About some extended Erlang-Sevast’yanov queueing system and its convergence rate.

G. A. ZVERKINA
Moscow State University of Railway Engineering (MIIT),
V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences
zverkina@gmail.com

UDC 519.21

Key words: distribution of the state of the queueing system, rate of convergence, regenerative processes, coupling method.

Abstract

The upper bound for the convergence rate of the distribution of the state of a queueing system with infinitely many servers is obtained, in the case when the intensity of the incoming flow and the intensity of the service depend on the state of the system.

1 Introduction.

1.1 Motivation

Consider the queueing system with infinitely many servers. Let $t_1, t_2, \ldots, t_n$ be the times of first, second, \ldots, $n$-th customers input (denote $t_0 \equiv 0$). At the input time, the customer immediately begins to be serviced. The length of the time of the service of $i$-th customer is a random variable (r.v.) $\xi_i$ with cumulative distribution function (c.d.f) $G_i(s)$. The periods between customer inputs are $\eta_i \equiv t_i - t_{i-1}$ – r.v. with c.d.f. $F_i(s)$. For simplicity, suppose that the distributions of r.v. $\xi_i$ and $\eta_i$ are absolutely continuous (with respect to the Lebesgue measure). Denote $f_i(s) \equiv F_i'(s)$, $g_i(s) \equiv G_i'(s)$.

The distributions $F_i(s)$ and $G_i(s)$ can be described by the intensities.

Definition 1 Let $\mathcal{T}$ be a some random time period (with c.d.f. $\Phi(x)$) starting at the time $t = 0$. And let at the time $s > 0$ this time period in not ended. Then

$$P\{\mathcal{T} \in (s, s + \Delta) | \mathcal{T} > s\} = \frac{\Phi(s + \Delta) - \Phi(s)}{1 - \Phi(s)}.$$ 

If c.d.f. $\Phi(x)$ is absolutely continuous, then

$$\phi(s) \equiv \lim_{\Delta \downarrow 0} \frac{P\{\mathcal{T} \in (s, s + \Delta) | \mathcal{T} > s\}}{\Delta} = \frac{\Phi'(s + 0)}{1 - \Phi(s)}$$

is called the intensity of the end of the random period $\mathcal{T}$ at the time $s$ given $\mathcal{T} \geq s$.  

\[\blacksquare\]
Remark 1 Definition holds:
\[ \mathbb{P}\{ \mathcal{T} \in (s, s + \Delta | \mathcal{T} > s) \} = \phi(s)\Delta + o(\Delta). \]

Remark 2 C.d.f. and its density can be reconstructed by the intensity:
\begin{align*}
\Phi(s) &= 1 - \exp\left( - \int_0^s \phi(u) \, du \right); \quad (1) \\
\Phi'(s) &= \phi(s) \exp\left( - \int_0^s \phi(u) \, du \right). \quad (2)
\end{align*}

It is easy to see: if c.d.f. is exponential then its intensity is a constant.
If c.d.f. \( \Phi(s) \) is not continuous, then it can also use the intensity of c.d.f. \( \Phi(s) \), but the formulae (1) and (2) are not true.

So, the behaviour of considered queueing system with infinitely many servers, will be described on the intensities’ language. Namely.

1. The intensity of the input of next on line customer at the time moment \( t \) is \( \lambda(t) \), i.e.,
\[ \mathbb{P}\left\{ \text{Next on line customer comes in the time interval } (t, t + \Delta] \right\} = \lambda(t)\Delta + o(\Delta). \]

2. Denote \( h_i(t) \) – the service intensity (at the time \( t \)) of \( i \)-th (by the order of arrivals) customers given at the time moment \( t \) in the queueing system there are \( n \) customers, i.e.
\[ \mathbb{P}\left\{ \text{\( i \)-th customer being on the service at the time } t \text{ ends own service in the time interval } (t, t + \Delta] \right\} = h_i(t)\Delta + o(\Delta). \]

Here and everywhere the \( I \)-th customer being in the queueing system, is the \( I \)-th by order of arrival from all customers being in the queueing system.

In the case when \( \lambda(t) \equiv \lambda \) and \( h_i(t) \equiv \mu \), the considered queueing system is a queueing system \( \text{M|M|}\infty \). Firstly, this queueing system was studied in [6]; therein the following statement has been proved: if \( \mu < \lambda \), then the distribution of the quantity \( n_t \) of the customers being in the queueing system converges to the stationary distribution:
\[ \lim_{t \to \infty} \mathbb{P}\{n_t = k\} = \mathcal{P}_k = \mathcal{P}_0 \frac{\lambda^k}{\mu^k}, \text{ where } \mathcal{P}_0 = 1 - \frac{\lambda}{\mu}. \]

Later, this result was generalized for the case when the intensity of the input flow depends on the quantity \( n_t \) of the customers being in the queueing system.
\[ \lambda(t) \equiv \lambda_k \text{ for } n_t = k \text{ (see, e.g., [11, Ch. 4, §4-5])}: \]

\[ \lim_{t \to \infty} P\{n_t = k\} = \mathcal{P}_k = P_0 \frac{\prod_{i=0}^{k-1} \lambda_i}{\mu^k}, \quad \text{where } P_0 = \left( \sum_{k=0}^{\infty} \frac{\prod_{i=0}^{k-1} \lambda_i}{\mu^k} \right)^{-1}. \quad (3) \]

Moreover, if the intensity of the service of all customers depends on the quantity \( n_t \) of the customers in the queueing system (i.e. on \( n_t \)), i.e. \( h_i(t) \equiv \mu_k > 0 \text{ for } n_t = k \), then the formula (3) is true, but with replacing \( \mu^k \) by \( \prod \mu_i \). Clearly, if \( \lambda_k = 0 \text{ for some } k \), then (3) is useful for the queueing system with bounded quantity of servers, without places for the queue (without the buffer to wait for service).

In the case described above, when

\[ \lambda(t) \neq \lambda, \quad \mu(t) \neq \mu, \quad \lambda(t) = \lambda_n, \quad h_i(t) = \mu_n, \quad (4) \]

where \( n_t \) is a quantity of the customers being in the queueing system at the time moment \( t \), the (random) service times, and the input flow are not independent.

In [9] the queueing system with Poisson input and arbitrary absolutely continuous distribution of time of service (queueing system \( M|G|\infty \)) was studied. Here all customers have the same distribution of the service. The length of service and input flow are independent. In the language of the intensities, it means that if at the time moment \( t \) in the system there are \( n_t \) customers, and elapsed time of the service of \( i \)-th is \( x_i^{(i)} \text{ (} i = 1, 2, \ldots, n \text{) } \), then the intensity of service of \( i \)-th customer is equal \( h_i(t) = h\left(x_i^{(i)}\right) \text{ at the time moment } t \).\

Denote \( \mu \overset{\text{def}}{=} \left( \int_0^{\infty} s d G(s) \right)^{-1} \), where \( G(s) = 1 - \exp \left( -\int_0^s h(u) \, du \right) - \text{c.d.f. of the service length. In [9] the statement was made: for } k > 0, \)

\[ \lim_{t \to \infty} P\{n_t = k; x_t^{(1)} < \alpha_1, x_t^{(2)} < \alpha_2, \ldots, x_t^{(k)} < \alpha_k\} = \mathcal{P}(k; \alpha_1, \alpha_2, \ldots, \alpha_k) = P_0 \prod_{i=1}^{k} \lambda_i G(1 - \alpha_i), \]

where the normalizing multiplier \( P_0 \) is calculated in the same way as in (3).

This fact for queueing system \( M|G|m|0 \), but in more generalized form, without assumption of the absolutely continuity of c.d.f. of the service time \( G(s) \), was proved in [21] (see also [22]).
I.e. in [9] the studied queueing system had
\[ \lambda(t) \equiv \lambda, \quad h_i(t) = h_i \left( x_i(t) \right), \]
where \( x_i(t) \) is the elapsed time of service of the \( i \)-th by the order of the arrival customer \( (i = 1, 2, \ldots, n_t, \text{ where at the time } t, \text{ the queueing system contains } n_t \text{ customers}) \).

Later, the results based on [6], was obtained in many papers (see, e.g., [10, 24, 7]).

The natural generalization of the cases of (4) and (5) is the queueing system in which the intensity of input flow and the intensities of service depend not only of \( n_t \), but also they depend of the elapsed service times and of the time from the last customer arrival:
\[ \lambda(t) = \lambda \left( x_i(0); x_i(1), x_i(2), \ldots, x_i(n_t) \right), \quad h_i(t) = h_i \left( x_i(0); x_i(1), x_i(2), \ldots, x_i(n_t) \right), \]
where \( x_i(0) \) is the time from the last arrival of the customer (at the time \( t \)), and \( x_i(t) \) is the elapsed service time of \( i \)-th by the order of arrival customer staying in the queueing system at the time \( t \). For applications of the queueing theory, it is very importantly, to know the stationary distribution of the queueing system; this distribution is the base for calculation of many important parameters of the complex technical systems.

Also for applications, it is very important to know the rate of convergence of the parameters of some technical systems to the stationary value. For this aim, it is enough to know the convergence rate of the distribution of the queueing system to the stationary one.

Such problems was considered, e.g., in [15, 16, 18].

For example, in [15] the following fact was proved: for queueing system with one server and infinitely many places in queue, in the case, when the intensity of the service \( h \) and intensity of the input flow \( \lambda \) depend on the quantity of customers in the queueing system \( n_t \), and on the elapsed service time \( x_t \), the next Theorem 1 is proved.

**Theorem 1** Denote \( X_t \overset{\text{def}}{=} (n_t, x_t) \), where \( n_t \) is the quantity of customers in the queueing system at the time moment \( t \), \( x_t \) is the elapsed service time (of the customer being on the service) at the time moment \( t \). In the case \( n_0 = 0 \), the intensity of the input flow is constant, and it is equal to \( \lambda_0 > 0 \), and \( X_t \overset{\text{def}}{=} (0, 0) \).

If \( \lambda(n_t, x_t) \) and \( h(n_t, x_t) \) are Borel measurable and bounded, and
\[
\sup_{n, x: n > 0} \lambda(n, x) = \Lambda < \infty; \quad \inf_{n > 0, x > 0} h(n, x) \geq \frac{C_0}{1 + x};
\]
\[ C_0 > 4(1 + 2\Lambda), \quad \text{and for some } k > 1 \quad C_0 > 2^{k+1}(1 + 2\Lambda^k), \]
then the distribution of the process \( X_t = (n_t, x_t) \) converges to the stationary distribution \( \mathcal{P} \) on the state space \( \mathcal{X} = \{(0, 0)\} \cup \{(n, x): n \in \mathbb{N}, x \geq 0\} \), and there
exists \( m > k \) and \( C > 0 \), such that for all \( t \geq 0 \)
\[
\sup_{S \in \scr{M}(X)} |P(X_t \in S|X_0 = (n,x)) - \mathcal{P}(S)| \leq C \frac{(1 + n + x)^m}{(1 + x)^k},
\]
where \( \mathcal{P} \) is a stationary distribution of the process \( X_t \).

In the proof of this Theorem\(^1\) there is (not full) algorithm of the calculation of the constant \( C \).

But there exists many results about the type of the convergence rate of the distribution of the state of the queueing system to the stationary distribution (see, e.g., \[24, 20, 13, 1\] et al.). The main result of this works is following: some parameters of the queueing system \( \chi_t \) converge to the stationary value \( \tilde{\chi} \) with the rate \( \phi(t) \), i.e. there exists the constant \( K \) such, that for all \( t \), \( |\chi_t - \tilde{\chi}| < K\phi(t) \) is true (here \( \phi(t) \searrow 0 \)).

For example, for the queueing system studied in \[18\] (Theorem\(^1\)), for the convergence of the distribution of the process \( X_t \) to the stationary distribution with the rate less then \( \frac{K}{(1 + x)^k} \), \( k \in [0, C_0 - 1) \), it is sufficiently \( \Lambda > C_0 \), but without additional conditions of the Theorem\(^1\) there was no bounds for the constant \( K \).

Emphasize, the queueing system \( M|GI|\infty \) is well studied, and many its parameters are well known (see, e.g., \[7\]). For example, this is \( \mathcal{P}_k(t) \) (the probability that the system contains \( k \) customers in the service at the time moment \( t \)), the expectation and the variance of the busy period, and the Laplace transform of c.d.f. of the busy period (it at the time \( t = 0 \) the queueing system \( M|GI|\infty \) is idle).

Correspondingly, these results make it easy to estimate the convergence rate of \( \mathcal{P}_k(t) \) to their stationary values \( \mathcal{P}_k \) – see, for example, \[5\].

1.2 Well-known facts and the aim of this paper

Recall that the queueing system \( M|GI|\infty \) consists of the infinite number of servers operating independently, and the random time of service of all customers has the same distribution with c.d.f. \( G(s) \). Input flow is Poisson flow with constant parameter \( \Lambda \). At the initial time moment \( t = 0 \) the queueing system is idle (no customer on the service).

Our goal is a study of the behaviour of the queueing system with infinitely many servers, in which the input flow and (random) service times are described by the intensities depending on the full queueing system state. Therefore, the service times and input flow are dependent. Naturally, despite this dependence, it is assumed that the characteristics of r.v. “service times” and “the time interval between the arrivals of the customers” are in some boundaries. For example, the requirement that the incoming flow intensity does not be zero and bounded is naturally.

In this paper, the full queueing system state at the time \( t \) consists of time from the last arrival of the customer before time \( t \), and the elapsed times of the service of the customer being on the service at the time \( t \).
Suggested above dependence of the service times and input flow allows us to claim that the process of service such queueing systems is Markov and regenerating.

The regeneration points of this process are the time moments when all components of the full queueing system state are equal to zero. It is possible at the time moment when in the idle queueing system the new customer comes. I.e. the regeneration point is the time moment when the queueing system serves only one customer, and the elapsed time of its service is equal zero; also at this time moment, the time from the last arrival is equal zero.

2 Description of the studied queuing system

So, consider the queueing system, in which the intensity of input flow and intensities of service depend on the full queueing system state, which will be described below.

This full queueing system state at the time moment $t$ includes: the time $x_t^{(0)}$ from the last arrival of the customer. And elapsed times of service $x_t^{(i)}$ of $i$-th customer from the customer being in the queueing system at the time $t$; for convenience, also the full queueing system state includes the variable $n_t$ – the quantity of the customers in the queueing system at the time moment $t$. (The customers are numbered by the order of the busy periods arrivals.) So, the full queueing system state is a vector $X_t = (n_t, x_t^{(0)}; x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(n_t)})$. As $i$-th customer arrived before $(i + 1)$-th one, clearly, that $x_t^{(i)} \geq x_t^{(i+1)}$ for $i = 1, \ldots, n_t$. Also $x_t^{(0)} \leq x_t^{(n_t)}$, i.e. $n_t$-th customer appeared in the system no later than the time $t - x_t^{(0)}$.

The intensity of the input flow is function from $X_t$: $\lambda = \lambda(X_t)$ – as soon as the intensity of service of $i$-th customer is a function from $X_t$: $h_i = h_i(X_t)$. Recall, the $i$-th customer has the elapsed service time $x_t^{(i)}$. Hence, for little time interval $\Delta > 0$ given $X_t = (n_t, x_t^{(0)}; x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(n_t)})$ (where $x_t^{(i)} \geq x_t^{(i+1)}$) the formulae
are true:

\[
\begin{align*}
\mathbb{P} \left\{ n_{t+\Delta} = n_t + 1, \ x_{t+\Delta}^{(i)} = x_t^{(i)} \Delta + \Delta \text{ for all } i = 1, \ldots, n_t \right\} &= \lambda(X_t) \Delta + o(\Delta); \\
\mathbb{P} \left\{ n_{t+\Delta} = n_t - 1 \right\} &= \sum_{i=1}^{n_t} h_i(X_t) \Delta + o(\Delta); \\
\mathbb{P} \left\{ n_{t+\Delta} = n_t, \ x_{t+\Delta}^{(i)} = x_t^{(i)} + \Delta \text{ for all } i = 0, \ldots, n_t \right\} &= 1 - \left( \lambda(X_t) + \sum_{i=1}^{n_t} h_i(X_t) \right) \Delta + o(\Delta).
\end{align*}
\]

From (6) it follows, that the distribution of \( X_{t+\Delta} \) depends only on distribution of \( X_t \) and \( \Delta \). So, \( X_t \) is Markov process on the state space \( \mathcal{X} \overset{\text{def}}{=} \bigcup_{i=0}^{\infty} \mathcal{J}_i \), where

\[
\mathcal{J}_i = \{ i \} \times \prod_{j=0}^{i} \mathbb{R}_+, \ i \in \mathbb{Z}_+.
\]

Here the set \( \mathcal{J}_0 \) is the set of the idle states of the queueing system (when there no customers in the queueing system).

Denote \( \mathcal{P}^0 \) – the distribution the process \( X_t \) with the initial state \( X_0 \): \( \mathcal{P}^0(S) = \mathbb{P}\{X_t \in S \mid X_0 = x\}, \ x \in \mathcal{X}, \ S \in \mathcal{B}(\mathcal{X}) \).

Recall, that the time moments when the process \( X_t \) hits the state \((1, 0; 0)\), i.e. when the process \( X_t \) exits the set \( \mathcal{J}_0 \) (in the idle queueing system new customer comes), are the regeneration times for the process \( X_t \). Thus, if the time between two regeneration times has the finite expectation, then the process \( X_t \) is ergodic, and its distribution weakly converges to the some stationary (invariant) distribution.

**Remark 3** The weak class of the queueing system may be described by this way. For example, the queueing system \( M|\mathcal{M}|\infty, M|G|\infty, M|G|m|0, \) and other. –

For the well-studied queueing system \( M|G|\infty \), the finite expectation of the service time is enough for the finiteness of the regeneration period. And in this case the convergence rate of the parameters \( \mathcal{P}_k(t) \) to their stationary values \( \mathcal{P}_k \) increases if the number of finite moments of the service time increases – see, e.g., [3].

Accordingly, it is naturally to suppose that the (dependent) service times have finite moments, and the parameters of the input flow can be bounded by the parameter of Poisson input flow.
Suppose the conditions:

\[ h_i(X_t) \geq \frac{K}{1 + x_i^t}, \quad K > 2; \quad 0 < \lambda_0 \leq \lambda(X_t) \leq \Lambda < \infty. \]  

(7)

From the first condition of (7) holds that c.d.f. of the service time \( G_i(s) \) satisfies the inequality \( G_i(s) \geq 1 - \frac{1}{(1 + s)^K} \), i.e. the service time has \( k \in [0; K) \) finite moments (see Remark 2).

Also the density of c.d.f. of the intervals between arrivals of the customers \( F(s) \) satisfies the inequality \( f_i(s) \leq \Lambda e^{-\lambda_0 s} \), and distribution of the time intervals \( \eta_i \) between arrivals of the customers has any moments:

\[ \mathbb{E}(\eta_i)^k \leq \frac{k! \Lambda}{\lambda_0^{k+1}}, \]  

(8)

in particular, this inequality is correct for the idle periods of \( X_t \), i.e. time intervals, where \( n_t \equiv 0 \).

Our goal is to find the computable bounds of the rate of convergence of the distribution of the process \( X_t \) to the stationary distribution in conditions (7).

For this aim, the coupling method will be used – see Section 5.

Previously, recall some information from the queueing theory, and give some necessary inequalities.

### 3 Some information about the queueing system MI|GI|∞

Here there are well-known results about the queueing system \( MI|GI|\infty \) (see [7], [8] at al.).

The input flow of \( MI|GI|\infty \) is Poisson, its intensity is constant \( \Lambda \). The arrived \( i \)-th customer immediately became be serviced. Its service time is r.v. \( \xi_i \) with c.d.f. \( G(s) \). All r.v. (service times of and times between the customer arrivals) are independent mutually.

At the initial time \( t = 0 \) the queueing system is idle.

Suppose, that \( \mathbb{E}(\xi_i)^k = k \int_0^\infty s^{k-1}(1 - G(s)) \, ds = m_k < \infty \) for some \( K \geq 1 \). For example,

\[ G(s) = 1 - \frac{1}{(1 + s)^K}, \quad K > 2. \]  

(9)

The condition (9) guarantees that for all \( k \in [0, K) \), \( \mathbb{E}(\xi_i)^k < \infty \). It is obvious that \( \mathbb{E}\xi_i = \frac{1}{K-1}; \quad \mathbb{E}(\xi_i)^2 = \frac{2}{K^2 - 3K + 2}; \) etc.

The loading of the queueing system is \( \rho \defeq \Lambda m_1 \).
A new the behaviour of the process of the queueing system described by Markov process $\tilde{X}_t = \left(\tilde{n}_t, x^{(0)}_t, x^{(1)}_t, \ldots, x^{(n)}_t\right)$, where $\tilde{n}_t$ is a quantity of the customers in the queueing system at the time moment $t$, and $x^{(i)}_t$ is the time from a last arrival of the customers to the queueing system; $x^{(i)}_t$ is the elapsed time of the service of $i$-th customer being in the queueing system ($i = 1, 2, \ldots, \tilde{n}_t$). Here and everywhere else the sign $\circ$ over a random variable, process or numerical characteristics indicate that the speech goes about the “classic” queueing system $MI|GI|\infty$.

Denote $\tilde{P}_t$ – the distribution the process $\tilde{X}_t$: $\tilde{P}_t(S) = \mathbb{P}\{\tilde{X}_t \in S\}$ for all $S \in \mathcal{B}(\mathcal{X})$, and again $\mathcal{X} = \bigcup_{i=0}^{\infty} \mathcal{X}_i$, where $\mathcal{X}_i = \{i\} \times \prod_{j=0}^{i} \mathbb{R}_+$, $i \in \mathbb{Z}_+$.

In condition (9), it is known (see, e.g., [7]), that $P_k(t) \overset{\text{def}}{=} \mathbb{P}\{\tilde{n}_t = k\} = \tilde{P}_t(S_0) = e^{-\Lambda t} \sum_{k=1}^{\infty} c_n^*(x)$, where $c_n(x) \overset{\text{def}}{=} \Lambda \int_0^t \left[1 - G(s)\right] \exp\left[-s - \Lambda \int_0^t \left[1 - G(v)\right] \frac{d v}{d t}\right] d s$.

Remark 4 $\mathcal{G}(+\infty) = \Lambda m_1$, and $P_0(t) \geq e^{-\Lambda m_1} = e^{-\rho}$ for all $t \geq 0$.

Moreover, the distribution of the busy period of the queueing system $M|G|\infty$ is known (see, e.g., [8]). Denote $\zeta_i$ – $i$-th busy period of the queueing system $M|G|\infty$. This r.v. has c.d.f. $B(x) \overset{\text{def}}{=} \mathbb{P}\{\zeta_i \leq x\} = 1 - \frac{1}{\Lambda} \sum_{k=1}^{\infty} c^{n*}(x)$, where $c(x) \overset{\text{def}}{=} \Lambda (1 - G(x)) e^{-\mathcal{G}(x)}$, and $c^{n*}$ is $n$-th convolution of the function $c(x)$.

Also the Laplace transform of the distribution $B(x)$ is known (see, e.g., [8]):

$$
\mathcal{L}[B](s) = 1 + \frac{s}{\Lambda} - \frac{1}{\Lambda} \int_0^t \exp\left[-st - \Lambda \int_0^t \left[1 - G(v)\right] \frac{d v}{d t}\right] d t.
$$

Moreover, there exists the formulae for the moments of the busy period:

$$
E(\zeta_i)^n = (-1)^{n+1} \left\{ \frac{e^\rho}{\Lambda} n C^{(n-1)} - e^\rho \sum_{k=1}^{n-1} C_n^k E(\zeta_i)^{n-k} C^{(k)} \right\}, \quad n \in \mathbb{N},
$$

where $C^{(k)} = \int_0^\infty \left[1 - G_0(t)\right] \frac{d t}{d t}$.
From (10) and (11), it can obtain
\[
\mathbb{E}^{\circ} \zeta_i = e^\rho - 1, \quad \text{and} \quad \mathbb{E}^{\circ} (\zeta_i)^2 = \frac{2e^{2\rho}}{\Lambda} \int_0^\infty \left( \exp \left( -\Lambda \int_0^{1 - G(v)} dv \right) \right) dt + \frac{2e^\rho}{\Lambda^2} - \frac{2}{\Lambda^2} - 2\frac{e^\rho - 1}{\Lambda^2}.
\] (12)

But for calculation of \( \mathbb{E}^{\circ} (\zeta_i)^k, k > 2 \), the use of the formulae (10) and (11) is very complicated and hardly to apply in practice.

But for our aim, we need to use the bounds for \( \mathbb{E}^{\circ} (\zeta_i)^k \). For this goal, we use

**Proposition 1** If \( \mathbb{E}^{\circ} (\zeta_i)^k < \infty \) for some \( k \in \mathbb{N} \), then the following inequality is true:
\[
\mathbb{E}^{\circ} (\zeta_i)^k \leq \frac{\mathbb{E}^{\circ} (\zeta_i)^k}{1 - e^{-\rho}} \times \varphi((1 - e^{-\rho}), k - 1),
\] (13)

where \( \varphi(x, k) \overset{\text{def}}{=} = \sum_{n=1}^\infty n^{k-1}x^n = \left( x \frac{d}{dx} \right)^k \frac{1}{1 - x} \).

**Proof.** The function \( c(x) \) from (10) is nonnegative, and
\[
\int_0^\infty c(x) \, dx = \int_0^\infty \Lambda (1 - G(x)) \exp \left( -\Lambda \int_0^x (1 - G(s)) \, ds \right) \, dx =
\]
\[
= \left( 1 - \exp \left( -\Lambda \int_0^\infty (1 - G(s)) \, ds \right) \right) = \left( 1 - e^{-\Lambda m_1} \right) = (1 - e^{-\rho}) \overset{\text{def}}{=} = \rho,
\]
where again \( \rho = \Lambda m_1 \).

I.e. the function \( \zeta(x) \overset{\text{def}}{=} = \theta^{-1} c(x) \) is a density of c.d.f. of some r.v. \( \eta \), and \( \zeta^{n*}(x) \) is a density of c.d.f. of \( \sum_{j=1}^n \eta_j \) – the summ of \( n \) i.i.d. r.v. with the density of c.d.f. \( c(x) \).

Thus, \( \int_0^\infty \zeta^{n*}(x) \, dx = 1 \). Therefore,
\[
\mathbb{E}^{\circ} \zeta_i = \int_0^\infty (1 - B(x)) \, dx = \sum_{n=1}^\infty \frac{1}{\Lambda} \int_0^\infty c^{n*}(x) \, dx =
\]
\[
= \frac{1}{\Lambda} \sum_{n=1}^\infty \theta^n \int_0^\infty c^{n*}(x) \, dx = \frac{1}{\Lambda} \sum_{n=1}^\infty \theta^n = \frac{1}{\Lambda} \times \frac{\theta}{1 - \theta} = \frac{e^\rho - 1}{\Lambda}.
\]
Now, if \(m_{k+1} = \mathbb{E}\xi_{i}^{k+1} < \infty\) for some \(k \in \mathbb{N}\), then

\[
\mathbb{E}\vartheta^{k} = \int_{0}^{\infty} x^{k} \varphi(x) \, dx = \int_{0}^{\infty} x^{k} \varphi^{-1}(1 - G(x)) e^{-\varphi(x)} \, dx \leq \frac{\Lambda}{\varrho} \int_{0}^{\infty} x^{k}(1 - G(x)) \, dx = \frac{\Lambda \mathbb{E}\xi_{i}^{k+1}}{\varrho(k+1)},
\]

and

\[
\int_{0}^{\infty} x^{k} \varsigma^{n*}(x) \, dx = \mathbb{E}\left(\sum_{j=1}^{n} \vartheta_{j}\right)^{k} \leq n^{k} \mathbb{E}\vartheta^{k} \leq \frac{n^{k} \Lambda m_{k+1}}{\varrho(k+1)}.
\]

In the last inequality the Jensen’s inequality was used in the form: for \(k \geq 1\),

\[
(a_{1} + \ldots + a_{n})^{k} \leq n^{k-1}(a_{1}^{k} + \ldots + a_{n}^{k}).
\]

Now,

\[
\mathbb{E}(\xi_{i})^{k} = \int_{0}^{\infty} kx^{k-1}(1 - B(x)) \, dx = \frac{1}{\Lambda} \sum_{n=1}^{\infty} \int_{0}^{\infty} kx^{k-1} \varsigma^{n*}(x) \, dx = \frac{1}{\Lambda} \sum_{n=1}^{\infty} \varrho^{n} \mathbb{E}\left(\sum_{j=1}^{n} \vartheta_{j}\right)^{k-1} \leq \frac{1}{\Lambda} \sum_{n=1}^{\infty} \varrho^{n} n^{k-1} \Lambda \mathbb{E}\xi_{i}^{k} = \frac{\mathbb{E}\xi_{i}^{k}}{\varrho} \sum_{n=1}^{\infty} n^{k-1} \varrho^{n} = \frac{\mathbb{E}\xi_{i}^{k}}{\varrho} \times \varphi(\varrho, k-1).
\]

Proposition \[\] is proved. \[\]

4 Bounds for the process \(X_{t}\)

Return to the studied process \(X_{t}\), described in the Section 2. Suppose, that at the initial time \(t = 0, X_{0} = (0, 0)\). I.e. the queueing system is idle, and it waits the first customer. At the time of arrival of the first customer, the process \(X_{t}\) hits to the state \((1, 0); 0\); and this time moment is the regeneration time.

Recall, that the process \(\tilde{X}_{t}\) described in the Section 3 also starts from the state \((0, 0); \tilde{X}_{0} = (0, 0)\). At the time of arrival of the first customer, the first regeneration period begins.

**Definition 2** The random variable \(\eta\) does not exceed the random variable \(\theta\) by distribution, if for all \(s \in \mathbb{R}\) the inequality \(F_{\eta}(s) = \mathbb{P}\{\eta \leq s\} \geq \mathbb{P}\{\theta \leq s\} = F_{\theta}(s)\) is true. Hence, c.d.f. of r.v. \(\theta\) does not exceed c.d.f. of r.v. \(\eta\). This is order relation. Denote it by \(\eta < \theta\).
Proposition 2 \( \eta \prec \theta \) iff on the som probability space \( (\Omega, \mathcal{F}, P) \) there exists r.v. \( \eta' \) and \( \theta' \) such that \( \eta' \leq \theta' \) for all \( \omega \in \Omega \), and \( P\{\eta \leq s\} = P\{\eta' \leq s\} \) and \( P\{\theta \leq s\} = P\{\theta' \leq s\} \) for all \( s \in \mathbb{R} \).

Proof. see [23]. \( \blacksquare \)

Proposition 3 If \( X_0 = X_0 = (0, 0) \), and the condition (7) holds true (recall, that input flow for \( \overset{\circ}{X}_t \) is Poisson with the intensity \( \Lambda \), and the time of service has c.d.f. (9)), then for all \( t \geq 0 \) condition (15) is true:

\[
\overset{\circ}{n}_t \prec \overset{\circ}{n}_t . \tag{15}
\]

Proof. The proof is based on the method of one probability space. Namely, on the one probability space, it create the processes \( X_t \) and \( \overset{\circ}{X}_t \) (more specifically, their version) by such a way, that for all trajectories of both processes, the inequality \( n_t \leq \overset{\circ}{n}_t \) is true – for all \( t \geq 0 \).

This construction is possible by two reasons.

I. The input flow of the process \( X_t \) has the intensity, which does not exceed the intensity of the input flow of the process \( \overset{\circ}{X}_t \) (\( \lambda(X_t) \leq \Lambda \) from (7)).

Therefore, on the one probability space, it can create two input flows for both process: the input flow \( \overset{\circ}{\mathfrak{F}}_t \) with intensity \( \lambda(X_t) \) for the process \( X_t \), and the input flow \( \overset{\circ}{\mathfrak{F}}^+_t \) with intensity \( \Lambda - \lambda(X_t) \); the union of the flows \( \overset{\circ}{\mathfrak{F}}_t \) and \( \overset{\circ}{\mathfrak{F}}^+_t \) is the flow with intensity \( \Lambda \) for the process \( \overset{\circ}{X}_t \).

So, at any time \( t \geq 0 \), the quantity of the customers arriving to the queueing system with the process \( X_t \), does not exceed the quantity of the customers arriving to the queueing system with the process \( \overset{\circ}{X}_t \). Moreover, at the times of arrivals of the customers to the queueing system with the process \( X_t \), contemporaneously the customers arrive into the queueing system with the process \( \overset{\circ}{X}_t \); these customers are called “common”.

In other words, at the any time moment, the vector \( \left( x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(n_t)} \right) \) consists of the components of the vector \( \left( \overset{\circ}{x}_t^{(1)}, \overset{\circ}{x}_t^{(2)}, \ldots, \overset{\circ}{x}_t^{(n_t)} \right) \). I.e. the system described by the process \( \overset{\circ}{X}_t \), the vector \( \left( x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(n_t)} \right) \) consists of the elapsed service times of the customers. And these elapsed service times coincide with some components of the vector \( \left( \overset{\circ}{x}_t^{(1)}, \overset{\circ}{x}_t^{(2)}, \ldots, \overset{\circ}{x}_t^{(n_t)} \right) \).

II. The intensity of service of \( i \)-th customer being in the queueing system described by the process \( \overset{\circ}{X}_t \): \( \overset{\circ}{x}_t \), \( \overset{\circ}{x}_t^{(1)}, \overset{\circ}{x}_t^{(2)}, \ldots, \overset{\circ}{x}_t^{(n_t)} \), is equal \( \overset{\circ}{h}_i \).
\[
G'\left(\frac{\circ(t)}{x_t}\right) = \frac{K}{1 - \circ(t)} \quad 1 + \circ(t)
\]

Thereby, the intensity of the service of the “common”
customers in the process \(\circ\) does not exceed the intensity of the service of the
same customers in the process \(X\). This is the second reason for apply the method
of one probability space.

Considering that the time of service has c.d.f. which can be calculated by the
intensity of the service (see (1)–(2)), it can obtain the next fact. The residual time
of service of the “common” customers in the process \(X\) does not exceed \(\circ\) the
distribution the residual time of service of these customers in the process \(\circ\). So,
it can to construct on the one probability space this r.v. by such a way that the
“common” customers will complete the service in the process \(X\) no later than in
the process \(\circ\) – see Proposition 2.

Therefore, in the queueing system described by the process \(\circ\) the flow of
the customers is larger than in the queueing system described by the process \(X\),
and simultaneously arrived (common) customers earlier leave the queueing system
described by the process \(X\). This is the reason for the truth of the Proposition 3.

\[\blacksquare\]

**Corollary 1** From (15), it follows that the property of orderliness in the distribu-
tion (see Definition 2) not true for only \(n_t\) and \(\circ n_t\), but it is true also and for the
busy periods of the processes \(X_t\) and \(\circ X_t\).

The processes \(X_t\) and \(\circ X_t\) are regenerative (recall that the regeneration points
are the time points when the process is equal to (1,0; 0), i.e. the beginning of the
busy period). Hence, the distribution of the busy periods \(\zeta\) of the process \(X_t\) and
the busy periods \(\circ \zeta\) of the process \(\circ X_t\) not depend on their number \(i\). So, \(\zeta, \circ \zeta\),
and for all \(k > 0\) the inequality \(E(\zeta)^k \leq E(\circ \zeta)^k\) is true. It means that the bounds
(13) are true for \(\zeta_i\).

\[\blacksquare\]

**Corollary 2** From Remark 4 and (15) it follows that for all \(t \geq 0\) (and \(X_0 =
(0,0))\) the inequality

\[
P\{X_t \in S_0\} = P\{n_t = 0\} \geq \mathcal{P}_0(t) \geq e^{-\rho}.
\]

(16)

\[\blacksquare\]

5 Coupling method

The original coupling method for the Markov processes have been applied to the
Markov chains with a finite number of states (see [5], [13]). The main idea of this
method is as follows.
Let two homogeneous Markov processes $Y_t$ and $Z_t$ (on the probability space $(\Omega, \mathcal{F}, P)$ with the filtration $\mathcal{F}_t$) with the same transition function, have the different states at the initial time moment.

Denote $\tau = \tau(Y_0, Z_0) \overset{\text{def}}{=} \inf\{t > 0 : Y_t = Z_t\}$. So, $Y_\tau = Z_\tau$, and for all $t > \tau$, the distributions of the processes $Y_t$ and $Z_t$ coincide in accordance with the Markov property, i.e. $P\{Y_t \in S | t > \tau\} = P\{Z_t \in S | t > \tau\}$ for all $S \in \mathcal{B}(\mathcal{X})$, where $\mathcal{B}(\mathcal{X})$ is $\sigma$-algebra on the state space $\mathcal{X}$ of $Y_t$ and $Z_t$.

**The basic coupling inequality:**

\[
{|P\{Y_t \in S\} - P\{Z_t \in S\}| = |P\{Y_t \in S \& \tau > t\} - P\{Z_t \in S \& \tau > t\} + |P\{Y_t \in S \& \tau < t\} - P\{Z_t \in S \& \tau < t\}| = |P\{Y_t \in S \& \tau > t\} - P\{Z_t \in S \& \tau > t\}| \leq P\{\tau > t\}, \tag{17}
\]

as if $\tau < t$, then “coupling” of the processes $Y_t$ and $Z_t$ already occurred, and their distribution at the time $t$ are the same, and $P\{Y_t \in S | \tau < t\} = P\{Z_t \in S | \tau < t\}$; here $S \in \mathcal{B}(\mathcal{X})$.

Therefore if there exists some bounds $E\phi(\tau(Y_0, Z_0)) \leq C(X_0, Y_0)$ for some increasing positive function $\phi(t)$, then the Markov inequality is applicable:

\[
P\{\tau(Y_0, Z_0) > t\} = P\{\phi(\tau(Y_0, Z_0)) > \phi(t)\} \leq \frac{E\phi(\tau(Y_0, Z_0))}{\phi(t)} \leq \frac{C(X_0, Y_0)}{\phi(t)}. \tag{18}
\]

Moreover, if there exists the stationary invariant distribution $\mathcal{P}$, and the distributions of $Y_t$ and $Z_t$ weakly converges to $\mathcal{P}$ as $t \to \infty$, then from (17) and (18) holds:

\[
|P\{Y_t \in S\} - \mathcal{P}(S)| \leq \int_{\mathcal{X}} \frac{C(X_0, y)}{\phi(t)} \mathcal{P}(dy) = \frac{\bar{C}(X_0)}{\phi(t)}.
\]

consequently,

\[
\|P^{X_0}_t - \mathcal{P}\|_{TV} \overset{\text{def}}{=} \sup_{S \in \mathcal{B}(\mathcal{X})} |P^{X_0}_t(S) - \mathcal{P}(S)| \leq 2 \frac{\bar{C}(X_0)}{\phi(t)}. \tag{19}
\]

Recall, that initially, the coupling method has been proposed for Markov chains with a discrete set of states (15), but studied in this paper process $X_t$ with transition low (6) is the continuous process. For two independent processes $X_t$ and $X'_t$ satisfying (6), and with different initial states, $P\{\tau(X_0, x'_0) < \infty\} = 0$.

For this reason, the notion of **successful coupling** will be applied (see [4]).

**Definition 3** Let $X_t$ and $X'_t$ are two independent versions of one homogeneous Markov process with the state space $\mathcal{X}$: $X_t$ starts from the state $X_0$, and $X'_t$ starts from the state $x'_0$.

**The successful coupling** (see [4]) of the processes $X_t$ and $X'_t$ is the pair process $Z_t = (Y_t, Y'_t)$, defined on the same probability space, and this pair process satisfies the conditions:
(i) for all $s \geq 0$ and $S \in \mathcal{B}(\mathcal{X})$ the equalities
\[ P\{Y_t \in S\} = P\{X_t \in S\}, \quad P\{Y'_t \in S\} = P\{X'_t \in S\} \] are true. (Correspondingly, $Y_0 = X_0$ and $Y'_0 = X'_0$.)

Emphasize, that for all fixed time moment $t \geq 0$ the random variables $Y_t$, $Y'_t$, $X_t$ and $X'_t$ are considered only as some random variables, not as the stochastic processes! Here there no coincidence of the distribution of two processes, but only the coincidence of the marginal distributions.

(ii) for all $t > \tau = \tau(X_0, X'_0) = \tau(Y_0, Y'_0) \overset{\text{def}}{=} \inf\{t \geq 0 : Y_t = Y'_t\}$ the equality $Y_t = Y'_t$ is true.

(iii) For any initial states $X_0, X'_0 \in \mathcal{X}$, $P\{\tau(X_0, X'_0) < \infty\} = 1$.

For successful coupling, the basic coupling inequality (17):
\[
|P\{X_t \in S\} - P\{X'_t \in S\}| = |P\{Y_t \in S\} - P\{Y'_t \in S\}|
= |P\{Y_t \in S\} - P\{Y'_t \in S\}| \times (1(\tau > t) + 1(\tau \leq t)) \leq P\{\tau > t\}
\]
is true.

So now, our goal is the construction of the successful coupling for two version of the process $X_t$ ($X_t$ with initial state $X_0 = (0, 0)$ and $X'_t$ with initial state $X'_0 = (n_0, x'_t(0); x'_t(1), x'_t(2), \ldots, x'_t(n'_0))$), and both processes satisfy the conditions (6). From this construction of the successful coupling, the bounds for the convergence rate of the distribution of the process $X_t$ to the stationary distribution will be obtained.

For use the coupling method, the next Lemma is needful.

### 5.1 Basic coupling Lemma

This Lemma exists in many papers (see, e.g., [14], [12]). Here, this Lemma is in the simplest form.

**Definition 4** The common part of the distributions of two r.v. $\xi_1$ and $\xi_2$ having the c.d.f. $\Psi_j(s) = P\{\xi_j \leq s\}$, $s \in \mathbb{R}$ ($j = 1, 2$) correspondingly, is $\varkappa \overset{\text{def}}{=} \varkappa(\Psi_1(s), \Psi_2(s)) \overset{\text{def}}{=} \int_{-\infty}^{\infty} \min(\psi_1(u), \psi_2(u)) \, du$, where
\[
\psi_j(s) \overset{\text{def}}{=} \begin{cases} 
\Psi_j(s), & \text{if there exists } \Psi_j'(s), \\
0, & \text{otherwise}, \quad \text{here } j = 1, 2.
\end{cases}
\]

**Lemma 1** If $\varkappa > 0$, then on some probability space there exists two r.v. $\xi'_1$ and $\xi'_2$ such, that:
\[
P\{\xi'_j \leq s\} = P\{\xi_j \leq s\} = \Psi_j; \quad \text{for } j = 1, 2
\]

\[
P\{\xi'_1 = \xi'_2\} \geq \varkappa.
\]
Proof. Denote
\[
\psi(u) \overset{\text{def}}{=} \min(\psi_1(u), \psi_2(u)); \quad \Psi(s) \overset{\text{def}}{=} \int_{-\infty}^{s} \psi(u) \, du;
\]
\[
\hat{\Psi}(s) \overset{\text{def}}{=} \frac{\Psi(s)}{s}; \quad \hat{\Psi}_j(s) \overset{\text{def}}{=} \frac{\Psi_j(s) - \Psi(s)}{1 - s}, \quad j = 1, 2
\]
(if \(s = 1\), then \(\hat{\Psi}_j(s) \equiv 0\)).

Let \(\mathcal{U}_1, \mathcal{U}_2\) and \(\mathcal{U}_3\) be independent r.v. with the continuous uniform distribution on \([0; 1]\). Put \(\xi'_j \overset{\text{def}}{=} \hat{\Psi}^{-1}(\mathcal{U}_j) \times 1(\mathcal{U}_3 \leq \kappa) + \hat{\Psi}^{-1}(\mathcal{U}_j) \times 1(\mathcal{U}_3 > \kappa)\).

Recall that the inverse function of monotonous function \(f(s)\) is \(f^{-1}(u) \overset{\text{def}}{=} \inf\{s : f(s) \geq u\}\).

R.v. \(\xi'_j\) satisfy conditions (20)–(21).

Lemma is proved. \(\blacksquare\)

Remark 5 It at some time \(\vartheta, n_\vartheta = 0\) (i.e. the queueing system is idle), then the residual time of the idle period of the process \(X_t\) (described in the Section 2) has the density of distribution \(\varphi(s) = \lambda(X_\vartheta + s) \exp\left(\int_{0}^{s} \lambda(X_{\vartheta + u}) \, du\right) \in (\lambda_0 e^{-\lambda s}, \Lambda_0 e^{-\lambda s})\).

Thus, if two processes \(X_t\) and \(X'_t\) with different initial states \(X_0\) and \(X'_0\), at some time \(\vartheta\) are in the set \(\mathcal{J}_0\) (the queueing system is idle), then the common part \(\kappa_\vartheta\) of the distributions of their residual times of the staying in the set \(\mathcal{J}_0\) satisfies the inequality
\[
\hat{\kappa}_\vartheta \geq \int_{0}^{\infty} \lambda_0 e^{-\lambda s} \, ds = \frac{\lambda_0}{\Lambda} = \hat{\kappa}.
\]

6 Construction of successful coupling for two processes \(X_t, X'_t\)

Consider two independent processes: the process \(X_t\) with initial state \(X_0 = (0, 0)\) and the process \(X'_t\) with initial state \(X'_0 = (n_0', x'_0(0), x'_0(1), x'_0(2), \ldots, x'_0(n_0'))\).

Their successful coupling \(\mathcal{Z}_t = (Y_t, Y'_t)\) will be constructed on the probability space \((\Omega, \mathcal{F}, P) = \prod_{i=0}^{\infty} (\Omega_i, \mathcal{F}_i, P_i)\), where \(\Omega_i = [0; 1], \mathcal{F}_i = \mathcal{B}([0; 1]),\) and \(P_i\) is the Lebesgue measure on \([0; 1]\).

The procedure of the construction of the successful coupling will be described by steps. This steps operate at the time moments when the first component of one of the processes
\[ Y_t = \left( m_t, y_t^{(1)}, y_t^{(2)}, \ldots, y_t^{(m_t)} \right), \quad Y'_t = \left( m'_t, y'_t^{(0)}, y'_t^{(1)}, y'_t^{(2)}, \ldots, y'_t^{(m'_t)} \right) \]

changes. Denote these time moments by \( t_k \): \( t_0 \overset{\text{def}}{=} 0 \); \( t_k \overset{\text{def}}{=} \inf \{ t > t_{k-1} : m_t + m_t' \neq m_{t_0} + m_{t_0}' \} \).

In all steps, the residual virtual times of the service of all customers being in the queueing systems (in hypothesis: during this service, the first component \( n_t \) and \( n_t' \) of processes not change), and the residual virtual times until the arrival of the next customers. For this aim, in any step, some quantity \( \nu_k \) of exemplars \((\Omega_i, \mathcal{F}_i, P_i)\) is used, and in any step, new exemplars \((\Omega_i, \mathcal{F}_i, P_i)\)  are used.

**Algorithm of construction**

**A.** If \( m_{t_k} + m_{t_k}' > 0 \), for convenience put \( Y_{\nu_k} = (i, \alpha_0; \alpha_1, \alpha_2, \ldots, \alpha_\ell) \), \( Y'_{\nu_k} = (j, \beta_0; \beta_1, \beta_2, \ldots, \beta_\ell) \).

In \( k \)-th step, the probability spaces \((\Omega_i, \mathcal{F}_i, P_i)\), \( i = \ell_k + 1, \ell_k + 2, \ldots, \ell_{k+1} \) are used; \( \nu_k \overset{\text{def}}{=} \ell_{k+1} - \ell_k = n_{\nu_k} + n_{\nu_k}' + 2 = i + j + 2 \). So, before \( k \)-th step, \( \ell_k = \sum_{i=0}^{k-1} \nu_i \) exemplars of the space \((\Omega_i, \mathcal{F}_i, P_i)\) was used.

At the time \( t_k \) virtual time until arrival of the next customer \( \xi \) (given in this time no ends of service) for \( Y_t \) has c.d.f.

\[ F_1(s) = 1 - \exp \left( \int_0^s \lambda(t, \alpha_0 + u; \alpha_1 + u, \alpha_2 + u, \ldots, \alpha_\ell + u) \, du \right). \]

R.v. \( \xi \overset{\text{def}}{=} F_1^{-1}(\omega) \), \( \omega \in \Omega_{t_k+1} \) is created on the space \((\Omega_{t_k+1}, \mathcal{F}_{t_k+1}, P_{t_k+1})\).

At the time \( t_k \) virtual residual time \( \eta_j \) of the service of \( j \)-th \((j = 1, 2, \ldots, i)\) customer (given in this residual time no ends of other services and no new customers) for \( Y_t \) has c.d.f.

\[ G_{1,j}(s) = 1 - \exp \left( \int_0^s h_j(t, \alpha_0 + u; \alpha_1 + u, \alpha_2 + u, \ldots, \alpha_\ell + u) \, du \right). \]

R.v. \( \eta_j \overset{\text{def}}{=} G_{1,j}^{-1}(\omega) \), \( \omega \in \Omega_{t_k+j+1} \) is created on the space \((\Omega_{t_k+j+1}, \mathcal{F}_{t_k+j+1}, P_{t_k+j+1})\).

The construction of \( \xi' \) and \( \eta'_j \) is analogous \((i = 1, 2, \ldots, j)\):

\[ F_2(s) = 1 - \exp \left( \int_0^s \lambda(j, \beta_0 + u; \beta_1 + u, \beta_2 + u, \ldots, \beta_\ell + u) \, du \right), \quad \xi' \overset{\text{def}}{=} F_2^{-1}(\omega), \]

\( \omega \in \Omega_{t_k+j+1}; \)

\[ G_{2,j}(s) = 1 - \exp \left( \int_0^s h_j(j, \beta_0 + u; \beta_1 + u, \beta_2 + u, \ldots, \beta_\ell + u) \, du \right), \quad \eta'_j \overset{\text{def}}{=} G_{2,j}^{-1}(\omega), \]

\( \omega \in \Omega_{t_k+j+1+2}. \)

In \( k \)-th step, \( \nu_k = i + j + 2 \) exemplars \((\Omega_i, \mathcal{F}_i, P_i)\) was used; \( \ell_{k+1} = \ell_k + \nu_k \).
The time $t_{k+1} = t_k + \min\{\xi, \xi', \eta_j, \eta_j' \mid (j = 1, 2, \ldots, i = 1, 2, \ldots)\}$ is the time of the change of the first component of the process $Y$ or of the process $Y'$.

If again $m_{t_k} + m'_{t_k} > 0$, then the procedure $A$ is repeated.

B. If $m_{t_k} + m'_{t_k} = 0$, then $Y_{t_k} = (0, \alpha_0)$ and $Y'_{t_k} = (0, \beta_0)$. C.d.f. of the times $\xi$ and $\xi'$ till the arrival of the new customer for the process $Y$ or $Y'$ correspondingly, is

$$P\{\xi \leq s\} = F_1(s) = 1 - \exp\left(\int_0^s \lambda(0, \alpha_0 + u) \, du\right);$$

$$P\{\xi' \leq s\} = F_2(s) = 1 - \exp\left(\int_0^s \lambda(0, \beta_0 + u) \, du\right).$$

By Remark 5 and (22), the common part of these distributions is not less than $\hat{\kappa} = \lambda_0 \tilde{\alpha}$. Therefore, on the space $\prod_{i = t_k + 1}^t (\Omega_i, \mathcal{F}_i, P_i)$ (here $\nu_k = 3$) By Lemma 1, there exists $\xi$ and $\xi'$ such that $P\{\xi = \xi'\} \geq \tilde{\alpha}$.

At the time moment $t_{k+1} \overset{\text{def}}{=} t_k + \min\{\xi, \xi'\}$ the first component of at least one of the processes $Y$, $Y'$ increases, and the first components of both processes $Y$, $Y'$ will change simultaneously – with probability greater than $\tilde{\alpha}$, i.e. the event $Y_{t_{k+1}} = Y'_{t_{k+1}} = (1, 0; 0)$ will happen.

C. If the event $Y_{t_k} = Y'_{t_k} = (1, 0; 0)$ happened, then for $t > t_k$ put the processes $Y_t$ and $Y'_t$ equal. Namely, it can construct only one process $Y_t$ by procedure given in A. And $Y'_t \equiv Y_t$ for $t > t_{k+1}$.

The processes $Y_t$ and $Y'_t$ constructed by the procedures $A$-$C$ satisfy the conditions (i)-(ii) from Definition 3.

Let us check the condition (iii) from Definition 3.

Consider the process $Y_t \overset{\text{def}}{=} X_t$. If $n'_0 > 0$, then by comparison with the “classic” queueing system (see Sections 3 and 4), it can estimate the moments of the residual time $\tau_1$ of the first busy period of $Y_t$'s, it is easy to see that for $k \in [0; K]$, $E(\zeta_1)^k < \infty$.

Denote $\varphi_0 \overset{\text{def}}{=} \zeta_0$, $\varphi_k \overset{\text{def}}{=} \inf\{t > \varphi_k : Y'_t \in \mathcal{S}_0, Y'_{t-0} \in \mathcal{S}_0\}$, $\varphi_k$ is the time of end of $k$-th busy period of the process $Y'_t$. If $Y'_0 \in \mathcal{S}$, put $\varphi_0 \overset{\text{def}}{=} 0$.

As distribution the process $Y_t$ is equal to distribution of the process $X_t$, the inequality $P\{X_{\varphi_0} \in \mathcal{S}_0\} \geq e^{-\rho}$ is true – see (16).

In addition, if both processes $Y_t$ and $Y'_t$ hit to the set $\mathcal{S}$, then at the exit of this set, they coincide with probability greater then $\tilde{\alpha}$. Hence, again denote $\tau \overset{\text{def}}{=} \inf\{t > \varphi_k : Y'_t = Y'_{t-0}\}$, and $\eta_k$ is the length of an idle period of $Y'_t$ after $\varphi_k$, and we have: $P\{\tau > \varphi_k + \eta_k\} \leq (1 - \omega)^k$. Note, that $\tau = \tau(Y'_0) = \tau(X'_0)$ (as $X_0 = Y_0 = (0, 0; 0)$ is fixed).
Evidently, \( \vartheta_k + \eta_k = \vartheta_0 + \eta_0 + \sum_{i=1}^{k} (\zeta_i + \eta_i) \), where \( \zeta_i \) is \( i \)-th busy period, and \( \eta_i \) is the next free period. Denote \( \vartheta_k \text{ def } \{ Y_{\vartheta_k} \neq Y_{\vartheta_k}' \text{ & } Y_{\vartheta_k} + \eta_k = Y_{\vartheta_k} + \eta_k \} \) – it is the event “coincidence of the processes \( Y_t \) and \( Y_t' \) is immediately after \( k \)-th hit of \( Y_t' \)‘ the set \( \mathcal{J}_0 \”).

**Proposition 4** If \( r \in [1; K) \) then \( E \tau^r \leq C_1(\varpi, r)E \vartheta_0^r + C_1(\varpi, r)E \zeta_1^r + C_3(\varpi, r, \lambda) \), where the constats \( C_1, C_2, C_3 \) can be calculated by the formulae below.

**Proof.** Firstly, the bounds for \( E(\eta_k)^r |_{\vartheta_k} \). If the event \( \vartheta_k \) happened, then r.v. \( \eta_k \) (see Lemma [11]) has a density of distribution

\[
\varphi(s) = \min \left\{ \lambda(Y_{\vartheta_k} + s) \exp \left( - \int_{0}^{s} \lambda(Y_{\vartheta_k} + u) \, du \right) \right\},
\]

\[
\lambda(Y_{\vartheta_k}' + s) \exp \left( - \int_{0}^{s} \lambda(Y_{\vartheta_k}' + u) \, du \right) \leq \Lambda e^{-\lambda_0 s}.
\]

Hence, \( E(\eta_k)^r |_{\vartheta_k} \leq \frac{r! \Lambda}{\lambda_0^{r+1}} \). Also \( E(\eta_k)^r |_{\vartheta_k} \leq \frac{r! \Lambda}{\lambda_0^{r+1}} \); \( \zeta_k \) not depends on \( \vartheta_j \) if \( k \neq j \).

Now for estimate \( E \tau^r \) we use the Jensen’s inequality in the form ([14]):

\[
E \tau^r \leq \sum_{k=1}^{\infty} \left( \vartheta_0 + \eta_0 + \sum_{i=1}^{k} (\zeta_i + \eta_i) \right)^r 1(\vartheta_k) \leq \sum_{k=1}^{\infty} (1 - \varpi)^{k-1} \varpi (2k + 2)^{r-1} \left( E \vartheta_0^r + (k + 1) \frac{r! \Lambda}{\lambda_0^{r+1}} + k E \zeta_1^r \right) < 2^{r-1} \left( E \vartheta_0^r \sum_{k=1}^{\infty} (1 - \varpi)^{k-1}(k + 1)^{r-1} + \frac{r! \Lambda}{\lambda_0^{r+1}} \sum_{k=1}^{\infty} (1 - \varpi)^{k-1} (k + 1)^{r-1} + E \zeta_1^r \sum_{k=1}^{\infty} (1 - \varpi)^{k-1}(k + 1)^{r-1} \right).
\]

Proposition [4] is proved.

So now, the constructed pair \( 2 \mathcal{I} = (Y_t, Y_t') \) is a successful coupling of the processes \( X_t \) and \( X_t' \), and there is the explicit bounds for the coupling time:

\[
E(\tau(X_0'))^r \leq C(r, \lambda, \Lambda, E \vartheta_0^r).
\]

Therefore,

\[
|P\{X_t \in S\} - P\{X_t' \in S\}| \leq \frac{C(r, \lambda, \Lambda, E \vartheta_0^r)}{t^r},
\]

19
but here the next stage – integration with respect to the stationary measure \(19\) – is impossible, because the stationary distribution of the queueing system \(M_{|G|\infty}\) (the process \(X_t\)) is unknown, and for the studied process \(X_t\) also.

7 Process \(X_t\) as a regenerative process

**Definition 5** The stochastic process \(\{W_t, t \geq 0\}\), defined on the probability space \((\Omega, \mathcal{F}, P)\), measurable with respect to the filtration \(\{\mathcal{F}_t, t \geq 0\}\), with the state space \((\mathcal{W}, \mathcal{B}(\mathcal{W}))\), is called regenerative process, if there exists the sequence of the Markov times (stopping times) \(\{\theta_i\}_{i \in \mathbb{N}}\) such that

1. \(W_{\theta_i} = W_{\theta_j}\) for all \(i, j \in \mathbb{N}\);

2. Random elements \(\Xi_i \overset{\text{def}}{=} \{W_t, t \in [\theta_i, \theta_{i+1}]\} (i \in \mathbb{N})\) are i.i.d.

The intervals \([\theta_i, \theta_{i+1}]\) are called regeneration periods. Evidently, the length of the regeneration periods \(\chi_i \overset{\text{def}}{=} \theta_{i+1} - \theta_i\) are i.i.d.. Denote \(\Phi(s) \overset{\text{def}}{=} P\{\chi_i \leq s\}\) – c.d.f. of r.v. \(\chi_i\); it is unknown.

If \(E \chi_i < \infty\), then the distribution of the regenerative process \(W_t\) weakly converges to the stationary (invariant) distribution – it is well-known fact.

Note, for the stochastic process \(X_t = (n_t, x_t^{(0)}; x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(n_t)})\) described in Section 2, the regeneration periods are the sum of busy period and next idle period: \(\chi_k = \zeta_k + \eta_k\).

The values of \(E(\zeta_k)\) and \(E(\eta_k)\) can be estimated by (8), (12) and (13).

Note, the regeneration period is greater then its part – idle period having the distribution density \(\varphi(s) \geq \lambda_0 e^{-\Lambda s}\) (Remark 5).

So, the next bounds are true:

\[
E \chi_k^r \leq 2^{r-1}(E \zeta_k^r + E \eta_k^r) = M_r, \quad E \chi_i > \frac{\lambda_0}{\Lambda^2}. \tag{25}
\]

The Markov times \(\{\theta_i\}_{i \in \mathbb{N}}\) form the embedded renewal process \(N_t = \sum_{i=1}^{\infty} 1(\theta_i < t)\).

Consider the backward renewal time of the process \(N_t\). This is Markov processes \(B_t \overset{\text{def}}{=} t - \theta_{N_t}\) (with the state space \(\mathbb{R}_+\)). It is well-known (19, 20), that the distribution of the backward renewal time \(\mathcal{P}_t\) weakly converges to the stationary invariant distribution \(\mathcal{P}\) with c.d.f.

\[
\tilde{\Phi}(s) = \frac{\int_0^s (1 - \Phi(u)) \, du}{E \chi_i}. \tag{26}
\]
Also, this distribution (26) is a stationary distribution for the forward renewal process \( R_t \) \( \stackrel{\text{def}}{=} \theta_{N_{t+1}} - t \) (here the renewal processes \( N_t \)).

As the process \( X_t \) is regenerative, its distribution fully defined by the time elapsed from the last regeneration point (beginning of the regeneration period).

So, \( \mathcal{P}_t = \mathbb{P}(B_t) \).

Now, consider the process \( X_t \) with initial state \((0,0)\) and the process \( X'_t \) with arbitrary initial state \( X_0 \). Both these processes have the embedded renewal processes: \( N_t \) and \( N'_t \) correspondingly. Moreover, for the renewal processes \( N_t \) and \( N'_t \) the backward renewal processes \( R_t \) and \( R'_t \) correspond.

The construction of the successful coupling \( Z_t = (Y_t, Y'_t) \) for the pair of the processes \( X_t \) and \( X'_t \) gives simultaneously the successful coupling \( W_t = (D_t, D'_t) \) for the pair of the processes \( B_t \) and \( B'_t \). Denote \( \mathbb{P}_t \) and \( \mathbb{P}'_t \) - the distributions of the processes \( B_t \) and \( B'_t \) correspondingly (\( \mathbb{P}\{B_t \in S\} = \mathbb{P}_t(S) \) and \( \mathbb{P}\{B'_t \in S\} = \mathbb{P}'_t(S) \)).

From (23) and (24) momentarily we have
\[
|\mathbb{P}\{B_t \in S\} - \mathbb{P}\{B'_t \in S\}| \leq C(\varpi, r, \lambda_0, \Lambda, \Xi) \frac{t^r}{t^r},
\]
where \( \vartheta_0 = R'_0 \) is the backward renewal process of the renewal process \( N'_t \) at the time moment \( t = 0 \).

**Proposition 5**

\[
\|\mathbb{P}_t - \mathbb{P}\|_{TV} \leq 2 \left( C_1(\varpi, r) \frac{\mathbb{E} \chi^{r+1}}{r+1} + C_1(\varpi, r) \mathbb{E} \chi'_1 + C_3(\varpi, r, \lambda_0, \Lambda) \right) t^r \leq \frac{C_1(r, \lambda_0, \Lambda)}{t^r},
\]
where \( r \in [1, K - 1) \).

**Proof.** Consider the backward renewal process \( B_t \) with the initial state \( B_0 = 0 \) and the backward renewal process \( \tilde{B}_t \) with the initial stationary distribution \( \mathbb{P}_t \); for all \( t \geq 0 \)
\[
\mathbb{P}\{\tilde{B}_t \in S\} = \mathbb{P}(S) = \int_S \frac{1 - \Phi(u)}{\mathbb{E} \chi_i} \, du.
\]

By substitution in (27) \( \vartheta_0 = R'_0 \) by the stationary forward renewal process \( \bar{R}_0 \) with the distribution (23), the Proposition 3 is proved.

Considering (24), it can calculate the upper bound for the value \( C_1(r, \lambda_0, \Lambda) \).

**Theorem 2** In the conditions (7), the inequality \( \|\mathcal{P}_t - \mathbb{P}\|_{TV} \leq \frac{C_1(r, \lambda_0, \Lambda)}{t^r} \), is true for all \( r \in [1, K - 1) \).
Proof. It was said above that \( P_t = F(P_t) \). Accordingly, \( P_t = F(P_t) \). Recall that \( \chi_i \) is the length of the regeneration period, which includes absolutely continuous r.v. So, the distribution of r.v. \( \chi_i \) is absolutely continuous. Therefore, the distribution of the backward renewal process \( B_t \) has a density of c.d.f \( \psi_t(s) \). It is known \(^{[28]}\), the c.d.f. of the stationary distribution of the backward renewal process is known, its density is \( \psi(s) = 1 - \Phi(u) \). Denote \( \sigma_t = \{ s : \psi_t(s) > \psi(s) \} \).

Now,

\[
|P_t(S) - P(S)| = \\
= \left| \int_0^\infty P\{X_t \in S | B_t = u \} \psi_t(u) \, du - \int_0^\infty P\{X_t \in S | B_t = u \} \psi(u) \, du \right| \\
\leq \max \left\{ \int_{\sigma_t} (\psi_t(u) - \psi(u)) \, du, \int_{\mathbb{R}_+ \setminus \sigma_t} (\psi(u) - \psi_t(u)) \, du \right\} \\
= \max \{|\Psi_t(\sigma_t) - \Psi(\sigma_t)|, |\Psi_t(\mathbb{R}_+ \setminus \sigma_t) - \Psi(\mathbb{R}_+ \setminus \sigma_t)|\} \\
\leq \sup_{\sigma \in \mathcal{P}(\mathbb{R})} |\Psi_t(\sigma) - \Psi(\sigma)|,
\]

so \( \|P_t - P\|_{TV} \leq \|\Psi_t - \Psi\|_{TV} \), and it is the end of the Proof. \( \blacksquare \)

Remark 6 The obtained bounds are not optimal for the following reasons.

1. The application of the coupling method to processes in continuous time usually gives a suboptimal estimates.

2. In the construction of the successful coupling only times of hitting of the process \( X_t \) to the idle state has been used. In addition, it can consider the times of the hitting of the process \( X_t \) to the idle state.

3. In the conditions \(^{[7]}\), it is impossible to give the accurate estimate for \( (C' \text{ and } C'_1) \). If the distributions or intensities are given more accurate, these values can be estimated more exactly.

However, the resulting estimates may be useful in practical applications, because the simulation is not always able to provide sufficient qualitative bounds. \( \triangleright \)

The author thanks L. G. Afanasyeva, A. Yu. Veretennikov and A. D. Manita for important and useful comments in the discussion of this work. The work is supported by RFBR (project No 17-01-00633 A).
References

[1] Afanasyeva, L. G., Tkachenko, A.V. On the Convergence Rate for Queueing and Reliability Models Described by Regenerative Processes // Journal of Mathematical Sciences, October 2016, Volume 218, Issue 2, pp 119–136.

[2] Asmussen, S. Applied Probability and Queues. Second edition. New York: Springer-Verlag, 2003.

[3] Golovastova, E. On convergence rate for an infinite-channel queuing system with Poisson input flow // arXiv:1711.07075 [math.PR]

[4] Griffeath, D. A maximal coupling for Markov chains // Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete — 1975 — Volume 31 — Issue 2, P. 95–106.

[5] Doeblin, W. Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d’états // Rev. Math. de l’Union Interbalkanique — 1938 — 2, P. 77–105.

[6] Erlang, A. K.: Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. Elektroteknikeren 13, 1917, 5-13 (1917) (in Danish); Engl. transl. P.O. Elect. Eng. J. 10, 189-197 (1918).

[7] Takács, L. Introduction to the Theory of Queues, (Oxford University Press, 1962)

[8] Stadje, W. The busy period of the queueing system $M|G|\infty$ // Journal of Applied Probability — 1985 — 22, P. 697–704. http://dx.doi.org/10.2307/3213872

[9] Fortet, R.: Calcul des probabilités. CNRS, Paris (1950)

[10] Kalashnikov, V.V. Mathematical Methods in Queueing Theory. Dordrecht: Kluwer Academic Publishers Group, 1994.

[11] Karlin, S., Taylor, H.M. A Second Course in Stochastic Processes, San Diego at al., 2nd Ed., Academic Press, 1975.

[12] Kato, K. Coupling Lemma and Its Application to The Security Analysis of Quantum Key Distribution // Tamagawa University Quantum ICT Research Institute Bulletin Vol.4 No.1 : 23-30 (2014) P.23-30.

[13] Thorisson, H. Coupling, Stationarity, and Regeneration. Springer, 2000.

[14] Veretennikov A., Butkovsky O.A. On asymptotics for Vaserstein coupling of Markov chains // Stochastic Processes and their Applications — 123(9) — 2013, P. 3518–3541.
[15] Veretennikov, A. On convergence rate for Erlang–Sevastyanov type models with infinitely many servers // Theory of Stochastic Processes. 2017. No. 1, 88-102; https://arxiv.org/abs/1412.3849

[16] Veretennikov, A. On Recurrence and Availability Factor for Single–Server System With General Arrivals, Reliability: Theory and Applications (RT&A), 2016, vol.11, No3(42), 49-58. (http://www.gnedenko-forum.org/Journal/2016-3.htm)

[17] Veretennikov, A. Yu. On the rate of convergence for infinite server Erlang–Sevastyanov’s problem // Queueing Systems, 2014, Volume 76, Issue 2, pp 181-203; DOI: 10.1007/s11134-013-9384-4

[18] Veretennikov, A.Yu., Zverkina, G.A. Simple Proof of Dynkin’s Formula for Single-Server Systems and Polynomial Convergence Rates // Markov Processes Relat. Fields 20 (2014), 479-504.

[19] B. V. Gnedenko,Yu. K. Belyayev,A. D. Solovyev, Mathematical Methods of Reliability Academic Press, 2014

[20] B.?V. Gnedenko, I.?N. Kovalenko, Introduction to queueing theory 2nd ed., rev. and supplemented. Boston : Birkhauser, 1989

[21] B. A. Sevast’yanov. An Ergodic Theorem for Markov Processes and Its Application to Telephone Systems with Refusals, Theory Probab. Appl., 2:1 (1957), 104–112

[22] B. A. Sevast’yanov, Erlang’s formulas in telephony with an arbitrary distribution law for the duration of the conversation, Teor. Veroyatnost. i Primenen., 2:1 (1957), 135

[23] D. Stoyan, Qualitative Eigenschaften und Abschätzungen stochastischer Modelle Berlin, 1977

[24] Borovkov A.A., Asymptotic methods in queuing theory. John Wiley & Sons, Ltd., Chichester, 1984.
Об одной обобщённой системе Эрланга-Севастьянова и её скорости сходимости.

**Г. А. ЗВЕРКИНА**

Российский университет транспорта (МИИТ),
Институт проблем управления
им. В.А. Трапезникова РАН
zverkina@gmail.com

УДК 519.21

Ключевые слова: распределение состояния системы массового обслуживания, скорость сходимости, регенерирующие процессы, метод склеивания.

**Аннотация**

Получена оценка сверху для скорости сходимости распределения состояния бесконечнолинейной системы массового обслуживания в случае, когда интенсивность входящего потока и интенсивность обслуживания зависят от состояния системы.

**Abstract**

G. A. Zverkina. About some extended Erlang-Sevastyanov queueing system and its convergence rate. Fundamentalnaya i prikladnaya matematika, vol. **, no. *, pp. ** - **

The upper bound for the convergence rate of the distribution of the state of a queuing system with infinitely many servers is obtained, in the case when the intensity of the incoming flow and the intensity of the service depend on the state of the system.

1 Введение.

1.1 Мотивация

Рассматривается бесконечнолинейная система массового обслуживания (СМО), в которую в моменты времени $t_1, t_2, \ldots, t_n$ поступают заявки, которые сразу же начинают обслуживаться (полагаем $t_0 \overset{\text{def}}{=} 0$). Время обслуживания $i$-й заявки – случайная величина (сл.в.) $\xi_i$ с функцией распределения (ф.р.) $G_i(s)$, а промежутки между поступлениями требований $\eta_i \overset{\text{def}}{=} t_i - t_{i-1}$ – сл.в. с ф.р. $F_i(s)$. Для простоты предположим, что распределения сл.в. $\xi_i$ и $\eta_i$ абсолютно непрерывны; обозначим $f_i(s) \overset{\text{def}}{=} F_i'(s), g_i(s) \overset{\text{def}}{=} G_i'(s)$.

Мы будем описывать распределения $F_i(s)$ и $G_i(s)$ в терминах интенсивностей.

**Определение 1** Интенсивностью в момент времени $s > 0$ окончания временного периода $\mathcal{T}$, начавшегося в момент времени $t = 0$ и не завершившегося до момента времени $s > 0$, называется $\phi(s) \overset{\text{def}}{=} \lim_{\Delta \to 0} \frac{\mathbb{P}\{\mathcal{T} \in (s, s + \Delta) | \mathcal{T} > s\}}{\Delta}$.

Если $\mathcal{T}$ имеет абсолютно непрерывную ф.р. $\Phi(x)$, то...
\[ \phi(s) = \lim_{\Delta \to 0} \frac{\Phi(s + \Delta) - \Phi(s)}{\Delta(1 - \Phi(s))} = \Phi'(s + 0) \frac{1}{1 - \Phi(s)}. \]

Замечание 1 Из Определения 7 следует, что \( P\{T \in (s, s + \Delta) | T > s\} = \phi(s)\Delta + o(\Delta) \).

Замечание 2 Функция распределения и её плотность легко восстанавливаются по интенсивности:

\[ \Phi(s) = 1 - \exp\left( - \int_0^s \phi(u) \, du \right); \]

(1)

\[ \Phi'(s) = \phi(s) \exp\left( - \int_0^s \phi(u) \, du \right). \]

(2)

Очевидно, постоянная интенсивность соответствует экспоненциальному распределению. Интенсивность можно также рассмотривать и в случае, когда функция \( \Phi(s) \) имеет разрывы. Но в этом случае формулы (1) и (2) неприемлемы.

Итак, мы описываем поведение бесконечнолинейной СМО на языке интенсивностей, а именно.

1. Интенсивность поступления очередной заявки в момент времени \( t \) есть \( \lambda(t) \), т.е.

\[ P\left\{ \text{Очередная заявка поступит в интервале \((t, t + \Delta]\)} \right\} = \lambda(t)\Delta + o(\Delta). \]

2. Интенсивность обслуживания \( i \)-й по порядку поступления заявки из имеющихся \( n \) заявок в системе в момент времени \( t \) равна \( h_i(t) \), т.е.

\[ P\left\{ i\text{-е требование, обслуживающееся в момент времени } t, \text{ закончит обслуживание на промежутке \((t, t + \Delta]\)} \right\} =
\]

\[ = P\{\xi_i \in (s, s + \Delta) | \xi_i > s\} = h_i(t)\Delta + o(\Delta). \]

Здесь и везде далее под \( i \)-м требованием, находящимся в СМО, мы понимаем \( i \)-е по порядку поступления требование из находящихся в СМО.

В случае, когда \( \lambda(t) \equiv \lambda \) и \( h_i(t) \equiv \mu \), мы имеем дело с СМО \( M|M|\infty \), впервые изученной в [3]; было показано, что при условии \( \mu < \lambda \) распределение количества заявок в системе \( n_t \) сходится к стационарному распределению:

\[ \lim_{t \to \infty} P\{n_t = k\} = \mathcal{P}_k = \mathcal{P}_0 \frac{\lambda^k}{\mu^k}, \text{ где } \mathcal{P}_0 = 1 - \frac{\lambda}{\mu}. \]
Позднее этот результат был обобщен на случай, когда интенсивность входящего потока зависит только от числа требований в системе, т.е. $\lambda(t) \equiv \lambda_k$ при $n_t = k$ (см., например, [11, Ch.4, §4-5]):

$$
\lim_{t \to \infty} P\{n_t = k\} = \mathcal{P}_k = \frac{\prod_{i=0}^{k-1} \lambda_i}{\mu^k}, \quad \text{где} \quad \mathcal{P}_0 = \left(\sum_{k=0}^{\infty} \prod_{i=0}^{k-1} \lambda_i \right)^{-1}. \quad (3)
$$

Более того, если интенсивность обслуживания всех находящихся в СМО заявок зависит только от количества заявок, т.е. $h_i(t) \equiv \mu_k > 0$ при $n_t = k$, формула [3] применима, но с заменой $\mu^k$ на $\prod_{i=1}^{k} \mu_i$. Очевидно, если $\lambda_k = 0$ при некотором $k$, то [3] применима для системы с конечным числом серверов без очереди.

В описанном случае, когда

$$
\lambda(t) \neq \lambda, \quad \mu(t) \neq \mu, \quad h_i(t) = \mu_{n_t},
$$

где $n_t$ — количество требований в системе в момент времени $t$, времена обслуживания и входящих поток зависят между собой.

В [9] исследовалась СМО с пуассоновским входящим потоком и произвольным абсолютнопрерывным распределением времени обслуживания ($\text{СМО } M|G|\infty$), одинаковым для всех требований; предполагалось, что времена обслуживания и входящих поток независимы. В терминах интенсивностей это означает, что если в момент времени $t$ в системе находится $n_t$ требований, и $i$-е требование находилось на обслуживании в течение времени $x^{(i)}_t$ ($i = 1, 2, \ldots, n$), то интенсивность обслуживания $i$-го требования в момент времени $t$ равна $h_i(t) = h\left(x^{(i)}_t\right)$.

Обозначим $\mu \equiv \left(\int_0^\infty s \, dG(s)\right)^{-1}$, где $G(s) = 1 - \exp\left(-\int_0^s h(u) \, du\right)$ — ф.р. времени обслуживания. В [9] было высказано утверждение, что для $k > 0$

$$
\lim_{t \to \infty} P\{n_t = k; x^{(1)}_t < \alpha_1, x^{(2)}_t < \alpha_2, \ldots x^{(k)}_t < \alpha_k\} = \mathcal{P}(k; \alpha_1, \alpha_2, \ldots, \alpha_k) = \mathcal{P}_0 \prod_{i=1}^{k} \lambda_i G(1 - \alpha_i),
$$

где нормирующий множитель $\mathcal{P}_0$ вычисляется так же, как в [3].
Этот факт для СМО \( M|G|m(0) \) в более общем виде, без предположения об абсолютной непрерывности ф.р. времени обслуживания \( G(s) \), был доказан в \[21\] (см. также \[22\]).

То есть в \[9\] было предложено рассматривать СМО, в которой

\[
\lambda(t) \equiv \lambda, \quad h_i(t) = h_i \left( x_i^{(i)} \right), \quad (5)
\]

где \( x_i^{(i)} \) – пропущенное к моменту времени \( t \) время обслуживания \( i \)-го по порядку требования из \( n_t \) требований, находящихся в системе.

В дальнейшем результаты, основанные на \[6\], были получены многими исследователями (см., например, \[10, 2, 24, 7\]).

Естественным обобщением ситуаций \( 4 \) и \( 5 \) является СМО, в которой интенсивности входящего потока и обслуживания зависят не только от \( n_t \) и прошедших времён обслуживания находящихся на обслуживании требований, но и от того, когда произошло последнее изменение \( n_t \) – количества требований в системе:

\[
\lambda(t) = \lambda \left( x_t^{(0)} : x_t^{(1)} , x_t^{(2)} , \ldots , x_t^{(n_t)} \right) , \quad h_i(t) = h_i \left( x_t^{(0)} : x_t^{(1)} , x_t^{(2)} , \ldots , x_t^{(n_t)} \right) ,
\]

где \( x_t^{(0)} \) – время, пропущенное с последнего изменения количества требований в системе, а \( x_t^{(i)} \) – прошедшее к моменту времени \( t \) время обслуживания \( i \)-го по порядку поступления требования из \( n_t \) требований, находящихся в системе.

Знание стационарного распределения состояния СМО чрезвычайно важно для приложений ТМО, на основании этих распределений рассчитываются технические характеристики сложных систем.

Такого рода СМО были рассмотрены, например, в \[15, 16, 18\].

Например, \[18\] был доказан следующий факт: для СМО с одним обслуживающим прибором и бесконечной очередью, в случае, когда интенсивность обслуживания \( h \) и интенсивность входящего потока \( \lambda \) зависят от количества заявок в системе \( n_t \) и прошедшего времени обслуживания \( x_t \), верна следующая

**Теорема 1** Обозначим \( X_t \equiv (n_t, x_t) \), где \( n_t \) – число находящихся в СМО требований в момент времени \( t \), \( x_t \) – время, в течение которого находилось на обслуживании требование, которое в момент времени \( t \) находится на обслуживании. В случае \( n_t = 0 \) интенсивность входящего потока постоянна и равна \( \lambda_0 > 0 \) и \( X_t \equiv (0,0) \).

Если \( \lambda(n_t, x_t) \) и \( h(n_t, x_t) \) измеримы по Борелю и ограничены, выполнены условия

\[
\sup_{n,x: n>0} \lambda(n,x) = \Lambda < \infty; \quad \inf_{n>0, x>0} h(n,x) \geq \frac{C_0}{1 + x},
\]

\[
C_0 > 4(1 + 2\Lambda), \quad \text{и для некоторого} \quad k > 1 \quad C_0 > 2^{k+1}(1 + \Lambda 2^k),
\]

то распределение процесса \( X_t = (n_t, x_t) \) сходится к стационарному распределению \( \mathcal{P} \) на пространстве состояний \( \mathcal{X} = \{(0,0)\} \cup \{(n,x): n \in \mathbb{N}, x \geq 0\} \).
и найдутся такое \( m > k \) и \( C > 0 \), что при всех \( t \geq 0 \)

\[
\sup_{S \in \mathcal{S}(x)} |P\{X_t \in S|X_0 = (n,x)\} - \mathcal{P}(S)| \leq C \frac{(1 + n + x)^m}{(1 + x)^k},
\]

где \( \mathcal{P} \) – стационарное распределение процесса \( X_t \).

В доказательстве этой Теоремы приводится (неполный) алгоритм вычисления константы \( C \).

Однако существует достаточно большое количество результатов о порядке скорости сходимости распределения состояния СМО к стационарному распределению (см., например, [24, 20, 13, 1] и др.). В этих работах доказано, что некоторые характеристики СМО \( \chi_t \) сходятся к стационарному значению \( \tilde{\chi} \) со скоростью \( \phi(t) \), т.е. существует такая константа \( K \), что при всех \( t \) выполняется

\[
|\chi_t - \tilde{\chi}| < K \phi(t),
\]

где \( \phi(t) \to 0 \).

Так, например, в СМО, исследованной в [18] (Теорема 1), для сходимости распределения процесса \( X_t \) к стационарному распределению со скоростью не превосходящей

\[
\frac{K}{(1 + x)^k}, k \in [0, C_0 - 1),
\]

достаточно \( \Lambda > C_0 \), но без дополнительных условий Теоремы 1 оценить константу \( K \) ранее не удавалось.

Отметим, что для СМО \( MI|GI|\infty \) хорошо известны (см., например, [7]) такие характеристики, как \( \mathcal{P}_k(t) \) – вероятность того, что в системе занято \( k \) приборов в момент времени \( t \), известны средняя длина периода занятости и его дисперсия, а также преобразование Лапласа ф.р. периода занятости (если в начальный момент времени система свободна).

Соответственно, эти результаты позволяют легко оценить скорость сходимости величин \( \mathcal{P}_k(t) \) к их стационарным значениям \( \mathcal{P}_k \) – см., например, [3].

1.2 Известные факты и цель работы

Напомним, что СМО \( MI|GI|\infty \) состоит из бесконечного числа приборов, работающих независимо, и случайное время обслуживания попавшего на прибор требования (заявки) одинаково распределено для всех приборов и имеет ф.р. \( G(s) \); входящий поток пуссоновский с постоянным параметром \( \Lambda \). Предполагается, что в начальный момент времени система свобода (нет требований на обслуживании).

Цель настоящей работы – исследование поведения бесконечно цикличной СМО, в которой входящий поток и времена обслуживания задаются интенсивностями, зависящими от полного состояния системы; при этом все времена обслуживания находящиеся в системе требований некоторым образом зависят между собой, а также они зависят и от входящего потока требований, который может менять свою интенсивность в зависимости от полного состояния системы. При этом естественно предполагать, что, несмотря на такую зависимость, некоторые характеристики времени обслуживания и входящего потока находятся в определённых границах (например, естественно требовать, чтобы
интенсивность входящего потока была ограничена, а интенсивность обслуживания не обращалась в нуль.

В работе исследуется ситуация, когда интенсивность входящего потока, а также интенсивность обслуживания зависят от полного состояния системы, а именно – от того времени, которое прошло с момента поступления последнего пришедшего требования и от того времени, которое имеющиеся в системе требования находились на обслуживании.

Как уже говорилось, несмотря на наличие указанной зависимости случайных величин (времена между поступлениями требований и времени обслуживания требований), мы предполагаем некоторые ограничения для интенсивности поступления требований в систему и интенсивностей их обслуживания.

Предложенный выше вариант зависимости времён обслуживания и входящего потока позволяют утверждать, что процесс обслуживания такой системы – марковский и регенерирующий. Точками регенерации такого процесса являются моменты времени, когда все входящие в полное состояние системы величины обнуляются. Это может произойти тогда, когда в свободную систему приходит требование, т.е. полное состояние системы содержит нулевое время обслуживания единственного требования, и нулевое время ожидания поступления следующего требования.

2 Описание изучаемой СМО

Итак, мы рассматриваем систему массового обслуживания, в которой интенсивность входящего потока и интенсивность обслуживания зависят от полного состояния системы, описание которого будет дано ниже.

Это полное состояние системы в момент времени \( t \) включает: время \( x_t^{(0)} \), прошедшее с момента прихода последней заявки, и времена \( x_t^{(i)} \), прошедшие с момента начала обслуживания \( i \)-го требования из общего количества находящихся (требования нумеруются в соответствии с порядком их поступления); также для удобства мы будем использовать переменную \( n_t \) – число заявок в системе в момент времени \( t \). Таким образом, полное состояние системы описывается вектором \( X_t = (n_t, x_t^{(0)}, x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(n_t)}) \) – поскольку \( i \)-я заявка поступила ранее \((i + 1)\)-й, ясно, что \( x_t^{(i)} \geq x_t^{(i+1)} \) для \( i = 1, \ldots, n_t \). Легко видеть, что \( x_t^{(0)} \leq x_t^{(n_t)} \), то есть заявка с номером \( n_t \) появилась в системе не позднее времени \( t - x_t^{(0)} \).

Мы предполагаем, что интенсивность входящего потока есть функция от \( X_t \): \( \lambda = \lambda(X_t) \) – так же, как и интенсивность обслуживания \( i \)-й заявки (которая уже обслуживалась в течение времени \( x_t^{(i)} \)) – это функция \( h_i = h_i(X_t) \). Это значит, что для малого промежутка времени \( \Delta > 0 \) при заданном \( X_t = \)
\[(n_t, x_t^{(0)}; x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(n_t)})\] — напомним, что \(x_t^{(i)} \geq x_t^{(i+1)}\), — верно следующее:

\[
P \left\{ \begin{array}{l}
  n_{t+\Delta} = n_t + 1, \quad x_t^{(n_t)} = x_t^{(n_t+\Delta)} \in (0; \Delta), \\
  x_t^{(i)} = x_t^{(i+\Delta)} \quad \text{для всех } i = 1, \ldots, n_t \\
  x_t^{(n_t+\Delta)} \in (0; \Delta)
\end{array} \right. \}

= \lambda(X_t)\Delta + o(\Delta)

P\{n_{t+\Delta} = n_t - 1\} = \sum_{i=1}^{n_t} h_i(X_t)\Delta + o(\Delta);

1. \text{для всех } j < i, \quad x_t^{(j)} = x_t^{(j_i)} + \Delta,
2. \text{для всех } j > i, \quad x_t^{(j)} = x_t^{(j_i)} + \Delta

P\{n_{t+\Delta} = n_t; x_t^{(i)} = x_t^{(i)} + \Delta \quad \text{для всех } i = 0, \ldots, n_t\} =

1 - \left(\lambda(X_t) + \sum_{i=1}^{n_t} h_i(X_t)\right)\Delta + o(\Delta).

Из (6) следует, что распределение \(X_{t+\Delta}\) зависит только от распределения \(X_t\), т.е. \(X_t\) — марковский процесс на пространстве состояний \(\mathcal{X} \overset{\text{def}}{=} \bigcup_{i=0}^{\infty} \mathcal{F}_i\), где

\(\mathcal{F}_i = \{i\} \times \prod_{j=0}^{i} \mathbb{R}_+, \quad i \in \mathbb{Z}_+\); множество \(\mathcal{F}_0\) — это множество свободных состояний СМО (когда заявки в системе отсутствуют).

Обозначаем \(\mathcal{X}_0\) — распределение процесса \(X_t\) с начальным состоянием \(X_0\):

\(\mathcal{X}_0(S) = P\{X_t \in S | X_0 = x\}, \quad x \in \mathcal{F}, \quad S \in \mathcal{B}(\mathcal{F})\).

Как уже говорилось, моменты попадания процесса \(X_t\) в состояние \((1, 0; 0), \text{т.е. выхода процесса } X_t \text{ из множества } \mathcal{F}_0 \text{ (в свободную систему поступило требование)} \text{ являются моментами регенерации процесса } X_t\). Поэтому, если время между моментами регенерации имеет конечное математическое ожидание, то процесс \(X_t\) — эргодический, и его распределение слabo сходится к некоторому стационарному (инварианному) распределению.

Замечание 3. Предложенным способом может быть описан широкий класс СМО, например, рассмотренные выше СМО \(M|M|\infty, M|G|\infty, M|G|n|0\), и другие.

Для того, чтобы среднее периода регенерации хорошо изученной СМО \(M|G|\infty\) было конечным, достаточно, чтобы среднее времени обслуживания было конечным. При этом скорость сходимости величин \(\mathcal{P}_k(t)\) к их стационарным значениям \(\mathcal{P}_k\) тем выше, чем больше конечных моментов у распределения времени обслуживания — см., например, [3].

Естественно поэтому рассматривать ситуацию, когда (занятые) времена обслуживания имеют конечные моменты, а характеристики входящего потока могут быть сравнены с характеристиками пуассоновского потока.
Наложим на интенсивности следующие условия:

\[ h_i(X_t) \geq \frac{K}{1 + x_i(t)}, \quad K > 2; \quad 0 < \lambda_0 \leq \lambda(X_t) \leq \Lambda < \infty. \quad (7) \]

Из первого условия следует, что ф.р. времени обслуживания \( G_i(s) \) удовлетворяет неравенству \( G_i(s) \geq 1 - \frac{1}{(1 + s)^K} \), т.e. время обслуживания имеет \( k \in [0; K) \) конечных моментов (см. Замечание 2).

Также плотность ф.р. интервалов времени поступления заявок \( F(s) \) удовлетворяет неравенству \( f_i(s) \leq \Lambda e^{-\lambda_0 s} \), и распределение времён \( \eta \) между поступлениями заявок имеет любые степенныё моменты: \( \mathbb{E}(\eta_i)^k \leq \frac{k!\Lambda}{\lambda_0^{k+1}} \), в частности, эта оценка верна для свободных периодов \( X_t \), т.e. промежутков времени, где \( n_t \equiv 0 \).

Наша цель – найти вычислимые оценки скорости сходимости распределения процесса \( X_t \) к стационарному распределению в условиях (7).

Для этого мы воспользуемся методом склеивания – см. Раздел 5.

Однако сначала мы напомним некоторые сведения из теории массового обслуживания и получим необходимые в дальнейшем оценки.

3 Некоторые сведения о системе \( M|GI|\infty \)

Здесь мы приведём известные результаты о системе \( M|GI|\infty \) (см. [7], [8] и др.).

В систему \( M|GI|\infty \) поступает пуассоновский поток заявок с постоянной интенсивностью \( \Lambda \). Пришедшая заявка сразу начинает обслуживание. Её время обслуживания является случайной величиной (обозначаемой \( \xi \)) с ф.р. \( G(s) \).

Все случайные величины (времена обслуживания и времена между поступлениями заявок) независимы в совокупности.

В начальный момент времени СМО свободна.

Предполагается, что существует некоторое количество конечных моментов

сл.в. \( X_i: \mathbb{E}(\xi_i)^k = k \int_0^\infty s^{k-1}(1 - G(s)) \, ds = m_k < \infty; \quad K \geq 1. \) В дальнейшем положим

\[ G(s) = 1 - \frac{1}{(1 + s)^K}, \quad K > 2. \quad (8) \]

Условие (8) гарантирует наличие \( K - \varepsilon \) (\( \varepsilon > 0 \)) конечных моментов; очевидно, что \( \mathbb{E} \xi_i = \frac{1}{K - 1} \); \( \mathbb{E} (\xi_i)^2 = \frac{2}{K^2 - 3K + 2} \) и т.д.

Нагрузкой СМО назовём величину \( \rho \overset{\text{def}}{=} \Lambda m_1 \).
Опять описываем поведение процесса обслуживания марковским процессом \( \bar{X}_t = \left( \bar{n}_t, \bar{x}_t^{(0)} ; \bar{x}_t, \bar{x}_t^{(1)} , \ldots , \bar{x}_t^{(n_t)} \right) \), где \( \bar{n}_t \) — число требований в СМО в момент времени \( t \), а \( \bar{x}_t^{(i)} \) — время, пройденное с момента поступления последнего требования; \( \bar{x}_t \) — время, в течение которого \( i \)-я по порядку поступления заявки из находящихся в системе (\( i = 1, 2, \ldots , n_t \)) находилась на обслуживании. (Здесь и везде в дальнейшем знак \( \bar{\circ} \) над случайной величиной, процессом или их числовыми характеристиками обозначает, что речь идёт о "классической" СМО \( MI|G|\infty \).)

Распределение процесса \( \bar{X}_t \) обозначаем \( \bar{\mathcal{P}}_t \): \( \bar{\mathcal{P}}_t (S) = P\{ \bar{X}_t \in S \} \) для всех \( S \in \mathcal{B}(\mathbb{X}) \), и снова \( \mathcal{X} \overset{\text{def}}{=} = \bigcup_{i=0}^{\infty} \mathcal{A}_i \), и \( \mathcal{R}_i \overset{\text{def}}{=} \{ i \} \times \prod_{j=0}^{i} \mathbb{R}_+ , i \in \mathbb{Z}_+ \).

В этих предположениях известно (см., например, [7]), что \( P_{k}(t) \overset{\text{def}}{=} P\{ \bar{n}_t = k \} = \mathcal{P}_t (\mathcal{F}_0) = e^{-\Phi(t)} \left( \Phi(t) \right)^k \frac{1}{k!} \), где \( \Phi(t) \overset{\text{def}}{=} \Lambda \int_0^t (1 - G(s)) \, ds \).

Замечание 4 \( \Phi(\infty) = \Lambda m_1 \), и \( P_{0}(t) \geq e^{-\Lambda m_1} = e^{-\rho} \) при \( t \geq 0 \).

Кроме этого, известно распределение периода занятости СМО \( M|G|\infty \) (см., например, [8]). Обозначим \( \bar{\zeta}_i \) — \( i \)-й период занятости СМО \( M|G|\infty \). Эта случайная величина имеет ф.р. \( B(x) \overset{\text{def}}{=} P\{ \bar{\zeta}_i \leq x \} = 1 - \frac{1}{\Lambda} \sum_{k=1}^{\infty} c^{n*}(x) \), где \( c(x) \overset{\text{def}}{=} \Lambda (1 - G(x)) e^{-\Phi(x)} \), а \( c^{n*} \) — это \( n \)-я свёртка функции \( c(x) \).

Также можно выписать преобразование Лапласа для \( B(x) \) (см., например, [8]):

\[
L[B](s) = 1 + \frac{s}{\Lambda} - \frac{1}{\Lambda} \int_0^\infty \exp \left( -st - \Lambda \int_0^t [1 - G(v)] \, dv \right) \, dt. \tag{9}
\]

Впрочем, моменты периода занятости можно вычислять и по формулам

\[
\mathbb{E} (\bar{\zeta}_i)^n = (-1)^{n+1} \left\{ \frac{e^{\rho}}{\Lambda} n C^{(n-1)} - e^{\rho} \sum_{k=1}^{n-1} C_n^{i-k} \mathbb{E} (\bar{\zeta}_i)^{n-k} C^{(k)} \right\}, \quad n \in \mathbb{N}, \tag{10}
\]

где \( C^{(k)} = \int_0^\infty (-t)^k \left( \exp \left( -\Lambda \int_0^\infty [1 - G_0(v)] \, dv \right) \right) \Lambda [1 - G_0(t)] \, dt \).
Используя (9) и (10), можно получить значения
\[ E \circ \zeta_i = e^\rho - 1 \quad \text{и} \quad E (\zeta_i)^2 = \]
\[ \frac{2e^\rho}{\Lambda} \int_0^\infty \left( \exp \left( -\Lambda \int_0^t [1 - G(v)] \, dv \right) \right) \, dt + \frac{2e^\rho}{\Lambda^2} - \frac{2e^\rho - 1}{\Lambda^2}. \] (11)

Но для более высоких моментов \( \zeta \), вычисления по формулам (9) и (10) очень сложны и трудно применимы на практике.
Однако нам в дальнейшем потребуются оценки моментов \( \zeta \). Для этого используем

Утверждение 1 Если \( \mathbb{E} \xi_k < \infty \) при некотором \( k \in \mathbb{N} \), то верна оценка
\[ E (\zeta_i)^k \leq \frac{\mathbb{E} \xi_k}{1 - e^{-\rho}} \times \varphi((1 - e^{-\rho}), k - 1), \] (12)
где \( \varphi(x, k) \overset{\text{def}}{=} \sum_{n=1}^\infty n^{k-1}x^n = \left( x \frac{d}{dx} \right)^k \frac{1}{1 - x}. \)

Доказательство. Функция \( c(x) \), используемая в (9), неотрицательна, и
\[ \int_0^\infty c(x) \, dx = \int_0^\infty \Lambda (1 - G(x)) \exp \left( -\Lambda \int_0^x (1 - G(s)) \, ds \right) \, dx = \]
\[ = \left( 1 - \exp \left( -\Lambda \int_0^\infty (1 - G(s)) \, ds \right) \right) = \left( 1 - e^{-\Lambda m_1} \right) = (1 - e^{-\rho}) \overset{\text{def}}{=} \varrho, \]
где опять \( \rho = \Lambda m_1 \).
Т.е. функция \( c(x) \overset{\text{def}}{=} \varrho^{-1} c(x) \) является плотностью распределения некой сл.в. \( \vartheta \), а \( \varsigma^{n*}(x) \) – это плотность распределения \( \sum_{j=1}^n \vartheta_j \) – суммы \( n \) одинаково распределенных и независимых сл.в. с тем же распределением, что и \( \vartheta \).
Поэтому \[ \int_0^\infty \varsigma^{n*}(x) \, dx = 1. \] Следовательно,
\[ \mathbb{E} \circ \zeta_i = \int_0^\infty (1 - B(x)) \, dx = \frac{1}{\Lambda} \int_0^\infty e^{\rho x} (x) \, dx = \]
\[ = \frac{1}{\Lambda} \sum_{n=1}^\infty \vartheta^n \int_0^\infty \varsigma^{n*}(x) \, dx = \frac{1}{\Lambda} \sum_{n=1}^\infty \vartheta^n = \frac{1}{\Lambda} \times \frac{\vartheta}{1 - \vartheta} = \frac{e^\rho - 1}{\Lambda}. \]
Далее, если $m_{k+1} = \mathbb{E}_k^{k+1} < \infty$ при некотором $k \in \mathbb{N}$, то

$$
\mathbb{E} \varphi^k = \int_0^\infty x^k \nu(x) \, dx = \int_0^\infty x^k e^{-x} \Lambda(1 - G(x)) e^{-\varphi(x)} \, dx \leq \frac{\Lambda}{\varphi} \int_0^\infty x^k (1 - G(x)) \, dx = \frac{\Lambda \mathbb{E}_k^{k+1}}{\varphi(k + 1)},
$$

и

$$
\int_0^\infty x^k n^*(x) \, dx = \mathbb{E} \left( \sum_{j=1}^n \varphi_j \right)^k \leq n^k \mathbb{E} \varphi^k \leq \frac{n^k \Lambda m_{k+1}}{\varphi(k + 1)}.
$$

В последней оценке мы использовали неравенство Йенсена (для $k \geq 1$) в таком виде:

$$(a_1 + \ldots + a_n)^k \leq n^{k-1} (a_1^k + \ldots + a_n^k). \quad (13)$$

Теперь оценим $\mathbb{E} (\xi)^k$:

$$
\mathbb{E} (\xi)^k = \int_0^\infty k x^{k-1} (1 - B(x)) \, dx = \frac{1}{\Lambda} \sum_{n=1}^\infty \int_0^\infty k x^{k-1} e^{n^*(x)} \, dx = \frac{1}{\Lambda} \sum_{n=1}^\infty \varphi^n \mathbb{E} \left( \sum_{j=1}^n \varphi_j \right)^{k-1} \leq \frac{1}{\Lambda} \sum_{n=1}^\infty \varphi^n \Lambda \mathbb{E}_n^{k} \frac{\mathbb{E}_n^k}{\varphi} = \frac{\mathbb{E}_n^k}{\varphi} \sum_{n=1}^\infty n^{k-1} \varphi^n = \frac{\mathbb{E}_n^k}{\varphi} \times \varphi(\varphi, k - 1).
$$

Предложение доказано.

4 Откычены для процесса $X_t$

Вернёмся к исследуемому нам процессу $X_t$, описанному в Разделе 2. Предполагаем, что в начальный момент времени $X_0 = (0, 0)$. Т.е. система свободна и начинается отсчётом времени до поступления первого требования. В момент поступления первого требования процесс $X_t$ попадает в состояние $(1, 0; 0)$ — это момент регенерации.

Напомним, что описанный в разделе процесс $\hat{X}_t$ также стартует из этого состояния: $\hat{X}_0 = (0, 0)$. В момент поступления первого требования начинается первый период регенерации.

Определение 2 Случайная величина $\eta$ не превосходит случайную величину $\theta$ по распределению, если при всех $s \in \mathbb{R}$ верно неравенство $F_\eta(s) = P(\eta \leq \theta) \leq P(\theta < \eta)$. 

11
случайное событие, что ф.р. сл.в. θ не превосходит ф.р. сл.в. η. Это — отношение порядка, будем обозначать его η < θ.

Утверждение 2 η < θ тогда и только тогда, когда на некотором вероятностном пространстве (Ω, ℜ, P) можно построить сл.в. η' и θ' такие, что η' ⩽ θ' при всех ω ∈ Ω, и P{η' ⩽ s} = P{θ' ⩽ s} и P{θ ⩽ s} = P{θ' ⩽ s} для всех s ∈ R.

Доказательство. См. [23]. ■

Утверждение 3 Если $X_0 = \tilde{X}_0 = (0,0)$, и выполнено условие (2) (напомним, что входящий поток для $\tilde{X}_t$ — пуссоновский и имеет интенсивность Λ, а время обслуживания имеет ф.р. (3)), то при т ⩾ 0 выполнено соотношение:

\[ n_t < \tilde{n}_t. \] (14)

Доказательство. Для доказательства используется метод одного вероятностного пространства, который заключается в том, что на одном вероятностном пространстве строится процесс $X_t$ и $\tilde{X}_t$ (а точнее говоря, их версии) таким образом, что на всех траекториях процессов выполнено неравенство $n_t < \tilde{n}_t$ — при всех т ⩾ 0.

Эта конструкция возможна по двум причинам.

I. Входящий поток для процесса $X_t$ имеет интенсивность, не превосходящую интенсивность входящего потока для $\tilde{X}_t$ ($\lambda(X_t) \leq\Lambda$ из условия (7)).

Поскольку можно строить на одном вероятностном пространстве два входящих потока для обоих процессов: поток $\tilde{X}_t$ с интенсивностью $\lambda(X_t)$ для процесса $X_t$, и поток $\tilde{X}_t$ с интенсивностью $\Lambda - \lambda(X_t)$; объединение потоков $\tilde{X}_t$ и $\tilde{X}_t$ — это поток с интенсивностью $\Lambda$ для процесса $\tilde{X}_t$.

Таким образом, в каждый момент времени t число заявок, поступивших в СМО с процессом $X_t$ не превосходит число заявок, поступивших в СМО с процессом $\tilde{X}_t$. Это — отправная точка для построения таких заявок "общими".

Иначе говоря, в каждый момент времени вектор $\left(\tilde{x}_t^{(1)}, \tilde{x}_t^{(2)}, \ldots, \tilde{x}_t^{(n_t)}\right)$ состоит из компонент вектора $\left(\tilde{x}_t^{(1)}, \tilde{x}_t^{(2)}, \ldots, \tilde{x}_t^{(n_t)}\right)$, т.е. $\left(\tilde{x}_t^{(1)}, \tilde{x}_t^{(2)}, \ldots, \tilde{x}_t^{(n_t)}\right)$ состоит из обслуженных времени находящихся на обслуживании требований, совпадающих с некоторыми из элементов вектора $\left(\tilde{x}_t^{(1)}, \tilde{x}_t^{(2)}, \ldots, \tilde{x}_t^{(n_t)}\right)$. 12
II. Интенсивность обслуживания $i$-го требования СМО, описываемой процессом $\tilde{X}_t = (\tilde{n}_t, \tilde{x}_t^{(0)}; \tilde{x}_t^{(1)}, \tilde{x}_t^{(2)}, \ldots, \tilde{x}_t^{(\tilde{n}_t)})$, равна $h_i (\tilde{X}_t) = \frac{G'(\tilde{x}_t^{(i)})}{1 - G(\tilde{x}_t^{(1)})} = \frac{K}{1 + \tilde{x}_t^{(i)}}$.

Таким образом, интенсивность обслуживания “общих” заявок у процесса $\tilde{X}_t$ не превосходит интенсивность обслуживания этих же заявок у процесса $X_t$, и это – второе обстоятельство, позволяющее применить метод одного вероятностного пространства.

Учитывая, что время обслуживания имеет ф.р., которую можно восстановить по интенсивности обслуживания (см. (1)–(2)), получаем, что остаточное время обслуживания “общих” заявок у процесса $\tilde{X}_t$ не превосходит по распределению остаточное время обслуживания этих заявок у процесса $X_t$. Поэтому можно построить на одном вероятностном пространстве эти случайные величины так, что “общие” заявки завершат обслуживание у процесса $X_t$ не позднее, чем у процесса $\tilde{X}_t$ – см. Предложение 2.

Стало быть, в СМО с процессом $\tilde{X}_t$ поступает больший поток заявок, чем в СМО с процессом $X_t$, и одновременно поступившие (общие) заявки покидают СМО с процессом $X_t$ раньше. Из этого следует утверждение предложения 3.

Следствие 1 Из (14) следует, что свойство упорядоченности по распределению верно не только для $n_t$ и $\tilde{n}_t$, но и для периодов занятости процессов $X_t$ и $\tilde{X}_t$. Учитывая, что процессы $X_t$ и $\tilde{X}_t$ регенерирующие (напомним, моментом регенерации является момент, когда процесс принимает значение $(1, 0; 0)$, т.е. момент начала периода занятости), распределение периодов занятости $\zeta_i$ процесса $X_t$ и периодов занятости $\zeta_i$ процесса $\tilde{X}_t$ не зависят от номера $i$. То есть $\zeta_i \prec \zeta_i$, и для всех $k > 0$ верно неравенство $E(\zeta_i)^k \leq E(\zeta_i)^k$.

То есть оценки (13) верны и для $\tilde{\zeta}_i$.

Следствие 2 Из Замечания 4 и (14) следует, что при всех $t \geq 0$ (и $X_0 = (0, 0)$) верно неравенство

$$P\{X_t \in \mathcal{X}_0\} = P\{n_t = 0\} \geq P_0(t) \geq e^{-\rho}. \quad (15)$$

5 Метод склеивания

Первоначально метод склеивания марковских процессов был применён к цепям Маркова с конечным числом состояний (см. [5], [13]). Основная идея этого...
метода состоит в следующем.

Пусть два однородных марковских процесса $Y_t$ и $Z_t$ с одинаковой переходной функцией в начальный момент времени имеют различные состояния; они заданы на $(\Omega, \mathcal{F}, \mathbb{P})$, и измеримы относительно фильтрации $\mathcal{F}_t$.

Обозначим $\tau = \tau(Y_0, Z_0) \overset{\text{def}}{=} \inf\{t > 0 : Y_t = Z_t\}$. То есть $Y_\tau = Z_\tau$, и при $t > \tau$ распределения процессов $Y_t$ и $Z_t$ совпадают в соответствии с марковским свойством, т.е. $\mathbb{P}\{Y_t \in S|t > \tau\} = \mathbb{P}\{Z_t \in S|t > \tau\}$ для всех $S \in \mathcal{B}(\mathcal{X})$, $Y_t$ and $Z_t$, здесь $\mathcal{B}(\mathcal{X})$ – $\sigma$-алгебра на пространстве состояний $\mathcal{X}$ процессов $Y_t$ и $Z_t$.

Имеет место основное неравенство склеивания:

$$|\mathbb{P}\{Y_t \in S\} - \mathbb{P}\{Z_t \in S\}| =$$

$$= |\mathbb{P}\{Y_t \in S \& \tau > t\} - \mathbb{P}\{Z_t \in S \& \tau > t\}| +$$

$$+ |\mathbb{P}\{Y_t \in S \& \tau < t\} - \mathbb{P}\{Z_t \in S \& \tau < t\}| =$$

$$= |\mathbb{P}\{Y_t \in S \& \tau > t\} - \mathbb{P}\{Z_t \in S \& \tau > t\}| \leq \mathbb{P}\{\tau > t\}, \quad (16)$$

поскольку при $\tau < t$ “склеивание” процессов $Y_t$ и $Z_t$ уже произошло, их распределения в момент $t$ одинаковы, и $\mathbb{P}\{Y_t \in S|\tau < t\} = \mathbb{P}\{Z_t \in S|\tau < t\}$; здесь $S \in \mathcal{B}(\mathcal{X})$.

Поэтому если для некоторой возрастающей положительной функции $\phi(t)$ удастся найти оценку $\mathbb{E}\phi(\tau(Y_0, Z_0)) \leq C(X_0, Y_0)$, то можно применить неравенство Маркова:

$$\mathbb{P}\{\tau(Y_0, Z_0) > t\} = \mathbb{P}\{\phi(\tau(Y_0, Z_0)) > \phi(t)\} \leq \frac{\mathbb{E}\phi(\tau(Y_0, Z_0))}{\phi(t)} \leq \frac{C(X_0, Y_0)}{\phi(t)}, \quad (17)$$

Далее, если существует стационарное инвариантное распределение $\mathcal{P}$, к которому (слабо) сходятся распределения $Y_t$ и $Z_t$ при $t \to \infty$, то из (16) и (17) получаем

$$\begin{align*}
|\mathcal{P}\{Y_t \in S\} - \mathcal{P}(S)| &\leq \int_{\mathcal{X}} \frac{C(X_0, y)}{\phi(t)} \mathcal{P}(\text{d}y) = \frac{\hat{C}(X_0)}{\phi(t)}, \\
\text{откуда} \quad \|\mathcal{P}^{X_0}_\tau - \mathcal{P}\|_{TV} \overset{\text{def}}{=} \sup_{S \in \mathcal{B}(\mathcal{X})} |\mathcal{P}^{X_0}_\tau(S) - \mathcal{P}(S)| &\leq 2 \frac{\hat{C}(X_0)}{\phi(t)}. \quad (18)
\end{align*}$$

Изначально метод склеивания был предложен для цепей Маркова с дискретным множеством состояний ([5]), однако изучаемый нами процесс $X_t$ с переходными вероятностями ([3]) – это процесс в непрерывном времени, и для двух независимых процессов $X_t$ и $Y_t$, удовлетворяющих ([6]), но с различными начальными состояниями $\mathbb{P}\{\tau(X_0, x_0') < \infty\} = 0$.

Поэтому мы будем использовать понятие успешной склейки, предложенное в ([4]).
Определение 3 Пусть $X_t$ и $X'_t$ – две независимые версии одного марковского процесса с пространством состояний $X$: $X_t$ с начальным состоянием $X_0$, и $X'_t$ с начальным состоянием $X'_0$.

Успешной склейкой (см. [4]) процессов $X_t$ и $X'_t$ называется парный процесс $Z_t = (Y_t, Y'_t)$, заданный на некотором вероятностном пространстве, такой, что

(i) Для всех $s \geq 0$ и $S \in \mathcal{B}(X)$ выполнены равенства
$$P\{Y_t \in S\} = P\{X_t \in S\}, \quad P\{Y'_t \in S\} = P\{X'_t \in S\}. \quad \text{(Соответственно, } Y_0 = X_0 \text{ и } Y'_0 = X'_0\text{.)}$$

При этом для любого фиксированного момента $t \geq 0$ случайные величины $Y_t, Y'_t, X_t$ и $X'_t$ рассма-
триваются как некоторые случайные величины, но не как случайные процессы! Т.е. речь идёт не о совпадении распределений двух процессов, под которым понимается совпадение всех конечномерных распределений.

(ii) Для всех $t > \tau = \tau(X_0, X'_0) = \tau(Y_0, Y'_0) \overset{def}{=} \inf\{t \geq 0 : Y_t = Y'_t\}$ выполняется равенство $Y_t = Y'_t$.

(iii) Для любых начальных значений $X_0, X'_0 \in X$ верно:
$$P\{\tau(X_0, X'_0) < \infty\} = 1.$$  

Для успешной склейки можно также использовать основное неравенство склейки (16):
$$|P\{X_t \in S\} - P\{X'_t \in S\}| = |P\{Y_t \in S\} - P\{Y'_t \in S\}|$$
$$= |P\{Y_t \in S\} - P\{Y'_t \in S\}| \times (1(\tau > t) + 1(\tau \leq t)) \leq P\{\tau > t\}.$$  

Итак, наша цель – сконструировать успешную склейку для двух версий процесса $X_t$ ($X_t$ с начальным состоянием $X_0 = (0, 0)$ и $X'_t$ с начальным состоянием $X'_0 = (n'_0, x'_t(0), x'_t(1), x'_t(2), \ldots, x'_t(n'_0))$, причём каждый из процессов удовлетворяет условиям [3], и из этого получить оценку скорости сходимости распределения процесса $X_t$ к стационарному распределению.

Для применения метода склейивания нам понадобится следующая

5.1 Основная лемма склейивания

Этот факт изложен во многих публикациях (см., например, [11, 12]), но здесь мы приведём его простейшую форму.

Определение 4 Общей частью распределения двух случайных величин $\xi_1$ и $\xi_2$ с функциями распределения $\Psi_j(s) = P\{\xi_j \leq s\} \ s \in \mathbb{R}$ ($j = 1, 2$), назовём
$$\kappa \overset{def}{=} \kappa(\Psi_1(s), \Psi_2(s)) \overset{def}{=} \int_{-\infty}^s \min(\psi_1(u), \psi_2(u)) \, du,$$  

где
ψ_j(s) \overset{\text{def}}{=} \begin{cases} \Psi_j'(s), & \text{если существует } \Psi_j'(s), \\ 0, & \text{в противном случае}, \end{cases} \quad \text{для } j = 1, 2.

Лемма 1 Если $\kappa > 0$, то на некотором вероятностном пространстве можно построить две случайные величины $\xi_1'$ и $\xi_2'$ такие, что:

\begin{align}
\mathbb{P}\{\xi_j' \leq s\} &= \mathbb{P}\{\xi_j \leq s\} = \Psi_j; \\
\mathbb{P}\{\xi_1' = \xi_2'\} &\geq \kappa.
\end{align}

Доказательство. Обозначим

\begin{align}
\psi(u) &\overset{\text{def}}{=} \min(\psi_1(u), \psi_2(u)); \\
\Psi(s) &\overset{\text{def}}{=} \int_{-\infty}^s \psi(u) \, du; \\
\hat{\Psi}(s) &\overset{\text{def}}{=} \frac{\Psi(s)}{\kappa}; \\
\hat{\Psi}_j(s) &\overset{\text{def}}{=} \frac{\Psi_j(s) - \Psi(s)}{1 - \kappa}, \quad j = 1, 2
\end{align}

(если $\kappa = 1$, то $\hat{\Psi}_j(s) \equiv 0$).

Пусть $\mathcal{U}_1$, $\mathcal{U}_2$ и $\mathcal{U}_3$ – независимые случайные величины, равномерно распределённые на отрезке $[0; 1]$. Положим $\xi_j' \overset{\text{def}}{=} \hat{\Psi}_j^{-1}(\mathcal{U}_j) \times 1(\mathcal{U}_3 \leq \kappa) + \hat{\Psi}_j^{-1}(\mathcal{U}_j) \times 1(\mathcal{U}_3 > \kappa)$.

Напомним, что под обратной функцией монотонной функции $f(s)$ мы понимаем $f^{-1}(u) \overset{\text{def}}{=} \inf\{s : f(s) \geq u\}$.

Случайные величины $\xi_j'$ удовлетворяют условиям (19)–(20).

Лемма доказана. ■

Замечание 5 Если $n_0 = 0$ в некоторый момент времени $\vartheta$ (СМО свобода), то остаточная длина свободного периода процесса $X_t$ (описанного в Разделе 2) имеет плотность распределения $\varphi(s) = \lambda(X_{\vartheta+s}) \exp\left(\int_0^s \lambda(X_{\vartheta+u}) \, du\right) \in (\lambda_0 e^{-\Lambda s}; \Lambda_0 e^{-\Lambda s})$. Поэтому, если рассматривается два процесса $X_t$ и $X'_t$ с различными начальными состояниями $X_0$ и $X'_0$, и в некоторый момент времени $\vartheta$ они оба оказались в множестве $\mathcal{A}_0$ (СМО свободны), то общая часть $\mathcal{A}_0$ распределений их остаточных времен пребывания в множестве $\mathcal{A}_0$ удовлетворяет неравенству

\[ \mathcal{K}_\vartheta \geq \int_0^\infty \lambda_0 e^{-\Lambda s} \, ds = \frac{\Lambda_0}{\Lambda} \overset{\text{def}}{=} \mathcal{K}. \]
6 Конструирование успешной склейки для пары процессов $X_t, X'_t$

Рассмотрим два независимых процесса: процесс $X_t$ с начальным состоянием $X_0 = (0, 0)$ и процесс $X'_t$ с начальным состоянием $X'_0 = \big(m'_0, x'_0(0), x'_0(1), x'_0(2), \ldots, x'_0(\nu'_0)\big)$.

Будем строить их склейку $\mathcal{F}_t = (Y_t, Y'_t)$ на вероятностном пространстве $(\Omega, \mathcal{F}, \mathbb{P}) = \prod_{i=0}^{\infty} (\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, где $\Omega_i = [0; 1]$, $\mathcal{F}_i = \mathcal{B}([0; 1])$, и $\mathbb{P}_i$ — мера Лебега на $[0; 1]$.

Опишем процедуру конструирования склейки по шагам, которые производятся в моменты изменения первой компоненты любого из процессов $Y_t = \big(m_t, y_t(0) : y_t(1), y_t(2), \ldots, y_t(\nu_t)\big)$, $Y'_t = \big(m'_t, y'_t(0) : y'_t(1), y'_t(2), \ldots, y'_t(\nu'_t)\big)$ — обозначим эти моменты $t_k$; $t_0 \overset{\text{def}}{=} 0$; $t_k \overset{\text{def}}{=} \inf\{t > t_{k-1} : m_t + m'_t \neq m_{t+k} + m'_{t+k} \}$.

На каждом шаге строятся остаточные виртуальные времена обслуживания находящихся в СМО заявок (в предположении, что до окончания обслуживания данной заявки первые компоненты $n_t$ и $n'_t$ — обон процессов не изменяются) и остаточные виртуальные времена до прихода следующей заявки. Для этого используется некоторое количество $\nu_k$ экземпляров $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ — так, что на каждом шаге используются новые $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$.

Опишем алгоритм конструирования.

А. Если $m_t + m'_t > 0$, для удобства положим $Y_{t_0} = (t, \alpha_0; \alpha_1, \alpha_2, \ldots, \alpha_t)$, $Y'_{t_0} = (j, \beta_0; \beta_1, \beta_2, \ldots, \beta_j)$.

На $k$-том шаге строятся сл.в. на пространствах $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i = \ell_k + 1, \ell_k + 2, \ldots, \ell_{k+1}$; при этом $\nu_k \overset{\text{def}}{=} \ell_{k+1} - \ell_k = n_{\ell_k} + n'_{\ell_k} + 2 = \lambda + 2$. К этому моменту времени было использовано $\ell_k = \sum_{i=0}^{k-1} \nu_i$ экземпляров пространства $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$.

В момента $t_k$ виртуальное время до поступления очередного требования $\xi$ (при условии, что не закончилось обслуживание какого-то требования) для $Y_t$ имеет ф.р. $F_1(s) = 1 - \exp\left(\int_0^s \lambda(u, \alpha_0 + u; \alpha_1 + u, \alpha_2 + u, \ldots, \alpha_t + u) \, du\right)$. Строим на пространстве $(\Omega_{\ell_k+1}, \mathcal{F}_{\ell_k+1}, \mathbb{P}_{\ell_k+1})$ сл.в. $\xi \overset{\text{def}}{=} F_1^{-1}(\omega), \omega \in \Omega_{\ell_k+1}$.

В момент $t_k$ виртуальное время до окончания обслуживания $j$-го $(j = 1, 2, \ldots, i)$ требования $\eta_j$ (при условии, что не закончилось обслуживание какого-то другого требования и не поступил новых требований) $Y'_t$ имеет ф.р. $G_{1,j}(s) = 1 - \exp\left(\int_0^s h_j(t, \alpha_0 + u; \alpha_1 + u, \alpha_2 + u, \ldots, \alpha_t + u) \, du\right)$.

Строим на пространстве $(\Omega_{\ell_k+j+1}, \mathcal{F}_{\ell_k+j+1}, \mathbb{P}_{\ell_k+j+1})$ сл.в. $\eta_j \overset{\text{def}}{=} G_{1,j}^{-1}(\omega)$.
произойдет изменение первой компоненты одного из процессов $\omega$ прихода следующей заявки для процессов $\omega$ меньше $\hat{\nu}$. В соответствии с Замечанием 5 и (21), общая часть этих распределений не соответствует Лемме 1 строим $\xi$ и $\eta_j$ $(j = 1, 2, \ldots, i)$ прихода следующей заявки для процессов $Y_t$ и $Y_t'$ соответственно имеют вид:

$$P\{\xi \leq s\} = F_1(s) = 1 - \exp \left( \int_0^s \lambda(0, \alpha_0 + u) \, du \right);$$

$$P\{\xi' \leq s\} = F_2(s) = 1 - \exp \left( \int_0^s \lambda(0, \beta_0 + u) \, du \right).$$

В соответствии с Замечанием 3 и (21), общая часть этих распределений не меньше $\tilde{x} = \frac{\lambda_0}{\lambda}$. Поэтому на пространстве $\prod_{i=t_k+1}^{t_k+3} (\Omega_i, \mathcal{F}_i, P_i)$ (здесь $\nu_k = 3$) в соответствии с Леммой строим $\xi$ и $\xi'$ таким образом, что $P\{\xi = \xi'\} \geq \tilde{x}$.

В момент времени $t_k+1 \defeq t_k + \min\{\xi, \xi'\}$ первая компонента хотя бы одного из процессов $Y_t$, $Y_t'$ возрастёт, и с вероятностью большей чем $\tilde{x}$ первые компоненты процессов $Y_t$, $Y_t'$ возрастают одновременно, т.е. произойдет событие $Y_t+1 = Y_t+1 = (1, 0; 0)$.

**B.** Если случилось событие $Y_t = Y_t' = (1, 0; 0)$, при $t > t_k$ полагаем процесс $Y_t$ и $Y_t'$ равными, и строим процесс $Y_t$ (и равный ему $Y_t'$) по процедуре, описанной в пункте **A**.

Несложно заметить, что сконструированные по алгоритму **A-B** процессы $Y_t$ и $Y_t'$ удовлетворяют условиям (i)-(ii) Определения 3.

Покажем, что выполнено также условие (iii) Определения 3.

Рассмотрим процесс $Y_t' \not\equiv X_t'$. Если $n_0 > 0$, то пользуясь теми же соображениями, что и в Разделах 3 и 4 можно показать, что остаточное время $\zeta_t$ первого периода занятости процесса $Y_t'$, зависящее от начального состояния $Y_0'$, имеет $k \in [0, K]$ конечных моментов.
Обозначим $\vartheta_0 \triangleq \zeta_0$, $\vartheta_k \triangleq \inf\{t > \vartheta_k : Y_t' \in \mathcal{S}_0, Y_{t-0} \notin \mathcal{S}_0\}$: $\vartheta_k$ – это момент окончания $k$-го периода занятости процесса $Y_t'$. В случае $Y_0' \in \mathcal{S}_0$ положим $\vartheta_0 \triangleq 0$.

Поскольку распределение процесса $Y_t$ совпадает с распределением процесса $X_t$, верно неравенство $P\{X_{\vartheta_k} \in \mathcal{S}_0\} \geq e^{-\rho}$ – см. (15).

И, как мы уже видели, после совместного попадания процессов $Y_t$ и $Y_t'$ в множество $\mathcal{S}$ с вероятностью большей, чем $\hat{\kappa}$, при выходе из этого множества процессы “склеиваются”. Поэтому, снова обозначив $\tau \triangleq \inf\{t > 0 : Y_t = Y_t'\}$, а $\eta_k$ – длина свободного периода после $\vartheta_k$, получаем, что $\tau = \tau(Y_0') = \tau(X_0')$ (поскольку $X_0 = Y_0 = (0, 0)$ зафиксирован). Опишем, что

$$\vartheta_k + \eta_k = \vartheta_0 + \eta_0 + \sum_{i=1}^{k} (\zeta_i + \eta_i), \quad \text{где } \zeta_i - i-й период занятости, а } \eta_i - \text{ следующий за ним свободный период.}$$

Обозначим $\delta_k \triangleq \{Y_{\vartheta_k} \neq Y_{\vartheta_k}' \& Y_{\vartheta_k} + \eta_k = Y_{\vartheta_k} + \eta_k\}$ – событие, заключающееся в том, что склеивание произошло сразу после $k$-го попадания $Y_t'$ в множество $\mathcal{S}_0$.

Утверждение 4. При $r \in [1; K]$ выполнено $E\tau^r \leq C_1(\varpi, r)E \vartheta_k + C_1(\varpi, r)E \zeta_1 + C_3(\varpi, r, \lambda_0, \Lambda)$, константы $C_1$, $C_2$, $C_3$ вычисляются по приведённым ниже формулам.

Доказательство. Сначала оценим $E|\eta_k|$, $E|\zeta_k|$, $E|\delta_k|$, $E|\vartheta_k|$. В случае события $\delta_k$ сл.в. $\eta_k$ (в соответствии с Леммой [1]) имеет плотность распределения

$$\varphi(s) = \min \left\{ \lambda(Y_{\vartheta_k} + s) \exp \left( -\int_0^s \lambda(Y_{\vartheta_k} + u) \, du \right), \right.$$  

$$\lambda(Y_{\vartheta_k}' + s) \exp \left( -\int_0^s \lambda(Y_{\vartheta_k}' + u) \, du \right) \right\} \leq \Lambda e^{-\lambda_0 s}.$$  

Поэтому $E|\eta_k| < \frac{r!A}{\lambda_0^{r+1}}$. Легко видеть, что и $E|\zeta_k| \leq \frac{r!A}{\lambda_0^{r+1}}$. $\zeta_k$ от $\delta_j$ не зависят при $k \neq j$. 

19
Оценим $E \tau^r$ (используя неравенство Йенсена в форме (13)):

$$E \tau^r \leq \sum_{k=1}^{\infty} E \left( \vartheta_0 + \eta_0 + \sum_{i=1}^k (\zeta_i + \eta_i) \right) 1(\vartheta_k) \leq$$

$$\leq \sum_{k=1}^{\infty} (1 - \varpi)^{k-1} \varpi (2k + 2)^{r-1} \left( E \vartheta_0^r + (k + 1) \frac{r! \Lambda}{\lambda_0^{r+1}} + k E \zeta_i^r \right) <$$

$$< 2^{r-1} \left( E \vartheta_0^r \sum_{k=1}^{\infty} (1 - \varpi)^{k-1} (k + 1)^{r-1} + \frac{r! \Lambda}{\lambda_0^{r+1}} \sum_{k=1}^{\infty} (1 - \varpi)^{k-1} (k + 1)^r + E \zeta_i^r \sum_{k=1}^{\infty} (1 - \varpi)^{k-1} (k + 1)^{r-1} k \right).$$

Предложение 4 доказано. ■

Итак, сконструированная выше пара $Z_t = (Y_t, Y'_t)$ – успешная склейка процессов $X_t$ и $X'_t$, и получена явная оценка времени склеивания:

$$E (\tau(X'_0))^r \leq C(r, \lambda_0, \Lambda, E \vartheta_0^r).$$

(22)

То есть

$$|P\{X_t \in S\} - P\{X'_t \in S\}| \leq C(r, \lambda_0, \Lambda, E \vartheta_0^r),$$

(23)

однако следующий шаг – интегрирование (18) мы произвести не можем, поскольку стационарное распределение СМО $M|G|\infty$ (процесс $\overset{\circ}{X}_t$) и, тем более, $G|G|\infty$ (процесс $X_t$), неизвестно.

7 Процесс $X_t$ как регенерирующий процесс

Напомним определение регенерирующего процесса.

Определение 5 Случайный процесс $\{W_t, t \geq 0\}$, заданный на вероятностном пространстве $(\Omega, \mathcal{F}, \mathbb{P})$, измеримый относительно фильтрации $\mathcal{F}_t, t \geq 0$, и принимающий значения в измеримом пространстве состояний $\mathcal{W}, B(\mathcal{W})$ называется регенерирующим процессом, если существует такая последовательность Марковских по отношению к фильтрации $\mathcal{F}_t$ моментов $\{\theta_i\}_{i \in \mathbb{N}}$, таких, что:

1. $W_{\theta_i} = W_{\theta_j}$ при всех $i, j \in \mathbb{N}$;

2. Случайные элементы $\Xi_i \overset{def}{=} \{W_t, t \in [\theta_i, \theta_{i+1})\}$ ($i \in \mathbb{N}$) одинаково распределены и независимы в совокупности.
Назовем периодами регенерации промежутки $[\theta_i, \theta_{i+1}]$; очевидно, длины периодов регенерации $\chi_i \overset{\text{def}}{=} \theta_{i+1} - \theta_i$ одинаково распределены и независимы в совокупности. Обозначим $\Phi(s) \overset{\text{def}}{=} \Pr \{ \chi_i \leq s \}$ – (неизвестная нам) ф.р. слв. $\chi_i$

Известно, что если $\mathbb{E} \chi_i < \infty$, то распределение регенерирующего процесса $W_t$ слабо сходит к стационарному (инвариантному) распределению.

Заметим, что для введенного в Раздел 2 процесса $X_t = \left( n_t, x_t^{(0)}, x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(n_t)} \right)$ периоды регенерации разделяются на период занятости и свободный период: $\chi_k = \zeta_k + \eta_k$. Моменты периода занятости оцениваются с помощью (11) и (12). Моменты свободных периодов $\eta_i$, следующих за периодами занятости $\zeta_i$, оценены в Предложении 4. Кроме того, период регенерации заведомо больше своей части – свободного периода, а плотность распределения свободного периода $\varphi(s) \geq \lambda_0 e^{-\Lambda s}$ (Замечание 5).

Таким образом, известны оценки

$$\mathbb{E} \chi_k \leq 2^{r-1}(\mathbb{E} \zeta_k + \mathbb{E} \eta_k) = M_r, \quad \mathbb{E} \chi_i > \frac{\lambda_0}{\Lambda^2}.$$ (24)

Марковские моменты $\{\theta_i\}_{i \in \mathbb{N}}$ образуют вложенный процесс восстановления $N_t = \sum_{i=1}^{\infty} 1(\theta_i < t)$.

Рассмотрим перескок процесса $N_t$, а именно Марковский процесс $B_t \overset{\text{def}}{=} t - \theta_{N_t}$ (с пространством состояний $\mathbb{R}_+$), распределение которого $\Psi_t$, как известно (17), (20), слабо сходит к стационарному инвариантному распределению $\Psi$ с ф.р. $\Phi(s) = \int_0^s (1 - \Phi(u)) \, du \mathbb{E} \chi_i$. (25)

К этому же распределению (25) слабо сходит распределение недоскока $R_t \overset{\text{def}}{=} \theta_{N_{t+1}} - t$ процесса восстановления $N_t$.

Поскольку процесс $X_t$ – регенерирующий, его распределение полностью определяется временем, прошедшим с момента начала периода регенерации, т.е. $\mathcal{P}_t = \Psi(B_t)$.

Теперь рассмотрим процесс $X_t$ с начальным состоянием $(0, 0)$ и процесс $X'_t$ с произвольным начальным состоянием $X_0$. Этим процессам соответствуют вложенные процессы восстановления $N_t$ и $N'_t$, а процессам $N_t$ и $N'_t$ соответственно процессы перескоков $B_t$ и $B'_t$ и процессы недоскоков $R_t$ и $R'_t$.

Сконструировав успешную склейку $Z_t = (Y_t, Y'_t)$ для процессов $X_t$ и $X'_t$, одновременно строится успешная склейка $W_t = (D_t, D'_t)$ процессов $B_t$ и $B'_t$. Обозначим $\mathcal{P}_r$ и $\mathcal{P}'_r$ – распределение процессов $B_t$ и $B'_t$ соответственно ($\Pr \{ B_t \in S \} = \mathcal{P}_r(S)$ и $\Pr \{ B'_t \in S \} = \mathcal{P}'_r(S)$).
Из (22) и (23) сразу следует

\[ |P\{B_t \in S\} - P\{B'_t \in S\}| \leq \frac{C(\varpi, r, \lambda_0, \Lambda, \vartheta_0)}{t^r}, \]

где \( \vartheta_0 = R_0' \) — недоскок процесса \( N'_t \) в момент времени \( t = 0 \).

**Утверждение 5**

\[ \|P - \mathfrak{P}\|_{TV} \leq 2 \left( \frac{C_1(\varpi, r)}{t^r} \frac{\mathbb{E} \chi_{i+1}}{r+1} + C_1(\varpi, r) \mathbb{E} \zeta_i + C_3(\varpi, r, \lambda_0, \Lambda) \right) = \frac{C_1(r, \lambda_0, \Lambda)}{t^r}, \]

где \( r \in [1, K - 1) \).

**Доказательство.** Рассмотрим процесс перескока \( B_t \) с начальным условием \( B_0 = 0 \) и процесс перескока \( \tilde{B}_t \), имеющий стационарное начальное распределение \( \mathfrak{P} \); при всех \( t \geq 0 \)

\[ P\{\tilde{B}_t \in S\} = \mathfrak{P}(S) = \int_S 1 - \Phi(u) \frac{d u}{\mathbb{E} \chi_i}. \]

Подставляя в (26) вместо \( \vartheta_0 = R_0' \) стационарный недоскок \( \tilde{R}_0 \) с распределением \( (25) \), получаем утверждение Предложения [3].

Учитывая (24), можно вычислить верхнюю оценку величины \( C_1(r, \lambda_0, \Lambda) \). □

**Теорема 2** В условиях (7) верно неравенство \[ \|\mathcal{P}_t - \mathfrak{P}\|_{TV} \leq \frac{C_1(r, \lambda_0, \Lambda)}{t^r}, \]

где \( r \in [1, K - 1) \).

**Доказательство.** Как уже говорилось, \( \mathcal{P}_t = F(\Psi_t) \). Соответственно, \( \mathcal{P} = F(\mathfrak{P}) \). Напомним, что \( \chi_i \) — длина периода регенерации, включающего в себя абсолютно непрерывные величины, т.е. распределение сл.в. \( \chi_i \) имеет плотность. Поэтому имеет плотность \( \psi_i(s) \) перескок \( B_i \); как уже говорилось (24), известна ф.р. стационарного перескока, и её плотность равна \( \psi(s) = \frac{1 - \Phi(u)}{\mathbb{E} \chi_i} \).

Обозначим \( \sigma_t = \{s : \psi_i(s) > \psi(s)\} \).
Оценим

\[ |\mathcal{P}_t(S) - \mathcal{P}(S)| = \int_0^\infty P\{X_t \in S|B_t = u\} \psi_t(u) \, du - \int_0^\infty P\{X_t \in S|B_t = u\} \psi(u) \, du \leq \]

\[ \leq \max \left\{ \int_{\sigma_t} (\psi_t(u) - \psi(u)) \, du, \int_{R_+ \setminus \sigma_t} (\psi(u) - \psi_t(u)) \, du \right\} = \]

\[ = \max \{ |\mathcal{P}_t(\sigma_t) - \mathcal{P}(\sigma_t)|, |\mathcal{P}_t(R_+ \setminus \sigma_t) - \mathcal{P}(R_+ \setminus \sigma_t)| \} \leq \sup_{\sigma \in \mathcal{R}(\mathbb{R})} |\mathcal{P}_t(\sigma) - \mathcal{P}(\sigma)|, \]

откуда \( \|\mathcal{P}_t - \mathcal{P}\|_{TV} \leq \|\mathcal{P}_t - \mathcal{P}\|_{TV} \), что завершает доказательство. ■

Замечание 6 Полученная оценка неоптимальна по следующим причинам.

1. Применение метода склеивания для процессов в непрерывном времени, как правило, даёт неоптимальную оценку.

2. При конструировании успешной склейки использовались только моменты попадания процесса \( X'_t \) в свободное состояние; также можно было бы учесть и моменты попадания в свободное состояние процесса \( X_t \).

3. Условия (7) не позволяют априорно оценить полученные в ходе вычислений величины \( (C \text{ и } C_1) \). При исследовании конкретных прикладных задач эти величины могут быть оценены точнее.

Тем не менее, полученная оценка может быть полезна в практических приложениях, поскольку имитационное моделирование не всегда способно дать достаточно качественные оценки.

Автор благодарит Л.Г. Афанасьеву, А.Ю. Веретенникова и А.Д. Маниту за важные и полезные замечания при обсуждении представленной работы. Работа выполнена при финансовой поддержке Российского научного фонда (проект № 17-01-00633 А).

Список литературы

[1] Afanasyeva, L. G., Tkachenko, A. V. On the Convergence Rate for Queueing and Reliability Models Described by Regenerative Processes // Journal of Mathematical Sciences, October 2016, Volume 218, Issue 2, pp 119–136.

[2] Asmussen, S. Applied Probability and Queues. Second edition. New York: Springer-Verlag, 2003.
[3] Golovastova, E. On convergence rate for an infinite-channel queuing system with Poisson input flow // arXiv:1711.07075 [math.PR]

[4] Griffeath, D. A maximal coupling for Markov chains // Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete — 1975 — Volume 31 — Issue 2, P. 95–106.

[5] Doeblin, W. Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d’états // Rev. Math. de l’Union Interbalkanique — 1938 — 2, P. 77–105.

[6] Erlang, A. K.: Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. Elektroteknikeren 13, 1917, 5-13 (1917) (in Danish); Engl. transl. P.O. Elect. Eng. J. 10, 189-197 (1918).

[7] Takács, L. Introduction to the Theory of Queues, (Oxford University Press, 1962)

[8] Stadje, W. The busy period of the queueing system $M|G|\infty$ // Journal of Applied Probability — 1985 — 22, P. 697–704. http://dx.doi.org/10.2307/3213872

[9] Fortet, R.: Calcul des probabilités. CNRS, Paris (1950)

[10] Kalashnikov, V.V. Mathematical Methods in Queueing Theory. Dordrecht: KluwerAcademic Publishers Group, 1994.

[11] Karlin, S., Taylor, H.M. A Second Course in Stochastic Processes, San Diego at al., 2nd Ed., Academic Press, 1975.

[12] Kato, K. Coupling Lemma and Its Application to The Security Analysis of Quantum Key Distribution // Tamagawa University Quantum ICT Research Institute Bulletin Vol.4 No.1 : 23-30 (2014) P.23–30.

[13] Thorisson, H. Coupling, Stationarity, and Regeneration. Springer, 2000.

[14] Veretennikov A., Butkovsky O.A. On asymptotics for Vaserstein coupling of Markov chains // Stochastic Processes and their Applications — 123(9) — 2013, P. 3518–3541.

[15] Veretennikov, A. On convergence rate for Erlang–Sevastyanov type models with infinitely many servers // Theory of Stochastic Processes. 2017. No. 1, 88-102; https://arxiv.org/abs/1412.3849

[16] Veretennikov, A. On Recurrence and Availability Factor for Single–Server System With General Arrivals, Reliability: Theory and Applications (RT&A), 2016, vol.11, No3(42), 49-58. [http://www.gnedenko-forum.org/Journal/2016-3.htm]
[17] Veretennikov, A. Yu. On the rate of convergence for infinite server Erlang–Sevastyanov’s problem // Queueing Systems, 2014, Volume 76, Issue 2, pp 181-203; DOI: 10.1007/s11134-013-9384-4

[18] Veretennikov, A.Yu., Zverkina, G.A. Simple Proof of Dynkin’s Formula for Single-Server Systems and Polynomial Convergence Rates // Markov Processes Relat. Fields 20 (2014), 479–504.

[19] Гнеденко Б.В., Беляев Ю.К., Соловьев А.Д. Математические методы в теории надежности. М.: “НАУКА”, 1965. — 524 с.

[20] Гнеденко Б.В., Коваленко И.Н. Введение в теорию массового обслуживания. М.: Наука, 1966. — 432 с.

[21] Б. А. Севастьянов, Эргодическая теорема для марковских процессов и ее приложение к телефонным системам с отказами. // ТВП, 2:1 (1957), 106-116

[22] Б. А. Севастьянов, Формулы Эрланга в телефонии при произвольном законе распределения длительности разговора //ТВП, 2:1 (1957), 135

[23] Штойян Д. Качественные свойства и оценки стохастических моделей. М.: Мир, 1979. – 271 с.

[24] Боровков, А.А. Асимптотические методы в теории массового обслуживания. М.: Физматгиз, 1980. Engl. transl.: Asymptotic methods in queueing theory. John Wiley & Sons, Ltd., Chichester, 1984.