considered in [ANI 92a], and also for non-Markov models.

5.2.6. *A system with impatient calls*

As another example of application of Theorem 5.1 we consider a time-homogenous 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 the notation of Theorem 5.1, $\alpha(q) \equiv 1$, $q \geq 0$, $\gamma(q) = 1$, $\beta(q) = -1$, 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 (5.1), (5.3) can be re-written for this case.

Consider a particular case, when $\lambda(q) \equiv \lambda$, $q \geq 0$, $\mu(q) \equiv \mu$, $\nu(q) \equiv \nu$, $q > 0$. Then equations (5.1), (5.3) have the form:

$$\mathrm{d}s(t) = \big(\lambda - \mu - \nu s(t)\big)\mathrm{d}t,$$

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

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

$$s(t) = \nu^{-1}(\lambda - \mu) + \big(s_0 - \nu^{-1}(\lambda - \mu)\big)e^{-\nu t},$$

$$\zeta(t) = e^{-\nu t}\big(\zeta_0 + w\big(\psi(t)\big)\big),$$

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

If $\lambda \geq \mu$, then in the same way as was proved for system $M/M/\infty$ we can show that equation (5.11) holds for any $T > 0$. In this case we have a quasi-stationary point $s^* = \nu^{-1}(\lambda - \mu)$, i.e., as $n \to \infty$ and $t \to \infty$, $n^{-1}Q_n(nt) \xrightarrow{\mathrm{P}} s^*$.

If $\lambda < \mu$, then equation (5.11) holds in the interval $[0, T]$ for any T such that $T < \nu^{-1}\ln((\mu - \lambda + \nu s_0)/(\mu - \lambda))$.

5.3. Non-Markov queueing models

5.3.1. *System* $GI/M_Q/1/\infty$

For the illustration of the approach we consider first a rather simple queueing system $GI/M_Q/1/\infty$ with recurrent input and exponential service with a rate depending on the state of the queue in the system and AP and DA study in the overloaded case.

Assume that the calls enter the system one at a time at the times $t_1 < t_2 < \cdots$ of the events of the renewal flow (the variables $\{t_{k+1} - t_k\}$, $k = 1, 2, \ldots$, are independent identically distributed variables). Suppose that the distribution of inter-arrival times $t_{k+1} - t_k$ coincides with the distribution of variable τ. Let the non-negative function $\mu(\alpha)$, $\alpha \geq 0$, be given. There is one server and an infinite number of waiting places. If a call enters the system at time t_k and the number of calls in the system becomes equal to Q, then the service rate in interval $[t_k, t_{k+1})$ is $\mu(Q/n)$. After service completion the call leaves the system. Let Q_{n0} be the initial number of calls, and let $Q_n(t)$ be the total number of calls in the system at time t. Assuming that corresponding expressions exist, denote

$$m = \mathbf{E}\tau, \quad b(\alpha) = \big(1 - \mu(\alpha)m\big),$$
$$d^2 = \mathbf{Var}\tau, \quad D^2(\alpha) = m\mu(\alpha) + d^2/m^2.$$

STATEMENT 5.3. *Suppose that* $m > 0$, *function* $\mu(\alpha)$ *is locally Lipschitz and has no more than linear growth and* $n^{-1}Q_n(0) \xrightarrow{\text{P}} s_0 > 0$. *Then relation (5.11) holds where*

$$\mathrm{d}s(t) = \big(m^{-1} - \mu\big(s(t)\big)\big)\mathrm{d}t, \quad s(0) = s_0,$$

and T *is any positive value such that* $s(t) > 0$, $t \in [0, T]$.

Suppose in addition that function $\mu(\alpha)$ *is continuously differentiable and*

$$n^{-1/2}\big(Q_n(0) - s_0\big) \xRightarrow{\text{w}} \gamma_0.$$

Then the sequence of processes $\gamma_n(t) = n^{-1/2}(Q_n(nt) - ns(t))$ *J-converges in interval* $[0, T]$ *to the diffusion process* $\gamma(t)$*:* $\gamma(0) = \gamma_0$,

$$\mathrm{d}\gamma(t) = -\mu'\big(s(t)\big)\gamma(t)\mathrm{d}t + \sqrt{\mu\big(s(t)\big) + d^2/m^3}\,\mathrm{d}w(t).$$

Proof. We represent a queueing process $Q_n(\cdot)$ in the system as a process with semi-Markov switching using relations (4.13) and (4.14). Let us choose the switching times as the times t_k when the new calls enter the system and construct an auxiliary embedded RPSM $\widetilde{S}_n(t)$. If in the time interval (t_k, t_{k+1}) the queue is positive, then the output process in this interval follows a Poisson process with rate $\mu(n^{-1}Q)$ where

$Q = Q_n(t_k + 0)$. Therefore, let us define process $\widetilde{S}_n(t)$ according to relations (4.13) and (4.14) where $\tau_{n1}(z) = \tau$ and $\xi_{n1}(n\alpha) = 1 - \Pi_{\mu(\alpha)}(\tau)$, where $\Pi_\mu(t)$ stands for a Poisson process with parameter μ.

It is easy to see that $\mathbf{E}\xi_1(n\alpha) = 1 - \mu(\alpha)m$ and using Theorem 4.3 it is not difficult to prove relation (5.11) for the process $\widetilde{S}_n(t)$. Furthermore, according to relations (4.30) and (4.31) we can calculate that $D^2(\alpha) = m\mu(\alpha) + d^2/m^2$, $q_n(\alpha, z) \to -\mu'(\alpha)$ and prove for the process $\widetilde{S}_n(t)$ DA using the result of Theorem 4.4.

Let us now define the embedded queueing process $\widetilde{Q}_n(t)$ constructed by the times of jumps t_{nk} as follows:

$$\widetilde{Q}_n(t) = Q_n\big(t_{nk} + 0\big) \quad \text{as } t_{nk} < t < t_{n,k+1},\ t > 0. \tag{5.33}$$

Following the lines of proof of Theorem 5.1 we see that the trajectories of the processes $S_n(nt)/n$ and $\widetilde{Q}_n(nt)/n$ asymptotically coincide in the interval $[0, T]$ where $s(t) > 0$.

Now note that the queueing process is monotonically decreasing in each interval (t_k, t_{k+1}). Thus, according to Statements 4.1 and 4.2, J-convergence in interval $[0, T]$ of the embedded process $\widetilde{Q}_n(nt)/n$ to $s(t)$ automatically implies J-convergence of $Q_n(nt)/n$ and also implies J-convergence of the process $\gamma_n(t)$ to $\gamma(t)$. This finally proves Statement 5.3. □

Statement 5.3 is also valid when the service rate at time t has the form $\mu(x_k, n^{-1}Q(nt))$ (may depend on the current value of the queue at time t).

Note that condition $s(t) > 0$, $t \in [0, T]$, means heavy traffic conditions (system is overloaded). This is always true if $\mu(\alpha) < 1/m$, $\alpha > 0$.

5.3.2. *Semi-Markov system* $SM/M_{SM,Q}/1/\infty$

Now we study a more general overloaded queueing system $SM/M_{SM,Q}/1/\infty$ with semi-Markov input and Markov-type service where the service rate depends on the state of the system and the state of a semi-Markov process. Let $x(t)$, $t \geq 0$, be an SMP with values in X which stands for an external random environment. Denote by $\tau(x)$ a sojourn time in state x. Let the non-negative function $\mu(x, \alpha)$, $x \in X$, $\alpha \geq 0$, be given. There is one server and an infinite number of waiting places. Assume that the calls enter the system one at a time at the times $t_1 < t_2 < \cdots$ of the jumps of process $x(t)$. Denote $x_k = x(t_k + 0)$. If a call enters the system at time t_k and the number of calls in the system becomes equal to Q, then the service rate in the interval $[t_k, t_{k+1})$ is $\mu(x_k, n^{-1}Q)$. After service completion the call leaves the system. Let Q_{n0} be the initial number of calls, and $Q_n(t)$ be the total number of calls in the system at time t.

Consider the case when the embedded MP x_k, $k \geq 0$, does not depend on parameter n and is uniformly ergodic with stationary measure $\pi(A)$, $A \in \mathcal{B}_X$. Assuming that corresponding expressions exist, denote

$$\begin{aligned}
m(x) &= \mathbf{E}\tau(x), \quad m = \int_X m(x)\pi(\mathrm{d}x), \quad c(\alpha) = \int_X \mu(x,\alpha)m(x)\pi(\mathrm{d}x),\\
b(\alpha) &= \big(1 - c(\alpha)\big)m^{-1}, \quad G(\alpha) = c'(\alpha),\\
g(x,\alpha) &= 1 - m(x)\big(1 - c(\alpha) + \mu(x,\alpha)m\big)m^{-1},\\
d^2(x) &= \mathbf{Var}\tau(x), \quad d^2 = \int_X d^2(x)\pi(\mathrm{d}x),\\
e_1(\alpha) &= \int_X \mu^2(x,\alpha)d^2(x)\pi(\mathrm{d}x), \quad e_2(\alpha) = \int_X \mu(x,\alpha)d^2(x)\pi(\mathrm{d}x),\\
D^2(\alpha) &= c(\alpha) + e_1(\alpha) + 2\big(1 - c(\alpha)\big)e_2(\alpha)m^{-1} + \big(1 - c(\alpha)\big)^2 d^2 m^{-2}.
\end{aligned}$$

STATEMENT 5.4. *Suppose that $m > 0$, function $\mu(x,\alpha)$ is locally Lipschitz with respect to α uniformly in $x \in X$, function $c(\alpha)$ has no more than linear growth and $n^{-1}Q_n(0) \xrightarrow{\mathrm{P}} s_0 > 0$. Then relation (5.11) holds where*

$$\mathrm{d}s(t) = m^{-1}\big(1 - c\big(s(t)\big)\big)\mathrm{d}t, \quad s(0) = s_0,$$

and T is any positive value such that $s(t) > 0$, $t \in [0,T]$.

Suppose in addition that variables $\tau(x)^2$ are uniformly integrable, function $c(\alpha)$ is continuously differentiable, and

$$n^{-1/2}\big(Q_n(0) - s_0\big) \xrightarrow{\mathrm{W}} \gamma_0.$$

Then the sequence of processes $\gamma_n(t) = n^{-1/2}(Q_n(nt) - ns(t))$ J-converges in interval $[0,T]$ to the diffusion process $\gamma(t)$: $\gamma(0) = \gamma_0$,

$$\mathrm{d}\gamma(t) = -m^{-1}G\big(s(t)\big)\gamma(t)\mathrm{d}t + m^{-1/2}\Big(D^2\big(s(t)\big) + B^2\big(s(t)\big)\Big)^{1/2}\mathrm{d}w(t),$$

where

$$B^2(a) = \mathbf{E}\left(g\big(x_0,\alpha\big)^2 + 2\sum_{k=1}^{\infty} g\big(x_0,\alpha\big)g\big(x_k,\alpha\big)\right),$$

and $P\{x_0 \in A\} = \pi(A)$, $A \in \mathcal{B}_X$.

Proof. At first we represent a queueing process $Q_n(\cdot)$ in the system as a process with semi-Markov switching. Let us construct again an auxiliary embedded RPSM $\widetilde{S}_n(t)$ using relations (4.47) and (4.48) where the variables $\tau_{nk}(x,S)$ do not depend on S

and the times t_k (switching times) are the times of jumps of process $x(t)$. Note that in the time interval (t_k, t_{k+1}) where the queue is positive the output process follows a Poisson process with rate $\mu(x_k, n^{-1}Q_{nk})$ where $x_k = x(t_k + 0)$, $Q_{nk} = Q_n(t_k + 0)$. Therefore, let us define the process $\widetilde{S}_n(t)$ according to relations (4.47) and (4.48) where $\tau_{n1}(x, n\alpha) = \tau(x)$ and $\xi_{n1}(x, n\alpha) = 1 - \Pi_{\mu(x,\alpha)}(\tau(x))$. It is easy to see that $\mathbf{E}\xi_1(x, n\alpha) = 1 - \mu(x, \alpha)m(x)$ and using the result of Theorem 4.5 we can easily prove AP. Furthermore, we can calculate other characteristics according to relations (4.66) and prove DA using the result of Theorem 4.6.

Let us now define the embedded queueing process $\widetilde{Q}_n(t)$ constructed by the times of jumps t_k according to relation (5.33). Following the lines of proof of Theorem 5.1 we see that the trajectories of the processes $S_n(nt)/n$ and $\widetilde{Q}_n(nt)/n$ asymptotically coincide in interval $[0, T]$ where $s(t) > 0$.

Note also that the queueing process is monotonically decreasing in each interval (t_k, t_{k+1}). Thus, according to Statements 4.1 and 4.2, J-convergence of the embedded process $\widetilde{Q}_n(nt)/n$ to $s(t)$ in interval $[0, T]$ automatically implies J-convergence of $Q_n(nt)/n$ and also implies J-convergence of process $\gamma_n(t)$ to $\gamma(t)$. This finally proves Statement 5.4. □

Note that the condition $s(t) > 0$, $t \in [0, T]$, corresponds to a heavy traffic condition. This is always true if $c(\alpha) < 1$, $\alpha > 0$.

Statement 5.4 is also valid when the service rate has the form $\mu(x_k, n^{-1}Q(nt))$ (may depend on the current value of the queue at time t).

Consider a particular case when $\mu(x, \alpha) \equiv \alpha\mu$. Then the system above is equivalent to a system $SM/M/\infty$ with semi-Markov input and exponential service. In this case $c(\alpha) = \alpha\mu m$, $b(\alpha) = 1/m - \alpha\mu$, $G(\alpha) = -\mu$, $D(\alpha)^2 = \mu m\alpha + d^2/m^2$, $g(x, \alpha) = 1 - m(x)/m$, and function $s(t)$ has the form $s(t) = (\mu m)^{-1} - ((\mu m)^{-1} - s_0)e^{-\mu t}$.

5.3.3. *System* $M_{SM,Q}/M_{SM,Q}/1/\infty$

Now consider a queueing system $M_{SM,Q}/M_{SM,Q}/1/\infty$ where the input and service rates depend on the state of an external SMP and the value of the queue. Let $x(t)$, $t \geq 0$, be an SMP with values in $X = \{1, 2, \ldots, d\}$, and let $\tau(i)$ be a sojourn time in state i. Let the family of non-negative functions $\{\lambda(i, \alpha), \mu(i, \alpha),\ \alpha \geq 0\}$, $i \in X$, also be given. There is one server and an infinite number of waiting places. The instantaneous input and service rates depend on the state $x(\cdot)$, the value of the queue and the normalizing factor n in the following way: if at time t, $x(t) = i$ and $Q_n(t) = Q$, then the input rate is $\lambda(i, Q/n)$ and the service rate is $\mu(i, Q/n)$. Calls enter the system one at a time. Note that we consider the times t_k as the switching times, but at these times there are no additional jumps of input flow or completion service.

Denote by x_k, $k \geq 0$, the embedded MP for the process $x(t)$. Assume that x_k is irreducible with the stationary distribution π_i, $i \in X$. Let

$$\begin{aligned} m(i) &= \mathbf{E}\tau(i), \quad m = \sum_{i \in X} m(i)\pi_i, \\ b(\alpha) &= \sum_i \big(\lambda(i,\alpha) - \mu(i,\alpha)\big)m(i)\pi_i/m. \end{aligned} \tag{5.34}$$

STATEMENT 5.5. *Suppose that functions $\lambda(i,\alpha), \mu(i,\alpha)$ are locally Lipschitz with respect to α, $m > 0$, function $b(\alpha)$ has no more than linear growth and $n^{-1}Q_n(0) \xrightarrow{\mathrm{P}} s_0 > 0$. Then relation (5.11) holds where $s(t)$ is a unique solution of equation (5.8) and T is any positive value such that $s(t) > 0$ in the interval $[0, T]$.*

The proof follows the same lines as the proof of Statement 5.4. We can define an auxiliary process $\widetilde{S}_n(t)$ according to relations (4.47) and (4.48) where $\tau_{n1}(i, n\alpha) = \tau(i)$ and $\xi_{n1}(i, n\alpha) = \Pi_{\lambda(i,\alpha)}(\tau(i)) - \Pi_{\mu(i,\alpha)}(\tau(i))$. It is easy to see that $\mathbf{E}\xi_1(i, n\alpha) = (\lambda(i,\alpha) - \mu(i,\alpha))m(i)$ and AP follows from Theorem 4.5.

DA can be formulated in a similar way and we leave this for the readers.

COROLLARY 5.3. *If $x(t)$ is an irreducible MP with the stationary distribution ρ_i, $i \in X$, then function $b(\alpha)$ in equation (5.34) has the form $b(\alpha) = \sum_i (\lambda(i,\alpha) - \mu(i,\alpha))\rho(i)$.*

5.3.4. *System* $SM_Q/M_{SM,Q}/1/\infty$

Consider a non-Markov system $SM_Q/M_{SM,Q}/1/\infty$ which is in some sense a generalization of the system $SM/M_{SM,Q}/1/\infty$ considered in section 5.3.2. There is one server and an infinite number of waiting places. Let x_k, $k > 0$, be a geometrically ergodic MP with values in X and the stationary measure $\pi(A)$. In addition, let $\{\tau_k(x,q),\ q > 0,\ x \in X\}$, $k > 0$, be the independent families of non-negative random variables with distributions not depending on k, and let $\mu(x,q)$ be the non-negative functions.

The calls enter the system one at a time. If a call enters the system at time t_k and the number of calls in the system becomes Q, then the next call enters the system at the time

$$t_{k+1} = t_k + \tau_k\big(x_k, Q/n\big), \quad k \geq 0, \tag{5.35}$$

and the service rate in the interval $[t_k, t_{k+1})$ is $\mu(x_k, Q/n)$.

This system is not a Markov or semi-Markov type system, but the queueing process has a recurrent structure and can be described in terms of SP. Put $m(x, q) = \mathbf{E}\tau_1(x, q)$, and denote

$$
\begin{aligned}
m(q) &= \int_X m(x, q)\pi(\mathrm{d}x), \quad c(q) = \int_X \mu(x, q)m(x, q)\pi(\mathrm{d}x), \\
d^2(x, q) &= \mathbf{Var}\tau_1(x, q), \quad d^2(q) = \int_X d^2(x, q)\pi(\mathrm{d}x), \\
e_1(q) &= \int_X \mu^2(x, q)d^2(x, q)\pi(\mathrm{d}x), \quad e_2(q) = \int_X \mu(x, q)d^2(x, q)\pi(\mathrm{d}x), \\
D^2(q) &= c(q) + e_1(q) + 2\big(\big(1 - c(q)\big)/m(q)\big)e_2(q) \\
&\quad + \big(\big(1 - c(q)\big)/m(q)\big)^2 d^2(q), \quad g(q) = \big(1 - c(q)\big)/m(q), \\
\alpha(x, q) &= 1 - \big(m(x, q)/m(q)\big)\big(1 - c(q) + \mu(x, q)m(q)\big).
\end{aligned}
$$

STATEMENT 5.6. *Assume that*

$$
n^{-1}Q_{n0} \xrightarrow{\mathrm{P}} s_0, \tag{5.36}
$$

functions $m(x, q)$ and $\mu(x, q)$ uniformly in x satisfy local Lipschitz condition with respect to q, and for some $T > 0$ there exists an interval $[0, A]$ such that the equation

$$
\mathrm{d}\eta(t) = \big(1 - c\big(\eta(t)\big)\big)\mathrm{d}t, \quad \eta(0) = s_0,
$$

has a unique solution, $\eta(t) > 0$, $t \in (0, A)$, and

$$
\int_0^A m\big(\eta(t)\big)\mathrm{d}t > T.
$$

Then relation (5.11) holds where function $s(t)$ satisfies the differential equation

$$
\mathrm{d}s(t) = \big(1 - c\big(s(t)\big)\big)m\big(s(t)\big)^{-1}\mathrm{d}t, \quad s(0) = s_0.
$$

In addition, let functions $c(q)$ and $m(q)$ be continuously differentiable, the variables $\tau^2(x, q)$ be uniformly integrable with respect to q, functions $d^2(x, q)$ be continuous in q and

$$
n^{-1/2}\big(Q_n(0) - ns_0\big) \xRightarrow{\mathrm{w}} \gamma_0.
$$

Then the sequence of processes $\gamma_n(t) = n^{-1/2}(Q_n(nt) - ns(t))$ J-converges in interval $[0, T]$ to the diffusion process $\gamma(t)$ satisfying the following SDE:

$$
\begin{aligned}
\mathrm{d}\zeta(t) &= g'\big(\eta(t)\big)\zeta(t)\mathrm{d}t + m\big(\eta(t)\big)^{-1}\big(D^2\big(\eta(t)\big) + B^2\big(\eta(t)\big)\big)^{-1/2}\mathrm{d}w(t), \\
\zeta(0) &= \zeta_0,
\end{aligned} \tag{5.37}
$$

where

$$B^2(q) = \mathbf{E}\left(\alpha^2(x_0, q_0) + 2\sum_{k=1}^{\infty} \alpha(x_0, q_0)\alpha(q, x_k)\right)$$

and the expectation is calculated given that $P(x_0 \in A) = \pi(A)$, $A \in B_X$.

Proof. First we construct an auxiliary SP $(x_n(t), \zeta_n(t))$ according to relations (1.3) and (1.4). Let us choose the switching times t_{nk}, $k > 0$, using relation (5.35). Let $\{\Pi^{(k)}_{\mu(x,q)}(t),\ k \geq 0\}$, be the family of Poisson processes independent in k. Denote $\zeta_{nk}(t, x, q) = 1 - \Pi^{(k)}_{\mu(x,q)}(t)$, and define the recurrent sequences in the following way: $t_{n0} = 0$, $S_{n0} = Q_{n0}$, and

$$\begin{aligned} t_{n,k+1} &= t_{nk} + \tau_k\big(x_k, S_{nk}/n\big), \quad k \geq 0, \\ S_{n,k+1} &= S_{nk} + \xi_{nk}\big(x_k, S_{nk}/n\big), \quad k \geq 0, \end{aligned} \tag{5.38}$$

where $\xi_{nk}(x, q) = \zeta_{nk}(\tau_k(x, q), x, q)$. Put

$$\begin{aligned} \zeta_n(t) &= S_{nk} + \zeta_{nk}\big(t - t_{nk}, x_k, S_{nk}\big), \\ x_n(t) &= x_k, \quad \text{as } t_{nk} \leq t < t_{n,k+1},\ t \geq 0. \end{aligned} \tag{5.39}$$

Also define the embedded RPSM $S_n(t)$:

$$S_n(t) = S_{nk} \quad \text{as } t_{nk} \leq t < t_{n,k+1},\ t \geq 0. \tag{5.40}$$

By definition, if we construct the trajectories of process $\zeta_n(t)$ and queueing process $Q_n(t)$ on the same probability space, then in interval $[0, T]$, where $Q_n(\cdot) > 0$, process $Q_n(t)$ satisfies the same recurrent relations (5.38) and (5.39). Thus, both trajectories coincide in this interval. Therefore, using the approach described in Theorem 5.1, we first prove AP and DA for the normalized embedded process $S_n(t)$. Furthermore, as the trajectory of the queueing process in each interval $(t_{nk}, t_{n,k+1})$ is monotonically decreasing, Statement 4.2 implies that the normalized trajectories of $S_n(t)$ and $\zeta_n(t)$ asymptotically have the same behavior. Finally, according to Theorem 5.1, in any interval $[0, T]$ such that $s(t) > 0$, $0 \leq t \leq T$, the normalized processes $\zeta_n(t)$ and $Q_n(t)$ asymptotically have the same behavior.

To prove AP for process $S_n(t)$ we need to check the conditions of Theorem 4.5. It is easy to see that $\mathbf{E}\xi_{nk}(nq, x) = 1 - \mu(x, q)m(x, q)$. Calculating other characteristics in relations (4.49) we can see that all conditions of Theorem 4.5 are satisfied and relation (5.11) is true. The second part of Theorem 5.6 on the DA follows from Theorem 4.6. □

Consider some particular examples.

EXAMPLE 5.1. Let $\tau_1(x,q) = \tau(x)$, $\mu(x,q) = q\mu$. Then the system above is equivalent to the system $SM/M/\infty$ with semi-Markov input, exponential service with rate μ and an infinite number of servers. In this case $m(q) = m$ and it is easy to calculate that $s(t) = 1/(m\mu) - (1/(m\mu) - s_0)e^{-\mu t}$, and $g'(q) = -\mu$, $D^2(q) = \mu mq + d^2/m^2$, $\alpha(x,q) = (m - \mathbf{E}\tau(x))/m$. As $s(t) > 0$ for any $t > 0$, then the convergence holds in any interval.

EXAMPLE 5.2. Let $\tau_1(x,q) = \tau(x)$, $\mu(x,q) = \mu$. Then the system above is equivalent to the system $SM/M/1/\infty$ with semi-Markov input, one server with service rate μ and an infinite number of waiting places. In this case it is easy to calculate that $s(t) = s_0 + (1/m - \mu)t$. In the overloaded case $(1/m \geq \mu)$ the average input rate is no less than the service rate and the convergence holds in any interval. If $1/m < \mu$, the convergence holds only in the interval $[0, T_*)$ where $T_* = s_0(\mu - 1/m)^{-1}$.

5.3.5. *System* $G_Q/M_Q/1/\infty$

Consider a system $G_Q/M_Q/1/\infty$ with the recurrent input depending on the value of the queue. In fact this system is a simplification of the system $SM_Q/M_{SM,Q}/1/\infty$ considered in section 5.3.4 as we omit the additional Markov environment. There is one server and an infinite number of waiting places. Function $\mu(\alpha), \alpha \geq 0$, and the family of non-negative random variables $\{\tau(\alpha), \alpha \geq 0\}$ are given. The characteristics of the system depend on the scaling factor n in the following way: if a call enters the system at time t_{nk} and $Q_n(t_{nk} + 0) = Q$, then the next call enters the system at time

$$t_{nk+1} = t_{nk} + \tau(Q/n)$$

and the service rate in the interval (t_{nk}, t_{nk+1}) is $\mu(Q/n)$.

Suppose that there exists a function $d^2(q) = E\tau^2(q)$, $q \geq 0$. Put

$$m(q) = E\tau(q), \quad d^2(q) = \mathbf{Var}\tau(q), \quad b(q) = 1 - \mu(q)m(q),$$

$$\tilde{b}(q) = b(q)/m(q), \quad D^2(q) = \mu(q)m(q) + d^2(q)/m(q)^2.$$

STATEMENT 5.7. *Let condition (5.7) hold, functions $m(q)$ and $\mu(q)$ satisfy a local Lipschitz condition and for some $T > 0$ there exists an interval $[0, A]$ such that the equation*

$$\mathrm{d}\eta(t) = b\big(\eta(t)\big)\mathrm{d}t, \quad \eta(0) = s_0,$$

has a unique solution, $\eta(t) > 0$, $t \in (0, A)$, and

$$\int_0^A m\big(\eta(t)\big)\mathrm{d}t > T.$$

Then relation (5.11) holds where

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

If functions $b(q)$ and $m(q)$ are continuously differentiable, variables $\tau(q)^2$ are uniformly integrable with respect to q, function $d(q)$ is continuous in q and equation (5.12) holds, then the sequence of processes $\gamma_n(t) = n^{-1/2}(Q_n(nt) - ns(t))$ J-converges in interval $[0, T]$ to the diffusion process $\gamma(t)$ satisfying the following SDE:

$$\mathrm{d}\gamma(t) = \tilde{b}'\big(s(t)\big)\gamma(t)\mathrm{d}t + m\big(s(t)\big)^{-1/2}D\big(s(t)\big)\mathrm{d}w(t), \quad \gamma(0) = \gamma_0.$$

Proof. We use the same approach as in the previous theorems and construct an auxiliary SP $(x_n(t), \zeta_n(t))$ according to relations (1.3), (1.4). The switching times are the values t_{nk}, variables $\tau_{nk}(nq)$ have the same distribution as variable $\tau(q)$, and variables $\xi_{nk}(nq)$ have the same distribution as variable

$$\xi(q) = 1 - \Pi_{\mu(q)}\big(\tau(q)\big).$$

It is obvious that $\mathbf{E}\xi(q) = b(q)$. Calculating other characteristics and following the lines of proof of Theorem 5.1 and Statement 5.6 we prove Statement 5.7. □

Note that the result of Statement 5.7 is also valid if the service rate at each time t may depend on the normalized value of the queue in the form $\mu(n^{-1}Q_n(t))$.

5.3.6. *A system with unreliable servers*

To illustrate the wide range of possibilities of the suggested approach let us consider a system $GI/M_Q/r/\infty$ with unreliable servers. Calls enter the system one at a time according to a renewal process where the interarrival times are iidrv τ_k, $k \geq 1$. There are r identical servers which are subject to random failures and an infinite number of waiting places. Let the non-negative functions $\{\mu(q),\ q > 0\}$, and the values $\{\nu_i,\ i = \overline{1, r},\ \kappa_i,\ i = \overline{0, r-1}\}$ be given. Denote by $Q_n(t)$, $t \geq 0$, the 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 each operating server in the interval (t_k, t_{k+1}) has a service rate $\mu(Q/n)$.

Let $y(t)$ be a number of operating (not failed) servers at time t. If $y(t) = i$, then each operating server has a failure rate ν_i. If at the failure instant there is a call on service, then this call goes back to the queue. Each failed server has a repair rate κ_i. After repair a server immediately takes a call for service if there are calls waiting in the queue. By construction, process $y(t)$ does not depend on the value of the queue and on index n and is a Birth-and-Death process with state space $\{0, 1, \ldots, r\}$ and birth and death rates $(r - i)\kappa_i$ and $i\nu_i$, respectively. 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 $m = \mathbf{E}\tau_1$, $\widehat{\rho} = \sum_{i=1}^{r} i\rho_i$.

STATEMENT 5.8. *Suppose that 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*

$$\mathrm{d}s(t) = \big(m^{-1} - \widehat{\rho}\mu\big(s(t)\big)\big)\mathrm{d}t, \quad s(0) = s_0 \tag{5.41}$$

exists in an interval $[0, T]$, and $s(t) > 0$, $t \in [0, T]$. Then equation (5.11) holds.

Proof. Denote by $y_i(t)$ a Birth-and-Death process $y(t)$ with the initial state i. Let $\Pi_i(t, y_i(\cdot), q)$, $t \geq 0$, be a Poisson process modulated by $y_i(t)$ in the following way: $\Pi_i(0, y_i(\cdot), q) = 0$, and if at time t, $y_i(t) = j$, then the local rate of a jump at time t is $j\mu(q)$. Denote by $x(t)$, $t \geq 0$, an embedded SMP with state space $\{0, 1, \ldots, r\}$ which is constructed with the help of $y(t)$ in the following way: the times of jumps are chosen as the arrival times of calls t_k, $k \geq 0$. If $y(t_k + 0) = i$, then we put $x(t_k + 0) = i$. Sojourn times in any state have the same distribution as variable τ_1, and transition probabilities p_{ij} of the embedded 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 in index $k \geq 0$ random processes $\zeta_{nk}(t, i, Q)$, having the same distribution as the process $1 - \Pi_i(t, y_i(\cdot), Q/n)$, $t \geq 0$, $i = \overline{0, r}$, $Q > 0$.

Now let $(x(t), \widetilde{Q}_n(t))$, $t \geq 0$, be an auxiliary RPSM which is constructed with the help of $x(t)$ and processes $\{\zeta_{nk}(t, i, Q);\ t \geq 0\}$, $k > 0$, according to equation (1.14). By definition a trajectory of queue $Q_n(t)$ coincides with $\widetilde{Q}_n(t)$ in the domain $\widetilde{Q}_n(t) > 0$, $t \in [0, T]$. Then, using the same arguments as in the proof of Theorem 5.1, we see that it is enough to prove the *AP* for $\widetilde{Q}_n(t)$. Let us use Theorem 4.5. In our case $\xi_{n1}(i, nq) = 1 - \Pi_i(\tau_1, y_i(\cdot), q)$. It is easy to calculate, that the stationary distribution of the embedded MP $x_k = y(t_k + 0)$ is also ρ_i, $i = \overline{0, r}$, and for any $t > 0$,

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

Therefore, $\sum_{i=0}^{r} \rho_i \mathbf{E}\xi_{n1}(i, nq) = 1 - m\widehat{\rho}\mu(q)$. All other conditions of Theorem 4.5 are satisfied, and equation (4.57) has the form of equation (5.41). □

If a service rate depends on the number of operating devices (equal to $\mu_i(q)$ when $y(t) = i$), then equation (5.11) also holds, where in equation (5.41) the expression $\widehat{\rho}\mu(s(t))$ should be changed to $\widehat{\mu}(s(t)) = \sum_{i=1}^{r} \rho_i i\mu_i(s(t))$.

Using Theorem 4.6 we can also prove DA for $Q_n(t)$.

Results of Statement 5.8 can be extended to the system $G_Q/M_Q/r/\infty$ with inter-arrival times and rates ν_i, κ_i, also depending on the current size of the queue $Q_n(t)/n$. 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. However, it is possible to use a general representation of the queue in terms of SP and also use AP for so-called quasi-ergodic MP (e.g. section 3.3 and [ANI 92a]).

More examples of non-Markov and even non-semi-Markov models of the types $G_Q/M_Q/1/\infty$, $SM_Q/M_Q/1/\infty$ and $(G_Q/M_Q/1/\infty)^r$ are considered in [ANI 93, ANI 94a, ANI 95, ANI 99b, ANI 00, ANI 02, ANI 92b].

5.3.7. *Polling systems*

Consider a polling system defined in section 2.2.4. The system consists of r stations and a single moving server. Suppose that the service rates depend on the normalized size of the queue in the following way: if upon the arrival at station j at time t_k a server sees Q_j calls waiting there, then the service rate in the time interval of the length $\kappa_k(j)$ is $\mu_j(Q_j/n)$. 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 notations of section 2.2.4. Suppose that an MP with transition probabilities p_{ij}, $i, j = \overline{1,r}$, is ergodic with stationary distribution π_i.

STATEMENT 5.9. *Assume that for any* $i = \overline{1,r}$, *the values* $m_i = \mathbf{E}\kappa_1(i)$ *and* $\widetilde{m}_i = \mathbf{E}\widetilde{\kappa}_1(i)$ *exist, the functions* $\mu_i(q)$, $q > 0$, *satisfy a local Lipschitz condition,* $\overline{Q}_n(0) \xrightarrow{\text{P}} \overline{s}_0 = (s_{01}, \dots, s_{0r})$, *a unique solution of the equation*

$$\mathrm{d}s_i(t) = \big(\lambda_i - \widehat{\rho}_i\mu_i\big(s_i(t)\big)\big)\mathrm{d}t, \quad s_i(0) = s_{0i}, \tag{5.42}$$

exists in an interval $[0,T]$ *at each* $i = \overline{1,r}$, *and* $s_i(t) > 0$, $t \in [0,T]$, *where*

$$\widehat{\rho}_i = \pi_i m_i \left(\sum_{j=1}^{r} \pi_j \big(m_j + \widetilde{m}_j\big) \right)^{-1}.$$

Then relation (5.4) holds with $\overline{s}(t) = (s_i(t),\ i = \overline{1,r})$.

Proof. Let us construct an auxiliary PSMS. Let $\Pi_k(t,i,\lambda_i)$ and $\widetilde{\Pi}_k(t,i,\mu_i)$ be the Poisson processes, which are independent at different k, i, 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:

$$\zeta_{nk}^{(i)}\big(t,i,Q_i\big) = \Pi_k\big(t,i,\lambda_i\big) - \widetilde{\Pi}_k\big(t,i,\mu_i\big(Q_i/n\big)\big), \text{ as } 0 \le t \le \kappa_k(i);$$

$$\zeta_{nk}^{(i)}\big(t,i,Q_i\big)=\Pi_k\big(t,i,\lambda_i\big)-\widetilde{\Pi}_k\big(\kappa k(i),i,\mu_i\big(Q_i/n\big)\big),\ \text{as}\ \kappa_k(i)<t\le\kappa_k(i)+\widetilde{\kappa}_k(i);$$

$$\zeta_{nk}^{(j)}\big(t,i,Q_j\big)=\Pi_k\big(t,j,\lambda_j\big),\ \text{as}\ 0\le t\le\kappa_k(i)+\widetilde{\kappa}_k(i),\ j=\overline{1,r},\ j\ne i.$$

Denote by $(x(t),\widetilde{Q}_n(t))$, $t\ge 0$, an auxiliary PSMS constructed according to equation (1.14) by the introduced processes and SMP $x(t)$, introduced in section 2.2.4. Following the lines of proof of Theorem 5.1 we see that by the construction, if in 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 queueing process $\overline{Q}_n(t)$. Therefore, it is enough to prove AP for $\widetilde{Q}_n(t)$.

Now we can use Theorem 4.5 and Corollary 4.4. In this case $\overline{\xi}_n(i,\overline{Q})=\overline{\zeta}_{n1}(\kappa_1(i)+\widetilde{\kappa}_1(i),i,\overline{Q})$ and $\tau_n(i)=\kappa_1(i)+\widetilde{\kappa}_1(i)$. It is not so hard to check all conditions of Theorem 4.5 and calculate that the function $m^{-1}b(\overline{q})$ in equation (4.57) has the form $(\lambda_i-\widehat{\rho}_i\mu_i(q_i),\ i=\overline{1,r})$. Thus, relation (5.4) is proved. □

The interval $[0,T]$ where the convergence holds, depends on the representation of the service rate. For example, if $\mu_i(q)=\alpha_i+\mu_i q$ and $\widehat{\rho}_i\alpha_i<\lambda_i, i=\overline{1,r}$, then relation (5.4) holds for any $T>0$, and equation (5.42) has a point of stability $\overline{s}^*=((\lambda_i-\widehat{\rho}_i\alpha_i)/\mu_i,\ i=\overline{1,r})$.

NOTE 5.3. Using Theorem 4.6 and Corollary 4.5 we can also prove that the sequence $\overline{\gamma}_n(t)=n^{-1/2}(\overline{Q}_n(nt)-n\overline{s}(t))$ J-converges in $\mathcal{D}_T$ to the diffusion process $\overline{\gamma}(t)$ satisfying the equation:

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

where $G(\overline{q})$ is a diagonal matrix with elements $-\widehat{\rho}_i\mu_i'(q_i)$, and matrix $B^2(\overline{q})$ is calculated by vectors $\overline{\xi}_n(i,\overline{Q})$ and the embedded MP, $x(t_k+0)$, $k\ge 0$, according to relations (4.66), (4.70).

Note that using this 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 [ANI 95, ANI 97].

5.4. Retrial queueing systems

Retrial queues are comparatively a new direction in queueing models. Over recent years there have appeared many publications concerning the development of approximating methods and analysis of steady-state behavior for different classes of retrial queueing models (see reviews by Yang and Templeton [YAN 87], Falin [FAL 90] and Kulkarni and Liang [KUL 97], the book by Falin and Templeton [FAL 97] and other papers [ART 99, FAL 95, ART 96, MAR 95].

This is a wide and fruitful direction of applications of limit theorems for SP. Note that AP and DA type theorems for different 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 a state-dependent Markov arrival process, general or exponential service and an asymptotically small rate of retrial calls are studied in [ANI 99a, ANI 99c, ANI 01].

In retrial systems customers finding the server busy may join the special retrial queue and repeat their attempts for service after some random time. Consider several examples of retrial queueing systems.

5.4.1. *Retrial system* $M_Q/G/1/w.r$

This system is described in section 2.2.5. There is one server and an infinite number of waiting places. Calls enter the system one at a time. If the server is free it immediately takes the call for service. If the server is busy the call will wait in a queue.

Suppose that system characteristics depend on a parameter n, $n \to \infty$. Let $\lambda(q)$, $\nu(q)$, $q \geq 0$, be given non-negative functions. Denote by $Q_n(t)$ the number of calls in the queue at time t. Let $t_{n1} < t_{n2} < \cdots$ be sequential points of service completion, $t_{n0} = 0$. Denote $Q_{nk} = Q_n(t_{nk} + 0)$ and assume that in the interval $[t_{nk}, t_{nk+1})$ an input flow is a Poisson flow with parameter $\lambda(Q_{nk}/n)$ and each call in the queue independently of other calls with local rate $n^{-1}\nu(Q_{nk}/n)$ may re-apply for service. If the server is free, it immediately takes the call for service. If the server is busy, the call remains in the queue and repeats its attempts for service in the same way. A service time κ_n does not depend on the type of a call (if the call appears from the input flow or from the orbit), and has a general distribution function $B_n(x) = \mathbf{P}(\kappa_n \leq x)$ with finite moments of the first and second order m_n and $m_n^{(2)}$. Denote $\Lambda(q) = \lambda(q) + q\nu(q)$.

THEOREM 5.3. *1) If functions $\lambda(q)$, $\nu(q)$ are locally Lipschitz, $\lambda(q) > 0$, $\nu(q) > 0$, $q \geq 0$, $\Lambda(q) \leq L(1 + q)$, as $n \to \infty$, $m_n \to m$, $n^{-1}Q_n(0) \xrightarrow{\text{P}} s_0$, and variables κ_n are uniformly integrable, that means,*

$$\lim_{L\to\infty} \limsup_{n\to\infty} \mathbf{E}\kappa_n \chi\big(\kappa_n > L\big) = 0,$$

then for any $T > 0$, relation (5.11) is true, where function $s(t)$ satisfies the equation:

$$\mathrm{d}s(t) = \Big(\lambda\big(s(t)\big) - m\big(s(t)\big)^{-1}\Big)\mathrm{d}t, \quad s(0) = s_0, \tag{5.43}$$

with $m(q) = m + (\lambda(q) + q\nu(q))^{-1}$.

2) If in addition functions $\lambda(\cdot)$ and $\nu(\cdot)$ are continuously differentiable and as $n \to \infty$,

$$\sqrt{n}\big(m_n - m\big) \longrightarrow 0, \quad m_n^{(2)} \longrightarrow m^{(2)}, \quad n^{-1/2}\big(Q_n(0) - ns_0\big) \xrightarrow{\text{w}} \zeta_0,$$

and variables κ_n^2 are uniformly integrable, that means,

$$\lim_{L\to\infty} \limsup_{n\to\infty} \mathbf{E}\kappa_n^2\chi\big(\kappa_n > L\big) = 0,$$

then the sequence of processes $\zeta_n(t) = n^{-1/2}(Q_n(nt) - ns(t))$ for any $T > 0$ J-converges in $\mathcal{D}_T$ to a diffusion process $\zeta(t)$ satisfying the following SDE:

$$\mathrm{d}\zeta(t) = g\big(s(t)\big)\zeta(t)\mathrm{d}t + m\big(s(t)\big)^{-1/2} D\big(s(t)\big)\mathrm{d}w(t), \quad \zeta(0) = \zeta_0.$$

Here $g(q) = \frac{d}{dq}(\lambda(q) - m(q)^{-1})$,

$$D^2(q) = \lambda(q)m(q) + \sigma(q)^2 m(q)^{-2} - 2\lambda(q)m(q)^{-1}\big(\lambda(q) + q\nu(q)\big)^{-2},$$
$$\sigma^2(q) = \sigma^2 + \big(\lambda(q) + q\nu(q)\big)^{-2}, \quad \sigma^2 = m^{(2)} - m^2.$$

NOTE 5.4. As function $s(t)$ has a continuous derivative and for any t such that $s(t) = 0$, relation $s'(t) = \lambda(0) - \lambda(0)(1 + \lambda(0)m)^{-1} > 0$ is true, then the assumption $s_0 \geq 0$ implies that the solution of equation (5.43) is strictly positive in any interval $[0, T]$.

Proof. We represent the process $Q_n(t)$ as an SP as it was described in section 2.2.5. Let us choose times t_{nk} as switching times. Denote $\tau_{nk}(nq) = t_{nk+1} - t_{nk}$ given that $Q_{nk} = nq$. Then

$$\mathbf{P}\big(\tau_{nk}(nq) \leq x\big) = \mathbf{P}\big(\eta\big(\Lambda(q)\big) + \kappa_n \leq x\big),$$

where $\eta(\Lambda(q))$ is an independent of κ_n exponentially distributed random variable with parameter $\Lambda(q)$. Let $\xi_{nk}(Q_{nk}) = Q_{nk+1} - Q_{nk}$. Then

$$\mathbf{P}\big(\xi_{nk}\big(Q_{nk}\big) \leq x \mid Q_{nk} = nq,\ \kappa_n = z\big)$$
$$= \big(\lambda(q) + q\nu(q)\big)^{-1}\big(q\nu(q)\mathbf{P}\big(\Pi_{\lambda(q)}(z) - 1 \leq x\big) + \lambda(q)\mathbf{P}\big(\Pi_{\lambda(q)}(z) \leq x\big)\big),$$

where $\Pi_b(t)$ stands for a Poisson process with parameter b.

Now we can construct a PSMS $\widetilde{Q}_n(t)$ with the help of processes $\zeta_{nk}(t, q)$ defined according to relations (2.6). This process is equivalent to the queueing process $Q_n(t)$ in the region $Q_n(t) > 0$ if we construct both processes on the same probability space in the same way as was done in Theorem 5.1.

Now let us introduce an embedded process $\widehat{Q}_n(t) = Q_{nk}$ as $t_{nk} \leq t < t_{nk+1}$. Process $\widehat{Q}_n(t)$ is an RPSM constructed with the help of variables $\tau_{nk}(nq), \xi_{nk}(nq)$. It is easy to calculate the first and second moments of these variables and check the conditions of Theorem 4.3. Thus, relation (5.11) holds for the process $\widehat{Q}_n(t)$. Furthermore, as in our case the trajectory of $Q_n(t)$ in each interval (t_{nk}, t_{nk+1}) is monotonically increasing, according to Statement 4.2, the normalized trajectories of $\widetilde{Q}_n(t)$ and $\widehat{Q}_n(t)$ asymptotically have the same behavior. Finally, according to Theorem 5.1, in any interval $[0, T]$ such that $s(t) > 0$, $0 \le t \le T$, the normalized processes $\widetilde{Q}_n(t)$ and $Q_n(t)$ asymptotically have the same behavior. This finally implies the statement of the first part of Theorem 5.3. To prove the second part of Theorem 5.3 we need to use the notation used in equations (4.30) and (4.31). It is easy to check conditions (4.32) and (4.33). Condition (4.35) follows from the condition of the uniform integrability of the variables κ_n. Finally this implies the statement of the second part of Theorem 5.3. □

EXAMPLE 5.3. If $\lambda(q) \equiv \lambda$, $\nu(q) \equiv \nu$, then our system is equivalent to a classical retrial system with Poisson input, constant retrial rate and general service time. In particular, if $\lambda m < 1$, there exists a stationary point s^* of equation (5.43): as $t \to \infty$, $s(t) \to s^* = \lambda^2 m((1-\lambda m)\nu)^{-1}$. In a stationary case (when $s_0 = s^*$), $s(t) \equiv s^*$, and process $\zeta(t)$ satisfies the equation:

$$\mathrm{d}\zeta(t) = -(1-\lambda m)^2 \nu \zeta(t)\mathrm{d}t + \lambda\sqrt{\lambda\sigma^2 + 2m - \lambda m^2}\,\mathrm{d}w(t), \quad \zeta(0) = \zeta_0. \quad (5.44)$$

This is an Ornstein-Uhlenbeck process. Solving equation (5.44) we obtain:

$$\zeta(t) = e^{-at}\zeta_0 + b\int_0^t e^{-a(t-u)}\mathrm{d}w(u),$$

where $a = (1-\lambda m)^2\nu$, $b^2 = \lambda^2(\lambda\sigma^2 + 2m - \lambda m^2)$.

Note that as $t \to \infty$, the distribution of $\zeta(t)$ weakly converges to a Gaussian distribution with parameters $(0, b^2(2a)^{-1})$. Thus, in a stationary case at large n and t we can use the approximation:

$$Q_n(nt) \approx ns^* + \sqrt{n}b(2a)^{-1/2}\mathcal{N}(0, 1).$$

The result of Theorem 5.3 is also true if functions $\lambda(\cdot)$, $\nu(\cdot)$ depend on the number of calls $Q_n(t)$ at current time t.

Using Theorems 4.5, 4.6 these results can be extended to the case when there are additional Markov switches at times t_{nk}, and to the case when the server is not reliable [ANI 91, ANI 94b].

Similar results (AP and DA) for a system $\overline{M}_Q/\overline{G}/\overline{1}/w.r$ which is described as a one-server system with multiple Poisson input (a call of type i has an input rate λ_i, $i = 1, \ldots, r$, $r < \infty$), general service depending on the type of a call, and rates of repeated calls in the queue $\{\nu_i(\overline{q}),\ \overline{q} \in \mathcal{R}_+^r\}$ depending on the current vector of all waiting calls are considered in the following section.

5.4.2. *System* $\bar{M}/\bar{G}/\bar{1}/w.r$

Let us consider a one-server system with multiple Poisson input (a call of type i has the input rate λ_i, $i = \overline{1,r}$, $r < \infty$). Let a family of distribution functions $\{F_i(x),\ i = \overline{1,r}\}$, $(F_i(0) = 0)$, the values $\{q_i,\ i = \overline{1,r}\}$, $(0 \le q_i \le 1)$, and a family of continuous functions $\{\nu_i(\bar{s}),\ i = \overline{1,r},\ \bar{s} \in R^r_+\}$ also be given. The service discipline is organized in the following way. If a call of type i enters the system and finds the server idle, then it goes directly to the server and the service time is an independent random variable κ_i with distribution function $F_i(x)$. Calls waiting to try the service again are said to be in "orbit". If the incoming call finds the server busy it goes directly into "orbit".

Denote $\bar{Q}_n(t) = \{Q_n^{(i)}(t),\ i = \overline{1,r}\}$, where $Q_n^{(i)}(t)$ is the number of calls of the type i in the orbit. If $\bar{Q}_n(t) = n\bar{s}$, then in the small interval $[t, t+h]$ each call in orbit, independently of others, can re-apply for service with probability $\frac{1}{n}\nu(\bar{s})h + o(h)$. If a call finds the server idle, then the server immediately takes the call and the service time is κ_i. If a call finds the server busy, then it returns to orbit.

Let $\bar{s} = (s_1, s_2, \ldots, s_r)$ be a column-vector. By symbol $\bar{s}^*$ we denote a conjugate vector. Suppose that there exist the means $\mathbf{E}\kappa_i = m_i$, $i = \overline{1,r}$. Let us introduce the following values:

$$\lambda(\bar{s}) = \sum_{i=1}^{r} \big(\lambda_i + s_i\nu_i(\bar{s})\big), \quad \widehat{m} = \sum_{i=1}^{r} m_i\lambda_i. \tag{5.45}$$

Let $\bar{\lambda}$ and $\bar{\alpha}(\bar{s})$ also be the column-vectors with entries λ_i and $m_i\nu_i(\bar{s})$, respectively. Let us introduce matrix $M = \bar{\lambda}\bar{\alpha}(\bar{s})^*$. Denote by I and G the diagonal matrices with the elements 1 and $\nu_i(\bar{s})$, respectively, and put

$$g(\bar{s}) = 1 + \widehat{m} + \big(\bar{\alpha}(\bar{s}), \bar{s}\big), \quad \bar{A}(\bar{s}) = \widehat{m}\bar{\lambda} + (M - G)\bar{s}. \tag{5.46}$$

Now we prove two theorems on the asymptotic behavior of vector $\bar{Q}_n(nt)$.

THEOREM 5.4 (AP). *Suppose that* $n^{-1}\bar{Q}_n(0) \xrightarrow{\text{P}} \bar{s}_0$, *functions* $\nu_i(\bar{s})$ *satisfy the local Lipschitz condition and*

$$m_i > 0, \quad i = \overline{1,r}. \tag{5.47}$$

Then, as $n \to \infty$, *for any* $T > 0$,

$$\sup_{0\le t\le T} \left|n^{-1}\bar{Q}_n(nt) - \bar{s}(t)\right| \xrightarrow{P} 0, \tag{5.48}$$

where $\bar{s}(t)$ satisfies the system of differential equations

$$\mathrm{d}\bar{s}(t) = g\big(\bar{s}(t)\big)^{-1}\bar{A}\big(s(t)\big)\,\mathrm{d}t, \quad \bar{s}(0) = \bar{s}_0, \tag{5.49}$$

and a solution to equation (5.49) exists in each interval and is unique.

Proof. Let us represent process $\bar{Q}_n(t)$ as an SP. Denote by $t_{n1} < t_{n2} < t_{n3} < \cdots$ the sequential times of service completion. We consider these times as switching times. Denote $\bar{Q}_{nk} = \bar{Q}_n(t_{nk})$, $k > 0$, and introduce the family of random variables $\tau_n(\bar{s})$ such that $\mathbf{P}\{\tau_n(\bar{s}) \le x\} = \mathbf{P}\{t_{nk+1} - t_{nk} \le x \mid \bar{Q}_{nk} = n\bar{s}\}$.

If the server is idle, given that $\bar{Q}_{nk} = n\bar{s}$, we have two flows of calls at the server: in the first flow the call of type i has the rate λ_i, the second flow is formed by the calls from orbit and the call of type i can appear with the rate $s_i\nu_i(\bar{s})$. Using the properties of the minimum of independent exponential random variables we can represent variable $\tau_n(\bar{s})$ in the form

$$\tau_n(\bar{s}) = \eta\big(\lambda(\bar{s})\big) + \kappa(\bar{s}), \tag{5.50}$$

where $\eta(\lambda(\bar{s}))$ is an exponential random variable with parameter $\lambda(\bar{s})$ and $\kappa(\bar{s})$ is an independent of $\eta(\lambda(\bar{s}))$ variable and can be represented in the form:

$$\kappa(\bar{s}) = \Big\{\kappa_j \quad \text{with probability } \lambda(\bar{s})^{-1}\big(\lambda_j + s_j\nu_j(\bar{s})\big), \quad j = \overline{1,r}.$$

Introduce indicators $\chi_{j1}(\bar{s})$ ($\chi_{j2}(\bar{s})$) of the following events: after an idle period a call of type j which comes from the input flow (from orbit, respectively) takes the server. This means,

$$\mathbf{P}\big\{\chi_{j1}(\bar{s}) = 1\big\} = 1 - \mathbf{P}\big\{\chi_{j1}(\bar{s}) = 0\big\} = \lambda_j\lambda(\bar{s})^{-1},$$
$$\mathbf{P}\big\{\chi_{j2}(\bar{s}) = 1\big\} = 1 - \mathbf{P}\big\{\chi_{j2}(\bar{s}) = 0\big\} = s_j\nu_j(\bar{s})\lambda(\bar{s})^{-1}.$$

According to these notations we can write that

$$\tau_n(\bar{s}) = \eta\big(\lambda(\bar{s})\big) + \sum_{j=1}^{r}\big(\chi_{j1}(\bar{s}) + \chi_{j2}(\bar{s})\big)\kappa_j. \tag{5.51}$$

Let us construct an auxiliary SP describing the behavior of $Q_n(t)$. In this case there is no discrete component $x(t)$. Let us define the family of processes $\zeta_{nk}(t,\bar{s})$. By definition, the orbit changes according to the following processes: in the idle interval there is no change, and in the busy period there is a Poisson flow with parameter λ_i of calls of type i. Denote by $\bar{\Pi}^{(k)}_{a_i}(t) = (\Pi^{(k)}_{a_i}(t),\ i = \overline{1,r})$, $k \ge 0$, the vector-valued jointly independent at different k Poisson processes, where components $\Pi^{(k)}_{a_i}(t)$ are independent Poisson processes with parameters a_i. Suppose without loss of generality that at time $t_{n0} = 0$ the server is idle. Let us introduce the following process in the interval $[t_{n0}, t_{n1})$

$$\begin{aligned} &\bar{\zeta}_{n0}(t, \bar{s}) = 0 \quad \text{as } t < \eta\big(\lambda(\bar{s})\big), \\ &\bar{\zeta}_{n0}(t, \bar{s}) = -\sum_{j=1}^{r} \bar{e}_j \chi_{j2}(\bar{s}) + \sum_{j=1}^{r} \big(\chi_{j1}(\bar{s}) + \chi_{j2}(\bar{s})\big)\bar{\Pi}^{(j)}_{\lambda_i}(t) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{as } \eta\big(\lambda(\bar{s})\big) < t \le \tau_n(\bar{s}), \end{aligned} \tag{5.52}$$

where $\bar{e}_j$ is the column vector with jth entry equal to 1 and others equal to 0. In each following interval $[t_{nk}, t_{nk+1})$ process $\bar{\zeta}_{nk}(t, \bar{s})$ is constructed in the same way.

Then process $\bar{Q}_n(t)$ is equivalent to an SP constructed by the families $\bar{\zeta}_{nk}(t, \bar{s})$ according to formulae (1.6) and (1.7). Let us introduce the family of vector-valued variables $\bar{\xi}_n(\bar{s})$ as follows:

$$\mathbf{P}\big\{\bar{\xi}_n(\bar{s}) \le x\big\} = \mathbf{P}\big\{\bar{Q}_{nk+1} - \bar{Q}_{nk} \le x \mid \bar{Q}_{nk} = n\bar{s}\big\}.$$

We can then represent $\bar{\xi}_n(\bar{s})$ in the form

$$\bar{\xi}_n(\bar{s}) = -\sum_{j=1}^{r} \bar{e}_j \chi_{j2}(\bar{s}) + \sum_{j=1}^{r} \big(\chi_{j1}(\bar{s}) + \chi_{j2}(\bar{s})\big)\bar{\Pi}^{(j)}_{\lambda_i}\big(\kappa_j\big). \tag{5.53}$$

Now we use Theorem 4.9. For simplicity we omit index k and index n where it is possible. It is easy to calculate that

$$\begin{aligned} m(\bar{s}) = \mathbf{E}\tau_n(\bar{s}) = \lambda(\bar{s})^{-1}\left(1 + \sum_{i=1}^{r} \big(\lambda_i + s_i \nu_i(\bar{s})\big) m_i\right) = \lambda(\bar{s})^{-1} g(\bar{s}), \\ \bar{b}(\bar{s}) = \mathbf{E}\bar{\xi}_n(\bar{s}) = a(\bar{s})^{-1}\big(\widehat{m}\hat{\lambda} + (M - G)\bar{s}\big). \end{aligned} \tag{5.54}$$

Then in our case $g_n(\bar{s}) \le |\xi_n(\bar{s})| + 1$ and condition (4.103) automatically takes place according to Statement 4.1 as the trajectories of each component of the process $\bar{\zeta}_{nk}(t, \bar{s})$ are monotonically increasing. Now let us prove that the convergence in equation (5.48) takes place for any $T > 0$. It is easy to see that

$$m(\bar{s}) \ge \lambda(\bar{s})^{-1} + \min_i m_i \ge \min_i m_i$$

and according to condition (5.47), $\int_0^\infty m(\eta(u))\mathrm{d}u = +\infty$.

Note that the function $\bar{A}(\bar{s}) = m(\bar{s})^{-1} b(\bar{s})$ satisfies the local Lipschitz condition and has no more than linear growth. This means that the solution to equation (5.49) exists in each interval and is unique, and that finally implies Theorem 5.4. □

Let us consider as an example a one-dimensional case (only one type of call). We keep the previous notations and just omit index i and symbol bar$\bar{}$. Then

$$\begin{aligned} \lambda(s) = \lambda + s\nu, \quad m(s) = \lambda(s)^{-1} + m, \\ g(s) = 1 + \lambda m + \nu m s, \quad A(s) = \lambda^2 m + (\lambda m - 1)s\nu(s). \end{aligned} \tag{5.55}$$

COROLLARY 5.4. *Suppose that* $\mathbf{E}\kappa = m > 0$, $n^{-1}Q_n(0) \xrightarrow{\mathrm{P}} s_0$, *and function* $\nu(s)$ *satisfies a local Lipschitz condition. Then relation (5.48) takes places where in equation (5.49) the functions* $g(s)$, $A(s)$ *are given by expression (5.55).*

If $\lambda m < 1$ and $s\nu(s) \to \infty$ as $s \to \infty$, then equation (5.49) has a point of stability s_* which is the minimal solution of equation $s\nu(s) = (1 - \lambda m)^{-1}\lambda^2 m$. In particular if $\nu(s) \equiv \nu$, then $s_* = (1 - \lambda m)^{-1}\nu^{-1}\lambda^2 m$ and $s(t) \to s_*$ as $t \to \infty$.

In the case where $\lambda m = 1$ we obtain unusual behavior for $s(t)$:

$$s(t) = (m\nu)^{-1}\Big(\sqrt{2\lambda^2 m^2 \nu t + \big(1 + \lambda m + m\nu s_0\big)^2} - 1 - \lambda m\Big).$$

Now consider the diffusion approximation. Let us keep the notation of Theorem 5.4. Suppose that there exist second moments $\mathbf{E}\kappa_i^2$, $i = \overline{1,r}$. Put $\sigma_i^2 = \mathbf{Var}\kappa_i$ and introduce the following variables:

$$\begin{aligned} \widehat{m}(\bar{s}) = \sum_{i=1}^{r} m_i s_i \nu_i, \quad \alpha_j(\bar{s}) = \lambda(\bar{s})^{-1}\big(\lambda_j + s_j\nu_j\big), \\ \widehat{\sigma}^2(\bar{s}) = \sum_{j=1}^{r} \alpha_j(\bar{s})\left(\sigma_j^2 + \left(m_j - \sum_{i=1}^{r}\alpha_i(\bar{s})m_i\right)^2\right). \end{aligned} \tag{5.56}$$

Denote $f(\bar{s}) = \lambda(\bar{s})^{-1}(\widehat{m} + \widehat{m}(\bar{s}))$ and introduce vectors:

$$\bar{j}(\bar{s}) = -g(\bar{s})^{-1}f(\bar{s})\bar{\lambda} + \lambda(\bar{s})^{-1}G\bar{s}, \quad \bar{a}(\bar{s}) = g(\bar{s})^{-1}(\bar{\lambda} + G\bar{s}).$$

Let $\bar{\beta}(\bar{s})$ and $\bar{\beta}_m(\bar{s})$ be the column vectors with components $\beta_i(\bar{s})$ and $\beta_i(\bar{s})m_i$, respectively, where $\beta_i(\bar{s}) = \lambda(\bar{s})^{-1}s_i\nu_i(\bar{s})$. We put $B(\bar{s}) = \bar{j}(\bar{s})\bar{\beta}(\bar{s})^* + \bar{a}(\bar{s})\bar{\beta}_m(\bar{s})^*$. In addition, let Λ and $\Lambda_1(\bar{s})$ be diagonal matrices with elements on the diagonal λ_i and $p_i\lambda_i + s_i\nu_i$, respectively. Denote

$$\begin{aligned} D^2(\bar{s}) = g(\bar{s})^{-2}\widehat{\sigma}^2(\bar{s})(\bar{\lambda} + G\bar{s})(\bar{\lambda} + G\bar{s})^* + \lambda(\bar{s})^{-2}G\bar{s}\bar{s}^*G \\ + g(\bar{s})^{-2}\lambda(\bar{s})^{-2}\big(\lambda(\bar{s})f(\bar{s})\bar{\lambda} - G\bar{s}\big)\big(\lambda(\bar{s})f(\bar{s})\bar{\lambda} - G\bar{s}\big)^* \\ - B(\bar{s}) - B(\bar{s})^* + \lambda(\bar{s})^{-1}\Lambda_1(\bar{s}) + f(\bar{s})\Lambda. \end{aligned} \tag{5.57}$$

Furthermore, suppose that functions $\nu_i(\bar{s})$ are continuously differentiable and denote by $Q(\bar{s}) = (g(\bar{s})^{-1}A(\bar{s}))'$ a matrix derivative of the vector $g(\bar{s})^{-1}A(\bar{s})$. Put

$$\bar{\gamma}_n(t) = n^{-1/2}\big(\bar{Q}_n(nt) - n\bar{s}(t)\big), \quad t \in [0,T].$$

THEOREM 5.5 (DA). *Suppose that the conditions of Theorem 5.4 hold and* $n^{-1/2}(\bar{Q}_n(0) - \bar{s}_0) \overset{\mathrm{w}}{\Longrightarrow} \bar{\gamma}_0$.

Then the sequence of processes $\bar{\gamma}_n(t)$ J-converges in any interval $[0,T]$ to the diffusion process $\bar{\gamma}(t)$ satisfying the following stochastic differential equation:

$$\mathrm{d}\bar{\gamma}(t) = Q\big(\bar{s}(t)\big)\bar{\gamma}(t)\mathrm{d}t + D\big(\bar{s}(t)\big)m\big(\bar{s}(t)\big)^{-1/2}\mathrm{d}\bar{w}(t), \quad \bar{\gamma}(0) = \bar{\gamma}_0, \qquad (5.58)$$

where function $\bar{s}(\cdot)$ satisfies equation (5.49), function $m(\bar{s})$ is given in equation (5.54) and $\bar{w}(t)$ is the standard Wiener process in R^r.

Proof. We use the same representation for variables $\tau_n(\bar{s})$ and $\xi_n(\bar{s})$ (see equations (5.51), (5.53)) and Theorems 4.10 and 4.3. Let us calculate the variance of the variable

$$\bar{\rho}_n(\bar{s}) = \bar{\xi}_n(\bar{s}) - \bar{b}_n(\bar{s}) - g(\bar{s})^{-1}\bar{A}(\bar{s})\big(\tau_n(\bar{s}) - m(\bar{s})\big)$$

(see equation (5.54)). For convenience we can split $\bar{\rho}_n(\bar{s})$ into two independent parts: $\bar{\rho}_n^{(1)}(\bar{s}) = -g(\bar{s})^{-1}\bar{A}(\bar{s})(\eta_n(\lambda(\bar{s})) - \lambda(\bar{s})^{-1})$, and the remaining part. After calculations we obtain that $\mathbf{E}\bar{\rho}_n(\bar{s})\bar{\rho}_n(\bar{s})^* = D^2(\bar{s})$. This implies the statement of Theorem 5.5. □

NOTE 5.5. If equation (5.49) has the point of stability $\bar{s}_*$ and $\bar{s}_0 = \bar{s}_*$, then we have a so-called quasi-stationary regime where $\bar{s}(t) \equiv \bar{s}_*$ and process $\bar{\gamma}(t)$ in Theorem 5.5 satisfies the equation:

$$\mathrm{d}\bar{\gamma}(t) = Q\big(\bar{s}_*\big)\bar{\gamma}(t)\mathrm{d}t + D\big(\bar{s}_*\big)m\big(\bar{s}_*\big)^{-1/2}\mathrm{d}w(t), \quad \bar{\gamma}(0) = \bar{\gamma}_0,$$

where $\bar{\gamma}(t)$ is an Ornstein-Uhlenbeck process.

Note that similar models for the one-dimensional case were studied in [ANI 91, ANI 94b].

5.4.3. *Retrial system* $M/M/m/w.r$

Now consider a retrial system with m identical servers and service rate μ. An input is a Poisson flow of identical calls with parameter λ. Denote by $R_n(t)$ a number of busy servers at time t. Let the families $\{p_i(s), q_i(s), r_i(s),\ i = 0, 1, \ldots, m\}$, and $\{\nu(s), \alpha(s), g(s)\}$, $s \geq 0$, of continuous non-negative functions be given. Here for any $s \geq 0$, $i = 0, 1, \ldots, m$, $p_i(s) + q_i(s) + r_i(s) = 1$, $\alpha(s) + g(s) = 1$.

Denote by $Q_n(t)$ the number of waiting calls at time t (calls in orbit). The service process in the system is described in the following way: if a call enters the system at time t and $(R_n(t), Q_n(t)) = (i, nq)$ $(i < m)$, then with probability $p_i(q)$ it is immediately taken for service, with probability $q_i(q)$ the call goes to the queue, and with probability $r_i(q)$ the call gets a refusal and leaves the system (if $i = m$, then we put $p_m(q) \equiv 0$). If $Q_n(t) = nq$, each call in the queue independently of others can re-apply for service with local rate $n^{-1}\nu(q)$. If a call finds a idle server, the service immediately begins. If a call finds all servers busy, then it either returns to the orbit with probability $\alpha(q)$ or with probability $g(q)$ the call leaves the system.

We study the AP for the process $Q_n(nt)/n$. Note that the process $(R_n(t), Q_n(t))$ is an MP and we can represent it as an RPSM. Denote

$$\rho(j,s) = \widehat{c}(s)^{-1} \frac{1}{j!\mu^j} \prod_{i=0}^{j-1} \big(p_i(s)\lambda + s\nu(s)\big), \quad j = \overline{0,m}, \tag{5.59}$$

where

$$\widehat{c}(s) = \sum_{j=0}^{m} \frac{1}{j!\mu^j} \prod_{i=0}^{j-1} \big(p_i(s)\lambda + s\nu(s)\big)$$

and we set $\prod_{i=0}^{-1} = 1$. Let us define the function

$$\widehat{b}(s) = \lambda \sum_{i=0}^{m} \rho(i,s) q_i(s) - s\nu(s)\big(1 - \big(1 - g(s)\big)\rho(m,s)\big). \tag{5.60}$$

THEOREM 5.6. *Suppose that $n^{-1}\bar{Q}_n(0) \xrightarrow{\text{P}} \bar{s}_0$, the functions $p_i(s), q_i(s), g(s), \nu(s)$ satisfy the local Lipschitz condition, for any $s > 0$, $\nu(s) > 0$, and the function $\nu(s)$ is bounded. Then as $n \to \infty$, for any $T > 0$,*

$$\sup_{0 \le t \le T} \big| n^{-1} Q_n(nt) - s(t) \big| \xrightarrow{P} 0 \tag{5.61}$$

where

$$s(0) = s_0, \quad \mathrm{d}s(t) = \widehat{b}\big(s(t)\big)\,\mathrm{d}t, \tag{5.62}$$

and a solution of equation (5.62) exists in each interval and is unique.

Proof. At first we represent a process $(R_n(t), Q_n(t))$ as an SP. In our case the process $(R_n(t), Q_n(t))$ is an MP with values in $\{0, 1, \ldots, m\} \times \{0, 1, \ldots\}$. Denote by $t_{n1} < t_{n2} < \cdots$ the sequential times of any transition in the system. Note that some transitions may not cause a change of state but they are related to some service processes (for example loss of an input call). We consider times $t_{n1} < t_{n2} < \cdots$ as switching times. By the constriction, process $(R_n(t), Q_n(t))$ is an RPSM with feedback between both components. In scale of time nt the first component is quickly varying and we use Theorem 4.7, section 4.5.

We can always define process $(R_n(t), Q_n(t))$ as a right-continuous process. Let us calculate its transition rates. Put $\lambda_i(s) = \lambda + i\mu + s\nu(s)$, $i \le m$. If $Q_n(t) = ns$, then transition rates do not depend on n and we omit index n for simplicity. Let $\lambda((i,s),(j,y))$ denote the transition rate from state $(R_n(t), Q_n(t)) = (i, ns)$ to state (j, ny). Then:

$$\lambda\big((i,s),(j,y)\big) = \begin{cases} i\mu & \text{if } j = i-1,\ y = s; \\ p_i(s)\lambda & \text{if } j = i+1,\ y = s; \\ r_i(s)\lambda & \text{if } j = i,\ y = s; \\ q_i(s)\lambda & \text{if } j = i,\ y = s+1; \\ s\nu(s) & \text{if } j = i+1,\ y = s-1; \\ 0 & \text{otherwise} \end{cases} \qquad 0 \le i < m;$$

$$\lambda\big((m,s),(j,y)\big) = \begin{cases} m\mu & \text{if } j = m-1,\ y = s; \\ r_m(s)\lambda & \text{if } j = m,\ y = s; \\ q_m(s)\lambda & \text{if } j = m,\ y = s+1; \\ g(s)s\nu(s) & \text{if } j = m,\ y = s-1; \\ \alpha(s)s\nu(s) & \text{if } j = m,\ y = s; \\ 0 & \text{otherwise.} \end{cases}$$

Let us introduce the family of random variables $\xi(i,s)$ such that

$$\begin{aligned} &\mathbf{P}\{\xi(i,s) \in C\} \\ &\quad = \mathbf{P}\{Q_n(t_{n2}) - Q_n(t_{n1}) \in C \mid (R_n(t_{n1}), Q_n(t_{n1})) = (i, ns)\}. \end{aligned}$$

Then variable $\xi(i,s)$ can be represented in the form: for $i < m$,

$$\xi(i,s) = \begin{cases} 1 & \text{with probability } \lambda_i(s)^{-1} q_i(s); \\ -1 & \text{with probability } \lambda_i(s)^{-1} s\nu(s); \\ 0 & \text{otherwise,} \end{cases}$$

and at $i = m$,

$$\xi(m,s) = \begin{cases} 1 & \text{with probability } \lambda_m(s)^{-1} q_m(s); \\ -1 & \text{with probability } \lambda_m(s)^{-1} g(s)s\nu(s); \\ 0 & \text{otherwise.} \end{cases}$$

Now we can describe process $(R_n(t), Q_n(t))$ as an SP (see relations (4.92) and (4.93)). In this case variable $\tau_{nk}(i,s)$ has an exponential distribution with parameter

$\lambda_i(s)$ and variable $\xi_{nk}(i,s)$ has the same distribution as variable $\xi(i,s)$ introduced above. Furthermore, at each fixed $s \geq 0$, denote by $\tilde{x}_k(s)$, $k \geq 0$, an MP in discrete time with transition probabilities

$$p_{ij}(s) = \begin{cases} \lambda_i(s)^{-1} i\mu & \text{if } j = i-1; \\ \lambda_i(s)^{-1}(r_i(s) + q_i(s))\lambda & \text{if } j = i; \\ \lambda_i(s)^{-1}(p_i(s)\lambda + s\nu(s)) & \text{if } j = i+1; \\ 0 & \text{otherwise} \end{cases} \qquad i = \overline{0, m-1};$$

$$p_{mj}(s) = \begin{cases} \lambda_m(s)^{-1} m\mu & \text{if } j = m-1; \\ \lambda_m(s)^{-1}\big(\lambda + s\nu(s)\big) & \text{if } j = m; \\ 0 & \text{otherwise.} \end{cases}$$

As can be seen, at any $s > 0$, the state space of $\tilde{x}_k(s)$ forms one essential class. Denote by $\{\pi(i,s),\ i = \overline{0,m}\}$, a stationary distribution for $\tilde{x}_k(s)$, $k \geq 0$. It is easy to see that in any bounded region $\{\delta \leq s \leq L\}$ ($\delta > 0$) process $\tilde{x}_k(s)$ is uniformly ergodic. Let us introduce the functions:

$$\begin{aligned} m(s) &= \sum_{i=0}^{m} \pi(i,s)\lambda_i(s)^{-1}, \quad \widehat{q}(s) = \sum_{i=0}^{m} \pi(i,s) q_i(s)\lambda_i(s)^{-1}, \\ b(s) &= \lambda\widehat{q}(s) - s\nu(s)m(s) + s\nu(s)\big(1 - g(s)\big)\pi(m,s)\lambda_m(s)^{-1}. \end{aligned} \tag{5.63}$$

Denote $\tilde{b}(s) = m(s)^{-1}b(s)$. According to Theorem 4.7, relation (5.61) holds where $s(t)$ is a solution to equation (5.62) (see also (4.95)). Furthermore, as function $\nu(s)$ is bounded, then for some $c_0 > 0$, $\liminf_{s\to\infty} sm(s) > c_0$. This relation implies that

$$\int_0^\infty m\big(\eta(u)\big)\mathrm{d}u = +\infty$$

and the convergence in equation (5.61) holds for any $T > 0$.

Now let us calculate the function $\tilde{b}(s)$ in the explicit form. Note that the values $m(s)^{-1}\pi(i,s)\lambda(i,s)^{-1}$, $i = \overline{0,m}$, at each fixed s are the stationary probabilities of an MP in continuous time $\tilde{x}(t,s)$, $t \geq 0$, which is given by transition rates $\tilde{a}_{ij}(s) = \lambda(i,s)p_{ij}(s)$ (here we allow transitions back to the same state). However, process $\tilde{x}(t,s)$, $t \geq 0$, is equivalent to the Birth-and-Death process $x(t,s)$ with birth and death rates in the state i, $c_i(s)$ and $d_i(s)$, respectively, where $c_i(s) = p_i\lambda + s\nu(s)$, $i < m$, $d_i(s) = i\mu$, $i \leq m$. Therefore, the stationary probabilities of $\tilde{x}(t,s)$ are defined by expression (5.59), and after some algebra we find that $\tilde{b}(s) = \widehat{b}(s)$. This finally proves Theorem 5.6. □

Let us study the cases when equation (5.62) has a point of stability.

CASE 1. Suppose that

$$\inf_{s\geq 0} g(s) = g_0 > 0, \quad \inf_{s\geq 0} \nu(s) = \nu_0 > 0, \quad \sum_{i=0}^{m} q_i(0) > 0.$$

This means that there exists a flow of lost calls in the state m (when all servers are busy). Then $\widehat{b}(0) > 0$ and $\widehat{b}(s) \leq \lambda - s\nu_0 g_0$. Thus, $\widehat{b}(s) \to -\infty$ as $s \to \infty$. Denote by s_* the minimal root of the equation

$$\widehat{b}(s) = 0 \tag{5.64}$$

in the region $(0, \infty)$ which exists according to continuity of the function $\widehat{b}(s)$. In some small neighborhood of s_*, $\widehat{b}(s) > 0$ as $s < s_*$ and $\widehat{b}(s) < 0$ as $s > s_*$. This means that point s_* is the point of stability for the solutions with the initial value s_0 in a neighborhood of s_*.

Note that in this case the stable solution exists for any values of λ and μ. This fact can be explained in the following way: if s is large, then the flow of lost calls also has a large rate no less than $s\nu_0 g_0$.

CASE 2. Suppose that $g(s) \equiv 0$, $q_m(s) \equiv 1$, $\sum_{i=0}^{m} q_i(0) > 0$, $s\nu(s) \to \infty$ as $s \to \infty$, and

$$\lambda < m\mu. \tag{5.65}$$

This means that if a call finds all servers busy, it goes with probability one to the orbit and there is no flow of lost calls in state m. It is not difficult to calculate that:

$$\lim_{s\to\infty} \rho(m, s) = 1, \quad \lim_{s\to\infty} \widehat{b}(s) = \lambda - m\mu. \tag{5.66}$$

As $\widehat{b}(0) > 0$, relations (5.65) and (5.66) imply that the minimal root of equation (5.64) exists and it is the point of stability.

In particular, if $m = 1$, functions $p_i(\cdot), q_i(\cdot), \nu_i(\cdot)$ do not depend on s, and $g(s) \equiv 0$, then:

$$\widehat{b}(s) = \frac{\lambda^2 + s\nu(\lambda - \mu)}{\lambda + \mu + s\nu},$$

which is in agreement with equation (5.55) for the case $q = 1$, $p = 0$, $m = \lambda^{-1}$.

These results show that the technique based on limit theorems of AP and DA types for SP provides us with the new effective approach for studying transient and stable operating regimes for rather complex retrial queueing systems in overloading conditions.

Using this approach, AP for the number of calls in orbit in the overloading case for the Markov multiserver retrial queues with negative arrivals was proved in [ANI 01].

5.5. Queueing networks

5.5.1. *State-dependent Markov network* $(M_Q/M_Q/1/\infty)^r$

Consider a queueing network $(M_Q/M_Q/1/\infty)^r$ which consists of r nodes with one server in each node and an infinite number of waiting places. Denote by $Q_n(i,t)$ the number of calls in the ith node at time t and let $\overline{Q}_n(t) = (Q_n(i,t),\ i = \overline{1,r})$ be the column vector.

We assume that the time goes to infinity in the scale nt and consider the AP and DA for the normalized vector-valued queueing process $\overline{Q}_n(nt)$. Let the functions $\{\lambda_i(\overline{q}), \mu_i(\overline{q}), p_{ij}(\overline{q}),\ i = \overline{1,r},\ j = \overline{0,r},\ \overline{q} \in [0,\infty)^r\}$ be given. The network is operating in the following way. If at some time-point u, $\overline{Q}_n(u) = \overline{Q}$, then the local input rate in the ith node is $\lambda_i(\overline{Q}/n)$ and the local service rate is $\mu_i(\overline{Q}/n)$. If at this time a call has completed service in node i, then either with probability $p_{ij}(\overline{Q}/n)$ it goes from ith to jth node, $j = \overline{1,r}$ or with probability $p_{i0}(\overline{Q}/n)$ it leaves the network. This network belongs to Jackson type networks.

Let $\overline{\lambda}(\overline{q}) = (\lambda_1(\overline{q}), \ldots, \lambda_r(\overline{q}))$, $\overline{\mu}(\overline{q}) = (\mu_1(\overline{q}), \ldots, \mu_r(\overline{q}))$ be the column vector-valued functions,

$$P(\overline{q}) = \big\|p_{ij}(\overline{q})\big\|_{i,j=\overline{1,r}}, \quad a(\overline{q}) = \sum_{i=1}^{r} \big(\lambda_i(\overline{q}) + \mu_i(\overline{q})\big),$$

and for a vector-valued function $\overline{f}(\overline{q}) = (f_1(\overline{q}), \ldots, f_r(\overline{q}))$ with $\overline{q} = (q_1, \ldots, q_r)$ we denote by $f'(\overline{q})$ the matrix derivative: $f'(\overline{q}) = \|\partial f_i(\overline{q})/\partial q_j\|_{i,j=\overline{1,r}}$.

THEOREM 5.7. *Assume that as $n \to \infty$,*

$$n^{-1}\overline{Q}_n(0) \xrightarrow{\text{P}} \overline{s}_0, \tag{5.67}$$

functions $\overline{\lambda}(\overline{q}), \overline{\mu}(\overline{q}), P(\overline{q})$, satisfy a local Lipschitz condition, and there exists $A > 0$ such that the system of differential equations

$$\mathrm{d}\overline{\eta}(t) = \big(\overline{\lambda}\big(\overline{\eta}(t)\big) + \big(P^*\big(\overline{\eta}(t)\big) - I\big)\overline{\mu}\big(\overline{\eta}(t)\big)\big)a\big(\overline{\eta}(t)\big)^{-1}\mathrm{d}t, \quad \overline{\eta}(0) = \overline{s}_0,$$

has a unique solution $\overline{\eta}(t)$ such that $\overline{\eta}(t) > 0$ in each component, $t \in (0,A)$, and $\int_0^A a(\overline{\eta}(t))^{-1}\mathrm{d}t > T$.

Then

$$\sup_{0\le t\le T} \big|n^{-1}\overline{Q}_n(nt) - \overline{s}(t)\big| \xrightarrow{\text{P}} 0, \tag{5.68}$$

where the vector-valued function $\overline{s}(t)$ satisfies the equation:

$$\mathrm{d}\overline{s}(t) = \big(\overline{\lambda}\big(\overline{s}(t)\big) + \big(P^*\big(\overline{s}(t)\big) - I\big)\overline{\mu}\big(s(t)\big)\big)\mathrm{d}t, \quad \overline{s}(0) = \overline{s}_0. \tag{5.69}$$

If in addition

$$n^{-1/2}\big(\overline{Q}_n(0) - n\overline{s}_0\big) \stackrel{\mathrm{w}}{\Longrightarrow} \overline{\gamma}_0, \tag{5.70}$$

and functions $\overline{\lambda}(\overline{q}), \overline{\mu}(\overline{q}), P(\overline{q})$ *are continuously differentiable, then the sequence of processes*

$$\overline{\gamma}_n(t) = n^{-1/2}\big(\overline{Q}_n(nt) - n\overline{s}(t)\big)$$

J*-converges in interval* $[0,T]$ *to a multidimensional diffusion process* $\overline{\gamma}(t)$ *satisfying SDE*

$$\mathrm{d}\overline{\gamma}(t) = G\big(\overline{s}(t)\big)\overline{\gamma}(t)\mathrm{d}t + B\big(\overline{s}(t)\big)\mathrm{d}\overline{\omega}(t), \quad \overline{\gamma}(0) = \overline{\gamma}_0. \tag{5.71}$$

Here

$$G(\overline{q}) = \big(\lambda(\overline{q}) + \big(P^*(\overline{q}) - I\big)\mu(\overline{q})\big)', \quad B(\overline{q})^2 = \big\|b_{ij}(\overline{q})\big\|_{i,j=\overline{1,r}},$$

$$b_{ij}(\overline{q}) = -\mu_i(\overline{q})p_{ij}(\overline{q}) - \mu_j(\overline{q})p_{ji}(\overline{q}), \quad i \neq j,$$

$$b_{ii}(\overline{q}) = -2\mu_i(\overline{q})p_{ii}(\overline{q}) + \lambda_i(\overline{q}) + \mu_i(\overline{q}) + \sum_k \mu_k(\overline{q})p_{ki}(\overline{q}), \quad i = \overline{1,r},$$

and P^* *is a transposed matrix.*

Proof. In this case the process $\overline{Q}_n(t)$ is an MP which can be represented as a simple RPSM. Switching times t_{nk} are the times of sequential changes of the states of process $\overline{Q}_n(t)$. Then variable $\tau_{n1}(n\overline{q})$ has an exponential distribution with parameter $a(\overline{q})$ and the vector-valued variable $\overline{\xi}_{n1}(n\overline{q})$ does not depend on $\tau_{n1}(n\overline{q})$ and can be represented in the form:

$$\xi_{n1}(n\overline{q}) = \begin{cases} \overline{e}_j & \text{with probab. } \lambda_j(\overline{q})a(\overline{q})^{-1}, \\ \overline{e}_j - \overline{e}_i & \text{with probab. } \mu_i(\overline{q})p_{ij}(\overline{q})a(\overline{q})^{-1}, \quad i,j = \overline{1,m},\ i \neq j, \\ -\overline{e}_i, & \text{with probab. } \mu_i(\overline{q})p_{i0}(\overline{q})a(\overline{q})^{-1}, \end{cases}$$

where $\overline{e}_i$ is the column-vector with ith component equal to 1 and other components equal to 0.

Calculating the characteristics of these variables, following the lines of proof of Theorem 5.1 and using the results of Theorems 4.3 and 4.4 we prove the statements of both parts of the theorem. □

Let us consider an example, when $\lambda_i(\overline{q}) = \lambda_i$, $\mu_i(\overline{q}) = \mu_i q_i$, $p_{ij}(\overline{q}) = p_{ij}$, $i = \overline{1,r}$, $j = \overline{0,r}$. This network is equivalent to the classical network $(M/M/\infty)^r$. In that case equation (5.69) has the form

$$\mathrm{d}\overline{s}(t) = \big(\overline{\lambda} + (P^* - I)A\overline{s}(t)\big)\mathrm{d}t, \tag{5.72}$$

where $\overline{\lambda} = (\lambda_1, \ldots, \lambda_r)$, and A is a diagonal matrix with entries μ_i, $i = \overline{1,r}$.

Suppose that matrix $P^* - I$ is invertible. Iterating equation (5.72) we obtain the representation

$$\overline{s}(t) = \overline{q}_0 + \exp\big\{(P^* - I)At\big\}\big(s_0 - q_0\big),$$

where $\overline{q}_0$ is the stationary point: $\overline{q}_0 = A^{-1}(I - P^*)^{-1}\overline{\lambda}$.

Equation (5.71) has the form

$$\mathrm{d}\overline{\gamma}(t) = (P^* - I)A\overline{\gamma}(t)\mathrm{d}t + B\big(\overline{s}(t)\big)\mathrm{d}\overline{w}(t). \tag{5.73}$$

If $\overline{s}_0 = \overline{q}_0$, then for all $t > 0$, $\overline{s}(t) = \overline{q}_0$ and we have a quasi-stationary regime in the sense that for any $t > 0$, $\overline{Q}_n(nt)/n \approx \overline{q}_0$. In this case equation (5.73) has the stationary form with the matrix of diffusion $B = B(\overline{q}_0)$.

These results can be extended to the networks with impatient calls, unreliable servers and batch entry and service.

5.5.2. *Markov state-dependent networks with batches*

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 a 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 variables be given:

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

$$m_i(\overline{q}) = \mathbf{E}\delta_i(\overline{q}), \quad g_i(\overline{q}) = \mathbf{E}\gamma_i(\overline{q}), \quad e_i(\overline{q}) = \mathbf{E}\beta_i(\overline{q}) - 1,$$

$$\Lambda(\overline{q}) = \sum_{i=1}^{r} \big(\lambda_i(\overline{q}) + \mu_i(\overline{q}) + q_i\nu_i(\overline{q})\big), \quad a_i^2(\overline{q}) = \mathbf{E}\delta_i^2(\overline{q}),$$

$$c_i^2(\overline{q}) = \mathbf{E}\gamma_i^2(\overline{q}), \quad d_i^2(\overline{q}) = \mathbf{E}\big(\beta_i(\overline{q}) - 1\big)^2, \quad i = \overline{1,r}.$$

Let us introduce the following column-vector functions:

$$\overline{m}(\overline{q}) = \big(\lambda_1(\overline{q})m_1(\overline{q}), \ldots, \lambda_r(\overline{q})m_r(\overline{q})\big), \quad \overline{g}(\overline{q}) = \big(\mu_1(\overline{q})g_1(\overline{q}), \ldots, \mu_r(\overline{q})g_r(\overline{q})\big),$$

$$\overline{e}(\overline{q}) = \big(q^{(1)}\nu_1(\overline{q}), \ldots, q^{(r)}\nu_r(\overline{q})\big), \quad \overline{b}(\overline{q}) = \overline{m}(\overline{q}) - \big(I - P_0(\overline{q})^*\big)\overline{g}(\overline{q}) + \overline{e}(\overline{q}),$$

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

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 a system of differential equations

$$\overline{s}(0) = \overline{s}_0, \quad \mathrm{d}\overline{s}(t) = \overline{b}\big(\overline{s}(t)\big)\mathrm{d}t. \tag{5.74}$$

THEOREM 5.8. *1) Suppose that in any bounded and closed domain in* $\operatorname{int}\{\mathcal{R}_+^r\}$ *the variables* $\delta_i(\overline{q})$, $\gamma_i(\overline{q})$, $\beta_i(\overline{q})$, $i = \overline{1,r}$, *are integrable uniformly in* $\overline{q}$, *functions* $\lambda_i(\overline{q})$, $\mu_i(\overline{q})$, $\nu_i(\overline{q})$, $m_i(\overline{q})$, $g_i(\overline{q})$, $e_i(\overline{q})$, $i = \overline{1,r}, P(\overline{q})$ *satisfy a local Lipschitz condition,* $\Lambda(\overline{q}) > 0$, $n^{-1}\overline{Q}_n(0) \xrightarrow{\mathrm{P}} \overline{s}_0 > \overline{0}$, *the equation*

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

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

$$\sup_{0 \le t \le T} \left|n^{-1}\overline{Q}_n(nt) - \overline{s}(t)\right| \xrightarrow{\mathrm{P}} 0. \tag{5.75}$$

2) If in addition in any bounded and closed domain in $\operatorname{int}\{\mathcal{R}_+^r\}$ *the 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

$$n^{-1/2}\big(\overline{Q}_n(0) - n\overline{s}_0\big) \overset{\mathrm{w}}{\Longrightarrow} \overline{\gamma}_0,$$

then the sequence $\overline{\gamma}_n(t) = n^{-1/2}(\overline{Q}_n(nt) - n\overline{s}(t))$ J-converges in $\mathcal{D}_T^r$ to the multidimensional diffusion process $\overline{\gamma}(t)$:

$$\mathrm{d}\overline{\gamma}(t) = G\big(\overline{s}(t)\big)\overline{\gamma}(t)\mathrm{d}t + B\big(\overline{s}(t)\big)\mathrm{d}\overline{w}(t), \quad \overline{\gamma}(0) = \overline{\gamma}_0, \tag{5.76}$$

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 define an auxiliary RPSM $\widetilde{Q}_n(t)$ by analogy to Theorem 5.1. This is an MP which is defined by the families of random variables $\{(\tau_{nk}(n\overline{q}), \overline{\xi}_{nk}(n\overline{q}))\}$, $k \geq 0$. 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

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

Calculating moment characteristics of these variables and using Theorems 4.3, 4.4 on AP and DA, we obtain the statement of Theorem 5.8 for process $\widetilde{Q}_n(t)$.

Now we follow the same lines as in Theorem 5.1 and use the equivalence of trajectories of $\widetilde{Q}_n(t)$ and $\overline{Q}_n(t)$. Note that the RPSM defined above is equivalent to the value of queue only in the region $Q_n(t) \geq 0$. But the condition $\eta(t) > 0$, $t \in (0, A)$, together with relation (5.75) implies that in each interval $[\alpha, \beta]$, $0 < \alpha$, $\beta < A$,

$$\mathbf{P}\big\{Q_n(nt) > 0,\ \alpha \leq t \leq \beta\big\} \longrightarrow 1$$

and therefore the process $Q_n(t)$ is asymptotically equivalent to the value of queue. □

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

EXAMPLE 5.4. 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 $\overline{q} \geq \overline{0}$, $i = \overline{1,r}$, $j = \overline{1,r+1}$. Then our network is equivalent to a classical network $(M/M/\infty)^r$. In this case

$$\mathrm{d}\overline{s}(t) = \big(\overline{\lambda} + \big(P_0^* - I\big)A\overline{s}(t)\big)\mathrm{d}t, \tag{5.77}$$

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 matrix $P_0^* - I$ is invertible. Then, iterating equation (5.77), we obtain a representation $\overline{s}(t) = \overline{q}_* + \exp\{(P_0^* - I)At\}(\overline{s}_0 - \overline{q}_*)$, where $\overline{q}_* = A^{-1}(I - P_0^*)^{-1}\overline{\lambda}$ ($\overline{q}_*$ is the stationary point). Equation (5.76) has the form

$$\mathrm{d}\overline{\gamma}(t) = \big(P_0^* - I\big)A\overline{\gamma}(t)\mathrm{d}t + B\big(\overline{s}(t)\big)\mathrm{d}\overline{w}(t). \tag{5.78}$$

If $\overline{s}_0 = \overline{q}_*$, then for all $t > 0$, $\overline{s}(t) \equiv \overline{q}_*$, and we have a quasi-stationary regime with the stationary form for equation (5.78) where the matrix of diffusion $B(\cdot) = B(\overline{q}_*)$.

The general construction of our network provides us with the opportunity to consider networks with impatient calls and also unreliable servers. Some other examples of Markov state-dependent models are studied in [ANI 92b]. Markov models with state-dependent routing (overloaded and heavy traffic conditions) are considered in [BAS 89]. In the case, when calls arrive and are served one at a time without transformation, equations (5.74) and (5.76) are in agreement with the results [MAN 98b], where state-dependent networks $(M_\xi/M_\xi/1)^K$ in heavy traffic conditions are studied.

5.6. Non-Markov queueing networks

We now consider a fluid limit (AP) and DA for some classes of non-Markov models considered in section 2.3.2. The method of analysis involves several stages. First, we represent a queueing process as an equivalent SP by choosing in the appropriate way switching times and constructing corresponding processes in switching intervals. As in general we have a truncation by the level zero, these processes in most cases are not so simple. At the next stage we construct an auxiliary SP which is asymptotically equivalent to the queueing process, and the basic processes in switching intervals are constructed without truncation. At the last stage we prove AP and DA for the auxiliary SP using limit theorems for SP.

5.6.1. *A network* $(M_{SM,Q}/M_{SM,Q}/1/\infty)^r$ *with semi-Markov switching*

Consider a queueing network $(M_{SM,Q}/M_{SM,Q}/1/\infty)^r$ described in section 2.3.2.1. Suppose that characteristics of the network depend on parameter n in the following way. An SMP $x(t)$ and the variables introduced in the section do not depend on n. However, if at the time t, $x(t) = x$ and $\overline{Q}_n(t)/n = \overline{q}$, 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 the notations given in section 2.3.2.1. Denote as before by $t_1 < t_2 < \cdots$ the times of sequential jumps of $x(t)$. Suppose that the embedded MP $x_k = x(t_k)$, $k \geq 0$, 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 (size of queue) at node i at time t, and $\overline{Q}_{n0}$ be the initial value. Let us put $\overline{Q}_n(t) = (Q_n^{(1)}(t), \ldots, Q_n^{(r)}(t))$, $t \geq 0$, and for any $x \in X$, $i = \overline{1,r}$, $\overline{q} \in \mathcal{R}^r$, introduce the following functions:

$$m(x) = \mathbf{E}\tau(x), \quad P_0(x,\overline{q}) = \|p_{ij}(x,\overline{q})\|_{i,j=\overline{1,r}}, \quad \overline{a}(x,\overline{q}) = \mathbf{E}\overline{\eta}(x,\overline{q}),$$

$$g_i(x,\overline{q}) = \mathbf{E}\kappa_i(x,\overline{q}), \quad \overline{g}(x,\overline{q}) = \big(\mu_1(x,\overline{q})g_1(x,\overline{q}),\dots,\mu_r(x,\overline{q})g_r(x,\overline{q})\big),$$

$$m = \sum_{x\in X} m(x)\pi_x, \quad \overline{c}(x,\overline{q}) = \lambda(x,\overline{q})\overline{a}(x,\overline{q}) + \big(P_0(x,\overline{q})^* - I\big)\overline{g}(x,\overline{q}),$$

$$\overline{b}(\overline{q}) = \sum_{x\in X} m(x)\overline{c}(x,\overline{q})\pi_x, \quad d^2(x) = \mathbf{Var}\tau(x), \quad d_i^2(x,\overline{q}) = \mathbf{E}\kappa_i^2(x,\overline{q}),$$

$$J^2(x,\overline{q}) = \lambda(x,\overline{q})\mathbf{E}\eta(x,\overline{q})\eta(x,\overline{q})^*.$$

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

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

Denote

$$\begin{aligned} D^2(x,\overline{q}) &= d^2(x)\big(\overline{c}(x,\overline{q}) - m^{-1}\overline{b}(\overline{q})\big)\big(\overline{c}(x,\overline{q}) - m^{-1}\overline{b}(\overline{q})\big)^* \\ &\qquad + m(x)\big(F^2(x,\overline{q}) + J^2(x,\overline{q})\big), \\ D^2(\overline{q}) &= \sum_{x\in X} D^2(x,\overline{q})\pi_x, \end{aligned} \tag{5.79}$$

$$\overline{\gamma}(x,\overline{q}) = m(x)\big(\overline{c}(x,\overline{q}) - m^{-1}\overline{b}(\overline{q})\big). \tag{5.80}$$

Let matrix $B^2(\overline{q})$ be calculated using variables $\overline{\gamma}(x,\overline{q})$ with the help of MP x_k according to equations (4.71) and (4.72) in Theorem 4.6. 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 to the equation

$$\mathrm{d}\overline{s}(t) = m^{-1}\overline{b}\big(\overline{s}(t)\big)\mathrm{d}t, \quad \overline{s}(0) = \overline{s}_0. \tag{5.81}$$

THEOREM 5.9. *1) Assume that the 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 $\overline{q}\in \mathrm{int}\{\mathcal{R}_+^m\}$, and $\mathbf{E}\tau(x)^2<\infty$, $x\in X$. Also let $m>0$, for any bounded and closed domain $G\in \mathrm{int}\{\mathcal{R}_+^m\}$,*

$$\mathbf{E}\kappa_i(x,\overline{q})^2 \le C_G, \quad \mathbf{E}\big|\eta(x,\overline{q})\big|^2 \le C_G, \quad i=\overline{1,r},\ x\in X,\ \overline{q}\in G, \tag{5.82}$$

where $C_G<\infty$, 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

$$\sup_{0\le t\le T}\left|n^{-1}\overline{Q}_n(nt)-\overline{s}(t)\right| \xrightarrow{\ \mathrm{P}\ } 0. \tag{5.83}$$

where $\overline{s}(t)$ *is defined in equation (5.81).*

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

$$\mathbf{E}\kappa_i(x,\overline{q})^3\le C_G,\quad \mathbf{E}\left|\eta(x,\overline{q})\right|^3\le C_G,\quad i=\overline{1,r},\ x\in X,\ \overline{q}\in G. \tag{5.84}$$

Also let

$$n^{-1/2}\big(\overline{Q}_n(0)-n\overline{s}(0)\big)\stackrel{\mathrm{w}}{\Longrightarrow}\overline{\gamma}_0, \tag{5.85}$$

and function $H^2(\overline{q})$ *is continuous.*

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

$$\mathrm{d}\overline{\gamma}(t)=G\big(\overline{s}(t)\big)\overline{\gamma}(t)\mathrm{d}t+m^{-1/2}H\big(\overline{s}(t)\big)\mathrm{d}\overline{w}(t),\quad \overline{\gamma}(0)=\overline{\gamma}_0. \tag{5.86}$$

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 2.3.2.1, in the following way: in 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 $x(t_k)=x$ and $\overline{q}=\overline{Q}_n(t_k)/n$ at the initial point t_k. This means that, at given values $x(t_k)=x$, $\overline{Q}_n(t_k)/n=\overline{q}$, the parameters of the network in 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 the behavior of the normalized queue is equivalent to the behavior of the normalized queue in 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

$$\begin{aligned}\widetilde{\zeta}(t,x,\overline{q})=\ &\overline{\Pi}_{\lambda(x,\overline{q})}\big(t,\overline{\eta}(x,\overline{q})\big)\\ &+\sum_{i,j=1}^{r}\Pi_{\mu_i(x,\overline{q})p_{ij}(x,\overline{q})}\big(t,\kappa_i(x,\overline{q})\big)\big(\overline{e}_j-\overline{e}_i\big)\\ &-\sum_{i=1}^{r}\Pi_{\mu_i(x,\overline{q})p_{i,r+1}(x,\overline{q})}\big(t,\kappa_i(x,\overline{q})\big)\overline{e}_i,\quad t\ge 0,\end{aligned} \tag{5.87}$$

where all introduced Poisson processes are independent. Let us introduce a family of processes $\widetilde{\zeta}_{nk}(t, x, n\overline{q})$, $k > 0$ which are independent at different k, such that their distributions coincide with the distribution of $\widetilde{\zeta}(t, x, \overline{q})$. Denote by $\{(x(t), \widetilde{Q}_n(t)), t \geq 0\}$ 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 (1.14). By analogy to the proof of Theorem 5.1 we can define $\widetilde{Q}_n(t)$ in the whole space $\mathcal{R}^r$. First we prove AP for $\widetilde{Q}_n(t)$. Let us check the conditions of Theorem 4.5 and Corollary 4.4. In our notation the distribution of $\xi_n(x, n\overline{q})$ coincides with that of $\widetilde{\zeta}(\tau(x), x, \overline{q})$. Conditions (4.51) and (4.52) 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 $\mathbf{E}\Pi_a^2(\tau, \xi) \leq a\mathbf{E}\tau\mathbf{E}|\xi|^2 + a^2(\mathbf{E}\xi)^2\mathbf{E}\tau^2$. Using Chebyshev's inequality we obtain

$$\begin{aligned} n\mathbf{P}\left(n^{-1} \sup_{t \leq \tau} \left|\Pi_a(t, \xi)\right| > \varepsilon\right) &\leq n\mathbf{P}\big(\Pi_a\big(\tau, |\xi|\big) > n\varepsilon\big) \\ &\leq n(n\varepsilon)^{-2}\mathbf{E}\Pi_a^2\big(\tau, |\xi|\big) \longrightarrow 0, \end{aligned}$$

for any $\varepsilon > 0$ as $n \to \infty$. This implies condition (4.109). As is easy to calculate, $\mathbf{E}\widetilde{\zeta}(\tau(x), x, \overline{q}) = \overline{c}(x, \overline{q})m(x)$. Using Theorem 4.5 we find that $\widetilde{Q}_n(t)$ satisfies relation (5.83) with $\overline{s}(t)$ defined in relation (5.81). Now, following the same lines as in the proof of Theorem 5.1 we find that the multidimensional process generated by the queue in system $\widetilde{QN}$ also satisfies relation (5.83).

Now let us consider the initial network. First, we introduce independent families of multi-dimensional MP $\{\overline{\gamma}_{nk}(t, x, n\overline{q}),\ t \geq 0,\ x \in X,\ \overline{q} \in R_+^r\}$, $k \geq 0$, with values in R_+^r in the same way as was done in section 2.3.2.1. 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 equation (2.7). 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 (1.14). Again we can define this process in the whole space $\mathcal{R}^r$. Note that by construction the trajectory of $\widetilde{\widetilde{Q}}_n(t)$ coincides with the trajectory of 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 equation (5.83) with $\overline{s}(t)$ defined in equation (5.81). Again we need to check the conditions of Theorem 4.5 and Corollary 4.4. Note that $\xi_n(x, n\overline{q}) = \overline{\zeta}_{n1}(\tau(x), x, n\overline{q})$. Let us follow the same steps as in [ANI 96] [proof of Theorem 1 and Lemmas 1,2]. Using condition (5.82) we can prove that for any $\overline{q} \in \mathcal{R}^r$, $\mathbf{E}\xi_n(x, n\overline{q}) \to \mathbf{E}\widetilde{\zeta}_1(\tau(x), x, \overline{q})$ (see (5.87)) and check other conditions of Theorem 4.5. This implies equation (5.83) for $\widetilde{\widetilde{Q}}_n(t)$. Now, by analogy to the proof of Theorem 5.1, we get that the asymptotic behavior of the queueing process $\overline{Q}_n(t)$ and RPSM $\widetilde{\widetilde{Q}}_n(t)$ is the same, and the first part of Theorem 5.9 is proved.

To prove DA we use Theorem 4.6 and Corollary 4.5. In the same way we see that it is enough to calculate the characteristics of the auxiliary RPSM $\widetilde{\widetilde{Q}}_n(t)$ defined above. To find the function $D_n^2(x, \alpha)$ (see (4.66)) we can again use Theorem 1 and Lemmas 1, 2 in [ANI 96]. Using condition (5.84) we can prove that for any $\overline{q} \in \mathcal{R}^r$ as $n \to \infty$,

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

To calculate $D^2(x, \overline{q})$ we note that in notations of Theorem 5.9,

$$\rho_{n1}(\cdot) = \widetilde{\zeta}\big(\tau(x), x, \overline{q}\big) - m(x)\overline{c}(x, \overline{q}) - m^{-1}\overline{b}(\overline{q})\big(\tau(x) - m(x)\big)$$

(see equation (5.87)). Therefore, we can calculate that

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

and after some algebra we obtain the expression for $D^2(x, \overline{q})$ given in (5.79). All other conditions of Theorem 4.6 are also satisfied. □

Note that in the case when $\lambda(x, \overline{q}) \equiv 0$ and $p_{i,r+1}(x, \overline{q}) \equiv 0$ for all $i = \overline{1, r}$, $x \in X$, $\overline{q}$, this is a closed network.

Consider as an example a state-dependent system $M_{SM,Q}/M_{SM,Q}/1/\infty$ with semi-Markov switching. Let $x(t), t \geq 0$, be an SMP with state space $X = \{1, 2, \ldots, d\}$ and $\tau(x)$ be the sojourn time in state $x \in X$. Suppose that the embedded Markov chain is ergodic and denote by π_x, $x = 1, \ldots, r$, its stationary distribution. Let non-negative functions $\{\lambda(x, q), \mu(x, q),\ x \in X,\ q \geq 0\}$, also 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 $x(t) = x$, $Q(t)/n = q$, the arrival rate is $\lambda(x, q)$ and the service rate is $\mu(x, q)$. This means that, we have a semi-Markov arrival process and a semi-Markov service. After service completion a call leaves the system. Denote

$$m(x) = \mathbf{E}\tau(x), \quad m = \sum_{x \in X} m(x)\pi_x, \qquad b(q) = \sum_{x \in X} m(x)\big(\lambda(x, q) - \mu(x, q)\big)\pi_x,$$

$$d^2(x) = \mathbf{Var}\tau(x), \quad \gamma(x, q) = m(x)\big(\lambda(x, q) - \mu(x, q) - m^{-1}b(q)\big),$$

$$D^2(q) = \sum_{x \in X} \Big[d^2(x)\big(\lambda(x, q) - \mu(x, q) - m^{-1}b(q)\big)^2 + m(x)\big(\lambda(x, q) + \mu(x, q)\big)\Big]\pi_x.$$

Let function $B^2(q)$ be calculated by variables $\gamma(x,q)$ with the help of MP x_k according to relations (4.72). Denote $H^2(q) = D^2(q) + B^2(q)$ and let $s(t)$ be a solution to the equation

$$\mathrm{d}s(t) = m^{-1}b\big(s(t)\big)\mathrm{d}t, \quad s(0) = s_0. \tag{5.88}$$

COROLLARY 5.5. *Suppose that for any $x \in X$ 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 (5.61) holds with $s(t)$ defined in relation (5.88).*

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

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

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 $\lambda = m^{-1}\sum_{x\in X} m(x)\lambda(x)\pi_x$, $\mu = m^{-1}\sum_{x\in X} m(x)\mu(x)\pi_x$, then equation (5.88) has the form $\mathrm{d}s(t) = (\lambda - \mu s(t))\mathrm{d}t$, which coincides with equation (5.16) for the system $M/M/\infty$ (see section 5.2.2, applications of Theorem 5.2, case 2).

NOTE 5.6. In the same way it is possible to study systems and networks of the type $SM/M_{SM,Q}/1/\infty$ and $(SM/M_{SM,Q}/1/\infty)^r$, where the calls may arrive at the times of jumps of some SMP $x(t)$, and the instant service rate may depend on $x(t)$ and the current number of calls (or on the amount of work) in the system. Note that the diffusion approximation of the system $GI/M/1/\infty$ was considered in [WHI 82].

5.6.2. *State-dependent network with recurrent input*

Consider a network $(G_Q/M_Q/1/\infty)^r$ consisting of r nodes with one server in each node and an infinite number of waiting places. Let $Q_n(i,t)$ be a number of calls in the ith node at time t and $\bar{Q}_n(t) = (Q_n(i,t),\ i = \overline{1,r})$ be a column-vector. Functions $\mu_i(\bar{\alpha})$, $q_i(\bar{\alpha})$, $p_{ij}(\bar{\alpha})$, $i = \overline{1,r}$, $j = \overline{1,r+1}$, $\bar{\alpha} \in \mathcal{R}^r_+$, and the family on non-negative variables $\tau(\bar{\alpha})$ are given. Here $\sum_{j=1}^{r} q_j(\bar{\alpha}) = \sum_{j=1}^{r+1} p_{ij}(\bar{\alpha}) = 1$ for each $i = \overline{1,r}$, $\bar{\alpha} \in \mathcal{R}^r_+$. If a call enters the system at time t_{nk} and $\bar{Q}_n(t_{nk}+0) = \bar{Q}$, then this call with probability $q_j(\bar{Q}/n)$ enters the jth node, the next call will enter the system at the time

$$t_{nk+1} = t_{nk} + \tau(\bar{Q}/n)$$

and the service rate in the ith node in time interval (t_{nk}, t_{nk+1}) is $\mu_i(\bar{Q}/n)$, $i = \overline{1,r}$. In addition if a call in this interval completes its service in the ith node, then this call either with probability $p_{ij}(\bar{Q}/n)$ goes to the jth node, $j = \overline{1,r}$, or with probability $p_{i,r+1}(\bar{Q}/n)$ leaves the network.

Let $\bar{q}(\bar{\alpha}) = (q_1(\bar{\alpha}), \ldots, q_r(\bar{\alpha}))$, $\bar{\mu}(\bar{\alpha}) = (\mu_1(\bar{\alpha}), \ldots, \mu_r(\bar{\alpha}))$, be column-vectors,

$$P(\bar{\alpha}) = \left\| p_{ij}(\bar{\alpha}) \right\|_{i,j=\overline{1,r}}.$$

Denote $m(\bar{\alpha}) = \mathbf{E}\tau(\bar{\alpha})$, $\sigma^2(\bar{\alpha}) = \mathbf{Var}\tau(\bar{\alpha})$,

$$\bar{b}(\bar{\alpha}) = \bar{q}(\bar{\alpha}) + \big(P^*(\bar{\alpha}) - I\big)\bar{\mu}(\bar{\alpha}), \quad G(\bar{\alpha}) = \bar{b}'(\bar{\alpha}).$$

THEOREM 5.10. *Let functions $m(\bar{\alpha})$, $\bar{q}(\alpha)$, $\bar{\mu}(\bar{\alpha})$, $P(\bar{\alpha})$ satisfy a local Lipschitz condition, $n^{-1}\bar{Q}_n(0) \xrightarrow{\mathrm{P}} \bar{s}_0$, variables $\tau(\bar{\alpha})$ be uniformly integrable in each bounded region and for some $T > 0$ let there exist an interval $[0, A]$ such that the equation*

$$\mathrm{d}\bar{\eta}(t) = \bar{b}\big(\bar{\eta}(t)\big)\mathrm{d}t, \quad \bar{\eta}(0) = s_0$$

has a unique solution. Let in addition, $\bar{\eta}(t) > 0$ in each component, $t \in (0, A)$ and $\int_0^A m(\bar{\eta}(u))\mathrm{d}u > T$.

Thus, relation (5.83) holds with

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

If functions $\bar{q}(\bar{\alpha})$, $\bar{\mu}(\bar{\alpha})$, $P(\bar{\alpha})$ are continuously differentiable, variables $\tau(\bar{\alpha})^2$ are uniformly integrable in each bounded region, function $\sigma(\bar{\alpha})$ is continuous and equation (5.85) holds, then the sequence of processes

$$\bar{\gamma}_n(t) = n^{-1/2}\big(\bar{Q}_n(nt) - n\bar{s}(t)\big)$$

J-converges in interval $[0, T]$ to diffusion process $\bar{\gamma}(t)$ satisfying the following stochastic differential equation:

$$\mathrm{d}\bar{\gamma}(t) = G\big(\bar{s}(t)\big)\bar{\gamma}(t)\mathrm{d}t + m\big(\bar{s}(t)\big)^{-1/2}D\big(\bar{s}(t)\big)\mathrm{d}\bar{w}(t), \quad \bar{\gamma}(0) = \bar{\gamma}_0.$$

Here $D(\bar{\alpha})^2 = \|d_{ij}(\bar{\alpha})\|_{i,j=\overline{1,r}}$,

$$\begin{aligned} d_{ij}(\bar{\alpha}) &= -m(\bar{\alpha})\big(\mu_i(\bar{\alpha})p_{ij}(\bar{\alpha}) + \mu_j(\bar{\alpha})p_{ji}(\bar{\alpha})\big) \\ &\qquad + q_i(\bar{\alpha})q_j(\bar{\alpha})\big(\sigma^2(\bar{\alpha})m(\bar{\alpha})^{-2} - 1\big), \quad i \neq j, \\ d_{ii}(\bar{\alpha}) &= m(\bar{\alpha})\left(\sum_{k=1}^{r} \mu_k(\bar{\alpha})p_{ki}(\bar{\alpha}) + \mu_i(\bar{\alpha}) - 2\mu_i(\bar{\alpha})p_{ii}(\bar{\alpha})\right) + q_i(\bar{\alpha}) \\ &\qquad + q_i(\bar{\alpha})^2\big(\sigma(\bar{\alpha})^2 m(\bar{\alpha})^{-2} - 1\big), \quad i = \overline{1,r}. \end{aligned}$$

Proof. Let us represent process $Q_n(t)$ as an SP. The switching times t_{nk} are the times of arrivals of calls, variables $\tau_{nk}(n\bar{\alpha})$ are equivalent to $\tau(\bar{\alpha})$, and variables $\xi_{nk}(n\bar{\alpha})$ can be represented in the form

$$\bar{\xi}(\bar{\alpha}) = \delta(\bar{\alpha}) - \sum_{k=0}^{r} \pi_k + \sum_{j=1}^{r} \nu_j,$$

where $\delta(\bar{\alpha}) = e_i$ with probability $q_i(\bar{\alpha})$, $i = \overline{1,r}$, and $\pi_k = (\pi_{ik}(\tau(\bar{\alpha})),\ i = \overline{1,r})$, $\nu_j = (\pi_{ij}(\tau(\bar{\alpha})),\ i = \overline{1,r})$, where $\pi_{ij}(t)$ are the Poisson processes with parameters $\mu_i(\bar{\alpha})p_{ij}(\bar{\alpha})$ which are independent at different indexes and independent of τ.

It is easy to see that an RPSM constructed by variables $\{\tau(\bar{\alpha}), \xi(\bar{\alpha})\}$ is asymptotically equivalent to $\bar{Q}_n(t)$ in the region $\{\bar{Q}_n(t) \geq 0\}$. Calculating the moment functions according to Theorems 4.3, 4.4, we obtain that $\mathbf{E}\bar{\xi}(\bar{\alpha}) = \bar{b}(\bar{\alpha})$ and $\mathbf{E}\bar{\rho}(\bar{\alpha})\bar{\rho}(\bar{\alpha})^* = D(\bar{\alpha})^2$, where

$$\bar{\rho}(\bar{\alpha}) = \bar{\xi}(\bar{\alpha}) - \bar{b}(\bar{\alpha}) - \frac{\bar{b}(\bar{\alpha})}{m(\bar{\alpha})}\big(\tau(\bar{\alpha}) - m(\bar{\alpha})\big).$$

Our conditions imply the conditions of Theorems 4.3 and 4.4 and finally prove the statement of Theorem 5.10. □

The results of Theorem 5.10 are also valid when the service rates at each time t are the functions of the form $\mu_i(\bar{Q}_n(t)/n)$.

Note that when the variables $\tau(\bar{\alpha})$ have the exponential distributions, $\sigma(\bar{\alpha})^2 = m(\bar{\alpha})^2$ and we obtain the result of Theorem 5.7 for Markov networks $(M_Q/M_Q/1/\infty)^r$.

Using Theorems 4.5, 4.6, these results can be extended to the state-dependent networks with semi-Markov input of the type $(SM/M_Q/1/\infty)^r$ and also to the models with batch entry and service. The general form of representation of our network also makes it possible to consider networks with impatient calls and unreliable servers.

These results show that a technique based on limit theorems of AP and DA types for SPs provides us with a new analytic approach in the approximate modeling of transient and stable regimes for rather complex queueing models under heavy traffic conditions. Instead of a direct simulation, we can use a simpler approximate analytic relation

$$Q_n(nt) \approx ns(t) + \sqrt{n}\zeta(t),$$

where the function $\bar{s}(t)$ and the process $\zeta(t)$ satisfy the differential and stochastic differential equations with coefficients that are calculated using the first and second order moment functions of the increments of original processes in switching intervals. This approach also provides us with the possibility to estimate various reliability characteristics and cost functionals of the system.

5.7. Bibliography

[ANI 77] ANISIMOV V., "Switching processes", *Cybernetics*, vol. 13, no. 4, p. 590–595, 1977.

[ANI 78] ANISIMOV V., "Limit theorems for switching processes and their applications", *Cybernetics*, vol. 14, no. 6, p. 917–929, 1978.

[ANI 90a] ANISIMOV V. and ALIEV A., "Limit theorems for recurrent processes of semi-Markov type", *Theor. Prob. and Math. Stat.*, vol. 41, p. 7–13, 1990.

[ANI 90b] ANISIMOV V. and CHABANYUK V., "On applying of Skorokhod reflecting problem at diffusion approximation of queueing networks", *Soviet-Phys. Dokl.*, vol. 35, no. 6, p. 505–506, 1990.

[ANI 91] ANISIMOV V. and ATADZHANOV H., "Diffusion approximation of systems with repeated calls", *Theor. Prob. and Math. Stat.*, vol. 44, p. 3–8, 1991.

[ANI 92a] ANISIMOV V., "Averaging principle for switching processes", *Theor. Probab. and Math. Stat.*, vol. 46, p. 1–10, 1992.

[ANI 92b] ANISIMOV V. and LEBEDEV E., *Stochastic Queueing Networks. Markov Models*, Kiev University (Russian), Kiev, Ukraine, 1992.

[ANI 93] ANISIMOV V., "Averaging principle for the processes with fast switching", *Random Oper. and Stoch. Eqv.*, vol. 1, no. 2, p. 151–160, 1993.

[ANI 94a] ANISIMOV V., "Limit theorems for processes with semi-Markov switching and their applications", *Random Oper. and Stoch. Eqv.*, vol. 2, no. 4, p. 333–352, 1994.

[ANI 94b] ANISIMOV V. and ATADZHANOV H., "Diffusion approximation of systems with repeated calls and unreliable server", *J. of Math. Sci.*, vol. 72, no. 2, p. 3032–3034, 1994.

[ANI 95] ANISIMOV V., "Switching processes: averaging principle, diffusion approximation and applications", *Acta Applicandae Mathematicae*, vol. 40, p. 95–141, 1995.

[ANI 96] ANISIMOV V., "Averaging principle for near-critical branching processes with semi-Markov switching", *Theor. Probab. and Math. Stat.*, vol. 52, p. 13–26, 1996.

[ANI 97] ANISIMOV V., "Asymptotic analysis of switching queueing systems in conditions of low and heavy loading", in CHAKRAVARTHY S. and ALFA A., Eds., *Matrix-Analytic Methods in Stochastic Models*, vol. 183 of *Lecture Notes in Pure and Appl. Math.*, p. 241–260, Dekker, New York, 1997.

[ANI 99a] ANISIMOV V., "Averaging methods for transient regimes in overloading retrial queuing systems", *Mathematical and Computing Modelling*, vol. 30, no. 3/4, p. 65–78, 1999.

[ANI 99b] ANISIMOV V., "Diffusion approximation for processes with semi-Markov switches and applications in queueing models", in JANSSEN J. and LIMNIOS N., Eds., *Semi-Markov Models and Applications*, p. 77–101, Kluwer Acad. Publ., Dordrecht, 1999.

[ANI 99c] ANISIMOV V., "Switching stochastic models and applications in retrial queues", *Top*, vol. 7, no. 2, p. 169–186, 1999.

[ANI 00] ANISIMOV V., "Asymptotic analysis of reliability for switching systems in light and heavy traffic conditions", in LIMNIOS N. and NIKULIN M., Eds., *Recent Advances in Reliability Theory: Methodology, Practice, and Inference*, p. 119–133, Birkhäuser Boston, Massachusetts, 2000.

[ANI 01] ANISIMOV V. and ARTALEJO J., "Analysis of Markov multiserver retrial queues with negative arrivals", *Queueing Systems*, vol. 39, no. 2/3, p. 157–182, 2001.

[ANI 02] ANISIMOV V., "Diffusion approximation in overloaded switching queueing models", *Queueing Systems*, vol. 40, no. 2, p. 141–180, 2002.

[ART 96] ARTALEJO J. and FALIN G., "On the orbit characteristics of the M/G/1 retrial queue", *Naval Research Logistics*, vol. 43, p. 1147–1161, 1996.

[ART 99] ARTALEJO J., "A classified bibliography of research on retrial queues: progress in 1990–1999", *Top*, vol. 7, no. 2, p. 187–211, 1999.

[BAS 89] BASHARIN G., BOCHAROV P. and KOGAN J., *Analysis of Queues in Computing Networks (Russian)*, Nauka, Moscow, 1989.

[BIL 68] BILLINGSLEY P., *Convergence of Probability Measures*, Wiley, New York, 1968.

[BRA 98] BRAMSON M., "State space collapse with applications to heavy traffic limits for multiclass queueing networks", *Queueing Systems*, vol. 30, no. 1,2, p. 89–148, 1998.

[BRA 01] BRAMSON M. and DAI J., "Heavy traffic limits for some queueing networks", *Ann. Appl. Prob.*, vol. 11, p. 49–90, 2001.

[CHE 00] CHEN H. and ZHANG H., "Diffusion approximations for some multiclass queueing networks under FIFO disciplines", *Math. Oper. Res.*, vol. 25, no. 4, p. 679–707, 2000.

[DAI 95] DAI J. and KURTZ T., "A multiclass station with Markovian feedback in heavy traffic", *Math. Oper. Res.*, vol. 20, p. 721–742, 1995.

[DAI 99] DAI J. and DAI W., "A heavy traffic limit theorems for a class of open queueing networks with finite buffers", *Queueing Systems*, vol. 32, p. 5–40, 1999.

[ETH 86] ETHIER S. and KURTZ T., *Markov Processes, Characterization and Convergence*, John Wiley & Sons, New York, 1986.

[FAL 90] FALIN G., "A survey of retrial queues", *Queueing Systems*, vol. 7, p. 127–168, 1990.

[FAL 95] FALIN G. and ARTALEJO J., "Approximation for multiserver queues with balking/retrial discipline", *OR Spektrum*, vol. 17, p. 239–244, 1995.

[FAL 97] FALIN G. and TEMPLETON J., *Retrial Queues*, Chapman & Hall, London, 1997.

[GIK 72] GIKHMAN I. and SKOROKHOD A., *Stochastic Differential Equations and their Applications*, Springer-Verlag, New York, 1972.

[HAR 95] HARRISON J., "Balanced fluid models of multiclass queueing networks: a heavy traffic conjecture", in KELLY F. and WILLIAMS R., Eds., *Stochastic Networks*, vol. 71 of *IMA Vol. Math. Appl.*, p. 1–20, Springer, New York, 1995.

[HAR 96] HARRISON J. and WILLIAMS R., "A multiclass closed queueing network with unconventional heavy traffic behavior", *Ann. Appl. Prob.*, vol. 6, no. 1, p. 1–47, 1996.

[IGL 65] IGLEHART D., "Limit diffusion approximation for the many server queue and the repairman problem", *J. Appl. Prob.*, vol. 2, p. 429–441, 1965.

[KRI 88] KRICHAGINA E., LIPTSER R. and PUHALSKY A., "Diffusion approximation for the system with arrival process depending on queue and arbitrary service distribution", *Theory Prob. and Appl.*, vol. 33, p. 124–135, 1988.

[KUL 97] KULKARNI V. and LIANG H., "Retrial queues revisited", in DSHALALOW J., Ed., *Frontiers in Queueing. Models and Applications in Science and Engineering*, p. 19–34, CRC Press, Florida, 1997.

[LIP 89] LIPTSER R. and SHIRYAEV A., *Theory of Martingales*, Kluwer, Dordrecht, 1989.

[MAN 95] MANDELBAUM A. and MASEY W., "Strong approximation for time-dependent queues", *Math. Oper. Res.*, vol. 20, no. 1, p. 33–64, 1995.

[MAN 98a] MANDELBAUM A., MASEY W. and REIMAN M., "Strong approximation for Markovian service networks", *Queueing Systems*, vol. 30, no. 1,2, p. 149–202, 1998.

[MAN 98b] MANDELBAUM A. and PATS G., "State-dependent stochastic networks. Part I: Approximations and applications with continuous diffusion limits", *Ann. Appl. Prob.*, vol. 8, no. 2, p. 569–646, 1998.

[MAR 95] MARTIN M. and ARTALEJO J., "Analysis of an M/G/1 queue with two types of impatient units", *Advances in Applied Probability*, vol. 27, p. 840–861, 1995.

[REI 84] REIMAN M., "Open queueing networks in heavy traffic", *Math. Oper. Res.*, vol. 9, no. 3, p. 441–458, 1984.

[REI 88] REIMAN M., "A multiclass feedback queue in heavy traffic", *Adv. in Appl. Prob.*, vol. 20, p. 179–207, 1988.

[SKO 56] SKOROKHOD A., "Limit theorems for random processes", *Theory Prob. Appl.*, vol. 1, p. 289–319, 1956.

[SKO 62] SKOROKHOD A., "Stochastic equations for diffusion processes with boundaries. I, II", *Theory Prob. Appl.*, vol. 6, 7, p. 287–298, 5–25, 1961, 1962.

[WHI 82] WHITT W., "On the heavy-traffic limit theorem for $GI/G/\infty$ queues", *Adv. in Appl. Prob.*, vol. 14, p. 171–190, 1982.

[WIL 96] WILLIAMS R., "On the approximation of queueing networks in heavy traffic", in KELLY F., ZACHARY S. and ZIEDINS I., Eds., *Stochastic Networks. Theory and Applications*, p. 35–56, Oxford University Press, Oxford, 1996.

[WIL 98] WILLIAMS R., "Diffusion approximation for open multiclass queueing networks: sufficient conditions involving state space collapse", *Queueing Systems*, vol. 30, no. 1-2, p. 27–88, 1998.

[YAN 87] YANG T. and TEMPLETON J., "A survey on retrial queues", *Queueing Systems*, vol. 2, p. 203–233, 1987.

Chapter 6

Systems in Low Traffic Conditions

6.1. Introduction

Real life mathematical models of computing systems, telephone and communication networks usually have a complex hierarchical structure and operate at different scales of time. Even for Markov models, exact analytic solutions can be obtained only for special rare cases. Therefore, asymptotic methods and approximating techniques play a very important role in investigation and modeling.

In many models of practical interest, usually "small parameters" are present, e.g., the rate of incoming calls in a system is much smaller than the rate of service (in queueing theory it is called "low loading" or "fast service"). These small parameters give rise to the analysis of so-called flows of rare events in reliability and queueing theory. In applications rare events usually mean different types of failures, exit times from a particular region, losses of calls, exceeding some level, etc. Any reader interested in this can find a survey of results devoted to the analysis of rare events in queueing systems in [KOV 94].

When studying various classes of computer and telecommunication systems we often encounter cases when there are different scales of time corresponding to different processes in the system. That means the corresponding stochastic process may have transition rates or transition probabilities of different orders. To analyze these classes of models, a novel approach is developed which is based on using the so-called S-sets (asymptotically connected set) introduced by the author in [ANI 70, ANI 74]. The method of S-sets allows us to study the asymptotic behavior of the first exit time from a subset of states of Markov and semi-Markov models. In queueing applications this means that we can study the asymptotic behavior of the time of the first loss of a call for Markov and semi-Markov type queueing models with a finite number of states and in the case of "fast" service or "low" loading. Various applications of this method can be found in Anisimov *et al.* [ANI 87], Anisimov and Sztrik [ANI 89b, ANI 89a, ANI 89c], Sztrik and Kouvatsos [SZT 91], Anisimov [ANI 97, ANI 00a], Anisimov and Kurtulus [ANI 01] and Sztrik [SZT 92].

This chapter is devoted to the asymptotic analysis of the special type rare events for Markov and semi-Markov processes with finite state space – first exit time from a subset of states, and applications to the analysis of the first loss of a call in queueing models with transition rates of a different order. First, we study the asymptotic behavior of the first exit time from the fixed subset of states of a Markov or semi-Markov process. Then we investigate the asymptotic behavior of the time of the first loss of a call for some types of Markov and semi-Markov queueing models and for several classes of Markov multiserver retrial queueing models. Note that some results devoted to the asymptotic analysis and Poisson approximation of flows of rare events on trajectories of the processes with discrete components are also obtained in [ANI 70, ANI 74].

The basic assumptions are that the rate of service is large ("fast" service of "low" traffic), or the retrial rate is large. We assume that the characteristics of the system depend on a particular scaling parameter n and we analyze the system as $n \to \infty$. The method of S-sets and the properties of a special type of S-set called a "monotone structure" are used to prove the exponential approximation of the time of a first loss of a call under the appropriate scaling and the Poisson approximation of the flow of lost calls.

6.2. Analysis of the first exit time from the subset of states

The presentation of the following sections mainly follows the author's papers [ANI 70, ANI 74, ANI 97]. An important notion of an S-set (asymptotically connected set) is introduced and an exponential approximation for the first exit time from the S-set is considered. A special class of hierarchical type S-sets, a *monotone structure*, is studied. The results of this part provide the reader with the analytical technique for the analysis and simulation of reliability characteristics of hierarchical Markov and semi-Markov models, and, in particular, queueing models.

6.2.1. *Definition of S-set*

Let x_{nk}, $k \geq 0$, for each $n = 1, 2, \ldots$ be an MP with finite state space $X = \{0, 1, 2, \ldots, r\}$ given by a matrix of one-step transition probabilities $P_n = \|p_n(i,j)\|_{i,j=\overline{0,r}}$. Let X_0 be a particular fixed subset of X. Without loss of generality we can take $X_0 = \{1, 2, \ldots, r\}$. Denote by

$$\nu_n(i) = \min\big\{k : k > 0,\ x_{nk} \notin X_0\big\}, \quad i \in X_0, \tag{6.1}$$

the first exit time from X_0 starting from state $i \in X_0$.

DEFINITION 6.1. *The subset X_0 is called an S-set if for any $i, j \in X_0$, as $n \to \infty$,*

$$\mathbf{P}\big\{\textit{there exists } k,\ k < \nu_n(i) \textit{ such that } x_{nk} = j \mid x_{n0} = i\big\} \longrightarrow 1.$$

This means that as $n \to \infty$ the probabilities of exit from the states of a subset X_0 tend to zero and the probabilities of transitions between the states in X_0 are changing in such a way that starting from any state the process will visit all states in X_0 before exit with probability tending to one.

6.2.2. *An asymptotic behavior of the first exit time*

Now consider the first exit time from the subset of states of an SMP. Let $x_n(t)$ be an SMP with finite state space $X = \{0, 1, \dots, r\}$ given by the embedded MP x_{nk} and by the family of sojourn times $\{\tau_n(i),\ i \in X\}$. Let $X_0 = \{1, 2, \dots, r\}$. Given that $x_n(0) = i \in X_0$, denote by

$$\Omega_n(i) = \inf\big\{t : t > 0,\ x_n(t) \notin X_0\big\} \tag{6.2}$$

the first exit time from the subset X_0 starting from state $i \in X_0$.

Consider the limit behavior of the variables $\nu_n(i)$ and $\Omega_n(i)$. Let us construct an auxiliary MP $\widetilde{x}_{nk}$ with state space X_0 and matrix of transition probabilities $\widetilde{P}_n(X_0) = \|\widetilde{p}_n(i,j)\|$, $i, j \in X_0$, where

$$\widetilde{p}_n(i,j) = p_n(i,j)p_n\big(i, X_0\big)^{-1}, \quad i, j \in X_0,$$

$$p_n\big(i, X_0\big) = \sum_{l \in X_0} p_n(i,l).$$

Suppose that the subset X_0 forms an S-set. Denote by $\widetilde{\pi}_n(i)$, $i \in X_0$, a stationary distribution for MP $\widetilde{x}_{nk}$ (which exists at least at large enough n) and define

$$g_n\big(X_0\big) = \sum_{i \in X_0} \widetilde{\pi}_n(i)\big(1 - p_n\big(i, X_0\big)\big), \tag{6.3}$$

where $g_n(X_0)$ means the stationary (or aggregated) probability of exit from X_0.

THEOREM 6.1. *Let the set X_0 form an S-set and there exist a normalizing factor β_n and functions $a_i(\theta)$ ($a_i(\pm 0) = 0$) such that as $n \to \infty$,*

$$g_n\big(X_0\big)^{-1}\big(1 - \mathbf{E}\exp\big\{-\beta_n\theta\tau_n(i)\big\}\big) \longrightarrow a_i(\theta), \quad i \in X_0. \tag{6.4}$$

Then for any initial state $i_0 \in X_0$,

$$\lim_{n\to\infty} \mathbf{E}\exp\big\{-\beta_n\theta\Omega_n\big(i_0\big)\big\} = \big(1 + A(\theta)\big)^{-1},$$

where

$$A(\theta) = \lim_{n\to\infty} \sum_{i\in X_0} \widetilde{\pi}_n(i)a_i(\theta). \quad (6.5)$$

The proof can be found in [ANI 70, ANI 74], Anisimov *et al.* [ANI 87]. It is based on the asymptotic analysis of the matrix equation for the characteristic function of the normalized vector $\{\beta_n\Omega_n(i),\ i \in X_0\}$, the recurrent aggregation technique, and uses the representation

$$\big(I - \widetilde{P}_n(X_0)\big)^{-1} = g_n(X_0)^{-1}\widetilde{\Pi}_n(X_0),$$

where I is the unit matrix, and $\widetilde{\Pi}_n(X_0) = \|\widetilde{\pi}_n(i)(1 + o_{ij}(1))\|$, $i, j \in X_0$.

An algorithm on how to check whether a particular subset forms an S-set or not is also given in these papers.

Note that the asymptotic distribution of the exit time from the S-set does not depend on the initial state. This fact allows us to consider an asymptotic aggregation of the state space of S-sets and will be used in the following chapters.

Consider the exit time $\nu_n(i)$ from a subset of an MP. In this case we can take $\tau_n(i) \equiv 1$ and, as a consequence, the following result is true:

COROLLARY 6.1. *If the set X_0 forms an S-set, then for any $i_0 \in X_0$,*

$$\lim_{n\to\infty} \mathbf{P}\{g_n(X_0)\nu_n(i_0) > t\} = \exp\{-t\}, \quad t > 0. \quad (6.6)$$

This means that the exponential approximation of the first exit time from any subset forming an S-set is valid with the natural rate which is the stationary probability of exit. In [ANI 88a] the proximity estimates of the rate of convergence in (6.6) are also given.

Now consider the exponential approximation of the exit time $\Omega_n(i_0)$ from a subset of an SMP.

COROLLARY 6.2. *Suppose that the expectations of sojourn times $m_n(i) = \mathbf{E}\tau_n(i)$ exist, $m_n(i) \to m_i$, $i \in X_0$, and for any $i \in X_0$, $\mathbf{E}\tau_n(i)^2 < C < \infty$. Denote*

$$M = \lim_{n\to\infty} \sum_{i\in X_0} \widetilde{\pi}_n(i)m_n(i).$$

If the set X_0 forms an S-set, then for any $i_0 \in X_0$ the variable $g_n(X_0)\Omega_n(i_0)$ weakly converges to the exponential random variable with parameter M^{-1}.

The proof follows from relations (6.4), (6.5) where in this case $a_i(\theta) = m_i\theta$, $i \in X_0$.

Now consider as important for queueing applications the case when $x_n(t)$, $t \geq 0$, is a continuous time MP.

COROLLARY 6.3. *Suppose that the process $x_n(t)$, $t \geq 0$, is a continuous time MP given by the embedded MP with matrix of transition probabilities P_n and by exit rates $\lambda_n(i)$, $i = \overline{1, r}$, the set X_0 forms an S-set and, as $n \to \infty$,*

$$\min_i \lambda_n(i) \not\to 0, \quad \sum_{i \in X_0} \widetilde{\pi}_n(i)/\lambda_n(i) \not\to 0.$$

Then the distribution of $\beta_n \Omega_n(i_0)$ weakly converges to the exponential distribution with parameter 1 *where*

$$\beta_n = g_n\big(X_0\big)\left(\sum_{i \in X_0} \widetilde{\pi}_n(i)/\lambda_n(i)\right)^{-1}.$$

Note that in this case β_n is asymptotically equivalent to the expression

$$\beta_n \approx \widetilde{\Lambda}_n\big(X_0\big) = \sum_{i \in X_0} \widetilde{\rho}_n(i) \sum_{k \notin X_0} \lambda_n(i, k),$$

where $\lambda_n(i, j)$, $i, j \in X_0$, $i \neq j$, are the transition rates of $x_n(\cdot)$ and $\widetilde{\rho}_n(i)$, $i \in X_0$, is the stationary distribution of the auxiliary continuous time MP with state space X_0 and transition rates $\lambda_n(i, j)$, $i, j \in X_0$, $i \neq j$.

Note that the value $\widetilde{\Lambda}_n(X_0)$ represents the stationary (aggregated) rate of exit from X_0.

These results show that for finding a normalizing coefficient and the parameter in exponential approximation of exit time from the subset we need to estimate the main order of stationary probabilities $\widetilde{\pi}_n(i)$, $i \in X_0$, for the auxiliary MP $\widetilde{x}_{nk}$. In practical applications it can be a separate, rather complicated problem.

Suppose now that after leaving the subset X_0 the system returns again to X_0 after a random time. Denote by $Y_n(t)$ the number of exits (which corresponds to the number of lost calls) in the interval $[0, t]$. Using the asymptotic exponentiality of exit time the following statement can be proved:

STATEMENT 6.1. *If the return time to X_0 is stochastically bounded uniformly in n, then for any initial state $i \in X_0$ the process $Y_n(\beta_n^{-1}t)$ J-converges to an ordinary Poisson process. The parameter of this process is the same as the parameter of the exponential distribution in the approximation of exit time.*

6.2.3. *State space forming a monotone structure*

Now we consider an important class of MPs where the state space forms a special monotone structure introduced in [ANI 87] (see also [ANI 97, ANI 00a]). In this case it is possible to derive the explicit formulae for the main order terms of the stationary probabilities of the S-set. Models of this type usually appear at the asymptotic analysis of wide classes of queueing models and reliability models with transition rates of different orders (fast service, low input, etc.).

Let x_{nk}, $k \geq 0$, be an MP with finite state space X. Suppose that X can be represented in the form $X = \cup_{s=0}^{m+1}(X_s, s)$, where X_s, $s = 0, 1, \ldots, m+1$, are some subsets. The individual states can be represented in the form $\{(l, q)\}$. Consider a subset of states $Z = \{(i, s),\ i \in X_s,\ s = \overline{0, m}\}$. Denote by $p_n((i, s), (j, q))$, the one-step transition probabilities.

DEFINITION 6.2. *The subset* $Z = \{(i, s),\ i \in X_s,\ s = \overline{0, m}\}$ *is called a* monotone structure *of the order* m *if the following asymptotic relations hold:*

1. $p_n((i, s), (j, s+1)) = \varepsilon_n(s) a_{ij}(s)(1 + o(1))$, $i \in X_s$, $j \in X_{s+1}$, $s = \overline{0, m}$, *where* $\varepsilon_n(s) \to 0$ *for any* s*;*

2. $p_n((i, s), (j, s+k)) = 0$, $i \in X_s$, $j \in X_{s+k}$, $s = \overline{0, m-2}$, $k > 1$*;*

3. $p_n((i, s), (j, s)) = p_{ij}(s)(1 + o(1))$, $i, j \in X_s$, $s = \overline{0, m}$,

where for each $s = \overline{1, m}$ *the matrix* $I - P(s)$ *is invertible, and* $P(0)$ *is an irreducible matrix with stationary distribution* π_i, $i \in X_0$ *(here* $P(s) = \|p_{ij}(s)\|$, $i, j \in X_s$*),* I *is the unit matrix and* $o(1) \to 0$ *as* $n \to \infty$.

Figure 6.1 provides an illustration of the state space of the monotone structure.

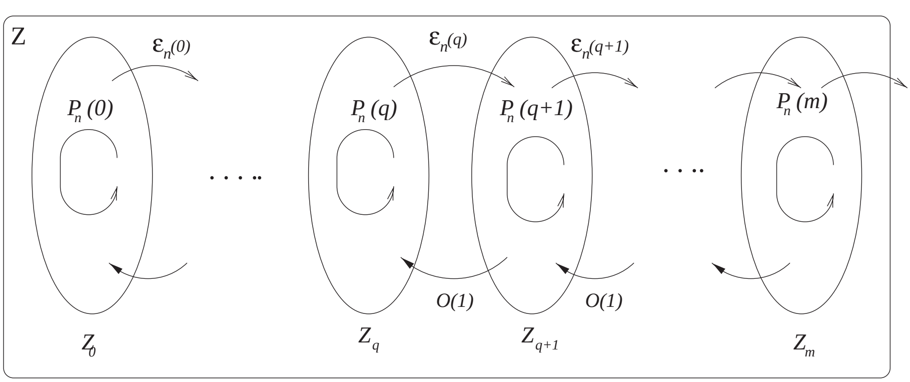

Figure 6.1. *Monotone structure*

Let us call the subset of states $Z_q = \{(i,q),\ i \in X_q\}$ a q-level. Denote

$$p_n\big((i,s), Z\big) = \sum_{(l,g)\in Z} p_n\big((i,s),(l,g)\big),$$

and

$$b_i = \sum_{k\in X_{m+1}} a_{ik}(m).$$

Denote also by $\overline{\pi}_n(s) = (\pi_n(i,s),\ i \in X_s)$, $s = \overline{0,m}$, $\overline{\pi} = (\pi_i,\ i \in X_0)$ and $\overline{b} = (b_i,\ i \in X_m)$ the row-vectors, where $\pi_n(i,s)$ is a stationary probability of the state (i,s) for the MP with state space Z and matrix of transition probabilities

$$\widetilde{P}_n(Z) = \Big\| p_n\big((i,s),(j,q)\big) p_n\big((i,s),Z\big)^{-1} \Big\|, \quad (i,s),(j,q) \in Z.$$

THEOREM 6.2. *If the state space $Z = \{(i,s),\ i \in X_s,\ s = \overline{0,m}\}$ forms a monotone structure, then it also forms an S-set and for any $q = \overline{1,m}$ the following representations hold:*

$$\overline{\pi}_n(q) = \overline{\pi}\left(\prod_{j=0}^{q-1} A(j)\big(I - P(j+1)\big)^{-1} \varepsilon_n(j)\right)\big(1 + o(1)\big), \tag{6.7}$$

and

$$g_n(Z) = \overline{\pi}\left(\prod_{j=0}^{m-1} A(j)\big(I - P(j+1)\big)^{-1} \varepsilon_n(j)\right)\varepsilon_n(m)\overline{b}^*\big(1 + o(1)\big), \tag{6.8}$$

where $A(s) = \|a_{ij}(s)\|$, $i,j \in X_s$, $\overline{b}^*$ is the transposed vector to $\overline{b}$, $\prod_{j=k}^{s} C(j) = C(k)C(k+1)\cdots C(s)$ as $k \le s$, and $o(1) \to 0$.

The proof is provided recursively to the order of the monotone structure. The main problem is in the approximation of the stationary probabilities. First, it can be shown that as $n \to \infty$,

$$\pi_n(i,q) = O\left(\prod_{s=0}^{q-1} \varepsilon_n(s)\right), \quad i = \overline{1,r},\ q > 0.$$

Then from the matrix equation for stationary probabilities

$$\overline{\pi}_n(q) = \overline{\pi}_n(q)P_n(q) + \overline{\pi}_n(q-1)\varepsilon_n(q-1)A(q-1) + O\left(\prod_{s=0}^{q} \varepsilon_n(s)\right), \tag{6.9}$$

where $P_n(q) = \|p_n((i,q),(j,q))\|$, $i,j \in X_q$, we obtain

$$\overline{\pi}_n(q) = \pi_n(q-1)A(q-1)\big(I - P_n(q)\big)^{-1}\varepsilon_n(q-1)\big(1 + o(1)\big),$$

and this implies (6.7). The expression for $g_n(Z)$ follows from (6.3).

6.2.4. *Exit time as the time of first jump of the process of sums with Markov switching*

Now let us consider an equivalent representation of the first exit time in terms of the first time of occurrence of the Bernoulli event on the trajectory of some auxiliary MP, which is useful for future exposition. We introduce an auxiliary MP $\widetilde{x}_{nk}$ with state space X_0 and matrix of transition probabilities $\widetilde{P}_n(X_0) = \|\widetilde{p}_n(i,j)\|$, $i,j \in X_0$, where

$$\widetilde{p}_n(i,j) = p_n(i,j)p_n\big(i,X_0\big)^{-1}, \quad i,j \in X_0, \tag{6.10}$$

and $p_n(i,X_0)$ is defined as $p_n(i,X_0) = \sum_{l\in X_0} p_n(i,l)$.

Let us define a family of jointly independent random indicators $\{\chi_n(i,k),\ i \in X_0,\ k \geq 0\}$ such that $\mathbf{P}(\chi_n(i,k) = 1) = 1 - \mathbf{P}(\chi_n(i,k) = 0) = 1 - p_n(i,X_0)$, and construct a process of Bernoulli variables on the trajectory of $\widetilde{x}_{nk}$ in the form:

$$y_n(m) = \sum_{k=0}^{m} \chi_n\big(\widetilde{x}_{nk},k\big), \quad m \geq 0. \tag{6.11}$$

Let $\widetilde{\nu}_n(i)$ be the time of the first jump of $y_n(m)$ given that $\widetilde{x}_{n0} = i \in X_0$:

$$\widetilde{\nu}_n(i) = \min\big\{m : m > 0,\ y_n(m-1) = 1\big\} \tag{6.12}$$

Comparing the distributions $\nu_n(i)$ and $\widetilde{\nu}_n(i)$, we can straightforwardly prove:

STATEMENT 6.2. *For any $i \in X_0$, the distributions of the variables $\nu_n(i)$ and $\widetilde{\nu}_n(i)$ coincide.*

A similar result is valid for SMP. Let us define an accumulating process $\eta_n(m)$ on the trajectory of the auxiliary embedded MP $\widetilde{x}_{nk}$ given by transition probabilities $\tilde{p}_n(i,j), i,j \in X_0$ (see (6.10)) using the sojourn times of the SMP $z_n(t)$. For this purpose we introduce a family of jointly independent random variables $\{\tau_n(i,k),\ i \in X_0,\ k \geq 0\}$ such that for any i,k the distribution of $\tau_n(i,k)$ coincides with the distribution of a sojourn time $\tau_n(i)$. Let

$$\eta_n(m) = \sum_{k=0}^{m} \tau_n\big(\widetilde{x}_{nk},k\big), \quad m \geq 0, \tag{6.13}$$

and let $\widetilde{\Omega}_n(i)$ be the value of $y_n(\cdot)$ stopped at the time $\widetilde{\nu}_n(i)$ (see (6.12)) of the first jump of $y_n(m)$ given that $\widetilde{x}_{n0} = i \in X_0$:

$$\widetilde{\Omega}_n(i) = \eta_n\big(\widetilde{\nu}_n(i)\big).$$

Comparing the distributions of the variables $\Omega_n(i)$ in (6.2) and $\widetilde{\Omega}_n(i)$, we can straightforwardly prove the following statement (see [ANI 87]).

STATEMENT 6.3. *For any $i \in X_0$, the distributions of the variables $\Omega_n(i)$ and $\widetilde{\Omega}_n(i)$ coincide.*

These results mean that the problems of asymptotic analysis of the variables $\Omega_n(j)$ and $\tilde{\Omega}_n(j)$ are equivalent and open the possibility of using limit theorems for accumulative type processes with Markov switching for studying the asymptotic behavior of exit time. For example, the asymptotic analysis of flows of rare events constructed on the trajectories of stochastic systems (or switched by some random environment) is given in Chapter 3 (see also [ANI 83, ANI 88b, ANI 00a]). In the case when the switching system satisfies an asymptotically mixing condition, the Poisson approximation is proved. This implies the exponential approximation for the time of the first jump.

These results allow us to study the asymptotic behavior of the time of first loss of a call for wide classes of queueing systems and networks with a finite number of states and fast service or low loading (see [ANI 87, ANI 89b, ANI 89a, ANI 89c, ANI 00a, ANI 97]). Note that the asymptotic behavior of first exit time from a fixed subset of states for Markov and semi-Markov processes was studied independently by the author [ANI 70, ANI 74, ANI 87] and by Korolyuk and his pupils [KOR 69, KOR 93].

In addition we note that, as follows from Theorem 6.1, the asymptotic behavior of the exit time from the S-set does not depend on the initial state. This provides us with the possibility of studying models of the asymptotic aggregation of state space in Chapter 8 (see also [ANI 73, ANI 87]).

6.3. Markov queueing systems with fast service

In this section we consider several examples of queueing systems in the conditions of "fast" service (or "low" traffic) illustrating the use of the S-sets technique. We study the time of the first loss of a call in the system and derive the parameter of the approximating exponential distribution using the notion of a monotone structure.

6.3.1. $M/M/s/m$ *systems*

Consider a traditional Markov queueing system with losses. The calls arrive at the system according to a Poisson flow with rate λ. There are s independent identical servers with service rate μ_n where the service rate depends on a parameter n. The system has m waiting places. The call entering the system occupies the server or joins the queue if the server is busy. If all servers and all waiting positions are busy, the call is lost. On service completion the call leaves the system and the server immediately takes the next call from the queue if there are some waiting calls. Otherwise it waits for the next arriving call.

Suppose that the service is asymptotically fast, i.e., $\mu_n = n\mu$ where $n \to \infty$. Denote by $Q_n(t)$ the number of calls in the system at time t. Let $\Omega_n(q)$ be the time of the first loss of a call given that $Q_n(0) = q$. We study the asymptotic behavior of $\Omega_n(q)$ as $n \to \infty$.

It is easy to see that the process $Q_n(t)$ is a Birth and Death process with state space $Z = \{0, 1, \dots, s+m\}$ and birth and death rates in state q, λ and $\min(q, s)\mu_n$, respectively. Consider an auxiliary system with an infinite number of waiting places and denote by $\widehat{Q}_n(t)$ the number of calls in this system at time t. Then $\widehat{Q}_n(t)$ is a Birth-and-Death process with state space $Z = \{0, 1, \dots\}$. Given that $Q_n(0) = q$, we can represent $\Omega_n(q)$ as the first exit time from a subset $\{0, 1, \dots, s+m\}$:

$$\Omega_n(q) = \min\big\{t : t > 0,\ \widehat{Q}_n(t) > s+m\big\}.$$

Let $p_n(q, s)$ be the transition probabilities of the embedded MP. Then $p_n(q, q+1) = \lambda(\lambda + \min(q, s)\mu_n)^{-1}$, and for $q > 0$, $p_n(q, q+1) \approx n^{-1}\lambda/(\min(q, s)\mu)$. Therefore, the state space Z forms a monotone structure of the order $s+m-1$, where 0-level is the subset $Z_0 = \{0, 1\}$, q-level is the subset $Z_q = \{q+1\}$, $0 < q \le s+m-1$, $a_{ij}(q) = a(q) = \lambda/(\min(q+1, s)\mu)$, $q = 0, 1, \dots$, and $\varepsilon_n(q) = n^{-1}$.

Let $\pi_n(k)$, $k \ge 0$, be the stationary distribution for the embedded MP. We can easily see that limits $\lim_{n\to\infty} \pi_n(k) = \pi_k$, $k = 0, 1$, exist and $\pi_0 = \pi_1 = 1/2$. Then, using the matrix relation in Theorem 6.2, we obtain

$$\pi_n(k) = \frac{1}{n^{k-1}}\pi_n(1)\frac{\lambda^{k-1}}{(k-1)!\mu^{k-1}}\big(1 + o(1)\big), \quad \text{if } 1 < k \le s,$$

$$\pi_n(k) = \frac{1}{n^{k-1}}\pi_n(1)\frac{\lambda^{k-1}}{s!s^{k-s}\mu^{k-1}}\big(1 + o(1)\big), \quad \text{if } k > s,$$

and

$$g_n(Z) = \frac{1}{n^{s+m}}G\big(1 + o(1)\big),$$

where:

$$G = \frac{\lambda^{s+m}}{2s!s^m\mu^{s+m}}.$$

Finally, Theorem 6.1 implies that for any $i_0 \in Z$,

$$n^{-(s+m)}G\Omega_n(i_0) \stackrel{\mathrm{w}}{\Longrightarrow} \widehat{M}\eta_1,$$

where

$$\widehat{M} = \lim_{n\to\infty}\sum_{i\in Z_0}\pi_n(i)\mathbf{E}\tau_n(i),$$

η_1 is exponentially distributed with parameter 1, and $\tau_n(i)$ is the sojourn time in state i which is exponentially distributed with parameter $\lambda + n\min(i,s)\mu$. It is easy to calculate that $\widehat{M} = 1/(2\lambda)$. Rearranging the terms we get $n^{-(s+m)}\Omega_n(i) \approx G^{-1}M\eta_1$. From the above relations, by setting $\beta_n = n^{-(s+m)}$, we obtain the following result:

COROLLARY 6.4. *For the system $M/M/s/m$ under the assumption of fast service for any $q \leq s+m$,*

$$\lim_{n\to\infty} \mathbf{P}\{n^{-m-s}\Omega_n(q) \geq t\} = \exp\{-\Lambda t\}, \quad t > 0,$$

where $\Lambda = (s!s^m)^{-1}\lambda\rho^{s+m}$, $\rho = \lambda/\mu$.

Correspondingly, the flow of lost calls in the interval $[0, n^{m+s}t]$ weakly converges to a Poisson process with parameter Λ.

6.3.1.1. *System $M_M/M/\bar{l}/m$ in a Markov environment*

As an illustration of the wide possibilities of this technique which also allows us to consider an asymptotic aggregation of the state space of a system, consider an example given in [ANI 97]. Let $x(t)$, $t \geq 0$, be a continuous time MP with finite state space $\{1, 2, \ldots, r\}$ given by the transition rates $\{a_{ij},\ i = \overline{1,r},\ j = \overline{1,r},\ i \neq j\}$. Denote $a_{ii} = \sum_{j\neq i} a_{ij}$. Let the non-negative values $\{\lambda_k,\ k = \overline{1,r}\}$ also be given. The arrival process is switched by an MP $x(t)$: as $x(t) = k$, the instantaneous arrival rate is λ_k. We can also call this a Markov modulated arrival process [NEU 89]. The system has l labeled servers and the server i has the service rate $\mu_n(i)$, $i = \overline{1,l}$. The system also has m waiting places. The call entering the system either occupies the free server with a minimal label or joins the queue if no servers are available. After service completion, the first waiting call in the queue immediately goes for service if there are waiting calls. Otherwise, the server waits for the next call to arrive. If all servers and all waiting places are busy, the new arriving call is lost. We suppose that the service is asymptotically fast in the sense that

$$\mu_n(i) = n\mu_i, \quad i = \overline{1,l}, \tag{6.14}$$

where $n \to \infty$. Denote by $Q_n(t)$ the number of calls in the system at time t. Let $\Omega_n(k,j)$ be the time of the first loss of a call given that $x(0) = k$, $Q_n(0) = j$. We study the asymptotic behavior of $\Omega_n(k,j)$ as $n \to \infty$.

To describe the system as a multicomponent MP, we introduce the indicator of the state of the ith server: $\delta_i(t) = 1$, if at time t, ith server is occupied, and $\delta_i(t) = 0$ otherwise. Consider a multicomponent process $z_n(t) = (x(t), Q_n(t), \delta_i(t),\ i = \overline{1,l})$, $t \geq 0$. This process is a homogenous MP in continuous time with state space $I = \{(i, q, j_1, \ldots, j_l),\ i \in \{1, 2, \ldots, r\},\ q \in \{0, \ldots, l+m\}\}$, where the components j_i,

$i = \overline{1, l}$, take values 0 or 1. Introduce subsets:

$$I_s = \left\{ \big(i, s, j_1, \ldots, j_l\big) : i \in \{1, 2, \ldots, r\}, \ \sum_{k=1}^{l} j_k = s \right\} \quad \text{as } s \le l,$$

$$I_s = \big\{(i, s, 1, 1, \ldots, 1) : i \in \{1, 2, \ldots, r\}\big\} \quad \text{as } l < s \le l + m.$$

If $j_1 = j_2 = \cdots = j_s = 1$, $j_{s+1} = j_{s+2} = \cdots = j_l = 0$, and $s \le l$, denote for simplicity the state $(i, q, 1, 1, \ldots, 1, 0, \ldots, 0, 0)$ as (i, s). The state $(i, s, 1, 1, \ldots, 1)$ at $l < s \le l + m$ is also denoted as (i, s).

Then we can calculate the transition probabilities of the embedded MP between states (i, s): if $0 < s < l + m$

$$p_n\big((i, s), (j, s)\big) = a_{ij} c_n(s); \quad p_n\big((i, s), (i, s+1)\big) = \lambda_i c_n(s);$$

$$p_n\big((i, s), (i, s-1)\big) = n \left(\sum_{k=1}^{s_l} \mu_k \right) c_n(s),$$

where

$$c_n(s) = \left(\lambda_i + a_{ii} + n \sum_{k=1}^{s_l} \mu_k \right)^{-1}, \quad s_l = \min(s, l).$$

It is not difficult to calculate the transition probabilities between other states and to see that the state space I forms a monotone structure of the order $l + m - 1$, where 0-level is the subset $Z_0 = I_0 \cup I_1$, q-level is the subset $Z_q = I_{q+1}$, $0 < q \le l + m$, and we can choose $\varepsilon_n(q) = 1/n$.

It is also easy to see that for $0 < q \le l + m$ the transition probabilities at any level Z_q tend to zero. This means that in Definition 6.2 of a monotone structure $P(q) = 0$ as $0 < q \le l + m$. Let $\{\pi_n(i, s, j_1, \ldots, j_l)\}$ be the stationary probabilities of the embedded MP. Using the structure of transition probabilities we can calculate in a recursive way in $q = 1, 2, \ldots$ that for each level Z_q, $q > 0$, for the states in the form (i, q), $i = \overline{1, r}$, $\pi_n(i, q) = O(n^{-q+1})$, and for other states at this level $\pi_n(i, q, \ldots) = O(n^{-q})$. This means that the subset of states $\{(i, q), \ i = \overline{1, r}\}$ accumulates the main order of stationary probabilities at the q-level. As from 0-level to the 1-level the process can only come through the states $(i, 1, 1)$ (we denoted them as $(i, 1)$), we obtain from (6.7) that for $q > 1$, $i = \overline{1, r}$,

$$\pi_n(i, q) = n^{-q+1} \pi_n(i, 1) \lambda_i^{q-1} \prod_{s=1}^{q-1} \left(\sum_{k=1}^{s_l} \mu_k \right)^{-1} \big(1 + o(1)\big). \tag{6.15}$$

As the level Z_0 forms in a limit one essential class, then the limits

$$\lim_{n\to\infty} \pi_n(i,s) = \pi(i,s), \quad i=\overline{1,r},\ s=0,1,$$

exist and satisfy the system of equations:

$$\begin{aligned} \pi(i,0) &= \sum_{k\neq i} \pi(k,0)\frac{a_{ki}}{\lambda_k + a_{kk}} + \pi(i,1); \\ \pi(i,1) &= \pi(i,0)\frac{\lambda_i}{\lambda_i + a_{ii}}, \quad i=\overline{1,r}. \end{aligned} \tag{6.16}$$

These equations yield

$$\pi(i,0) = B\pi_i\big(\lambda_i + a_{ii}\big); \quad \pi(i,1) = B\pi_i\lambda_i, \quad i=\overline{1,r},$$

where $B = \left(\sum_{k=1}^{r} \pi_k(2\lambda_k + a_{kk})\right)^{-1}$, and π_i, $i=\overline{1,r}$, is the stationary distribution of an MP $x(t)$. Finally (6.15) implies:

$$\pi_n(i,q) = n^{-q+1}B\pi_i\lambda_i^q \prod_{s=1}^{q-1}\left(\sum_{k=1}^{s_l}\mu_k\right)^{-1}\big(1+o(1)\big), \quad i=\overline{1,r},\ q>0,$$

and

$$g_n(Z) = n^{-l-m}B\left(\sum_{k=1}^{l}\mu_k\right)^{-m-1}\prod_{s=1}^{l-1}\left(\sum_{j=1}^{s}\mu_j\right)^{-1}\sum_{i=1}^{r}\pi_i\lambda_i^{l+m+1}\big(1+o(1)\big), \tag{6.17}$$

where $\prod_1^0 = 1$. Set $\beta_n = n^{-l-m}$. As in each state of the type $(i,q,\ldots)$, $q>0$, the sojourn time has an exponential distribution with parameter tending to ∞, using the Corollary 6.2 and formulae (6.15)–(6.17) we prove the following result:

COROLLARY 6.5. *For the system $M_M/M/\bar{l}/m$ under assumptions (6.14) of fast service for any $i=\overline{1,r}$, $q\le l+m$, the variable $n^{-l-m}\Omega_n(i,q)$ weakly converges to the exponential random variable with parameter*

$$\Lambda = \left(\sum_{k=1}^{l}\mu_k\right)^{-m-1}\prod_{s=1}^{l-1}\left(\sum_{j=1}^{s}\mu_j\right)^{-1}\sum_{i=1}^{r}\pi_i\lambda_i^{l+m+1},$$

and the flow of lost calls $y_n(n^{l+m}t)$, $t\ge 0$, weakly converges to a Poisson process with parameter Λ.

In particular for the system $M/M/l/m$ with Poisson arrival process with parameter λ and l identical servers with service rate $n\mu$ we obtain

$$\Lambda = l^{-m}(l!)^{-1}\lambda(\lambda/\mu)^{l+m}.$$

This is in agreement with the result of Corollary 6.4.

6.3.2. *Semi-Markov queueing systems with fast service*

Consider a system $SM/M/l/m$ with semi-Markov type input and losses. Let $x(t)$, $t \geq 0$, be the right-continuous SMP with finite number of states $\{1, 2, \ldots, r\}$ given by the family of sojourn times $\{\tau_i,\ i = \overline{1, r}\}$ and embedded MP x_k, $k \geq 0$, with transition probability matrix $P = \|p_{ij}\|_{i,j=\overline{1,r}}$. Also let $0 = t_0 \leq t_1 \leq t_2 \leq \cdots$ be the times of sequential jumps of $x(t)$. Suppose that the calls enter the system one at a time at the times t_k. The system has l servers (for simplicity we suppose that they are identical), m waiting places and exponential service with parameter $n\mu$, $\mu > 0$.

Denote as before by $Q_n(t)$ the number of calls in the system at time t and study the asymptotic behavior of $\Omega_n(k, q)$ (the time of the first loss of a call given that $x(0) = k$, $Q_n(0) = q$). Let us consider the two-component process $\xi_n(t) = (x(t), Q_n(t))$. This process can be represented as an SP (more general example was considered in section 2.2.2.1). Since calls can be lost only at times t_k we can consider more simple embedded SMP $z_n(t) = (x(t), \bar{Q}_n(t))$, where in the interval $t \in [t_k, t_{k+1})$, $\bar{Q}_n(t) = Q_n(t_k + 0)$.

For the process $z_n(t)$ we introduce the embedded MP $z_{nk} = (x_k, Q_{nk})$, where $Q_{nk} = Q_n(t_k + 0)$. Let $p_n((i, s), (j, q))$ be the transition probability from state (i, s) to (j, q) for the process z_{nk} and $\tau_n((i, s), (j, q))$ be the time of this transition for the process $z_n(t)$. Denote $G_i(x) = P\{\tau_i < x\}$, $i = \overline{1, r}$. We suppose that

$$G_i(+0) = 0, \quad i = \overline{1, r}. \tag{6.18}$$

Let us introduce the values

$$v_n(i, s) = \int_0^\infty \exp\big(-n\mu s_l x\big) \mathrm{d}G_i(x), \quad i = \overline{1, r},\ s > 0,$$

where $s_l = \min(s, l)$. It is easy to see that $\lim_{n\to\infty} v_n(i, s) = 0$, $s > 0$, and

$$\begin{aligned} p_n\big((i, s), (j, s+1)\big) &= p_{ij} v_n(i, s), \quad i, j = \overline{1, r},\ s > 0; \\ \lim_{n\to\infty} p_n\big((i, s), (j, s)\big) &= 0, \quad i, j = \overline{1, r},\ s > 1; \\ p_n\big((i, s), (j, s+2)\big) &\equiv 0, \quad i, j = \overline{1, r},\ s \geq 0; \\ p_n\big((i, 1), (j, 1)\big) &= p_{ij}\big(1 - v_n(i, 1)\big), \quad i, j = \overline{1, r}. \end{aligned} \tag{6.19}$$

Thus, the state space of the process z_{nk} forms a monotone structure of the order $l+m-1$ and the level q is a subset $Z_q = \{(i, q+1),\ i = \overline{1,r}\}$, $q = 0, 1, \ldots, l+m-1$. Suppose that there exist constants $b(i, s)$ and a normalizing factor $\gamma_n \to 0$ such that, as $n \to \infty$,

$$v_n(i, s) = \gamma_n b(i, s)\big(1 + o(1)\big), \quad i = \overline{1, r},\ s > 0. \tag{6.20}$$

Thus, (6.19) implies

$$\pi_n(i, 1) = \pi_i\big(1 + o(1)\big), \quad i = \overline{1, r},$$

where $\pi_n(i, s)$ is the stationary distribution for the process z_{nk} and π_i is the stationary distribution for the MP x_k.

Further we denote $A(s) = \|p_{ij} b(i, s)\|$, $i, j = \overline{1, r}$, $s > 0$. Then according to Theorem 6.2 we obtain that

$$\bar{\pi}_n(s) = \gamma_n^{s-1} \bar{\pi} A(1) \cdots A(s-1)\big(1 + o(1)\big), \quad s > 0,$$

and

$$g_n = \gamma_n^{l+m} \bar{\pi} A(1) \cdots A(l-1) A(l)^{m+1} \bar{1}\big(1 + o(1)\big), \quad s > 0,$$

where $\bar{\pi}_n(s) = (\pi_n(i, s),\ i = \overline{1, r})$, $\bar{\pi} = (\pi_i,\ i = \overline{1, r})$ are row-vectors, $\bar{1}$ is the unit column-vector.

Now we need to check conditions (6.4). Suppose that there exist $0 < \alpha \leq 1$ and constants c_i, $i = \overline{1, r}$, such that

$$E \exp\big(- \theta \tau_i\big) = 1 - c_i \theta^{\alpha} + o\big(\theta^{\alpha}\big), \quad i = \overline{1, r}. \tag{6.21}$$

Using the inequality

$$\int_{-\infty}^{\infty} f(x) g(x) \mathrm{d}F(x) \leq \int_{-\infty}^{\infty} f(x) \mathrm{d}F(x) \int_{-\infty}^{\infty} g(y) \mathrm{d}F(y),$$

which is true for any distribution function $F(x)$ and any functions $f(x) \geq 0$, $g(x) \geq 0$, where $f(x)$ and $g(x)$ are non-increasing functions, it is possible to prove that as $n \to \infty$,

$$\mathbf{E} \exp\big(- \beta_n \theta \tau_n\big((i, 1), (j, 1)\big)\big) = 1 - \gamma_n^{l+m} c_i \theta^{\alpha} + o\big(\gamma_n^{l+m}\big),$$

and

$$p_n\big((i, s), (j, q)\big)\big(1 - \mathbf{E} \exp\big(\beta_n \theta \tau_n\big((i, s), (j, q)\big)\big)\big) = O\big(\gamma_n^{l+m}\big)$$

for any $i = \overline{1, r}$, $j = \overline{1, r}$, $0 \leq q \leq s + 1$, where $\beta_n = \gamma_n^{\frac{l+m}{\alpha}}$.

Thus, according to Theorem 6.1 we obtain the following result.

THEOREM 6.3. *Under the assumptions (6.18), (6.20), (6.21) for any* $k = \overline{1,r}$, $q \le l+m$, *as* $n \to \infty$, *the asymptotic relation*

$$\mathbf{E}\exp\big(-\theta\gamma_n^{\frac{l+m}{\alpha}}\Omega_n(k,q)\big) \longrightarrow \big(1+M\theta^{\alpha}\big)^{-1},$$

is true, where

$$M = \left(\sum_{i=1}^{r}\pi_i c_i\right)\big(\bar{\pi}A(1)\cdots A(l-1)A(l)^{m+1}\bar{1}\big)^{-1},$$

and the flow of lost calls in the interval $[0, \gamma_n^{-\frac{l+m}{\alpha}}t]$ *weakly converges to a recurrent flow with the Laplace transformation of the interval-arrival time* $(1+M\theta^{\alpha})^{-1}$.

In particular for the system $GI/M/1/m$ with recurrent input $M = cb^{-m-1}$.

Some others models of renewable systems with fast service or fast repair were considered in [ANI 87, ANI 89a, ANI 89b, ANI 89c]. Note that Theorem 6.1 implies that the asymptotic distribution of the sojourn time in an S-set does not depend on the initial state. This allows us to study the models of asymptotic aggregation of state space [ANI 73, ANI 78, ANI 87, ANI 88b]. We also acknowledge other techniques of studying rare events in queueing systems (e.g. [SOL 71, KOV 80, KOV 94]).

In conclusion of this section we stress that the asymptotic behavior of the time of first loss of a call is not invariant concerning the mean characteristics of the input flow and service time. In fact, this behavior is determined by the normalizing coefficient γ_n and by the value M. The value M is determined by the behavior of the functions $G_i(x)$ as $x \to \infty$. If the mean values of the variables τ_i exist, then M is invariant concerning the means. However, the order of γ_n is determined by the behavior of $G_i(x)$ in the vicinity of zero. For example, if $G_i(x) \sim x^{\alpha}$ as $x \to 0$, then $\gamma_n \sim c/n^{\alpha}$. Therefore, it is not difficult to construct three queueing models with the same means of the times τ_i such that for one system the variable $\Omega_n(k,q)/n$ converges to the exponential distribution with a parameter λ ($0 < \lambda < \infty$), and for two other systems $\Omega_n(k,q)/n$ converges to zero and to ∞, respectively.

6.4. Single-server retrial queueing model

In the following sections we consider the applications of the method of S-sets to the asymptotic analysis of retrial queueing systems. The presentation follows [ANI 01, ANI 00b].

Consider a single-server retrial queueing model with m waiting places for calls in the retrial group and losses. It operates in the following way. The arrival process is a Poisson process with rate λ. Arriving calls are identified as primary calls. If the server is free, then the arriving call immediately goes to service. After service completion the call leaves the system, otherwise, if the server is busy and there are less than

m calls waiting in the retrial group, the arriving call joins a retrial group (orbit) and becomes a source of repeated calls – secondary calls. On the other hand, if a primary arriving call finds the server and all of the waiting positions in the orbit occupied, the call will be lost. Each call in the orbit produces a Poisson process of repeated attempts with rate ν. If a secondary call finds the server free, the service immediately starts and after completion of service the call leaves the system. We assume for simplicity that the distribution of service time is exponential with parameter μ for both primary and secondary calls and the input flow of primary calls, attempts for service of different retrial calls and service times are jointly independent. This model is called an $M/M/1/m/wr$ retrial model. A similar system with m servers was described in section 5.4.3.

This section deals with the analysis of the time of first loss of a call in $M/M/1/m/wr$ queueing model. The method of analysis is based on the results of the asymptotic behavior of the first exit time from the fixed subset of states forming a monotone structure given in the previous sections. First, we consider the asymptotic behavior of the system under the assumption of fast service. Then we consider the system under the assumption of both a fast service and large retrial rate. Finally, we consider the system operating in a Markov environment and under the assumption of fast service. We derive the expressions for the parameter of exponential distribution in asymptotic approximation of the time of first loss of a call for these models.

Suppose that the service rate $\mu = \mu_n$ and the retrial rate of secondary calls $\nu = \nu_n$ may depend on some scaling factor n, $n \to \infty$. Without loss of generality we assume that the arrival rate λ does not depend on n. Consider the following cases:

Case 1: $\mu_n = n\mu$ (fast service), $\nu_n = \nu$ (usual retrial rate), $n \to \infty$.

Case 2: $\mu_n = n\mu$ (fast service), $\nu_n = V_n\nu$ (large retrial rate), $n \to \infty$, $V_n \to \infty$.

Denote by $Q_n(t)$, $t \geq 0$, the number of sources of repeated calls (the number of calls in the orbit) at time t, and let the component $\delta_n(t)$ denote the state of service at time t ($\delta_n(t) = 1$ if at time t the server is occupied and $\delta_n(t) = 0$ otherwise).

Let $\Omega_n(j, q)$ be the time of first loss of a call given the $Q_n(0) = q$ and $\delta_n(0) = j$, $0 \leq q \leq m$, $j = 0, 1$, and $Y_n(t)$ be the number of lost calls on the interval $[0, t]$. We study the asymptotic behavior of $\Omega_n(j, q)$ as $n \to \infty$.

6.4.1. *Case 1: fast service*

THEOREM 6.4. *For the model described above under the assumption of fast service (case 1), if $\lambda > 0$, $\mu > 0$, $\nu > 0$, then for any initial state (j, q), $0 \leq q \leq m$, $j = 0, 1$, the distribution of the normalized random variable $n^{-m-1}\Omega_n(j, q)$ weakly converges to the exponential distribution:*

$$\lim_{n\to\infty} \mathbf{P}\big\{n^{-m-1}\Omega_n(j, q) \geq t\big\} = \exp\{-\Lambda t\}, \quad t > 0,$$

where

$$\Lambda = \frac{\lambda\rho^{m+1}}{m!\nu^m}\prod_{k=1}^{m}(\lambda + k\nu), \quad \rho = \lambda/\mu. \tag{6.22}$$

Proof. Consider a multicomponent process $z_n(t) = (\delta_n(t), Q_n(t))$, $t \geq 0$. Process $z_n(t)$ is a continuous time homogenous MP with state space

$$Z = \big\{(j,q),\ j = 0,1,\ q = 0,1,\ldots,m\big\}$$

and describes the queueing process in the system. Denote by $\widehat{Q}_n(t)$ the number of retrial calls in the auxiliary system with infinite number of waiting places in the orbit and put $\widehat{z}_n(t) = (\delta_n(t), \widehat{Q}_n(t))$. Then $\widehat{z}_n(t)$ is an MP with state space $\{0,1\} \times \{0,1,\ldots\}$ and $\Omega_n(j,q)$ is the first exit time of $\widehat{z}_n(t)$ from the subset Z:

$$\Omega_n(j,q) = \min\big\{t : t > 0,\ \widehat{Q}_n(t) > m \text{ given } \widehat{Q}_n(0) = q,\ \delta_n(0) = j\big\}.$$

Transition rates of the process $z_n(t)$ can be easily calculated and it can be seen that the subset Z forms a monotone structure (see Definition 6.2) where at each fixed $q = 0,1,\ldots,m$, the subset $Z_q = \{(j,q),\ j = 0,1\}$ forms q-level. The monotone structure for the model and corresponding transition probabilities are shown in Figure 6.2 where α_q, β_q and $\varepsilon_n(q)$ are defined as

$$\alpha_q = \frac{q\nu}{\lambda + q\nu}, \quad \beta_q = \frac{\lambda}{\lambda + q\nu}, \quad \varepsilon_n(q) = \frac{1}{n}\frac{\lambda}{\mu}.$$

A sojourn time in state (j,q) for the process $\widehat{z}_n(t)$ has an exponential distribution with parameter

$$\Lambda_n(j,q) = \begin{cases} \lambda + n\mu & \text{if } j = 1, \\ \lambda + q\nu & \text{if } j = 0. \end{cases}$$

The transition probabilities are:

$$p_n\big((1,q),(1,q+1)\big) \approx \frac{1}{n}\frac{\lambda}{\mu} \longrightarrow 0, \quad p_n\big((0,q),(1,q-1)\big) = \frac{q\nu}{\lambda + q\nu},$$
$$p_n\big((0,q),(1,q)\big) = \frac{\lambda}{\lambda + q\nu}, \quad p_n\big((1,q),(0,q)\big) = \frac{n\mu}{\lambda + n\mu} \longrightarrow 1.$$

Now we can directly apply matrix relations of Theorem 6.2. Denote by $\overline{\pi}_n(q) = (\pi_n(0,q), \pi_n(1,q))$ the stationary distribution of the embedded Markov process for $z_n(t)$ and let $\pi_i = \pi(i,0), i = 0,1$ ($\overline{\pi} = (\pi_0, \pi_1)$) be the limiting stationary distribution for the states of $Z_0 = \{(0,0),(1,0)\}$ (0-level). Since Z_0 forms one essential class in the limit, these limiting probabilities exist and we can easy calculate that $\pi_0 = \pi_1 = 1/2$.

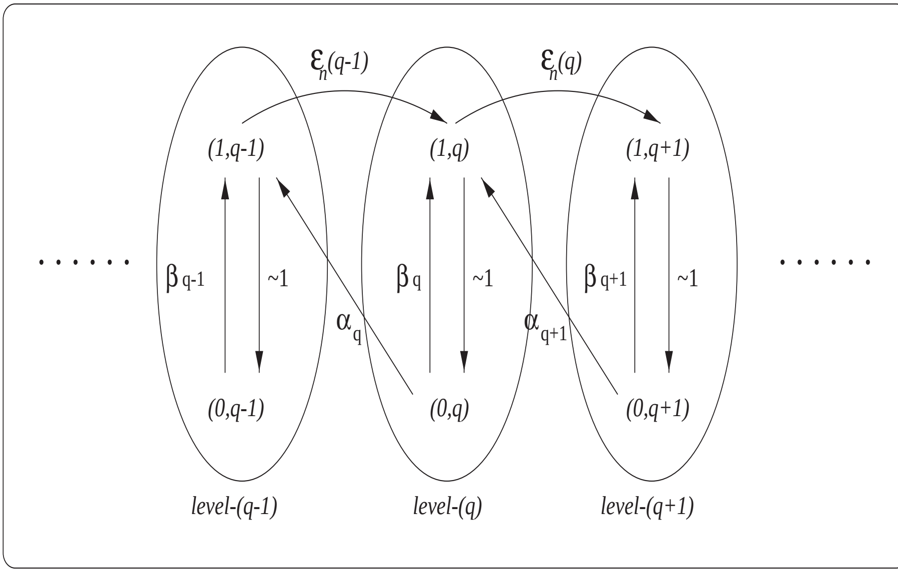

Figure 6.2. *Monotone structure for single-server model with fast service*

In matrix relations (6.7) of Theorem 6.2, the matrices $A(j)$ and $P(j)$ are

$$A(j) = \begin{bmatrix} 0 & 0 \\ 0 & \lambda/\mu \end{bmatrix}, \quad P(j) = \begin{bmatrix} 0 & \dfrac{\lambda}{\lambda + j\nu} \\ 1 & 0 \end{bmatrix}.$$

Therefore, the stationary distribution of the states of the embedded MP for $z_n(t)$ can be represented as

$$\overline{\pi}_n(q) = \overline{\pi} \prod_{j=0}^{q-1} \left(\begin{bmatrix} 0 & 0 \\ 0 & \lambda/\mu \end{bmatrix} \left(I - \begin{bmatrix} 0 & \dfrac{\lambda}{\lambda + (j+1)\nu} \\ 1 & 0 \end{bmatrix} \right)^{-1} \frac{1}{n} \right) \big(1 + o(1)\big).$$

Rearranging terms, we obtain

$$\overline{\pi}_n(q) = \overline{\pi} \prod_{j=0}^{q-1} \left(\begin{bmatrix} 0 & 0 \\ 0 & \lambda/\mu \end{bmatrix} \frac{\lambda + (j+1)\nu}{(j+1)\nu} \begin{bmatrix} 1 & \dfrac{\lambda}{\lambda + (j+1)\nu} \\ 1 & 1 \end{bmatrix} \frac{1}{n} \right) \big(1 + o(1)\big).$$

After some algebra we obtain

$$\overline{\pi}_n(q) = \frac{1}{n^q} \pi_1 \frac{\rho^q}{q!\nu^q} \prod_{j=1}^{q} (\lambda + j\nu) \overline{e} \big(1 + o(1)\big), \quad q = 1, 2, \ldots,$$

where $\rho = \lambda/\mu$ and $\overline{e}$ is the unit vector. Note that

$$\overline{\pi}_n(q) = O\left(\prod_{s=0}^{q-1} \varepsilon_n(s)\right).$$

The expression for $g_n(Z)$ can be obtained in the same way:

$$g_n(Z) = \frac{1}{n^{m+1}} \pi_1 \frac{\rho^{m+1}}{m!\nu^m} \prod_{j=1}^{m} (\lambda + j\nu)\big(1 + o(1)\big) = \frac{1}{n^{m+1}} G\big(1 + o(1)\big).$$

If we set the normalizing coefficient $\beta_n = n^{-m-1}$ in relation (6.4), then we obtain:

$$a_{(j,q)}(\theta) = \begin{cases} 0 & \text{as } q > 0,\ j = 0, 1, \\ 0 & \text{as } q = 0,\ j = 1, \\ G^{-1}\lambda^{-1}\theta & \text{as } q = 0,\ j = 0. \end{cases}$$

Therefore, for the value $A(\theta)$ in (6.5) we obtain: $A(\theta) = \pi_0(G\lambda)^{-1}\theta$. This means that the limiting distribution is exponential with parameter $\Lambda = \pi_0(G\lambda)^{-1}$. After simple transformations we obtain expression (6.22). □

6.4.1.1. *State-dependent case*

These results can be extended to the case when the arrival, service and retrial rates may depend on the size of the queue. This means, the values $\lambda(k)$, $\mu(k)$, $k = \overline{0, m}$, and $\nu(k)$, $k = \overline{1, m}$, are given. At $Q_n(t) = k$, the instantaneous arrival rate is $\lambda(k)$, the instantaneous service rate is $n\mu(k)$ as $k = \overline{0, m}$, and the instantaneous retrial rate is $\nu(k)$ as $k = \overline{1, m}$. In this case using the same technique we obtain the following statement:

STATEMENT 6.4. *Assume that* $\prod_{k=0}^{m}(\lambda(k)\mu(k)) \prod_{j=1}^{m} \nu(j) > 0$. *Then for the state-dependent retrial model* $M/M1/m/wr$ *in the case of fast service the distribution of the variable* $n^{-m-1}\Omega_n(j, q)$ *weakly converges to the exponential distribution with parameter*

$$\Lambda = \lambda(0) \frac{1}{m!} \prod_{k=0}^{m} \frac{\lambda(k)}{\mu(k)} \prod_{k=0}^{m-1} \frac{\lambda(k) + (k+1)\nu(k+1)}{\nu(k+1)}.$$

Denote by $Y_n(t)$ the number of lost calls.

NOTE 6.1. In both cases (homogenous or state-dependent model) process $Y(n^{m+1}t)$ weakly converges to an ordinary Poisson process with parameter Λ.

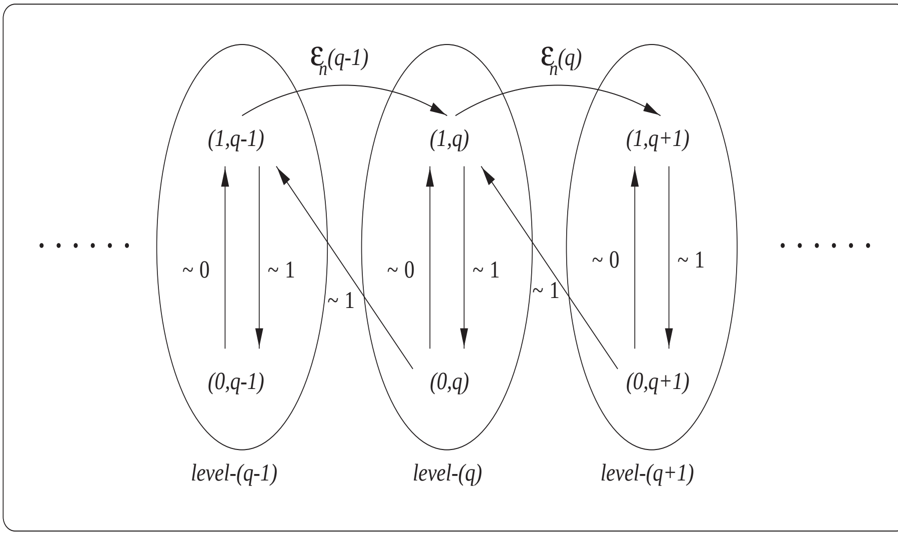

Figure 6.3. *Monotone structure for a single-server model with large service and retrial rates*

6.4.2. *Case 2: fast service and large retrial rate*

Now we consider the system with a single server and m waiting places in orbit and suppose that the service and retrial rates are both large in the sense that $\mu_n = n\mu$ and $\nu_n = V_n\nu$. Again we study the asymptotic behavior of the time $\Omega_n(j,q)$ of the first lost of a call as $n \to \infty$, $V_n \to \infty$.

THEOREM 6.5. *For the system described above (case 2), under the assumption that the service and retrial rates are asymptotically large, if $\lambda > 0$, $\mu > 0$, $\nu > 0$, then for any initial state the distribution of the normalized variable $n^{-m-1}\Omega_n(j,q)$ weakly converges to the exponential distribution with parameter Λ:*

$$\lim_{n\to\infty} \mathbf{P}\big\{n^{-m-1}\Omega_n(j,q) > t\big\} = \exp\{-\Lambda t\}, \quad t \geq 0,$$

where $\Lambda = \lambda(\lambda/\mu)^{m+1}$.

Proof. The proof is similar to the proof of Theorem 6.4. Consider an auxiliary multicomponent MP $\widehat{z}_n(t) = (\delta_n(t), \widehat{Q}_n(t))$, $t \geq 0$. It can easily be seen that the subset Z forms a monotone structure and at each fixed $q = 0, 1, \ldots, m$ the subset $Z_q = \{(j,q),\ j = 0, 1\}$ forms q-level. Figure 6.3 illustrates the monotone structure of the model and corresponding transition probabilities.

In each state (j,q) the process $\widehat{z}_n(t)$ spends an exponential time with parameter

$$\Lambda_n(j,q) = \begin{cases} \lambda + n\mu & \text{if } j = 1, \\ \lambda + qV_n\nu & \text{if } j = 0. \end{cases}$$

Denote as before $\rho = \lambda/\mu$. The transition probabilities for the process are as follows:

$$p_n\big((1,q),(1,q+1)\big) \approx \frac{1}{n}\rho \longrightarrow 0, \quad p_n\big((0,q),(1,q-1)\big) = \frac{qV_n\nu}{\lambda + qV_n\nu} \longrightarrow 1,$$

$$p_n\big((0,q),(1,q)\big) = \frac{\lambda}{\lambda + qV_n\nu}, \quad p_n\big((1,q),(0,q)\big) = \frac{n\mu}{\lambda + n\mu} \longrightarrow 1.$$

Denote by $\overline{\pi}_n(q) = (\pi_n(0,q), \pi_n(1,q))$, $q \geq 0$, the stationary distribution of the embedded Markov process for $\widehat{z}_n(t)$, and let $\pi_j = \pi_0(j,0)$, $j = 0,1$, ($\overline{\pi} = (\pi_0, \pi_1)$) be the limiting stationary probabilities of the level Z_0. Again we can easily calculate that $\pi_0 = \pi_1 = 1/2$.

In the matrix relations of Theorem 6.2, matrices $A(j)$ and $P(j+1)$ are:

$$A(j) = \begin{bmatrix} 0 & 0 \\ 0 & \rho \end{bmatrix}, \qquad P(j+1) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

After some algebra the expressions for $\overline{\pi}_n(q)$ and $g_n(Z)$ are obtained as

$$\overline{\pi}_n(q) = \pi_1 \frac{1}{n^q} \rho^q \overline{e}\big(1 + o(1)\big),$$

$$g_n(Z) = \frac{1}{n^{m+1}} \pi_1 \rho^{m+1}\big(1 + o(1)\big).$$

Setting $\beta_n = n^{-m-1}$ we obtain the parameter of the limiting exponential distribution $\Lambda = \lambda\rho^{m+1}$. □

Note that in this case the answer does not depend on the value $\nu > 0$. However, if $\nu = 0$ or $\nu = \nu_n \to 0$, the answer will be different.

Now consider a state-dependent case. Suppose that as $Q_n(t) = q$, the instantaneous arrival rate is $\lambda(q)$, the instantaneous service rate is $n\mu(q)$, $q = \overline{0,m}$, and at $q = \overline{1,m}$ the instantaneous retrial rate is $V_n\nu(q)$. Here $n \to \infty$, $V_n \to \infty$.

NOTE 6.2. Suppose that $\prod_{k=0}^{m}(\lambda(k)\mu(k)) \prod_{j=1}^{m} \nu(j) > 0$. Then at the assumption of large service and retrial rates, the statement of Theorem 6.5 also holds where the parameter of exponential distribution becomes

$$\Lambda = \lambda(0) \prod_{k=0}^{m} \frac{\lambda(k)}{\mu(k)},$$

and the process $Y(n^{m+1}t)$ weakly converges in any finite interval to an ordinary Poisson process with parameter Λ.

In this case the result also does not depend on values $\nu(j) > 0$, $j = \overline{1,m}$.

6.4.3. *State-dependent model in a Markov environment*

Consider a Markov retrial queueing model of the type $M_M/M_M/1/m/wr$. It consists of one server and m waiting places in orbit. In addition to the system $M/M/1/m/wr$ described in the previous section, this system operates in a Markov environment $x(t)$, $t \geq 0$, where $x(t)$ is an ergodic MP with finite state space $X = \{1, 2, \ldots, r\}$ given by the initial state x_0 and transitions rates a_{ij}, $i, j \in X$, $i \neq j$. Denote by π_i, $i = \overline{1, r}$, the stationary distribution of $x(t)$. Let $\lambda(i, q)$, $\nu(i, q)$, $\mu(i, q)$, $i \in X$, $q = \overline{0, m}$, be given non-negative functions. Consider the case of fast service. Denote by $Q_n(t)$ the number of waiting calls in the retrial queue at time t. Fast service here means that if $x(t) = i$ and $Q_n(t) = q$, then $\lambda(i, q)$ is the instantaneous input rate, $\nu(i, q)$ is the instantaneous retrial rate and $n\mu(i, q)$ is the instantaneous service rate, where n is a scaling factor ($n \to \infty$). In addition, denote by $Y_n(t)$ the number of lost calls in the interval $[0, t]$.

Let $\Omega_n(i, j, q)$ be the time of the first loss of a call given $x(0) = i$, $Q_n(0) = q$ and $\delta_n(0) = j$. The asymptotic behavior of $\Omega_n(i, j, q)$ as $n \to \infty$ is studied.

THEOREM 6.6. *Suppose that:*

$$\max_{i=\overline{1,r}} \lambda(i, q) > 0,\ q = \overline{0, m}; \quad \max_{i=\overline{1,r}} \nu(i, q) > 0,\ q = \overline{1, m}; \qquad (6.23)$$
$$\min_{i=\overline{1,r},\ q=\overline{0,m}} \mu(i, q) > 0,$$

and the condition of fast service holds. Then for any initial state $(i, j, q) \in Z$,

$$\lim_{n\to\infty} \mathbf{P}\big\{n^{-m-1}\Omega_n(i, j, q) \geq t\big\} = \exp\{-\Lambda t\}, \quad t \geq 0, \qquad (6.24)$$

where

$$\Lambda = \overline{\pi}\Lambda G(0)\big(I - B(1) - \Lambda(1)\big)^{-1}(I - B(1))G(1) \cdots \qquad (6.25)$$
$$\cdots G(m-1)\big(I - B(m) - \Lambda(m)\big)^{-1}\big(I - B(m)\big)G(m)\overline{e},$$

$\overline{\pi}$ is a row vector $(\pi_1, \ldots, \pi_r)$, Λ and $G(q), q = \overline{0, m}$, are diagonal matrices with diagonal elements $\lambda(i, 0)$ and $\lambda(i, q)/\mu(i, q)$ correspondingly, $B(q)$ is defined as

$$B(q) = \left\| \frac{a_{ij}\big(1 - \delta_{ij}\big)}{\lambda(i, q) + a_{ii} + q\nu(i, q)} \right\|, \quad i, j = \overline{1, r},\ q = \overline{1, m},$$

where δ_{ij} is Kronecker symbol, $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$, and $\Lambda(q)$ is defined as

$$\Lambda(q) = \left\| \frac{\lambda(i, q)\delta_{ij}}{\lambda(i, q) + a_{ii} + q\nu(i, q)} \right\|, \quad i, j = \overline{1, r},\ q = \overline{1, m}.$$

Proof. Consider a multicomponent process $z_n(t) = (x(t), \delta_n(t), Q_n(t))$, $t \geq 0$. The process $z_n(t)$ is a homogenous MP in continuous time and the state space

$$Z = \{(i, j, q),\ i \in X,\ j = 0, 1,\ q = \overline{0, m}\}.$$

Denote by $\widehat{z}_n(t) = (x(t), \delta_n(t), \widehat{Q}_n(t))$, $t \geq 0$, an auxiliary MP, where $\widehat{Q}_n(t)$ is the number of waiting calls in the orbit for a system with an infinite number of waiting places. Then $\Omega_n(i, j, q)$ is the exit time of the process $\widehat{z}_n(t)$ from the subset Z.

The transitions rates for the process $\widehat{z}_n(t)$ can be easily calculated and it can be seen that the subset Z forms a monotone structure where at each fixed $q = 0, 1, \ldots, m$ the subset $Z_q = \{(i, j, q),\ i = \overline{1, r},\ j = 0, 1\}$ forms q-level. The monotone structure and corresponding transition probabilities for this model are shown in Figure 6.4.

In each state (i, j, q) the process $\widehat{z}_n(t)$ spends an exponential time with parameter

$$\Lambda_n(i, j, q) = \begin{cases} \lambda(i, q) + n\mu(i, q) + a_{ii} & \text{if } j = 1, \\ \lambda(i, q) + q\nu(i, q) + a_{ii} & \text{if } j = 0, \end{cases}$$

where $a_{ii} = \sum_{k \neq i} a_{ik}$. The transition probabilities for the embedded Markov process are as follows:

$$p_n\big((i, 1, q), (i, 1, q+1)\big) = \frac{\lambda(i, q)}{\lambda(i, q) + a_{ii} + n\mu(i, q)} \approx \frac{1}{n} \frac{\lambda(i, q)}{\mu(i, q)}, \quad i = \overline{1, r},$$

$$p_n\big((i, 0, q), (i, 1, q-1)\big) = \frac{q\nu(i, q)}{\lambda(i, q) + a_{ii} + q\nu(i, q)},$$

$$p_n\big((i, 0, q), (i, 1, q)\big) = \frac{\lambda(i, q)}{\lambda(i, q) + a_{ii} + q\nu(i, q)},$$

$$p_n\big((i, 0, q), (k, 0, q)\big) = \frac{a_{ik}}{\lambda(i, q) + a_{ii} + q\nu(i, q)}, \quad i \neq k,$$

$$p_n\big((i, 1, q), (i, 0, q)\big) = \frac{n\mu(i, q)}{\lambda(i, q) + a_{ii} + n\mu(i, q)} \longrightarrow 1,$$

$$p_n\big((i, 1, q), (k, 1, q)\big) = \frac{a_{ik}}{\lambda(i, q) + a_{ii} + n\mu(i, q)} \longrightarrow 0, \quad i \neq k.$$

Now we can directly apply matrix relations of Theorem 6.2. Denote by $\overline{\pi}_n(j, q) = (\pi_n(i, j, q),\ i \in X,\ j = 0, 1,\ q = 0, 1, \ldots, m)$ the stationary distribution of the embedded Markov process for $z_n(t)$. In Theorem 6.2 the matrices $A(j)$ and $P(j)$ are defined as:

$$A(j) = \begin{bmatrix} 0 & 0 \\ 0 & G(j) \end{bmatrix}, \quad P(j) = \begin{bmatrix} B(j) & \Lambda(j) \\ I & 0 \end{bmatrix}.$$

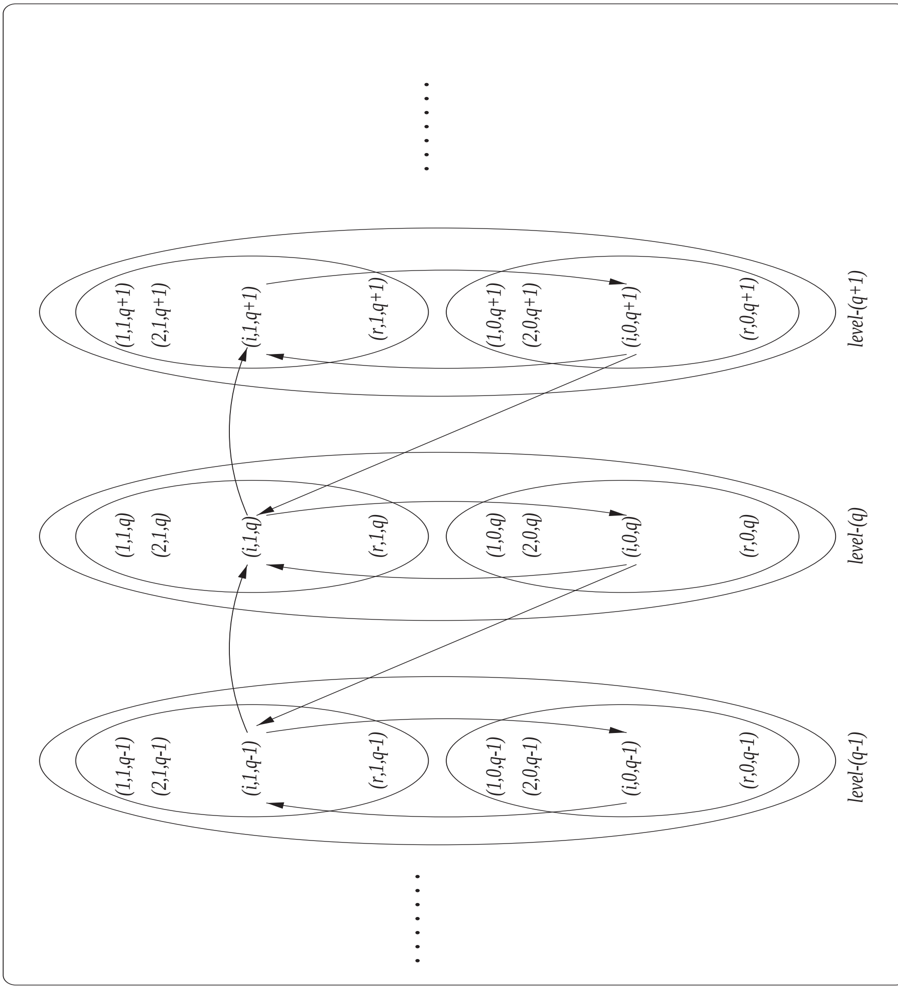

Figure 6.4. *Monotone structure for a single-server system in a Markov environment and fast service*

Denote $\overline{\pi}_n(q) = (\overline{\pi}_n(0,q), \overline{\pi}_n(1,q))$, where $\overline{\pi}_n(0,q) = (\pi(i,0,q),\ i = \overline{1,r})$ and $\overline{\pi}_n(1,q) = (\pi(i,1,q),\ i = \overline{1,r})$, $q = \overline{0,m}$, are row vectors.

Using the formula from [GAN 98] we obtain

$$\big(I - P(q)\big)^{-1} = \begin{bmatrix} \big(I - B(q) - \Lambda(q)\big)^{-1} & \big(I - B(q) - \Lambda(q)\big)^{-1}\Lambda(q) \\ \big(I - B(q) - \Lambda(q)\big)^{-1} & \big(I - B(q) - \Lambda(q)\big)^{-1}\big(I - B(q)\big) \end{bmatrix}.$$

Now applying Theorem 6.2 we can write the expression for the stationary probability of the exit in the matrix form

$$g_n(Z) = \frac{1}{n^{m+1}}\overline{\pi}(1,0)\left(\prod_{j=0}^{m-1} G(j)K(j+1)\big(I - B(j+1)\big)\right)G(m)\big(1+o(1)\big),$$

where $K(j) = (I - B(j) - \Lambda(j))^{-1}$ and $\overline{\pi}(1,0) = \lim_{n\to\infty}\overline{\pi}_n(1,0)$.

Since the level Z_0 forms in a limit one essential class, the stationary probabilities $\pi(i,j,0)$ of state $(i,j,0) \in Z_0$ exist and satisfy the system of equations:

$$\pi(i,0,0) = \sum_{k\neq i}\pi(k,0,0)\frac{a_{ki}}{\lambda(k,0)+a_{kk}} + \pi(i,1,0),$$

$$\pi(i,1,0) = \pi(i,0,0)\frac{\lambda(i,0)}{\lambda(i,0)+a_{ii}}, \quad i=\overline{1,r}.$$

Denote $B = \sum_{k=1}^{r}\pi_k(2\lambda(k,0)+a_{kk})$. It can be easily shown that

$$\pi(i,0,0) = B^{-1}\big(\lambda(i,0)+a_{ii}\big)\pi_i, \quad \pi(i,1,0) = B^{-1}\lambda(i,0)\pi_i, \quad i=\overline{1,r}.$$

Finally, setting $\beta_n = n^{-m-1}$, we obtain the parameter of exponential distribution in the form (6.25). □

Note that the process $Y(n^{m+1}t)$ weakly converges to an ordinary Poisson Process with parameter Λ.

We mention that if for some q-level, $\nu(i,q)=0$ for all $i=\overline{1,r}$, then Z does not form a monotone structure and the result will be different.

In the same way we can consider the case of large service and large retrial rates. This means that in the previous system, given that $x(t)=i$, $Q_n(t)=q$, the instantaneous retrial rate is $V_n\nu(i,q)$, $i=\overline{1,r}$, $q=0,1,\ldots$, where $V_n\to\infty$.

STATEMENT 6.5. *Suppose that* $\max_{i=\overline{1,r}}\lambda(i,q)>0$, $q=\overline{0,m}$, *and for all* $i=\overline{1,r}$, $\nu(i,q)>0$, $q=\overline{1,m}$*;* $\mu(i,q)>0$, $q=\overline{0,m}$.

Then in the case of large service and large retrial rates, relation (6.24) is true where

$$\Lambda = \overline{\pi}\Lambda G(0)G(1)\cdots G(m)\overline{e}.$$

Note that in this case the answer also does not depend on the values $\nu(i,q)>0$.

It is also possible to study the mixed case when some of values $\nu(i,q)$ may be equal to zero. Consider the model with large instantaneous retrial rate $V_n\nu(i,q)$, $i=\overline{1,r}$, $q=0,1,\ldots$.

STATEMENT 6.6. *Suppose that in our notation relation (6.23) holds. Then relation (6.24) is true, Λ is calculated according to (6.25) where corresponding matrices are calculated as follows: Λ and $G(q)$ are the same diagonal matrices, $B(q) = \|b_{ij}(q)\|$, where*

$$b_{ij}(q) = a_{ij}\big(1 - \delta_{ij}\big)\big(\lambda(i,q) + a_{ii}\big)^{-1}, \text{ if } \nu(i,q) = 0; \quad b_{ij}(q) = 0, \text{ if } \nu(i,q) > 0,$$

and $\Lambda(q)$ is a diagonal matrix with diagonal elements

$$\lambda_{ii}(q) = \lambda(i,q)\big(\lambda(i,q) + a_{ii}\big)^{-1}, \text{ if } \nu(i,q) = 0; \quad \lambda_{ij}(q) = 0, \text{ if } \nu(i,q) > 0.$$

In this case the answer also does not depend on values $\nu(i,q)$, but it depends on the structure of the set where $\nu(i,q) > 0$.

6.5. Multiserver retrial queueing models

Consider a model with s identical servers and m waiting places for repeated calls. An arrival process is a Poisson process with rate λ. If an arriving primary call finds a server free, it immediately occupies a server and after completion of service leaves the system. Otherwise, if all servers are engaged, an arriving primary call goes into orbit and produces a source of repeated calls with rate ν_n until it finds a free server after one or more attempts. However, if there are no free waiting places, the primary call is lost. Again we assume that the service rate is μ_n for both primary and secondary calls.

The operation of the system can be described using the two-component process $(N_n(t), Q_n(t))$, where $N_n(t)$ is the number of busy servers and $Q_n(t)$ is the number of calls in the retrial queue at time t. Under the above assumptions this process is an MP with state space $S = \{0, 1, \ldots, s\} \times \{0, 1, \ldots, m\}$.

We study a multiserver retrial queueing model of the type $M/M/s/m$ and derive the expression for the parameter of a limiting exponential distribution for the time of the first loss of a call. Consider again two cases:

Case 1: $\mu_n = n\mu$ (fast service), $\nu_n = \nu$ (usual retrial rate), $n \to \infty$.

Case 2: $\mu_n = n\mu$ (fast service), $\nu_n = V_n\nu$ (large retrial rate), $n \to \infty$, $V_n \to \infty$.

Let $\Omega_n(j,q)$ be the time of the first loss of a call given as $Q_n(0) = q$ and $N_n(0) = j$.

THEOREM 6.7. *For the model described above under the assumption of fast service (case 1) if $\lambda > 0$, $\mu > 0$, $\nu > 0$, then for any initial state the distribution of the variable $n^{-s-m}\Omega_n(j,q)$ weakly converges to the exponential distribution*

$$\lim_{n\to\infty} \mathbf{P}\big\{n^{-s-m}\Omega_n(j,q) \geq t\big\} = \exp\{-\Lambda t\}, \quad t > 0,$$

where $\Lambda = \lambda \frac{\rho^{s+m}}{s!s^m}$, $\rho = \lambda/\mu$.

Proof. Consider a two-component process $(N_n(t), Q_n(t))$. It is a homogenous MP in continuous time with state space $Z = \{(j,q),\ j = \overline{0,s},\ q = \overline{0,m}\}$. We again consider an auxiliary MP $(N_n(t), \widehat{Q}_n(t))$, where $\widehat{Q}_n(t)$ is the number of retrial calls in the orbit for the system with an infinite buffer, and mention that $\Omega_n(j,q)$ is the exit time of this process from the subset Z. In each state (j,q) the process spends an exponential time with parameter

$$\Lambda_n(j,q) = \begin{cases} \lambda + sn\mu & \text{if } j = s; \\ \lambda + jn\mu + q\nu & \text{if } j \in \{0,1,2,\ldots,s-1\}. \end{cases}$$

We can easily calculate the transition probabilities for an embedded MP and verify that the subset Z forms a monotone structure with the following levels: $Z_0 = \{(0,0), (1,0)\}$, $Z_1 = \{(2,0)\}, \ldots$, $Z_{s-1} = \{(s,0)\}$, and $Z_{s-1+q} = \{(j,q),\ j = \overline{0,s}\}$, $q = 1,2,\ldots,m$.

Figure 6.5 illustrates the monotone structure and corresponding probabilities for a two-server model, where α_q, β_q, a_q, b_q and $\varepsilon_n(q)$ are defined as: $\alpha_q = q\nu(\lambda+q\nu)^{-1}$, $\beta_q = \lambda(\lambda+q\nu)^{-1}$, $a_q = q\rho/n \to 0$, $b_q = \rho/n \to 0$, $\varepsilon_n(q) = \rho/(2n) \to 0$.

Denote by $\pi_n(i,q)$, $i \in \{0,\ldots,s\}$, $q \in \{0,\ldots,m\}$, the stationary distribution of the embedded MP for $(N_n(t), \widehat{Q}_n(t))$. Consider first 0-level Z_0. As it forms in a limit one essential class, we denote $\pi_i = \lim_{n\to\infty} \pi_n(i,0)$, $i = 0,1$, and easily obtain that $\pi_0 = \pi_1 = 1/2$. Furthermore, we mention that the set $\{(i,0),\ i = 0,1,\ldots,s\}$ also forms a monotone structure and from representation (6.7) we obtain

$$\pi_n(i,0) = \frac{1}{2}\frac{1}{n^{i-1}}\frac{\rho^{i-1}}{(i-1)!}\big(1+o(1)\big), \quad i = 2,3,\ldots,s.$$

Denote now $\overline{\pi}_n(q) = (\pi_n(i,q),\ i = 0,1,\ldots,s)$ for $q = s, s{+}1, \ldots, s{+}m$. Taking into account the structure of transition probabilities (see Figure 6.5) and relation (6.9) we obtain recursively for any $q = 1,2,\ldots,m$,

$$\begin{aligned} &\pi_n(s,q) \approx \frac{\rho}{ns}\pi_n(s,q-1), \\ &\pi_n(s,q) \approx \pi_n(s-1,q) \approx \cdots \approx \pi_n(2,q), \\ &\pi_n(1,q) \approx \pi_n(2,q) + \pi_n(0,q)\frac{\lambda}{\lambda+q\nu}, \\ &\pi_n(0,q) \approx \pi_n(1,q). \end{aligned} \tag{6.26}$$

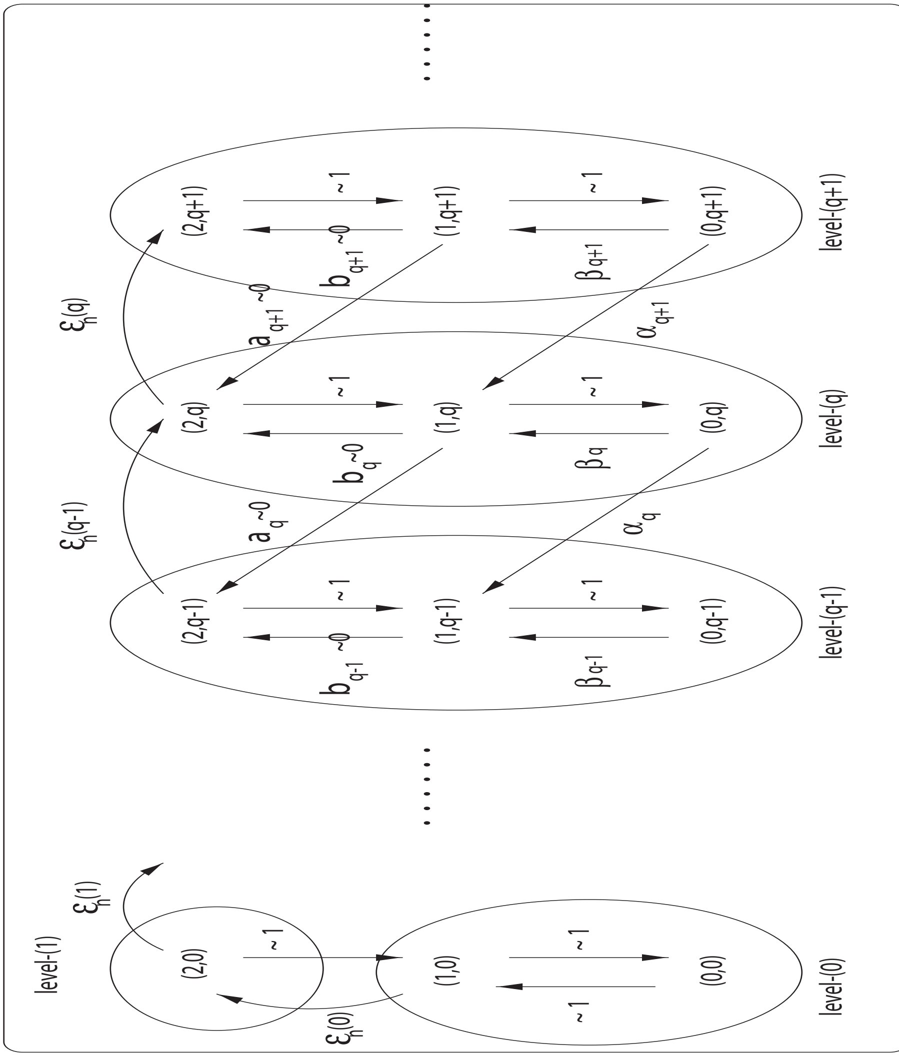

Figure 6.5. *Monotone structure for two-server model with fast service*

Finally, for any $q = 1, 2, \ldots, m$:

$$\pi_n(s,q) \approx \pi_n(s,0)\left(\frac{\rho}{ns}\right)^q = \frac{1}{2}\frac{1}{n^{s+q-1}}\frac{\rho^{s+q-1}}{s!s^{q-1}}\big(1+o(1)\big),$$

$$\pi_n(i,q) \approx \pi_n(s,q), \quad i = 2,\ldots,s-1, \tag{6.27}$$

$$\pi_n(0,q) \approx \pi_n(1,q) \approx \pi_n(s,q)\frac{\lambda+q\nu}{q\nu}.$$

As $g_n(Z) \approx \pi_n(s,m)\frac{\rho}{ns}$, we obtain

$$g_n(Z) = \frac{1}{2}\frac{1}{n^{s+m}}\frac{\rho^{s+m}}{s!s^m}\big(1+o(1)\big), \tag{6.28}$$

and setting $\beta_n = n^{-m-s}$, similar to the proof of Theorem 6.4 we obtain the parameter of exponential distribution as

$$\Lambda = \lambda\frac{\rho^{s+m}}{s!s^m}. \qquad \square$$

Now, consider the case where the service and retrial rates are large in the sense that $\mu_n = n\mu$ and $\nu_n = V_n\nu$ (case 2).

THEOREM 6.8. *As $\lambda > 0$, $\mu > 0$, $\nu > 0$, under the assumption of large service and retrial rates (case 2) for any initial state the distribution of the variable $n^{-s-m}\Omega_n(j,q)$ weakly converges to the exponential distribution with parameter $\Lambda = \lambda\frac{\rho^{s+m}}{s!s^m}$, where $\rho = \lambda/\mu$.*

Proof. The proof follows the same lines as in Theorem 6.7. We consider an auxiliary multicomponent process $(N_n(t), \widehat{Q}_n(t))$. Again it can be seen that the subset Z forms a monotone structure with the same levels as above for case 1. In each state (j,q) the process spends an exponential time with parameter

$$\Lambda(j,q) = \begin{cases} \lambda + sn\mu & \text{if } j = s, \\ \lambda + jn\mu + qn\nu & \text{if } j \in (0,1,\ldots,s-1). \end{cases}$$

In the same way we can prove relations (6.26), (6.27) and show that

$$\pi_n(s,q) \approx \pi_n(s-1,q) \approx \cdots \approx \pi_n(1,q) \approx \pi_n(0,q).$$

Finally we obtain expression (6.28) and the statement of Theorem 6.8 is true. $\square$

Note that the value of parameter Λ calculated in Theorem 6.8 is the same as the value of the parameter calculated in Theorem 6.7. In both cases the result does not depend on the retrial rate if $\nu_n \not\to 0$.

6.6. Bibliography

[ANI 70] ANISIMOV V., "Limit distributions of functionals of a semi-Markov process given on a fixed set of states up to the time of first exit", *Soviet Math. Dokl.*, vol. 11, no. 4, p. 1002–1004, 1970.

[ANI 73] ANISIMOV V., "Asymptotic consolidation of the states of random processes", *Cybernetics*, vol. 9, no. 3, p. 494–504, 1973.

[ANI 74] ANISIMOV V., "Limit theorems for sums of random variables in an array of sequences defined on a subset of states of a Markov chain up to the exit time", *Theor. Probab. and Math. Stat.*, , no. 4, p. 1–12, 1974.

[ANI 78] ANISIMOV V., "Limit theorems for switching processes and their applications", *Cybernetics*, vol. 14, no. 6, p. 917–929, 1978.

[ANI 83] ANISIMOV V., "Limit theorems for non-homogenous weakly dependent summation schemes", *Theor. Probab. and Math. Stat.*, vol. 27, p. 9–21, 1983.

[ANI 87] ANISIMOV V., ZAKUSILO O. and DONTCHENKO V., *The Elements of Queueing Theory and Asymptotic Analysis of Systems*, Visca Scola (Russian), Kiev, Ukraine, 1987.

[ANI 88a] ANISIMOV V., "Estimates for deviations of transient characteristics of non-homogenous Markov processes", *Ukrainian Math. J.*, vol. 40, no. 6, p. 588–592, 1988.

[ANI 88b] ANISIMOV V., *Random Processes with Discrete Component. Limit Theorems*, Kiev University (Russian), Kiev, Ukraine, 1988.

[ANI 89a] ANISIMOV V. and SZTRIK J., "Asymptotic analysis of some complex renewable system operating in random environment", *European Journal of Operations Research*, vol. 41, p. 162–168, 1989.

[ANI 89b] ANISIMOV V. and SZTRIK J., "Asymptotic analysis of some controlled finite-source queueing systems", *Acta Cybernetica*, vol. 9, no. 1, p. 27–38, 1989.

[ANI 89c] ANISIMOV V. and SZTRIK J., "Reliability analysis of a complex renewable system with fast repair", *J. of Information Processing and Cybernetics*, vol. 25, no. 11/12, p. 573–583, 1989.

[ANI 97] ANISIMOV V., "Asymptotic analysis of switching queueing systems in conditions of low and heavy loading", in CHAKRAVARTHY S. and ALFA A., Eds., *Matrix-Analytic Methods in Stochastic Models*, vol. 183 of *Lecture Notes in Pure and Appl. Math.*, p. 241–260, Dekker, New York, 1997.

[ANI 00a] ANISIMOV V., "Asymptotic analysis of reliability for switching systems in light and heavy traffic conditions", in LIMNIOS N. and NIKULIN M., Eds., *Recent Advances in Reliability Theory: Methodology, Practice, and Inference*, p. 119–133, Birkhäuser Boston, Massachusetts, 2000.

[ANI 00b] ANISIMOV V. and KURTULUS M., Asymptotic analysis of highly reliable Markov retrial queueing models, Report no. IEOR-2002, Dep. of Industrial Engineering, Bilkent Univ., Ankara, 2000, p. 1–28.

[ANI 01] ANISIMOV V. and KURTULUS M., "Some Markovian queueing retrial systems under light-traffic conditions", *Cybernetics and System Analysis*, vol. 37, no. 6, p. 876–886, 2001.

[GAN 98] GANTMACHER F., *The Theory of Matrices*, AMS, Rhode Island, 1998.

[KOR 69] KOROLYUK V., "The asymptotic behavior of the sojourn time of a semi-Markov process in a subset of the states", *Ukrainian Math. J.*, vol. 21, p. 705–707, 1969.

[KOR 93] KOROLYUK V. and TURBIN A., *Mathematical Foundation of the State Lumping of Large Systems*, Kluwer, Dordrecht, 1993.

[KOV 80] KOVALENKO I., *Rare Events Analysis in the Estimation of Systems Efficiency and Reliability*, Sov. Radio (Russian), Moscow, 1980.

[KOV 94] KOVALENKO I., "Rare events in queueing systems, a survey", *Queueing Systems*, vol. 16, p. 1–49, 1994.

[NEU 89] NEUTS M., *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York & Basel, 1989.

[SOL 71] SOLOVIEV A., "Asymptotic behavior of the first occurrence time of a rare event in a regenerative process", *Izvest. Akad. Nauk. SSSR Tekhn. Kibern.*, vol. 6, p. 79–89, 1971.

[SZT 91] SZTRIK J. and KOUVATSOS D., "Asymptotic analysis of a heterogeneous multiprocessor system in a randomly changing environment", *IEEE Transactions on Software Engineering*, vol. 17, no. 10, p. 1069–1075, 1991.

[SZT 92] SZTRIK J., "Asymptotic analysis of a heterogeneous renewable complex system with random environments", *Microelectronics and Reliability*, vol. 32, p. 975–986, 1992.

Chapter 7

Flows of Rare Events in Low and Heavy Traffic Conditions

7.1. Introduction

At the investigation of complex technical systems and computing networks various events related to failures, changes of the operational regime, exceeding some level, etc., usually have small probabilities (or rates) and may depend on the current state of the system. This implies a significance of the analysis of flows of rare events. Different asymptotic approaches to reliability analysis of various classes of stochastic systems are studied in Korolyuk and Turbin [KOR 93], Kovalenko [KOV 80], Anisimov *et al.* [ANI 87], and Anisimov [ANI 88]. An asymptotic analysis of wide classes of regenerative queueing models is considered by Soloviev [SOL 70, SOL 71]. A survey of results devoted to the analysis of rare events in queueing systems is given by Kovalenko [KOV 94].

In this chapter we consider the asymptotic analysis of flows of rare events switched by a random environment. The environment may be non-homogenous in time and not regenerative. In a case when the environment satisfies an asymptotically mixing condition, an approximation by non-homogenous Poisson flows is proved. The approximation of the time of the first jump by the exponential (in the homogenous case) or generalized exponential distribution is proved and the proximity estimates are considered. Therefore, the results of this chapter extend some of the results from section 3.2.1 on the behavior of flows of rare events with switching. Special attention is given to the case of finite MP with transition rates of different orders. The notions of S-set and "monotone" structure introduced in Chapter 6 are extended to the non-homogenous case. The asymptotic behavior of a flow of rare events defines on the Markov processes with a state space forming S-sets and the exit time from the S-set are studied.

In the homogenous case corresponding results for S-sets using another technique are obtained in [ANI 70, ANI 73, ANI 74]. In a case when the environment is a non-homogenous MP satisfying conditions of the asymptotic aggregation (merging) of a state space, an approximation of flows of rare events by Poisson flows with Markov switching is proved.

Note that the asymptotic aggregation of the state space of Markov and semi-Markov processes and algorithms of the sequential aggregation are studied in [ANI 70, ANI 73, ANI 74]. For an homogenous MP in continuous time with transition rates of different order similar results (the analysis of stationary and transient probabilities in case of the asymptotic aggregation of the state space) are obtained in [COU 77, BOB 86]. Using the technique of linear operators which are perturbed on the spectrum, the models of the asymptotic consolidation of homogenous MP and SMP are studied in [KOR 93]. In heavy traffic conditions an averaging principle with Poisson approximation for flows of rare events is proved. The method of proof is based on the results of [ANI 90, ANI 94, ANI 95] on AP and DA for processes with semi-Markov switching (see Chapter 4).

Applications to the reliability analysis of state-dependent Markov and semi-Markov queueing systems of the $M_{SM,Q}/M_{SM,Q}/m/k$ type in light and heavy traffic conditions are considered. The models of the asymptotic consolidation and the case of highly reliable servers are studied as well.

These results provide us with the approximative analytic approach in modeling of reliability characteristics of rather complex queueing models in transient and stable regimes under light and heavy traffic conditions.

7.2. Flows of rare events in systems with mixing

In various models the analysis of reliability is essentially related to analysis of flows of rare events on the trajectory of a system. In some cases rare events can appear only in some region of the state space and this region may be accessible with small probability.

Let for any $n > 0$, $x_n(t)$, $t \geq 0$, be a random process with state space X and $\{q_n(x,t),\ x \in X,\ t \geq 0\}$ be non-negative functions such that $q_n(x,t) \equiv 0$ as $x \notin Z_n$, where $Z_n \subset X$. Denote by $(x_n(t), \Pi_n(t))$, $t \geq 0$, a two-component process, where $\Pi_n(t)$ is a Poisson process switched by $x_n(t)$. This means that if at time t, $x_n(t) = x$, then the instantaneous rate of jump of $\Pi_n(t)$ is $q_n(x,t)$. Therefore, the process $\Pi_n(t)$ is a Poisson process with random rate $q_n(x_n(t),t)$ and belongs to the class of doubly stochastic Poisson processes or Cox processes [COX 80]. The rate $q_n(x,t)$ can also be interpreted as the rate of failure in state x at time t. Then $\Pi_n(t)$ is the total number of failures in the interval $[0,t]$.

Let us study the behavior of $\Pi_n(t)$. Suppose that $x_n(t)$ satisfies an asymptotically mixing condition in Z_n and introduce a strong mixing coefficient:

$$\begin{aligned}\varphi_n\big(u, Z_n\big) = \sup_{t\geq 0} \sup_{A_1, A_2 \subset Z_n} \big|&\mathbf{P}\big\{x_n(t) \in A_1, x_n(t+u) \in A_2\big\} \\ &- \mathbf{P}\big\{x_n(t) \in A_1\big\}\mathbf{P}\big\{x_n(t+u) \in A_2\big\}\big|.\end{aligned} \tag{7.1}$$

Put

$$\Lambda_n(t) = \mathbf{E} \int_0^{V_n t} q_n\big(x_n(v), v\big)\mathrm{d}v,$$

and let $\widehat{\Pi}_n(t)$ be a non-homogenous Poisson process with cumulative rate $\Lambda_n(t)$, i.e.,

$$\mathbf{E} \exp\big\{i\theta\widehat{\Pi}_n(t)\big\} = \exp\big\{\big(e^{i\theta} - 1\big)\Lambda_n(t)\big\}, \quad t > 0.$$

For a fixed $T > 0$, denote $q_n = \sup_{t\in[0,T]} \sup_{x\in Z_n} q_n(x, t)$.

THEOREM 7.1. *Assume that as $n \to \infty$, $V_n \to \infty$ and*

$$\alpha_n = V_n q_n^2 \int_0^{V_n T} \varphi_n\big(u, Z_n\big)\mathrm{d}u \longrightarrow 0 \tag{7.2}$$

then the finite-dimensional distributions of processes $\Pi_n(V_n t)$ and $\widehat{\Pi}_n(t)$ in the interval $[0, T]$ are asymptotically equivalent.

Proof. According to the construction of $\Pi_n(\cdot)$, for any $t > 0$,

$$\mathbf{E} \exp\big\{ - \theta\Pi_n\big(V_n t\big)\big\} = \mathbf{E} \exp\left\{\big(e^{-\theta} - 1\big) \int_0^{V_n t} q_n\big(x_n(v), v\big)\mathrm{d}v\right\}.$$

Taking into account the inequality

$$\left|\int_Y f(y)P(\mathrm{d}y) - \int_Y f(y)Q(\mathrm{d}y)\right| \leq \sup_{y\in Y} f(y) \sup_{A\in\mathcal{B}_Y} \big|P(A) - Q(A)\big| \tag{7.3}$$

valid for any non-negative bounded function $f(y)$ and any non-negative measures $P(\cdot), Q(\cdot)$ on Y, we easily obtain that as $u < v$,

$$\big|\mathbf{E}q_n\big(x_n(u), u\big)q_n\big(x_n(v), v\big) - \mathbf{E}q_n\big(x_n(u), u\big)\mathbf{E}q_n\big(x_n(v), v\big)\big| \leq q_n^2\varphi_n\big(v - u, Z_n\big).$$

Using the inequality $|e^{-a} - e^{-b}| \leq |a - b|$, $a, b > 0$, and condition (7.2), we obtain

$$\begin{aligned}&\big|\mathbf{E} \exp\big\{ - \theta\Pi_n\big(V_n t\big)\big\} - \exp\big\{\big(e^{-\theta} - 1\big)\Lambda_n(t)\big\}\big| \\ &\leq \big|e^{-\theta} - 1\big|^2\left(\mathbf{E}\left(\int_0^{V_n t} q_n\big(x_n(v), v\big)\mathrm{d}v - \Lambda_n(t)\right)^2\right)^{1/2} = O\big(\sqrt{\alpha}_n\big) \longrightarrow 0.\end{aligned}$$

This relation proves the equivalence of one-dimensional distributions of $\Pi_n(V_n t)$ and $\widehat{\Pi}_n(t)$. By analogy the equivalence of finite-dimensional distributions can be proved. □

NOTE 7.1. In particular if $q_n = O(1/V_n)$, then (7.2) is satisfied if there exists such a non-random sequence r_n, that, as $n \to \infty$,

$$V_n^{-1} r_n \longrightarrow 0, \quad \sup_{u \geq r_n} \varphi_n\big(u, Z_n\big) \longrightarrow 0. \tag{7.4}$$

NOTE 7.2. Let (7.2) hold and there exist a continuous function $\Lambda_0(t)$ such that for any $t \in [0, T]$,

$$\lim_{n\to\infty} \Lambda_n(t) = \Lambda_0(t).$$

Then the sequence of processes $\Pi_n(V_n t)$ J-converges in $[0, T]$ to the Poisson process $\Pi_0(t)$ with cumulative rate $\Lambda_0(t)$.

This means that the flows of rare events in systems with mixing can be approximated by Poisson processes with average cumulative rates. Denote now by ν_n the time of the first jump of $\Pi_n(t)$ (this can be interpreted as the time of the first failure). It is also possible to investigate the rate of convergence of ν_n to the generalized exponential distribution and obtain low and upper boundaries.

STATEMENT 7.1. *For any $t > 0$,*

$$\exp\big\{-\Lambda_n(t)\big\} \leq \mathbf{P}\big\{\nu_n > V_n t\big\} \leq \exp\big\{-\Lambda_n(t)\big\}\big(1 + Q\big(\Lambda_n(t)\big)\alpha_n(t)t\big), \tag{7.5}$$

where $Q(a) = (a^2/2)^{-1}(e^a - 1 - a)$ and $\alpha_n(t) = V_n q_n^2 \int_0^{V_n t} \varphi_n(u, Z_n)\mathrm{d}u$.

Proof. It is true that

$$\mathbf{P}\big\{\nu_n > V_n t\big\} = \mathbf{E}\exp\left\{-\int_0^{V_n t} q_n\big(x_n(v), v\big)\mathrm{d}v\right\}. \tag{7.6}$$

Note that $e^x \leq 1 + x + x^2/2$ as $x \leq 0$, and $e^x \leq 1 + x + x^2 Q(a)/2$ as $0 \leq x \leq a$. Since $e^x \geq 1 + x$ and $Q(a) \geq 1$ as $a > 0$, it follows that as $-\infty < x \leq a$,

$$1 + x \leq e^x \leq 1 + x + x^2 Q(a)/2. \tag{7.7}$$

Therefore, for any non-negative random variable ζ such that $\mathbf{E}\zeta = a$, and $\mathbf{Var}\zeta = \sigma^2$, we obtain from (7.7), by putting $x = a - \zeta$, taking expectation of both parts and multiplying by e^{-a}, that

$$e^{-a} \leq \mathbf{E}e^{-\zeta} \leq e^{-a}\big(1 + \sigma^2 Q(a)/2\big).$$

Using representation (7.6), considering $\zeta = \int_0^{V_n t} q_n(x_n(v), v)\mathrm{d}v$ and putting $a = \mathbf{E}\zeta = \Lambda_n(t)$ we obtain (7.5). Note that in this case, $\mathbf{Var}\zeta \leq 2\alpha_n(t)t$. □

If

$$\alpha_n(t)t/\Lambda_n(t)^2 \longrightarrow 0, \tag{7.8}$$

as $n \to \infty$, then relation (7.5) implies

$$\sup_{t\geq 0} \left|\mathbf{P}\{\nu_n > V_n t\} - \exp\{-\Lambda_n(t)\}\right| \longrightarrow 0. \tag{7.9}$$

Condition (7.8) is valid for wide classes of Markov models satisfying a strong mixing condition. In this case as usual for some $q < 1$, $\varphi_n(u) \leq Cq^u$, therefore relation (7.4) is true and the integral $\int_0^\infty \varphi_n(u)\mathrm{d}u < C < \infty$. Thus, in this case $\alpha_n(t) < CV_n q_n^2$.

Similar results can be obtained for discrete time models.

The main problem in applications is to evaluate the function $\Lambda_n(\cdot)$ and the coefficient $\varphi_n(u, Z_n)$. In the following sections we consider homogenous and non-homogenous finite Markov processes and show that $\Lambda_n(\cdot)$ can be replaced by an equivalent function calculated with the help of stationary or quasi-stationary distribution.

7.3. Asymptotically connected sets (V_n-S-sets)

Now consider the extension of the notion of the S-set given in section 6.2. We now introduce the important notion of V_n-S-set. Let $x_n(t)$, $t \geq 0$, be an MP in discrete or continuous time with finite state space $X = \{1, 2, \ldots, r\}$. Let $X_0 \subset X$ be a fixed subset.

DEFINITION 7.1. *The subset X_0 is called a V_n-S-set if as $n \to \infty$ for any $i \in X_0$,*

$$\mathbf{P}\{x_n(t) \in X_0 \textit{ for all } t \leq V_n \mid x_n(0) = i\} \longrightarrow 1,$$

and for any $i, j \in X_0$,

$$\mathbf{P}\{\textit{there exists } u,\ u < V_n \textit{ such that } x_n(u) = j \mid x_n(0) = i\} \longrightarrow 1.$$

This means that in the interval $[0, V_n]$ the process remains in X_0 with a probability close to one and all states in X_0 asymptotically communicate. In particular, the total state space X may form a V_n-S-set. In this case (7.4) is satisfied.

7.3.1. *Homogenous case*

Now consider discrete time and suppose that x_{nk}, $k \geq 0$, is a homogenous MP with finite state space X. Let $\{\chi_{nk}(i),\ i \in X\}$, $k \geq 0$, be the jointly independent families of rare indicators, i.e. $\mathbf{P}(\chi_{nk}(i) = 1) = 1 - \mathbf{P}(\chi_{nk}(i) = 0) = q_n(i)$, where $q_n(i) \to 0$, $i \in X$. Let

$$\Pi_n(V_n t) = \sum_{k=0}^{[V_n t]} \chi_{nk}(x_{nk}).$$

Suppose that X forms a V_n-S-set. Denote by $\pi_n(i)$, $i \in X$, the stationary distribution of x_{nk} which exists at large n given this assumption. Put $A_n = \sum_{i \in X} \pi_n(i) q_n(i)$. As a consequence of Theorem 7.1 we can prove the following.

STATEMENT 7.2. *If* $\limsup_{n\to\infty} V_n A_n < \infty$, *then the finite-dimensional distributions of the process* $\Pi_n(V_n t)$ *and the Poisson process with parameter* $V_n A_n$ *are asymptotically equivalent.*

Moreover, it can also be proved using the sequential algorithm of testing that if a set forms an V_n-S-set [ANI 70, ANI 74], then $V_n \pi_n(i) \to \infty$, $i \in X$.

These results also provide us with the possibility of studying the exit time from the subset as the exit time can be represented as the time of the first jump of the auxiliary stepwise process of the sum of indicators.

Let X_0 be a fixed subset of X. Let us recall Definition 6.1 of an S-set. Consider an auxiliary MP $\tilde{x}_{nk}$ with state space X_0 and matrix of transition probabilities $\tilde{P}_n(X_0) = \|p_n(i,j) p_n(i, X_0)^{-1}\|$, $i, j \in X_0$, where $p_n(i, X_0) = \sum_{l \in X_0} p_n(i, l)$. Denote by $\tilde{\pi}_n(i)$, $i \in X_0$, its stationary distribution (which exists at least at large enough n) and define the stationary probability of exit $g_n(X_0) = \sum_{i \in X_0} \tilde{\pi}_n(i)(1 - p_n(i, X_0))$ (see also section 6.2, relation (6.3)).

It is useful to know that if the subset X_0 is an S-set for the initial process x_{nk}, then it forms a $g_n(X_0)^{-1}$-S-set for the auxiliary MP $\tilde{x}_{nk}$. It is also always possible to find V_n such that $V_n g_n(X_0) \to 0$ and X_0 forms a V_n-S-set for x_{nk}.

Let us recall the notion of a monotone structure for the process in discrete time in section 6.2.3 and consider a similar definition for the process in continuous time. Let $x_n(t)$, $t \geq 0$, be an MP with finite state space Z which can be represented in the form: $Z = \{(i, s),\ i \in X_s,\ s = \overline{0, r}\}$, with transition rates $\mu_n((i, s), (j, q))$.

DEFINITION 7.2. *The state space* Z *is called a "monotone structure" of the order* r *if as* $n \to \infty$ *the following asymptotic relations hold:*

1. $\mu_n((i, s), (j, s+1)) = \varepsilon_n(s) a_{ij}(s)(1 + o(1))$, $i \in X_s$, $j \in X_{s+1}$, *where* $\varepsilon_n(s) \to 0$, $s = \overline{0, r-1}$*;*

2. $\mu_n((i, s), (j, s+k)) = 0$, $i \in X_s$, $j \in X_{s+k}$, $s = \overline{0, r-2}$, $k > 1$*;*

3. $\mu_n((i, s), (j, k)) = \mu_{ij}(s, k)(1 + o(1))$, $i \in X_s$, $j \in X_k$, $s = \overline{0, r}$, $k \leq s$*;*

4. for each $s = \overline{1,r}$ the matrix $G(s) - M(s)$ is invertible, where $G(s)$ is a diagonal matrix with elements $\mu_i^{(s)} = \sum_{m \le s,\ j \in X_m} \mu_{ij}(s,m)$, and $M(s) = \|\mu_{ij}(s,s)\|$, $i,j \in X_s$, $i \neq j$, where $\mu_{ii}(s,s) \equiv 0$, $i \in X_s$;

5. an auxiliary MP with state space $\{(i,0),\ i \in X_0\}$ and transition rates $\mu_{ij}(0,0)$ is irreducible with stationary distribution ρ_i, $i \in X_0$.

We call a subset of states $Z_q = \{(i,q),\ i \in X_q\}$ a q-level, $q = \overline{0,r}$.

Let $\overline{\rho}_n(s) = (\rho_n(i,s),\ i \in X_s)$, $s = \overline{0,m}$, and $\overline{\rho} = (\rho_i,\ i \in X_s)$ be the row-vectors, where $\rho_n(i,s)$ is the stationary probability of state (i,s). We put $\delta_n(s) = \prod_{j=0}^{s-1} \varepsilon_n(j)$.

THEOREM 7.2. *If Z forms a monotone structure, then for any $v_n \to \infty$ such that $v_n \delta_n(r)^{-1} \to \infty$ it also forms $v_n\delta_n(r)^{-1}$-S-set and for any $q = \overline{1,r}$ the following representation holds:*

$$\overline{\rho}_n(q) = \delta_n(q)\overline{\alpha}(q)\big(1 + o(1)\big),$$

where $\overline{\alpha}(q) = \overline{\rho} \prod_{j=0}^{q-1} A(j)(G(j+1) - M(j+1))^{-1}$, $A(s) = \|a_{ij}(s)\|$, $i,j \in X_s$, $\prod_{j=k}^{s} C(j) = C(k)C(k+1)\cdots C(s)$.

The proof is provided recursively to the order of the monotone structure and follows similar lines to the discrete case in section 6.2.3.

NOTE 7.3. If Z forms a monotone structure, then for any level q and for some $0 < a < 1$, $\varphi_n(u, Z_q) \le C\delta_n(q)a^u$.

Using Note 7.3 we can study flows of rare events of different orders defined on the monotone structure. Denote by $\Pi_n(t)$ a Poisson process switched by $x_n(t)$ in the following way: in state (i,s) the rate is $q_n(i,s)$, where

$$q_n(i,s) = q_n b_i(s) \prod_{j=s}^{r-1} \varepsilon_n(j)\big(1 + o(1)\big),$$

$b_i(s)$ are some given values and $q_n \to 0$ (we set $\prod_r^{r-1} = 1$). Put $V_n = (q_n\delta_n(r))^{-1}$.

STATEMENT 7.3. *If the state space Z forms a monotone structure, then $\Pi_n(V_n t)$ J-converges to the Poisson process with parameter $A = \sum_{s=0}^{r} (\overline{\alpha}(s), \overline{b}(s))$, where $\overline{b}(s)$ is a column vector with elements $b_i(s)$, $i \in X_s$ and the vector $\overline{\alpha}(s)$ is defined in Theorem 7.2.*

In particular, if Z is a subset of the state space and the rate of exit from state (i,s) is equal to $q_n(i,s)$, then using Statement 7.3 we can prove that for any initial state $(i_0, s_0) \in Z$ the distribution of variable $V_n^{-1}\Omega_n(i_0,s_0)$ weakly converges to the exponential distribution with parameter A, where $\Omega_n(i_0,s_0)$ is the exit time from Z starting from state (i_0,s_0).

7.3.2. *Non-homogenous case*

Let us now extend the notion of the monotone structure to the case where $x_n(t)$ is a non-homogenous MP. Suppose that $x_n(t)$ takes values in $Z = \{(i,s),\ i \in X_s,\ s = \overline{0,r}\}$ and transition rates at time t are $\mu_n((i,s),(j,q),t)$. Assume that there exists a normalizing factor V_n such that at each fixed t the rates $\mu_n((i,s),(j,q),V_n t)$ satisfy all items of Definition 7.2 where the values $a_{ij}(s) = a_{ij}(s,t)$ and $\mu_{ij}(s,k) = \mu_{ij}(s,k,t)$ depend on t. This means for instance that $\mu_n((i,s),(j,q),V_n t) = \varepsilon_n(s) a_{ij}(s,t)(1 + o(1))$ for item 1 in Definition 7.2 of the monotone structure. We denote corresponding matrices as $G(s,t)$ and $M(s,t)$. Let $\rho_i(t)$, $i \in X_0$, be the stationary distribution of the auxiliary MP with state space $\{(i,0),\ i \in X_0\}$ and transition rates $\mu_{ij}(0,0,t)$, $i \neq j$.

Suppose also that the following condition is satisfied: functions $a_{ij}(s,t)$ and $\mu_{ij}(s,k,t)$ are piecewise continuous in t and if $a_{ij}(s,1) > 0$, then for some $c_0 > 0$, $a_{ij}(s,t) \geq c_0$ in an interval $[0,T]$ (the same for $\mu_{ij}(s,k,t)$).

Then the set Z in the scale of time V_n forms an inhomogenous in time monotone structure. Denote

$$\overline{\alpha}(q,t) = \overline{\rho}(t) \prod_{j=0}^{q-1} A(j,t)\big(G(j+1,t) - M(j+1,t)\big)^{-1}.$$

THEOREM 7.3. *If Z forms an inhomogenous in time monotone structure, then for any $q = \overline{1,r}$, $0 < t < T$,*

$$\begin{aligned} &\mathbf{P}\big(x_n\big(V_n t\big) = (i,q),\ i \in X_q\big) = \delta_n(q)\overline{\alpha}(q,t)(1 + o(1)), \\ &\mathbf{P}\big(x_n\big(V_n t\big) = (i,0)\big) = \rho_i(t)\big(1 + o(1)\big), \quad i \in X_0. \end{aligned} \tag{7.10}$$

Note that the right-hand side in (7.10) stands for quasi-stationary probabilities.

Using this result we can study the behavior of inhomogenous flows of rare events by analogy to Statement 7.3. Assume that the family of non-negative functions $q_n(i,s,t)$, $(i,s) \in Z$, $t \geq 0$, is given. Let $\Pi_n(t)$ be a Poisson type process switched by a process $x_n(t)$: at time t the instantaneous rate of jump if $q_n(x_n(t),s,t)$. Also let

$$q_n\big(i,s,V_n t\big) = q_n b_i(s,t) \prod_{j=s}^{r-1} \varepsilon_n(j),$$

where $b_i(s,t)$ are given values and $V_n = (q_n \delta_n(r))^{-1} \to \infty$.

STATEMENT 7.4. *If Z forms a monotone structure, then $\Pi_n(V_n t)$ converges to a Poisson process with instantaneous rate $\widehat{\lambda}(t) = \sum_{s=0}^{r}(\overline{\alpha}(s,t), \overline{b}(s,t))$, where $\overline{b}(s,t)$ is a vector with elements $b_i(s,t)$, $i \in X_s$.*

In particular, if Z is a subset of the state space and the rate of exit from state (i,s) at time t is equal to $q_n(i,s,t)$, then

$$\mathbf{P}\big\{V_n^{-1}\Omega_n(i,s) > t\big\} \longrightarrow \exp\bigg\{-\int_0^t \widehat{\lambda}(u)\mathrm{d}u\bigg\}. \qquad (7.11)$$

7.4. Heavy traffic conditions

In heavy traffic conditions the trajectory of a system in most cases is non-stable as the value of the queue usually goes to infinity. We consider the behavior of a flow of rare events on the trajectory of a switching type system, which satisfies the averaging principle.

Let for each $n > 0$, $\mathcal{F}_{nk} = \{\zeta_{nk}(t,x,z),\ t \geq 0,\ x \in X,\ z \in \mathcal{R}^r\}$, $k \geq 0$, be jointly independent families of random processes in D_∞^r, $x_n(t)$, $t \geq 0$, be an independent of $\mathcal{F}_{nk}$ SMP in X which plays the role of a switching environment and S_{n0} be the initial value. Denote by $0 = t_{n0} < t_{n1} < \cdots$ the times of sequential jumps of $x_n(\cdot)$, and put $x_{nk} = x_n(t_{nk})$, $k \geq 0$. We construct a process with semi-Markov switching (PSMS) in the following way. Consider a sequence $S_{nk+1} = S_{nk} + \xi_{nk}$, where $\xi_{nk} = \zeta_{nk}(\tau_{nk}, x_{nk}, S_{nk})$, $\tau_{nk} = t_{nk+1} - t_{nk}$, and let

$$\zeta_n(t) = S_{nk} + \zeta_{nk}\big(t - t_{nk}, x_{nk}, S_{nk}\big) \quad \text{as } t_{nk} \leq t < t_{nk+1},\ t \geq 0.$$

Then the process $(x_n(t), \zeta_n(t))$, $t \geq 0$, is a PSMS (see section 1.2.5).

Let $q_n(x,z)$, $x \in X$, $z \in \mathcal{R}^r$, be a non-negative function. We construct a Poisson type process $\Pi_n(t)$ switched by $(x_n(t), \zeta_n(t))$ in the following way: if at time t, $x_n(t) = x$, $\zeta_n(t) = z$, then the instantaneous rate of jump of $\Pi_n(t)$ is $q_n(x,z)$.

Consider for simplicity the homogenous case (the distributions of the processes $\zeta_{nk}(\cdot)$ do not depend on the index $k \geq 0$). Let $\tau_n(x)$ be the sojourn time in state x for the process $x_n(\cdot)$. Denote $\xi_n(x,z) = \zeta_{n1}(\tau_n(x), x, z)$, and

$$g_n(x,z) = \sup\big\{|\zeta_{n1}(t,x,z)| : t < \tau_n(x)\big\}, \quad x \in X,\ z \in \mathcal{R}^r.$$

Suppose that the embedded MP x_{nk}, $k \geq 0$, has at each $n \geq 0$ the stationary measure $\pi_n(A)$, $A \in \mathcal{B}_X$, and put $m_n(x) = \mathbf{E}\tau_n(x)$, $b_n(x,z) = \mathbf{E}\xi_n(x,nz)$,

$$m_n = \int_X m_n(x)\pi_n(\mathrm{d}x), \quad b_n(z) = \int_X b_n(x,z)\pi_n(\mathrm{d}x),$$

$$q_n(z) = \int_X q_n(x,nz)\pi_n(\mathrm{d}x).$$

THEOREM 7.4. *Suppose that* $n^{-1}S_{n0} \xrightarrow{\text{P}} s_0$, *there exists a sequence of integers* r_n *such that* $n^{-1}r_n \to 0$, $\sup_{k\ge r_n}\varphi_n(k,X) \to 0$, *where* $\varphi_n(k,X)$ *is a strong mixing coefficient for* x_{nk} *(see (7.1)), for any* $N>0$, $\varepsilon>0$,

$$\lim_{n\to\infty}\sup_{|z|\le N}\sup_x n\mathbf{P}\big\{n^{-1}g_n(x,nz)>\epsilon\big\}=0,$$

$$\limsup_{n\to\infty}\sup_{|z|\le N}\sup_x\big\{\mathbf{E}\tau_{n1}(x)\chi\big(\tau_{n1}(x)>L\big)+\mathbf{E}\big|\xi_{n1}(x,nz)\big|\chi\big(\big|\xi(x,nz)\big|>L\big)\big\}\longrightarrow 0,$$

as $L\to\infty$, *for any* x *as* $\max(|z_1|,|z_2|)\le N$,

$$\big|b_n\big(x,z_1\big)-b_n\big(x,z_2\big)\big|\le C_N\big|z_1-z_2\big|+\alpha_n(N),$$

$$\big|nq_n\big(x,nz_1\big)-nq_n\big(x,nz_2\big)\big|\le C_N\big|z_1-z_2\big|+\alpha_n(N),$$

where C_N *are some constants,* $\alpha_n(N)\to 0$ *uniformly in* $|z_1|\le N$, $|z_2|\le N$, *also there exist the functions* $b(z), q(z)$ *and a constant* m *such that for any* $z\in\mathcal{R}^r$,

$$b_n(z)\longrightarrow b(z),\quad nq_n(z)\longrightarrow q(z),\quad m_n\longrightarrow m>0,$$

and $b(z)$ *has no more than linear growth.*

Then the sequence $(n^{-1}\zeta_n(nt),\Pi_n(nt))$ *J-converges in* $[0,T]$ *to the process* $(s(t),\Pi_0(t))$, *where the function* $s(t)$ *satisfies a differential equation*

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

and $\Pi_0(t)$ *is a non-homogenous Poisson process with local rate at time* t, $q(s(t))$.

The proof is essentially based on averaging principle type theorems for processes with semi-Markov switching (see section 4.6).

Going forward we mention that if the state space of $x_n(t)$ satisfies conditions of the asymptotic aggregation of states (see Chapter 8, section 8.4.4, (8.51)), then the limiting process can be represented as the process $(s(t),\Pi_0(t))$ switched by MP $y(t)$ (see Theorem 8.8).

7.5. Flows of rare events in queueing models

In this section, as the applications of the results above, we consider the flows of rare events for the basic state-dependent queueing systems with Markov or semi-Markov switching in light and heavy traffic conditions.

7.5.1. *Light traffic analysis in models with finite capacity*

Consider a non-homogenous state-dependent Markov type queueing model $M_{M,Q}/M_{M,Q}/m/k$ switched by an external Markov environment. This model is similar to the one introduced in section 2.2.1.1 in the homogenous case. Let $x_n(t)$,

$t \geq 0$, be a non-homogenous MP with values in $X = \{1, 2, \ldots, r\}$ and transition rates $c_n(i, j, t)$, $i, j \in X$, $i \neq j$, $t \geq 0$. Let the family of non-negative functions $\{\lambda_n(i, t, q),\ \mu_n(i, t, q),\ q \geq 0,\ i \in X,\ t \geq 0\}$ be given. There are m servers and k waiting places. Denote by $Q_n(t)$ the number of calls in the system at time t. The system operates in the following way. Calls enter the system one at a time. If at time t, $x_n(t) = i$ and $Q_n(t) = q$, then the instantaneous input rate is $\lambda_n(i, t, q)$ and the incoming call takes an idle server if there is one. If all servers are busy and there are no more than $k + m - 1$ calls in the system, the call joins the queue. Otherwise, this call is lost. An instantaneous service rate for each busy server is $\mu_n(i, t, q)$.

Suppose that the system is in light traffic conditions and in the asymptotic sense the rates slowly depend on t, i.e.:

$$\begin{aligned} \lambda_n\big(i, V_n t, q\big) &= \varepsilon_n \lambda_0(i, t, q)\big(1 + o(1)\big), \quad i \in X, \\ \mu_n\big(i, V_n t, q\big) &= \mu_0(i, t, q)\big(1 + o(1)\big), \quad i \in X, \\ c_n\big(i, j, V_n t\big) &= c_{ij}(t)\big(1 + o(1)\big), \quad i, j \in X,\ i \neq j, \end{aligned} \tag{7.12}$$

where $\varepsilon_n \to 0$, $V_n = \varepsilon_n^{-k-m-1}$, functions $\lambda_0(i, t, q)$, $\mu_0(i, t, q)$, $c_{ij}(t)$ are continuous, and the values $o(1) \to 0$ uniformly in t in an interval $[0, T]$.

Consider at each fixed t an auxiliary homogenous MP $x^{(t)}(u)$ with transition rates $c_{ij}(t)$, $i, j \in X$, and suppose that at each t this process is ergodic with stationary distribution $\rho_i(t)$, $i \in X$. Let $A(t, q)$ and $G(q, t)$ be diagonal matrices with elements $\lambda_0(i, t, q)$ and $\min(q, m)\mu_0(i, t, q)$, respectively, $C(t) = \|c_{ij}(t)\|$, $i, j \in X$, $i \neq j$, where we assume $c_{ii}(t) = -\sum_{j \neq i} c_{ij}(t)$. Denote by $\overline{\rho}(t)$ and $\overline{1}$ row vectors with elements $\rho_i(t)$ and 1, respectively, and put

$$\widehat{\lambda}(t) = \overline{\rho}(t)\left(\prod_{q=0}^{m+k-1} A(q, t)\big(G(q + 1, t) - C(t)\big)^{-1}\right)A(m + k, t)\overline{1}.$$

Let $\Omega_n(i, s)$ be the time of the first loss of a call given that $x_n(0) = i$, $Q_n(0) = s$, and $Y_n(t)$ be the number of lost calls in the interval $[0, t]$.

Conditions (7.12) and the ergodicity of the process $x^{(t)}(u)$ imply that the state space of the system forms a monotone structure in the non-homogenous case. Moreover, as the rates $c_n(\cdot)$ satisfy conditions (7.12), then the process $x_n(t)$ is quasi-ergodic (see section 3.3 and papers [ANI 83, ANI 88]). Using the results of section 7.2 and also Theorem 7.3 and Statement 7.4 (see also [ANI 83]) we can prove the following.

STATEMENT 7.5. *If conditions (7.12) are satisfied, then relation (7.11) is true and the process $Y_n(V_n t)$ J-converges to the Poisson process with local rate $\widehat{\lambda}(t)$.*

In particular, if there is no Markov switching,

$$\widehat{\lambda}(t) = \left(\prod_{q=0}^{m+k-1} \lambda_0(q,t)\mu_0(q+1,t)^{-1} \right) \lambda_0(m+k,t).$$

NOTE 7.4. If $x_n(t)$ satisfies the conditions of asymptotic consolidation in the scale of time V_n (see section 8.4.4, (8.51), (8.52)), then we can prove J-convergence of $Y_n(V_n t)$ to a Poison process $\Pi(t)$ switched by limiting aggregated MP. Correspondingly, $V_n^{-1}\Omega_n(i,s)$ weakly converges to the time of the first jump of $\Pi(t)$. Some results in this direction are proved in Chapter 8.

7.5.2. *Heavy traffic analysis*

Now consider the system $M_{SM,Q}/M_{SM,Q}/1/\infty$ in heavy traffic conditions discussed in section 5.3.3. For simplicity we study the homogenous case and suppose that the parameters of the model do not depend on n. The system is switched by a homogenous SMP $x(t)$, $t \geq 0$, with values in $X = \{1,2,\ldots,r\}$, where $x(t)$ stands for the external environment. There is one server and an infinite number of waiting places. Denote by $Q(t)$ the number of calls in the system at time t. If $x(t) = i$ and $Q(t)/n = z$, then input and service rates are $\lambda(i,z)$ and $\mu(i,z)$, respectively.

Suppose that the calls are impatient. This means that each call in the queue independently of others may get a refusal (be lost) with the local rate $n^{-1}q(x(t), n^{-1}Q(t))$, where $q(i,z)$ is a continuous function. Let $Y_n(t)$ be the number of lost calls in the interval $[0,t]$. Suppose that $x(t)$ is ergodic with stationary distribution ρ_i, $i \in X$. Denote

$$\widehat{\lambda}(z) = \sum_{i\in X} \lambda(i,z)\rho_i, \quad \widehat{\mu}(z) = \sum_{i\in X} \mu(i,z)\rho_i, \quad \widehat{q}(z) = \sum_{i\in X} q(i,z)\rho_i.$$

STATEMENT 7.6. *If $Q(0) = nq_0$, functions $\lambda(i,z), \mu(i,z), q(i,z)$ are locally Lipschitz with respect to z, and the function $\widehat{\lambda}(z) - \widehat{\mu}(z)$ has no more than linear growth, then $Y_n(nt)$ J-converges in $[0,T]$ to the Poisson process with local rate $\widehat{q}(s(t))$, where the function $s(\cdot)$ satisfies the following differential equation*

$$s(0) = q_0, \quad \mathrm{d}s(t) = \big(\widehat{\lambda}\big(s(t)\big) - \widehat{\mu}\big(s(t)\big)\big)\mathrm{d}t,$$

and T is any positive value such that $s(t) > 0$ in the interval $[0,T]$.

The proof uses the result of Theorem 7.4 and Statement 5.5, section 5.3.3 and is based on the convergence of the process $Q(nt)/n$ to $s(t)$.

Similar results can be proved for semi-Markov queueing networks.

7.6. Bibliography

[ANI 70] ANISIMOV V., "Limit distributions of functionals of a semi-Markov process given on a fixed set of states up to the time of first exit", *Soviet Math. Dokl.*, vol. 11, no. 4, p. 1002–1004, 1970.

[ANI 73] ANISIMOV V., "Asymptotic consolidation of the states of random processes", *Cybernetics*, vol. 9, no. 3, p. 494–504, 1973.

[ANI 74] ANISIMOV V., "Limit theorems for sums of random variables in an array of sequences defined on a subset of states of a Markov chain up to the exit time", *Theor. Probab. and Math. Stat.*, , no. 4, p. 1–12, 1974.

[ANI 83] ANISIMOV V., "Limit theorems for non-homogenous weakly dependent summation schemes", *Theor. Probab. and Math. Stat.*, vol. 27, p. 9–21, 1983.

[ANI 87] ANISIMOV V., ZAKUSILO O. and DONTCHENKO V., *The Elements of Queueing Theory and Asymptotic Analysis of Systems*, Visca Scola (Russian), Kiev, Ukraine, 1987.

[ANI 88] ANISIMOV V., *Random Processes with Discrete Component. Limit Theorems*, Kiev University (Russian), Kiev, Ukraine, 1988.

[ANI 90] ANISIMOV V. and ALIEV A., "Limit theorems for recurrent processes of semi-Markov type", *Theor. Prob. and Math. Stat.*, vol. 41, p. 7–13, 1990.

[ANI 94] ANISIMOV V., "Limit theorems for processes with semi-Markov switching and their applications", *Random Oper. and Stoch. Eqv.*, vol. 2, no. 4, p. 333–352, 1994.

[ANI 95] ANISIMOV V., "Switching processes: averaging principle, diffusion approximation and applications", *Acta Applicandae Mathematicae*, vol. 40, p. 95–141, 1995.

[BOB 86] BOBBIO A. and TRIVEDI K., "An Aggregation Technique for the Transient Analysis of Stiff Markov Chains", *IEEE Transactions on Computers*, vol. 35, no. 9, p. 803–814, 1986.

[COU 77] COURTOIS P., *Decomposability: Queueing and Computer Systems Applications*, Academic Press, New York, 1977.

[COX 80] COX D. and ISHAM V., *Point Processes*, Chapman & Hall, London, 1980.

[KOR 93] KOROLYUK V. and TURBIN A., *Mathematical Foundation of the State Lumping of Large Systems*, Kluwer, Dordrecht, 1993.

[KOV 80] KOVALENKO I., *Rare Events Analysis in the Estimation of Systems Efficiency and Reliability*, Sov. Radio (Russian), Moscow, 1980.

[KOV 94] KOVALENKO I., "Rare events in queueing systems, a survey", *Queueing Systems*, vol. 16, p. 1–49, 1994.

[SOL 70] SOLOVIEV A., "A redundancy with fast repair", *Izvest. Akad. Nauk. SSSR Tekhn. Kibern.*, vol. 1, p. 56–71, 1970.

[SOL 71] SOLOVIEV A., "Asymptotic behavior of the first occurrence time of a rare event in a regenerative process", *Izvest. Akad. Nauk. SSSR Tekhn. Kibern.*, vol. 6, p. 79–89, 1971.

Chapter 8

Asymptotic Aggregation of State Space

8.1. Introduction

Queueing models describing realistic computer networks usually have as usual rather quite complex structures and high dimensions. Therefore, during the investigation of such complex stochastic systems the following question is of a definite interest: to find the conditions when a complicated system can be approximated in some sense by a system with fewer states and a simpler structure, for example, by a Markov or semi-Markov system. We call this a problem of “asymptotic aggregation of state space” and “decreasing dimension”.

One of the main problems here is a possible loss of information which may occur after state aggregation and loss of Markovian property. For example, let us consider a homogenous Markov chain x_k with finite state space $I = \{1, 2, \dots, r\}$ and a matrix of one-step transition probabilities $P = \|p_{ij}\|_{i,j \in I}$. Assume a partition of the set I on the subsets I_k is given, where $\cup_k I_k = I$ and $I_k \cap I_m = \emptyset$ as $k \neq m$. Let us construct an aggregated process $\widehat{x}_k$, $k \geq 0$, as follows:

$$\widehat{x}_k = j \quad \text{if } x_k \in I_j,\ k \geq 0.$$

It is well-known [KEM 76] that $\widehat{x}_k$ is again a Markov process if and only if for any subsets I_i and I_j,

$$p\big(s, I_j\big) = p\big(q, I_j\big), \quad \text{for any } s, q \in I_i, \tag{8.1}$$

where $p(s, I_j) = \sum_{l \in I_j} p_{sl}$. In this case, if we denote for any $s \in I_i$, $p(I_i, I_j) = p(s, I_j)$, then the values $p(I_i, I_j)$ are the transition probabilities of the aggregated Markov chain $\widehat{x}_k$.

It is difficult to expect that condition (8.1) can be satisfied for many real systems. Therefore, the aggregated process realistically can only be close in some sense to a Markov process. However, in the analysis of highly reliable systems and systems with small probabilities of particular transitions, for example, small probabilities of failures or the loss of the call, we can come to the problem of asymptotic aggregation (merging or consolidation) of the state space, which can be formulated as follows.

Assume that the characteristics of the system depend on a small parameter ε in such a way that the state space can be subdivided on the subsets/regions such that the transition probabilities between regions tend to zero as $\varepsilon \to 0$. The problem is in the investigation of the conditions when the aggregated process obtained by coupling the states of each region into one state converges as $\varepsilon \to 0$ to a Markov or semi-Markov process, and also when different accumulative functionals also converge to corresponding functionals of the aggregated process. From a practical point of view this means that instead of the initial rather complicated system we can consider the system which is described by the limiting process and investigate its the characteristics which mimic with sufficient accuracy the characteristics of the initial system.

The results devoted to the problem of asymptotic aggregation can be conditionally divided in three directions. The first direction is the analytic approach based on the theory of the asymptotic expansions of the perturbed on the spectrum linear operators. This approach is being systematically developed by Korolyuk and co-authors [KOR 69, KOR 93, KOR 94, KOR 99, KOR 00, KOR 04, KOR 05] with applications to asymptotic analysis of time homogenous Markov and semi-Markov processes. A similar technique is also used by Yin and Zhang [YIN 03] and in [IL' 99, YIN 03]. A martingale technique is used in [KOR 94] (for finite MPs in [YIN 00]). Another approach based on direct probabilistic methods is developed by Kovalenko [KOV 75] where he obtained the proximity estimates of the aggregated processes to the Markov processes.

The author developed another approach based on the convergence of switching processes with rare switching [ANI 73, ANI 78, ANI 87a, ANI 88c, ANI 00a, ANI 00b, ANI 02, ANI 04, ANI 87b]. The advantage of this approach is that it allows us to study non-homogenous in time processes and even some classes of non-semi-Markov models for which analytic approaches are not yet developed.

Results on the weak convergence of aggregated processes for Markov and semi-Markov models are proved in [ANI 75, ANI 78] using constructive methods and limit theorems for SPs. The weak convergence of accumulative processes defined on asymptotically aggregated Markov or SMPs to two-component MPs with Markov or semi-Markov switching was first proved in [ANI 78] for a finite state space and then extended to a general state space and non-homogenous in time models in [ANI 88c].

Aggregation models for queueing systems in light and heavy traffic conditions are also studied in [ANI 98, ANI 00a]; the convergence of SPs in the asymptotic aggregation scheme to switched diffusion processes is studied in [ANI 00b].

8.2. Aggregation of finite Markov processes (stationary behavior)

In this section we consider the behavior of a stationary distribution for a finite MP in discrete and continuous time with asymptotically aggregated state space and prove using the matrix analytic technique that it can be approximated as a product of stationary probability of the aggregated state (region) of the limiting process and the stationary probability of a particular state inside the aggregated region.

8.2.1. *Discrete time*

In the previous sections we consider the dependence of the parameters of the system on n where $n \to \infty$. When we talk about small probabilities or low rates, sometimes it might be expedient to consider the dependence on a small parameter $\varepsilon \to 0$. In fact, this is an equivalent setting as, without loss of generality, we can always consider in a triangular scheme the dependence on a parameter $n \to \infty$, and consider a small parameter ε depending on n in such a way that $\varepsilon = \varepsilon_n \to 0$. In this section for better transparency of the model we consider the dependence on a small parameter $\varepsilon \to 0$ and omit for simplicity the dependence $\varepsilon = \varepsilon_n$.

Let at each $\varepsilon > 0$, $x_\varepsilon(k)$, $k \geq 0$, be a homogenous MP in discrete time with a finite state space $X = \{1, 2, \dots, r\}$ and the matrix of transition probabilities

$$P_\varepsilon = \left\|p_\varepsilon(i,j)\right\|, \quad i, j \in X. \tag{8.2}$$

Suppose that $P_\varepsilon = P_0 + \varepsilon B + o(\varepsilon)$, where the matrix P_0 is reducible with m classes of essential states $X_1, X_2, \dots, X_m$, and $\varepsilon \to 0$, i.e.,

$$P_0 = \begin{pmatrix} P^{(1)} & 0 & 0 & \cdots \\ 0 & P^{(2)} & 0 & \cdots \\ 0 & \cdots & \cdots & 0 \\ \cdots & 0 & 0 & P^{(m)} \end{pmatrix}, \tag{8.3}$$

each matrix $P^{(k)} = \| p_0(i,j) \|$, $i, j \in X_k$, is irreducible, and as $\varepsilon \to 0$,

$$p_\varepsilon(i,j) = p_0(i,j) + \varepsilon b_{ij} + o(\varepsilon), \quad i, j = 1, \dots, r, \tag{8.4}$$

where for all k, $p_0(i,j) = 0$ if $i \in X_k$, $j \notin X_k$. Denote by $\pi^{(k)}(i)$, $i \in X_k$, a stationary distribution for matrix $P^{(k)}$:

$$\pi^{(k)}(i) = \sum_{j \in X_k} \pi^{(k)}(j) p_0(j,i), \quad i \in X_k, \tag{8.5}$$

where $\sum_{i\in X_k}\pi^{(k)}(i)=1$ for any k, and put

$$A_{ks}=\sum_{i\in X_k}\pi^{(k)}(i)\sum_{j\in X_s}b_{ij},\quad k,s=1,\dots,m,\ k\neq s,$$
$$A_i=\sum_{l\neq i}A_{il}. \tag{8.6}$$

Let $y(t)$, $t\geq 0$, be a continuous time MP with state space $\{1,\dots,m\}$ and transition rates A_{ks}, $k\neq s$. Suppose that $y(t)$ is irreducible with stationary distribution ρ_i, $i=1,\dots,m$, i.e.

$$\rho_i A_i=\sum_{j\neq i}\rho_j A_{ji},\quad \sum_{i=1}^{m}\rho_i=1. \tag{8.7}$$

Denote by $\pi_\varepsilon(i)$, $i\in X$, a stationary distribution for matrix P_ε.

THEOREM 8.1. *At our assumptions $\pi_\varepsilon(i)$, $i\in X$, exists at rather small ε and for any $k=1,\dots,m$, as $i\in X_k$,*

$$\pi_\varepsilon(i)=\rho_k\cdot\pi^{(k)}(i)+O(\varepsilon),\quad i\in X_k,\ k=1,\dots,m. \tag{8.8}$$

NOTE 8.1. This result allows us to decrease the dimension of the initial MP at the calculation of the stationary distribution: instead of finding a solution of the system of linear equations of the rank r we can solve m systems of ranks r_k, $k=1,\dots,m$, respectively, (r_k is the number of states in X_k) and one system of rank m.

Proof. The values $\pi_\varepsilon(i)$, $i\in X$, satisfy the system of linear equations (8.5). Using (8.4) we can re-write it in the form: for any k and $i\in X_k$,

$$\pi_\varepsilon(i)=\sum_{j\in X_k}\pi_\varepsilon(j)p_\varepsilon(j,i)+\varepsilon\sum_{s\neq k}\sum_{j\in X_s}\pi_\varepsilon(j)b_{ji}+o(\varepsilon). \tag{8.9}$$

Now let us seek a solution for $\pi_\varepsilon(i)$, $i\in X_k$, in the form

$$\pi_\varepsilon(i)=c_k\pi^{(k)}(i)+\varepsilon z^{(k)}(i)+o(\varepsilon), \tag{8.10}$$

where $\sum_{k\in X}c_k=1$. It is clear that $\sum_k\sum_{i\in X_k}z^{(k)}(i)=0$. Combining together (8.9) and (8.10) we obtain

$$c_k\pi^{(k)}(i)+\varepsilon z^{(k)}(i)=\sum_{j\in X_k}\big(c_k\pi^{(k)}(j)\big(p_0(j,i)+\varepsilon b_{ji}\big)+\varepsilon z^{(k)}(j)p_0(j,i)\big)$$
$$+\,\varepsilon\sum_{s\neq k}\sum_{j\in X_s}c_s\pi^{(s)}(j)b_{ji}+o(\varepsilon). \tag{8.11}$$

Taking into account (8.5) we obtain from (8.11) the system of equations for values $z^{(k)}(i)$:

$$
\begin{aligned}
z^{(k)}(i) = c_k \sum_{j\in X_k} \pi^{(k)}(j) b_{ji} + \sum_{j\in X_k} z^{(k)}(j) p_0(j,i) \\
+ \sum_{s\neq k} \sum_{j\in X_s} c_s \pi^{(s)}(j) b_{ji}, \quad i \in X_k,\ k = 1, 2, \dots, m.
\end{aligned}
\tag{8.12}
$$

From (8.4) it follows that for any i, $\sum_{j\in X} b_{ij} = 0$. This means that

$$
-\sum_{j\in X_k} b_{ij} = \sum_{s\neq k} \sum_{l\in X_s} b_{il}. \tag{8.13}
$$

Taking a sum of both parts in (8.12) by $i \in X_k$ we obtain

$$
\begin{aligned}
\sum_{i\in X_k} z^{(k)}(i) = c_k \sum_{j\in X_k} \pi^{(k)}(j) \sum_{i\in X_k} b_{ji} + \sum_{j\in X_k} z^{(k)}(j) \\
+ \sum_{s\neq k} \sum_{j\in X_s} c_s \pi^{(s)}(j) \sum_{i\in X_k} b_{ji}.
\end{aligned}
\tag{8.14}
$$

Finally, using notation (8.6) and property (8.13), we obtain from (8.14):

$$
c_k A_k = \sum_{s\neq k} c_s A_{sk}, \quad k = 1, \dots, m. \tag{8.15}
$$

However, this system is equivalent to system (8.7) for the stationary distribution of MP $y(\cdot)$. As $\sum_k c_k = 1$, and the stationary distribution is unique, this implies $c_k = \rho_k$, $k = 1, \dots, m$, and finally proves (8.8). □

It is also possible to take the next step and find the 2nd terms $z^{(k)}(i)$ in (8.8) from system (8.12).

Note that the asymptotic expansions of the stationary distribution for homogenous MP with the arbitrary state space using the technique of linear operators perturbed on the spectrum obtained in [KOR 93].

8.2.2. *Hierarchic asymptotic aggregation*

These results can be extended to the hierarchic aggregation models (regions X_k in the limit can be reducible and also consist of several subregions). In this case it is possible to use the algorithm given in [ANI 70, ANI 73, ANI 74] and recursively calculate a stationary distribution in each region step by step providing the aggregation at each step.

Suppose that instead of (8.4) the following representation holds:

$$p_\varepsilon(i,j) = p_\varepsilon^{(1)}(i,j) + \varepsilon b_{ij} + o(\varepsilon), \tag{8.16}$$

where $p_\varepsilon^{(1)}(i,j) = 0$ if $i \in X_k$, $j \notin X_k$, $\sum_{l \in X_k} p_\varepsilon^{(1)}(i,l) = 1$, $i \in X_k$, and each region X_k can also be divided over subregions $X_s^{(k)}$, $s = 1, \dots, d_k$, such that

$$p_\varepsilon^{(1)}(i,j) = p_0(i,j) + \varepsilon_1 b_{ij}^{(1)} + o(\varepsilon_1),$$

where $p_0(i,j) = 0$ if $i \in X_s^{(k)}$, $j \notin X_s^{(k)}$, each subregion $X_s^{(k)}$ is irreducible and $\varepsilon/\varepsilon_1 \to 0$. This means that there is a fast transition process in each subregion $X_s^{(k)}$, a slow transition process between regions $X_s^{(k)}$, $s = 1, \dots, d_k$, inside each region $X^{(k)}$, and an even slower transition process between regions $X^{(k)}$, $k = 1, \dots, m$.

Denote by $\pi(i,s,k)$, $i \in X_s^{(k)}$, a stationary distribution for matrix

$$P(s,k) = \big\| p_0(i,j) \big\|, \quad i,j \in X_s^{(k)},$$

and put

$$A_{sq}^{(k)} = \sum_{i \in X_s^{(k)}} \pi(i,s,k) \sum_{j \in X_q^{(k)}} b_{ij}^{(1)},$$

$$A_s^{(k)} = \sum_{q \neq s} A_{sq}^{(k)}, \quad s,q = 1, \dots, d_k.$$

Let for any $k = 1, \dots, m$, $\rho_s^{(k)}$, $s = 1, \dots, d_k$, be the stationary distribution satisfying the system

$$\rho_s^{(k)} A_s^{(k)} = \sum_{q \neq s} \rho_q^{(k)} A_{qs}^{(k)}, \quad s = 1, \dots, d_k.$$

Denote

$$A_{kn} = \sum_{s=1}^{d_k} \rho_s^{(k)} \sum_{i \in X_s^{(k)}} \pi(i,s,k) \sum_{j \in X_n} b_{ij}, \quad k,n = 1, \dots, m,\ k \neq n,$$

$$A_k = \sum_{n \neq k} A_{kn},$$

and let ρ_k, $k = 1, \dots, m$, be the stationary distribution satisfying the system

$$\rho_k A_k = \sum_{n \neq k} \rho_n A_{nk}.$$

Then in a similar way we can prove the following asymptotic formula for the hierarchic models: if $i \in X_s^{(k)}$ then

$$\pi_\varepsilon(i) = \rho_k \rho_s^{(k)} \pi(i, s, k) + O\big(\varepsilon/\varepsilon_1\big). \tag{8.17}$$

This formula can be extended to any level of hierarchy in a similar way.

8.2.3. *Continuous time*

Similar results can be proved for an MP in continuous time. Let $x_\varepsilon(t)$, $t \geq 0$, be a homogenous MP with finite state space $X = \{1, \ldots, r\}$ and transition rates $a_\varepsilon(i, j)$. Suppose that

$$X = \bigcup_{k \in Y} X_k, \quad X_k \cap X_j = \emptyset \quad \text{as } k \neq j, \tag{8.18}$$

$$a_\varepsilon(i, j) = a_0(i, j) + \varepsilon b_{ij} + o(\varepsilon), \quad i, j \in X, \tag{8.19}$$

for any k, $a_0(i, j) = 0$ if $i \in X_k$, $j \notin X_k$, and an MP defined by transition rates $a_0(i, j)$ in each region X_k is irreducible. Denote $a_0(i) = \sum_{j \neq i} a_0(i, j)$. Let $\pi^{(k)}(i)$, $i \in X_k$, be a stationary distribution in region X_k:

$$\pi^{(k)}(i) a_0(i) = \sum_{j \in X_k,\, j \neq i} \pi^{(k)}(j) a_0(j, i), \tag{8.20}$$

and put

$$\widehat{a}_{ks} = \sum_{i \in X_k} \pi^{(k)}(i) \sum_{j \in X_s} b_{ij}, \quad k \neq s, \tag{8.21}$$

$$\widehat{a}_k = \sum_{s \neq k} \widehat{a}_{ks}. \tag{8.22}$$

Let $\rho_k, k = 1, \ldots, m$, be the stationary distribution of the aggregated MP $y(t)$ with transition rates $\widehat{a}_{ks}$ constructed on the coupled regions X_k:

$$\widehat{a}_k \rho_k = \sum_{s \neq k} \rho_s \widehat{a}_{ks}, \quad k = 1, \ldots, m. \tag{8.23}$$

This exists and is unique if $y(t)$ is irreducible. Let $\pi_\varepsilon(i)$, $i \in X$, be the stationary distribution for the initial MP:

$$\pi_\varepsilon(i) a_\varepsilon(i) = \sum_{j \neq i} \pi_\varepsilon(j) a_\varepsilon(j, i), \quad i \in X,$$

where

$$a_\varepsilon(i) = \sum_{j \neq i} a_\varepsilon(i, j).$$

THEOREM 8.2. *At our assumptions, for any $k = 1, \dots, m$, and $i \in X_k$,*

$$\pi_\varepsilon(i) = \rho_k \pi^{(k)}(i) + O(\varepsilon). \tag{8.24}$$

Proof. We search $\pi_\varepsilon(i)$ in the form:

$$\pi_\varepsilon(i) = c_k \pi^{(k)}(i) + \varepsilon z^{(k)}(i), \quad i \in X_k,$$

where $\sum_{k=1}^m c_k = 1$. Taking into account system (8.20) and using a similar method as in the proof of Theorem 8.1 we obtain a system of equations for variables c_k:

$$\widehat{a}_k c_k = \sum_{s \neq k} c_s \widehat{a}_{sk}.$$

However, this system is equivalent to system (8.23) for the stationary distribution of an MP $\widehat{y}(\cdot)$. As $\sum_k c_k = 1$, and the stationary distribution is unique, this implies $c_k = \rho_k$, $k = 1, \dots, m$, and finally proves (8.8). This implies the statement of Theorem 8.2. □

This result can also be extended to hierarchic models as in section 8.2.1. Note that the method of sequential aggregation of states of an MP in a triangular scheme was first proposed in [ANI 70, ANI 73, ANI 74].

8.3. Convergence of switching processes

In the previous sections we studied the asymptotic expansion of the stationary distribution for an MP satisfying the conditions of asymptotic aggregation of state space. Now we consider the convergence of random processes switched by an MP satisfying the conditions of asymptotic aggregation. The method of investigation is based on the limit theorems on the convergence in the class of SPs. For this purpose we prove a quite general theorem that establishes the conditions when a sequence of SPs aggregated by the first component J-converges to a regular limiting SP. In various applications a limiting process has a simpler structure and may depend on fewer parameters. Therefore this theorem provides us with a new approach to the problems of asymptotic decreasing dimension and aggregation of the state space in complex stochastic systems.

Let at each $n > 0$,

$$\mathcal{F}_{nk} = \big\{ \big(\zeta_{nk}(t, x, z), \tau_{nk}(x, z), \beta_{nk}(x, z)\big),\ t \geq 0,\ x \in X,\ z \in \mathcal{Z} \big\}, \quad k \geq 0,$$

be the jointly independent parametric families of random processes $\zeta_{nk}(t, x, z)$ with values in a space $\mathcal{Z}$ and random variables $(\tau_{nk}(x, z), \beta_{nk}(x, z))$ with values in $[0, \infty) \times X$, where X is some measurable set. We assume that $\mathcal{Z} \subset \mathcal{R}^r$ or $\mathcal{Z} \subset \{0, \pm 1, \pm 2, \dots\}$ (a discrete set). Let (x_{n0}, S_{n0}) be the initial value. These families define a sequence of SP $(\kappa_n(t), \zeta_n(t))$, $t \geq 0$, with values in $(X, \mathcal{Z})$ according to relations (1.3), (1.4). Let for simplicity Y be a discrete set, $Y = \{y_1, y_2, \ldots\}$, and $K(\cdot) : X \to Y$ be a map from X to Y. Denote $X_y = K^{-1}(y)$, $y \in Y$. Suppose that $(y_0(t), \zeta_0(t))$, $t \geq 0$, is a regular SP with values in $(Y, \mathcal{Z})$, which is defined by the families

$$\widetilde{F}_k = \bigl\{\bigl(\widetilde{\zeta}_k(t, y, z), \widetilde{\tau}_k(y, z), \widetilde{\beta}_k(y, z)\bigr),\ t \geq 0,\ y \in Y,\ z \in \mathcal{Z}\bigr\}, \quad k \geq 0,$$

and the initial value (y_0, S_0). We study the conditions of the convergence of the sequence of SP $(K(\kappa_n(\cdot)), \zeta_n(\cdot))$ which are aggregated in the first component to $(y_0(\cdot), \zeta_0(\cdot))$. This result is related to asymptotic averaging or aggregation of the states of the initial SP. For any $m \geq 1$ denote,

$$\begin{aligned} &\psi_{nk}\bigl(\lambda_1, \ldots, \lambda_m, t_1, \ldots, t_m, \theta, f(\cdot), x, z\bigr) \\ &\quad = \mathbf{E} \exp\left\{ i \sum_{l=1}^{m} \lambda_l \zeta_{nk}\bigl(t_l, x, z\bigr) - \theta \tau_{nk}(x, z) \right\} f\bigl(K\bigl(\beta_{nk}(x, z)\bigr)\bigr), \end{aligned} \tag{8.25}$$

where $i = +\sqrt{-1}$, $\lambda_l \in (-\infty, \infty)$, $l = \overline{1, m}$, $0 \leq t_1 < \cdots < t_m$, $x \in X$, $z \in \mathcal{Z}$, $\theta \geq 0$, $f(\cdot)$ is some bounded function in Y. Let the function $\widetilde{\psi}_k(\lambda_0, \ldots, \lambda_j, t_1, \ldots, t_j, \theta, f(\cdot), y, a)$ be determined by expression (8.25) for the families $\widetilde{F}_k$, $k \geq 0$, where the last factor in the right-hand side is $f(\widetilde{\beta}_k(y, z))$.

DEFINITION 8.1. *We say that the sequence of random processes $\zeta_n(\cdot)$ J-converges to the process $\zeta_0(\cdot)$ in $[0, \infty)$ if there exists a sequence of intervals $[0, T_m]$, $T_m \nearrow \infty$, such that for any $m > 0$, $\zeta_n(\cdot)$ J-converges to $\zeta_0(\cdot)$ in the interval $[0, T_m]$.*

The following theorem is a modification of Theorems 1,4 [ANI 78] that are directed at the applications in the asymptotic aggregation models; see also [ANI 88b].

THEOREM 8.3. *Let $(K(x_{n0}), S_{n0}) \stackrel{\mathrm{w}}{\Longrightarrow} (y_0, S_0)$ and the following conditions hold:*

1. there exists an everywhere dense set $D \subset [0, \infty)$ such that for any $m \geq 1$, $y \in Y$, $z \in \mathcal{Z}$, any deterministic sequences (x_n, z_n) such that $x_n \in X_y$, $z_n \to z$, and for any $k \geq 0$, $\lambda_1, \ldots, \lambda_m$, $t_1, \ldots, t_m \in D, \theta, f(\cdot)$,

$$\begin{aligned} &\lim_{n\to\infty} \psi_{nk}\bigl(\lambda_1, \ldots, \lambda_m, t_1, \ldots, t_j, \theta, f(\cdot), x_n, z_n\bigr) \\ &\qquad = \widetilde{\psi}_k\bigl(\lambda_1, \ldots, \lambda_m, t_1, \ldots, t_m, \theta, f(\cdot), y, z\bigr); \end{aligned}$$

2. for any $y \in Y$, $z \in \mathcal{Z}$, $k \geq 0$, the sequence of measures generated by the sequence of processes $\zeta_{nk}(\cdot, K(x_n), z_n)$ as $x_n \in X_y$, $z_n \to z$ is relatively compact in Skorokhod space $\mathcal{D}_\infty(\mathcal{Z})$;

3. for any $y \in Y$, $z \in \mathcal{Z}$, $k \geq 0$, the variable $\widetilde{\tau}_k(y, z)$ is almost sure (a.s.) the point of continuity in t of the process $\widetilde{\zeta}_k(t, y, z)$, and $\mathbf{P}\{\widetilde{\tau}_k(y, z) > 0\} = 1$.

Then the sequence $(K(\kappa_n(\cdot)), \zeta_n(\cdot))$*, as* $n \to \infty$*, J-converges in* $[0, \infty)$ *to an SP* $(y_0(\cdot), \zeta_0(\cdot))$.

Proof. Denote by $\{t_{nk}, x_{nk}, S_{nk},\ k \geq 0\}$ and $\{\widetilde{t}_k, \widetilde{y}_k, \widetilde{S}_k,\ k \geq 0\}$ the sequences defined according to relations (1.3), (1.4) using the families $\mathcal{F}_{nk},\ k \geq 0$, and $\widetilde{\mathcal{F}}_k$, $k \geq 0$, respectively. Using the definition of J-convergence [SKO 56, BIL 68] it is not hard to prove the following statement: if $f_n(t),\ t \in [0, T]$, is a sequence of deterministic functions J-convergent to $f_0(t),\ t \in [0, T]$, and a_n is a deterministic sequence such that $a_n \to a_0 \in (0, T)$, where a_0 is a point of continuity of $f_0(t)$, then $f_n(a_n) \to f_0(a_0)$. Let us now fix k, x, z. Then, using the Skorokhod method of a common probability space, we obtain, according to conditions 1-3, that the sequence of variables $\xi_{nk}(x_n, z_n) = \zeta_{nk}(\tau_{nk}(x_n, z_n), x_n, z_n)$ as $x_n \in X_y$, $z_n \to z$, weakly converges to $\widetilde{\xi}_k(y, z) = \widetilde{\zeta}_k(\widetilde{\tau}_k(y, z), y, z)$. Also, according to condition 1, for any $m \geq 1,\ k \geq 0$, the finite dimensional distributions of the vector $(\zeta_{nk}(t_l, x_n, z_n),\ l = \overline{1, m},\ \xi_{nk}(x_n, z_n), \tau_{nk}(x_n, z_n), K(\beta_{nk}(x_n, z_n)))$ weakly converge to the corresponding distributions of the vector $(\widetilde{\zeta}_k(t_l, y, z),\ l = \overline{1, m},$ $\widetilde{\xi}_k(y, z), \widetilde{\tau}_k(y, z), \widetilde{\beta}_k(y, z))$. Now, as we have the convergence of the initial values at $k = 0$, from recurrent relations (1.3), (1.4) it follows that for any $m \geq 1$ the multidimensional distributions of the vector $(t_{nk}, K(x_{nk}), S_{nk},\ k = 0, \ldots, m)$ weakly converge to the multidimensional distributions of the vector $(\widetilde{t}_k, \widetilde{y}_k, \widetilde{S}_k,\ k = 0, \ldots, m)$.

Now consider a deterministic function $\varphi_n(t) = f_n(t - t_n, x_n, z_n)\chi(t \geq t_n)$, where $f_n(u, x, z)$ is a sequence of functions such that $f_n(u, x_n, z_n)$ J-converges in u to $f_0(u, y, z)$ in $[0, \infty)$ as $K(x_n) = y, z_n \to z$. Let $t_n \to t_0$. Then, by the definition of J-convergence, $\varphi_n(t)$ converges to $\varphi_0(t) = f_0(t - t_0, y, z)\chi(t \geq t_0)$ for all $t > t_0$ that are the points of continuity of $\varphi_0(t)$. Again using Skorokhod method of a common probability space, we obtain according to conditions 1 and 2 that the finite dimensional distributions of process $\varphi_n(t) = \zeta_{nk}(t - t_{nk}, x_{nk}, S_{nk})\chi(K(x_{nk}) = y,$ $t_{nk} \leq t)$ weakly converge to corresponding distributions of process $\varphi_0(t) = \widetilde{\zeta}_k(t - \widetilde{t}_k,$ $\widetilde{y}_k, \widetilde{S}_k)\chi(\widetilde{y}_k = y,\ \widetilde{t}_k \leq t)$ for all t that are a.s. the points of continuity of $\varphi_0(t)$. Now we can use the relation:

$$
\begin{aligned}
&\bigl|\mathbf{E}e^{i\lambda\zeta_n(t)}\chi\bigl(K\bigl(\kappa_n(t)\bigr) = y\bigr) - \mathbf{E}e^{i\lambda\widetilde{\zeta}_0(t)}\chi\bigl(\widetilde{y}_0(t) = y\bigr)\bigr| \\
&\quad \leq \sum_{k=0}^{N} \bigl|\mathbf{E}e^{i\lambda\zeta_n(t)}\chi\bigl(K\bigl(\kappa_n(t)\bigr) = y,\ t_{nk} \leq t < t_{n,k+1}\bigr) \\
&\qquad\qquad - \mathbf{E}e^{i\lambda\widetilde{\zeta}_0(t)}\chi\bigl(\widetilde{y}_0(t) = y,\ \widetilde{t}_k \leq t < \widetilde{t}_{k+1}\bigr)\bigr| \\
&\qquad + \mathbf{P}\bigl(t_{n,N+1} \leq t\bigr) + \mathbf{P}\bigl(\widetilde{t}_{N+1} \leq t\bigr),
\end{aligned}
\tag{8.26}
$$

Let $t > 0$ be the point of continuity of $(y_0(\cdot), \zeta_0(\cdot))$. Then at any fixed N the first sum in the right-hand side of (8.26) goes to zero as $n \to \infty$. Now, as

$y_0(\cdot)$ is regular, then $\mathbf{P}(\tilde{t}_{N+1} \le t) \to 0$ as $N \to \infty$, and also at any fixed N, $\mathbf{P}(t_{n,N+1} \le t) \approx \mathbf{P}(\tilde{t}_{N+1} \le t)$ as $n \to \infty$. Therefore, at large enough N the right-hand side of (8.26) is small as $n \to \infty$, which implies the convergence of one-dimensional distributions. In the same way we prove the convergence of finite dimensional distributions.

To prove J-convergence we again use the Skorokhod method of a common probability space. As the convergence of finite dimensional distributions holds, we can construct the families $\tilde{\tilde{\mathcal{F}}}_{nk}(\omega)$ and $\tilde{\tilde{\mathcal{F}}}_k(\omega)$ on the same probability space Ω such that as $t \in D$, $x_n \in X_y$, $z_n \to z$, sequence $(\tilde{\tilde{\zeta}}_{nk}(t, x_n, z_n, \omega), \tilde{\tilde{\tau}}_{nk}(x_n, z_n, \omega), \tilde{\tilde{\beta}}_{nk}(x_n, z_n, \omega))$, converges to $(\tilde{\tilde{y}}_k(t, y, z, \omega), \tilde{\tilde{\tau}}_k(y, z, \omega), \tilde{\tilde{\beta}}_k(y, z, \omega))$, and conditions 2 and 3 are satisfied for all $\omega \in \Omega' \subset \Omega$, where $\mathbf{P}(\Omega') = 1$. If $(\tilde{\tilde{\kappa}}_n(\cdot, \omega), \tilde{\tilde{\zeta}}_n(\cdot, \omega))$ and $(\tilde{\tilde{y}}_0(\cdot, \omega), \tilde{\tilde{\zeta}}_0(\cdot, \omega))$ are SPs constructed by introduced families, then the sequence $(K(\tilde{\tilde{\kappa}}_n(t, \omega)), \tilde{\tilde{\zeta}}_n(t, \omega))$ converges to $(\tilde{\tilde{y}}_0(t, \omega), \tilde{\tilde{\zeta}}_0(t, \omega))$ for all $t \in D$, $\omega \in \Omega'$. Let us choose some $\omega' \in \Omega'$. According to condition 3, $\tilde{\tilde{t}}_0(\omega') < \tilde{\tilde{t}}_1(\omega') < \cdots$. This means that $K(\tilde{\tilde{\kappa}}_n(\cdot, \omega'))$ also J-converges to $K(\tilde{\tilde{y}}_0(\cdot, \omega'))$. Consider the sequence $\tilde{\tilde{\zeta}}_n(\cdot, \omega')$. For any $k \ge 0$ according to condition 2, $\tilde{\tilde{\zeta}}_n(\cdot, \omega')$ J-converges to $\tilde{\tilde{\zeta}}_0(\cdot, \omega')$ in any interval $[\alpha, \beta] \subset (t_k(\omega'), t_{k+1}(\omega'))$. Let us consider the behavior of $\tilde{\tilde{\zeta}}_n(\cdot, \omega')$ in the neighborhood of the point $\tilde{\tilde{t}}_k(\omega')$. Using condition 3 and the fact that the function $\tilde{\tilde{\zeta}}_k(t, y, z, \omega')$ is right continuous at $t = 0$ and $\tilde{\tilde{\zeta}}_{nk}(\cdot, x_{nk}(\omega'), z, \omega')$ J-converges to $\tilde{\tilde{\zeta}}_k(\cdot, y, z, \omega')$ in any interval $[0, \varepsilon]$ such that ε is a point of continuity of $\tilde{\tilde{\zeta}}_k(\cdot, y, z, \omega')$, the values $\tilde{\tilde{\zeta}}_n(t, \omega')$, as $\tilde{\tilde{t}}_{nk}(\omega') - \varepsilon < t < \tilde{\tilde{t}}_{nk}(\omega')$, are close to $\tilde{\tilde{\zeta}}_n(t_{nk}(\omega') - 0, \omega')$. The values $\tilde{\tilde{\zeta}}_n(t, \omega')$, as $\tilde{\tilde{t}}_{nk}(\omega') \le t < \tilde{\tilde{t}}_{nk}(\omega') + \varepsilon$, are close to $\tilde{\tilde{\zeta}}_n(t_{nk}(\omega'), \omega')$. This means, by definition of convergence in J-topology, that the modulus of continuity in J-topology of the function $\tilde{\tilde{\zeta}}_n(t, \omega')$ in the neighborhood of each point $\tilde{\tilde{t}}_{nk}(\omega')$ is small. This implies the relative compactness of $\tilde{\tilde{\zeta}}_n(\cdot, \omega')$ concerning J-convergence and finally proves Theorem 8.3. □

Note that if $\zeta_0(\cdot)$ is stochastically continuous, then in condition 3 we need to verify only relation $\mathbf{P}\{\tilde{\tau}_k(y, z) > 0\} = 1$.

8.4. Aggregation of states in Markov models

Let us consider the applications of Theorem 8.3 to the models of asymptotic aggregation (merging, enlargement) of the state space for hierarchic in time Markov systems. We prove that if the state space of the process can be divided in the regions such that transition probabilities between them are small in some sense, then under rather general conditions the aggregated process obtained by coupling each region into one state can be approximated by an MP constructed on the aggregated states and correspondingly the accumulative processes can be approximated by processes with independent increments with Markov or semi-Markov switching.

8.4.1. *Convergence of the aggregated process to a Markov process (finite state space)*

In section 8.2 we investigated the asymptotic behavior of the stationary distribution of the MP in the asymptotic aggregation setting. Now we consider the weak convergence of the aggregated process to an MP defined on the aggregated state space. As it follows from section 6.2.2, Theorem 6.1 and Corollaries 6.1, 6.2, the asymptotic behavior of the first exit time from the S-set does not depend on the initial state. This provides the opportunity of studying the convergence of the aggregated (consolidated) processes to Markov and semi-Markov processes [ANI 73, ANI 78, ANI 87a, ANI 88c, ANI 87b, ANI 02, ANI 04].

Consider first for the illustration a simpler case of a homogenous MP in discrete time and with a finite state space. We consider again a triangular scheme and assume that the MP depends on a small parameter $\varepsilon_n \to 0$ as $n \to \infty$. Without loss of generality we can assume that $\varepsilon_n = 1/n$. Let at each n, x_{nk}, $k \geq 0$ be a homogenous MP in discrete time with a finite state space $X = \{1, 2, \ldots, r\}$ and the matrix of transition probabilities $P_n = \|p_n(i,j)\|$, $i, j \in X$. Suppose that the following representation holds:

$$X = \bigcup_{y \in Y} X_y, \quad \text{where } X_{y_1} \bigcap X_{y_2} = \emptyset \text{ as } y_1 \neq y_2, \tag{8.27}$$

and $P_n = P_0 + B/n + o(1/n)$, where matrix P_0 is reducible with d classes of essential states $X_1, X_2, \ldots, X_d$. Thus, relation (8.3) holds, where each matrix $P^{(y)} = \|p_0(i,j)\|$, $i, j \in X_y$, is irreducible, and

$$p_n(i,j) = p_0(i,j) + b_{ij}/n + o(1/n), \quad i, j = 1, \ldots, r, \tag{8.28}$$

where for all y, $p_0(i,j) = 0$ if $i \in X_y$, $j \notin X_y$. This means that X is subdivided in the non-intersected regions with small transition probabilities of the order $O(1/n)$ between them. We define the aggregated process $\widehat{x}_{nk}$ as follows: for any $k \geq 0$,

$$\widehat{x}_{nk} = y \quad \text{as } x_{nk} \in X_y, \ y \in Y. \tag{8.29}$$

If we define a map $K(\cdot)$ from X to Y such that:

$$K(x) = y \quad \text{for any } x \in X_y, \ y \in Y, \tag{8.30}$$

then $\widehat{x}_{nk} = K(x_{nk})$.

Let us keep the notation of section 8.2, denote by $\pi^{(y)}(i)$, $i \in X_y$, a stationary distribution for matrix $P^{(y)}$, and put

$$A_{ys} = \sum_{i \in X_y} \pi^{(y)}(i) \sum_{j \in X_s} b_{ij}, \quad y, s \in Y,\ y \neq s. \tag{8.31}$$

Let $y(t)$, $t \geq 0$, be a continuous time MP with state space Y and transition rates A_{ys}, $y \neq z$.

THEOREM 8.4. *Let $x_{n0} \in X_{y_0}$. Then, as $n \to \infty$, the sequence of processes $K(x_{n,[nt]})$ J-converges in any interval $[0,T]$ to an MP process $y(t)$ with the initial state $y(0) = y_0$.*

Proof. To prove this theorem we use Theorem 8.3. In this case the process $K(x_{nm})$ can be represented as an SP with state space Y where the switching times are the times of transitions between regions. As the exit time from the region is asymptotically approximated by the exponential distribution and does not depend on the initial state of the region, then the limiting process has the exponential sojourn times in the states and is thus an MP.

Let us provide formal proof. First, we represent $K(x_{nm})$ as an SP using the construction given in section 6.2.4. For each $y \in Y$, let $\widetilde{x}_n^{(y)}(k)$, $k \geq 0$, be an auxiliary MP with state space X_y and matrix of transition probabilities

$$\widetilde{P}_n(X_y) = \left\|\widetilde{p}_n^{(y)}(i,j)\right\|, \quad i,j \in X_y, \tag{8.32}$$

where

$$\begin{aligned} \widetilde{p}_n^{(y)}(i,j) &= p_n(i,j)p_n\big(i,\bar{X}_y\big)^{-1}, \quad i,j \in X_y, \\ p_n\big(i,\bar{X}_y\big) &= \sum_{s \notin X_y} p_n(i,s), \quad i \in X_y, \end{aligned} \tag{8.33}$$

($p_n(i,\bar{X}_y)$ is one-step probability to leave region X_y from state i).

For each $y \in Y$, let us define a family of jointly independent random indicators $\{\chi_n^{(y)}(i,k),\ i \in X_y,\ k \geq 0\}$ such that $\mathbf{P}(\chi_n^{(y)}(i,k) = 1) = 1 - \mathbf{P}(\chi_n^{(y)}(i,k) = 0) = p_n(i,\bar{X}_y)$, and construct the family of independent and identically distributed in index j Bernoulli processes $\{y_{nj}^{(y)}(i,m),\ i \in X_y,\ y \in Y,\ m = 0,1,2,\ldots\}$, $j = 0,1,2,\ldots$ defined on the trajectories of MPs $\widetilde{x}_n^{(y)}(k)$ in the following way: denote by $\widetilde{x}_n^{(y)}(i,k)$ process $\widetilde{x}_n^{(y)}(k)$ given that $\widetilde{x}_n^{(y)}(0) = i \in X_y$, and put

$$y_{n1}^{(y)}(i,m) = \sum_{k=0}^{m} \chi_n^{(y)}\big(\widetilde{x}_n^{(y)}(i,k),k\big), \quad m \geq 0. \tag{8.34}$$

Let $\widetilde{\nu}_{nj}(i, X_y)$ be the time of the first jump of $y_{nj}^{(y)}(i, m)$:

$$\widetilde{\nu}_{nj}\big(i, X_y\big) = \min\big\{m : m > 0,\ y_{nj}^{(y)}(i, m-1) = 1\big\}, \tag{8.35}$$

and let $\{\beta_{nk}(i, X_y),\ i \in X_y,\ k \geq 0\}$ be a family of independent in index k random variables with values in $X \backslash X_y$, where

$$\mathbf{P}\big(\beta_{nk}\big(i, X_y\big) = j\big) = p_n(i, j) p_n\big(i, \bar{X}_y\big)^{-1}, \quad j \notin X_y \tag{8.36}$$

(probability of a jump to state j at the time of exit from X_y).

Let us define an SP in discrete time $(\kappa_n(m), \zeta_n(m))$, $m = 0, 1, 2, \dots$ using the family of processes $\{y_{nj}^{(y)}(i, m)\}$ and the variables introduced above as follows: suppose that the initial value $x_n(0) = i_0 \in X_{y_0}$. Put $t_{n0} = 0$, $y_{n0} = y_0$, $i_{n0} = i_0$,

$$t_{nk+1} = t_{nk} + \widetilde{\nu}_{nk}\big(i_{nk}, X_{y_{nk}}\big), \quad \widetilde{i}_{nk} = \widetilde{x}_n^{(y_{nk})}\big(i_{nk}, \widetilde{\nu}_{nk}\big(i_{nk}, X_{y_{nk}}\big) - 1\big),$$

$$i_{nk+1} = \beta_{nk}\big(\widetilde{i}_{nk}, X_{y_{nk}}\big), \quad y_{nk} = K\big(i_{nk}\big),\ k \geq 0.$$

According to the results of section 6.2.4, variable $\widetilde{\nu}_{nj}(i, X_y)$ has the same distribution as the exit time from region X_y starting from state i. Therefore, the times t_{nk} are equivalent to the times of transitions between regions X_y, sequence i_{nk} shows the initial states in the regions (for example, i_{n1} is the initial state in the region $X_{y_{n1}}$ after the jump from region X_{y_0}), value $\widetilde{i}_{nk}$ shows the state just before exit from region $X_{y_{nk}}$, and y_{nk} is the sequence of the regions for the aggregated process. Denote

$$\kappa_n(m) = y_{nk}, \quad \zeta_n(m) = y_{nk}^{(y_{nk})}\big(i_{nk}, m - t_{nk}\big)$$
$$\text{as } t_{nk} \leq m < t_{nk+1},\ m = 0, 1, 2, \dots$$

Thus, by the definition, process $\kappa_n(m)$ is equivalent (in the sense of the equivalence of finite dimensional distributions) to the aggregated process $K(x_{nm})$.

Let us consider the limiting behavior of the introduced variables. At any y the state space X_y for the process $\widetilde{x}_n^{(y)}(k)$ forms in the limit one irreducible class. Therefore, using the results on the convergence of the weakly dependent stepwise processes (see Chapter 3) we see that as $n \to \infty$, for any $i \in X_y$ the process $y_{n1}^{(y)}(i, [nt])$ J-converges to a Poisson process with rate $A_y = \sum_{s \neq y} A_{ys}$. Thus, the variable $\widetilde{\nu}_{nj}(i, X_y)/n$ as the time of first jump of $y_{n1}^{(y)}(i, m)$ weakly converges to the exponential random variable with rate A_y. In a similar way we can prove that

$$\mathbf{P}\big(y_{n1} = z \mid y_{n0} = y,\ i_{n0} = i \in X_y\big) \longrightarrow A_{yz}/A_y, \quad z \neq y.$$

Therefore, according to Theorem 8.3 the sequence of SPs $(\kappa_n([nt]), \zeta_n([nt]))$ J-converges to an SP $(\kappa(t), \zeta(t))$, which is constructed by the family of Poisson processes $\Pi_y(t)$ with rates A_y and the family of variables $\beta(y)$, where

$$\mathbf{P}\big(\beta(y) = z\big) = A_{yz}/A_y, \quad z \neq y,$$

in the following way: if $\kappa(t) = y$, then a switching time in state y is the time of the first jump of the Poisson process $\Pi_y(t)$, and the next state z is chosen according to the value of $\beta(y)$. Then by definition the process $\kappa(t)$ is equivalent to an MP $y(t)$. □

This result can be extended using the same technique to the case when each subset X_j can form a V_n-s-set, where $V_n/n \to 0$. Assume that relation (8.27) holds and instead of relation (8.28) a more general representation is true:

$$p_n(i,l) = p_n^{(0)}(i,l) + n^{-1}h_n(i,l), \quad i,l = \overline{1,d}, \tag{8.37}$$

where $\limsup_{n\to\infty} \max_{i,l} |h_n(i,l)| < C$, and for any $j \in Y$, $p_n^{(0)}(i,l) \equiv 0$ at $i \in X_j$, $l \notin X_j$.

For any $j \in Y$ denote by $x_{nk}^{(j)}$, $k \geq 0$, an auxiliary MP with state space X_j and transition probabilities $p_n^{(0)}(i,l)$, $i,l \in X_j$. Consider the case when the subset X_j in the limit can be split over several classes of essential states. Introduce a uniformly strong mixing coefficient

$$\varphi_n^{(j)}(k) = \max_{i_1,i_2\in X_j,\ A\subset X_j} \left|\mathbf{P}\{x_{nk}^{(j)} \in A \mid x_{n0}^{(j)} = i_1\} - \mathbf{P}\{x_{nk}^{(j)} \in A \mid x_{n0}^{(j)} = i_2\}\right|.$$

Suppose that there exists a sequence of integers r_n and q, $0 \leq q < 1$, such that

$$n^{-1}r_n \longrightarrow 0, \quad \text{and for any } j \in Y, \quad \varphi_n^{(j)}(r_n) \leq q. \tag{8.38}$$

Note that condition (8.38) means that each subset X_j forms an n-s-set (see section 7.3 and [ANI 70, ANI 73, ANI 78]). In particular X_j may form a closed ergodic subset.

Denote by $\pi_n^{(j)}(i)$, $i \in X_j$, a stationary distribution for $x_{nk}^{(j)}$. For any $j \in Y$, $m \in Y$, $j \neq m$, we introduce the aggregated transition rates

$$\widehat{a}_n(j,m) = \sum_{i\in X_j} \pi_n^{(j)}(i) \sum_{l\in X_m} h_n(i,l). \tag{8.39}$$

Suppose that there exist the values $\widehat{a}(j,m)$ such that for any $j,m \in Y$, $j \neq m$,

$$\widehat{a}_n(j,m) \to \widehat{a}(j,m). \tag{8.40}$$

Denote by $y(t)$ an MP with values in Y and transition rates $\widehat{a}(j,m)$, $j,m \in Y$, $j \neq m$.

THEOREM 8.5. *Let $x_{n0} \in X_{j_0}$. Then as $n \to \infty$, the sequence of processes $K(x_{n,[nt]})$ J-converges in any interval $[0,T]$ to an MP process $y(t)$ with the initial state j_0.*

The proof follows the same lines as the proof of Theorem 8.4.

8.4.2. *Convergence of the aggregated process with a general state space*

Now consider the asymptotic aggregation of homogenous MPs with a general state space assuming that the aggregated regions may form V_n-S-sets (satisfy the asymptotic uniformly strong mixing conditions in a particular scale of time). Suppose that MP x_{nk} takes values in a measurable set X with Borel σ-algebra $\mathcal{B}_X$ which can be represented in the form:

$$X = \bigcup_{y \in Y} X_y, \quad \text{where } X_{y_1} \bigcap X_{y_2} = \emptyset \text{ as } y_1 \neq y_2, \tag{8.41}$$

where Y is a measurable set with Borel σ-algebra $\mathcal{B}_Y$. Assume that one-step transition probabilities $p_n(x, A) = \mathbf{P}(x_{n1} \in A \mid x_{n0} = x)$ are represented in the form

$$p_n(x, A) = p_n^{(0)}(x, A) + \frac{1}{n} h_n(x, A), \quad x \in X, \ A \in \mathcal{B}_X, \tag{8.42}$$

where $\limsup_{n\to\infty} \sup_{x,A} |h_n(x, A)| < C$, and for any $y \in Y$,

$$p_n^{(0)}(x, A) \equiv 0 \text{ as } x \in X_y, \quad A \bigcap X_y = \emptyset.$$

and $p_n^{(0)}(x, X_y) = 1$ as $x \in X_y$. For any $y \in Y$, let $x_{nk}^{(y)}$ be an auxiliary MP with state space X_y and transition probabilities $p_n^{(0)}(x, A)$, $x \in X_y$, $A \in \mathcal{B}_{X_y}$. Introduce its uniformly strong mixing coefficient:

$$\begin{aligned} \varphi_n^{(y)}(k) = \sup_{x_1, x_2 \in X_y,\ A \in B_{X_y}} & \big| \mathbf{P}\{x_{nk}^{(y)} \in A \mid x_{n0}^{(y)} = x_1\} \\ & - \mathbf{P}\{x_{nk}^{(y)} \in A \mid x_{n0}^{(y)} = x_2\} \big|. \end{aligned} \tag{8.43}$$

Suppose that there exists a sequence of integers r_n and q, $0 \leq q < 1$, such that

$$n^{-1} r_n \longrightarrow O, \quad \text{and for any } y \in Y, \quad \varphi_n^{(y)}(r_n) \leq q. \tag{8.44}$$

Denote by $\pi_n^{(y)}(A)$, $A \in \mathcal{B}_{X_y}$, a stationary measure for an MP $x_{nk}^{(y)}$ (it exists under assumption (8.44)). For any $y \in Y$, $C \in \mathcal{B}_Y$, $y \notin C$, introduce the aggregated transition probabilities within the set Y:

$$a_n(y, C) = \int_{X_y} h_n\left(x, \bigcup_{u \in C} X_u\right) \pi_n^{(y)}(\mathrm{d}x).$$

Suppose that the following condition holds:

A) there exists a family of finite measures $A(y, C)$, $y \in Y$, $C \in B_Y$, $y \notin C$, such that an MP $y(t)$ with values in Y and transition rates $A(y, C)$ for any initial state is regular and for any $y \in Y$ and any continuous bounded function $f(v)$, $v \in Y$, uniformly in $y \in Y$,

$$\int_{Y\setminus\{y\}} f(v)\big(a_n(y,\mathrm{d}v) - A(y,\mathrm{d}v)\big) \longrightarrow 0. \tag{8.45}$$

Providing the same steps as at the proof of Theorem 8.4 and using Theorem 8.3 we can prove the following result.

THEOREM 8.6. *Let $x_n(0) \in X_{y_0}$ and conditions (8.41), (8.42), (8.44) and condition A) be satisfied. Then as $n \to \infty$, the sequence of aggregated processes $K(x_{n,[nt]})$ J-converges in any interval $[0,T]$ to an MP process $y(t)$ with the initial state y_0.*

Note that this result can be easy extended to the case when the auxiliary processes $x_{nk}^{(y)}$ in each region X_y are quasi-ergodic Markov processes (see section 8.6).

8.4.3. *Accumulating processes in aggregation scheme*

In the same way we can investigate the behavior of accumulating processes and, in particular, the flows of rare events switched by an MP admitting the asymptotic aggregation of state space. Assume that x_{nk}, $k \geq 0$, is an MP satisfying the conditions of the asymptotic aggregation given either in Theorem 8.4, or Theorem 8.5 or Theorem 8.6. Let $K(x_{nk})$ be the aggregated process defined according to (8.29), (8.30). Consider a general state space X and Y and let us keep the notation of section 8.4.2. Let $f(x)$, $x \in X$, be a bounded measurable function in X. Also let $\{\chi_{nk}(x),\ x \in X\}$, $k = 0, 1, \ldots$ be the families of indicators of rare events, which are jointly independent in index k, with distributions not depending on k, where

$$\mathbf{P}\big(\chi_{nk}(x) = 1\big) = q(x)/n + o_n(x), \quad x \in X, \tag{8.46}$$

$q(x)$ is a bounded measurable function in X, and

$$\lim_{n\to\infty} \sup_{x\in X} n o_n(x) = 0. \tag{8.47}$$

Put

$$S_n(m) = \sum_{k=0}^{m} f\big(x_{nk}\big), \quad \zeta_n(m) = \sum_{k=0}^{m} \chi_{nk}\big(x_{nk}\big), \quad m = 0, 1, \ldots \tag{8.48}$$

Denote

$$\widehat{f}_n(y) = \int_{x\in X_y} f(x)\pi_n^{(y)}(\mathrm{d}x), \quad \widehat{q}_n(y) = \int_{x\in X_y} q(x)\pi_n^{(y)}(\mathrm{d}x), \quad y \in Y. \tag{8.49}$$

Assume that there exist measurable functions $\widehat{f}(y)$ and $\widehat{q}(y)$ such that

$$\lim_{n\to\infty} \sup_{y\in Y} \left\{ \left|\widehat{f}_n(y) - \widehat{f}(y)\right| + \left|\widehat{q}_n(y) - \widehat{q}(y)\right| \right\} = 0. \tag{8.50}$$

Let us define the limiting accumulating process $S_0(t) = \int_0^t \widehat{f}(y(u))\mathrm{d}u$, where $y(\cdot)$ is a limiting MP defined in Theorem 8.4 or Theorem 8.6. Also let $\Pi(t)$ be a doubly stochastic Poisson process switched by $y(t)$ with the instantaneous rate of jump at time t, $\widehat{q}(y(t))$.

THEOREM 8.7. *If $x_n(0) \in X_{y0}$, then as $n \to \infty$ under the conditions above in any interval $[0,T]$ the sequence of processes $(K(x_{n,[nt]}), S_n([nt])/n)$ J-converges to a two-component MP process $(y(t), S_0(t))$, and the sequence $(K(x_{n,[nt]}), \zeta_n([nt]))$ J-converges to a two-component MP $(y(t), \Pi(t))$, where $y(0) = y_0$.*

The proof of Theorem follows the same lines as Theorem 8.4.

Note that condition (8.50) can weaken: instead of taking supremum in Y, for any fixed T the condition should be satisfied for a sequence $Y_m \in Y$ such that $\mathbf{P}(y(u) \in Y_m,\ 0 \le u \le T) \to 1$ as $m \to \infty$.

8.4.4. *MP aggregation in continuous time*

The results above can be easy extended to MPs in continuous time. Consider for simplicity a finite state space. Let $x_n(t)$, $t \ge 0$, be a homogenous MP with values in X and transition rates $a_n(i,j)$, $i, i \in X$, $i \ne j$, $t \ge 0$. Suppose that the following representation is valid:

$$X = \bigcup_{y\in Y} X_y, \quad \text{where } X_{y_1} \bigcap X_{y_2} = \emptyset \text{ as } y_1 \ne y_2, \tag{8.51}$$

and

$$a_n(i,j) = a_n^{(0)}(i,j) + V_n^{-1} h_0(i,j)\big(1 + o(1)\big), \quad i,j \in X, \tag{8.52}$$

where for any $y \in Y$, $a_n^{(0)}(i,j) \equiv 0$ as $i \in X_y$, $j \notin X_y$, and $V_n \to \infty$ as $n \to \infty$. This means that X can be divided in the non-intersected regions with small transition rates of the order $O(1/V_n)$ between them.

Let the non-negative values $\{q_n(x),\ x \in X\}$ be given. Denote by $\Pi_n(t)$, $t \ge 0$, a doubly stochastic Poisson process switched by $x_n(t)$: if at time t, $x_n(t) = x$, then the rate of jump of $\Pi_n(t)$ is $q_n(x)$. Consider the asymptotic behavior of the aggregated process $K(x_n(t))$, where $K(\cdot)$ is the map from X to Y defining the subdivision in regions X_y (see relation (8.30)), and also the behavior of the two-component process $(K(x_n(t)), \Pi_n(t))$.

For any $y \in Y$ denote by $x_n^{(y)}(t)$, $t \geq 0$, an auxiliary MP with state space X_y and transition rates $a_n^{(0)}(i,j)$, $i,j \in X_y$, $i \neq j$. Introduce a uniformly strong mixing coefficient

$$\begin{aligned}\varphi_n^{(y)}(u) = \max_{x_1,x_2 \in X_y,\ A \subset X_y} \Big| \mathbf{P}\big\{x_n^{(y)}(u) \in A \mid x_n^{(y)}(0) = x_1\big\} \\ - \mathbf{P}\big\{x_n^{(y)}(u) \in A \mid x_n^{(y)}(0) = x_2\big\}\Big|.\end{aligned} \tag{8.53}$$

Suppose that there exists a sequence r_n and q, $0 \leq q < 1$, such that

$$V_n^{-1} r_n \to 0, \quad \text{and for any } y \in Y, \quad \varphi_n^{(y)}\big(r_n\big) \leq q. \tag{8.54}$$

Note that condition (8.54) means that each subset X_y for an MP $x_n^{(y)}(t)$ forms an V_n-s-set (see section 7.3 and [ANI 70, ANI 73, ANI 78]). In particular X_y may form a closed ergodic subset.

Denote by $\rho_n^{(y)}(x)$, $x \in X_y$, a stationary distribution of $x_n^{(y)}(t)$ which exists at large enough n due to condition (8.54). Put

$$\widehat{a}_n(y,z) = \sum_{x \in X_y} \rho_n^{(y)}(x) \sum_{s \in X_z} h_0(x,s).$$

Also assume that

$$q_n(x) = q_0(x) V_n^{-1} + o\big(V_n^{-1}\big), \quad x \in X,$$

and denote

$$\widehat{q}_0(y) = \sum_{x \in X_y} \rho_n^{(y)}(x) q_0(x).$$

Suppose that there exist the values $\widehat{a}(y,z)$ and $\widehat{q}_0(y)$ such that

$$\widehat{a}_n(y,z) \longrightarrow \widehat{a}_0(y,z), \quad \widehat{q}_n(y) \longrightarrow \widehat{q}_0(y), \quad y,z \in Y,\ y \neq z. \tag{8.55}$$

Denote by $y(t)$, $t \geq 0$, an MP in continuous time with state space Y and transition rates $\widehat{a}_0(y,z)$, $y \neq z$, and let $\Pi(t)$ be a doubly stochastic Poisson process switched by $y(t)$ with the instantaneous rate at time t, $\widehat{q}_0(y(t))$.

THEOREM 8.8. *If $x_n(0) \in X_{y_0}$, then as $n \to \infty$, in any interval $[0,T]$ the sequence of aggregated processes $K(x_n(V_n t))$ J-converges to an MP $y(t)$, where $y(0) = y_0$, and also the sequence of two-component processes $(K(x_n(V_n t)), \Pi_n(V_n t))$ J-converges to the process $(y(t), \Pi(t))$.*

These results are partially published in the books [ANI 87b, ANI 88c] and papers [ANI 78, ANI 98, ANI 00a, ANI 00b] and are extended to the schemes of aggregation for Markov and semi-Markov processes with a general state space and also to non-homogenous in time models.

Note also that using the results of Chapter 3 we can prove that under rather general assumptions the behavior of the accumulating stochastic processes switched by a Markov process, which admits the asymptotic aggregation of state space, is approximated by the processes with independent increments and Markov switching.

8.5. Asymptotic behavior of the first exit time from the subset of states (non-homogenous in time case)

Consider the asymptotic behavior of the first exit time from the subset of states for a non-homogenous MP. In most applications this time is usually interpreted as the time of the first loss of a call, or the time of a failure of the system. We prove the results on the generalized exponential approximation of this time and use it for the investigation of the models of the asymptotic aggregation of the state space in the next sections. The proof uses the properties of a quasi-ergodic MP introduced in section 3.3 and the results of section 7.2.

We study the continuous time as this is usually the case in queueing models. The results for discrete time are similar. Let $x_n(t)$, $t \geq 0$, be a non-homogenous in time MP with finite state space $X = \{0, 1, 2, \dots, d\}$ given by the family of instantaneous transition rates $\{a_n(i, l, t),\ i, l \in X,\ i \neq l,\ t \geq 0\}$. Let $X_0 = \{1, 2, \dots, d\}$ be a subset of X. Given that $x_n(0) = i_0$, $i_0 \in X_0$, denote the first exit time from X_0 by

$$\nu_n\big(i_0\big) = \inf\big\{t : t > 0,\ x_n(t) = 0\big\}. \tag{8.56}$$

Consider the asymptotic behavior of the variable $\nu_n(i_0)$ with the assumption that the rates of exit tend to 0 and the subset X_0 forms in a limit one subset of communicated states. Suppose that there exists a family of continuous in v functions $\{a_0(i, j, v),\ i, j \in X_0,\ i \neq j,\ v \geq 0\}$ and a sequence $k_n \to \infty$ such that for any fixed $T \geq 0$,

$$\lim_{n\to\infty} \sup_{v\leq T} \big|a_n\big(i, j, k_n v\big) - a_0(i, j, v)\big| = 0, \quad i, j \in X_0,\ i \neq j. \tag{8.57}$$

For each fixed $v \geq 0$ denote by $x_0^{(v)}(\cdot)$ an auxiliary homogenous MP with state space X_0 given by the family of transition rates $\{a_0(i, j, v),\ i, j \in X_0,\ i \neq j\}$. Let $\varphi^{(v)}(u)$ be its uniformly strong mixing coefficient. Suppose that there exists q, $0 \leq q < 1$, and for any $T > 0$ there exists a constant $r(T) > 0$ such that for any $v \leq T$,

$$\varphi^{(v)}\big(r(T)\big) \leq q. \tag{8.58}$$

This means that for any $T > 0$ an MP $x_0^{(v)}(\cdot)$ is ergodic uniformly in $u \leq T$. Let $\pi^{(v)}(i)$, $i \in X_0$, be its stationary distribution. Denote by

$$\widehat{a}_n\big(X_0, 0, t\big) = \sum_{i\in X_0} \pi^{(v)}(i) a_n(i, 0, t)$$

the averaged (quasi-stationary) rate of transition from X_0 to the state $\{0\}$ at time t.

THEOREM 8.9. *Suppose that there exists a sequence $k_n \to \infty$ such that relations (8.57) and (8.58) hold, and*

$$\limsup_{n\to\infty} \max_{i\in X_0} \sup_{u\le T} k_n a_n\big(i,0,k_n u\big) < C_T < \infty. \tag{8.59}$$

Then for any $i_0 \in X_0$,

$$\lim_{n\to\infty} \sup_{t\le T} \big|\mathbf{P}\big\{\nu_n\big(i_0\big) > k_n t\big\} - \exp\big\{-\Lambda_n(t)\big\}\big| = 0, \tag{8.60}$$

where $\Lambda_n(t) = k_n \int_0^t \widehat{a}_n(X_0,0,k_n v)\mathrm{d}v$.

Note that proximity estimates similar to Statement 7.1, section 7.2 can also be provided.

NOTE 8.2. In the homogenous case $a_n(i,l,t) \equiv a_n(i,l)$, e.g., the transition rates do not depend on time t, and we obtain

$$\Lambda_n(t) = tk_n \sum_{i\in X_0} \pi(i)a_n(i,0)$$

(the exponential approximation of the variable $\nu_n(i_0)$). In the non-homogenous case the value $\Lambda_n(t)$ stands for the cumulative rate.

Proof. Consider an auxiliary non-homogenous MP $\tilde{x}_n(\cdot)$ with state space X_0 and the instantaneous transition rates $\{a_n(i,l,t),\ i,l \in X_0,\ i \ne l,\ t \ge 0\}$. Furthermore, denote by $(\tilde{x}_n(t), \Pi_n(t))$, $t \ge 0$, a two-component MP such that $\Pi_n(t)$ is a doubly stochastic Poisson process switched by the process $\tilde{x}_n(t)$: the instantaneous rate of a jump at time t is $a_n(x_n(t),0,t)$. Given that $\tilde{x}_n(0) = i_0 \in X_0$, put

$$\widetilde{\nu}_n\big(i_0\big) = \inf\big\{t : t > 0,\ \tilde{\Pi}_n(t) \ge 1\big\}.$$

It is not hard to prove that at each $i_0 \in X_0$ the variables $\nu_n(i_0)$ (see (8.56)) and $\tilde{\nu}_n(i_0)$ have the same distribution. Furthermore, according to relations (8.57), (8.58) process $\tilde{x}_n(\cdot)$ is a quasi-ergodic process (see section 3.3) and Lemma 3.2 with relation (3.55) holds.

Let us use following representation:

$$\mathbf{P}\big\{\widetilde{\nu}_n\big(i_0\big) > k_n t\big\} = \mathbf{E}\exp\bigg\{-\int_0^{k_n t} a_n\big(\tilde{x}_n(u),0,u\big)\mathrm{d}u\bigg\}. \tag{8.61}$$

Denote $\tilde{\Lambda}_n(t) = \mathbf{E}\int_0^{k_n t} a_n(\tilde{x}_n(u),0,u)\mathrm{d}u$. Using the inequality $|e^{\alpha} - e^{\beta} - e^{\beta}(\alpha-\beta)| \le \frac{1}{2}|\alpha-\beta|^2$ that is true as $\alpha,\beta \le 0$, we obtain from (8.61):

$$\big|\mathbf{P}\big\{\nu_n\big(i_0\big) > k_n t\big\} - \exp\big\{-\tilde{\Lambda}_n(t)\big\}\big| \le \frac{1}{2}\mathbf{E}\bigg|\int_0^{k_n t} a_n\big(\tilde{x}_n(u),0,u\big)\mathrm{d}u - \tilde{\Lambda}_n(t)\bigg|^2. \tag{8.62}$$

Denote by $\varphi_n(u,v)$ the uniformly strong mixing coefficient defined for the non-homogenous in time process $\widetilde{x}_n(\cdot)$ in the interval $[u,v]$ by analogy to (8.53). Using conditions (8.57) and (8.58) and the results of [ANI 88a] we can prove that for any $T>0$ there exists the values q_1, $q<q_1<1$ and $r(T)$ such that

$$\sup_{u\leq T}\varphi_n\big(u,u+r(T)\big)\leq q_1. \tag{8.63}$$

Inequality (7.3) implies that for any $u<v$,

$$\begin{aligned}&\mathbf{E}\big|a_n\big(x_n(u),0,u\big)a_n\big(x_n(v),0,v\big)-\mathbf{E}a_n\big(x_n(u),0,u\big)\mathbf{E}a_n\big(x_n(v),0,v\big)\big|\\&\quad\leq\sup_{x,s}a_n(x,0,s)\varphi_n(u,v),\end{aligned}$$

Using this inequality it is not hard to prove that

$$\sup_{t\leq T}\mathbf{E}\left|\int_0^{k_nt}a_n\big(x_n(v),v\big)\mathrm{d}v-\tilde{\Lambda}_n(t)\right|^2\longrightarrow 0.$$

Furthermore, relation (8.63) together with (8.59) implies that uniformly in $t\leq T$, $\tilde{\Lambda}_n(t)-\Lambda_n(t)\to 0$. This completes the proof of Theorem 8.9. □

The results of Theorem 8.9 can be extended to the case of a general state space X using the results of sections 3.2.1.2 and 3.3.

Note that Theorem 8.9 extends to the non-homogenous case some results on the asymptotic behavior of the first exit time from a subset of states for homogenous MP and SMP studied independently in [ANI 70, ANI 73, ANI 74, ANI 87b, ANI 88c] and [KOR 69, KOR 93].

It is also possible to study the exit time from a subset $\widetilde{X}$ which satisfies the conditions of the asymptotic aggregation of the states in a particular scale of time. Using the same technique we can represent the exit time from $\widetilde{X}$ as the time of the first jump of $\Pi_n(t)$ and prove that asymptotically it is equivalent to the time τ of the first jump of process $\Pi_0(t)$ which is a doubly stochastic Poisson process switched by the aggregated MP. In this case τ has a PH-type distribution. This means that in complex systems with transition rates of a different order the distribution of the time of first failure in general is approximated by a PH-type distribution.

In queueing systems the value $\nu_n(i_0)$ usually means the time of the first loss of a call (or overfilling of the system). Note that for homogenous in time models the applications to the analysis of the behavior of complex renewable systems with fast repair were given in [ANI 78, ANI 97, ANI 98, ANI 00a, ANI 87b, ANI 89b, ANI 89a, ANI 89c, SZT 91].

8.6. Aggregation of states of non-homogenous Markov processes

Here we extend the results of sections 8.4.1, 8.4.2 and 8.4.4 to the models of asymptotic aggregation of a state space for hierarchic non-homogenous in time Markov systems. We prove that if the state space of an MP can be divided in the regions such that the transition probabilities between them are small in some sense, then under rather general conditions the accumulating processes can be approximated by non-homogenous in time processes with independent increments and Markov switching with the number of states equal to the number of regions.

Let for each $n > 0$, $x_n(t)$, $t \geq 0$, be a non-homogenous MP in continuous time with state space $X = \{1, 2, \ldots, d\}$ given by the family of instantaneous transition rates $a_n(i, j, t)$, $i, j \in X$, $i \neq j$. Suppose that X can be represented in the form:

$$X = \bigcup_{y \in Y} X_y, \quad \text{where } X_{y_1} \cap X_{y_2} = \emptyset \text{ as } y_1 \neq y_2, \tag{8.64}$$

and $K(\cdot)$ is a map from X to Y such that $K(x) = y$ for any $x \in X_y$. Consider the aggregated process $K(x_n(t))$, $t \geq 0$, and study the conditions of the convergence to an MP at the assumption that transition rates between regions X_y are asymptotically small. Assume that the transition rates are represented in the form

$$a_n(i, l, t) = a_n^{(0)}(i, l, t) + \frac{1}{n} b_n(i, l, t), \quad i, l \in X, \tag{8.65}$$

where for any $T > 0$,

$$\limsup_{n \to \infty} \max_{i,l} \sup_{t \leq nT} \left| b_n(i, l, t) \right| < C_T < \infty, \tag{8.66}$$

and for any $y \in Y$, $t > 0$,

$$a_n^{(0)}(i, l, t) \equiv 0 \quad \text{as } i \in X_y,\ l \notin X_y. \tag{8.67}$$

Let functions $a_n^{(0)}(i, l, t)$ regularly depend on the parameter t in the following way: there exists a family of continuous functions $\{a_0(i, l, u),\ i, l = \overline{1, d},\ i \neq l,\ u \geq 0\}$ such that for any $y \in Y$ and $T > 0$,

$$\lim_{n \to \infty} \sup_{u \leq T} \left| a_n^{(0)}(i, l, nu) - a_0(i, l, u) \right| = 0, \quad i, l \in X_y. \tag{8.68}$$

For any $y \in Y$ and fixed $v \geq 0$ denote by $x_0^{(y)}(t, v)$, $t \geq 0$, an auxiliary homogenous MP with state space X_y and transition rates $a_0(i, l, v)$, $i, l \in X_y$, $i \neq l$, and introduce a uniformly strong mixing coefficient

$$\begin{aligned} \varphi_0^{(y,v)}(u) = \max_{i_1, i_2 \in X_y,\ A \subset X_y} \Big| &\mathbf{P}\Big\{ x_0^{(y)}(u, v) \in A \mid x_0^{(y)}(0, v) = i_1 \Big\} \\ &- \mathbf{P}\Big\{ x_0^{(y)}(u, v) \in A \mid x_0^{(y)}(0, v) = i_2 \Big\} \Big|. \end{aligned} \tag{8.69}$$

Suppose that there exists q, $0 \leq q < 1$, and for any $T > 0$ there exists a value $r(T)$ such that for any $y \in Y$, $v \leq T$,

$$\varphi_0^{(y,v)}\big(r(T)\big) \leq q. \tag{8.70}$$

Note that conditions (8.68)–(8.70) mean that each subset X_y for the initial process $x_n(t)$ forms a quasi-ergodic set (see section 3.3). Furthermore, for each $v \geq 0$ denote by $\pi_0^{(y)}(i,v)$, $i \in X_y$, a stationary distribution of an MP $x_0^{(y)}(t,v)$ (this exists under the assumption (8.70)). For any $y \in Y$, $z \in Y$, $y \neq z$, we put

$$\hat{a}_n(y,z,v) = \sum_{i \in X_y} \pi_0^{(y)}(i,v) \sum_{l \in X_z} b_n(i,l,nv).$$

Suppose that for any $y, z \in Y$, $y \neq z$, and for any $t > 0$ there exist the limits:

$$\Lambda(y,z,t) = \lim_{n\to\infty} \int_0^t \hat{a}_n(y,z,u)\mathrm{d}u, \tag{8.71}$$

and the functions $\Lambda(y,z,t)$ can be represented in the form:

$$\Lambda(y,z,t) = \int_0^t \hat{\lambda}_0(y,z,u)\mathrm{d}u, \tag{8.72}$$

and $\hat{\lambda}_0(y,z,t)$ are some continuous with respect to t functions. Denote by $y(t)$ a non-homogenous MP with state space Y and the instantaneous transition rates at time t, $\hat{\lambda}_0(y,z,t)$, $j,z \in Y$, $y \neq z$.

THEOREM 8.10. *Suppose that conditions (8.64)–(8.68), (8.70)–(8.72) hold and* $x_n(0) \in Y_{y_0}$. *Then the sequence of aggregated processes* $K(x_n(nt))$ *J-converges in any interval* $[0,T]$ *to an MP* $y(t)$ *with the initial state* y_0, *and also for any* $t > 0$ *and* $i \in X_y$,

$$\lim_{n\to\infty} \mathbf{P}\big(x_n(nt) = i\big) = \mathbf{P}\big(y(t) = y\big)\pi_0^{(y)}(i,t).$$

Now we consider the convergence of the accumulating processes switched by $x_n(t)$. Let $\{f(i,t),\ i \in X,\ t \geq 0\}$ be a family of continuous functions. Consider the process

$$S_n(t) = \int_0^t f\big(x_n(nu),u\big)\,\mathrm{d}u.$$

Denote

$$\hat{f}(y,t) = \sum_{i \in X_y} \pi_0^{(y)}(i,t) f(i,t), \quad y \in Y,\ t \geq 0.$$

THEOREM 8.11. *Suppose that the conditions of Theorem 8.10 hold. Then the sequence* $(K(x_n(nt)), S_n(t)/n)$ *J-converges in any interval* $[0, T]$ *to the process* $(y(t), S(t))$ *where* $y(t)$ *is defined in Theorem 8.10 and*

$$S(t) = \int_0^t \hat{f}\big(y(u), u\big)\,du. \tag{8.73}$$

It is also possible to study stochastic additive functionals. For example, consider the behavior of flow of rare Poisson events switched by $x_n(\cdot)$. Let $\{q_n(i,t),\ i \in X,\ t \geq 0\}$ be a family of continuous non-negative functions and let $\Pi_n(t)$ be a doubly stochastic Poisson process switched by $x_n(\cdot)$: if $x_n(t) = i$ then the instantaneous rate of a jump is $q_n(i,t)$. Denote by

$$A_n^{(y)}(u) = \int_0^u \sum_{i \in X_y} \pi_0^{(y)}(i,v) q_n(i,nv)\mathrm{d}v$$

the cumulative rate of jump in region X_y, which is averaged by a quasi-ergodic distribution.

THEOREM 8.12. *Suppose that conditions of Theorem 8.10 hold, for any* $T > 0$

$$\limsup_{n\to\infty} \max_{i \in X} \sup_{u \leq T} nq_n(i,nu) < C_T < \infty,$$

and there exists a family of functions $\{\hat{q}(y,v),\ y \in Y,\ v \geq 0\}$*, which are continuous with respect to* v*, such that for any* $u > 0$*,*

$$\lim_{n\to\infty} nA_n^{(y)}(u) = \int_0^u \hat{q}(y,v)\mathrm{d}v.$$

Then the sequence of processes $(K(x_n(nt)), \Pi_n(nt))$ *J-converges in any interval* $[0, T]$ *to the process* $(y(t), \Pi(t))$*, where MP* $y(t)$ *is defined in Theorem 8.10 and* $\Pi(t)$ *is a Poisson process switched by* $y(\cdot)$*: if at the time* t*,* $y(t) = y$*, then the local rate of jump is* $\hat{q}(y,t)$*.*

Proof of Theorems 8.10–8.12. At first we represent the process $(K(x_n(t)), S_n(t))$ as an SP. In this case switching times are the times of transitions between different regions X_y and the corresponding process $\zeta_n(t,y)$ is constructed as an accumulating process given on the auxiliary process $\tilde{x}_n^{(y)}(\cdot)$ in the region X_y. Let us define it more formally. For any $y \in Y$, $l \in X_y$ denote by $\tilde{x}_n^{(y)}(t,l)$ an auxiliary MP with state space X_y, initial state l given by transition rates $a_n(i,k,t)$, $i, k \in X_y$, $i \neq k$. Also let $\Pi_n^{(y)}(t)$ be a non-homogenous compound Poisson type process switched by $\tilde{x}_n^{(y)}(t,l)$ with the instantaneous rate of a jump in the state i at the time t, $b_n^{(y)}(i,t)/n$, where $b_n^{(y)}(i,t) = \sum_{l \notin X_y} b_n^{(y)}(i,l,t)$, and the size of jump

$$\kappa_n^{(y)}(i,t) = \Big\{ l \quad \text{with probability } b_n^{(y)}(i,l,t) b_n^{(y)}(i,t)^{-1},\ l \notin X_y.$$

We consider a two-component process $(\tilde{x}_n^{(y)}(t,l), \Pi_n^{(y)}(t))$. Denote by $\tau_n^{(y)}(u,l)$ the time of the first jump of process $\Pi_n^{(y)}(t)$ in interval $[nu,\infty)$ and by $\beta_n^{(y)}(u,l)$ its size. Let us construct an SP $y_n(t)$ with values in Y using the families of variables $\{\tau_n^{(y)}(u,l),\ \beta_n^{(y)}(u,l),\ l \in X_y\}$, $y \in Y$. Suppose that $i_0 \in X_{y_0}$ is the initial value. Put $t_{n0} = 0$,

$$t_{nk+1} = t_{nk} + \tau_n^{(y_k)}\big(t_{nk}, i_{nk}\big), \quad i_{nk+1} = \beta_n^{(y_{nk})}\big(t_{nk}, i_{nk}\big),$$
$$y_{nk} = K\big(i_{nk}\big), \quad k \geq 0,$$

and denote

$$y_n(t) = y_{nk} \quad \text{as } t_{nk} \leq t < t_{nk+1},\ t \geq 0.$$

Then by definition process $y_n(\cdot)$ is equivalent to process $K(x_n(\cdot))$. Denote by $\Pi^{(y)}(t)$ a compound Poisson process with instantaneous rate of jump $\lambda_0(y,t) = \sum_{z \neq y} \lambda_0(y,z,t)$ (see (8.71), (8.72)) and the size of jump

$$\kappa_0(y,t) = \Big\{ z \quad \text{with probability } \lambda_0(y,z,t)\lambda_0(y,t)^{-1},\ z \neq y.$$

If $n \to \infty$, then according to Theorem 8.9 the variable $\tau_n^{(y)}(u,l)$ for any $l \in X_y$, $u \geq 0$, weakly converges to the variable $\tau^{(y)}(u)$, where $\tau^{(y)}(u)$ is the time of the first jump of the process $\Pi^{(y)}(t)$ in the interval (u,∞). The variable $K(\beta_n^{(y)}(u,l))$ weakly converges to the variable $\kappa_0(y,u)$, respectively. However, an SP constructed with the help of processes $\Pi^{(y)}(t)$ and variables $\kappa_0(y,t)$ is equivalent to an MP $y(t)$ defined in Theorem 8.10. Finally using the result of Theorem 8.3 in section 8.3 we obtain the statement of Theorem 8.10.

In the case of Theorem 8.11 we put $\zeta_n(t,y,i) = n^{-1}\int_0^t f(\tilde{x}_n^{(y)}(nu,i),u)\,\mathrm{d}u$. Then the process $(K(x_n(t)), S_n(t)/n)$ by analogy can be represented as an SP using the processes $\zeta_n(t,y,i)$ and previous notation. As in each region X_y the process $\tilde{x}_n^{(y)}(t,l)$ satisfies the uniformly strong mixing condition, then the process $\zeta_n(t,y,i)$ for any $i \in X_y$ converges to the deterministic function $\int_0^t \hat{f}(y,u)\mathrm{d}u$ and the limiting process for $S_n(t)/n$ which is constructed using the average characteristics corresponds to the expression (8.73). Similar conclusions are made at the proof of Theorem 8.12. □

Similar results can be proved for a non-homogenous in time MP in discrete time satisfying condition (8.64) and a condition of the (8.28) type, where the local transition rates $a_n(i,l,t)$ should be replaced by one-step transition probabilities at step $[nt]$ and the condition of form (8.68) should be satisfied.

8.7. Averaging principle for RPSM in the asymptotically aggregated Markov environment

In this section we study the averaging principle for RPSM introduced in section 1.2.3 with the additional Markov switching for the case when a switching MP is asymptotically aggregated in the scale of time nt. This means that condition (4.51) of Theorem 4.5 in section 4.4 is not true and the states of MP x_{nk} do not asymptotically communicate in the scale of time nt.

8.7.1. *Switching MP with a finite state space*

We keep the notation of section 4.4. Let $(x_n(t), S_n(t))$, $t \geq 0$, be an RPSM with additional Markov switching defined according to relations (4.47), (4.48) and functions $m_n(x, a)$ and $b_n(x, a)$ are defined according to (4.49). Suppose first for simplicity that MP x_{nk} has a finite state space $X = \{1, 2, \dots, d\}$ and keep the previous notations. Let the following representation hold:

$$X = \bigcup_{j \in Y} X_j, \quad \text{where } X_{y_1} \bigcap X_{y_2} = \emptyset \text{ as } y_1 \neq y_2, \tag{8.74}$$

and one-step transition probabilities $p_n(i, j) = \mathbf{P}\{x_{n1} = j \mid x_{n0} = i\}$ are represented in the form

$$p_n(i, j) = p_n^{(0)}(i, j) + n^{-1} h_n(i, j), \quad i, j \in X, \tag{8.75}$$

where $\limsup_{n \to \infty} \max_{i,j} |h_n(i, j)| < C$, and for any $y \in Y$, $p_n^{(0)}(i, j) \equiv 0$ as $i \in X_y$, $j \notin X_y$ and $\sum_{l \in X_y} p_n^{(0)}(i, l) = 1$, $i \in X_y$.

For any $y \in Y$ denote by $x_{nk}^{(y)}$, $k \geq 0$, an auxiliary MP with state space X_y and transition probabilities $p_n^{(0)}(i, j)$, $i, j \in X_y$. Introduce a uniformly strong mixing coefficient

$$\varphi_n^{(y)}(k) = \max_{i_1, i_2 \in X_y,\, A \subset X_y} \left| \mathbf{P}\left\{x_{nk}^{(y)} \in A \mid x_{n0}^{(y)} = i_1\right\} - \mathbf{P}\left\{x_{nk}^{(y)} \in A \mid x_{n0}^{(y)} = i_2\right\}\right|.$$

Suppose that there exists a sequence of integers r_n and q, $0 \leq q < 1$, such that

$$n^{-1} r_n \longrightarrow 0, \quad \text{and for any } y \in Y, \quad \varphi_n^{(y)}(r_n) \leq q. \tag{8.76}$$

Note that condition (8.76) means that each subset X_y forms an n-s-set for process $x_{nk}^{(y)}$ (see section 7.3 and [ANI 70, ANI 73, ANI 78]). In particular X_y may form a closed ergodic subset. Denote by $\pi_n^{(y)}(i)$, $i \in X_y$, a stationary distribution for $x_{nk}^{(y)}$. Furthermore, for any $y \in Y$, $z \in Y$, $y \neq z$, we introduce the aggregated characteristics:

$$\begin{aligned} \widehat{\lambda}_n(y, z) &= \sum_{i \in X_y} \pi_n^{(y)}(i) \sum_{l \in X_z} h_n(i, l), \\ \widehat{m}_n(y, \alpha) &= \sum_{i \in X_y} \pi_n^{(y)}(i) m_n(i, \alpha), \\ \widehat{b}_n(y, \alpha) &= \sum_{i \in X_y} \pi_n^{(y)}(i) b_n(i, \alpha). \end{aligned} \tag{8.77}$$

Suppose that there exist the values $\widehat{\lambda}(y,z)$ and continuous with respect to α functions $\widehat{m}(y,\alpha)$, $\widehat{b}(y,\alpha)$ such that for any $\alpha \in \mathcal{R}^r$, $y,z \in Y$, $y \neq z$,

$$\begin{gathered} \widehat{\lambda}_n(y,z) \longrightarrow \widehat{\lambda}(y,z), \quad \widehat{m}_n(y,\alpha) \longrightarrow \widehat{m}(y,\alpha) > 0, \\ \widehat{b}_n(y,\alpha) \longrightarrow \widehat{b}(y,\alpha). \end{gathered} \tag{8.78}$$

Denote by $y(t,y_0)$ an MP with values in Y, transition rates $\widehat{\lambda}(y,z)$, $y,z \in Y$, $y \neq z$, and the initial state y_0. Let $\eta(u,y_0,s_0)$ be a solution of a differential equation with a random function in the right-hand side:

$$\mathrm{d}\eta\big(u,y_0,s_0\big) = \widehat{b}\big(y(u,y_0),\eta\big(u,y_0,s_0\big)\big)\mathrm{d}u, \quad \eta\big(0,y_0,s_0\big) = s_0. \tag{8.79}$$

Denote

$$z\big(u,y_0,s_0\big) = \int_0^u \widehat{m}\big(y\big(v,y_0\big),\eta\big(v,y_0,s_0\big)\big)\mathrm{d}v, \tag{8.80}$$

and define the processes

$$\begin{aligned} \zeta\big(t,y_0,s_0\big) &= \eta\big(z^{-1}\big(t,y_0,s_0\big),y_0,s_0\big), \\ \kappa\big(t,y_0,s_0\big) &= y\big(z^{-1}\big(t,y_0,s_0\big),y_0\big), \quad t \geq 0, \end{aligned} \tag{8.81}$$

where $z^{-1}(t,y_0,s_0) = \inf\{u : u \geq 0,\ z(u,y_0,s_0) = t\}$.

Here we assume that $z(u,y_0,s_0) \xrightarrow{\mathrm{P}} \infty$ as $u \to \infty$, i.e., for any $t > 0$, $\mathbf{P}(z^{-1}(t,y_0,s_0) < \infty) = 1$. Thus, the variable $z^{-1}(t,y_0,s_0)$ exists and is a proper random variable.

Consider the aggregated process $K(x_n(t))$ where $K(\cdot)$ is a map from X to Y such that $K(x) = y$ as $x \in X_y$. Note that the original process $x_n(t)$ is not generally an MP or even an SMP (see section 1.2.3). Moreover, neither is process $K(\cdot)$.

THEOREM 8.13. *Let $(x_n(t),S_n(t))$ be an RPSM defined according to relations (4.47),(4.48), functions $m_n(x,a)$ and $b_n(x,a)$ are defined according to (4.49) and conditions (4.52), (4.53) of Theorem 4.5 hold. Also let $x_n(0) \in X_{y_0}$ and $S_n(0)/n \xrightarrow{\mathrm{P}} s_0$ as $n \to \infty$, conditions (8.74)–(8.76) and (8.78) hold. Then the sequence of processes $(K(x_n(t)),S_n(t)/n)$ J-converges in any interval $[0,T]$ to the process $(\kappa(t,y_0,s_0),\zeta(t,y_0,s_0))$ defined above by (8.81).*

Proof. We use AP for RPSM with Markov switching (Theorem 4.5, section 4.4) and Theorem 8.3, section 8.3. The proof follows the lines similar to the proof of Theorems 8.4, 8.7, section 8.4 on the convergence of accumulating processes switched by the asymptotically aggregated MP. First, let us introduce the auxiliary random processes following the notation similar to proof of Theorem 4.5. Denote

$$g_n(u) = K\big(x_{nk}\big), \quad \eta_n(u) = S_{nk}/n, \quad z_n(u) = t_{nk} \\ \text{as } k/n \le u < (k+1)/n,\ u \ge 0, \tag{8.82}$$

where process $z_n(\cdot)$ corresponds to $y_n(u)$ in Theorem 4.5.

Put $\mu_n(t) = \inf\{u : u > 0,\ z_n(u) \ge t\}$. The following representations hold:

$$S_n(nt)/n = \eta_n\big(\nu_n(t)/n\big) = \eta_n\big(\mu_n(t) - 1/n\big),$$
$$K\big(x_n(t)\big) = g_n\big(\mu_n(t) - 1/n\big).$$

In this way process $(K(x_n(t)), S_n(t)/n)$ is represented as a superposition of two processes: $(g_n(u), \eta_n(u))$ and $\mu_n(t)$. First, we study the behavior of the processes $(g_n(u), \eta_n(u))$ jointly with $z_n(u)$, then $\mu_n(t)$ and their superposition.

Consider process $(g_n(u), \eta_n(u), z_n(u))$, $u \ge 0$. We can represent it as an SP. The formal proof can be provided following similar lines as at the proof of Theorem 8.4. Let us avoid many technical details and provide the explanation of the basic steps. The switching times are chosen as the times of transitions between regions X_y. According to Statement 7.1 the exit time from any region is asymptotically approximated by the exponential distribution and does not depend on the initial state of this region. Notice that $g_n(u)$ represents the aggregated process for the initial MP x_{nk}. Therefore, as it follows from Theorem 8.5, given that $g_n(0) \in X_{y_0}$, process $g_n(t)$ weakly converges to an MP $y(t, y_0)$. Moreover, while process x_{nk} is in region X_y, it behaves like process $x_{nk}^{(y)}$. Let us define for any X_y an auxiliary RPSM $(\eta_n^{(y)}(u), z_n^{(y)}(u))$ which is defined by relation (8.82) where x_{nk} is replaced by $x_{nk}^{(y)}$. Then process $(\eta_n(u), z_n(u))$ can be represented as an SP constructed by processes $(\eta_n^{(y)}(u), z_n^{(y)}(u))$ where the switching times are the transition times between regions.

For any $y \in Y$ let us introduce the auxiliary processes $z^{(y)}(u, s)$ and $\eta^{(y)}(u, s)$, where $\eta^{(y)}(u, s)$ is a solution of a differential equation

$$\mathrm{d}\eta^{(y)}(u,s) = \widehat{b}(y, \eta^{(y)}(u,s))\mathrm{d}u,\ \eta^{(y)}(0,s) = s,$$
$$z^{(y)}(u,s) = \int_0^u \widehat{m}(y, \eta^{(y)}(v,s))\mathrm{d}v.$$

Assume that at instant nt_1 process x_{nk} jumps into region X_y and the value of S_{nk} at this time is ns_1. Then according to Theorem 4.5, section 4.4, process $(\eta_n^{(y)}(u),$ $z_n^{(y)}(u))$ in the interval $[t_1, t_2]$ J-converges to process $(\eta^{(y)}(u, s_1), z^{(y)}(u, s_1))$ and its distribution asymptotically does not depend on the initial state of $x_{nk}^{(y)}$ in region X_y. As it is mentioned above the process $g_n(t)$ weakly converges to an MP $y(t, y_0)$. Therefore, using Theorem 8.3 we can prove that under the conditions of Theorem 8.13 process $(g_n(u), \eta_n(u), z_n(u))$ in any interval $[0, A]$ J-converges to $(y(u, y_0), \eta(u, y_0, s_0), z(u, y_0, s_0))$. As $\widehat{m}(y, \alpha) > 0$, process $z(u, y_0, s_0)$ is strictly monotonically increasing. Thus, process $z^{-1}(t, y_0, s_0)$ exists, and process $(g_n(u), \eta_n(u), \mu_n(u))$ also J-converges to $(y(u, y_0), \eta(u, y_0, s_0), z^{-1}(u, y_0, s_0))$. Note that process $\eta(u, y_0, s_0)$ is continuous, but $y(u, y_0)$ is discontinuous. Based on the results on the U-convergence of the superposition of random functions [BIL 68] for the process $\eta_n(u)$ and on the results about J-convergence of the superposition of discontinuous processes [ANI 79, ANI 88c] we can state that the superposition of processes $(g_n(u), \eta_n(u))$ and $\mu_n(t)$ J-converges to the superposition of limiting processes which implies the statement of Theorem 8.13. □

8.7.2. *Switching MP with a general state space*

Using the results of section 8.4.2 these results can be easily extended to the case where X and Y have a general state space. Assume that an MP x_{nk} satisfies the conditions of asymptotic aggregation in Theorem 8.6, section 8.4.2. Define the values $m_n(x, \alpha)$ and $b_n(x, \alpha)$ by relation (4.49) and in each region X_y similar to (8.77) denote

$$\begin{aligned} \widehat{m}_n(y, \alpha) &= \int_{X_y} m_n(x, \alpha)\pi_n^{(y)}(\mathrm{d}x), \\ \widehat{b}_n(y, \alpha) &= \int_{X_y} b_n(x, \alpha)\pi_n^{(y)}(\mathrm{d}x). \end{aligned} \tag{8.83}$$

Assume that variables $(\tau_{nk}(\cdot), \xi_{nk}(\cdot))$ satisfy conditions (4.52), (4.53) of Theorem 4.5, section 4.4 and there exist functions $\widehat{m}(y, \alpha) > 0$ and $\widehat{b}(y, \alpha)$ which are continuous with respect to α such that for any $\alpha \in \mathcal{R}^r$,

$$\sup_{y \in Y} \left\{ \left|\widehat{m}_n(y, \alpha) - \widehat{m}(y, \alpha)\right| + \left|\widehat{b}_n(y, \alpha) - \widehat{b}(y, \alpha)\right| \right\} \longrightarrow 0. \tag{8.84}$$

Let us define a regular MP $y(t, y_0)$ with state space Y, transition rates $A(y, B)$, $y \in Y$, $B \in \mathcal{B}_Y$, $y \notin B$, given by (8.45) and the initial state y_0. Define also the processes $\eta(u, y_0, s_0)$ and $z(u, y_0, s_0)$ according to relations (8.79) and (8.81).

THEOREM 8.14. *Let $(x_n(t), S_n(t))$ be an RPSM defined according to relations (4.47), (4.48), the functions $m_n(x, a)$ and $b_n(x, a)$ defined according to (4.49) and the conditions (4.52), (4.53) of Theorem 4.5 hold. Let MP x_{nk} satisfy the conditions of asymptotic aggregation in Theorem 8.6, section 8.4.2. Also let $x_n(0) \in X_{y_0}$ and $S_n(0)/n \xrightarrow{\mathrm{P}} s_0$ as $n \to \infty$ and (8.84) holds.*

Then the sequence of processes $(K(x_n(t)), S_n(t)/n)$ J-converges in any interval $[0, T]$ to process $(\kappa(t, y_0, s_0), \zeta(t, y_0, s_0))$ defined above by relations (8.79)–(8.81).

Similar results can be proved for non-homogenous in time MP.

NOTE 8.3. If for any y the process $x_{nk}^{(y)}$ in each region Y is a quasi-ergodic MP with the corresponding family of quasi-ergodic measures $\pi^{(y)}(t, A)$, then the result of Theorem 8.14 is also valid, where the functions $\widehat{m}_n(y, \alpha)$ and $\widehat{b}_n(y, \alpha)$ in (8.83) should be replaced by $\widehat{m}_n(y, \alpha, t)$ and $\widehat{b}_n(y, \alpha, t)$, and the functions $\widehat{m}(y, \alpha)$ and $\widehat{b}(y, \alpha)$ in (8.84) should be replaced by $\widehat{m}(y, \alpha, t)$ and $\widehat{b}(y, \alpha, t)$, respectively, and also the transition rates of the limiting process $y(t, y_0)$ at time t have the form $A(y, B, t)$.

8.7.3. *Averaging principle for accumulating processes in the asymptotically aggregated semi-Markov environment*

Consider a special case of the accumulating process, switched by a semi-Markov process admitting the asymptotic aggregation of state space. This case is important in various applications. Assume that the distributions of variables $(\xi_{nk}(x, z), \tau_{nk}(x, z))$ in (4.46) do not depend on the argument z. Therefore, introduce the family of jointly independent random variables

$$F_{nk} = \big\{\big(\xi_{nk}(x), \tau_{nk}(x)\big),\ x \in X\big\}, \quad k \geq 0, \tag{8.85}$$

with values in $\mathcal{R}^r \times [0, \infty)$ and distributions not depending on index k. In this case the process $x_n(t)$ defined by (4.48) is an SMP given by the embedded MP x_{nk} and the sojourn time in state x is $\tau_{nk}(x)$. Correspondingly, process $S_n(t)$ is a sum of random variables $\xi_{nk}(x)$ defined on the trajectory of an SMP $x_n(\cdot)$ in the interval $[0, t]$. Let us keep the notation of section 4.4.1. Denote

$$m_n(x) = \mathbf{E}\tau_{n1}(x), \quad b_n(x) = \mathbf{E}\xi_{n1}(x). \tag{8.86}$$

Assume that MP x_{nk} satisfies the conditions of asymptotic aggregation in Theorem 8.6, section 8.4.2. For any region X_y denote

$$\widehat{m}_n(y) = \int_{X_y} m_n(x)\pi_n^{(y)}(\mathrm{d}x), \quad \widehat{b}_n(y) = \int_{X_y} b_n(x)\pi_n^{(y)}(\mathrm{d}x). \tag{8.87}$$

Assume that there exist functions $\widehat{m}(y) > 0$ and $\widehat{b}(y)$ such that

$$\sup_{y\in Y} \big\{\big|\widehat{m}_n(y) - \widehat{m}(y)\big| + \big|\widehat{b}_n(y) - \widehat{b}(y)\big|\big\} \longrightarrow 0. \tag{8.88}$$

Let $y(t, y_0)$ be a regular MP with state space Y, transition rates $A(y, B)$, $y \in Y$, $B \in \mathcal{B}_Y$, $y \notin B$, defined by (8.45), and the initial state y_0. Define the processes $\eta(u, y_0, s_0)$ and $z(u, y_0, s_0)$ according to relations (8.79)–(8.81). In our case,

$$\eta\big(u, y_0, s_0\big) = s_0 + \int_0^u \widehat{b}\big(y\big(u, y_0\big)\big)\mathrm{d}u,$$

$$z\big(u, y_0, s_0\big) = z\big(u, y_0\big) = \int_0^u \widehat{m}\big(y\big(v, y_0\big)\big)\mathrm{d}v.$$

Assume that $\eta(u, y_0, s_0) \xrightarrow{\text{P}} \infty$ as $u \to \infty$. Then for any $t > 0$ the value $z^{-1}(t, y_0)$ is a proper random variable and it is easy to check that the process $\kappa(t, y_0) = y(z^{-1}(t, y_0), y_0)$ is also an MP with transition rates $A(y, B)/\widehat{m}(y), y \in Y$, $B \in \mathcal{B}_Y$. Assume that $\kappa(t, y_0)$ is regular (e.g. with probability one has a finite number of jumps in any finite interval). Then process $\zeta(t, y_0, s_0) = \eta(z^{-1}(t, y_0), y_0, s_0)$ after changing time using transformation $u = z^{-1}(v, y_0)$ can be represented in the form

$$\zeta(t, y_0, s_0) = s_0 + \int_0^t \widetilde{b}(\kappa(u, y_0))\mathrm{d}u, \tag{8.89}$$

where $\widetilde{b}(y) = \widehat{b}(y)/\widehat{m}(y)$.

As a consequence of Corollary 4.2, section 4.4.1, Theorem 8.6, section 8.4.2 and Theorem 8.14, section 8.7 we obtain the following result.

COROLLARY 8.1. *Let $S_n(t)$ be a stepwise process of sums of random variables defined by the family of random variables F_{nk} (8.85) on a trajectory of an SMP $x_n(\cdot)$ according to relations (4.47), (4.48) and the following conditions are satisfied:*

1) MP x_{nk} satisfies conditions of asymptotic aggregation of Theorem 8.6, section 8.4.2;

2) functions $m_n(x)$ and $b_n(x)$ defined according to (8.86), condition (4.88) of Corollary 4.2 section 4.4.1 holds, and (8.88) is true;

3) $x_n(0) \in X_{y_0}$ and $S_n(0)/n \xrightarrow{\text{P}} s_0$ as $n \to \infty$.

Then the sequence of processes $(K(x_n(t)), S_n(t)/n)$ J-converges in any interval $[0, T]$ to the process $(\kappa(t, y_0), \zeta(t, y_0, s_0))$, where $\kappa(t, y_0)$ is an MP with transition rates $A(y, B)/\widehat{m}(y)$, $y \notin B$, and the initial state y_0, and process $\zeta(t, y_0, s_0)$ defined by (8.89) is an accumulating process given on the trajectory of $\kappa(t, y_0)$.

8.8. Diffusion approximation for RPSM in the asymptotically aggregated Markov environment

Now consider the diffusion approximation of RPSM switched by an MP which admits the asymptotic aggregation of state space in the scale of time nt. Assuming that $\widehat{b}(y, \alpha) \equiv 0$, it is possible to prove that $\zeta_n(t)$ weakly converges to a diffusion process with Markov switching. Note that the case of a finite state space is studied in [ANI 00b]. We consider a general case.

As according to Theorem 8.13, $S_n(t)/n$ J-converges to process $\zeta(t, y_0, s_0)$, then there is no sense in considering the normalized process $(S_n(t) - n\zeta(t, y_0, s_0))/\sqrt{n}$ as function $\zeta(t, y_0, s_0)$ is random. However, if $\widehat{b}(y, \alpha) \equiv 0$, then $\zeta(t, y_0, s_0) \equiv 0$ and we can investigate the convergence of process $S_n(t)/\sqrt{n}$.

We keep the notation of section 8.7.2. Let at each $n \geq 0$, $F_{nk} = \{(\xi_{nk}(x, z), \tau_{nk}(x, z)),\ x \in X,\ z \in \mathcal{R}^r\}$, $k \geq 0$, be jointly independent families of random variables with values in $R^r \times [0,\infty)$ and distributions not depending on $k \geq 0$, and let x_{ni}, $i \geq 0$, be an independent of F_{nk}, $k \geq 0$, homogenous MP with values in a measurable space $(X, \mathcal{B}_X)$, S_{n0} be the initial value. Let $(x_n(t), S_n(t))$ be an RPSM defined according to relations (4.47), (4.48). Denote

$$m_n(x,\alpha) = \mathbf{E}\tau_{n1}(x,\sqrt{n}\alpha), \quad b_n(x,\alpha) = \mathbf{E}\xi_{n1}(x,\sqrt{n}\,\alpha), \tag{8.90}$$

using a normalizing factor $\sqrt{n}$ and assume that

$$b_n(x,\alpha) \equiv 0 \quad \text{for any } x \in X,\ \alpha \in R^r. \tag{8.91}$$

Let MP x_{nk} satisfy the conditions of asymptotic aggregation in Theorem 8.6, section 8.4.2 and the values $\widehat{m}_n(y,\alpha)$ are defined by relations (8.83). Suppose also that there exists a matrix function $B_n^2(x,\alpha) = \mathbf{E}\xi_{n1}(x,\alpha\sqrt{n})\xi_{n1}(x,\alpha\sqrt{n})^*$, and for any $y \in Y$ denote

$$\widehat{B}_n^2(y,\alpha) = \int_{X_y} B^2(x,\alpha)\pi_n^{(y)}(\mathrm{d}x).$$

Suppose that there exist functions $\widehat{m}(y,\alpha)$, $\widehat{B}^2(y,\alpha)$ such that for any $\alpha \in R^r$,

$$\sup_y \left\{ \left|\widehat{m}_n(y,\alpha) - \widehat{m}(y,\alpha)\right| + \left|\widehat{B}_n^2(y,\alpha) - \widehat{B}^2(y,\alpha)\right| \right\} \longrightarrow 0. \tag{8.92}$$

Denote by $y(t,y_0)$ a limiting MP for the aggregated process $K(x_{nk})$ with state space Y, transition rates $A(y,B)$, $y \in Y$, $B \in \mathcal{B}_Y$, $y \notin B$, defined by (8.45) and the initial state y_0. Let $\rho(t,y_0,s_0)$, $y_0 \in Y$, $s_0 \in \mathcal{R}^r$, be a solution of a stochastic differential equation with a random function in the right-hand side:

$$\rho\big(0,y_0,s_0\big) = s_0, \quad \mathrm{d}\rho\big(t,y_0,s_0\big) = \widehat{B}\big(y(t,y_0),\rho\big(t,y_0,s_0\big)\big)\mathrm{d}w(t), \tag{8.93}$$

where $w(t)$ is a standard Wiener process in $\mathcal{R}^r$. We can say that $\rho(t,y_0,s_0)$ is a diffusion process with Markov switching. Also define process $z(t,y_0,s_0)$:

$$z\big(t,y_0,s_0\big) = \int_0^t \widehat{m}\big(y\big(v,y_0\big),\rho\big(v,y_0,s_0\big)\big)\mathrm{d}v. \tag{8.94}$$

Denote

$$\begin{aligned} \vartheta\big(t,y_0,s_0\big) &= \rho\big(z^{-1}\big(t,y_0,s_0\big),y_0,s_0\big), \\ \kappa\big(t,y_0,s_0\big) &= y\big(z^{-1}\big(t,y_0,s_0\big),y_0\big), \quad t \geq 0, \end{aligned} \tag{8.95}$$

where we assume that $z(u,y_0,s_0) \xrightarrow{\mathrm{P}} \infty$ as $u \to \infty$, i.e., for any $t > 0$, $\mathbf{P}(z^{-1}(t,y_0,s_0) < \infty) = 1$. Thus, the variable $z^{-1}(t,y_0,s_0)$ exists and is a proper random variable.

Consider the asymptotic behavior of process $(K(x_n(nt)), S_n(nt)/\sqrt{n})$.

THEOREM 8.15. *Let $(x_n(t), S_n(t))$ be an RPSM defined according to relations (4.47), (4.48), functions $m_n(x,a)$ and $b_n(x,a)$ defined according to (8.90), relation (8.91) holds, function $m_n(x,\alpha)$ satisfies condition (4.53) of Theorem 4.5 and for any $N > 0$ the following conditions are satisfied:*

$$\lim_{L\to\infty} \limsup_{n\to\infty} \sup_{|\alpha|\le N} \sup_{x} \mathbf{E}\tau_{n1}(x,\sqrt{n}\alpha)\chi\big(\tau_{n1}(x,\sqrt{n}\alpha) > L\big) = 0; \tag{8.96}$$

$$\lim_{L\to\infty} \limsup_{n\to\infty} \sup_{|\alpha|\le N} \sup_{x} \mathbf{E}\big|\xi_{n1}(x,\sqrt{n}\alpha)\big|^2\chi\big(\big|\xi_{n1}(x,\sqrt{n}\alpha)\big| > L\big) = 0; \tag{8.97}$$

and as $\max(|\alpha_1|, |\alpha_2|) \le N$,

$$\big|B_n^2\big(x,\alpha_1\big) - B_n^2\big(x,\alpha_2\big)\big| \le C_N\big|\alpha_1 - \alpha_2\big| + \alpha_n(N), \tag{8.98}$$

where $\alpha_n(N) \to 0$ uniformly in $|\alpha_1| \le N$, $|\alpha_2| \le N$;

Let MP x_{nk} satisfy the conditions of asymptotic aggregation in Theorem 8.6, section 8.4.2, and relation (8.92) holds, where for any $\alpha \in \mathcal{R}^r$, uniformly with respect to $y \in Y$, $|\widehat{B}(y,\alpha)| \le C(1+|\alpha|)$. Also let $x_n(0) \in X_{y_0}$ and $S_n(0)/\sqrt{n} \xrightarrow{\text{P}} s_0$ as $n \to \infty$.

Then the sequence of processes $(K(x_n(t)), S_n(t)/\sqrt{n})$ J-converges in any interval $[0,T]$ such that $z(+\infty, y_0, s_0) > T$ with probability one, to process $(\kappa(t, y_0, s_0), \vartheta(t, y_0, s_0))$ defined above by relations (8.93)–(8.95).

The proof follows the same lines as the proof on Theorem 8.13. We keep relation (8.82) where instead of process $\eta_n(u)$ we introduce the process

$$\rho_n(u) = S_{nk}/\sqrt{n}, \quad \text{as } k/n \le u < (k+1)/n,\ u \ge 0.$$

Let us define for any X_y an auxiliary RPSM $(\rho_n^{(y)}(u), z_n^{(y)}(u))$ which is defined on the trajectory of process $x_{nk}^{(y)}$. Then process $(\rho_n(u), z_n(u))$ can be represented as an SP constructed by processes $(\rho_n^{(y)}(u), z_n^{(y)}(u))$ where the switching times are the transition times between regions. Now we introduce for any $y \in Y$ the auxiliary processes $z^{(y)}(u,s)$ and $\rho^{(y)}(u,s)$, where $\rho^{(y)}(u,s)$ is a solution of a stochastic differential equation

$$\mathrm{d}\rho^{(y)}(u,s) = \widehat{B}\big(y, \rho^{(y)}(u,s)\big)\mathrm{d}w(u), \quad \rho^{(y)}(0,s) = s,$$

and $z^{(y)}(u,s) = \int_0^u \widehat{m}(y, \rho^{(y)}(v,s))\mathrm{d}v$.

According to Theorem 4.6, section 4.4, process $(\rho_n^{(y)}(u), z_n^{(y)}(u))$ in any interval $[t_1, t_2]$ given that $\rho_n^{(y)}(t_1) = s_1$, J-converges to the process $(\rho^{(y)}(u,s_1), z^{(y)}(u,s_1))$. Then following the final lines of the proof of Theorem 8.13 we obtain our statement.

NOTE 8.4. As mentioned in Corollary 8.1, in the case when functions $\widehat{m}(y, \alpha)$ do not depend on argument α, process $\kappa(t, y_0, s_0)$ does not depend on s_0 and is an MP with values in Y and transition rates $\widehat{m}(y)^{-1}A(y, B)$. In general $\kappa(t, y_0, s_0)$ is not a Markov or even a semi-Markov process.

Note that when the variables $\{(\xi_{nk}(x, \alpha), \tau_{nk}(x, \alpha))\}$ do not depend on parameter α, (see (8.85)), then process $x_n(t)$ is an SMP given by the embedded MP x_{nk} and the sojourn time in state x $\tau_{nk}(x)$. Correspondingly, the process $S_n(t)$ is a sum of random variables $\xi_{nk}(x)$ defined on the trajectory of an SMP $x_n(\cdot)$ in the interval $[0, t]$ (see section 8.7.3). Therefore, as a consequence of Theorem 8.15 we can formulate similar to Corollary 8.1 the result on the diffusion approximation of processes in the asymptotically aggregated semi-Markov environment.

The results of this section can easily be extended to the case when the auxiliary processes $x_{nk}^{(y)}$ in each region X_y are quasi-ergodic Markov processes.

Note that AP and DA for sums of random variables defined on homogenous Markov or semi-Markov processes with either ergodic or asymptotically aggregated state space are studies in [KOR 04, KOR 05] using another analytic technique based on the theory of the linear operators that are perturbed on the spectrum.

8.9. Aggregation of states in Markov queueing models

In the following sections we consider the applications of the results on the analysis of rare events and asymptotic aggregation of state space to queueing systems with different types of rare events or rare transitions.

8.9.1. *System $M_Q/M_Q/r/\infty$ with unreliable servers in heavy traffic*

Consider a state-dependent system $M_Q/M_Q/r/\infty$ similar to the system introduced in section 2.2.1.1. We assume that there are r servers which are subject to random failures. The system is described by the family of rates $\{\lambda(q), \mu(q), a(q), c(q), q \geq 0\}$ and random variables $\{\eta(q), \kappa(q),\ q \geq 0\}$. Suppose that the system operates in the following way. Denote by $Q_n(t)$ the total number of calls in the system at time t (or volume of information). If $Q(t)/n = q$, then the local arrival rate is $\lambda(q)$ and a batch of $\eta(q)$ calls may enter the system. Correspondingly, the local service rate is $\mu(q)$ and a batch of $\min\{\kappa(q), q\}$ calls may complete service and leave the system. Suppose also that each server is subject to random failures and repairs and assume that the rates of failure and repair are small (of the order $O(1/n)$). This means, if at time t, $n^{-1}Q(t) = q$, then the instantaneous rate of failure for each busy server is $a(q)/n$, and the rate of repair for each failed server is $c(q)/n$.

For this system we naturally obtain the model which enables the asymptotic aggregation of states. Denote by $R_n(t)$ the number of servers "on" at time t. Let $g(q) = \mathbf{E}\eta(q)$, $v(q) = \mathbf{E}\kappa(q)$.

STATEMENT 8.1. *Suppose, that* $Q_n(0) = nq_0$, $R_n(0) = r_0$, *variables* $\eta(q)$, $\kappa(q)$ *are integrable uniformly in q in any bounded region, functions* $\lambda(q)$, $\mu(q)$, $g(q)$, $v(q)$, $a(q)$, $c(q)$ *are locally Lipschitz with respect to q and no longer have linear growth.*

Then the sequence of processes $(R_n(nt), Q(nt)/n)$ *J-converges in* $[0,T]$ *to an MP* $(R(t), z(t))$ *with values in* $\{0,1,\ldots,r\} \times [0,\infty)$ *such that* $R(0) = r_0$, *and at fixed* $z(t) = q$ *process* $R(t)$ *is a Birth-and-Death process in interval* $[0,r]$ *with the following transition rates: from k to* $k+1$ *the rate is* $(r-k)c(q)$ *and from k to* $k-1$ *the rate is* $ka(q)$. *Correspondingly, process* $z(t)$ *satisfies the differential equation with the random right-hand side:*

$$z(0) = q_0, \quad \mathrm{d}z(t) = \big(g\big(z(t)\big)\lambda\big(z(t)\big) - R(t)v\big(z(t)\big)\mu\big(z(t)\big)\big)\mathrm{d}t.$$

Here T *is any positive value such that* $z(t) > 0$ *in* $[0,T]$ *with probability one.*

Proof. To prove this result we note that when $R_n(t) = k$, process $Q_n(\cdot)$ behaves as the queue in system $M_Q/M_Q/k/\infty$ with k available servers. Therefore, according to Theorem 5.2, section 5.2, process $Q_n(t)/n$ in the interval where $R_n(\cdot) = k$ is approximated by process $z^{(k)}(\cdot)$ such that

$$\mathrm{d}z^{(k)}(t) = \big(g\big(z^{(k)}(t)\big)\lambda\big(z^{(k)}(t)\big) - kv\big(z^{(k)}(t)\big)\mu\big(z^{(k)}(t)\big)\big)\mathrm{d}t.$$

Correspondingly, the process $R_n(nt)$ at fixed $Q_n(t) = nq$ locally behaves as a Birth-and-Death process with the rates $(r-k)b(q)$ and $ka(q)$, respectively, and in some sense it plays the role of the environment for $Q_n(t)$.

Therefore, we can represent process $(R_n(nt), Q(nt)/n)$ as an SP where the switching times are the subsequent times of failures and repairs of the servers and use Theorem 8.3, section 8.3 and Theorem 5.2, section 5.2. □

Note that similar results can be obtained when we have an additional semi-Markov environment, and also for queueing networks $(M_{SM,Q}/M_{SM,Q}/k_i/\infty)^r$ with unreliable servers with rare failures and repairs.

8.9.2. *System* $M_{M,Q}/M_{M,Q}/1/\infty$ *in heavy traffic*

Consider a system $M_{M,Q}/M_{M,Q}/1/\infty$ (see section 5.3.3, Corollary 5.3) and assume that a switching MP $x_n(t)$, $t \geq 0$, with values in $X = \{1,2,\ldots,r\}$ satisfies the conditions of the asymptotic aggregation of state space (8.18) and (8.19) in section 8.2.3. Process $x_n(t)$ stands for the switching environment of the queueing model. Assume that an MP defined by transition rates $a_0(i,j)$ in each region X_k is irreducible. Denote by $\pi^{(k)}(i)$, $i \in X_k$, a stationary distribution in region X_k given by (8.20). Define by $K(x_n(t))$ the aggregated process where $K(\cdot)$ is a map from X to Y: $K(i) = y$ if $i \in X_y$.

Also let the family of non-negative functions $\{\lambda(i,q), \mu(i,q),\ q \geq 0\}$, $i \in X$, be given. A queueing model consists of one server and an infinite number of waiting places. Calls enter the system one at a time. The instantaneous input and service rates depend on the state $x(\cdot)$, the value of the queue and the normalizing factor n in the following way: if at time t, $x_n(t) = i$ and $Q_n(t) = Q$, then the input rate is $\lambda(i, Q/n)$ and the service rate is $\mu(i, Q/n)$, where $Q_n(t)$ is the number of calls in the system at time t. For any $y \in Y$, denote

$$\widehat{\lambda}(y,q) = \sum_{i \in X_y} \lambda(i,q)\pi^{(y)}(i), \quad \widehat{\mu}(y,q) = \sum_{i \in X_y} \mu(i,q)\pi^{(y)}(i),$$

and let $\widehat{b}(y,q) = \widehat{\lambda}(y,q) - \widehat{\mu}(y,q)$.

STATEMENT 8.2. *Suppose that functions $\lambda(i,q)$, $\mu(i,q)$ are locally Lipschitz with respect to q, for any $y \in Y$ function $\widehat{b}(y,q)$ has no more than linear growth, $n^{-1}Q_n(0) \xrightarrow{\text{P}} s_0 > 0$ and $x_n(0) \in X_{y_0}$.*

Then the sequence of processes $(K(x_n(t)), Q_n(nt)/n)$ J-converges in the interval $[0,T]$ to the process $(y(t,y_0), q(t,s_0))$, where $y(t)$ is an MP with values in Y, transition rates $\widehat{a}_{ks}$ defined in (8.21) and the initial state y_0, and the process $q(t,s_0)$ is a solution of a differential equation with random right-hand side:

$$\mathrm{d}q\big(t,s_0\big) = \widehat{b}\big(y\big(t,y_0\big), q\big(t,s_0\big)\big)\mathrm{d}t, \quad q\big(0,s_0\big) = s_0,$$

where T is any positive value such that $q(t,s_0) > 0$ in the interval $[0,T]$ with probability one.

The explanation of this result is the following: the transitions between regions occur rarely and the sojourn time in each region is of the order n and is approximated by an exponential distribution. However, in the interval of the order n in each region, the normalized queue $Q(nu)/n$ due to AP is approximated by a solution of differential equation with the parameters averaged by a quasi-stationary distribution in this region.

The formal proof follows the lines of the proof of Theorem 8.4. Let us give the basic ideas without detailed constructions given in this theorem. Note that we can represent the process $(K(x_n(t)), Q_n(nt)/n)$ as an SP where the switching times are the times of transitions between regions X_y. The convergence of the aggregated process $K(x_n(t))$ to an MP $y(t,y_0)$ directly follows from Theorem 8.8, section 8.4.4. Furthermore, in the interval where $x_n(t) \in X_y$, the queueing process asymptotically behaves as the queue in the system with input and service rates $\lambda(i,q)$, $\mu(i,q)$, $i \in X_y$, switched by the auxiliary MP with values in X_y and transition rates $a_0(i,j)$, $i,j \in X_y$. Therefore, according to Corollary 5.3, section 5.3.3 process $Q_n(nt)/n$ in region X_y is approximated by a solution of a differential equation

$$\mathrm{d}q^{(y)}(t) = \widehat{b}\big(y, q^{(y)}(t)\big)\mathrm{d}t.$$

Finally, we use Theorem 8.3, section 8.3.

8.10. Aggregation of states in semi-Markov queueing models

Similar results can be proved for semi-Markov queueing models $SM/M_{SM,Q}/1/\infty$ and $M_{SM,Q}/M_{SM,Q}/1/\infty$. Note that AP for these models is considered in sections 5.3.2 and 5.3.3.

8.10.1. *System* $SM/M_{SM,Q}/1/\infty$

Let $x_n(t)$, $t \geq 0$, be an SMP with values in some finite set X given by the embedded MP x_{nk} and by the family $\{\tau_n(x),\ x \in X\}$ of sojourn times. Also let non-negative functions $\mu_n(x,q)$, $x \in X$, $q \geq 0$, be given. There is one server and an infinite number of waiting places. The calls enter the system one at a time at the instants of jumps $u_{n1} < u_{n2} < \cdots$ of process $x_n(t)$. Put $x_{nk} = x(u_{nk}+0)$. If a call enters the system at time u_{nk} and the number of calls in the system becomes equal to Q, then the service rate in the interval $[u_{nk}, u_{n,k+1})$ is $\mu_n(x_{nk}, Q/n)$. After service completion the call leaves the system. Let Q_{n0} be the initial number of calls, and $Q_n(t)$ be the number of calls in the system at time t.

Note that AP and DA, for the case when the embedded MP x_k, $k \geq 0$, does not depend on parameter n and is uniformly ergodic, was studied in section 5.3.2.

Denote $m_n(x) = \mathbf{E}\tau_n(x), x \in X$. Assume that the embedded MP x_{nk} satisfies the conditions of the asymptotic aggregation (8.27) and (8.28) or more general conditions (8.37) and (8.38) of section 8.4.1. Let $\pi_n^{(j)}(i)$, $i \in X_j$, be a stationary distribution of the auxiliary MP $x_{nk}^{(j)}$ in region X_j. Also let $y(t)$ be an MP with values in Y and aggregated transition rates $\widehat{a}(j,m)$, $j,m \in Y$, $j \neq m$, defined by relations (8.39) and (8.40). For any $j \in Y$ denote

$$\widehat{m}_n(j) = \sum_{i \in X_j} m_n(i)\pi_n^{(j)}(i), \quad \widehat{c}_n(j,q) = \sum_{i \in X_j} \mu_n(i,q) m_n(i) \pi_n^{(j)}(i). \tag{8.99}$$

Assume that the functions $\mu_n(i,q)$ satisfy condition (4.53) of Theorem 4.5,

$$\lim_{L\to\infty} \limsup_{n\to\infty} \max_{x\in X} \mathbf{E}\tau_{n1}(x)\chi\big(\tau_{n1}(x) > L\big) = 0, \tag{8.100}$$

there exist values $\widehat{m}(j) > 0$ and functions $\widehat{c}(j,q)$ such that for any $j \in Y$, $q \geq 0$,

$$\widehat{m}_n(j) \longrightarrow \widehat{m}(j), \quad \widehat{c}_n(j,q) \longrightarrow \widehat{c}(j,q), \tag{8.101}$$

and $S_n(0) \xrightarrow{\text{P}} s_0$. Then similar to Statement 8.2 the following result can be proved:

STATEMENT 8.3. *If* $\mathbf{P}(x_n(0) \in X_{j_0}) \to 1$, *then under the conditions above the sequence* $(K(x_n(nt)), Q_n(nt)/n)$ *J-converges in the interval* $[0,T]$ *to the process* $(y(t,j_0), q(t,j_0,s_0))$, *where* $y(t,j_0)$ *is an MP defined above with the initial state* j_0

and $q(t, j_0, s_0)$ *is a solution of a differential equation with a random right-hand side:* $q(0, j_0, s_0) = s_0$ *and*

$$\mathrm{d}q\big(t, j_0, s_0\big) = \widehat{m}\big(y\big(t, j_0\big)\big)^{-1}\big(1 - \widehat{c}\big(y\big(t, j_0\big), q\big(t, j_0, s_0\big)\big)\big)\mathrm{d}t,$$

and T *is any positive value such that* $q(t, j_0, s_0) > 0$ *for all* $t \in [0, T]$ *with probability one.*

8.10.2. *System* $M_{SM,Q}/M_{SM,Q}/1/\infty$

Now consider a queueing system with semi-Markov modulated arrival flow investigated in section 5.3.3. Let $x_n(t)$, $t \geq 0$, be an SMP with values in $X = \{1, 2, \ldots, r\}$ and sojourn times $\tau_n(x)$, $x \in X$. We keep the notation of section 8.10.1. Let the family of non-negative functions $\{\lambda_n(i, q),\ \mu_n(i, q),\ q \geq 0\}$, $i \in X$, be given. There is one server and an infinite number of waiting places. Calls enter the system one at a time and if at time t, $x(t) = i$ and $Q_n(t) = Q$, then the instantaneous input rate is $\lambda_n(i, Q/n)$ and the service rate is $\mu_n(i, Q/n)$.

Note that the case when the embedded MP x_k, $k \geq 0$, does not depend on parameter n and is irreducible is studied in section 5.3.3. In this section we assume that the embedded MP x_{nk} satisfies the conditions of the asymptotic aggregation (8.27) and (8.28) or more general conditions (8.27), (8.37) and (8.38) of section 8.4.1. Let $\pi_n^{(j)}(i)$, $i \in X_j$, be a stationary distribution of the auxiliary MP $x_{nk}^{(j)}$ in region X_j. Also let $y(t)$ be an MP with values in Y and aggregated transition rates $\widehat{a}(j, m)$, $j, m \in Y$, $j \neq m$, defined by relations (8.39) and (8.40). Let the values $\widehat{m}_n(j)$ be defined in (8.99). Denote for any $j \in Y$

$$\widehat{b}_n(j, q) = \sum_{i \in X_j} \big(\lambda_n(i, q) - \mu_n(i, q)\big) m_n(i) \pi_n^{(j)}(i).$$

Assume that the functions $\lambda_n(i, q)$ and $\mu_n(i, q)$ satisfy condition (4.53) of Theorem 4.5, and there exist values $\widehat{m}(j) > 0$ and functions $\widehat{b}(j, q)$ such that for any $j \in Y$, $q \geq 0$,

$$\widehat{m}_n(j) \longrightarrow \widehat{m}(j), \quad \widehat{b}_n(j, q) \longrightarrow \widehat{b}(j, q), \tag{8.102}$$

$S_n(0) \xrightarrow{\mathrm{P}} s_0$ and condition (8.100) holds. Then the following result is true.

STATEMENT 8.4. *If* $x_n(0) \in X_{j_0}$, *then under the conditions above the sequence* $(K(x_n(nt)), Q_n(nt)/n)$ *J-converges in the interval* $[0, T]$ *to the process* $(y(t, j_0), q(t, j_0, s_0))$, *where* $y(t, j_0)$ *is an MP defined above with the initial state* j_0 *and* $q(t, j_0, s_0)$ *is a solution of a differential equation with a random right-hand side:* $q(0, j_0, s_0) = s_0$ *and*

$$\mathrm{d}q\big(t, j_0, s_0\big) = \widetilde{b}\big(y\big(t, j_0\big), q\big(t, j_0, s_0\big)\big)\mathrm{d}t, \tag{8.103}$$

where $\widetilde{b}(j,q) = \widehat{b}(j,q)/\widehat{m}(q)$*, and* T *is any positive value such that* $q(t, j_0, s_0) > 0$ *for all* $t \in [0,T]$ *with probability one.*

NOTE 8.5. Consider an auxiliary SMP $x_n^{(j)}(t)$ with values in X_j given by the embedded MP $x_{nk}^{(j)}$ and sojourn times $\tau_n(i)$, $i \in X_j$. Denote by $\rho_n^{(j)}(i)$, $i \in X_j$, its stationary distribution. It is known that $\rho_n^{(j)}(i) = m_n(i)\pi_n^{(j)}(i)/\widehat{m}_n(j)$. Therefore function $\widetilde{b}(j,q)$ in (8.103) can be presented in the form:

$$\widetilde{b}(j,q) = \lim_{n\to\infty} \sum_{i\in X_j} \big(\lambda_n(i,q) - \mu_n(i,q)\big)\rho_n^{(j)}(i).$$

Similar results can be proved for the system with batch arrivals and service. Let the family of random variables $\{\eta_n(i,q), \kappa_n(i,q),\ i \in X,\ q \ge 0\}$ be given representing the batches of arriving calls and service batches. This means that if at time t, $x(t) = i$ and $Q_n(t) = Q$, then with the input rate $\lambda_n(i, Q/n)$ a batch of size $\eta_n(i, Q/n)$ can enter the system. Correspondingly, with the rate $\mu_n(i, Q/n)$ a batch of $\min\{Q, \kappa_n(i, Q/n)\}$ calls can complete service and leave the system. The sizes of arriving batches can be positive or negative representing the flows of negative customers, for example. Denote $g_n(i,q) = \mathbf{E}\eta_n(i,q)$, $v_n(i,q) = \mathbf{E}\kappa_n(i,q)$.

NOTE 8.6. Assume that the conditions of Statement 8.4 hold, variables $\eta_n(i,q)$ and $\kappa_n(i,q)$ are uniformly integrable in each bounded in q region, i.e., for any $N > 0$,

$$\lim_{L\to\infty} \limsup_{n\to\infty} \sup_{|q|\le N} \max_i \big\{\mathbf{E}\eta_n(i,q)\chi\big(\eta_n(i,q) > L\big) + \mathbf{E}\kappa_n(i,q)\chi\big(\kappa_n(i,q) > L\big)\big\} = 0,$$

functions $g_n(i,q)$ and $v_n(i,q)$ satisfy condition (4.53) of Theorem 4.5 and there exists the limiting function $G(j,q)$, $j \in Y$, $q \ge 0$, such that for any j, q,

$$G(j,q) = \lim_{n\to\infty} \sum_{i\in X_j} \big(\lambda_n(i,q)g_n(i,q) - \mu_n(i,q)v_n(i,q)\big)\rho_n^{(j)}(i).$$

Then the result of Statement 8.4 holds where in (8.103) function $\widetilde{b}(j,q)$ should be replaced by $G(j,q)$.

Using the same technique we can apply these results to retrial queues and state-dependent semi-Markov type queueing networks $(SM/M_{SM,Q}/1/\infty)^r$, $(M_{SM,Q}/M_{SM,Q}/1/\infty)^r$ with input flow and service depending on the state of some additional Markov or semi-Markov process admitting the asymptotic aggregation of the state space.

8.11. Analysis of flows of lost calls

To show other possibilities of the application of the results of this chapter let us consider the behavior of flows of lost calls in Markov queueing system $M_M/M_M/s/m$ with losses and Markov switching admitting the asymptotic aggregation of state space. Suppose that the family of non-negative functions $\lambda(i), \mu(i)$, $i \in X$, be given. Let $x_n(t)$ be a homogenous MP with values in X and transition rates $a_n(i,j)$, $i, i \in X$, $i \neq j$, $t \geq 0$. The system has s identical servers and m waiting places. The calls enter the system one at a time and if $x_n(t) = i$, then the instantaneous input rate is $\lambda(i)$ and the instantaneous service rate for each busy server is $\mu(i)$. Denote by $L_n(t)$ the probability of a loss of the call which entered the system at time t. Let $Z_n(t)$ be the total number of calls lost on the interval $[0,t]$.

Assume that the state space X of $x_n(t)$ is subdivided in the non-intersected regions X_y (see (8.51)) with small transition rates of the order $O(1/n)$ between them:

$$a_n(i,j) = a_0(i,j) + n^{-1}h_0(i,j)\big(1 + o_n(1)\big), \quad i,j \in X, \tag{8.104}$$

where for any $y \in Y$, $a_0(i,j) \equiv 0$ as $i \in X_y$, $j \notin X_y$. For any $y \in Y$ denote by $x^{(y)}(t)$, $t \geq 0$, an auxiliary MP with state space X_y and transition rates $a_0(i,j)$, $i,j \in X_j$, and assume that this process is irreducible with stationary distribution $\rho^{(y)}(i)$, $i \in X_y$. Define aggregated transition rates between regions

$$\widehat{a}_n(y,z) = \sum_{i \in X_y} \rho_n^{(y)}(i) \sum_{j \in X_z} h_0(x,j), \quad y \neq z,$$

and denote by $y(t, y_0)$, $t \geq 0$, an MP in Y with transition rates $\widehat{a}_n(y,z)$ and the initial state y_0.

For the analysis of this system let us introduce an auxiliary queueing system $M_M^{(y)}/M_M^{(y)}/s/m$ with losses given by the family of input and service rates $\{\lambda(i), \mu(i),\ i \in X_y\}$ and switched by an MP $x^{(y)}(t)$ as described above. Denote for this system by $G(y)$ a stationary probability for a call to be lost. As for this system the process $(x^{(y)}(t), Q^{(y)}(t))$ is an MP where $Q^{(y)}(t)$ is the value of queue, then $G(y)$ can be calculated by solving a system of linear equations for the stationary distribution of the system. Using the results on the convergence of the accumulating processes in aggregation scheme (see section 8.4.3) we can prove the following:

STATEMENT 8.5. *If $x_n(0) \in X_{y_0}$, then under our assumptions for any $t > 0$, $L_n(nt) \to \mathbf{E}\, G(y(t,y_0))$, and for any $T > 0$ the process $(K(x_n(nt)), Z_n(nt)/n)$ J-converges in the interval $[0,T]$ to the process $(y(t,y_0), \int_0^t G(y(u,y_0))\mathrm{d}u)$.*

Now consider the case when the system $M_M/M_m/s/m$ is operating in "fast" service conditions. This means that for any i, $\mu(i) = \mu_n(i)$, and $\mu_n(i) \to \infty$. Suppose that $\mu_n(i) = n^{1/(s+m)}c(i)$, where $c(i) > 0$, $i \in X$. Assume that $x_n(\cdot)$ satisfies the assumptions above and $\rho^{(y)}(i)$, $i \in X_y$ is a stationary distribution of the auxiliary MP

$x^{(y)}(t)$, $t \geq 0$, defined above in the previous section. For any $y \in Y$ denote

$$\Lambda^{(y)} = \frac{1}{s!s^m} \sum_{i \in X_y} \rho^{(y)}(i)\lambda(i)\left(\frac{\lambda(i)}{c(i)}\right)^{s+m},$$

where $\Lambda^{(y)}$ has a meaning of a stationary rate of lost calls in the region X_y.

Let $y(t, y_0)$ be an MP introduced above and $Z(t, y_0)$ be a doubly stochastic Poisson process switched by $y(t, y_0)$: if at time t, $y(t, y_0) = y$, then the instantaneous rate of jump is $\Lambda^{(y)}$.

STATEMENT 8.6. *Under our assumptions for any* $T > 0$ *process* $(K(x_n(nt)), Z_n(nt))$ *J-converges in interval* $[0, T]$ *to process* $(y(t, y_0), Z(t, y_0))$ *and also for any* $t > 0$, $nL_n(nt) \to \mathbf{E}G(y(t, y_0))$, *where*

$$G(y) = \frac{1}{s!s^m} \sum_{i \in X_y} \rho^{(y)}(i)\left(\frac{\lambda(i)}{c(i)}\right)^{s+m}.$$

The proof follows from the results on the convergence of accumulating processes in the aggregation scheme (section 8.4.3) and the results of sections 6.3.1 and 6.3.1.1.

These results mean that if a queueing system is switched by a Markov process admitting the asymptotic aggregation of its state space, then under appropriate scaling of time, different functionals defined on this system (queueing processes, flows of lost calls, etc.) can be approximated by corresponding functionals defined on a simpler queueing system switched by a Markov process with the aggregated state space and averaged transition probabilities.

Results for queueing systems can be extended using the same technique to state-dependent and also non-homogenous in time queueing networks. These results provide us with a new approach in analytic modeling of wide classes of queueing models with a hierarchic stochastic structure. For example, instead of simulation of the initial system we can use the approximate relation:

$$Q(nt) \approx nq(t, j_0, s_0) + \sqrt{n}\zeta(t, j_0),$$

and model the solutions of differential equations or diffusion processes in a Markov environment. Note also that the state space of a limiting switching environment corresponds to the regions of the initial environment. This provides the opportunity to decrease the dimension of the model.

8.12. Bibliography

[ANI 70] ANISIMOV V., "Limit distributions of functionals of a semi-Markov process given on a fixed set of states up to the time of first exit", *Soviet Math. Dokl.*, vol. 11, no. 4, p. 1002–1004, 1970.

[ANI 73] ANISIMOV V., "Asymptotic consolidation of the states of random processes", *Cybernetics*, vol. 9, no. 3, p. 494–504, 1973.

[ANI 74] ANISIMOV V., "Limit theorems for sums of random variables in an array of sequences defined on a subset of states of a Markov chain up to the exit time", *Theor. Probab. and Math. Stat.*, , no. 4, p. 1–12, 1974.

[ANI 75] ANISIMOV V., "Limit theorems for random processes and their applications to discrete summation schemes", *Theor. Probab. Appl.*, vol. 20, 1975.

[ANI 78] ANISIMOV V., "Limit theorems for switching processes and their applications", *Cybernetics*, vol. 14, no. 6, p. 917–929, 1978.

[ANI 79] ANISIMOV V., "Limit theorems for the composition of random processes", *Theor. Probab. and Math. Stat.*, vol. 17, p. 5–22, 1979.

[ANI 87a] ANISIMOV V., "Approximation of asymptotically consolidatable Markov processes", *Theor. Probab. and Math. Stat.*, vol. 34, p. 1–11, 1987.

[ANI 87b] ANISIMOV V., ZAKUSILO O. and DONTCHENKO V., *The elements of queueing theory and asymptotic analysis of systems*, Visca Scola (Russian), Kiev, Ukraine, 1987.

[ANI 88a] ANISIMOV V., "Estimates for deviations of transient characteristics of nonhomogenous Markov processes", *Ukrainian Math. J.*, vol. 40, no. 6, p. 588–592, 1988.

[ANI 88b] ANISIMOV V., "Limit theorems for switching processes", *Theor. Probab. and Math. Stat.*, vol. 37, p. 1–5, 1988.

[ANI 88c] ANISIMOV V., *Random Processes with Discrete Component. Limit Theorems*, Kiev University (Russian), Kiev, Ukraine, 1988.

[ANI 89a] ANISIMOV V. and SZTRIK J., "Asymptotic analysis of some complex renewable system operating in random environment", *European Journal of Operations Research*, vol. 41, p. 162–168, 1989.

[ANI 89b] ANISIMOV V. and SZTRIK J., "Asymptotic analysis of some controlled finite-source queueing systems", *Acta Cybernetica*, vol. 9, no. 1, p. 27–38, 1989.

[ANI 89c] ANISIMOV V. and SZTRIK J., "Reliability analysis of a complex renewable system with fast repair", *J. of Information Processing and Cybernetics*, vol. 25, no. 11/12, p. 573–583, 1989.

[ANI 97] ANISIMOV V., "Asymptotic analysis of switching queueing systems in conditions of low and heavy loading", in CHAKRAVARTHY S. and ALFA A., Eds., *Matrix-Analytic Methods in Stochastic Models*, vol. 183 of *Lecture Notes in Pure and Appl. Math.*, p. 241–260, Dekker, New York, 1997.

[ANI 98] ANISIMOV V., "Asymptotic analysis of stochastic models of hierarchic structure and applications in queueing models", in CHAKRAVARTHY S. and ALFA A., Eds., *Advances in Matrix Analytic Methods for Stochastic Models*, p. 237–259, Notable Publications, New Jersey, 1998.

[ANI 00a] ANISIMOV V., "Asymptotic analysis of reliability for switching systems in light and heavy traffic conditions", in LIMNIOS N. and NIKULIN M., Eds., *Recent Advances in Reliability Theory: Methodology, Practice, and Inference*, p. 119–133, Birkhäuser Boston, Massachusetts, 2000.

[ANI 00b] ANISIMOV V., "J-convergence for switching processes with rare perturbations to diffusion processes with Poisson type jumps", in KOROLYUK V., PORTENKO N. and SYTA H., Eds., *Skorokhod's Ideas in Probability Theory*, p. 81–98, Inst. of Math. Nat. Acad. Sci. of Ukraine, Kiev, 2000.

[ANI 02] ANISIMOV V., "Averaging in Markov models with fast Markov switches and applications to queueing models", *Annals of Operations Research*, vol. 112, no. 1, p. 63–82, 2002.

[ANI 04] ANISIMOV V., "Averaging in Markov models with fast semi-Markov switches and applications", *Communications in Statistics - Theory and Methods*, vol. 33, no. 3, p. 517–531, 2004.

[BIL 68] BILLINGSLEY P., *Convergence of Probability Measures*, Wiley, New York, 1968.

[IL' 99] IL'IN A., KHASMINSKII R. and YIN G., "Singularly Perturbed Switching Diffusions: Rapid Switchings and Fast Diffusions", *Journal of Optimization Theory and Applications*, vol. 102, no. 3, p. 555–591, 1999.

[KEM 76] KEMENY J. and SNELL J., *Finite Markov Chains*, Springer-Verlag, New York, 1976.

[KOR 69] KOROLYUK V., "The asymptotic behavior of the sojourn time of a semi-Markov process in a subset of the states", *Ukrainian Math. J.*, vol. 21, p. 705–707, 1969.

[KOR 93] KOROLYUK V. and TURBIN A., *Mathematical Foundation of the State Lumping of Large Systems*, Kluwer, Dordrecht, 1993.

[KOR 94] KOROLYUK V. and SWISHCHUK A., *Random Evolutions*, Kluwer, Dordrecht, 1994.

[KOR 99] KOROLYUK V. and KOROLYUK V., *Stochastic Models of Systems*, Kluwer, Dordrecht, 1999.

[KOR 00] KOROLYUK V. and LIMNIOS N., "Evolutionary systems in an asymptotic split phase space", in LIMNIOS N. and NIKULIN M., Eds., *Recent Advances in Reliability Theory: Methodology, Practice and Inference*, p. 145–161, Birkhäuser Boston, Massachusetts, 2000.

[KOR 04] KOROLYUK V. and LIMNIOS N., "Average and diffusion approximation for evolutionary systems in an asymptotic split phase state", *Ann. Appl. Prob.*, vol. 14, no. 1, p. 489–516, 2004.

[KOR 05] KOROLYUK V. and LIMNIOS N., *Stochastic Systems in Merging Phase Space*, World Scientific, Singapore, 2005.

[KOV 75] KOVALENKO I., *Investigation on the Relliability Analysis of Complex Systems*, Naukova Dumka (Russian), Kiev, 1975.

[SKO 56] SKOROKHOD A., "Limit theorems for random processes", *Theory Prob. Appl.*, vol. 1, p. 289–319, 1956.

[SZT 91] SZTRIK J. and KOUVATSOS D., "Asymptotic analysis of a heterogeneous multiprocessor system in a randomly changing environment", *IEEE Transactions on Software Engineering*, vol. 17, no. 10, p. 1069–1075, 1991.

[YIN 00] YIN G., ZHANG Q. and BADOWSKI G., "Asymptotic properties of a singularly perturbed Markov chain with inclusion of transient states", *Annals of Applied Probability*, vol. 10, no. 2, p. 549–572, 2000.

[YIN 03] YIN G., ZHANG Q. and BADOWSKI G., "Discrete-time singularly perturbed Markov chains: Aggregation, occupation measures, and switching diffusion limit", *Advances in Applied Probability*, vol. 35, p. 449–476, 2003.

Chapter 9

Aggregation in Markov Models with Fast Markov Switching

In this chapter we study the approximation of Markov type queueing models with fast Markov switching by Markov models with averaged transition rates. First, we prove an averaging principle for the two-component Markov process $(x_n(t), \zeta_n(t))$ in the following form: if a component $x_n(\cdot)$ has fast switching and satisfies the asymptotic mixing conditions, then component $\zeta_n(\cdot)$ J-converges in Skorokhod space to a Markov process with transition rates averaged by some quasi-stationary measures constructed by $x_n(\cdot)$. The convergence of a stationary distribution of $(x_n(\cdot), \zeta_n(\cdot))$ is studied as well. The presentation is based on [ANI 02a].

These results differ from the results on averaging principle for SP in Chapter 4 as component $\zeta_n(\cdot)$ is not growing and averaging is provided for transition rates of a switching component. This setting also differs from the models of the asymptotic aggregation of MP and SMP, as component $x_n(\cdot)$ itself is not generally an MP. To prove the main results we use Theorem 8.3, section 8.3 about the weak convergence in the class of SPs. The results are used in the approximation of state-dependent queueing systems of the type $M_{M,Q}/M_{M,Q}/m/N$ with fast Markov switching by simpler models without switching environment.

9.1. Introduction

At the investigation of hierarchic state-dependent queueing models, communication and manufacturing systems, etc., we come to a necessity to study models operating in different scales of time (slow and fast) and possibly under the influence of a random environment. Different examples can be found for example in [BOL 98]. Taking into account a high dimension and a complex structure, exact analytic solutions for these systems can be obtained only for special exceptional cases. Many of the results related to the analysis of Markov type queueing models use the technique of quasi-Birth-and-Death processes [NEU 89, NEU 89] and deal with matrix-analytic equations for queueing characteristics in a steady-state regime, for example [LUC 94a, LUC 94b, RAM 98]. These equations in most cases have a rather complicated form, and numerical methods and inversion algorithms are used in order to obtain some queueing characteristics. Therefore, asymptotic methods play an important role at the investigation and approximate analytic modeling.

In this chapter we develop an asymptotic approach, which can be efficiently applied to the approximation of queueing models with fast Markov switching by Markov queuing models of a simpler structure with averaged transition rates. This approach provides us with the opportunity of reducing essentially the dimension and the complexity of the initial model and studying transient and steady-state regimes, as well.

We study a sequence of two-component MP $(x_n(t), \zeta_n(t))$, $t \geq 0$, where $n \to \infty$ at the assumption that the transition rates of the first component $x_n(\cdot)$ are considerably larger compared to the transition rates of the second component $\zeta_n(\cdot)$ (we say that $x_n(\cdot)$ has fast switching). If in addition $x_n(\cdot)$ satisfies some form of asymptotic mixing condition, it is proved that the component $\zeta_n(\cdot)$ J-converges in Skorokhod space to an MP with transition rates averaged by a quasi-stationary distribution constructed by $x_n(\cdot)$. In particular, this model describes the behavior of an MP with local rates depending on a fast MP $x_n(\cdot)$ (switching by a fast Markov environment). The convergence of a stationary distribution for $(x_n(\cdot), \zeta_n(\cdot))$ is studied as well. The method of investigation uses the equivalent representation of a two-component MP as a process with Markov switching and the asymptotic technique for Markov processes with mixing conditions.

In section 9.4 these results are effectively applied to the approximation of Markov type queueing models with transition rates of different order, in particular, state-dependent queueing models in a fast Markov environment. Some classes of models of this type are studied in the remaining sections of this chapter.

Related results based on the approximating aggregation technique for the analysis of a stationary distribution of nearly decomposed MPs with applications to queues are considered in [COU 77, BOL 98]. Analysis of transient probabilities is given in [BOB 86]. Our approach deals with the convergence of the queueing processes. This also gives the possibility of studying various functionals and performance characteristics and the behavior of transition and stationary probabilities, as well. The convergence of aggregated processes is rigorously proved. This approach is simpler in some sense compared to the one developed in [COU 77], because averaging by stationary probabilities is provided using transition rates only within each aggregated block. Moreover, each block in our case can also be nearly decomposed in another scale of time. Some numerical results and the comparison of these approaches are provided in section 9.6.

Note that the transient behavior of the processes generated by queues is mainly investigated in heavy traffic and overloaded regimes, for instance see survey [WIL 96, ANI 02b]. Approximation, presented in this chapter, has another nature, because in our case the queueing process in not asymptotically growing. Related in some sense models on averaging of dynamic systems with fast Markov switches using a martingale technique are considered in the books [ETH 86, KUS 90, SKO 89]. The case of fast semi-Markov switching using asymptotic methods for switching processes is studied in [ANI 95]. Some specific results dealing with the convergence of aggregated processes for Markov systems in the fast ergodic Markov environment with applications to queueing systems are considered in [ANI 78, ANI 98].

9.2. Markov models with fast Markov switching

Consider a class of MP with Markov switching which later will be used in queueing applications.

9.2.1. *Markov processes with Markov switching*

Let $\{x_k(t,i),\ t \geq 0\}$, $i \in Z$, $k \geq 0$, be the jointly independent at different i, k, families of homogenous MPs in continuous time with transition rates not depending on k. Process $x_k(\cdot, i)$ takes values in a discrete set X_i and is defined by transition rates $b^{(i)}(x,y)$, $x, y \in X_i$, $x \neq y$. Z is also a discrete set, $Z = \{0, 1, 2, \ldots\}$. We assume that

$$b^{(i)}(x) = \sum_{y \neq x} b^{(i)}(x,y) < \infty, \quad x \in X_i,\ i \in Z.$$

Also let the family of non-negative functions $\{a(x,i,y,j),\ x \in X_i,\ y \in X_j,\ i,j \in Z,\ i \neq j\}$ and $\{c(x,i,y,j),\ x \in X_i,\ y \in X_j,\ i,j \in Z,\ j \neq i\}$ be given. Assume that for any $x \in X_i$, $i \in Z$,

$$\sum_{y \in X_j,\ j \neq i} a(x,i,y,j) = a(x,i) < \infty, \qquad \sum_{y \in X_j,\ j \neq i} c(x,i,y,j) = c(x,i) \leq 1.$$

Using introduced families we construct a two-component MP $(x(t), \zeta(t))$, $t \geq 0$, with values in the space $\{(x,i),\ x \in X_i,\ i \in Z\}$ in the following way: a component $x(\cdot)$ in the interval between kth and $k+1$th jumps of process $\zeta(\cdot)$ is operating as an MP $x_k(\cdot,\cdot)$ depending on the current state of $\zeta(\cdot)$, and $\zeta(\cdot)$ is operating as an MP switched by $x(\cdot)$, where the jumps of $\zeta(\cdot)$ may happen in the intervals between jumps of $x(\cdot)$ and at the times of jumps of $x(\cdot)$, as well. More specifically, let the initial value $(x(0), \zeta(0)) = (x_0, i_0)$ be given. While $\zeta(\cdot) = i_0$, component $x(\cdot)$ is operating as an MP $x_0(t, i_0)$ with the initial state x_0. If in the interval $[0, t]$ component

$\zeta(\cdot)$ has no jumps and at time t, $(x(t), \zeta(t)) = (x, i_0)$, then in the interval $[t, t+h]$ with probability $a(x, i_0, x_1, i_1)h + o(h)$ process $(x(\cdot), \zeta(\cdot))$ can jump to state (x_1, i_1), $i_1 \neq i_0$. If this happens, then in the next time interval $\zeta(\cdot) = i_1$ and component $x(\cdot)$ is operating as an MP $x_1(t, i_1)$ starting from state x_1 until the next jump time of $\zeta(\cdot)$, and so on. Furthermore, denote by $t_1 < t_2 < \cdots$ the times of sequential jumps of $x_0(t, i_0)$. Then, while the component $x(\cdot)$ is operating as an MP $x_0(t, i_0)$, at any instant of time t_k with probability $c(x(t_k - 0), i_0, x_1, i_1)$ process $(x(\cdot), \zeta(\cdot))$ can jump to state (x_1, i_1), $i_1 \neq i_0$. If this happens, then in the next time interval component $x(\cdot)$ is operating as an MP $x_1(t, i_1)$ with the initial state x_1, and in this interval $\zeta(\cdot) = i_1$ until the next jump time of $\zeta(\cdot)$, and so on.

It is easy to prove that by definition process $(x(\cdot), \zeta(\cdot))$ is equivalent to a two-component MP with state space $\{(x, i),\ x \in X_i,\ i \in Z\}$ and the following transition rates from state (x, i) to (y, j):

$$b\big((x,i),(y,j)\big) = \begin{cases} b^{(i)}(x,y)(1 - c(x,i)), & j = i,\ y \neq x; \\ a(x,i,y,j) + b^{(i)}(x)c(x,i,y,j), & j \neq i. \end{cases}$$

The proof is based on the following elementary fact. Consider an MP $y(t)$ with three states $\{1, 2, 3\}$. Suppose that $y(0) = 1$, in state $\{1\}$, $y(\cdot)$ spends an exponential time with parameter λ and then jumps either with probability p to state $\{2\}$, or with probability $q = 1 - p$ to state $\{3\}$, respectively, where the states $\{2, 3\}$ are absorbing states. Then this MP is equivalent to an MP $\widetilde{y}(t)$ with states $\{1, 2, 3\}$, initial state $\{1\}$ and transition rates $\lambda_{12} = \lambda p$, $\lambda_{13} = \lambda q$, where $\{2, 3\}$ are absorbing states.

We call $(x(t), \zeta(t))$, $t \geq 0$, an MP with Markov switching (MPMS). It is a special subclass of switching processes and the processes with Markov switching, see section 1.2.5. If for any i, $b^{(i)}(x, y) = b(x, y)$ (the rates do not depend on index i), then component $\zeta(\cdot)$ corresponds to an MP in a Markov environment with transition rates $b(x, y)$. This is a special case of Markov random evolutions [PIN 75].

Now consider a two-component MP $(\widetilde{x}(t), \widetilde{\zeta}(t))$, $t \geq 0$, with state space $\{(x, i),\ x \in X_i,\ i \in Z\}$ and transition rates $b((x, i), (y, j))$ from state (x, i) to (y, j), where $y \neq x$ as $j = i$ and show how to represent this process as an equivalent MPMS. At each i denote by $x_k^{(i)}(t), t \geq 0$, an auxiliary MP with state space X_i and transition rates $b^{(i)}(x, y) = b((x, i), (y, i))$, $x, y \in X_i$, $y \neq x$. Put $a(x, i, y, j) = b((x, i), (y, j))$, $j \neq i$. Let us construct an MPMS $(x(\cdot), \zeta(\cdot))$ as above by the family of MPs $x_k^{(i)}(\cdot)$ and transition rates $a(x, i, y, j)$ for the component $\zeta(\cdot)$ (in this case $c(x, i) \equiv 0$, $x \in X_i$, $i \in Z$). If the initial values $(\widetilde{x}(0), \widetilde{\zeta}(0))$ and $(x(0), \zeta(0))$ have the same distributions, then by definition processes $(\widetilde{x}(\cdot), \widetilde{\zeta}(\cdot))$ and $(x(\cdot), \zeta(\cdot))$ are equivalent (have the same finite-dimensional distributions).

A representation of $(\widetilde{x}(\cdot), \widetilde{\zeta}(\cdot))$ in the terms of the equivalent MPMS is essentially used in the asymptotic investigation. If the transition rates of $\widetilde{\zeta}(\cdot)$ are asymptotically small, then we can investigate transitions of $\widetilde{\zeta}(\cdot)$ by analyzing the flow of rare events constructed in each domain X_i on the auxiliary MP $x_k^{(i)}(\cdot)$. This representation essentially reduces the dimension and in the case where $x_k^{(i)}(\cdot)$ satisfies the asymptotically mixing condition, provides the possibility of using asymptotic results for accumulative processes (in particular, flows of events) defined on MPs. If $b((x,i),(y,j)) = 0$ as $|j-i| > 1$, then $(\widetilde{x}(\cdot), \widetilde{\zeta}(\cdot))$ is a quasi-Birth-and-Death process [NEU 81, NEU 89].

9.2.2. *Markov queueing systems with Markov type switching*

Consider as an example a state-dependent system $M_{M,Q}/M_{M,Q}/1/\infty$ with Markov type switching which is defined as follows. There is one server with an infinite buffer. Calls arrive one at a time and wait in the queue according to the FIFO discipline. Let non-negative functions $\{\lambda(x,i), \mu(x,i), \alpha_A(x,i), \alpha_S(x,i), b^{(i)}(x,y), x \in X_i,\ y \in X_i,\ x \neq y,\ i \geq 0\}$ be given. Here $\alpha_A(x,i) + \alpha_S(x,i) \leq 1$, $x \in X_i$, $i \geq 0$, and X_i are some discrete sets. Let $Q(t)$, $t \geq 0$, be the total number of calls in the system at time t. The system operates as a two-component MP $(x(t), Q(t))$, $t \geq 0$, as follows. Let the initial value $(x(0), Q(0))$ be given. Further, if at time t, $(x(t), Q(t)) = (x,i)$, then the local arrival rate is $\lambda(x,i)$, the local service rate is $\mu(x,i)$ ($\mu(x,0) = 0$), and process $x(t)$ has a local transition rate $b^{(i)}(x,y)$ from state x to state y, where $x, y \in X_i$, $y \neq x$. If $Q(\cdot)$ jumps to state j, then transition rates of $x(t)$ immediately change to $b^{(j)}(x,y)$, $x, y \in X_j$, $y \neq x$. Let t_1 be the time of the first jump of $x(\cdot)$. If $(x(t_1 - 0), Q(t_1 - 0)) = (x,i)$, then at time t_1 either an additional call may enter the system with probability $\alpha_A(x,i)$, or a call on service may complete service with probability $\alpha_S(x,i)$ (no changes with probability $1 - \alpha_A(x,i) - \alpha_S(x,i)$). After service completion a call leaves the system. We assume here that $\alpha_S(x,0) = 0$ and there are no additional transitions of $x(\cdot)$ at times of arrivals and completion service.

Note that this system is a generalization of the system $M_{M,Q}/M_{M,Q}/1/\infty$ considered in section 2.2.1.1 and differs from other Markov queueing systems with Markov switching considered in Chapter 5, as in this case the transition rates of the switching process $x(\cdot)$ also depend on the value of queue, therefore this is the case of feedback and $x(\cdot)$ in general is not an MP.

Denote $b^{(i)}(x) = \sum_{y \neq x} b^{(i)}(x,y)$. Suppose that for any $x \in X_i$, $i \in Z$, $b^{(i)}(x) \leq C_i < \infty$. By definition, $(x(\cdot), Q(\cdot))$ is an MP with the following transition rates from state (x,i) to (y,j):

$$b\big((x,i),(y,j)\big) = \begin{cases} b^{(i)}(x,y)\big(1 - \alpha_A(x,i) - \alpha_S(x,i)\big), & j = i,\ y \neq x; \\ \lambda(x,i) + b^{(i)}(x)\alpha_A(x,i), & j = i+1,\ y = x; \\ \mu(x,i) + b^{(i)}(x)\alpha_S(x,i), & j = i-1,\ y = x; \\ 0, & \text{otherwise,} \end{cases}$$

(at $i = 0$, $\mu(x, 0) = 0$, $\alpha_S(x, 0) = 0$). Note that the process $(x(t), Q(t))$, $t \geq 0$, can be represented as an MPMS (see the previous section).

In particular, if for all $i \in Z$, $X_i = X$ and $b^{(i)}(x, y) \equiv b(x, y)$, $x, y \in X$, $x \neq y$, then $x(\cdot)$ is an MP with state space X and transition rates $b(x, y)$, $x \neq y$, and we obtain the queueing model in a Markov environment [NEU 81].

In a similar way we can describe state-dependent networks with Markov switches, batch Markov arrival process and service, some classes of state-dependent retrial models with Markov switches, etc.

9.2.3. *Averaging in the fast Markov type environment*

Now we study a two-component MP $(x(t), \zeta(t))$, $t \geq 0$, with the assumption that component $x(\cdot)$ has fast switching (large transition rates) compared to $\zeta(\cdot)$. Consider in a triangular scheme a sequence of two-component MPs $(x_n(t), \zeta_n(t))$, $t \geq 0$, with state space $\{(x, i),\ x \in X_i,\ i \in Z\}$ and transition rate from state (x, i) to (y, j), $b_n((x, i), (y, j))$, where $y \neq x$ as $j = i$ (transition rates depend on some scaling factor n, $n \to \infty$). For any $i \in Z$ denote $b_n^{(i)}(x, y) = b_n((x, i), (y, i))$, $x, y \in X_i$, $x \neq y$ (transition rates at level i). Consider an auxiliary MP $x_n^{(i)}(t)$, $t \geq 0$, with state space X_i and transition rates $b_n^{(i)}(x, y)$, $x, y \in X_i$, $x \neq y$. For any fixed $x \in X_i$, $i \in Z$, put $b_n^{(i)}(x) = \sum_{y \in X_i,\ y \neq x} b_n^{(i)}(x, y)$. Let us introduce a uniformly strong mixing coefficient for MP $x_n^{(i)}(\cdot)$: for any $u > 0$ denote

$$\varphi_n^{(i)}(u) = \sup_{x,y,A} \Big| \mathbf{P}\Big\{x_n^{(i)}(u) \in A \mid x_n^{(i)}(0) = x\Big\} - \mathbf{P}\Big\{x_n^{(i)}(u) \in A \mid x_n^{(i)}(0) = y\Big\}\Big|.$$

Assume that $x_n^{(i)}(\cdot)$ satisfies the following condition: there exist a scaling factor V_n, $V_n \to \infty$ as $n \to \infty$, and constants L_i, $i \in Z$, such that for some q, $0 \leq q < 1$, and for any $i \in Z$,

$$\varphi_n^{(i)}\big(L_i/V_n\big) \leq q,\ n > 0. \tag{9.1}$$

Condition (9.1) means that for all $x \in X_i$ the exit rates $b_n^{(i)}(x)$ are asymptotically large (fast switching) and $x_n^{(i)}(\cdot)$ is asymptotically mixing in any fixed interval.

EXAMPLE 9.1. Suppose that X_i is a finite set, $b_n^{(i)}(x, y) = V_n(\widetilde{b}_0^{(i)}(x, y) + o_n(1))$, $x, y \in X_i$, where $o_n(1) \to 0$, $\sum_{y \in X_i,\ y \neq x} \widetilde{b}_0^{(i)}(x, y) > 0$ for each $x \in X_i$, and an MP $\widetilde{x}_0^{(i)}(\cdot)$ with transition rates $\widetilde{b}_0^{(i)}(x, y)$ is irreducible. Then (9.1) is satisfied.

EXAMPLE 9.2. Let $X_i = \{1, 2, 3, 4\}$, and the generator of the process $x_n^{(i)}(\cdot)$ is represented in the form:

$$B_n^{(i)} = \begin{pmatrix} -b_n(1) & V_n^2 b_{12} & V_n b_{13} & V_n b_{14} \\ V_n^2 b_{21} & -b_n(2) & V_n b_{23} & V_n b_{24} \\ V_n b_{31} & V_n b_{32} & -b_n(3) & V_n^2 b_{34} \\ V_n b_{41} & V_n b_{42} & V_n^2 b_{43} & -b_n(4) \end{pmatrix},$$

where entries $b_n(j)$ are the sums of other elements in the corresponding row. If $b_{12}b_{21} > 0$, $b_{34}b_{43} > 0$, $b_{13} + b_{14} + b_{23} + b_{24} > 0$, $b_{31} + b_{32} + b_{41} + b_{42} > 0$, then condition (9.1) is also satisfied. In this case X_i consists of two classes $\{1, 2\}$ and $\{3, 4\}$ with transition rates $O(V_n^2)$ in each class, and transition rates between these classes $O(V_n)$.

More general cases, when the state space forms S-set (asymptotically connected set, see section 6.2), are considered in [ANI 98, ANI 00].

If condition (9.1) is satisfied, then the number of jumps of component $x_n(\cdot)$ in each finite interval tends in probability to infinity. In this case we can asymptotically average the rates of $\zeta_n(\cdot)$ in each domain X_i by a corresponding quasi-stationary distribution and prove that the component $\zeta_n(\cdot)$ converges in each finite interval $[0, T]$ to an MP with averaged transition rates. Put

$$b_n(x, i) = \sum_{y \in X_j,\ j \in Z,\ j \neq i} b_n\big((x, i), (y, j)\big), \quad x \in X_i,\ i \in Z. \tag{9.2}$$

Let there exist constants C_i such that for any $n > 0$,

$$\sup_{x \in X_i} b_n(x, i) \leq C_i < \infty, \quad i \in Z. \tag{9.3}$$

Condition (9.3) means that the transition rates between domains X_i are bounded. Together with condition (9.1) this means that transitions between states within each domain X_i happen considerably faster rather than transitions between different domains X_i.

If condition (9.1) holds, then MP $x_n^{(i)}(t)$ at each $n > 0$ has an ergodic (stationary) distribution $\rho_n^{(i)}(x)$, $x \in X_i$. Denote

$$b_n\big((x, i), j\big) = \sum_{y \in X_j} b_n\big((x, i), (y, j)\big),$$

$$\widehat{a}_n(i, j) = \sum_{x \in X_i} \rho_n^{(i)}(x) b_n\big((x, i), j\big), \quad j \neq i,$$

$$\widehat{a}_n(i) = \sum_{j \in Z, j \neq i} \widehat{a}_n(i, j).$$

Assume that, as $n \to \infty$, the following condition holds:

A) there exist finite values $a_0(i,j)$, $a_0(i)$, $i,j \in Z$, $j \neq i$, such that $a_0(i) = \sum_{j\neq i} a_0(i,j)$, and for any $i,j \in Z$, $i \neq j$,

$$\widehat{a}_n(i,j) \longrightarrow a_0(i,j), \quad \widehat{a}_n(i) \longrightarrow a_0(i).$$

Let $\zeta_0(t,i_0)$, $t \geq 0$, be an MP in Z with transition rates $a_0(i,j)$, $i \neq j$, and the initial state i_0. According to relation $a_0(i) = \sum_{j\neq i} a_0(i,j)$ MP $\zeta_0(\cdot,i_0)$ is conservative (for any state, the sum of probabilities of jumps to all other states in Z is equal to one).

We say that a process is regular, if it almost surely has a finite number of jumps in any finite interval. The following theorem holds (the proof is given in section 9.3).

THEOREM 9.1. *Assume that* $(x_n(0), \zeta_n(0)) = (x_0, i_0)$, $x_0 \in X_{i_0}$, *the process* $\zeta_0(\cdot, i_0)$ *is regular and conditions A), (9.1), (9.3) hold. Then in any interval* $[0,T]$, $\zeta_n(\cdot)$ *J-converges to* $\zeta_0(\cdot, i_0)$ *as* $n \to \infty$.

9.2.4. *Approximation of a stationary distribution*

Results of section 9.2.3 deal with the approximation of the distributions of the process in any finite interval $[0,T]$. We now study the approximation of the stationary distribution. Consider the sequence of MPs $(x_n(t), \zeta_n(t))$, $t \geq 0$, introduced in section 9.2.3 with values in $\{(x,i),\ x \in X_i,\ i \in Z\}$ and transition rates from state (x,i) to (y,j), $b_n((x,i),(y,j))$, ($y \neq x$ as $j = i$). We keep all previous notation. Let $\zeta_0(t)$, $t \geq 0$, be a regular MP given by transition rates $a_0(i,j)$, $i,j \in Z$, $i \neq j$. Denote by $\{\rho_n(x,i),\ x \in X,\ i \in Z\}$ a stationary distribution of $(x_n(\cdot), \zeta_n(\cdot))$ (if it exists). Assume first that X_i and Z are finite sets.

THEOREM 9.2. *Let conditions A), (9.1), (9.3) hold, for any* $i \in Z$ *there exist limits*

$$\rho_0^{(i)}(x) = \lim_{n\to\infty} \rho_n^{(i)}(x), \quad x \in X_i,$$

and exist values $c_i > 0$ *such that* $\min_{x\in X_i} \liminf_{n\to\infty} b_n(x,i) \geq c_i$. *Also let MP* $\zeta_0(\cdot)$ *be ergodic with stationary distribution* $\Pi_0(i)$, $i \in Z$. *Then at large enough* n, $(x_n(\cdot), \zeta_n(\cdot))$ *is ergodic and*

$$\lim_{n\to\infty} \rho_n(x,i) = \rho_0^{(i)}(x)\Pi_0(i), \quad x \in X_i,\ i \in Z. \tag{9.4}$$

The proof of Theorem 9.2 is given in section 9.3. Note that a multiplicative form of a limiting stationary distribution in Theorem 9.2 is in agreement with the results on the aggregation of finite MP [COU 77].

If we assume that X_i are infinite sets and, in addition, for any $i \in Z$, as $n \to \infty$, $\widehat{a}_n(i,(y,j)) \to a_0(i,(y,j))$, where $\sum_{y\in X_j} a_0(i,(y,j)) = a_0(i,j)$, then the result of Theorem 9.2 is also valid.

Now consider the case when Z is infinite but sets X_i are finite. Denote by $T_{n1} < T_{n2} < \cdots$ the times of sequential jumps of $\zeta_n(\cdot)$ and put $(X_{nk}, Z_{nk}) = (x_n(T_{nk}), \zeta_n(T_{nk}+0))$, $k \geq 1$. Then (X_{nk}, Z_{nk}) is the embedded Markov chain for $(x_n(\cdot), \zeta_n(\cdot))$. Let us fix a state i_0 and denote by $\nu_n(x_0, i_0)$ a return time to domain (X_{i_0}, i_0) for the process (X_{nk}, Z_{nk}) given that $(X_{n1}, Z_{n1}) = (x_0, i_0)$. The proof of the following theorem is given in section 9.3.

THEOREM 9.3. *Let there exist constants $c > 0$, $C > 0$, such that the conditions of Theorem 9.2 hold, where for any $i \in Z$, $c_i \geq c$, and for any n, $\mathbf{E}\nu_n(x_0, i_0)^2 < C < \infty$. Then at large enough n, $\zeta_n(\cdot)$ is ergodic and relation (9.4) holds.*

Condition $\mathbf{E}\nu_n(x_0, i_0)^2 < C$ can be verified in particular applications, for instance see [ANI 01]. In the following section we check this condition for a system $M_{M,Q}/M_{M,Q}/1/\infty$.

9.3. Proofs of theorems

9.3.1. *Proof of Theorem 9.1*

First, we represent $(x_n(\cdot), \zeta_n(\cdot))$ as an MPMS using the auxiliary MPs $x_n^{(i)}(\cdot)$ as was shown in section 9.2.1. In our case $a(x, i, y, j) = b_n((x, i), (y, j))$, $x \in X_i$, $y \in X_j$, $j \neq i$, and for all x, i, $c(x, i) = 0$ (there are no jumps of component $\zeta_n(\cdot)$ at the times of jumps of $x_n(\cdot)$). Denote by $T_{n1} < T_{n2} < \cdots$ the times of sequential jumps of $\zeta_n(\cdot)$ and put $(X_{nk}, Z_{nk}) = (x_n(T_{nk}), \zeta_n(T_{nk}+0))$, $k \geq 1$, where (X_{nk}, Z_{nk}) is the embedded Markov chain for $(x_n(\cdot), \zeta_n(\cdot))$. Denote

$$P_n(x, i, t, j) = \mathbf{P}\big(T_{n2} - T_{n1} \leq t,\ Z_{n2} = j \mid \big(X_{n1}, Z_{n1}\big) = (x, i)\big), \quad j \neq i,\ t > 0.$$

First, we prove that for any $i, j \in Z$, $i \neq j$, $t > 0$,

$$\sup_{x \in X_i} \big|P_n(x, i, t, j) - \big(1 - e^{-a_0(i)t}\big)a_0(i, j)/a_0(i)\big| \longrightarrow 0. \tag{9.5}$$

Given that $x_n(0) = x$, denote $A_n(t, x, i) = \int_0^t b_n(x_n^{(i)}(u), i)\mathrm{d}u$. Then for any $t > 0$, $j \neq i$, we have a representation:

$$P_n(x, i, t, j) = \mathbf{E}\int_0^t \exp\big\{-A_n(u, x, i)\big\} b_n\big(\big(x_n^{(i)}(u), i\big), j\big)\mathrm{d}u. \tag{9.6}$$

Now we prove the following auxiliary statement: if conditions (9.1), (9.3) hold and $x_n^{(i)}(0) = x$, then for any bounded measurable function $f(x)$, $x \in X_i$, for any $t \geq 0$ uniformly in $x \in X_i$,

$$G_n(x, t) = \int_0^t f\big(x_n^{(i)}(u)\big)\mathrm{d}u - t\sum_{y \in X_i} \rho_n^{(i)}(y) f(y) \xrightarrow{\text{P}} 0. \tag{9.7}$$

If for some $r > 0$, $\varphi_n^{(i)}(r) \le q$, then, as is known (for example [DOO 53]), for any $t > 0$, $\varphi_n^{(i)}(t) \le q^{t/r-1}$. Thus, condition (9.1) implies

$$\varphi_n^{(i)}(t) \le q^{tV_n/L_i-1}, \quad t > 0. \tag{9.8}$$

Denote $K_f = \sup_y |f(y)|$. Using the inequality $|\int f(y)P(\mathrm{d}y) - \int f(y)Q(\mathrm{d}y)| \le 2K_f \sup_A |P(A) - Q(A)|$, which is true for any bounded real function $f(\cdot)$ and probability measures $P(\cdot), Q(\cdot)$, and properties of $\varphi_n^{(i)}(\cdot)$ we obtain the following relations:

$$\begin{aligned}
&\Big|\mathbf{E}f\big(x_n^{(i)}(t)\big) - \sum_{y\in X_i} \rho_n^{(i)}(y)f(y)\Big| \le 2K_f\varphi_n^{(i)}(t), \quad t > 0;\\
&\Big|\mathbf{E}\big\{\big(f\big(x_n^{(i)}(u)\big) - \mathbf{E}f\big(x_n^{(i)}(u)\big)\big)\big(f\big(x_n^{(i)}(v)\big) - \mathbf{E}f\big(x_n^{(i)}(v)\big)\big)\big\}\Big|\\
&\qquad \le 8K_f\varphi_n^{(i)}(v-u), \quad u < v.
\end{aligned} \tag{9.9}$$

Therefore, after some algebra we obtain

$$\begin{aligned}
\big|\mathbf{E}G_n(x,t)\big| &\le 2K_f \int_0^t \varphi_n^{(i)}(u)\mathrm{d}u \le C_n(f) \longrightarrow 0,\\
\mathbf{Var}G_n(x,t) &\le 16K_f \int_{0\le u\le v\le t} \varphi_n^{(i)}(v-u) \le 8tC_n(f) \longrightarrow 0,
\end{aligned}$$

where $C_n(f) = 2K_f\alpha^{-1}V_n^{-1}(1 - e^{-\alpha V_n t})$, $\alpha = -\ln q/L_i$. These relations imply (9.7). Note that as the function $G_n(x,t)$ is continuous in t uniformly in n, then (9.7) holds uniformly in t in any bounded region.

Using (9.7) and condition A), we find that for any $t > 0$ uniformly in $x \in X_i$, $u \le t$, $A_n(u,x,i) \xrightarrow{\mathrm{P}} a_0(i)u$. Now, using a stepwise approximation, we can prove that for any continuous function $h(t)$,

$$\int_0^t h(u)b_n\big(\big(x_n^{(i)}(u),i\big),j\big)\mathrm{d}u - \int_0^t h(u)\widehat{a}_n(i,j)\mathrm{d}u \xrightarrow{\mathrm{P}} 0, \quad t > 0. \tag{9.10}$$

It follows from the relations above that

$$\begin{aligned}
&\Big|\int_0^t \exp\big\{-A_n(u,x,i)\big\}b_n\big(\big(x_n^{(i)}(u),i\big),j\big)\mathrm{d}u - \int_0^t \exp\big\{-a_0(i)u\big\}a_0(i,j)\mathrm{d}u\Big|\\
&\quad\le \int_0^t \big|\exp\big\{-A_n\big(u,x,i\big)\big\} - \exp\big\{-a_0(i)u\big\}\big|b_n\big(\big(x_n^{(i)}(u),i\big),j\big)\mathrm{d}u\\
&\qquad + \Big|\int_0^t \exp\big\{-a_0(i)u\big\}\big(b_n\big(\big(x_n^{(i)}(u),i\big),j\big) - a_0(i,j)\big)\mathrm{d}u\Big| \xrightarrow{\mathrm{P}} 0.
\end{aligned}$$

As the function e^{-z} is continuous and bounded in the region $z \geq 0$ and relation (9.3) holds, then the convergence in probability implies the convergence of expectations and relation (9.5) is proved.

Relation (9.5) means that the distribution of the variable $T_{n2} - T_{n1}$ given that $(X_{n1}, Z_{n1}) = (x, i)$ weakly converges to the exponential distribution, and it does not depend on state $\{x\}$ and on the next transition of $\zeta_n(\cdot)$. This implies the weak convergence of finite dimensional distributions of $\zeta_n(\cdot)$ to corresponding distributions of an MP $\zeta_0(\cdot, i_0)$ as follows from Theorem 8.3 in section 8.3 (see also [ANI 78]). As $\zeta_0(\cdot, i_0)$ almost surely has no simultaneous jumps, the weak convergence of finite dimensional distributions also implies J-convergence and finally Theorem 9.1 is proved.

9.3.2. *Proof of Theorem 9.2*

Consider the embedded MP (X_{nk}, Z_{nk}), $k \geq 1$, introduced above. Denote $m_n(x,i) = \mathbf{E}[T_{n2} - T_{n1} \mid (X_{n1}, Z_{n1}) = (x,i)]$. Then

$$m_n(x,i) = \mathbf{E}\int_0^\infty u \exp\big\{ - A_n(u,x,i)\big\} b_n\big(x_n^{(i)}(u), i\big)\mathrm{d}u, \tag{9.11}$$

given that $x_n^{(i)}(0) = x$. According to the conditions of Theorem 9.2, the tail of the integral in the domain $\{u > L\}$ at $\varepsilon < c_i$ and large enough n can be approximated by the value $\int_L^\infty u \exp\{-(c_i - \varepsilon)u\} C_i \mathrm{d}u$, which is small at large L. According to Theorem 9.1, the integral in the domain $\{u \leq L\}$ converges to $\int_0^L u \exp\{-a_0(i)u\} a_0(i)\mathrm{d}u$. Thus, the right-hand side in (9.11) converges to $a_0(i)^{-1}$.

Let $x_n^{(i)}(0) = x$. As $A_n(t,x,i) \xrightarrow{\mathrm{P}} a_0(i)t$ and $A_n(t,x,i) > (c_i - \varepsilon)t$ at large n and $\varepsilon < c_i$, then, by adding and subtracting the term $e^{-a_0(i)t}\mathbf{P}(x_n^{(i)}(t) = y)$, we obtain that for any $x \in X_i$, $\delta > 0$ uniformly in $t \geq \delta$,

$$\mathbf{E}e^{-A_n(t,x,i)}\chi\big(x_n^{(i)}(t) = y\big) - \mathbf{E}e^{-A_n(t,x,i)}\mathbf{P}\big(x_n^{(i)}(t) = y\big) \longrightarrow 0. \tag{9.12}$$

Relation (9.9) implies that for any $i \in Z$, $y \in X_i$, uniformly in $t \geq \delta$, $\mathbf{P}(x_n^{(i)}(t) = y) - \rho_n^{(i)}(y) \to 0$. Denote $\widetilde{\tau}_n(t) = \max\{T_{nk} : T_{nk} < t\}$ ($\widetilde{\tau}_n(t)$ is the last time of jump of $\zeta_n(\cdot)$ before t). Thus

$$\begin{aligned}
&\mathbf{P}\big(\big(x_n(t), \zeta_n(t)\big) = (y,i)\big) \\
&\quad = \sum_{x \in X_i} \int_0^t \mathbf{P}\big(\widetilde{\tau}_n(t) \in \mathrm{d}u,\ x_n\big(\widetilde{\tau}_n(t) + 0\big) = x,\ \zeta_n\big(\widetilde{\tau}_n(t) + 0\big) = i\big) \\
&\qquad \times \mathbf{E}e^{-A_n(t-u,x,i)}\chi\big(x_n^{(i)}(t-u) = y\big).
\end{aligned} \tag{9.13}$$

Using relations (9.12), (9.13) we obtain after some algebra that as $n \to \infty$,

$$\mathbf{P}\big((x_n(t), \zeta_n(t)) = (y, i)\big) - \mathbf{P}\big(\zeta_n(t) = i\big)\rho_n^{(i)}(y) \longrightarrow 0 \tag{9.14}$$

uniformly in $t \geq \delta$. Consider the process $(x_n(t), \zeta_n(t)), t \geq 0$. According to condition (9.1) the states in X_i communicate (except possibly a number of transient states). Correspondingly, if $a_0(i,j) > 0$, then there exist states (x,i) and (y,j) such that $b_n((x,i),(y,j)) > 0$ at large enough n. This means, if the process $\zeta_0(\cdot)$ is ergodic, then at large n the states of $(x_n(\cdot), \zeta_n(\cdot))$ communicate (except possibly a number of transient states) and $(x_n(\cdot), \zeta_n(\cdot))$ is also ergodic (limiting probabilities exist but some of them can be zeros). Denote by $\{\pi_n(x,i),\ x \in X_i,\ i \in Z\}$ and $\{\Pi_n(i),\ i \in Z\}$ the stationary distributions of the embedded MP (X_{nk}, Z_{nk}) and component $\zeta_n(\cdot)$, respectively. From relation (9.14) it follows, as $t \to \infty$, that

$$\rho_n(y,i) - \Pi_n(i)\rho_n^{(i)}(y) \longrightarrow 0, \quad i \in Z,\ y \in X_i. \tag{9.15}$$

Let us find the limit of $\Pi_n(i)$ as $n \to \infty$. Denote

$$\begin{aligned} p_n\big((x,i),(y,j)\big) &= \mathbf{P}\big((X_{n2}, Z_{n2}) = (y,j) \mid (X_{n1}, Z_{n1}) = (x,i)\big), \\ \widehat{a}_n\big(i,(y,j)\big) &= \sum_{x \in X_i} \rho_n^{(i)}(x) b_n\big((x,i),(y,j)\big), \quad j \neq i. \end{aligned} \tag{9.16}$$

Without loss of generality we may assume that the limits $\lim_{n\to\infty} \widehat{a}_n(i,(y,j)) = a_0(i,(y,j))$ exist (otherwise we may assume the existence of partial limits and show that the final result does not depend on the subsequence $n_k \to \infty$). Then, using the representation similar to (9.6) and relation (9.10), we obtain that

$$\lim_{n\to\infty} p_n\big((x,i),(y,j)\big) = a_0\big(i,(y,j)\big)/a_0(i), \tag{9.17}$$

where $\sum_{y \in X_j} a_0(i,(y,j)) = a_0(i,j)$. Consider an MP (X_k, Z_k) with state space (x,i), $x \in X_i$, $i \in Z$, and transition probabilities $a_0(i,(y,j))/a_0(i)$ from state (x,i) to (y,j). It is easy to see that component Z_k forms an MP with transition probabilities $a_0(i,j)/a_0(i)$. This process is the embedded MP for $\zeta_0(\cdot)$ and is ergodic. Denote by $\widehat{\pi}_0(i)$, $i \in Z$, its stationary distribution. It is easy to verify that the values $\pi_0(x,i) = \sum_{j \neq i} \widehat{\pi}_0(j) a_0(i,(x,j))/a_0(i)$ are the stationary probabilities for (X_k, Z_k) and $\widehat{\pi}_0(i) = \sum_{x \in X_i} \pi_0(x,i)$. According to the ergodic theorem for semi-Markov renewal processes, for any $i \in Z$,

$$\Pi_n(i) = \sum_{x \in X} \pi_n(x,i) m_n(x,i) \left(\sum_{x \in X,\ j \in Z} \pi_n(x,j) m_n(x,j) \right)^{-1}, \tag{9.18}$$

and, as $n \to \infty$,

$$\Pi_n(i) \longrightarrow \widehat{\pi}_0(i) a_0(i)^{-1} \left(\sum_{j \in Z} \widehat{\pi}_0(j) a_0(j)^{-1} \right)^{-1}, \quad i \in Z. \tag{9.19}$$

As we can see, the right-hand side in (9.19) is equal to $\Pi_0(i)$. Finally, from relations (9.15), (9.19) we obtain that $\rho_n(x,i) \to \Pi_0(i)\rho_0^{(i)}(x)$ and Theorem 9.2 is proved.

9.3.3. *Proof of Theorem 9.3*

If $\mathbf{E}\nu_n(x_0,i_0)^2 < C$, then MP (X_{nk}, Z_{nk}), $k \geq 1$, is positive recurrent, the expectation of the return time to state (x_0,i_0) for $(x_n(\cdot), \zeta_n(\cdot))$ is finite. Then, as at our conditions the values $m_n(x,i)$ are uniformly bounded (see (9.18)), then component $\zeta_n(\cdot)$ is ergodic with stationary distribution $\Pi_n(i)$. According to Theorem 9.2, for each state (x,i) transition probabilities and the expectation of a sojourn time converge to $a_0(i,(y,j))/a_0(i)$ (see (9.17)) and $1/a_0(i)$, respectively. As $\nu_n(x_0,i_0)$ is uniformly integrable, then $\mathbf{E}\nu_n(x_0,i_0)$ converges to the expectation of the return time to state i_0 for process $\zeta_0(\cdot)$ with the initial state i_0, because we have the convergence of the expectation of a sum of occupation times on any finite sequence of states of (X_{nk}, Z_{nk}). Thus, $\Pi_n(i) \to \Pi_0(i)$, $i \in Z$. Together with (9.15) this proves the result of Theorem 9.3.

9.4. Queueing systems with fast Markov type switching

In this section as the applications of Theorems 9.1, 9.2, 9.3 we consider averaging in Markov queueing models with fast switching.

9.4.1. *System* $M_{M,Q}/M_{M,Q}/1/N$

9.4.1.1. *Averaging of states of the environment*

Consider a queueing system in a fast Markov environment. A system consists of one server and N waiting places. Consider a general case, when the calls arrive according to a state-dependent Poisson process with Markov switching and also at the times of jumps of a switching MP. If the system is full, an arriving call is lost. Let $x_0(t)$, $t \geq 0$, be an ergodic MP with values in $X = \{x_1, \ldots, x_r\}$. Denote by $\rho(x)$, $x \in X$, its stationary distribution. Let $b(x)$ be the exit rate from state x, $x \in X$. We define a Markov environment with fast switching as follows: $x_n(t) = x_0(V_n t)$, $t \geq 0$, where V_n is a scaling factor, $V_n \to \infty$.

Denote by $\varphi_0(\cdot)$ and $\varphi_n(\cdot)$ uniformly strong mixing coefficients for processes $x_0(\cdot)$ and $x_n(\cdot)$, respectively. According to the ergodicity of $x_0(\cdot)$, there exist $q < 1$ and $L > 0$ such that $\varphi_0(L) \leq q$. Then $\varphi_n(L/V_n) = \varphi_0(L) \leq q$, and condition (9.1) holds.

Let $\{\lambda(x,i), \mu(x,i), \alpha_A(x,i), \alpha_S(x,i),\ x \in X,\ i \geq 0\}$ be the non-negative functions. Denote by $Q_n(t)$, $t \geq 0$, the total number of calls in the system at time t. The system is switched by the process $(x_n(t), Q_n(t))$ as follows: if $(x_n(t), Q_n(t)) = (x,i)$, then the local arrival rate is $\lambda(x,i)$, and the local service rate is $\mu(x,i)$. Moreover, if at the time t_{nk} of the kth jump of $x_n(t)$, $(x_n(t_{nk}-0), Q_n(t_{nk}-0)) = (x,i)$, *Proof.* Let T_{nk}, $k \geq 1$, be the times of sequential jumps of $Q_n(t)$. Put $(X_{nk}, Z_{nk}) = (x_n(T_{nk}+0), Q_n(T_{nk}+0))$, $k \geq 1$. Then (X_{nk}, Z_{nk}) is the embedded MP. Let $p_n((x,i),(y,j))$ be defined by (9.16). Denote $\widehat{v}(i) = \widehat{\lambda}(i) + \widehat{\mu}(i)$. As functions $\lambda(\cdot)$, $\mu(\cdot)$ are uniformly bounded, using Theorem 9.1 we can prove that $p_n((x,i),(y,j)) \to p_0(i,(y,j))$ uniformly in $i \geq 1$, $x \in X$, where $p_0(i,(y,j)) = 0$, as $|i-j| > 1$, and $p_0(i,(y,i+1)) = \rho(y)\lambda(y,i)\widehat{v}(i)^{-1}$, $p_0(i,(y,i-1)) = \rho(y)\mu(y,i)\widehat{v}(i)^{-1}$, and also uniformly in $i \geq 1$, $x \in X$,

$$\begin{aligned} \mathbf{E}\big[T_{n,k+1} - T_{nk} \mid \big(X_{nk}, \widetilde{Z}_{nk}\big) = (x,i)\big] &\longrightarrow \widehat{v}(i)^{-1}, \\ \mathbf{E}\big[Z_{n,k+1} - Z_{nk} \mid \big(X_{nk}, Z_{nk}\big) = (x,i)\big] &\longrightarrow \big(\widehat{\lambda}(i) - \widehat{\mu}(i)\big)\widehat{v}(i)^{-1}. \end{aligned} \tag{9.21}$$

This means that for some n_0, (X_{nk}, Z_{nk}) is irreducible as $n > n_0$. Also for some $\varepsilon > 0$ the left-hand side in (9.21) at large enough n (i.e. $n > n_0$) is no greater than $-\varepsilon$ as $i > L$. Then, according to the classic Foster criterion, at $n > n_0$ the process $(\widetilde{X}_{nm}, \widetilde{Z}_{nm})$ is positive recurrent. Consider a finite domain $D = X \times \{0, \ldots, L\}$ and denote by $\nu_n(x, L+1, D)$ a return time to D for $(\widetilde{X}_{nm}, \widetilde{Z}_{nm})$ given that $(\widetilde{X}_{n0}, \widetilde{Z}_{n0}) = (x, L+1)$. In the same way as it was done in Theorem 4.1 [ANI 01], we can prove that for some α, $0 < \alpha < 1$, $\mathbf{P}(\nu_n(x, L+1, D) > k) \leq \alpha^k$, $k > 1$, as $n > n_0$. This implies uniformly in $n > n_0$ the existence of the 2nd moment for $\nu_n(x, L+1, D)$ and for the return time to the domain $(X, 0)$, respectively. Thus, our result follows from Theorem 9.3. □

EXAMPLE 9.3. Consider a system $M_M/M_M/1/\infty$ which is switched by the fast MP $x_n(t) = x_0(V_n t)$ with values in $X = \{x_1, \ldots, x_r\}$, where as $x_n(t) = x$, the arrival and service rates are $\lambda(x)$ and $\mu(x)$, respectively. Let $x_0(t)$ be ergodic with stationary distribution $\rho(x)$, and $\lambda(x) + \mu(x) > 0$, $x \in X$. Denote $\widehat{\lambda} = \sum_{x \in X} \lambda(x)\rho(x)$, $\widehat{\mu} = \sum_{x \in X} \mu(x)\rho(x)$. Suppose that $\widehat{\mu} > \widehat{\lambda}$. Put $g = \widehat{\lambda}/\widehat{\mu}$. Let $Q_n(0) = i_0$. Then Statement 9.1 holds, where $Q(\cdot)$ is a Birth-and-Death process with constant birth and death rates $\widehat{\lambda}$ and $\widehat{\mu}$, $Q(0) = i_0$, and $\rho_n(x,i) \to \rho(x)(1-g)g^i$, $x \in X$, $i \geq 0$.

Note that for this system the approximation of a stationary distribution can also be obtained using matrix-analytic relations [NEU 81, NEU 89].

9.4.2. *Batch system* $BM_{M,Q}/BM_{M,Q}/1/N$

Consider the case of batch arrival and service. In general, batch systems even in Markov case are very difficult for analytic study, because they do not belong to the class of quasi-Birth-and-Death processes [NEU 81]. We keep all notations from section 9.4.1.1 and suppose for simplicity that $\alpha_A(\cdot) \equiv 0$, $\alpha_S(\cdot) \equiv 0$. Let in addition the families of non-negative integer random variables $\{\xi(x,i), \eta(x,i),\ x \in X,\ i \geq 0\}$, be given. The system operates as follows: if $(x_n(t), Q_n(t)) = (x,i)$, then with rate $\lambda(x,i)$ a batch of $\min(\xi(x,i), N+1-i)$ calls may enter the system, or with rate

$\mu(x,i)$ a batch of $\min(\eta(x,i),i)$ calls may complete service and leave the system ($\mu(x,0) \equiv 0$). Put

$$a^+(x,i,m) = \lambda(x,i)\mathbf{P}\big(\xi(x,i)=m\big), \quad a^-(x,i,m) = \mu(x,i)\mathbf{P}\big(\eta(x,i)=m\big),$$

$$\widehat{a}^+(i,m) = \sum_{x\in X}\alpha^+(x,i,m)\rho(x), \quad \widehat{a}^-(i,m) = \sum_{x\in X}\alpha^-(x,i,m)\rho(x),$$

$$\widehat{a}^+(i) = \sum_{m=0}^{\infty}\widehat{a}^+(i,m), \quad \widehat{a}^-(i) = \sum_{m=0}^{\infty}\widehat{a}^-(i,m), \quad i \ge 0,\ m \ge 0,$$

and let $\widehat{\xi}(i)$, $\widehat{\eta}(i)$ be random variables such that $\mathbf{P}(\widehat{\xi}(i) = m) = \widehat{a}^+(i,m)/\widehat{a}^+(i)$, $\mathbf{P}(\widehat{\eta}(i) = m) = \widehat{a}^-(i,m)/\widehat{a}^-(i)$, $i \ge 0, m \ge 0$.

Consider a state-dependent, approximating queueing system $BMAP_Q/BM_Q/1/N$ with averaged characteristics operating as follows. Denote by $Q(t)$ the total number of calls in the system at time t. If $Q(t) = i$, then with rate $\widehat{a}^+(i)$ a batch of $\min(\widehat{\xi}(i), N+1-i)$ calls may enter the system, and with rate $\widehat{a}^-(i)$ a batch of $\min(\widehat{\eta}(i), i)$ calls may complete service and leave the system.

STATEMENT 9.4. *If process $Q(\cdot)$ is regular and $Q_n(0) = q_0$, then for any $N \le \infty$ the process $Q_n(\cdot)$ J-converges in any finite interval $[0,T]$ to $Q(\cdot)$ ($Q(0) = q_0$). If $N < \infty$, $\lambda(x,i) + \mu(x,i) > 0$, $x \in X$, $i = 0,\dots,N+1$, and $Q(\cdot)$ is ergodic with stationary distribution $\Pi(i)$, then $Q_n(\cdot)$ is also ergodic and*

$$\rho_n(x,i) \longrightarrow \rho(x)\Pi(i), \quad x \in X,\ i = 0,\dots,N+1.$$

The proof follows from Theorem 9.1. Note that these results can easily be extended to multiserver models.

9.4.3. *System $M/M/s/m$ with unreliable servers*

Consider a system $M/M/s/m$ with s identical servers. Assume that no more than m calls can be in the system at one time. Suppose that each idle server is subject to failure and after failure it is immediately taken for repair. Let failure and repair rates be considerably smaller compared to arrival and service rates. More specifically, suppose that the arrival flow is a Poisson flow with rate $V_n\lambda$, service time is exponential with rate $V_n\mu$ ($V_n \to \infty$). Each idle server may fail with rate κ, and each failed server has a repair rate ν. If an arriving call sees m calls in the system, this call is lost. After completion of service a call leaves the system.

This can be a realistic model for a hospital with a number of ambulances which may serve calling patients. In this case patients arrive considerably more often compared to the failures of ambulances.

Denote by $Q_n(t)$ a number of calls in the system at time t and by $R_n(t)$ a number of failed servers. In this case $Q_n(\cdot)$ is a fast process with transition rates depending on the current value of $R_n(\cdot)$. Assume that $s \le m$. For each $i = 0, \dots, s$, denote by $x^{(i)}(t)$ an auxiliary Birth-and-Death process with values in $\{0, 1, \dots, m\}$, birth and death rates in state k, λ and $\min(k, s-i)\mu$, respectively, and stationary distribution $\rho_k^{(i)}$, $k = 0, \dots, m$. Note that the values $\rho_k^{(i)}$ can be explicitly calculated (as $i = s$, $\rho_m^{(s)} = 1$, $\rho_k^{(s)} = 0$, $k < m$). Let $R(t)$ be the approximating Birth-and-Death process with values in $\{0, 1, \dots, s\}$ and birth and death rates in state i, $\widehat{\lambda}(i)$ and $\widehat{\mu}(i)$, respectively, where

$$\widehat{\mu}(i) = i\nu, \quad i = 1, \dots, s,$$

$$\widehat{\lambda}(i) = \kappa \sum_{k=0}^{s-i-1} (s-i-k)\rho_k^{(i)}, \quad i = 0, \dots, s-1.$$

Suppose that $s < \infty$, $m \le \infty$, and $R_n(0) = r_0$.

STATEMENT 9.5. *If $\lambda\mu\kappa\nu > 0$, then for any $m \le \infty$, $R_n(\cdot)$ J-converges in any finite interval $[0, T]$ to $R(\cdot)$, where $R(0) = r_0$. If $m < \infty$, then the processes $R(\cdot)$ and $(Q_n(\cdot), R_n(\cdot))$ are both ergodic and for any $k = 0, \dots, m$, $i = 0, \dots, s$, $\rho_n(k, i) \to \rho_k^{(i)}\Pi(i)$, where $\rho_n(k, i)$ and $\Pi(i)$ are stationary distributions of $(Q_n(\cdot), R_n(\cdot))$ and $R(\cdot)$, respectively,*

Proof. It is easy to calculate that the transition rates $b_n((k, i), (l, j))$ from state (k, i) to (l, j) for $i = 0, \dots, s$ have the following form:

$$b_n(\cdot, \cdot) = \begin{cases} V_n\lambda & \text{if } j = i,\ l = k+1,\ k = 0, \dots, m-1; \\ V_n \min(k, s-i)\mu & \text{if } j = i,\ l = k-1,\ k = 1, \dots, m; \\ \min(0, s-i-k)\kappa & \text{if } j = i+1,\ l = k,\ k = 0, \dots, m; \\ i\nu & \text{if } j = i-1,\ l = k,\ k = 0, \dots, m; \end{cases}$$

and $b_n(\cdot, \cdot) = 0$ otherwise. Now Statement 9.5 follows from Theorems 9.1 and 9.2. □

9.4.4. *Priority model* $M_Q/M_Q/m/s, N$

Consider the model with two types of priorities and different arrival and service rates. Assume that there are m identical servers, s waiting places (buffer) for the 1st priority calls (first type) and no more than N places for the second type calls. A flow of first type calls is fast, and a flow of second type calls is slow. We may interpret, for instance, the first type as the flow of internal tasks in a computer service system, and the second type as an external flow of users to the system.

Let the first type of flow be a Poisson flow with rate $V_n a$, $V_n \to \infty$, and service rates for first type be $V_n b$. Denote by $Q_n^{(j)}(t)$ the total number of calls of type j in the system at time t, where $j = 1, 2$. The system operates as follows: if a new first type call enters the system, it either takes any idle server, if any, or it goes to one of the servers, if any, which is occupied by the second type call and interrupts its service. Otherwise, it goes to the queue of first type calls, or leaves the system, if all m servers have first type calls on service and s calls of first type are waiting in the buffer. If the first type call completes the service and there are waiting first type calls, then one of these calls immediately goes for service. If the service of a second type call was interrupted, then it goes back to the queue. If a server becomes idle and there are no first type calls in the buffer, then one of second type calls from the queue (if any) immediately goes for service. The arrival flow of second type calls is state-dependent and constructed as follows: if at time t, $Q_n^{(2)}(t) = i$ and there are k idle servers, then the local arrival rate for second type calls is $\lambda(k, i)$, $i = 0, \dots, N-1$, ($\lambda(k, N) \equiv 0$, $k = 0, \dots, m$), so there cannot be more than N of the second type calls in the system. At this time the service rate for each second type call on service is $\mu(i)$. All calls after completion of service leave the system.

Let $x_0(t)$, $t \geq 0$, be a Birth-and-Death process with values in $\{0, \dots, m+s\}$ and birth and death rates $a_k = a$ and $b_k = \min\{k, m\}b$, respectively. Let g_k, $k = 0, \dots, m+s$, be its stationary distribution (it can be calculated in a closed form). Put $\rho_k = g_{m-k}$, $k = 1, \dots, m$, $\rho_0 = \sum_{l=m}^{m+s} \rho_l$,

$$\widehat{\lambda}(i) = \sum_{k=0}^{m} \lambda(k, i)\rho_k, \quad i = 0, \dots, N-1,$$

$$\widehat{\mu}(i) = \mu(i) \sum_{k=1}^{\min\{i,m\}} k\rho_k, \quad i = 1, \dots, N.$$

Consider an approximating Birth-and-Death process $Q(t)$ with values in $\{0, \dots, N\}$ and birth and death rates $\widehat{\lambda}(i)$ and $\widehat{\mu}(i)$, respectively.

STATEMENT 9.6. *If $a > 0, b > 0$, $Q(\cdot)$ is regular and $Q_n^{(2)}(0) = q_0$, then for any $N \leq \infty$, process $Q_n^{(2)}(\cdot)$ J-converges in any finite interval $[0, T]$ to $Q(\cdot)$ ($Q(0) = q_0$). If $N < \infty$, $\lambda(l, i) + \mu(i) > 0$, $l = 0, \dots, m$, $i = 0, \dots, N$, and $Q(\cdot)$ is ergodic with stationary distribution $\Pi(i)$, $i = 0, \dots, N$, then $Q_n^{(2)}(\cdot)$ is also ergodic and for any $k = 0, \dots, m+s$, $i = 0, \dots, N$,*

$$\lim_{n\to\infty} \lim_{t\to\infty} \mathbf{P}\big(Q_n^{(1)}(t) = k,\ Q_n^{(2)}(t) = i\big) = g_k \Pi(i).$$

Proof. Consider MP $(Q_n^{(1)}(t), Q_n^{(2)}(t))$, $t \geq 0$. As first type calls have the absolute priority, then $Q_n^{(1)}(\cdot)$ is a Birth-and-Death process with fast switching and it forms a fast Markov environment for $Q_n^{(2)}(\cdot)$. Transition rates $b_n((k,i),(l,j))$ have the following form:

$$b_n(\cdot,\cdot) = \begin{cases} V_n a, & \text{if } j = i,\ l = k+1;\ i = 0,\dots,N, \\ & \quad k = 0,\dots,m+s-1; \\ V_n b_k & \text{if } j = i,\ l = k-1;\ i = 0,\dots,N, \\ & \quad k = 1,\dots,m+s; \\ \lambda(\min(0,m-k),i) & \text{if } j = i+1,\ l = k;\ i = 0,\dots,N-1, \\ & \quad k = 0,\dots,m+s; \\ \min(m-k,i)\mu(i) & \text{if } j = i-1,\ l = k;\ i = 1,\dots,N, \\ & \quad k = 0,\dots,m+s, \end{cases}$$

and $b_n(\cdot,\cdot) = 0$ otherwise. Then Statement 9.6 follows from Theorems 9.1 and 9.2. □

To this end, note that the approach suggested provides us with the opportunity to approximate various functionals of queueing processes. In particular, as a hitting time to a domain for a process with discrete state space is a continuous functional concerning J-convergence, we can easily prove the following result related to the models considered above. Denote by $H_n(i)$ a busy period for $Q_n(\cdot)$ given that $Q_n(0) = i$. Let $H(i)$ be a busy period for the approximated system $Q(\cdot)$. Suppose that $N < \infty$. If $Q_n(\cdot)$ J-converges to $Q(\cdot)$ in any interval $[0,T]$ and $H(i)$ is a proper random variable (for instance, all states of $Q(\cdot)$ communicate), then for any i, $H_n(i)$ converges in distribution to $H(i)$ and also $\mathbf{E}H_n(i) \to \mathbf{E}H(i)$.

A separate and interesting problem is to find an error of approximation. For general models it is still an open question. If $x_n(t) = x_0(V_n t)$, where $x_0(t)$ is a finite ergodic MP, and a Markov queueing system is also finite, it is possible to prove using the results of [ANI 88] that for any $t > 0$, $\mathbf{P}(Q_n(t) = j) - \mathbf{P}(Q(t) = j) = O(1/V_n)$, and the error in the approximation of the stationary distribution is also $O(1/V_n)$.

9.5. Non-homogenous in time queueing models

To show wide possibilities of applications of these results let us consider the asymptotic behavior of non-homogenous in time Markov queueing systems with fast Markov type switching for the cases when a switching MP is either ergodic or allows the asymptotic aggregation of the state space. These results are of a different type from those considered in Chapter 4, as we do not normalize the queueing process $Q(t)$ and consider it in a usual scale of time, but we assume that switching is very fast and in each finite interval $[a,b]$ the number of switches tends to infinity as $n \to \infty$.

9.5.1. *System* $M_{M,Q,t}/M_{M,Q,t}/s/m$ *with fast switching – averaging of states*

Let the families of continuous in argument t, non-negative functions $\{a(i,l,t,q), \lambda(i,t,q), \mu(i,t,q),\ i,l \in X,\ i \neq l,\ q \in \{0,1,2,\ldots\}\}$ be given, where $X = \{1, 2,\ldots,r\}$ is a finite set. The system consists of s identical servers and m waiting places. Denote by $x_n(t)$ a fast non-homogenous in time Markov type environment for the system with transitions also depending on the value of the queue. Let us describe the evolution of the system and the environment.

Calls enter the system one at a time. If at the time t the total number of calls in the system is Q, and $x_n(t) = i$, then the instantaneous input rate is $\lambda(i,t,Q)$, the instantaneous service rate for any busy server is $\mu(i,t,Q)$ and process $x_n(t)$ may jump from state i to state l with rate $na(i,l,t,Q)$. After completion of service the call leaves the system. This means that process $x_n(t)$ has asymptotically fast transitions and we also assume that the states of $x_n(t)$ communicate at each t and Q in the following sense. Consider for each fixed (v,q) an auxiliary homogenous MP $x(t,v,q)$, $t \geq 0$, with values in X given by the transition rates $\{a(i,l,v,q),\ i,l \in X,\ i \neq l\}$. Let $\varphi(u,v,q)$ be its uniformly strong mixing coefficient in the interval $[0,u]$:

$$\begin{aligned}\varphi(u,v,q) = \max_{i_1,i_2 \in X,\ A \subset X} \Big| \mathbf{P}\big\{x(u,v,q) \in A \mid x(0,v,q) = i_1\big\} \\ - \mathbf{P}\big\{x(u,v,q) \in A \mid x(0,v,q) = i_2\big\}\Big|.\end{aligned}$$

Suppose that there exists d, $0 \leq d < 1$, and for any $T > 0$ there exists a value $r(T) > 0$ such that for any $v \leq T$ and any $q \geq 0$,

$$\varphi(r(T),v,q) \leq d. \tag{9.22}$$

This condition implies that given Q, process $x_n(t)$ is a quasi-ergodic process (see section 3.3). Denote by $\{\pi(i,v,q),\ i \in X\}$ the stationary distribution of the process $x(t,v,q)$, $t \geq 0$, and put

$$\widehat{\lambda}(t,q) = \sum_{i \in X} \lambda(i,t,q)\pi(i,t,q), \quad \widehat{\mu}(t,q) = \sum_{i \in X} \mu(i,t,q)\pi(i,t,q). \tag{9.23}$$

Also denote by $Q_n(t)$ the total number of calls in the system at the time t.

Let $M_{Q,t}/M_{Q,t}/s/m$ be the approximating state-dependent queueing system with the local input and service rates at time t, $\widehat{\lambda}(t,Q)$ and $\widehat{\mu}(t,Q)$, respectively, given that $Q(t) = Q$. Suppose that process $Q(t)$ is regular.

STATEMENT 9.7. *If* $Q_n(0) \stackrel{\mathrm{w}}{\Longrightarrow} Q_0$*, then with the assumption above, process* $Q_n(t)$ *J-converges in each finite interval to the process* $Q(t)$ *with* $Q(0) = Q_0$.

This means that the queueing process in the initial system is approximated by the queueing process in the system with input and service rates that are averaged by a quasi-stationary distribution.

Proof. We consider a two-component MP $(x_n(t), Q_n(t))$ and describe it as an SP. In this case the component $Q_n(\cdot)$ plays the role of a slow environment, $x_n(\cdot)$ is a process of Markov type with fast switching and it is quasi-ergodic at any fixed value of $Q_n(\cdot)$. Therefore, the proof follows directly from Theorem 9.1 under the natural modification accounting for non-homogenity in time. □

9.5.2. *System* $M_{M,Q}/M_{M,Q}/s/m$ *with fast switching – aggregation of states*

Now we consider the previous system in the case when process $x_n(\cdot)$ allows an asymptotic aggregation of state space. Consider for transparency a homogenous case and assume that the transition rates of the environment do not depend on the values of the queue.

Let for each $n > 0$, $x_n(t)$, $t \geq 0$, be a fast homogenous MP with state space $X = \{1, 2, \ldots, r\}$, given by the family of transition rates $\{na_n(i,l),\ i,l \in X,\ i \neq l\}$. Also let the family of non-negative values $\{\lambda(i,q), \mu(i,q),\ i \in X,\ q \geq 0\}$ be given. Process $x_n(\cdot)$ forms a switching environment: if $x_n(t) = i$ and $Q_n(t) = Q$, then the instantaneous input rate is $\lambda(i,Q)$ and the instantaneous service rate for each busy server is $\mu(i,Q)$. We assume that $x_n(t)$ satisfies the conditions of asymptotic aggregation of state space (8.18) and (8.19) in section 8.2.3 with $\varepsilon = 1/n$. Assume that for any $y \in Y$ an MP $x^{(y)}(t)$, defined by transition rates $a_0(i,j)$ in the region X_y is irreducible. Denote by $\pi^{(y)}(i)$, $i \in X_y$, its stationary distribution given by (8.20) and put

$$\widehat{a}_{jm} = \sum_{i \in X_j} \pi^{(j)}(i) \sum_{l \in X_m} b_{il}, \quad j \neq m.$$

Let $y(t, y_0)$ be an MP with values in Y, transition rates $\widehat{a}_{jm}$ and initial state y_0. Denote by $K(x_n(t))$ the aggregated process where $K(\cdot)$ is a map from X to Y: $K(i) = y$ if $i \in X_y$. Put

$$\widehat{\lambda}(j,q) = \sum_{i \in X_j} \lambda(i,q)\pi^{(j)}(i), \quad \widehat{\mu}(j,q) = \sum_{i \in X_j} \mu(i,q)\pi^{(j)}(i).$$

Denote by $Q_n(t)$ the total number of calls in the system at time t.

Let $M_M/M_M/s/m$ be the approximating queueing system switched by process $y(\cdot)$: if $y(t) = j$ and $Q(t) = Q$, then the input rate is $\widehat{\lambda}(j,Q)$ and the service rate for each busy server is $\widehat{\mu}(j,Q)$. Suppose that process $Q(t)$ is regular.

STATEMENT 9.8. *Assume that* $x_n(0) \in X_{y_0}$ *and* $Q_n(0) \stackrel{\text{w}}{\Longrightarrow} Q_0$. *Then under our assumptions process* $(K(x_n(t)), Q_n(t))$ *J-converges in each finite interval to process* $(y(t, y_0), Q(t))$ *with* $Q(0) = Q_0$.

The proof uses the representation of the process $(K(x_n(t)), Q_n(t))$ as an SP where the switching times are the times of transitions between regions X_j. The approximation of the queueing process in each region by the queueing process with averaged characteristics follows directly from Theorem 9.1 (see also Statement 9.7). Convergence of the aggregated process $K(x_n(t))$ to an MP $y(t, y_0)$ follows from Theorem 8.8, and finally the convergence of a two-component process $(K(x_n(t)), Q_n(t))$ to $(y(t, y_0), Q(t))$ follows from Theorem 8.3.

Note that in this case the limiting queueing system is operating in the Markov environment with aggregated space state and averaging of local characteristics is made in each asymptotically connected region.

9.6. Numerical examples

Consider a system $M/M/s/m$ with unreliable servers investigated in section 9.4.3. Put $m = s = 2$. This system has 9 states (k, i), $k, i = 0, 1, 2$. In the two tables below the reader can see the results of exact and approximate calculations of stationary probabilities and macro-state probabilities (for the component $Q(\cdot)$) for two different cases:

1) the ratio $\lambda/\kappa = 200$ ($V_n = 200$);

2) the ratio $\lambda/\kappa = 50$ ($V_n = 50$).

Column “E” shows the exact values of stationary probabilities; “A” - shows approximate results using the method suggested above in this section; “C” - shows approximate results using the method suggested in [COU 77]; ε(E,A) and ε(C,A) show the differences between column “E” and columns “A” and “C”, respectively.

The tables show that in the first case ($V_n = 200$) the error of approximation is much less, and also column “A” provides in general better results comparatively to column “C”. This illustrates the effectiveness of the proposed approach compared to the well-known Courtois method.

State	E	A	C	ε(E,A)	ε(E,C)
(0,0)	0.1791	0.1795	0.1801	-0.0004	-0.001
(2,0)	0.0627	0.0623	0.0615	0.0004	0.0012
(0)	0.3917	0.3913	0.3893	0.0004	0.0024
(1,1)	0.1675	0.1677	0.1662	-0.0002	0.0013
(2,1)	0.1398	0.1397	0.1362	0.0001	0.0035
(1)	0.5080	0.5086	0.5079	-0.0006	-0.0001
(2,2)	0.0984	0.1006	0.095	-0.0022	0.0034
(2)	0.1004	0.1006	0.1028	-0.0002	-0.0024

Table 9.1. $\lambda = 200$, $\mu = 240$, $\kappa = \nu = 1$

State	E	A	C	ε(E,A)	ε(E,C)
(0,0)	0.1780	0.1795	0.1801	-0.0015	-0.0001
(2,0)	0.0641	0.0623	0.0615	0.0018	0.0026
(0)	0.3931	0.3913	0.3893	0.0018	0.0038
(1,1)	0.1671	0.1677	0.1662	-0.0006	0.0009
(2,1)	0.1400	0.1397	0.1362	0.0003	0.0038
(1)	0.5070	0.5086	0.5079	-0.0016	-0.0009
(2,2)	0.0924	0.1006	0.095	-0.0082	0.0026
(2)	0.0999	0.1006	0.1028	-0.0007	-0.0029

Table 9.2. $\lambda = 50$, $\mu = 60$, $\kappa = \nu = 1$

9.7. Bibliography

[ANI 78] ANISIMOV V., "Limit theorems for switching processes and their applications", *Cybernetics*, vol. 14, no. 6, p. 917–929, 1978.

[ANI 88] ANISIMOV V., "Estimates for deviations of transient characteristics of non-homogenous Markov processes", *Ukrainian Math. J.*, vol. 40, no. 6, p. 588–592, 1988.

[ANI 95] ANISIMOV V., "Switching processes: averaging principle, diffusion approximation and applications", *Acta Applicandae Mathematicae*, vol. 40, p. 95–141, 1995.

[ANI 98] ANISIMOV V., "Asymptotic analysis of stochastic models of hierarchic structure and applications in queueing models", in CHAKRAVARTHY S. and ALFA A., Eds., *Advances in Matrix Analytic Methods for Stochastic Models*, p. 237–259, Notable Publications, New Jersey, 1998.

[ANI 00] ANISIMOV V., "Asymptotic analysis of reliability for switching systems in light and heavy traffic conditions", in LIMNIOS N. and NIKULIN M., Eds., *Recent Advances in Reliability Theory: Methodology, Practice, and Inference*, p. 119–133, Birkhäuser Boston, Massachusetts, 2000.

[ANI 01] ANISIMOV V. and ARTALEJO J., "Analysis of Markov multiserver retrial queues with negative arrivals", *Queueing Systems*, vol. 39, no. 2/3, p. 157–182, 2001.

[ANI 02a] ANISIMOV V., "Averaging in Markov models with fast Markov switches and applications to queueing models", *Annals of Operations Research*, vol. 112, no. 1, p. 63–82, 2002.

[ANI 02b] ANISIMOV V., "Diffusion approximation in overloaded switching queueing models", *Queueing Systems*, vol. 40, no. 2, p. 141–180, 2002.

[BOB 86] BOBBIO A. and TRIVEDI K., "An Aggregation Technique for the Transient Analysis of Stiff Markov Chains", *IEEE Transactions on Computers*, vol. 35, no. 9, p. 803–814, 1986.

[BOL 98] BOLCH G., GREINER S., DE MEER H. and TRIVEDI K., *Queueing Networks and Markov Chains*, John Wiley & Sons, New York, 1998.

[COU 77] COURTOIS P., *Decomposability: Queueing and Computer Systems Applications*, Academic Press, New York, 1977.

[DOO 53] DOOB J. L., *Stochastic Processes*, Wiley, New York, 1953.

[ETH 86] ETHIER S. and KURTZ T., *Markov Processes, Characterization and Convergence*, John Wiley & Sons, New York, 1986.

[KUS 90] KUSHNER H., *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser Boston, Massachusetts, 1990.

[LUC 94a] LUCANTONI D., CHOUDHURY G. and WHITT W., "The transient BMAP$/G/1$ queue", *Comm. Statist. Stochastic Models*, vol. 10, no. 1, p. 145–182, 1994.

[LUC 94b] LUCANTONI D. and NEUTS M., "Some steady-state distributions for the $MAP/SM/1$ queue", *Comm. Statist. Stochastic Models*, vol. 10, no. 3, p. 575–598, 1994.

[NEU 81] NEUTS M., *Matrix-Geometric Solutions in Stochastic Models*, John Hopkins University Press, Baltimore, 1981.

[NEU 89] NEUTS M., *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York & Basel, 1989.

[PIN 75] PINSKY M., "Random evolutions", in *Probabilistic Methods in Differential Equations*, vol. 451 of *Lecture Notes in Math.*, p. 89–99, Springer, Berlin, 1975.

[RAM 98] RAMASWAMI V., "The generality of Quasi Birth-and-Death processes", in CHAKRAVARTHY S. and ALFA A., Eds., *Advances in Matrix Analytic Methods for Stochastic Models*, p. 93–113, Notable Publications, New Jersey, 1998.

[SKO 89] SKOROKHOD A., *Asymptotic Methods in the Theory of Stochastic Differential Equations*, Amer. Math. Soc., Rhode Island, 1989.

[WIL 96] WILLIAMS R., "On the approximation of queueing networks in heavy traffic", in KELLY F., ZACHARY S. and ZIEDINS I., Eds., *Stochastic Networks. Theory and Applications*, p. 35–56, Oxford University Press, Oxford, 1996.

Chapter 10

Aggregation in Markov Models with Fast Semi-Markov Switching

This chapter extends in some sense the results of Chapter 9 (see also [ANI 02]) to semi-Markov models. We study in a triangular scheme the convergence in a finite interval (transient behavior) of a piece-wise MP $\zeta_n(\cdot)$ with local rates depending on a fast SMP $x_n(\cdot)$ (fast semi-Markov switching). First, we suppose that $x_n(\cdot)$ satisfies the asymptotic mixing property. Then $\zeta_n(\cdot)$ in each finite interval weakly converges in Skorokhod space (J-converges) to an MP with transition rates averaged by the stationary measure of $x_n(\cdot)$. The convergence of a stationary distribution for $(x_n(\cdot), \zeta_n(\cdot))$ is studied as well. Note that these results differ from the results of Chapters 4 and 5 as we do not normalize processes $\zeta_n(\cdot)$ and $Q_n(\cdot)$ and understand averaging principle in the sense of the convergence to the limiting process with average transition characteristics. Related in some sense results on averaging of dynamic systems with fast Markov switching using a martingale technique are considered in the books [ETH 86, KUS 90, SKO 89]. Some results on averaging of dynamic systems with fast semi-Markov switches using asymptotic methods for SP are obtained in [ANI 92, ANI 95].

The next step in the investigation is when SMP $x_n(\cdot)$ satisfies the conditions of the asymptotic aggregation (see section 8.4), so that the state space can be divided in the disjoint domains with small transition probabilities between them. Consider the aggregated process $K(x_n(\cdot))$ defined on the domains. Then the conditions of the convergence in Skorokhod space D_T of the pair $(K(x_n(\cdot)), \zeta_n(\cdot))$ to a two-component MP $(y_0(\cdot), \zeta_0(\cdot))$ with averaged transition rates are studied, where $y_0(\cdot)$ is the limiting aggregated MP, and $\zeta_0(\cdot)$ is an MP switched by $y_0(\cdot)$. The method of investigation in this chapter uses the results on the convergence of SPs developed in sections 8.3

and 8.4.1 (see also [ANI 77, ANI 78]). Note that the aggregation models for queueing systems in light and heavy traffic conditions are studied also in [ANI 98, ANI 00].

The results obtained show that under quite general assumptions an MP in a fast semi-Markov environment can be approximated by an MP or an MP in a Markov environment with the aggregated state space and the averaged transition rates. This approach gives the possibility to reduce the dimension and the complexity of the initial model. These results are illustrated on the example of a state dependent queueing model $M_{SM,Q}/M_{SM,Q}/1/N$ with fast semi-Markov switching.

10.1. Markov processes with fast semi-Markov switches

In this section we study two types of limit theorems for Markov processes with semi-Markov switching (MPSMS). Let characteristics of the process depend on a scaling factor n, $n \to \infty$. Suppose that $x_n(\cdot)$ is an SMP with stands for the switching environment. Assume that at given $x_n(t) = x$ the process $\zeta_{nk}(\cdot, x, z)$ is a stepwise MP and the sojourn times in states of an SMP $x_n(\cdot)$ are asymptotically small (fast switching). First, we study the case when the embedded MP for $x_n(\cdot)$ satisfies an asymptotically strong mixing condition. In this case all states of $x_n(\cdot)$ can be asymptotically averaged and the component $\zeta_n(\cdot)$ converges on each finite interval $[0, T]$ to an MP with averaged rates. An approximation of a stationary distribution is studied as well. In the second case the embedded MP satisfies the conditions of the asymptotic aggregation of state space (state space can be divided in disjoint domains with small transition probabilities between them). Then $\zeta_n(\cdot)$ converges to an MP switched by a new MP, which is a weak limit of the aggregated process constructed on the domains (limiting aggregated process).

10.1.1. *Averaging of a semi-Markov environment*

Let the families of non-negative functions $\{a_n(x,i,j),\ i,j \in Z,\ i \neq j,\ x \in X\}$, an SMP $x_n(t)$, $t \geq 0$, with values in X, the family of non-negative functions $\{c_n(x,i,j),\ i,j \in Z,\ i \neq j,\ x \in X\}$, and the initial value (x_{n0}, i_{n0}) be given. Assume that Z is a discrete set, $Z = \{0, 1, \ldots\}$. Let $(x_n(\cdot, x_{n0}), \zeta_n(\cdot, i_{n0}))$, $t \geq 0$, be an MPSMS constructed by the introduced families in the following way: $x_n(0, x_{n0}) = x_{n0}$, $\zeta_n(0, i_{n0}) = i_{n0}$. In the interval where $x_n(t) = x$, process $\zeta_n(\cdot, i_{n0})$ operates as an MP with transition rates $a_n(x,i,j)$. Furthermore, denote by t_{nk}, $k \geq 0$, the times of sequential jumps of $x_n(\cdot)$. If at time t_{nk}, $x_n(t_{nk} - 0) = x$, $\zeta_n(t_{nk} - 0, i_{n0}) = i$, then process $\zeta_n(\cdot, i_{n0})$ can make a transition to state j with probability $c_n(x,i,j)/V_n$, $j \in Z$, $j \neq i$, or remain in state i with probability $1 - c_n(x,i)/V_n$, where $c_n(x,i) = \sum_{j \neq i} c_n(x,i,j)$, and $V_n \to \infty$ is a normalizing factor.

Introduce for convenience a family of random variables $\{\gamma_{nk}(x,i),\ x \in X,\ i \in Z\}$, such that $\mathbf{P}(\gamma_{nk}(x,i) = j - i) = c_n(x,i,j)/V_n,\ j \in Z, j \neq i$, and $\mathbf{P}(\gamma_{nk}(x,i) = 0) = 1 - c_n(x,i)/V_n$. By definition, we can say that if at time t_{nk}, $x_n(t_{nk} - 0) = x$, $\zeta_n(t_{nk} - 0, i_{n0}) = i$, then process $\zeta_n(\cdot, i_{n0})$ may have a jump of size $\gamma_{nk}(x,i)$.

Put $a_n(x,i) = \sum_{j\in Z,\, j\neq i} a_n(x,i,j)$. Let x_{nk}, $k \geq 0$, be the embedded MP, and $\theta_n(x)$ be the sojourn time in state x for SMP $x_n(\cdot)$. Suppose that the variables $\theta_n(x)$ are of the order $O(1/V_n)$. Denote $\widetilde{\theta}_n(x) = V_n\theta_n(x)$, $x \in X$. Suppose that for any $n > 0$, $i \in Z$,

$$\sup_{x\in X} a_n(x,i) \leq C_i^{(1)} < \infty, \qquad\qquad \sup_{x\in X} \sum_{l\neq i} a_n(x,i,l)a_n(x,l) \leq C_i^{(2)} < \infty, \tag{10.1}$$

$$\sup_{x\in X} c_n(x,i) \leq C_i^{(3)} < \infty. \tag{10.2}$$

First, we study AP for the switched component $\zeta_n(\cdot)$. Introduce as in previous chapters a uniformly strong mixing coefficient for an MP x_{nk}:

$$\varphi_n(k) = \sup_{x,y\in X,\ A\in\mathcal{B}_X} \big|\mathbf{P}\big\{x_{nk}\in A \mid x_{n0}=x\big\} - \mathbf{P}\big\{x_{nk}\in A \mid x_{n0}=y\big\}\big|, \quad k>0. \tag{10.3}$$

Let there exist a sequence r_n such that for some q, $0 \leq q < 1$,

$$\varphi_n\big(r_n\big) \leq q,\ n>0, \quad \text{and} \quad r_n/V_n \to 0. \tag{10.4}$$

Given condition (10.4), MP x_{nk} at each $n > 0$ has a stationary measure $\pi_n(C)$, $C \in \mathcal{B}_X$. Denote

$$m_n(x) = \mathbf{E}\widetilde{\theta}_n(x), \quad \widehat{m}_n = \int_X m_n(x)\pi_n(\mathrm{d}x), \tag{10.5}$$

$$\widehat{a}_n(i,j) = \int_X a_n(x,i,j)m_n(x)\pi_n(\mathrm{d}x), \quad \widehat{a}_n(i) = \sum_{j\in Z, j\neq i} \widehat{a}_n(i,j), \qquad \widehat{c}_n(i,j) = \int_X c_n(x,i,j)\pi_n(\mathrm{d}x), \quad \widehat{c}_n(i) = \sum_{j\in Z,\ j\neq i} \widehat{c}_n(i,j). \tag{10.6}$$

Suppose that as $n \to \infty$ the following condition holds:

A) there exist values $a_0(i,j)$, $a_0(i)$, $c_0(i,j)$, $c_0(i)$, $i,j \in Z$, $j \neq i$, and a value $m > 0$, such that $\widehat{m}_n \to m$, and for any $i,j \in Z$, $i \neq j$,

$$\widehat{a}_n(i,j) \longrightarrow a_0(i,j), \quad \widehat{a}_n(i) \longrightarrow a_0(i), \quad \widehat{c}_n(i,j) \longrightarrow c_0(i,j), \quad \widehat{c}_n(i) \longrightarrow c_0(i),$$

$$a_0(i) = \sum_{j\neq i} a_0(i,j), \quad c_0(i) = \sum_{j\neq i} c_0(i,j).$$

Denote $\lambda(i) = m^{-1}(a_0(i) + c_0(i))$, and let $\xi(i)$ be a random variable such that

$$\mathbf{P}\big(\xi(i) = j\big) = \big(a_0(i) + c_0(i)\big)^{-1}\big(a_0(i,j) + c_0(i,j)\big), \quad i, j \in Z,\ j \neq i.$$

Let $\zeta_0(t, i_0)$, $t \geq 0$, be an MP in Z with the exit rate $\lambda(i)$ in state $\{i\}$, the size of a jump $\xi(i)$, and with the initial state i_0. This means, $\zeta_0(\cdot, i_0)$ has transition rates $\lambda(i,j) = m^{-1}(a_0(i,j) + c_0(i,j))$, $j \neq i$.

THEOREM 10.1. *Let $\zeta_0(\cdot, i_0)$ be regular, conditions A), (10.1), (10.4) hold, $i_{n0} \stackrel{\mathrm{w}}{\Longrightarrow} i_0$, and*

$$\lim_{L\to\infty} \limsup_{n\to\infty} \sup_x \mathbf{E}\widetilde{\theta}_n(x)\chi\big(\widetilde{\theta}_n(x) > L\big) = 0. \tag{10.7}$$

Then for any $x_{n0} \in X$, $T > 0$, process $\zeta_n(\cdot, i_{n0})$ J-converges in $[0, T]$ to an MP $\zeta_0(\cdot, i_0)$.

Proof. First let us represent process $(x_n(\cdot, x_{n0}), \zeta_n(\cdot, i_{n0}))$ as an SP. To do this we need to choose a sequence of switching times, a switching component and the elementary processes $\zeta_{nk}(\cdot, x, i_{n0})$. For convenience we separate the jumps of $\zeta_n(\cdot)$ into two different processes: the jumps of $\zeta_n(\cdot)$ in the intervals between the times t_{nk} of jumps of $x_n(\cdot)$, and the jumps at times t_{nk}. Take $k = 1$, denote for simplicity $(x_{n0}, i_{n0}) = (x_0, i_0)$ and suppose that (x_0, i_0) is non-random. Let $\xi_n(t, x_0, i_0)$ be an MPSMS with the initial state i_0 constructed on the trajectory of $x_n(\cdot, x_0)$ by rates $a_n(x, i, j)$ without additional jumps at times t_{nk}. Denote by $\Omega_n(x_0, i_0)$ the time of the first jump of $\xi_n(\cdot, x_0, i_0)$ and put $\widetilde{\Omega}_n(x_0, i_0) = \min\{t_{nl} : t_{nl} > \Omega_n(x_0, i_0)\}$. Thus, this is the time of the next jump of SMP $x_n(\cdot)$ after time $\Omega_n(x_0, i_0)$. Let $\{\Sigma_n(t, x_0, i_0),\ t \geq 0\}$ be another process constructed on $x_n(\cdot, x_0)$ as follows. Put $\nu_n = \min\{t_{nl} : t_{nl} > 0,\ \gamma_{nl}(x_{nl}, i_0) \neq 0\}$. Then $\Sigma_n(t, x_0, i_0) = i_0$, $0 \leq t < t_{n,\nu_n}$, and if at time t_{n,ν_n}, $x_{n,\nu_n} = y$, then $\Sigma_n(\cdot, x_0, i_0)$ can make a jump of the size $j - i_0$ with probability $c_n(y, i_0, j)/c_n(y, i_0)$. After the jump it operates in the same manner until the next time when the variable $\gamma_{nk}(x_{nk}, \cdot) \neq 0$. Denote $\tau_{n1}(x_0, i_0) = \min\{\widetilde{\Omega}_n(x_0, i_0), t_{n,\nu_n}\}$. It is clear that $\tau_{n1}(x_0, i_0)$ is a Markov time for $(x_n(\cdot, x_0), \zeta_n(\cdot, i_0))$. Now we choose the interval $[0, \tau_{n1}(x_0, i_0))$ as the first switching interval for $(x_n(\cdot, x_0), \zeta_n(\cdot, i_0))$. The elementary process $\zeta_{n1}(t, i_0)$ in $[0, \tau_{n1}(x_0, i_0))$ is constructed as follows: $\zeta_{n1}(t, i_0) = \xi_n(t, x_0, i_0)$, $0 \leq t \leq \tau_{n1}(x_0, i_0)$, given that $\widetilde{\Omega}_n(x_0, i_0) < t_{n,\nu_n}$, and $\zeta_{n1}(t, i_0) = \Sigma_n(t, x_0, i_0)$, $0 \leq t \leq \tau_{n1}(x_0, i_0)$, given that $\widetilde{\Omega}_n(x_0, i_0) \geq t_{n,\nu_n}$.

The value $(x_1, i_1) = (x_n(\tau_{n1}(x_0, i_0), x_0), \zeta_n(\tau_{n1}(x_0, i_0) + 0, i_0))$ is a starting point for the process $(x_n(\cdot, x_0), \zeta_n(\cdot, i_0))$ in the next switching interval, and so on. Then we can represent $(x_n(t, x_0), \zeta_n(t, i_0))$ as an SP $(\kappa_n(t), (x_n(t, x_0), \zeta_n(t, i_0)),$ where the component $\kappa_n(t) = (x_0, i_0)$, $0 \leq t < \tau_{n1}(x_0, i_0))$, $\kappa_n(t) = (x_1, i_1)$ on the next switching interval, and so on, and the variable $\beta(\cdot)$ in the definition of an SP in (1.3), section 1.2.1 has the form: $\beta_{n1}(x_0, i_0) = (x_n(\tau_{n1}(x_0, i_0), x_0), \zeta_n(\tau_{n1}(x_0, i_0)+0, i_0))$ (the starting value in the next switching interval).

As asymptotically the states of $x_n(\cdot)$ are averaging, we consider a map K: $K(x, i) = i, x \in X$, and use Theorem 8.3, section 8.3. According to this theorem we need to prove the convergence of finite-dimensional distributions of $\zeta_{n1}(\cdot, i_{n0})$ jointly with the time of the first jump $\tau_{n1}(x_0, i_0)$ and the value $K(\beta_{n1}(x_0, i_0))$. Note that the interval $(\Omega_n(x_0, i_0), \widetilde{\Omega}_n(x_0, i_0))$ is asymptotically small, so with a probability close to one there are no additional jumps of $\xi_n(t, x_0, i_0)$ in this interval. This means that the starting point in the next switching interval is asymptotically determined either by the first jump of $\xi_n(\cdot, x_0, i_0)$ (according to transitions governed by the rates $a_n(x_0, i_0, j)$), or by the first jump of $\Sigma_n(\cdot, x_0, i_0)$.

As for any $t < \tau_{n1}(x_0, i_0)$, $\zeta_{n1}(t, x_0, i_0) \approx i_0$, then actually we need to prove only the convergence of the joint distribution of $\tau_{n1}(x_0, i_0)$ and $K(\beta_{n1}(x_0, i_0))$. This is the advantage of the use of Theorem 8.3. It is also possible to use the general results of Ethier and Kurtz [ETH 86, pp. 225–238], on the convergence to MPs. However, in our case the initial process is non-Markov, and in this way we will meet the same difficulties as when proving the convergence of transition probabilities of $\zeta_n(t, i_0)$ in any interval $[s, u]$. Therefore, it is easier to check the conditions of Theorem 8.3.

Given that $x_n(0) = x_0$, denote $A_n(t, x_0, i_0) = \int_0^t a_n(x_n(u), i_0)\mathrm{d}u$. Let $P_n(x, m, j, u)$ be the probability that at fixed x an MP with transition rates $a_n(x, i, s)$ and the initial state m will reach the state j at time u. For any $\theta > 0$, $j \neq i_0$, we have a representation:

$$
\begin{aligned}
&\mathbf{E}\exp\big\{-\theta\tau_{n1}(x_0, i_0)\big\}\chi\big(K\big(\beta_{n1}(x_0, i_0)\big) = j\big) \\
&\quad = \sum_{k=0}^{\infty}\mathbf{E}\int_{t_{nk}}^{t_{n,k+1}} \exp\big\{-\theta t_{n,k+1} - A_n\big(t, x_0, i_0\big)\big\}\prod_{l=0}^{k+1}\big(1 - c_n\big(x_{nl}, i_0\big)/V_n\big) \\
&\qquad\quad \times \sum_{m\in Z} a_n\big(x_{nk}, i_0, m\big)P_n\big(x_{nk}, m, j, t_{n,k+1} - t\big)\mathrm{d}t \\
&\qquad + \sum_{k=0}^{\infty}\mathbf{E}\exp\big\{-\theta t_{n,k+1} - A_n\big(t_{n,k+1}, x_0, i_0\big)\big\} \\
&\qquad\qquad \times \prod_{l=1}^{k}\big(1 - c_n(x_{nl}, i_0\big)/V_n\big)c_n\big(x_{n,k+1}, i_0, j\big)/V_n = \Sigma_n(1) + \Sigma_n(2).
\end{aligned}
\tag{10.8}
$$

Note that at $m \neq j$,

$$
P_n(x, m, j, u) \leq \mathbf{P}\big(\text{at least one jump exists in } [0, u]\big) \leq 1 - e^{-a_n(x,m)u}.
$$

Correspondingly, $1-P_n(x,j,j,u) \leq 1-e^{-a_n(x,j)u}$. As $t_{n,k+1}-t_{nk} = \theta_n(x_{nk}) = \widetilde{\theta}_n(x_{nk})/V_n$, then, using the inequality $1-e^{-\alpha} \leq \alpha$ as $\alpha \geq 0$, it is not difficult to prove that for any $a > 0$, $L > 0$,

$$\mathbf{E}\int_0^{\theta_n(x)} \big(1-e^{-au}\big)\mathrm{d}u \leq \frac{aL^2}{2V_n^2} + \frac{1}{V_n}\mathbf{E}\widetilde{\theta}_n(x)\chi\big(\widetilde{\theta}_n(x) > L\big). \tag{10.9}$$

Using (10.1), (10.2), (10.9) and the relations above, we obtain that for any x,

$$\begin{aligned}
&\mathbf{E}\int_{t_{nk}}^{t_{n,k+1}} \sum_{m\in Z} a_n\big(x,i_0,m\big)\big|P_n\big(x,m,j,t_{n,k+1}-t\big) - a_n\big(x,i_0,j\big)\big|\,\mathrm{d}t \\
&\quad\leq \mathbf{E}\int_0^{\theta_n(x)} \Bigg(a_n\big(x,i_0,j\big)\big(1-e^{-a_n(x,j)u}\big) \\
&\qquad\qquad + \sum_{m\neq j} a_n\big(x,i_0,m\big)P_n\big(x,m,j,\theta_n(x)-u\big)\Bigg)\mathrm{d}u \\
&\quad\leq \frac{C_{i_0}^{(2)}L^2}{2V_n^2} + \frac{C_{i_0}^{(1)}}{V_n}\mathbf{E}\widetilde{\theta}_n(x)\chi\big(\widetilde{\theta}_n(x) > L\big).
\end{aligned} \tag{10.10}$$

Denote

$$\begin{aligned}
\widetilde{\Sigma}_n(1) = \sum_{k=0}^{\infty}\mathbf{E}\int_{t_{nk}}^{t_{n,k+1}} &\exp\big\{-\theta t - A_n\big(t,x_0,i_0\big)\big\}a_n\big(x_{nk},i_0,j\big)\mathrm{d}t \\
&\times\prod_{l=0}^{k+1}\big(1-c_n\big(x_{nl},i_0\big)/V_n\big).
\end{aligned} \tag{10.11}$$

Using relations (10.9), (10.10), we obtain after some algebra that for any fixed $\theta > 0$,

$$\begin{aligned}
&\big|\Sigma_n(1) - \widetilde{\Sigma}_n(1)\big| \\
&\quad\leq \left(\frac{C_1L^2}{V_n^2} + \frac{C_2}{V_n}\sup_x \mathbf{E}\widetilde{\theta}_n(x)\chi\big(\widetilde{\theta}_n(x) > L\big)\right)\sum_{k=0}^{\infty}\mathbf{E}\exp\big\{-\theta t_{nk}\big\},
\end{aligned} \tag{10.12}$$

where C_1, C_2 are some constants. Let $\theta_{nl}(x)$, $l \geq 0$, be jointly independent variables having the same distribution as $\theta_n(x)$. Then $t_{nk} = \sum_{l=0}^{k-1}\theta_{nl}(x_{nl})$, and

$$\begin{aligned}
\sum_{k=0}^{\infty}\mathbf{E}e^{-\theta t_{nk}} &\leq \sum_{k=0}^{\infty}\Big(\max_{x\in X}\mathbf{E}\exp\big\{-\theta\theta_n(x)\big\}\Big)^k \\
&= \Big(1-\max_{x\in X}\mathbf{E}\exp\big\{-\theta\theta_n(x)\big\}\Big)^{-1} = O\big(V_n\big),
\end{aligned} \tag{10.13}$$

as for any $x \in X$, $\mathbf{E}\exp\{-\theta\theta_n(x)\} = 1 - \theta V_n^{-1}m(x) + o(V_n^{-1})$. Relations (10.7), (10.12), (10.13) imply $\Sigma_n(1) - \widetilde{\Sigma}_n(1) \to 0$.

To find the limits of values $\widetilde{\Sigma}_n(1)$ and $\Sigma_n(2)$ we prove first an auxiliary lemma.

LEMMA 10.1. *If (10.4), (10.7) hold, then for any bounded measurable function $f(x)$, $x \in X$,*

$$\int_0^t f\big(x_n(u)\big)\mathrm{d}u - t\widehat{f}_n/\widehat{m}_n \xrightarrow{\mathrm{P}} 0, \quad t \geq 0, \tag{10.14}$$

where $\widehat{f}_n = \int_X f(x)m_n(x)\pi_n(\mathrm{d}x)$.

Proof. Denote $\nu_n(t) = \min\{k : k \geq 0,\ t_{n,k+1} > t\}$. Let $\{\widetilde{\theta}_{nk}(x),\ x \in X\}$, $k \geq 1$, be the jointly independent in k families such that $\widetilde{\theta}_{nk}(x)$ has the same distribution as $\widetilde{\theta}_n(x)$. Then

$$\begin{aligned}\int_0^t f\big(x_n(u)\big)\mathrm{d}u - t\widehat{f}_n/\widehat{m}_n = V_n^{-1}\sum_{k=1}^{\nu_n(t)}\big(f\big(x_{nk}\big) - \widehat{f}_n/\widehat{m}_n\big)\widetilde{\theta}_{nk}\big(x_{nk}\big) \\ + \big(f\big(x_{n,\nu_n(t)}\big) - \widehat{f}_n/\widehat{m}_n\big)\big(t - t_{n,\nu_n(t)}\big),\end{aligned} \tag{10.15}$$

where $\sum_1^0 = 0$. First we prove that the 1st term in the right-hand side of (10.15) tends in probability to 0. Put

$$s_{nj} = V_n^{-1}\sum_{k=1}^{j}\big(f\big(x_{nk}\big) - \widehat{f}_n/\widehat{m}_n\big)\widetilde{\theta}_{nk}\big(x_{nk}\big), \quad h_{nj} = V_n^{-1}\sum_{k=1}^{j}\widetilde{\theta}_{nk}\big(x_{nk}\big).$$

Consider the auxiliary processes: $s_n(u) = s_{nj}$, $h_n(u) = h_{nj}$, as $V_n^{-1}j \leq u < V_n^{-1}(j+1)$. Denote $\mu_n(t) = \inf\{u : u > 0,\ h_n(u) \geq t\}$. By the construction, $\mu_n(t) = V_n^{-1}(\nu_n(t)+1)$. First, let us prove that $\mu_n(t) - tm^{-1} \xrightarrow{\mathrm{P}} 0$. Using (10.7) we obtain:

$$\begin{aligned}\mathbf{E}e^{-\theta h_n(u)} &= \mathbf{E}\prod_{k=0}^{[V_n u]}\mathbf{E}\big[e^{-\theta V_n^{-1}\widetilde{\theta}_{nk}(x_{nk})} \mid x_{nk}\big] \\ &= \mathbf{E}\prod_{k=0}^{[V_n u]}\big(1 - \theta m_n\big(x_{nk}\big)/V_n + o_{nk}\big(1/V_n\big)\big) \\ &\approx \mathbf{E}\exp\Bigg\{-\theta V_n^{-1}\sum_{k=0}^{[V_n u]} m_n\big(x_{nk}\big)\Bigg\}.\end{aligned} \tag{10.16}$$

Here $[a]$ means the integer value of a, and $V_n\max_{k\leq[nu]}|o_{nk}(1/V_n)| \to 0$.

According to properties of $\varphi_n(\cdot)$ (for instance see [BIL 68, ANI 88]), we have: $|\mathbf{P}(x_{nk} \in A) - \pi_n(A)| \le \varphi_n(k)$, and (10.4) implies that $\varphi_n(k) \le q^{[k/r_n]}$. Using the known inequality: for any probability measures $P(A)$, $Q(A)$

$$\left| \int_X f(x)P(\mathrm{d}x) - \int_X f(x)Q(\mathrm{d}x) \right| \le 2 \sup_x \bigl|f(x)\bigr| \sup_A \bigl|P(A) - Q(A)\bigr|,$$

we obtain the following inequalities, where $C_n = \sup_x m_n(x)$. For $k < j$,

$$\bigl|\mathbf{E}m_n\bigl(x_{nk}\bigr) - \mathbf{E}m_n\bigl(x_{nj}\bigr)\bigr| \le 2C_n\varphi_n(j-k);$$

$$\bigl|\mathbf{E}m_n\bigl(x_{nk}\bigr) - \widehat{m}_n\bigr| \le 2C_n\varphi_n(k),$$

$$\bigl|\mathbf{E}\bigl(m_n\bigl(x_{nk}\bigr) - \widehat{m}_n\bigr)\bigl(m_n\bigl(x_{nj}\bigr) - \widehat{m}_n\bigr)\bigr| \le 8C_n^2\varphi_n(j-k) + 4C_n^2\varphi_n(k)\varphi_n(j).$$

Relation (10.7) implies that $\sup_x m_n(x) \le C < \infty$. Thus, (10.4) implies:

$$\begin{aligned}
&\mathbf{E}\Biggl(V_n^{-1}\sum_{k=0}^{[V_n u]}\bigl(m_n\bigl(x_{nk}\bigr) - \widehat{m}_n\bigr)\Biggr)^2 \\
&\quad \le 8V_n^{-2}\sum_{k=1}^{[V_n u]}\mathbf{E}\bigl(m_n\bigl(x_{nk}\bigr) - \widehat{m}_n\bigr)^2 + 4C^2\sum_{1\le k<j\le[V_n u]} q^{[k/V_n]}q^{[j/V_n]} \\
&\quad \le 8C(1-q)^{-1}V_n^{-2}\bigl[V_n u\bigr]r_n + 4C^2V_n^{-2}r_n^2 \longrightarrow 0.
\end{aligned}$$

This means that for any $u > 0$, $h_n(u) \xrightarrow{\text{P}} mu$. As $h_n(u)$ monotonically increases, then $h_n(u)$ converges in probability to mu uniformly in any finite interval, and $\mu_n(t)$ converges in probability to t/m. In particular, $V_n^{-1}\nu_n(t) \xrightarrow{\text{P}} t/m$, and $V_n^{-1}(\nu_n(t)+1) \xrightarrow{\text{P}} t/m$. This implies that $V_n^{-1}\widetilde{\theta}_{n,\nu_n(t)+1}(x_{n,\nu_n(t)+1}) \xrightarrow{\text{P}} 0$, and also $V_n^{-1}(t - t_{n,\nu_n(t)}) \xrightarrow{\text{P}} 0$. Now from (10.15) we see that the asymptotic behavior of the left-hand side of (10.14) and of the 1st term in the right-hand side of (10.15) is the same. In a similar way we prove that $s_n(u) \xrightarrow{\text{P}} 0$, uniformly in any interval $[0,T]$. Using the results on the convergence of a superposition of random functions [BIL 68], we obtain that $s_n(\mu_n(t)) \xrightarrow{\text{P}} 0$. Finally this implies (10.14) and proves Lemma 10.1. □

Now we return to the proof of Theorem 10.1. It follows from Lemma 10.1 that $A_n(t, x_{n0}, i_{n0}) - t\widehat{a}_n(i_{n0})\widehat{m}_n^{-1} \xrightarrow{\text{P}} 0$. Taking into account that the function e^{-z} is continuous and bounded in the domain $\{z \ge 0\}$ and $c_n(\cdot)$ satisfy condition (10.1), we obtain:

$$\begin{aligned}
\mathbf{E}\prod_{k=1}^{[V_n u]}\bigl(1 - c_n\bigl(x_{nk}, i_{n0}\bigr)/V_n\bigr) &\approx \mathbf{E}\exp\Biggl\{-V_n^{-1}\sum_{k=1}^{[V_n u]}c_n\bigl(x_{nk}, i_{n0}\bigr)\Biggr\} \\
&\approx \exp\bigl\{-u\widehat{c}_n\bigl(i_{n0}\bigr)\bigr\},
\end{aligned}$$

$$\mathbf{E}\prod_{t_{nk}\le t}\big(1-c_n\big(x_{nk},i_{n0}\big)/V_n\big)\approx\mathbf{E}\exp\Big\{-V_n^{-1}\sum_{k=1}^{\nu_n(t)}c_n\big(x_{nk},i_{n0}\big)\Big\}$$
$$\approx\exp\big\{-u\widehat{c}_n\big(i_{n0}\big)\widehat{m}_n^{-1}\big\},$$

where $\widehat{c}_n(i)$ are defined in (10.6). Consider $\widetilde{\Sigma}_n(1)$ in (10.11). First we note that $\widetilde{\Sigma}_n(1)=\Sigma_n(3)(1+O(C_i^{(3)}/V_n))$, where

$$\Sigma_n(3)=\mathbf{E}\int_0^\infty\exp\big\{-\theta t-A_n\big(t,x_{n0},i_{n0}\big)\big\}a_n\big(x_n(t),i_{n0},j\big)$$
$$\times\prod_{t_{nk}\le t}\big(1-c_n\big(x_{nk},i_{n0}\big)/V_n\big)\mathrm{d}t.$$

As $\theta>0$, the integral in the domain $t>L$ using (10.1) can be estimated uniformly in n by $Ce^{-\theta L}/\theta$. This value is small at large enough $L>0$. According to the convergence of $A_n(t,x_0,i_0)$ the integral in the domain $0\le t\le L$ after some algebra is asymptotically equivalent to the value

$$\int_0^L\exp\big\{-\theta t-\big(\widehat{a}_n\big(i_{n0}\big)+\widehat{c}_n\big(i_{no}\big)\big)\widehat{m}_n^{-1}t\big\}\mathbf{E}a_n\big(x_n(t),i_{n0},j\big)\,\mathrm{d}t$$
$$\approx\widehat{a}_n\big(i_{n0},j\big)\widehat{m}_n^{-1}\int_0^L\exp\big\{-\theta t-\big(\widehat{a}_n\big(i_{n0}\big)+\widehat{c}_n\big(i_{no}\big)\big)\widehat{m}_n^{-1}t\big\}\,\mathrm{d}t.$$

Finally, according to condition A), $\Sigma_n(3)$ converges to the expression

$$a_0\big(i_0,j\big)m^{-1}\int_0^\infty\exp\big\{-\theta t-\big(a_0\big(i_0\big)+c_0\big(i_0\big)\big)m^{-1}t\big\}\mathrm{d}t$$
$$=a_0\big(i_0,j\big)m^{-1}\big(\theta+\big(a_0\big(i_0\big)+c_0\big(i_0\big)\big)m^{-1}\big)^{-1}.$$

Now consider $\Sigma_n(2)$. Divide it on two sums, as $j\le[V_nL]$, and $j>[V_nL]$. The second sum is no greater than $\mathbf{E}\exp\{-\theta t_{n,[V_nL]}\}$. According to Lemma 10.1, $t_{n,[V_nL]}\approx L\widehat{m}_n$. Thus, the second sum is small at $\theta>0$ and large L. The first sum can be approximated by the expression:

$$\sum_{j=1}^{[V_nL]}\mathbf{E}\exp\big\{-\big(\theta\widehat{m}_n+\widehat{a}_n\big(i_{n0}\big)+\widehat{c}_n\big(i_{n0}\big)\big)j/V_n\big\}\widehat{c}_n\big(i_{n0},j\big)/V_n.$$

Using condition A), we can finally find the limit of $\Sigma_n(2)$ in the form:

$$\int_0^\infty\exp\big\{-\big(\theta m+a_0\big(i_0\big)+c_0\big(i_0\big)\big)t\big\}c_0\big(i_0,j\big)\,\mathrm{d}t$$
$$=c_0\big(i_0,j\big)m^{-1}\big(\theta+\big(a_0\big(i_0\big)+c_0\big(i_0\big)\big)m^{-1}\big)^{-1}.$$

Combining together limiting expressions for $\Sigma_n(3)$ and $\Sigma_n(2)$, we find that $\tau_{n1}(x_{n0}, i_{n0})$ and $\beta_{n1}(x_{n0}, i_{n0})$ are asymptotically independent, $\tau_{n1}(x_{n0}, i_{n0})$ weakly converges to the exponential variable with rate $\lambda(i_0)$, and $\beta_{n1}(x_{n0}, i_{n0})$ weakly converges to $\xi(i_0)$. This implies the weak convergence of finite dimensional distributions. As the limiting process is a stepwise process and the intervals between sequential jumps of $\zeta_n(\cdot)$ do not tend to zero, the weak convergence implies J-convergence. □

NOTE 10.1. If the state space for x_{nk} in the limit forms one irreducible aperiodic class, then condition (10.4) is satisfied and $r_n = r$ (some constant). Note that (10.4) is also satisfied for more general cases, when in limit the state space can be reducible and forms so-called V_n-s-set (all states in the time interval $[0, V_n]$ asymptotically communicate, see section 7.3, and [ANI 00]).

10.1.2. *Asymptotic aggregation of a semi-Markov environment*

Now consider the case, when the embedded MP x_{nk} is asymptotically reducible and allows an asymptotic aggregation of its state space. This means that the state space can be divided in the domains with asymptotically connected states with small transition probabilities between domains. To avoid complicated assumptions, suppose for simplicity that X and Z are discrete sets ($X=\{x_1, x_2, \ldots\}$, $Z=\{0, 1, \ldots\}$), and component $\zeta_n(\cdot)$ has no additional jumps at times t_{nk}. Let SMP $x_n(\cdot)$ be given by sojourn times $\{\theta_n(x),\ x \in X\}$, transition probabilities $\{p_n(x_1, x_2),\ x_1, x_2 \in X\}$ of the embedded MP x_{nk}, and the initial state x_{n0}. Then MPSMS $(x_n(t, x_{n0}), \zeta_n(t, i_{n0}))$ is constructed as above using transition rates $\{a_n(x, i, j),\ x \in X,\ i, j \in Z,\ i \neq j\}$, and the initial state (x_{n0}, i_{n0}). This means that if at time t, $x_n(t) = x$ and $\zeta_n(t, i_{n0}) = i$, then the instantaneous transition rate for $\zeta_n(t, i_{n0})$ from state i to j is $a_n(x, i, j)$, $j \neq i$. Let $P_n = \|p_n(x_1, x_2)\|_{x_1, x_2 \in X}$. Assume that

$$\begin{gathered} X = \cup_{y \in Y} X_y, \quad X_{y_1} \cap X_{y_2} = \emptyset, \quad y_1 \neq y_2, \\ P_n = P_n^{(0)} + V_n^{-1} B_n, \end{gathered} \tag{10.17}$$

where $P_n^{(0)}=\|p_n^{(0)}(x_1, x_2)\|_{x_1, x_2 \in X}$, is a stochastic matrix, $B_n=\|b_n(x_1, x_2)\|_{x_1, x_2 \in X}$, and $p_n^{(0)}(x_1, x_2) = 0$ as $x_1 \in X_{y_1}$, $x_2 \in X_{y_2}$, $y_1 \neq y_2$. This means that X is divided in disjoint domains with transition rates of the order $O(1/V_n)$ between them.

For each $y \in Y$ denote by $x_{nk}^{(y)}$, $k \geq 0$, an auxiliary MP with the state space X_y and the matrix of transition probabilities $P_n(y) = \|p_n^{(0)}(x_1, x_2)\|_{x_1, x_2 \in X_y}$. Let $\psi_n^{(y)}(k)$, $k > 0$, be the uniformly strong mixing coefficient for $x_{nk}^{(y)}$ (see (10.3)). Suppose that for any $y \in Y$, $\varphi_n^{(y)}(k)$ satisfies condition (10.4) with the same sequence r_n and the value $q < 1$. Let $\pi_n^{(y)}(x)$, $x \in X_y$, be the stationary distribution of $x_{nk}^{(y)}$. Denote as above $\widetilde{\theta}_n(x) = V_n \theta_n(x)$. In each domain X_y we average transition rates of $\zeta_n(\cdot)$ by the corresponding stationary measure similar to (10.5), (10.6) as follows:

$$m_n(x) = \mathbf{E}\widetilde{\theta}_n(x), \quad \widehat{m}_n^{(y)} = \sum_{x\in X_y} m_n(x)\pi_n^{(y)}(x), \tag{10.18}$$

$$\begin{aligned} \widehat{a}_n(y,i,j) &= \sum_{x\in X_y} a_n(x,i,j)m_n(x)\pi_n^{(y)}(x), \\ \widehat{a}_n(y,i) &= \sum_{j\neq i} \widehat{a}_n(y,i,j). \end{aligned} \tag{10.19}$$

Introduce the averaged transition rates between domains. Denote

$$\widehat{\Lambda}_n\big(y_1,y_2\big) = \big(m_n^{(y_1)}\big)^{-1} \sum_{x\in X_{y_1}} \pi_n^{(y_1)}(x) b_n\big(x, X_{y_2}\big), \quad y_1 \neq y_2, \tag{10.20}$$

where $b_n(x,A) = \sum_{x_1\in A} b_n(x,x_1)$, $A \in X$. Put $\widehat{\Lambda}_n(y) = \sum_{y_2\neq y} \widehat{\Lambda}_n(y,y_2)$. Suppose that the following conditions hold:

A1) $\widehat{m}_n^{(y)} \to m^{(y)} > 0$, $y \in Y$, and there exist values $a_0(y,i,j)$, $a_0(y,i)$, $y \in Y$, $i,j = 0,1,\ldots$, $i \neq j$, such that for any $y \in Y$,

$$\widehat{a}_n(y,i,j) \longrightarrow a_0(y,i,j), \quad \widehat{a}_n(y,i) \longrightarrow a_0(y,i), \quad i,j = 0,1,\ldots,\ i \neq j,$$

and $a_0(y,i) = \sum_{j\neq i} a_0(y,i,j)$.

B) for any $n > 0$, $\sup_{x,A} |b_n(x,A)| \leq C < \infty$, and there exist values $\Lambda_0(y_1,y_2)$, $\Lambda_0(y_1)$, $y_1,y_2 \in Y$, $y_1 \neq y_2$, such that

$$\widehat{\Lambda}_n\big(y_1,y_2\big) \longrightarrow \Lambda_0\big(y_1,y_2\big), \quad \widehat{\Lambda}_n\big(y_1\big) \longrightarrow \Lambda_0\big(y_1\big), \quad y_1,y_2 \in Y,\ y_1 \neq y_2,$$

and $\Lambda_0(y_1) = \sum_{y_2\neq y_1} \Lambda_0(y_1,y_2)$.

Denote by $y_0(t,y_0)$, $t \geq 0$, an MP with values in Y given by transition rates $\{\Lambda_0(y_1,y_2),\ y_1 \neq y_2\}$ and the initial state y_0. Let $(y_0(t,y_0),\zeta_0(t,i_0))$, $t \geq 0$, be a two-component MP with the initial state (y_0,i_0), where $\zeta_0(t,i_0)$ is an MP switched by $y_0(t,i_0)$ in the following way: if at time t, $y_0(t,y_0) = y$, then the local transition rates of $\zeta_0(t,i_0)$ are $a_0(y,i,j)/m^{(y)}$.

Denote by $K(\cdot)$ a map from X to Y such that $K(x) = y$ for all $x \in X_y$, $y \in Y$. Consider the aggregated process $K(x_n(\cdot,x_{n0}))$. Note that in general this process is not Markov.

THEOREM 10.2. *Suppose that the process* $(y_0(\cdot,y_0),\zeta_0(\cdot,i_0))$ *is regular, for any* $y \in Y$, $\varphi_n^{(y)}(k)$ *satisfies condition (10.4), conditions (10.1), (10.7), A1), B) hold and* $(K(x_{n0}),i_{n0}) \stackrel{\mathrm{w}}{\Longrightarrow} (y_0,i_0)$.

Then in any interval $[0,T]$ *process* $(K(x_n(t,x_{n0})),\zeta_n(t,i_{n0}))$ *J-converges to* $(y_0(t,y_0),\zeta_0(t,i_0))$.

Proof. We again use Theorem 8.3. Let us represent $(K(x_n(t, x_{n0})), \zeta_n(t, i_{n0}))$ as an SP. Denote by $0 = T_{n0} < T_{n1} < \cdots$ the sequential times of transitions between domains X_y for SMP $x_n(\cdot)$. They are constructed as follows: denote $y_{nk} = K(x_n(T_{nk}, x_{n0}))$, $k \geq 0$. Then recursively

$$T_{n1} = \inf\{t : t > 0,\ x_n(t, x_{n0}) \notin Y_{y_{n0}}\},$$

$$T_{n,k+1} = \inf\{t : t > T_{nk},\ x_n(t, x_{n0}) \notin Y_{y_{nk}}\}, \quad k > 0.$$

We represent T_{nk} as the times of jumps of sums of indicators constructed on an auxiliary MP. For each $y \in Y$ denote by $\widetilde{x}_{nk}^{(y)}(x_0)$, $k \geq 0$, an MP with the state space X_y, the initial state $x_0 \in X_y$ and the matrix of transition probabilities $\widetilde{P}_n(y) = \|\widetilde{p}_n(x_1, x_2)\|_{x_1, x_2 \in X_y}$, where

$$\widetilde{p}_n(x_1, x_2) = p_n(x_1, x_2) p_n(x_1, X_y)^{-1}, \quad p_n(x_1, X_y) = \sum_{x \in X_y} p_n(x_1, x).$$

Let for any $y \in Y$, $\{\chi_{nk}^{(y)}(x),\ x \in X_y\}$, $\{\beta_{nk}^{(y)}(x),\ x \in X_y\}$, and $\{\theta_{nk}(x),\ x \in X\}$, $k \geq 0$, be the jointly independent families of random variables such that

$$\mathbf{P}(\chi_{nk}^{(y)}(x) = 1) = 1 - \mathbf{P}(\chi_{nk}^{(y)}(x) = 0) = 1 - p_n(x, X_y),$$

$$\mathbf{P}(\beta_{nk}^{(y)}(x) = x_1) = p_n(x, x_1)(1 - p_n(x, X_y))^{-1}, \quad x_1 \notin X_y,$$

and $\mathbf{P}(\theta_{nk}(x) \leq z) = \mathbf{P}(\theta_n(x) \leq z)$, $z \geq 0$. Consider for any $y \in Y$ the auxiliary stepwise processes

$$\chi_n^{(y)}(m, x_0) = \sum_{k=0}^{m} \chi_{nk}^{(y)}(\widetilde{x}_{nk}^{(y)}(x_0)),$$

$$y_n^{(y)}(m, x_0) = \sum_{k=0}^{m} \theta_{nk}(\widetilde{x}_{nk}^{(y)}(x_0)), \quad x_0 \in X_y,\ m \geq 0.$$

Denote

$$\mu_n(y, x_0) = \min\{m : m \geq 0,\ \chi_n^{(y)}(m, x_0) > 0\},$$

$$\tau_n(X_y, x_0) = y_n^{(y)}(\mu_n(y, x_0), x_0), \quad x_0 \in X_y.$$

Also let $\widetilde{x}_n^{(y)}(t, x_0)$, $t \geq 0$, be an auxiliary SMP with values in X_y and the initial state x_0, which is constructed by the embedded MP $\widetilde{x}_{nk}^{(y)}(x_0)$ and sojourn times $\theta_{nk}(x)$.

We define an SP $\kappa_n(t)$, $t \geq 0$, as follows. If $x_n(0) = x_0 \in X_{y_0}$, we set $\kappa_n(0) = x_0$, and in the interval $[0, \tau_n(X_{y_0}, x_0))$, $\kappa_n(t) = \widetilde{x}_n^{(y_0)}(t, x_0)$. If $\widetilde{x}_n^{(y_0)}(\tau_n(X_{y_0}, x_0) - 0, x_0) = \widetilde{x}_0$, then $\mathbf{P}(\kappa_n(\tau_n(X_{y_0}, x_0)) = x_1) = \mathbf{P}(\beta_{n1}^{(y_0)}(\widetilde{x}_0) = x_1)$, where $x_1 \notin X_{y_0}$.

Furthermore, if $\beta_{n1}^{(y_0)}(\widetilde{x}_0) = x_1 \in X_{y_1}$, then in the next interval $[\tau_n(X_{y_0}, x_0), \tau_n(X_{y_0}, x_0) + \tau_n(X_{y_1}, x_1))$, $\kappa_n(t) = \widetilde{x}_n^{(y_1)}(t, x_1)$, and so on. It is not hard to check that the finite dimensional distributions of $\kappa_n(t)$ coincide with the corresponding distributions of the initial SMP $x_n(t, x_0)$.

Now we construct the process $\widetilde{\zeta}_n(t, i_{n0})$ with the initial state i_{n0} on the trajectory of $\kappa_n(t)$ as a locally MP with transition rates $a_n(x, i, j)$ in state $\kappa_n(t) = x$. Then the process $(\kappa_n(\cdot), \widetilde{\zeta}_n(\cdot, i_{n0}))$ has the same finite dimensional distributions as $(x_n(\cdot, x_{n0}), \zeta_n(\cdot, i_{n0}))$. Correspondingly, the aggregated process $(K(\kappa_n(\cdot)), \widetilde{\zeta}_n(\cdot, i_{n0}))$ is equivalent to $(K(x_n(\cdot, x_{n0})), \zeta_n(\cdot, i_{n0}))$.

The advantage of this construction is that $(\kappa_n(\cdot), \widetilde{\zeta}_n(\cdot, i_{n0}))$ is represented as an SP which is constructed by more simple processes defined in the domains X_y. Switching times here correspond to the times T_{nk} of transitions between domains X_y and are represented as the times of jumps of sums of indicators constructed on the auxiliary MPs defined in each domain. To study the limiting behavior of the initial complicated process we need to study the limiting behavior of the elementary processes in each domain.

Let us check first the asymptotic mixing property of MP $\widetilde{x}_{nk}^{(y)}(x_0)$. Consider a fixed domain X_y and denote:

$$P_n^{(0)}(x, B) = \sum_{u \in B} p_n^{(0)}(x, u), \quad \widetilde{P}_n(x, B) = \sum_{u \in B} \widetilde{p}_n(x, u), \quad x \in X_y,\ B \subset X_y.$$

Condition $\sup_{x,A} |b_n(x, A)| \leq C$ (see B)) implies that at large n,

$$\sup_{x,B} \left| P_n^{(0)}(x, B) - \widetilde{P}_n(x, B) \right| \leq C_1 / V_n,$$

where $C < C_1 < \infty$. Then for any $k > 1$ we obtain for k-step transition probabilities:

$$\sup_{x,B} \left| P_n^{(0)}(x, k, B) - \widetilde{P}_n(x, k, B) \right| \leq k C_1 / V_n$$

(see [ANI 88]). Denote by $\widetilde{\varphi}_n(k)$, $k > 0$, and $\varphi_n^{(0)}(k)$, $k > 0$, the uniformly strong mixing coefficients for $\widetilde{x}_{nk}^{(y)}$ and for an MP with values in X_y and transition probabilities $P_n^{(0)}(x, B)$, respectively (see (10.3)). Then, according to the relations above, $\widetilde{\varphi}_n(r_n) \leq \varphi_n^{(0)}(r_n) + 2 r_n C_1 / V_n$. So that, if $\varphi_n^{(0)}(k)$ satisfies condition (10.4), then for some q_1, $q < q_1 < 1$ at large enough n, $\widetilde{\varphi}_n(r_n) \leq q_1$. This means that we can apply Lemma 10.1 to an MP $\widetilde{x}_{nk}^{(y)}(x_0)$, as well. If $\mathbf{P}(x_{n0} \in X_y) = 1$, then according to Lemma 10.1 and condition A1) uniformly in u in each interval $[0, T]$,

$y_n^{(y)}([V_n u], x_{n0}) \xrightarrow{\mathrm{P}} m^{(y)} u$. Furthermore, according to condition B)

$$\begin{aligned}\mathbf{E} e^{-\theta \chi_n^{(y)}([V_n u], x_{n0})} &= \mathbf{E} \prod_{k=0}^{[V_n u]} \mathbf{E}\Big[e^{-\theta \chi_{nk}^{(y)}(\widetilde{x}_{nk}^{(y)}(x_{n0}))} \mid \widetilde{x}_{nk}^{(y)}(x_{n0})\Big] \\ &= \mathbf{E} \prod_{k=0}^{[V_n u]} \big(1 - p_n\big(\widetilde{x}_{nk}^{(y)}(x_{n0}), X_y\big)\big(1 - e^{-\theta}\big)\big) \\ &\approx \mathbf{E} \exp\left\{ -\frac{1}{V_n} \sum_{k=0}^{[V_n u]} b_n\big(\widetilde{x}_{nk}^{(y)}(x_{n0}), X_y\big)\big(1 - e^{-\theta}\big)\right\} \\ &\approx \mathbf{E} \exp\big\{ -u\Lambda_0(y)\big(1 - e^{-\theta}\big)\big\}.\end{aligned}$$

Writing the same relations for increments, we find that $\chi_n^{(y)}([V_n u], x_{n0})$ weakly converges to a Poisson process with parameter $\Lambda_0(y)$. Correspondingly, $\mu_n(y, x_{n0})$ weakly converges to the exponential random variable $\eta(\Lambda_0(y))$ with parameter $\Lambda_0(y)$, and $\tau_n(X_y, x_{n0})$, as a superposition of random functions, weakly converges to $\eta(\Lambda_0(y))m^{(y)}$.

This result shows that the time spent by $x_n(\cdot)$ in the domain X_y is asymptotically exponential with parameter $\Lambda_0(y)/m^{(y)}$. Using Lemma 10.1 we can also prove that the probability of a jump from X_y directly to X_{y_1} tends to $\Lambda_0(y, y_1)\Lambda_0(y)^{-1}$. This means that the aggregated process $K(x_n(\cdot))$ weakly converges to an MP $y_0(\cdot, y_0)$. As the transition rates are bounded, then $y_0(\cdot, y_0)$ has no simultaneous jumps a.s., which also implies J-convergence.

Finally, according to Lemma 10.1 and condition A1), process $\widetilde{\zeta}_n^{(y)}(\cdot, i_{n0})$, constructed on the trajectory of $\widetilde{x}_n^{(y)}(\cdot, x_{n0})$ using transition rates $a_n(x, i, j)$, J-converges to an MP with averaged transition rates $a_0(y, i, j)/m^{(y)}$. Using the representation of $(K(x_n(\cdot, x_{n0})), \zeta_n(\cdot, i_{n0}))$ as an SP and Theorem 8.3, we finally prove Theorem 10.2. □

Similar results can be proved if we have additional jumps at times t_{nk}.

Consider an example related to the convergence of the aggregated process. Let $V_n = n$ and matrix P_n in (10.17) has the form:

$$P_n = \begin{pmatrix} 1 - 3/n^\alpha & 1/n^\alpha & 2/n^\alpha \\ 2/n^\alpha & 1 - 2/n^\alpha - 1/n^\beta & 1/n^\beta \\ 3/n^\alpha & 2/n^\beta & 1 - 3/n^\alpha - 2/n^\beta \end{pmatrix},$$

where $\alpha > 0$, $\beta > 0$. If $\alpha < 1$, then $r_n = O(n^\alpha)$ and all states can be asymptotically averaged in the interval $[0, n]$ (we can use Theorem 10.1). If $\alpha = 1$, we are in the conditions of Theorem 10.2. If $\beta < 1$, then there are two domains: $X_1 = \{1\}$, $X_2 = \{2, 3\}$, and for domain X_2, $r_n = O(n^\beta)$. If $\beta = 1$, there are three domains: $X_1 = \{1\}$, $X_2 = \{2\}$, $X_3 = \{3\}$.

Consider as an example, the case $\alpha = 1$, $\beta < 1$. Then in the first domain X_1, $\pi_n^{(1)}(1) = 1$, and in X_2, $\pi_n^{(2)}(2) \approx 2/3$, $\pi_n^{(2)}(3) \approx 1/3$. Consider a map: $K(1) = y_1$, $K(2) = K(3) = y_2$. Suppose that $m_n(x) \to m(x)$, $x = 1, 2, 3$. In this case it is easy to calculate that $\widehat{m}^{(1)} = m(1)$, $\widehat{m}^{(2)} = (2m(2) + m(3))/3$, and using (10.20) we obtain: $\Lambda_0(y_1, y_2) = 3/m(1)$, $\Lambda_0(y_2, y_1) = 7/(2m(2) + m(3))$.

If for example $x_n(0) \in \{2, 3\}$, then $K(x_n(\cdot))$ J-converges to an MP $y_0(\cdot)$ with two states $\{y_1, y_2\}$, transition rates $\Lambda_0(y_i, y_j)$ and the initial state y_2.

Note that the results of sections 10.1.1, 10.1.2 can be extended to non-homogenous in time models (functions $a_n(\cdot)$ and process $x_n(\cdot)$ may depend on time also).

10.1.3. *Approximation of a stationary distribution*

Results of section 10.1.1 deal with the approximation in the finite interval $[0, T]$. We study now the approximation of a stationary distribution. Consider process $(x_n(t), \zeta_n(t))$, $t \geq 0$, introduced in section 10.1.1. To avoid complicated assumptions, suppose that $x_n(t) = x_0(V_n t)$, where $x_0(t)$, $t \geq 0$, is a right-continuous SMP with values in finite space $X = \{x_1, \dots, x_m\}$, $V_n \to \infty$, and we do not have additional jumps of $\zeta_n(\cdot)$ at times t_{nk} of jumps of $x_n(\cdot)$, so that $c_n(x, i, j) \equiv 0$, $i \neq j$. Process $\zeta_n(\cdot)$ is constructed as before on the trajectory $x_n(\cdot)$ as a local MP using the family of transition rates $\{a_n(x, i, j),\ i, j \in Z,\ i \neq j,\ x \in X\}$.

Denote by $\theta(x)$ a sojourn time of $x_0(\cdot)$ in state $x \in X$. Suppose that $\mathbf{E}\theta(x) = m(x) < \infty$, $x \in X$, the embedded MP for $x_0(\cdot)$ is irreducible and at least one of the variables $\theta(x)$ is non-lattice. Then $x_0(\cdot)$ is ergodic. Denote by $\rho(x)$, $x \in X$, its stationary distribution. Put

$$\widehat{\lambda}_n(i, j) = \sum_{x \in X} a_n(x, i, j)\rho(x), \quad \widehat{\lambda}_n(i) = \sum_{j \neq i} \widehat{\lambda}_n(i, j).$$

Suppose that for any $i, j \in Z$, $i \neq j$,

$$\widehat{\lambda}_n(i, j) \longrightarrow \lambda_0(i, j), \quad \widehat{\lambda}_n(i) \longrightarrow \lambda_0(i), \quad \lambda_0(i) = \sum_{j \neq i} \lambda_0(i, j). \tag{10.21}$$

Let $\zeta_0(t)$, $t \geq 0$, be a regular MP given by transition rates $\lambda_0(i, j)$, $i, j \in Z$, $i \neq j$. The following result follows from Theorem 10.1.

STATEMENT 10.1. *Suppose that values $a_n(x, i)$ satisfy (10.1), (10.21) holds and $\zeta_n(0) \stackrel{w}{\Longrightarrow} \zeta_0(0)$. Then in any interval $[0, T]$, process $\zeta_n(\cdot)$ J-converges to $\zeta_0(\cdot)$.*

Now we consider the approximation of a stationary distribution. Denote by $\{\rho_n(x,i),\ x \in X,\ i \in Z\}$ a stationary distribution of $(x_n(\cdot),\zeta_n(\cdot))$ (if it exists). Suppose first that Z is finite.

THEOREM 10.3. *Let values $a_n(x,i)$ satisfy (10.1), relation (10.21) hold, there exist values $c_i > 0$, $i \in Z$, such that for any i, $\min_{x\in X} \liminf_{n\to\infty} a_n(x,i) \geq c_i$, $m(x) > 0$, $x \in X$, and MP $\zeta_0(\cdot)$ be ergodic with stationary distribution $\Pi(i)$, $i \in Z$. Then at large n, $(x_n(\cdot),\zeta_n(\cdot))$ is also ergodic and for any $x \in X$, $i \in Z$, $\rho_n(x,i) \to \rho(x)\Pi(i)$.*

Proof. Consider the process $(x_n(\cdot),\zeta_n(\cdot))$. Suppose that the initial value (x_0,i_0) is given. We construct the following recurrent sequences. Let t_{nk}, $k \geq 1$, and Ω_{nm}, $m \geq 1$, be the times of sequential jumps of $x_n(\cdot)$ and $\zeta_n(\cdot)$, respectively. Denote $T_{n1} = \Omega_{n1}$, $\widetilde{T}_{n1} = \min\{t_{nk} : t_{nk} > T_{n1}\}$, $T_{n,m+1} = \min\{\Omega_{nk} : \Omega_{nk} > \widetilde{T}_{nm}\}$, $\widetilde{T}_{n,m+1} = \min\{t_{nk} : t_{nk} > T_{n,m+1}\}$, $m \geq 1$. Let $X_{nm} = x_n(\widetilde{T}_{nm})$, $Z_{nm} = \zeta_n(\widetilde{T}_{nm}+0)$, $m \geq 1$. Then (X_{nm},Z_{nm}), $m \geq 1$, is the embedded MP for $(x_n(\cdot)$, $\zeta_n(\cdot))$. For any $x_1,x_2 \in X$, $i,j \in Z$, denote

$$p_n\big((x_1,i),(x_2,j)\big) = \mathbf{P}\big((X_{n,m+1},Z_{n,m+1}) = (x_2,j) \mid (X_{nm},Z_{nm}) = (x_1,i)\big).$$

Take $m = 1$. According to homogenity in time, we can put for simplicity in calculations $\widetilde{T}_{n1} = 0$. Then, by analogy to (10.8), given that $(X_{n1},Z_{n1}) = (x,i)$ we can write the following representation:

$$\mathbf{E}\widetilde{T}_{n2} = \mathbf{E}\sum_{k=0}^{\infty} t_{n,k+1}\int_{t_{nk}}^{t_{n,k+1}} \exp\big\{-A_n(u,x,i)\big\}a_n\big(x_n(u),i\big)\mathrm{d}u, \tag{10.22}$$

$$\mathbf{E}T_{n,2} = \mathbf{E}\int_0^{\infty} u\exp\big\{-A_n(u,x,i)\big\}a_n\big(x_n(u),i\big)\mathrm{d}u. \tag{10.23}$$

According to conditions of Theorem 10.3, the tail of the 2nd integral in the domain $\{u > L\}$ at $\varepsilon < c_i$ and large enough n can be approximated by the value

$$\int_L^{\infty} u\exp\big\{-\big(c_i-\varepsilon\big)u\big\}\big(C_i+\varepsilon\big)\mathrm{d}u = \frac{C_i+\varepsilon}{c_i-\varepsilon}\exp\big\{-\big(c_i-\varepsilon\big)L\big\},$$

which is small at large L. According to Theorem 10.1, the integral in the domain $\{u \leq L\}$ converges to $\int_0^L u\exp\{-\lambda_0(i)u\}\lambda_0(i)\mathrm{d}u$, so that the right-hand side in (10.23) converges to $\lambda_0(i)^{-1}$. Denote $\theta_{nk} = t_{n,k+1} - t_{n,k}$, $k \geq 0$, $\delta_n(L) = C_1L^2/V_n^2 + (C_2/V_n)\sup_x \mathbf{E}\widetilde{\theta}_n(x)\chi(\widetilde{\theta}_n(x) > L)$. Then, given that $(X_{n1},Z_{n1}) = (x,i)$, we obtain, following the same steps as at the proof of (10.12), (10.13) and using an inequality similar to (10.9), that

$$\begin{aligned}\mathbf{E}\big(\widetilde{T}_{n2}-T_{n2}\big) &= \mathbf{E}\sum_{k=0}^{\infty}\int_{t_{nk}}^{t_{n,k+1}}\big(t_{n,k+1}-u\big)\exp\big\{-A_n(u,x,i)\big\}a_n\big(x_n(u),i\big)\mathrm{d}u \\ &\le \mathbf{E}\sum_{k=0}^{\infty}\theta_{nk}e^{-A_n(t_{nk},x,i)}\big(1-e^{-a_n(x_{nk},i)\theta_{nk}}\big) \\ &\le \delta_n(L)\sum_{k=0}^{\infty}\mathbf{E}\exp\big\{-c_i t_{nk}\big\} \\ &\le \delta_n(L)\sum_{k=0}^{\infty}\Big(\max_{x\in X}\mathbf{E}\exp\big\{-c_i V_n^{-1}\theta(x)\big\}\Big)^k \longrightarrow 0.\end{aligned}$$

Finally, according to Theorem 10.1, we have proved that for any $x \in X$ the distribution of the variable $\widetilde{T}_{n,m+1}-\widetilde{T}_{nm}$, given that $(X_{nm}, Z_{nm}) = (x, i)$, weakly converges to the exponential distribution with rate $\lambda_0(i)$ and its expectation converges to $1/\lambda_0(i)$.

Now consider the behavior of $p_n((x_1,i),(x_2,j))$. For any $t>0$, $\mathbf{P}(x_n(t,x_0)=x) \to \rho(x)$, $x \in X$. As according to Lemma 10.1, $A_n(t,x_0,i_0) \xrightarrow{\mathrm{P}} \lambda_0(i_0)t$, then, by adding and subtracting the term $\mathbf{E}e^{-\lambda_0(i_0)t}\mathbf{E}\chi(x_n(t,x_0)=x)$, we obtain:

$$\Big|\mathbf{E}e^{-A_n(t,x_0,i_0)}\chi\big(x_n\big(t,x_0\big)=x\big)-e^{-\lambda_0(i_0)t}\mathbf{E}\chi\big(x_n\big(t,x_0\big)=x\big)\Big| \longrightarrow 0. \quad (10.24)$$

Combining (10.24) with the result of Theorem 10.1 we obtain that for any $x_1, x_2 \in X$, $i, j \in Z$,

$$p_n\big(\big(x_1,i\big),\big(x_2,j\big)\big) \longrightarrow p_0\big(\big(x_1,i\big),\big(x_2,j\big)\big) = \rho\big(x_2\big)\lambda_0(i,j)/\lambda_0(i). \quad (10.25)$$

Denote by $\Delta_n(x,i)$ the total time spent by $\zeta_n(\cdot)$ in state $\{i\}$ in the interval $[\widetilde{T}_{n1},\widetilde{T}_{n,2})$ given that $(X_{n1}, Z_{n1}) = (x,i)$. By definition, $T_{n,2}-\widetilde{T}_{n1} \le \Delta_n(x,i) \le \widetilde{T}_{n,2}-\widetilde{T}_{n1}$. Then $\mathbf{E}\Delta_n(x,i) \to 1/\lambda_0(i)$. Let $\{\pi_n(x,i),\ x \in X,\ i \in Z\}$ be the stationary distribution of an MP with transition probabilities $p_n((x_1,i),(x_2,j))$. At large n it exists because transition probabilities $p_n(\cdot,\cdot)$ are close to $p_0(\cdot,\cdot)$ (see (10.25)), and a two-component MP constructed by $p_0(\cdot,\cdot)$ is irreducible, as $\zeta_0(\cdot)$ is ergodic. Let $\Pi_n(i)$, $i \in Z$, be the stationary distribution of the component $\zeta_n(\cdot)$. According to the ergodic theorem for semi-Markov renewal processes, for any $i \in Z$,

$$\Pi_n(i) = \sum_{x\in X}\pi_n(x,i)\mathbf{E}\Delta_n(x,i)\left(\sum_{x\in X,\ j\in Z}\pi_n(x,j)\mathbf{E}\Delta_n(x,j)\right)^{-1}, \quad (10.26)$$

and as $n \to \infty$,

$$\Pi_n(i) \longrightarrow \widehat{\pi}_0(i)\lambda_0(i)^{-1}\left(\sum_{j\in Z}\widehat{\pi}_0(j)\lambda_0(j)^{-1}\right)^{-1}, \quad i \in Z, \tag{10.27}$$

where $\widehat{\pi}_0(i) = \sum_{x\in X}\pi_0(x,i)$, and $\pi_0(x,i)$ is the stationary distribution of a two-component MP with transition probabilities $\rho(x_2)\lambda_0(i,j)/\lambda_0(i)$ from state (x_1,i) to (x_2,j).

Now consider an auxiliary MP $(\eta(t),\widetilde{\zeta}_0(t))$, $t \geq 0$, with values in $X \times Z$ and transition rates $\rho(x_2)\lambda_0(i,j)$ from state (x_1,i) to (x_2,j). Its embedded MP has transition probabilities $\rho(x_2)\lambda_0(i,j)/\lambda_0(i)$, and the sojourn time in state (x,i) has the expectation $1/\lambda_0(i)$. It is easy to check that the right-hand side in (10.27) is equal to the stationary distribution of MP $\widetilde{\zeta}_0(\cdot)$. As $\widetilde{\zeta}_0(\cdot)$ is equivalent to $\zeta_0(\cdot)$, this implies together with (10.24) the result of Theorem 10.3. □

Note that the multiplicative form of a limiting stationary distribution in Theorem 10.3 is in agreement with the results on the aggregation of a finite MP [COU 77].

If Z is infinite, we restrict our study to the case when $x_0(\cdot)$ is an MP, so that $\theta(x)$ has an exponential distribution with a rate $g(x)$, $0 < g(x) < \infty$, $x \in X$. In this case we obtain conditions which can be verified in particular applications. Let us keep the notation of Theorem 10.3. In this case we define switching times as the times of sequential jumps of $\zeta_n(\cdot)$. Put $(\widetilde{X}_{nm},\widetilde{Z}_{nm}) = (x_n(\Omega_{nm}+0),\zeta_n(\Omega_{nm}+0))$, $m \geq 1$. This MP is simpler than (X_{nm},Z_{nm}) constructed in Theorem 10.3. Given that $(\widetilde{X}_{n1},\widetilde{Z}_{n1}) = (x_0,i_0)$, denote by $\nu_n(x_0,i_0)$ a return time to domain (X,i_0) for $(\widetilde{X}_{nm},\widetilde{Z}_{nm})$.

THEOREM 10.4. *Suppose that values $a_n(x,i)$ satisfy (10.1), (10.21) holds, $x_0(\cdot)$ is ergodic with stationary distribution $\rho(x)$, $x \in X$, there exist values $c > 0$, $N > 0$, such that* $\min_{x\in X,\, i\in Z} a_n(x,i) \geq c$, *and for some (x_0,i_0), $\mathbf{E}\nu_n(x_0,i_0)^2 < C < \infty$ as $n > N$.*

Then $\zeta_0(\cdot)$ is ergodic and for any $x \in X$, $i \in Z$, $\rho_n(x,i) \to \rho(x)\Pi(i)$, where $\Pi(i)$, $i \in Z$, is the stationary distribution of $\zeta_0(\cdot)$.

Proof. If $\mathbf{E}\nu_n(x_0,i_0)^2 < C$, then (X_{nm},Z_{nm}), $m \geq 1$, is positive recurrent. According to our conditions, the expectation of a return time to state (x_0,i_0) for $(x_n(\cdot),\zeta_n(\cdot))$ is also finite, and the component $\zeta_n(\cdot)$ is ergodic with stationary distribution $\Pi_n(i)$ (see (10.26)), where $\Delta_n(x,i)$ should be replaced by the value $\Omega_{n,2} - \Omega_{n1}$ given that $(\widetilde{X}_{n1},\widetilde{Z}_{n1}) = (i,x)$. According to Theorem 10.3, for each state (x,i) transition probabilities and the expectation of a sojourn time converges to $p_0((x,i),\cdot)$ (see (10.25)) and $1/\lambda_0(i)$, respectively. As $\nu_n(x_0,i_0)$ is uniformly integrable, then $\mathbf{E}\nu_n(x_0,i_0)$ converge to the expectation of a return time to state i_0 for the process $\zeta_0(\cdot)$ with the initial state i_0 (as we have the convergence of the expectation of a sum of occupation times for any finite sequence of states of $(\widetilde{X}_{nm}, \widetilde{Z}_{nm})$). This means that $\Pi_n(i) \to \Pi(i)$, $i \in Z$. Together with (10.24) this proves the result of Theorem 10.4. □

Note that the condition $\mathbf{E}\nu_n(x_0, i_0)^2 < C$ can be verified for some classes of queueing models.

10.2. Averaging and aggregation in Markov queueing systems with semi-Markov switching

We illustrate the results of the previous section on the example of a state dependent system $M_{SM,Q}/M_{SM,Q}/1/\infty$ introduced in section 2.2.2 with fast semi-Markov switching.

10.2.1. *Averaging of states of the environment*

Suppose that $x_0(t)$, $t \geq 0$, is an ergodic SMP with values in a finite space $X = \{x_1, \ldots, x_r\}$ and a sojourn time $\theta(x)$ in state $\{x\}$. Denote by $\rho(x)$, $x \in X$, its stationary distribution and by $\pi(x)$, $x \in X$, the stationary distribution of the embedded MP. Let $\{\lambda(x,i), \mu(x,i), \alpha_A(x,i), \alpha_S(x,i),\ x \in X,\ i \geq 0\}$ be the families of non-negative functions and $V_n \to \infty$. We define a fast semi-Markov environment as $x_n(t) = x_0(V_n t)$. Let us denote by t_{nk}, $k \geq 0$, the sequential times of jumps of $x_n(t)$. Consider a queueing process $Q_n(t)$, $t \geq 0$, with switching governed by $x_n(t)$ in the following way: if $(x_n(t), Q_n(t)) = (x, i)$, then the local arrival rate is $\lambda(x,i)$ and the local service rate is $\mu(x,i)$. In addition, at any time t_{nk} of jump of $x_n(t)$ either an additional call may enter the system with probability $V_n^{-1}\alpha_A(x,i)$, or a call on service may complete service with probability $V_n^{-1}\alpha_S(x,i)$. Denote

$$
\begin{aligned}
\widehat{m} &= \sum_{x \in X} m(x)\pi(x), \\
\widehat{\lambda}(i) &= \sum_{x \in X} \lambda(x,i)m(x)\pi(x), \\
\widehat{\mu}(i) &= \sum_{x \in X} \mu(x,i)m(x)\pi(x), \\
\widehat{\alpha}_A(i) &= \sum_{x \in X} \alpha_A(x,i)\pi(x), \\
\widehat{\alpha}_S(i) &= \sum_{x \in X} \alpha_S(x,i)\pi(x), \\
A(i) &= \big(\widehat{\lambda}(i) + \widehat{\alpha}_A(i)\big)/\widehat{m}, \quad i \geq 0, \\
\Gamma(i) &= \big(\widehat{\mu}(i) + \widehat{\alpha}_S(i)\big)/\widehat{m}, \quad i \geq 1,\ \Gamma(0) = 0.
\end{aligned}
\tag{10.28}
$$

Note that $\rho(x) = m(x)\pi(x)/\widehat{m}$, $x \in X$. Let $M_Q/M_Q/1/\infty$ be the approximating state-dependent queueing system operating in the following way: as $Q(t) = i$, the local arrival rate is $A(i)$ and the service rate is $\Gamma(i)$, where $Q(t)$ is a number of calls in the system at time t. The following result is a consequence of Theorem 10.1.

STATEMENT 10.2. *Suppose that process $Q(\cdot)$ is regular and $Q_n(0) \stackrel{w}{\Longrightarrow} Q_0$. Then process $Q_n(t)$ J-converges in each finite interval $[0, T]$ to $Q(t)$ with $Q(0) = Q_0$.*

10.2.2. *Asymptotic aggregation of states of the environment*

Consider the system $M_{SM,Q}/M_{SM,Q}/1/\infty$ investigated in the previous section in the case when the process $x_n(\cdot)$ admits an asymptotic aggregation of the state space. Let $x_n(t)$, $t \geq 0$, be an SMP with state space $X = \{1, 2, \ldots, r\}$, given by the sojourn time $\theta(x)/V_n$ in state $\{x\}$ and the transition probability matrix $P_n = \|p_n(x_1, x_2)\|_{x_1,x_2 \in X}$ of the embedded MP x_{nk}. Assume that the conditions of asymptotic aggregation (10.17) hold, where for simplicity $P_n^{(0)} = P^{(0)}$ (does not depend on n), so that

$$p_n^{(0)}(x_1, x_2) = p_0(x_1, x_2), \quad x_1, x_2 \in X,$$

and also

$$b_n(x_1, x_2) \longrightarrow b_0(x_1, x_2), \quad x_1, x_2 \in X.$$

For each $y \in Y$ denote by $x_k^{(y)}$, $k \geq 0$, an MP with the state space X_y and transition probabilities $p_0(x_1, x_2)$, $x_1, x_2 \in X_y$. Let for any $y \in Y$, $x_k^{(y)}$ be irreducible with stationary distribution $\pi^{(y)}(x)$, $x \in X_y$. Put $m(x) = \mathbf{E}\theta(x)$ and denote

$$\widehat{m}^{(y)} = \sum_{x \in X_y} m(x)\pi^{(y)}(x),$$

$$\widehat{\lambda}(y, i) = \sum_{x \in X_y} \lambda(x, i)m(x)\pi^{(y)}(x),$$

$$\widehat{\mu}(y, i) = \sum_{x \in X_y} \mu(x, i)m(x)\pi^{(y)}(x),$$

$$\widehat{\alpha}_A(y, i) = \sum_{x \in X_y} \alpha_A(x, i)\pi^{(y)}(x),$$

$$\widehat{\alpha}_S(y, i) = \sum_{x \in X_y} \alpha_S(x, i)\pi^{(y)}(x),$$

$$A(y, i) = \widehat{\lambda}(y, i)/\widehat{m}^{(y)} + \widehat{\alpha}_A(y, i)/\widehat{m}^{(y)}, \quad i \geq 0,$$

$$\Gamma(y,i) = \widehat{\mu}(y,i)/\widehat{m}^{(y)} + \widehat{\alpha}_S(y,i)/\widehat{m}^{(y)}, \quad i \geq 1, \ \big(\Gamma(y,0) \equiv 0\big),$$

$$\Lambda(y,z) = \big(\widehat{m}^{(y)}\big)^{-1} \sum_{x_1 \in X_y} \pi^{(y)}\big(x_1\big) \sum_{x_2 \in X_z} b_0\big(x_1, x_2\big), \quad y \neq z.$$

Let $y(t)$, $t \geq 0$, be an MP with values in Y and transition rates $\{\Lambda(y,z), y \neq z\}$. Consider a state-dependent Markov system with Markov switching $M_{M,Q}/M_{M,Q}/1/\infty$, which is described by a two-component MP $(y(t), Q(t))$, $t \geq 0$, in the following way: if $(y(t), Q(t)) = (y, i)$, then the local arrival rate is $A(y,i)$ and the local service rate is $\Gamma(y,i)$. This system stands for the approximating queueing system for the initial system with fast semi-Markov switching. As we see, the approximating system is a Markov one and has a simpler structure as it is governed by a Markov process with the aggregated state space. Let $K(\cdot)$ be a map from X to Y such that $K(x) = y$, $x \in X_y$, $y \in Y$. The following result is a consequence of Theorem 10.2.

STATEMENT 10.3. *Suppose that the conditions above are satisfied and as* $n \to \infty$, $(K(x_n(0)), Q_n(0)) \stackrel{\mathrm{w}}{\Longrightarrow} (y_0, Q_0)$. *Then in any interval* $[0,T]$ *the process* $(K(x_n(\cdot)), Q_n(\cdot))$ J*-converges to* $(y(\cdot), Q(\cdot))$ *where* $(y(0), Q(0)) = (y_0, Q_0)$.

10.2.3. *The approximation of a stationary distribution*

Now we study the approximation of a stationary distribution. First, we investigate the finite system $M_{SM,Q}/M_{SM,Q}/1/N$ with losses and fast semi-Markov switching considered in section 10.2.1. We keep the notation in (10.28) and put $\lambda(x, N+1) = 0$, $x \in X$. Consider process $(x_n(t), Q_n(t))$ and denote by $\rho_n(x,i)$, $x \in X$, $i = 0, \ldots, N$, its stationary distribution. Let $\Pi(i)$, $i = 0, \ldots, N+1$, be the stationary distribution of the approximating queueing system $M_Q/M_Q/1/N$ with rates $A(i), \Gamma(i)$ (see (10.28)). The following result is a consequence of Theorem 10.3.

STATEMENT 10.4. *Suppose that* $A(i) > 0$, $\Gamma(i) > 0$, $m(x) > 0$, $\lambda(x,i)+\mu(x,i) > 0$, $x \in X$, $i = 0, \ldots, N+1$. *Then at large* n, $\rho_n(x,i)$ *exists and* $\rho_n(x,i) \to \rho(x)\Pi(i)$, $x \in X$, $i = 0, \ldots, N$.

When a system is infinite ($N = \infty$), we restrict our investigation to the case when $x_0(\cdot)$ is an ergodic MP with the purpose of obtaining more transparent conditions. Denote by $\rho(x)$ a stationary distribution of $x_0(\cdot)$. Suppose for simplicity that we do not have additional jumps at times t_{nk}, i.e. $\alpha_A(x,i) \equiv 0$, $\alpha_S(x,i) \equiv 0$. Consider a system $M_{M,Q}/M_{M,Q}/1/\infty$ with fast switching by an MP $x_n(t) = x_0(V_n t)$. Denote by $\rho_n(x,i)$, $x \in X$, $i = 0, \ldots$, the stationary distribution of $(x_n(\cdot), Q_n(\cdot))$. Let $\Pi(i)$, $i = 0, \ldots$, be the stationary distribution of the approximating system $M_Q/M_Q/1/\infty$ with rates $\widehat{\lambda}(i), \widehat{\mu}(i)$ (see (10.28)).

STATEMENT 10.5. *Suppose that there exist constants* $0 < c < C < \infty$ *such that for any* $x \in X$, $i \geq 0$, $c \leq \lambda(x,i) + \mu(x,i) \leq C$, *and for some* $\delta > 0$ *and* $L > 0$,

$\widehat{\mu}(i) - \widehat{\lambda}(i) \geq \delta$ as $i > L$. Then at large n, $\rho_n(x,i)$ exists and $\rho_n(x,i) \to \rho(x)\Pi(i)$, $x \in X$, $i = 0, \ldots$.

Proof. Let Ω_{nm}, $m \geq 1$, be the times of sequential jumps of $Q_n(t)$. Put $\widetilde{X}_{nm} = x_n(\Omega_{nm})$, $\widetilde{Z}_{nm} = Q_n(\Omega_{nm} + 0)$, $m \geq 1$. Then $(\widetilde{X}_{nm}, \widetilde{Z}_{nm})$ is an MP. Denote

$$p_n\big((x_1,i),(x_2,j)\big) = \mathbf{P}\big((\widetilde{X}_{n,m+1}, \widetilde{Z}_{n,m+1}) = (x_2,j) \mid (\widetilde{X}_{nm}, \widetilde{Z}_{nm}) = (x_1,i)\big).$$

As functions $\lambda(\cdot), \mu(\cdot)$ are uniformly bounded, using Theorem 10.1 we can prove that uniformly in $x \in X$, $i \geq 1$, $p_n((x_1,i),(x_2,j)) \to p_0(i,(x_2,j))$, where

$$p_0\big(i,(x_2,j)\big) = 0, \quad \text{as } |i-j| > 1,$$
$$p_0\big(i,(x_2,i+1)\big) = \rho(x_2)\widehat{\lambda}(i)\big(\widehat{\lambda}(i) + \widehat{\mu}(i)\big)^{-1},$$
$$p_0\big(i,(x_2,i-1)\big) = \rho(x_2)\widehat{\mu}(i)\big(\widehat{\lambda}(i) + \widehat{\mu}(i)\big)^{-1},$$

and

$$\mathbf{E}\big[(\Omega_{n,m+1} - \Omega_{nm}) \mid (\widetilde{X}_{nm}, \widetilde{Z}_{nm}) = (x,i)\big] \longrightarrow \big(\widehat{\lambda}(i) + \widehat{\mu}(i)\big)^{-1}.$$

Then uniformly in $x \in X$, $i \geq 1$,

$$\begin{aligned} &\mathbf{E}\big[\widetilde{Z}_{n,m+1} - \widetilde{Z}_{nm} \mid (\widetilde{X}_{nm}, \widetilde{Z}_{nm}) = (x,i)\big] \\ &\quad \longrightarrow \big(\widehat{\lambda}(i) - \widehat{\mu}(i)\big)\big(\widehat{\lambda}(i) + \widehat{\mu}(i)\big)^{-1}. \end{aligned} \tag{10.29}$$

Thus, for some n_0, $(\widetilde{X}_{nm}, \widetilde{Z}_{nm})$ is irreducible as $n > n_0$. As for some $L > 0$, $\widehat{\mu}(i) - \widehat{\lambda}(i) \geq \delta$ at $i > L$, this condition implies that for some $\varepsilon > 0$ the left-hand side in (10.29) is no greater than $-\varepsilon$ as $i > L$. Then, according to the classic Foster criterion, at $n > N$ process $(\widetilde{X}_{nm}, \widetilde{Z}_{nm})$ is positive recurrent. Consider a finite domain $D = X \times \{0, \ldots, L\}$ and denote by $\nu_n(x, L+1, D)$ a return time to D for $(\widetilde{X}_{nm}, \widetilde{Z}_{nm})$ given that $(\widetilde{X}_{n0}, \widetilde{Z}_{n0}) = (x, L+1)$. In the same way as was done in Theorem 4.1 [ANI 01], we can prove that for some q, $0 < q < 1$, $\mathbf{P}(\nu_n(x, L+1, D) > k) \leq q^k$, $k > 1$, as $n > N$. This implies uniformly in $n > N$ the existence of the 2nd moment for $\nu_n(x, L+1, D)$ and for the return time to domain $(X, 0)$, respectively, and our result follows from Theorem 10.4. □

As a conclusion, note that queueing models in a Markov and moreover in a semi-Markov environment are very difficult for analytic study (for example, see [NEU 89]). The results of this chapter provide us with an effective approach to the approximate analytic analysis and simulation of queueing models in a random environment given the assumption that the transitions of the environment are much faster than the transitions of the queueing process.

10.3. Bibliography

[ANI 77] ANISIMOV V., "Switching processes", *Cybernetics*, vol. 13, no. 4, p. 590–595, 1977.

[ANI 78] ANISIMOV V., "Limit theorems for switching processes and their applications", *Cybernetics*, vol. 14, no. 6, p. 917–929, 1978.

[ANI 88] ANISIMOV V., "Estimates for deviations of transient characteristics of non-homogenous Markov processes", *Ukrainian Math. J.*, vol. 40, no. 6, p. 588–592, 1988.

[ANI 92] ANISIMOV V., "Averaging principle for switching processes", *Theor. Probab. and Math. Stat.*, vol. 46, p. 1–10, 1992.

[ANI 95] ANISIMOV V., "Switching processes: averaging principle, diffusion approximation and applications", *Acta Applicandae Mathematicae*, vol. 40, p. 95–141, 1995.

[ANI 98] ANISIMOV V., "Asymptotic analysis of stochastic models of hierarchic structure and applications in queueing models", in CHAKRAVARTHY S. and ALFA A., Eds., *Advances in Matrix Analytic Methods for Stochastic Models*, p. 237–259, Notable Publications, New Jersey, 1998.

[ANI 00] ANISIMOV V., "Asymptotic analysis of reliability for switching systems in light and heavy traffic conditions", in LIMNIOS N. and NIKULIN M., Eds., *Recent Advances in Reliability Theory: Methodology, Practice, and Inference*, p. 119–133, Birkhäuser Boston, Massachusetts, 2000.

[ANI 01] ANISIMOV V. and ARTALEJO J., "Analysis of Markov multiserver retrial queues with negative arrivals", *Queueing Systems*, vol. 39, no. 2/3, p. 157–182, 2001.

[ANI 02] ANISIMOV V., "Averaging in Markov models with fast Markov switches and applications to queueing models", *Annals of Operations Research*, vol. 112, no. 1, p. 63–82, 2002.

[BIL 68] BILLINGSLEY P., *Convergence of Probability Measures*, Wiley, New York, 1968.

[COU 77] COURTOIS P., *Decomposability: Queueing and Computer Systems Applications*, Academic Press, New York, 1977.

[ETH 86] ETHIER S. and KURTZ T., *Markov Processes, Characterization and Convergence*, John Wiley & Sons, New York, 1986.

[KUS 90] KUSHNER H., *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser Boston, Massachusetts, 1990.

[NEU 89] NEUTS M., *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York & Basel, 1989.

[SKO 89] SKOROKHOD A., *Asymptotic Methods in the Theory of Stochastic Differential Equations*, Amer. Math. Soc., Rhode Island, 1989.

Chapter 11

Other Applications of Switching Processes

In this chapter we consider the applications of AP and DA for SP to several other problems such as self-organization in multicomponent interacting Markov models, and dynamic systems and random movements with fast Markov switching.

11.1. Self-organization in multicomponent interacting Markov systems

Consider the phenomena of self-organization that appear at the asymptotic analysis of some interactive multi-component Markov systems and at the analysis of migration models [WEI 88, WEI 91]. The presentation follows [ANI 95b]. First let us specify in which sense we understand the phenomenon of self-organization. Consider the differential equation (4.22) for the limiting function $s(\cdot)$ in AP for SP, Chapter 4, which has the form

$$\mathrm{d}s(t) = \widetilde{b}\big(s(t)\big)\mathrm{d}t, \quad s(0) = s_0, \tag{11.1}$$

where $\widetilde{b}(s) = b(s)/m(s)$. Let s_* be a stability point such that

$$\widetilde{b}\big(s_*\big) = 0, \tag{11.2}$$

and there exists a domain $D(s^*)$ such that if $s_0 \in D(s_*)$ then $s(t) \to s_*$ as $t \to \infty$.

This means that s_* is a stability point for equation (11.1) and relation (4.19) implies for any $\varepsilon > 0$:

$$\lim_{t\to\infty} \limsup_{n\to\infty} \mathbf{P}\big\{\big|S_n(nt)/n - s_*\big| > \varepsilon\big\} = 0. \tag{11.3}$$

This relation describes the phenomenon of self-organization in switching system.

Now we study this phenomenon in a model of asymptotically large number of interactive Markov systems with mutual influence between individual systems and the whole super-system. Consider n interacting Markov systems developing in the following way. The systems are identical with finite set of states $\{1, 2, \dots, r\}$. The family of non-negative functions $\lambda_{ij}(\overline{q})$, $i, j = \overline{1, r}$, $i \neq j$, $\overline{q} = (q_1, \dots, q_r)$, $q_i \geq 0$, $\sum_{i=1}^{r} q_i = 1$ is given. Let $\nu_n(i, t)$ be the total number of systems which are in state i in moment t. Denote by

$$\overline{\nu}_n(t) = \left(\frac{1}{n}\nu_n(i, t),\ i = \overline{1, r}\right)$$

the vector of proportions of the number of systems in particular states. If at time t, $\overline{\nu}_n(t) = \overline{q}$ and a system is in state i, then this system in a small interval $[t, t+h]$ independently of the other systems may pass to state j with probability $n^{-1}\lambda_{ij}(\overline{q})h + o(h)$, $j \neq i$. Let us introduce a column-vector

$$\widetilde{b}(\overline{q}) = \left(- q_i\lambda_i(\overline{q}) + \sum_{k\neq i} q_k\lambda_{ki}(\overline{q}),\ i = \overline{1, r}\right), \tag{11.4}$$

where $\lambda_i(\overline{q}) = \sum_{k\neq i} \lambda_{ik}(\overline{q})$, and put $\lambda(\overline{q}) = \sum_{i=1}^{r} q_i\lambda_i(\overline{q})$.

THEOREM 11.1. *Suppose that $\overline{\nu}_n(0) = \overline{s}_0$ and functions $\lambda_{ij}(\overline{q})$ satisfy local Lipschitz conditions. Then*

$$\sup_{t\leq T} \left|\overline{\nu}_n(nt) - \overline{s}(t)\right| \xrightarrow{\text{P}} 0, \tag{11.5}$$

where

$$\mathrm{d}\overline{s}(t) = \widetilde{b}\big(\overline{s}(t)\big)\mathrm{d}t, \quad \overline{s}(0) = \overline{s}_0, \tag{11.6}$$

and T is any positive value such that $y(+\infty) > T$, where

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

$$d\overline{\eta}(u) = \widetilde{b}\big(\overline{\eta}(u)\big)\lambda\big(\overline{\eta}(u)\big)^{-1}\mathrm{d}u, \quad \overline{\eta}(0) = \overline{s}_0.$$

Proof. In our case $\overline{\nu}_n(t)$ is a homogenous Markov process and it is easy to see that it can be described as a simple RPSM where $\tau_{n1}(n\overline{q})$ has the exponential distribution with parameter $\lambda(\overline{q})$ and

$$\xi_{n1}(n\overline{q}) = \left\{ \frac{1}{n}\overline{e}_{ij} \quad \text{with probability } q_i\lambda_{ij}(\overline{q})\lambda(\overline{q}){-}1, \quad i, j = \overline{1, r},\ i \neq j, \right. \tag{11.7}$$

where $\overline{e}_{ij}$ is a column-vector for which the ith component is equal to -1, the jth component is equal to 1 and other components are equal to 0. Thus, $\mathbf{E}\tau_{n1}(n\overline{q}) = \lambda(\overline{q})^{-1}$, $\mathbf{E}\xi_{n1}(n\overline{q}) = n^{-1}\widetilde{b}(\overline{q})\lambda(\overline{q})^{-1}$, and our statement follows directly from AP for RPSM in Theorem 4.3, section 4.3. □

Therefore, for analyzing self-organization phenomena we have to analyze the points of stability of equation (11.6). These results can be used for analyzing the problems of self-organization for multi-component Markov models of the society suggested by W. Weidlich and W. Weidlich and G. Haag in [WEI 88, WEI 91] and provide us with a new approach in the investigation of "synergetic" phenomena in these models.

Now let us formulate the theorem on the DA. Denote by symbol $\overline{f}'(\overline{a})$ a matrix of partial derivatives of the vector-function $\overline{f}(\overline{a})$:

$$\lim_{h\to 0} h^{-1}\big(\overline{f}(\overline{a}+h\overline{z})-\overline{f}(\overline{a})\big)=\overline{f}'(\overline{a})z.$$

Introduce a matrix $B^2(\overline{q})=\|b_{ij}(\overline{q})\|$, $i,j=\overline{1,r}$, where

$$b_{ij}(\overline{q})=-q_i\lambda_{ij}(\overline{q})-q_j\lambda_{ji}(\overline{q}),\quad i\neq j, \tag{11.8}$$

$$b_{ii}(\overline{q})=q_i\lambda_i(\overline{q})+\sum_{k\neq i}q_k\lambda_{ki}(\overline{q}). \tag{11.9}$$

THEOREM 11.2. *Let the conditions of Theorem 11.1 be satisfied, a continuous function* $Q(\overline{q})=\widetilde{b}'(\overline{q})$ *exists and* $\sqrt{n}(\overline{\nu}_n(0)-s_0)\stackrel{\mathrm{w}}{\Longrightarrow}\overline{\gamma}_0$.

Then the sequence of processes $\overline{\gamma}_n(t)=\sqrt{n}(\overline{\nu}_n(nt)-\overline{s}(t))$ *J-converges in the interval* $[0,T]$ *to the diffusion process* $\overline{\gamma}(t)$ *satisfying the following stochastic differential equation:*

$$\mathrm{d}\overline{\gamma}(t)=Q\big(\overline{s}(t)\big)\overline{\gamma}(t)\mathrm{d}t+B\big(\overline{s}(t)\big)\mathrm{d}\overline{w}(t),\quad \overline{\gamma}(0)=\overline{\gamma}_0,$$

where $\overline{w}(t)$ *is a standard Wiener process in* R^r.

Proof. Let us use DA for RPSM in Theorem 4.4, section 4.3. As $\tau_{n1}(\cdot)$ has an exponential distribution, according to Corollary 4.1, the matrix function $D_n^2(\alpha)$ in Theorem 4.4 has the form $D_n^2(\alpha)=\mathbf{E}\xi_{n1}(n\alpha)\xi_{n1}(n\alpha)^*$ and by the definition of $\xi_{n1}(n\alpha)$ (11.7) it is easy to calculate that in our case

$$\mathbf{E}\rho_n(\overline{q})\rho_n(\overline{q})^*=\mathbf{E}\xi_{n1}(nq)\xi_{n1}(nq)^*=\frac{1}{n^2}B^2(\overline{q})\lambda(\overline{q})^{-1},$$

where $\rho_n(\overline{q})=\xi_{n1}(n\overline{q})-E\xi_{n1}(n\overline{q})-\frac{1}{n}\widetilde{b}_n(\overline{q})(\tau_{n1}(n\overline{q})-\lambda(\overline{q})^{-1})$. Therefore our statement directly follows from Theorem 4.4 and Corollary 4.1, section 4.3. □

Note that if $\overline{q}_*$ is a stability point and $\overline{s}_0\in D(\overline{q}_*)$, then the finite dimensional distributions of process $\overline{\gamma}(t)$ at large t are close to the distributions of the stationary diffusion process $\widetilde{\gamma}(t)$ satisfying the following stochastic differential equation:

$$\mathrm{d}\widetilde{\gamma}(t)=Q\big(\overline{q}_*\big)\widetilde{\gamma}(t)+B\big(\overline{q}_*\big)d\overline{w}(t),\quad \widetilde{\gamma}(0)=\overline{\gamma}_0, \tag{11.10}$$

where $\overline{q}_*$ is a stationary point of equation (11.6).

Let us consider some examples.

EXAMPLE 11.1. Let $r = 2$, $\lambda_{12}(\bar{q}) \equiv \alpha$, $\lambda_{21}(\bar{q}) \equiv \beta$. Then $\lambda_1(\bar{q}) \equiv \alpha$, $\lambda_2(\bar{q}) \equiv \beta$ and system (11.2) with function $\widetilde{b}(\bar{q})$ defined by relation (11.4) has the form:

$$\begin{cases} -q_1\alpha + q_2\beta = 0 \\ -q_2\beta + q_1\alpha = 0 \end{cases} \tag{11.11}$$

with condition $q_1 + q_2 = 1$. From (11.11) we obtain:

$$q_1^* = \beta(\alpha+\beta)^{-1}, \quad q_2^* = \alpha(\alpha+\beta)^{-1}. \tag{11.12}$$

It can be verified that this solution is the point of stability of system (11.11). Denote $\widetilde{b}(\bar{q}) = (b_1(\bar{q}), b_2(\bar{q}))$. Then in our example according to (11.4)

$$b_1(\bar{q}) = -q_1\alpha + q_2\beta = \beta - (\alpha+\beta)q_1, \quad b_2(\bar{q}) = -b_1(\bar{q}).$$

Then system (11.6) is reduced to one equation

$$\mathrm{d}s_1(t) = \big(\beta - (\alpha+\beta)s_1(t)\big)\mathrm{d}t, \quad s_1(0) = s_0^{(1)}, \tag{11.13}$$

where $s_0^{(1)}$ is the 1st component of vector s_0. The solution of (11.13) has the form $s_1(t) = q_1^* + (q_1^* - s_0^{(1)})e^{-(\alpha+\beta)t}$, where $q_1^* = \beta(\alpha+\beta)^{-1}$. Correspondingly, $s_2^*(t) = q_2^* + (q_2^* - s_0^{(2)})e^{-(\alpha+\beta)t}$, where $q_2^* = \alpha(\alpha+\beta)^{-1}$, and $s(t) \to q^* = (q_1^*, q_2^*)$.

EXAMPLE 11.2. Now consider the case of a more general system but without feedback. Suppose that for any i, j, $i \neq j$, $\lambda_{ij}(\bar{q}) \equiv \lambda_{ij}$. This means that there is no influence between the individual systems and a super-system. Consider an MP $y(\cdot)$ with state space $\{1, \dots, r\}$ and transition rates λ_{ij}, $i \neq j$, and assume that this process is irreducible. Then the solution of system (11.2) where function $\widetilde{b}(\bar{q})$ is defined by relation (11.4), is a stationary probability distribution π^* for $y(\cdot)$. Correspondingly, $s(t)$ is a vector of transition probabilities for $y(\cdot)$ in interval $[0, t]$ with initial distribution s_0. Therefore, $s(t) \to \pi^*$ as $t \to \infty$.

EXAMPLE 11.3. Let us consider the case of linear influence between different subsystems. Assume that $\lambda_{ij}(\bar{q}) = q_i\lambda_{ij}$, $i, j = \overline{1, r}$, $i \neq j$. In that case system (11.2) has the form:

$$\lambda_i q_i^2 = \sum_{k \neq i} q_k^2 \lambda_{kj}, \quad i = \overline{1, r}, \tag{11.14}$$

where $\lambda_i = \sum_{k \neq i} \lambda_{ik}$. It can easily be verified that if an MP $y(\cdot)$ defined above in Example 11.2 is irreducible, then the solution of (11.14) is unique and has a form $q_i^* = \sqrt{\pi_i}$, $i = \overline{1, r}$, where $\pi^* = (\pi_1, \dots, \pi_r)$ is a stationary distribution for $y(\cdot)$.

Similar results can be obtained for the systems in discrete time.

11.2. Averaging principle and diffusion approximation for dynamic systems with stochastic perturbations

In the theory of storage processes, in various models describing physical and biological processes, etc. different processes are developing as the solutions of deterministic differential equations with stochastic perturbations. In this part we consider two cases. The first case deals with the perturbations at the times of a recurrent flow, and the second one deals with the semi-Markov perturbations. The presentation follows [ANI 95b].

11.2.1. *Recurrent perturbations*

Assume that $f(\alpha)$, $\alpha \in \mathcal{R}^r$, is a given deterministic function with values in $\mathcal{R}^r$, and the trajectories of a dynamic system $\zeta_n(t)$ with stochastic perturbations for any $n > 0$ are described in the following way: $\zeta_n(0) = \zeta_0$,

$$\begin{aligned} \mathrm{d}\zeta_n(t) &= f(\zeta_n(t))\,\mathrm{d}t \quad \text{as } t \in \big(t_{nk}, t_{nk+1}\big), \\ \zeta_n\big(t_{nk+1}+0\big) &= \zeta_n\big(t_{nk+1}-0\big) + \frac{1}{n}\delta_k\big(\zeta_n\big(t_{nk+1}-0\big)\big), \end{aligned} \tag{11.15}$$

where $\{t_{nk}, k \geq 0\}$ is a recurrent flow such that $t_{nk+1} - t_{nk} = \frac{1}{n}\tau_k$, $\{\tau_k, k \geq 0\}$ is a sequence of iidrv and $\{\delta_k(\alpha),\ \alpha \in R^r\}$, $k \geq 0$, are independent of $\{\tau_l, l \geq 0\}$ jointly independent in index k families of random variables with distributions not depending on index k.

THEOREM 11.3. *Suppose that* $\mathbf{E}\tau_1^2 < \infty$*, for any* $L > 0$*,*

$$\sup_{|\alpha|\leq L} \mathbf{E}\big|\delta_1(\alpha)\big|^2 < \infty, \tag{11.16}$$

functions $f(\alpha)$ *and* $g(\alpha) = \mathbf{E}\delta_1(\alpha)$ *are continuously differentiable and the solution of the equation*

$$\mathrm{d}s(t) = \big(f\big(s(t)\big) + m^{-1}g\big(s(t)\big)\big)\mathrm{d}t, \quad s(0) = \zeta_0, \tag{11.17}$$

where $m = \mathbf{E}\tau_1$*, exists and is unique in an interval* $[0, T]$*. Then*

$$\sup_{t\leq T} \big|\zeta_n(t) - s(t)\big| \xrightarrow{\mathrm{P}} 0. \tag{11.18}$$

THEOREM 11.4. *Let the conditions of Theorem 11.3 hold where* $\sqrt{n}(\zeta_n(0) - \zeta_0) \xRightarrow{\mathrm{w}} \gamma_0$*. Assume that for any* $L > 0$*,*

$$\lim_{N\to\infty} \sup_{|\alpha|\leq L} \mathbf{E}\big|\delta_1(\alpha)\big|^2 \chi\big(\big|\delta_1(\alpha)\big| > N\big) = 0,$$

and the function $\sigma^2(\alpha) = \mathbf{E}(\delta(\alpha) - g(\alpha))(\delta(\alpha) - g(\alpha))^*$ satisfies the local Lipschitz condition.

Then the sequence of processes $\kappa_n(t) = \sqrt{n}(\zeta_n(t) - s(t))$ J-converges in the space $\mathcal{D}_T^r$ to a diffusion process satisfying the following stochastic differential equation: $\kappa(0) = \gamma_0$, and

$$\mathrm{d}\kappa(t) = \big(f'\big(s(t)\big) + m^{-1}g'\big(s(t)\big)\big)\kappa(t)\mathrm{d}t + m^{-1/2}D\big(s(t)\big)\,\mathrm{d}w(t), \tag{11.19}$$

where $D^2(\alpha) = m^{-2}b^2 g(\alpha)g(\alpha)^* + \sigma^2(\alpha)$, $b^2 = \mathbf{Var}\tau_1$.

Proof. Denote by $v(t,\alpha)$, $t \geq 0$, a solution of the equation

$$\mathrm{d}v(t,\alpha) = f\big(v(t,\alpha)\big)\,\mathrm{d}t, \quad v(0,\alpha) = \alpha,$$

and put $\xi_{nk}(\alpha) = v(\tau_k/n,\alpha) - \alpha + \delta_k(v(\tau_k/n,\alpha))/n$, $k \geq 0$. The family $\{(\xi_{nk}(\alpha), \tau_k/n),\ \alpha \in R^r\}$, $k \geq 0$, defines an RPSM $S_n(t)$, $t \geq 0$, according to relations (4.13) where in our case variables τ_{nk} do not depend on argument α. Firstly, we suppose that

$$\sup_\alpha \big|f(\alpha)\big| \leq C < \infty.$$

Then $|v(u,\alpha) - \alpha| \leq Cu$. According to Theorem 4.8, section 4.6, to prove that in our case

$$\sup_{t\leq T}\big|\zeta_n(t) - S_n(t)\big| \xrightarrow{\mathrm{P}} 0, \tag{11.20}$$

it is sufficient to show that

$$n\mathbf{P}\big\{\tau_1 > n\varepsilon\big\} \longrightarrow 0, \quad n\sup_\alpha \mathbf{P}\big\{\big|\delta_1(\alpha)\big| > n\epsilon\big\} \longrightarrow 0. \tag{11.21}$$

However, relations (11.21) follow from conditions (11.16) and $\mathbf{E}\tau_1^2 < \infty$ in the same way as it was shown in the proof of Theorem 4.4 (see the lines following relation (4.45), section 4.4, or relations (4.85), (4.86), section 4.4).

Now we have to verify the conditions of Theorems 4.3 and 4.4. In our case $m_n(\alpha) \equiv n^{-1}m$, and it is easy to check that $b_n(\alpha) = n^{-1}(f(\alpha)m + g(\alpha)) + O(n^{-2})$, and

$$q_n(\alpha,z) \longrightarrow f'(\alpha) + m^{-1}g'(\alpha), \quad n^2\mathbf{E}\rho_n(\alpha)\rho_n(\alpha)^* \to D(\alpha)^2.$$

Since the solution of equation (11.17) is bounded in the interval $[0,T]$, it is sufficient to check all conditions in each bounded region $|\alpha| \leq N$. This observation together with the results of Theorems 4.3 and 4.4 completes the proof of Theorems 11.3 and 11.4. □

11.2.2. *Semi-Markov perturbations*

Let us now consider a more general model. Suppose that $x_n(t)$ at each $n > 0$ is a fast SMP in the space X which is defined by the families of transition probabilities $p(x, A)$, $x \in X$, $A \in \mathcal{B}_X$ and hitting times $\{\tau_n(x),\ x \in X\}$, where $\tau_n(x) = n^{-1}\tau(x)$. Also let $\{\delta_k(x, \alpha),\ x \in X,\ \alpha \in \mathcal{R}^r\}$, $k \geq 0$, be the jointly independent families of random variables with values in R^r and distributions not depending on k.

Denote by $0 = t_{n0} < t_{n1} < \cdots$ the sequential times of jumps of process $x_n(t)$ and put $x_{nk} = x_n(t_{nk}+0)$. Then process $\zeta_n(t)$, $t \geq 0$, which stands for a dynamic system with semi-Markov perturbations, is determined in the following way: $\zeta_n(0) = \zeta_0$,

$$\mathrm{d}\zeta_n(t) = f(x_{nk}, \zeta_n(t))\,\mathrm{d}t \quad \text{as } t \in (t_{nk}, t_{nk+1}),$$

$$\zeta_n\big(t_{nk+1} + 0\big) = \zeta_n\big(t_{nk+1} - 0\big) + \frac{1}{n}\delta_k\big(x_{nk}, \zeta_n\big(t_{nk+1} - 0\big)\big).$$

THEOREM 11.5. *Suppose that* $\sup_x \mathbf{E}\tau^2(x) < \infty$*, for any* $L > 0$*,*

$$\sup_{|\alpha| \leq L} \sup_x \mathbf{E}\big|\delta_1(x, \alpha)\big|^2 < \infty,$$

and in any region $|\alpha| \leq L$ *functions* $f(x, \alpha)$ *and* $g(x, \alpha) = \mathbf{E}\delta_1(x, \alpha)$ *are bounded and continuously differentiable in* α *uniformly in* $x \in X$*. Assume that an MP with transition probabilities* $p(x, A)$ *is uniformly ergodic with stationary measure* $\pi(A)$*,* $A \in \mathcal{B}_X$*, and the solution of the equation*

$$\mathrm{d}s(t) = m^{-1}\big(f\big(s(t)\big) + g\big(s(t)\big)\big)\mathrm{d}t, \quad s(0) = \zeta_0 \tag{11.22}$$

exists and is unique in an interval $[0, T]$*, where*

$$m = \int_X m(x)\pi(\mathrm{d}x), \quad m(x) = \mathbf{E}\tau(x),$$

$$f(\alpha) = \int_X m(x) f(x, \alpha)\pi(\mathrm{d}x), \quad g(\alpha) = \int_X g(x, \alpha)\pi(\mathrm{d}x).$$

Then relation (11.18) is true.

Proof. Let $\{\tau_k(x),\ x \in X\}$, $k \geq 0$, be the jointly independent families of random variables with the same distribution as the generic variable $\tau(x)$. Denote by $v(t, x, \alpha)$, $t \geq 0$, a solution of the equation

$$\mathrm{d}v(t, x, \alpha) = f\big(x, v(t, x, \alpha)\big)\,\mathrm{d}t, \quad v(0, x, \alpha) = \alpha.$$

As was noted in the proof of Theorem 11.3, without loss of generality we may assume that

$$\sup_{x,\alpha} \big| f(x,\alpha) \big| < C. \tag{11.23}$$

Put

$$\xi_{nk}(x,\alpha) = v\big(\tau_k(x)/n, x, \alpha\big) - \alpha + \delta_k\big(x, v\big(\tau_k(x)/n, x, \alpha\big)\big)/n,$$
$$\tau_{nk}(x,\alpha) = \tau_k(x)/n.$$

The families $\{(\xi_{nk}(x,\alpha), \tau_{nk}(x,\alpha)),\ \alpha \in R^r\}$, $k \geq 0$, define RPSM $S_n(t)$, $t \geq 0$. Using Theorem 4.3, section 4.3 and Theorem 4.8, section 4.6 we see that our conditions imply relation (11.20). Denote now

$$\widehat{\xi}_{nk}(x,\alpha) = \frac{1}{n}\tau_k(x) f(x,\alpha) + \frac{1}{n}\delta_k(x,\alpha). \tag{11.24}$$

According to relation (11.23) and local Lipschitz conditions for functions $f(x,\alpha)$ and $g(x,\alpha)$ (that follow from the continuously differentiability in α) we can prove that

$$n\big|\mathbf{E}\big(\xi_{nk}(x,\alpha) - \widehat{\xi}_{nk}(x,\alpha)\big)\big| \leq \frac{1}{n}C,$$
$$n^2\big|\mathbf{E}\big(\xi_{nk}(x,\alpha)\xi_{nk}(x,\alpha)^* - \widehat{\xi}_{nk}(x,\alpha)\widehat{\xi}_{nk}(x,\alpha)^*\big)\big| \leq \frac{1}{n}C.$$

Thus, it is sufficient to check the conditions of Theorem 11.5 for the variables $\widehat{\xi}_{nk}(x,\alpha)$ and $\tau_{nk}(x,\alpha) = \tau_k(x)/n$. After some algebra we obtain that

$$b(x,\alpha) = m(x)f(x,\alpha) + g(x,\alpha), \quad b(\alpha) = f(\alpha) + g(\alpha),$$
$$m(\alpha) \equiv m, \quad \widetilde{b}(\alpha) = m^{-1}b(\alpha),$$

and the results of our theorem follow from Theorem 4.5, section 4.4. □

Let us prove the diffusion approximation. Denote

$$\begin{aligned}
\rho(x,\alpha) &= \big(f(x,\alpha) - \tilde{b}(\alpha)\big)\big(\tau(x) - m(x)\big) + \delta_1(x,\alpha) - g(x,\alpha),\\
c^2(x) &= \mathbf{Var}\tau(x),\\
\sigma^2(x,\alpha) &= \mathbf{E}\big(\delta_1(x,\alpha) - g(x,\alpha)\big)\big(\delta_1(x,\alpha) - g(x,\alpha)\big)^*,\\
D^2(x,\alpha) &= \big(f(x,\alpha) - \tilde{b}(\alpha)\big)\big(f(x,\alpha) - \tilde{b}(\alpha)\big)^* c^2(x) + \sigma^2(x,\alpha),\\
D^2(\alpha) &= \int_X D^2(x,\alpha)\pi(\mathrm{d}x),
\end{aligned}$$

$$\gamma(x,\alpha) = g(x,\alpha) - m^{-1}m(x)g(\alpha) + m(x)\big(f(x,\alpha) - m^{-1}f(\alpha)\big),$$
$$B_1^2(\alpha) = \int_X \gamma(x,\alpha)\gamma(x,\alpha)^*\pi(\mathrm{d}x),$$
$$B_2^2(\alpha) = \sum_{k\geq 1}\mathbf{E}\gamma\big(x_0,\alpha\big)\gamma\big(x_k,\alpha\big)^*,$$

where $\mathbf{P}\{x_0 \in A\} = \pi(A)$, $A \in \mathcal{B}_X$, x_k, $k \geq 0$, is an MP in X with one-step transition probabilities $p(x,A)$, and put $B^2(\alpha) = B_1^2(\alpha) + B_2^2(\alpha) + B_2^2(\alpha)^*$.

THEOREM 11.6. *Let the conditions of Theorem 11.5 hold and*

$$\lim_{L\to\infty}\sup_x \mathbf{E}\tau(x)^2\chi\big(\tau(x) > L\big) = 0.$$

Assume that for any $N > 0$ the following conditions hold:

$$\lim_{L\to\infty}\sup_{|\alpha|\leq N}\sup_x\Big\{\mathbf{E}\big|\delta_1(x,\alpha)\big|^2\chi\big(\big|\delta_1(x,\alpha)\big| > L\big)\Big\} = 0,$$

and for all $x \in X$ as $\max(|a_1|,|a_2|) \leq N$,

$$\big|D^2\big(x,\alpha_1\big) - D^2\big(x,\alpha_2\big)\big| \leq C_N\big|\alpha_1 - \alpha_2\big|$$

and $\sqrt{n}(\zeta_n(0) - \zeta_0) \xrightarrow{\mathrm{w}} \kappa_0$.

Then the sequence of processes $\kappa_n(t) = \sqrt{n}(\zeta_n(t) - s(t))$ J-converges in the space $\mathcal{D}_T^r$ to the diffusion process $\kappa(t)$ satisfying the following stochastic differential equation:

$$\mathrm{d}\kappa(t) = m^{-1}\big(f'\big(s(t)\big) + g'\big(s(t)\big)\big)\kappa(t)\mathrm{d}t$$
$$+ m^{-1/2}\big(D^2\big(s(t)\big) + B^2\big(s(t)\big)\big)^{1/2}\mathrm{d}w(t), \quad \kappa(0) = \kappa_0,$$

Proof. It can be verified that under the assumptions of Theorem 11.6,

$$\sqrt{n}\sup_{t\leq T}\big|\zeta_n(t) - S_n(t)\big| \xrightarrow{\mathrm{P}} 0.$$

Furthermore, similar to the proof of Theorem 11.5 we can see that the behavior of process $\kappa_n(t)$ is asymptotically equivalent to the behavior of process $\sqrt{n}(\widehat{S}_n(t) - s(t))$ where RPSM $\widehat{S}_n(t)$ is constructed by the families $\{\xi_{nk}(x,\alpha), n^{-1}\tau_k(x)\}$ (see (11.24)). Thus, our result follows from Theorem 4.6, section 4.4. □

Note that the dynamical systems with fast Markov switching were independently studied using another technique in [TSA 93a, TSA 93b, KOR 05].

11.3. Random movements

Let us study AP and DA for a random movement with semi-Markov switching. A simple case is described in section 1.3.2, Chapter 1. Consider a more general case when the velocity of movement may depend on the current value of the trajectory. Some results in this direction are published in [ANI 99].

Let $\{v(i,\alpha),\ \alpha \in \mathcal{R}^r\}$, $i = 1, 2, \dots, d$, be a family of continuous, with respect to α, vector-valued functions in $\mathcal{R}^r$, and $x_n(t)$, $t \geq 0$, be an SMP with a finite number of states $X = \{1, 2, \dots, d\}$. We put $\zeta_k(t,i,\alpha) = tv(i,\alpha)$, $t \geq 0$, $i = \overline{1,d}$. Denote by $0 = t_{n0} < t_{n1} < \dots$ the times of sequential jumps for $x_n(t)$ and introduce the embedded MP $x_{nk} = x(t_{nk}+0)$, $k \geq 0$. Let us define a random movement in R^r with semi-Markov switching as a PSMS $(x_n(t), \zeta_n(t))$, $t \geq 0$, constructed by the family of processes $\{\zeta_k(t,i,\alpha),\ t \geq 0,\ \overline{1,r}\}$ and switching times t_{nk} in the following way. Denote $\nu_n(t) = \max\{k : k \geq 0,\ t_{nk} < t\}$,

$$\zeta_{n0} = \zeta_n(0), \quad \zeta_{n,k+1} = \zeta_{nk} + \big(t_{n,k+1} - t_{nk}\big)v\big(x_{nk}, \zeta_{nk}\big), \quad k \geq 0.$$

Then a position of movement $\zeta_n(t)$ at time t can be represented as:

$$\zeta_n(t) = \zeta_n(0) + \sum_{k=0}^{\nu_n(t)-1} \big(t_{n,k+1} - t_{nk}\big)v\big(x_{nk}, \zeta_{nk}\big) + \big(t - t_{n,\nu(t)}\big)v\big(x_{n,\nu(t)}, \zeta_{n,\nu(t)}\big).$$

Assume that SMP $x_n(t)$ has fast switching with a sojourn time in state i of the form $\tau_n(i) = \tau(i)/n$ and study two cases: the embedded MP is irreducible, or it admits an asymptotic aggregation of the state space.

11.3.1. *Ergodic case*

Suppose that the embedded MP $x_{nk} = x_k$ does not depend on index n and is irreducible with stationary distribution π_i, $i \in X$. Assume that the 2nd moments of sojourn times exist and denote $\mathbf{E}\tau(i) = m(i)$, $\mathbf{Var}\,\tau(i) = \sigma^2(i)$, $i = \overline{1,d}$. Put

$$m = \sum_{i=1}^{d} m(i)\pi_i > 0, \quad b(\alpha) = \sum_{i=1}^{d} v(i,\alpha)m(i)\pi_i,$$

$$B_2^2(\alpha) = \sum_{k\geq 1} \mathbf{E}m\big(x_0\big)m\big(x_k\big)\big(v\big(x_0,\alpha\big) - m^{-1}b(\alpha)\big)\big(v\big(x_k,\alpha\big) - m^{-1}b(\alpha)\big)^*,$$

$$B^2(\alpha) = \sum_{i=i}^{d} \pi_i m^2(i)\big(v(i,\alpha) - m^{-1}b(\alpha)\big)\big(v(i,\alpha) - m^{-1}b(\alpha)\big)^* + B_2^2(\alpha) + B_2^2(\alpha)^*,$$

$$D^2(\alpha) = \sum_{i=i}^{d} \pi_i\big(v(i,\alpha) - m^{-1}b(\alpha)\big)\big(v(i,\alpha) - m^{-1}b(\alpha)\big)^*\sigma^2(i),$$

where it is assumed that in the calculations of $B_2^2(\alpha)$, process x_k is in the stationary regime, i.e., $\mathbf{P}(x_0 = i) = \pi_i$, $i = \overline{1, d}$.

STATEMENT 11.1. *Let the functions $v(i, \alpha)$, $i \in X$, be locally Lipschitz with respect to α, have no more than linear growth and $\zeta_n(0) \xrightarrow{\mathrm{P}} s_0$. Then for any $T > 0$*

$$\sup_{0 \le t \le T} \left|\zeta_n(t) - s(t)\right| \xrightarrow{\mathrm{P}} 0,$$

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

If, in addition, $\sqrt{n}(\zeta_n(0) - s_0) \stackrel{\mathrm{w}}{\Longrightarrow} \gamma_0$, then the sequence $\sqrt{n}(\zeta_n(t) - s(t))$ J-converges to the diffusion process satisfying the following stochastic differential equation: $\gamma(0) = \gamma_0$,

$$\mathrm{d}\gamma(t) = m^{-1} b'\big(s(t)\big)\gamma(t) + m^{-\frac{1}{2}} \Big(D^2\big(s(t)\big) + B^2\big(s(t)\big)\Big)^{\frac{1}{2}} \mathrm{d}w(t), \tag{11.25}$$

where $w(t)$ is a standard Wiener process in $\mathcal{R}^r$, and the solution of (11.25) exists and is unique.

The proof follows straightforwardly from Theorems 4.5, 4.6, section 4.4.

11.3.2. *Case of the asymptotic aggregation of state space*

Suppose that the embedded MP also depends on parameter n in such a way that conditions (8.27), (8.28) hold. For simplicity assume that each region X_y forms in a limit one essential class. Let $x_k^{(y)}$ be an auxiliary MP in X_y with limiting transition probabilities $P^{(y)} = \|p_0(i,j)\|$, $i, j \in X_y$. Denote by $\pi^{(y)}(i)$, $i \in X_y$, its stationary distribution. Put

$$A_{ys} = \sum_{i \in X_y} \pi^{(y)}(i) \sum_{j \in X_s} b_{ij}, \quad y, s \in Y,\ y \neq s.$$

For any $y \in Y$ denote

$$\widehat{m}(y) = \sum_{i \in X_y} m(i)\pi^{(y)}(i), \quad \widehat{b}(y, \alpha) = \sum_{i \in X_y} v(i, \alpha) m(i) \pi^{(y)}(i). \tag{11.26}$$

Let $y(t, y_0)$ be an MP with values in Y, transition rates $A_{ys}/\widehat{m}(y)$, $y, s \in Y$, $y \neq s$, and the initial state y_0. Denote also by $z(t, y_0, s_0)$ a differential equation solution switched by process $y(t, y_0)$: $z(0, y_0, s_0) = s_0$,

$$\mathrm{d}z\big(t, y_0, s_0\big) = \widehat{m}\big(y(t, y_0)\big)^{-1} \widehat{b}\big(y(t, y_0), z\big(t, y_0, s_0\big)\big)\mathrm{d}t.$$

STATEMENT 11.2. *Let the functions* $v(i,\alpha)$, $i \in X$, *be locally Lipschitz with respect to* α, *have no more than linear growth and* $\zeta_n(0) \xrightarrow{P} s_0$. *Let in addition,* $\mathbf{P}(x_n(0) \in X_{y_0}) \to 1$ *as* $n \to \infty$, $\widehat{m}(y) > 0$, $y \in Y$, *the functions* $\widehat{b}(y,\alpha)$ *be locally Lipschitz and have no more than linear growth with respect to* α. *Then the sequence* $(K(x_n(t)), \zeta_n(t))$ *J-converges in any interval* $[0,T]$ *to the process* $(y(t,y_0), z(t,y_0,s_0))$.

The proof follows from Statement 11.1 and Theorem 8.3, section 8.3.

Now consider the case when in (11.26) $\widehat{b}(y,\alpha) \equiv 0$. In this case we can prove a DA. Note that as in this case the trajectory of $\zeta_n(\cdot)$ converges to zero, we need to calculate the limiting expressions only in point $\alpha = 0$. For any region X_y denote

$$\widehat{D}^2(y) = \sum_{i\in X_y} v(i,0)v(i,0)^*\sigma^2(i)\pi^{(y)}(i),$$

$$\widehat{B}_1^2(y) = \sum_{i\in X_y} m^2(i)v(i,0)v(i,0)^*\pi^{(y)}(i),$$

$$\widehat{B}_2^2(y) = \sum_{k\geq 1} \mathbf{E}m\big(x_0^{(y)}\big)m\big(x_k^{(y)}\big)v\big(x_0^{(y)},0\big)v\big(x_k^{(y)},0\big)^*,$$

where it is assumed that in calculations of $\widehat{B}_2^2(y)$ process $x_0^{(y)}$ is in stationary conditions, i.e., $\mathbf{P}(x_0^{(y)} = i) = \pi^{(y)}(i)$, $i \in X_y$. Put

$$\widehat{C}^2(y) = \widehat{D}^2(y) + \widehat{B}_1^2(y) + \widehat{B}_2^2(y) + \widehat{B}_2^2(y)^*.$$

STATEMENT 11.3. *If the conditions of Statement 11.2 hold, then the sequence* $(K(x_n(t)), \sqrt{n}\zeta_n(t))$ *J-converges to process* $(y(t,y_0), \gamma(t,y_0,s_0))$, *where*

$$\gamma\big(t,y_0,s_0\big) = \int_0^t \widehat{m}\big(y\big(t,y_0\big)\big)^{-1/2}\widehat{C}\big(y\big(t,y_0\big)\big)\,\mathrm{d}w(t).$$

This is a Wiener process with Markov switching.

The proof follows from Theorem 8.3, section 8.3 on the convergence of SP with rare switching and DA for RPSM in an asymptotically aggregated environment, Theorem 8.15, section 8.8.

11.4. Bibliography

[ANI 95a] ANISIMOV V., "Switching processes: averaging principle, diffusion approximation and applications", *Acta Applicandae Mathematicae*, vol. 40, p. 95–141, 1995.

[ANI 95b] ANISIMOV V., "Diffusion approximation in switching stochastic models and applications", in *Exploring Stochastic Laws*, SKOROKHOD A. V. and BOROVSKIKH YU. V. (eds.), p. 13–40, VSP, The Netherlands, 1995.

[ANI 99] ANISIMOV V., "Diffusion approximation for processes with seni-Markov switches and applications in queueing models", in *Semi-Markov Models and Applications*, JANSSEN J. and LIMNIOS M. (eds.), p. 77–101, Kluwer, Dordrecht, 1999.

[KOR 05] KOROLYUK V. and LIMNIOS N., *Stochastic Systems in Merging Phase Space*, World Scientific, Singapore, 2005.

[TSA 93a] TSARKOV J., "Averaging and stability of impulse systems with rapid Markov switchings", *Proc. Latv. Prob. Sem.*, vol. 2, p. 49–63, 1993.

[TSA 93b] TSARKOV J., "Limit theorems for impulse systems with rapid Markov switchings", *Proc. Latv. Prob. Sem.*, vol. 2, p. 74–96, 1993.

[WEI 88] WEIDLICH W. and HAAG G., *Interregional Migration-Dynamic Theory and Comparative Analysis*, Springer, Berlin, Heidelberg, New York, 1988.

[WEI 91] WEIDLICH W., "Physics and social science - the approach of synergetics", *Physics Reports*, vol. 204, p. 1–163, 1991.

Chapter 12

Simulation Examples

In this chapter we illustrate some theoretical results of the book using simulations in language R. R is a very convenient programming language and is freely distributed. The recurrent structure of SP is convenient for writing simulation codes. Simulation can be used for the illustration of asymptotic results and also for testing the models and providing other research experiments. R-codes are provided for all examples and can be used after corresponding modifications for simulation of similar models.

12.1. Simulation of recurrent sequences

Let us illustrate the results of section 4.2 for the averaging principle for switching recurrent sequences defined by (4.1). Consider the interval $[t_0, t_1]$, take a small step $h = 1/n$, consider a grid x_{hk}: $x_{hk} = t_0 + kh$, $k = 0, \dots, [(t_1 - t_0)/h]$, and represent (4.1) in the following form: $\xi_{0h} = y_0$,

$$\xi_{h,k+1} = \xi_{hk} + a\big(x_{hk}, \xi_{hk}\big)h + \beta_{hk}, \quad k = 0, \dots, \big[\big(t_1 - t_0\big)/h\big], \tag{12.1}$$

where $a(x, y)$ is a given function. Consider an example: $a(t, y) = f_1(t)y + f_2(t)$, where $f_1(t) = a/t$, $f_2(t) = b/t^g$, and $a + g \neq 1$. Then a solution of the equation

$$\mathrm{d}y(t) = a\big(t, y(t)\big)\mathrm{d}t, \quad y\big(t_0\big) = y_0, \tag{12.2}$$

has the form

$$y(t) = y_0\left(\frac{t}{t_0}\right)^a - \frac{bt^a}{a + g - 1}\big(t^{1-a-g} - t_0^{1-a-g}\big).$$

Let us take $a = -0.5$; $b = 1.5$; $g = 2$; $t_0 = 1$; $t_1 = 10$; $y_0 = 1$. Assume that the random sequence β_{hk} in (12.1) has the form: $\beta_{hk} = h(U_k(0,1) - 0.5)$, where $U_k(0,1)$ is the sequence of iidrv with uniform $(0,1)$ distribution. In this case $\mathbf{E}\beta_{hk} = 0$ and as $h \to 0$, relation (4.4) is satisfied for any $T > 0$ with $n = [1/h]$.

In Figure 12.1 a solid line denotes a solution $y(t)$ of equation (12.2), a dashed-dotted line is a sample path of a simulated trajectory of ξ_{kh} for $h = 0.02$ and a dashed line is a sample path for $h = 0.001$. Clearly for much smaller h the trajectory is much closer to $y(t)$ and this plot illustrates AP.

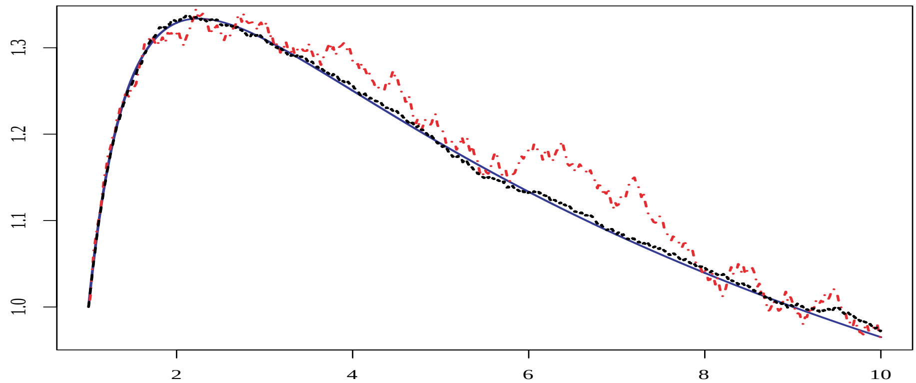

Figure 12.1. *Graph of the solution of the differential equation and two simulated trajectories of a recurrent sequence for* $h = 0.02$ *and* $h = 0.001$

R-code for simulation can be written as follows:

```
a=-0.5; b=1.5; g=2
f1 = function(x) a/x
f2 = function(x) b/x^g
f3 = function(x,y) f1(x)*y + f2(x) # function a(x,y)

# exact formula for the solution y(x)
ff4=function(x) {
y0*(x/t0)^a -(b/(a+g-1))*x^a*(x^(1-a-g)-t0^(1-a-g))
}
# Graph of the solution y(x) in interval [t_0,t_1]
hh=0.01
xx=seq(t0,t1,hh)
plot(xx,ff4(xx), type="l", col="blue",
lty=1, lwd=2, xlab=NA, ylab=NA, ylim=c(0.9,1.4))

# Simulation of the recurrent sequence
hh2=0.02;
xx2=seq(t0,t1,hh2)
kk2=length(xx2)
# Simulation of the random variables beta_k
ranxi=runif(kk2)-0.5
zz=numeric(kk2)
zz[1]=y0
for(i in 1:(kk2-1)){
zz[i+1]=zz[i]+hh2*f3(xx2[i],zz[i])+hh2*ranxi[i]
}
```

```
# Graph of the trajectory
lines(xx2, zz, type="l", lty=4, lwd=2, col="red")
```

The last part can be run for different hh2 in order to obtain different sample paths. If the solution of (12.2) cannot be found in the closed form, then we can calculate it numerically using the first order recurrent Runge-Kutta procedure:

$$y_{h,k+1} = y_{hk} + ha\big(x_{hk}, y_{hk}\big),\ k = 0, \dots, \big[\big(t_1 - t_0\big)/h\big], \quad y_{h0} = y_0, \qquad (12.3)$$

and defining the approximate solution as $y_h(t) = y_{hk}$ as $t_{hk} \le t < t_{h,k+1}$. It is well-known that the accuracy of approximation is $O(h)$, i.e.

$$\sup_{t\in[t_0,t_1]} \big|y(t) - y_h(t)\big| = O(h).$$

12.2. Simulation of recurrent point processes

Consider a simulation of a recurrent point process of the form

$$t_{k+1} = t_k + \tau\big(t_k\big), \quad k \ge 0,$$

where t_0 is given. Denote by $X(t)$ the number of events in interval $[0, t]$ starting from t_1. Consider the case where $\tau(t)$ has an exponential distribution with rate $\lambda(t) = 1 + a\sqrt{t}$, $t \ge 0$. Two simulated sample paths of $X(t)$ are shown in Figure 12.2.

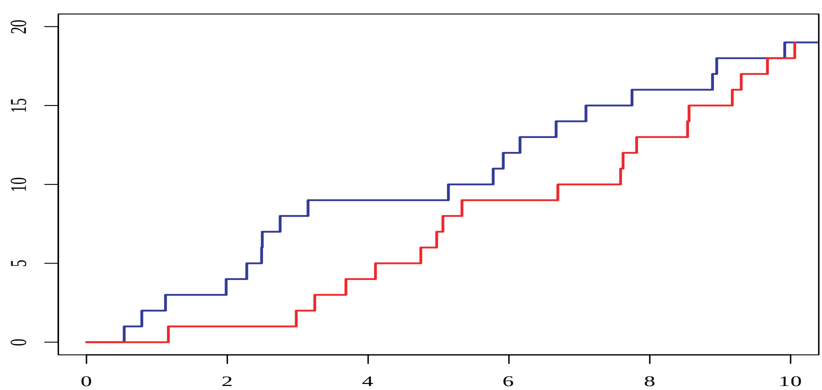

Figure 12.2. *Two sample trajectories of $X(t)$ in the interval $[0, 10]$ for $a = 0.5$*

R-code for simulation can be written as follows:

```
lla = function(x,a) 1+a*sqrt(x) ## rate function
# function for simulation of a sequence t_k, k \ge 1
frec2 = function(T,a) {
```

```
    tt=numeric()
    tt[1] = rexp(1,lla(0,a))
    k = 1
    while (tt[k] <= T) {
        tt[k+1] = tt[k] + rexp(1,lla(tt[k],a))
        k = k +1
    }
    return(tt)
}
# simulation of a sequence t_k, k \ge 0
T=10; a=0.5
ptt1=c(0,frec1(T,a))
llen1=length(ptt1)
kk1=seq(0,llen1-1,1)

# graph of X(t)
plot(ptt1,kk1, type="s", col="blue",
lty=1, lwd=2, xlab=NA, ylab=NA, xlim=c(0,T))
# second sample path
ptt2=c(0,frec1(T,a))
llen2=length(ptt2)
kk2=seq(0,llen2-1,1)
lines(ptt2,kk2, type="s", col="red", lty=1, lwd=2)
```

12.3. Simulation of RPSM

Consider a simulation of RPSM (simple case) defined in section 1.2.2. According to (1.8), (1.9) we introduce the following recurrent sequences:

$$t_0 = 0, \quad t_{h,k+1} = t_{hk} + h\tau_k\big(S_{hk}\big), \quad S_{h,k+1} = S_{hk} + h\xi_k\big(S_{hk}\big), \quad k \geq 0,$$

where the distributions of the family $\{(\tau_k(s), \xi_k(s))\}$ do not depend on index k, S_0 is given and h is a scaling coefficient which stands for $1/n$. Set $S_h(t) = S_{hk}$ as $t_{hk} \leq t < t_{h,k+1}, t \geq 0$. Consider as an example the case where $\tau(s)$ has an exponential distribution with rate $\lambda(s) = (1+s)^{-1},\ s \geq 0$, and $\xi(s)$ has a uniform distribution in interval $(c(s) - 2, c(s) + 2)$ where $c(s) = (1 + s)^{-1}$. These specific functions are chosen with the purpose of obtaining a closed-form expression of the limiting function $s(t)$ in (4.20). A sample path of process $S(t)$ is shown in Figure 12.3.

Let us illustrate the averaging principle and diffusion approximation for $S_h(t)$ at small h using Theorems 4.3, 4.4, section 4.3. In this case $m(s) = 1 + s$, $b(s) = (1 + s)^{-1}$, and as $h \to 0$, $\sup_{t\leq T} |S_h(t) - s(t)| \xrightarrow{\text{P}} 0$, where the function $s(t)$ satisfies the equation (4.20) in the form

$$\mathrm{d}s(t) = \big(1 + s(t)\big)^{-2}\mathrm{d}t, \quad s(0) = S_0,$$

with a solution $s(t) = \sqrt[3]{3(t + Q)} - 1$, where $Q = (1 + S_0)^3/3$.

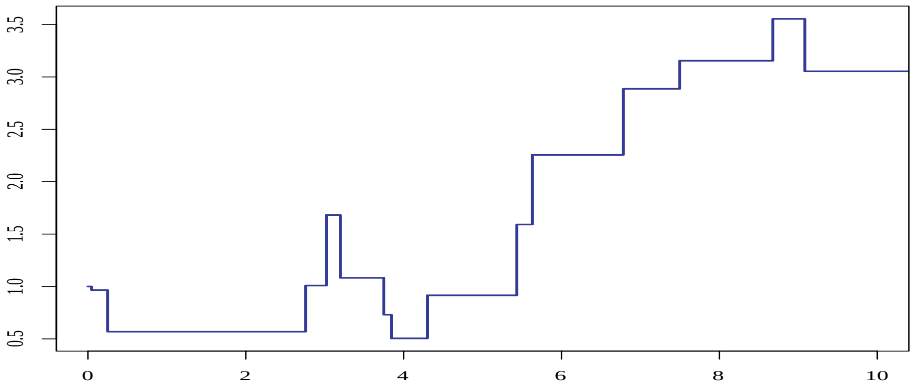

Figure 12.3. *Graph of a simulated trajectory of $S(t)$ in the interval $[0, 10]$ for $h = 1$*

Figure 12.4 illustrates AP and shows that at small h the trajectory of $S_h(t)$ is quite close to function $s(t)$ uniformly in the whole interval.

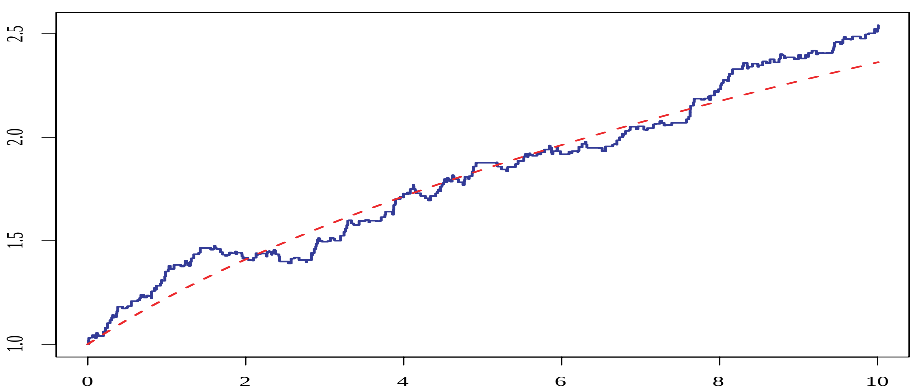

Figure 12.4. *Graphs of the simulated trajectory of $S_h(t)$ and the function $s(t)$, shown by the dashed line, in the interval $[0, 10]$ for $h = 0.01$ and $S_0 = 1$*

To illustrate DA we use Theorem 4.4 and find that $Q(s) = -2(1+s)^{-3}$, $g(s) = 0$, and $D^2(s) = \mathbf{E}\xi^2(s) = 4/3 + (1 + s)^{-2}$. Thus, equation (4.36) has the form:

$$\mathrm{d}\gamma(t) = -2\big(1 + s(t)\big)^{-3}\gamma(t)\mathrm{d}t + D\big(s(t)\big)\big(1 + s(t)\big)^{-1/2}\mathrm{d}w(t), \quad \gamma(0) = 0,$$

and therefore $\gamma(t)$ is a non-homogenous Ornstein-Uhlenbeck process. The graph in Figure 12.5 illustrates the behavior of the normalized difference $(S_h(t) - s(t))/\sqrt{h}$ which looks like a trajectory of a Brownian motion.

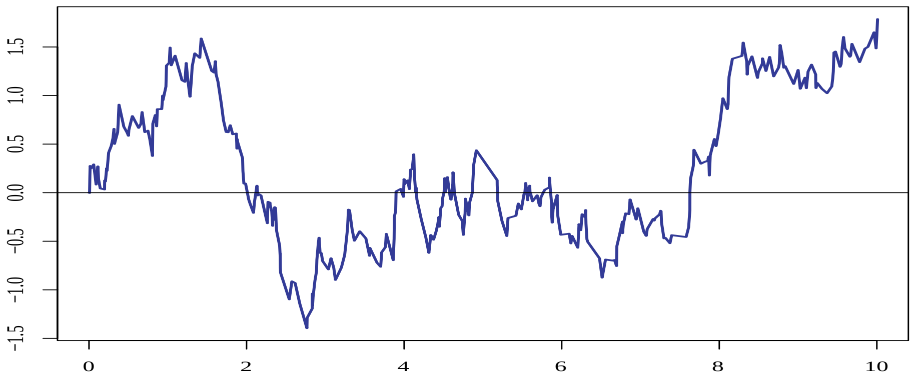

Figure 12.5. *Graph of the difference* $(S_h(t) - s(t))/\sqrt{h}$. *Illustration of DA for RPSM*

12.4. Simulation of state-dependent queueing models

Consider a queueing system $M_Q/M_Q/1/\infty$ with input rate $\lambda(q)$ and service rate $\mu(q)$ investigated in section 5.2.2. A general code for simulation can be written as below where h is a scaling factor which stands for $1/n$ and the functions $\lambda(q)$ and $\mu(q)$ should be defined. In order to be definite we take $\lambda(s) \equiv \lambda$, $\mu(s) = \mu s$, this corresponds to the system $M/M/\infty$. Then equation (5.16) implies: $s(t) = \lambda/\mu + (s_0 - \lambda/\mu)e^{-\mu t}$, $t \geq 0$.

```
funla = function(x, lam) lam  ## input rate
funmu = function(x, mu) mu*x  ## service rate
# probability of a jump up
ppx=function(x,lam,mu) funla(x,lam)/(funla(x,lam)+funmu(x,mu))

## Function for simulation of the queueing process Q(t)
funqueue <- function(T,lam,mu,Q0,h) {
tt=numeric()  ## sequence of jumps t_k
zz=numeric()  ## sequence Q_k
tt[1] =0
zz[1]=Q0
k = 1
while (tt[k] <= T) {
tt[k+1] = tt[k] + h*rexp(1,funla(zz[k],lam)+funmu(zz[k],mu))
zz[k+1]=zz[k]+h*(2*rbinom(1,1,ppx(zz[k],lam,mu))-1)
k = k +1
    }
return(c(tt,zz))
}
## Simulation of the queueing process Q(t)
Q0=5; T=10; h=1; la=1; mu=0.5
outque=funqueue(T,lam,mu,Q0,h)
len1=length(outque)
tt=outque[1:(len1/2)]
zz=outque[(len1/2+1):len1]
# Graph of Q(t)
plot(tt,zz,type="s",col="blue",lty=1,lwd=2,xlab=NA,ylab=NA,xlim=c(0,T))
```

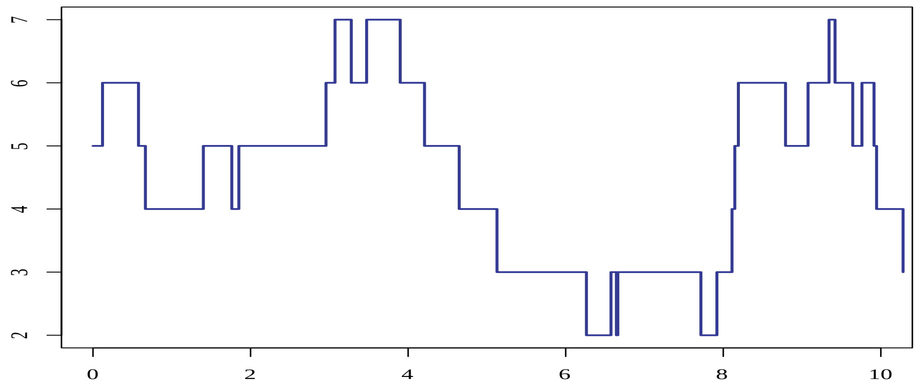

Figure 12.6. *Sample path of the process $Q(t)$ at $h = 1$, $Q_0 = 5$, $\lambda = 1$, $\mu = 0.5$*

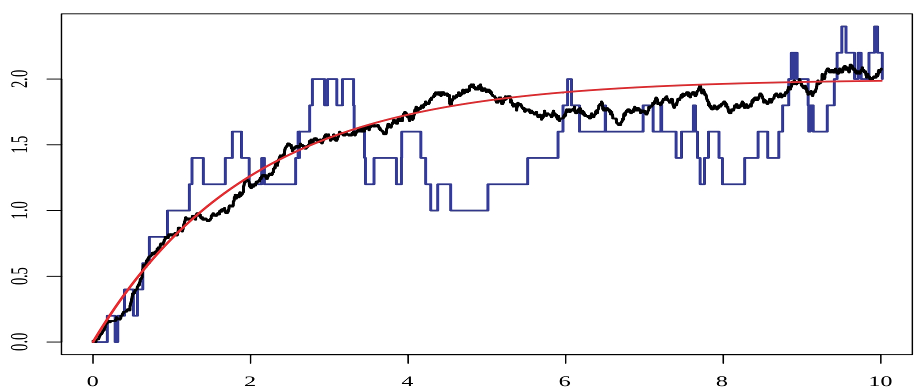

Figure 12.7. *Two sample paths of $Q(t)$ at $Q_0 = 5$, $\lambda = 1$, $\mu = 0.5$. Step-wise line corresponds to $h = 0.2$, wavy line – to $h = 0.01$. Graph of $s(t)$ is shown by a straight solid line*

A sample trajectory of the queueing process $Q(t)$ is shown in Figure 12.6. Figure 12.7 illustrates AP and DA for process $Q_h(t)$. At $h = 0.2$ we see quite a large deviation between $S_h(t)$ and $s(t)$, the deviation is much less at $h = 0.01$. R-code for simulation can be written as follows.

```
Q0=0; T=10; h=0.01; mu=0.5
## Simulation of the trajectory of Q(t)
outque=funqueue(T,lam,mu,Q0,h)
len1=length(outque)
tt=outque[1:(len1/2)]
zz=outque[(len1/2+1):len1]

# graph of Q(t)
plot(tt,zz,type="s",col="blue",lty=1,lwd=2,xlab=NA,ylab=NA,xlim=c(0,T))
```

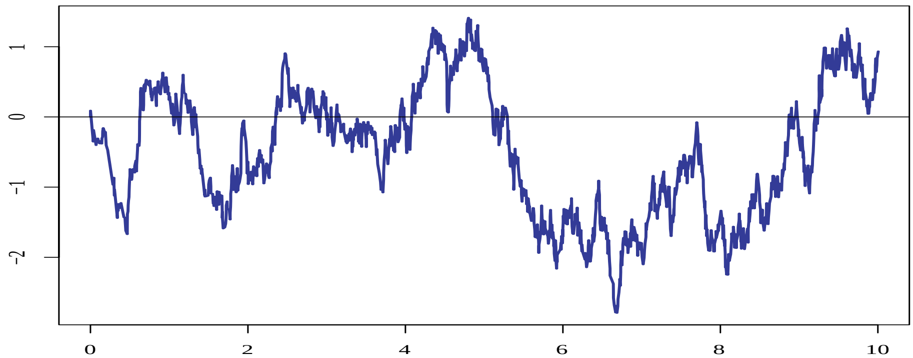

Figure 12.8. *Graph of* $\zeta_h(t) = (Q_h(t) - s(t))/\sqrt{h}$ *at* $h = 0.01$. *Illustration of a DA*

```
# function s(t)
funst =function(u,lam,mu) lam/mu+(Q0-lam/mu)*exp(-mu*u)
lines(tt,funst(tt,lam,mu), type="l", col="red", lty=2, lwd=2)
# graph of s(t)
lines(tt,funst(tt,lam,mu), type="l", col="red", lty=1, lwd=2)
```

DA is illustrated in Figure 12.8. In this case a limiting process for $\zeta_h(t)$ is described as a solution of a stochastic differential equation

$$\mathrm{d}\zeta(t) = -\mu\zeta(t)\mathrm{d}t + \big(\lambda + \mu s(t)\big)^{1/2}\mathrm{d}w(t), \quad \zeta(0) = 0, \tag{12.4}$$

which is an Ornstein-Uhlenbeck type process. R-code is given below.

```
plot(tt,(zz-funst(tt,lam,mu))/sqrt(0.01), type="s", col="blue",
lty=1, lwd=2, xlab=NA, ylab=NA, xlim=c(0,T))
abline(h=0)
```

We can also simulate a solution of equation (12.4) directly using the fact that at small h a solution is approximated by a stochastic recurrent sequence:

$$\zeta_{h,k+1} = \zeta_{hk} - \mu\zeta_{hk}h + \big(\lambda + \mu s(kh)\big)^{1/2} w_k\sqrt{h}, \quad \zeta_{h0} = 0,$$

where w_k is a sequence of iidrv with normal $\mathcal{N}(0,1)$ distribution. R-code for the simulation of the solution in the interval $[t_0, T]$ can be written as below and Figure 12.9 shows one simulated sample path which looks similar to the graph in Figure 12.8.

```
T=10; lam=1; mu=1; zet0=0; t0=0
hh=0.01; K=round(T/hh)
# calculation of grid t_k
tt=numeric(K)
tt[0]=t0
for(k in 1:(K-1)){
tt[k+1]=t0+k*hh
}
```

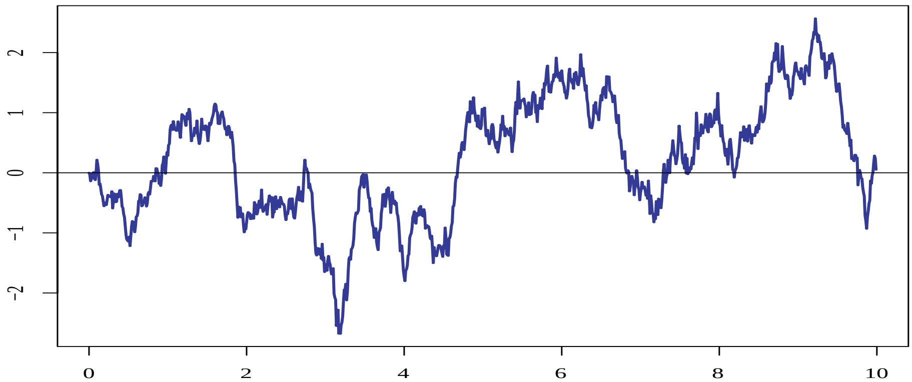

Figure 12.9. *A sample path of process $\zeta(t)$ (see (12.4))*

```
vnorm=rnorm(K)    ## simulation of a sequence w_k, k=1,..,K
# calculation of a sequence \zeta_{hk}
zet=numeric(K)
zet[0]=zet0
for(k in 1:(K-1)){
zet[k+1]=zet[k]-mu*zet[k]*hh+
(lam+mu*funst(tt[k],lam,mu))^(1/2)*vnorm[k]*sqrt(hh)
}
# graph of \zeta(t)
plot(tt,zet,type="l",col="blue",lty=1,lwd=2,xlab=NA,ylab=NA,xlim=c(0,T))
abline(h=0)
```

12.5. Simulation of the exit time from a subset of states of a Markov chain

Consider the simulation method of Markov models and let us illustrate the asymptotic exponentiality of the exit time from a subset of states proved in section 6.2 on the example of a Markov chain $Y(t)$ with 4 states $\{1, 2, 3, 4\}$. Consider a subset $X_0 = \{1, 2, 3\}$ with matrix of transition rates λ_{ij}:

$$\Lambda = \begin{pmatrix} 0 & 1 & 2 & \varepsilon \\ 1 & 0 & 1 & 2\varepsilon \\ 1 & 3 & 0 & 3\varepsilon \end{pmatrix}$$

where transition rates λ_{i4} to state 4 are small ($\lambda_{i4} = \varepsilon i$, $i = 1, 2, 3$). Let $\Omega_\varepsilon(x_0)$ be the exit time from X_0 starting from state $x_0 \in X_0$. Denote by ρ_i, $i \in X_0$, a stationary distribution of an MP with transition rates λ_{ij}, $i, j \in X_0$, $i \neq j$. It is easy to calculate that in this case the stationary distribution is $\{1/4, 1/2, 1/4\}$. Let $\beta_\varepsilon = \sum_{i \in X_0} \rho_i \lambda_{i4} = 2\varepsilon$ be the stationary exit rate. Then, according to Corollary 6.3, section 6.2, at small ε the variable $\beta_\varepsilon \Omega_\varepsilon(x_0)$ is approximated by an exponential distribution with rate 1. To illustrate this statement we first provide R-code for simulation of the exit time $\Omega_\varepsilon(x_0)$.

```
eps=0.01
matLa = matrix(nrow = 3, ncol=4) ## rate matrix \Lambda
matLa[1,]=c(0,1,2,eps)
matLa[2,]=c(1, 0, 1, 2*eps)
matLa[3,]=c(1,3,0, 3*eps)
# matrix of transition probabilities of the embedded MP
matP = matrix(nrow = 3, ncol=4)
for (i in 1:3){
matP[i,]=matLa[i,]/sum(matLa[i,])
}
# vector of the exit rates from the states 1,2,3
vecLa=numeric(3)
for(i in 1:3){
vecLa[i]=sum(matLa[i,])
}

# Function for simulation of exit time starting from state x_0
# xx[k] is the embedded MP, tt[k] is a sequence of jumps of Y(t)
x0=1
funMarkov <- function(x0) {
tt=numeric()
xx=numeric()
xx[1] =x0    ## initial state
tt[1]=0
k = 1
while (xx[k] < 4) {
tt[k+1] = tt[k] + rexp(1,vecLa[xx[k]])
xx[k+1]=c(1,2,3,4) %*% rmultinom(1,size=1,prob=matP[xx[k],])
k = k +1
    }
return(c(length(xx),tt[length(xx)]))
}
# length(xx) is the number of jumps in the subset X_0 before exit
# tt[length(xx)] - exit time
```

For the illustration of the asymptotic exponentiality of $\zeta_\varepsilon = \beta_\varepsilon \Omega_\varepsilon(x_0)$ we need to carry out many simulation runs of the exit time and use some statistical tests. Figure 12.10 shows a histogram of 10^5 simulation runs of the variable ζ_ε. The continuous line shows a graph of the function $\exp\{-x\}$, the probability density function of the exponential distribution with rate 1. As we see, the latter curve practically coincides with histogram. This result confirms the statement of Corollary 6.3, section 6.2. Similar results can be shown for the number of jumps before the exit.

R-code for simulation and statistical testing:

```
M=10^5  # number of simulation runs
matSim <- matrix(nrow = M, ncol=2)  ## output of simulation
for (i in 1:M){
matSim[i,]=funMarkov(x0)
}
# first column - simulated values of the number of jumps
# second column - simulated values of the exit time
```

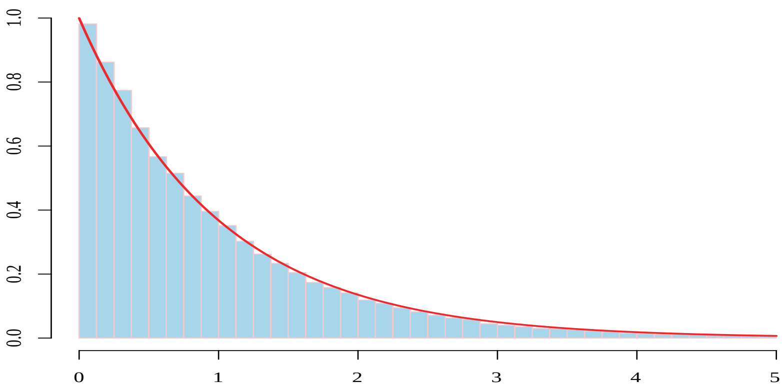

Figure 12.10. *Comparison of the histogram of the exit time and exponential probability density function with rate 1*

```
rh1=1/4; rh2=1/2; rh3=1/4;    ## stationary distribution
vecrho=c(rh1,rh2,rh3)
statrate=vecrho %*% matLa[,4]  ## stationary exit rate

# Histogram of 10^5 simulated exit times normalized by
# a stationary exit rate
hist(matSim[,2]*statrate,main=NA,xlab=NA,ylab=NA,
br=c(seq(0,5,0.125),max(matSim[,2]*statrate)+1),prob=TRUE,
xlim=c(0,5),col="lightblue", border="pink")
# graph of the exponential probability density function, rate=1
xxx=seq(0,5,0.01)
lines(xxx, exp(-xxx), type="l", col="red", lty=1, lwd=2)
```

Let us illustrate the so-called quasi-ergodic properties of S-sets. As $\varepsilon \to 0$, a subset X_0 forms an S-set and therefore as small ε, $\Omega_i(x_0)/\Omega(x_0) \approx \rho_i$, $i = 1, 2, 3$, where $\Omega_i(x_0)$ is the total time spent in state i up to the exit from X_0. This result can easily be verified by simulation. Let us take $\varepsilon = 0.0001$ and use the following function for simulation of sequences x_k and t_k up to the first exit time:

```
x0=1
funMarkov2 <- function(x0) {
tt=numeric()
xx=numeric()
xx[1] =x0
tt[1]=0
k = 1
while (xx[k] < 4) {
tt[k+1] = tt[k] + rexp(1,vecLa[xx[k]])
xx[k+1]=c(1,2,3,4)%*% rmultinom(1,size=1,prob=matP[xx[k],])
k = k +1
    }
Item=matrix(nrow = 2, ncol=length(xx))
Item[1,]=xx
Item[2,]=tt
```

```
return(Item)
}

## Simulation of the ratios \Omega_i(x_0)/\Omega(x_0)
Matout1=funMarkov2(x0)     ## simulation of sequences x_k, t_k
xxx=Matout1[1,]            ## sequence x_k
ttt=Matout1[2,]            ## sequence t_k
leng=length(xxx)           ## number of jumps before the exit
## sequence \tau_k of occupation times in states x_k
times=numeric(leng-1)
for (i in 1:(leng-1)){
times[i]=ttt[i+1]-ttt[i]
}
statdis=numeric(3)          ## ratios of \Om_i(x_0)/\Om(x_0)
for (i in 1:3){
statdis[i]=sum(times[i <= xxx & xxx <i+1])/ttt[leng]
}
```

For a particular simulation run, the length of vector xxx is 19041 and statdis $=$ $(0.248, 0.509, 0.243)$ which is very close to the exact values $(0.25, 0.5, 0.25)$. The same illustration can be performed for a stationary distribution of the embedded MP.

12.6. Aggregation of states in Markov models

Let us illustrate the results of Chapters 8 and 9 on the asymptotic aggregation of state space of Markov processes. Consider a Markov chain $Y_\varepsilon(t)$ with 4 states $\{1, 2, 3, 4\}$ and matrix of transition rates

$$\Lambda = \begin{pmatrix} 0 & 1 & 0.5\varepsilon & \varepsilon \\ 0.5 & 0 & \varepsilon & 2\varepsilon \\ 0.5\varepsilon & \varepsilon & 0 & 0.5 \\ \varepsilon & 2\varepsilon & 0.5 & 0 \end{pmatrix}.$$

As we see, there are two subsets of states $X_1 = \{1, 2\}$ and $X_2 = \{3, 4\}$ with small transition rates between subsets. $Y_\varepsilon(t)$ spends an asymptotically exponential time in each region (which was illustrated in the previous example) and then jumps to another region. This behavior is shown in Figure 12.11. R-code for simulation is given below.

```
eps=0.005
matLa <- matrix(nrow = 4, ncol=4)
matLa[1,]=c(0,1,0.5*eps,eps)
matLa[2,]=c(0.5, 0, eps, 2*eps)
matLa[3,]=c(0.5*eps,eps,0, 0.5)
matLa[4,]=c(eps,2*eps,0.5, 0)
# matrix of transition probabilities of the embedded MP
matP <- matrix(nrow = 4, ncol=4)
for (i in 1:4){
matP[i,]=matLa[i,]/sum(matLa[i,])
}
```

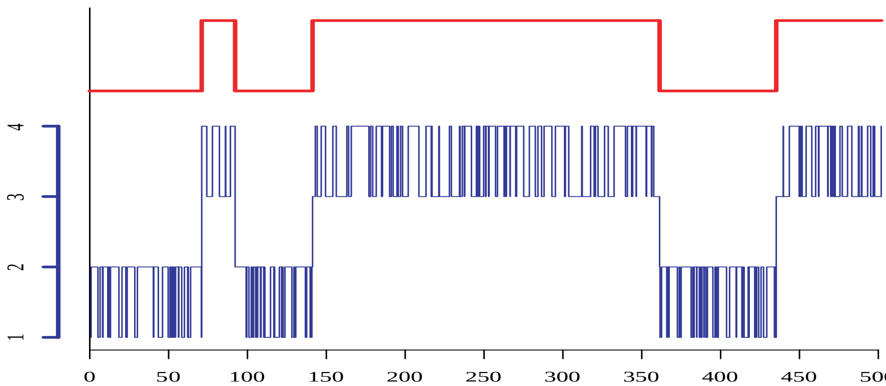

Figure 12.11. *One sample path of process $Y_\varepsilon(t)$ for $\varepsilon = 0.005$. The upper stepwise line shows the trajectory of the aggregated process $K(Y_\varepsilon(t))$. This line can be considered as a sample path of an MP with two states $\{X_1, X_2\}$ and aggregated transition rates:* $\widehat{\lambda}_{X_1,X_2} = 5\varepsilon/2$, $\widehat{\lambda}_{X_2,X_1} = 9\varepsilon/4$

```
vecLa=numeric(4)
for(i in 1:4){
vecLa[i]=sum(matLa[i,])
}
# Function for simulation of a MP in interval (0,T)
funMarkovAg <- function(T) {
tt=numeric()
xx=numeric()
xx[1] =x0
tt[1]=0
k = 1
while (tt[k] < T) {
tt[k+1] = tt[k] + rexp(1,vecLa[xx[k]])
xx[k+1]=c(1,2,3,4)%*% rmultinom(1,size=1,prob=matP[xx[k],])
k = k +1
}
Item=matrix(nrow = 2, ncol=length(xx))
Item[1,]=xx
Item[2,]=tt
return(Item)
}

## Simulation of the trajectory of Y_\e(t) in interval (0,500)
Matout1=funMarkovAg(500)
plot(Matout1[2,],Matout1[1,],type="s",col="blue",lty=1,lwd=1,
xlab=NA, ylab=NA,xlim=c(0,500),ylim=c(1,5.5),axes=FALSE)
axis(1, at=seq(0,500,by=50))
axis(2, at=seq(1,4,by=1), lwd=3, col="blue")
abline(v=0, lwd=1)
funK= function(j) ifelse(j <=2, 4.5,5.5) ## aggregation function
## adding trajectory of the aggregated process K(Y_\e(t))
lines(Matout1[2,],funK(Matout1[1,]),type="s",
col="red",lty=1,lwd=3)
```

Index

R

S

U