On recurrence and availability factor for single–server system with general arrivals

A. Yu. Veretennikov[2]

•

University of Leeds, UK; National Research University
Higher School of Economics, and Institute for Information
Transmission Problems, Moscow, Russia,

email: a.veretennikov @ leeds.ac.uk.

Abstract

Recurrence and ergodic properties are established for a single–server queueing system with variable intensities of arrivals and service. Convergence to stationarity is also interpreted in terms of reliability theory.

Keywords: *single-server system, arrivals, recurrence*

1 Introduction

In the last decades, queueing systems of $M/G/1/\infty$, or $M/G/1$, or more general $G/G/1$ type (cf. [10]) – one of the most important queueing systems – attracted much attention, see [1], [2], [4], [5], [6], [9], [12], [13], [15], and references therein, et al. In this paper a single–server system similar to [18, 19] is considered, in which *intensities* of new arrivals as well as of their service may depend on the "whole state" of the system, where the whole state includes the number of customers in the system – waiting and at service – *and* the elapsed time of the last service (that is, time from the beginning of this service), as well as the elapsed time from the *last arrival*. In queueing theory notations, the system under consideration may be denoted as $G/G/1/\infty$ with restrictions. Batch arrivals are not allowed. The model is not $GI/GI/1/\infty$ (here "I" stands for independence, as usual) because generally speaking periods between two consequent hits of idle state may be dependent, as well as by other reasons. (This is a slight abuse of notations because here "idle" is more than one state.) The generalisation in comparison to the standard $GI/GI/1/\infty$ model and to the models studied in [18, 19] is because of dependence of the intensities on time from the last arrival, due to which the moment of hitting the idle state cannot be considered as a regeneration. The restrictions mentioned above relate to the existence of intensities and to certain assumptions on them, see the details in the beginning of the next section. By the *m-availability factor* of the system we understand the probability of m customers in total on the server and in the queue. The problem addressed in the paper is how estimate convergence rate of characteristics of the system including the m-availability factors to their stationary values.

[2] The work was prepared within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program, and supported by the RFBR grant 14-01-00319-a.

The *elapsed* service time of a customer at service is assumed to be known at any moment, but the remaining service time is not; the same is true for the elapsed time from the last arrival and the remaining time to the next arrival. For definiteness, the discipline of serving is FIFO (first–in–first–out), although other disciplines may be also considered.

The paper consists of the Section 1 – Introduction, of the setting and main result in the Section 2, of the auxiliary lemmata in the Section 3 and of the outline of the proof of the main result in the Section 4. In the Theorem 1 we deliberately do not show exact inequalities on the characteristics in the assumptions, which would imply greater or lesser constant in the main assertion (7) because of various possibilities of dependences between them; however, some idea of what precisely may be assumed can be easily worked out from the calculus in the proof.

2 The setting and main result

2.1 Defining the process

Let us present the class of models under investigation in this paper. Here the state space is a union of subspaces,

$$\mathcal{X} = \{(0, y):\ y \geq 0\} \cup \bigcup_{n=1}^{\infty} \{(n, x, y):\ x, y \geq 0\}.$$

Functions of class $C^1(\mathcal{X})$ are understood as functions with classical continuous derivatives with respect to the variable x. Functions with compact support on $\mathcal{X}$ are understood as functions vanishing outside some domain bounded in this metric: for example, $C_0^1(\mathcal{X})$ stands for the class of functions with compact support and one continuous derivative. There is a generalised Poisson arrival flow with intensity $\lambda(X)$, where $X = (n, x, y)\ \ for\ any\ n \geq 1$, and $X = (0, y)\ \ for\ n = 0$. Slightly abusing notations, it is convenient to write $X = (n, x, y)$ for $n = 0$ as well, assuming that in this case $x \equiv 0$; after this identification, the "corrected" state space becomes

$$\mathcal{X} = \{(0,0, y):\ y \geq 0\} \cup \bigcup_{n=1}^{\infty} \{(n, x, y):\ x, y \geq 0\}.$$

If $n > 0$, then the server is serving one customer while all others are waiting in a queue. When the last service ends, immediately a new service of the next customer from the queue starts; recall that for definiteness the schedule of the service is assumed FIFO, although actually the result does not depend on it. If $n = 0$ then the server remains idle until the next customer arrival; the intensity of such arrival at state $(0, y) \equiv (0,0, y)$ may be variable depending on the value y, which stands for the elapsed time from the last end of service. Here n denotes the total number of customers in the system, and x stands for the elapsed time of the current service (except for $n = 0$, which was explained earlier), and y is the elapsed time from the last arrival. *Usually*, in the literature intensity of arrivals – if not independent – may depend on n (and in some cases on y), while intensity of service may depend on n (and in some cases also on x); however, we allow a more general dependence of all these variables together. Denote $n_t = n(X_t)$ – the number of customers corresponding to the state X_t, and $x_t = x(X_t)$, the second component of the process (X_t), and $y_t = y(X_t)$, the third component of the process (X_t) (the third if $n > 0$)). For any $X = (n, x, y)$, intensity of service $h(X) \equiv h(n, x, y)$ is defined; it is also convenient to assume that $h(X) = 0$ for $n(X) = 0$. Both intensities λ and h are understood in the following way, which is a definition: on any nonrandom interval of time $[t, t + \Delta)$, the conditional probability given X_t that the current service will *not* be finished and there will be no new arrivals reads,

$$\exp\left(-\int_0^{\Delta} (\lambda + h)(n_t, x_t + s, y_t + s)\, ds\right); \tag{1}$$

the conditional probability of exactly one arrival and no other events on this interval given X_t equals

$$\int_0^{\Delta} \exp\Big(-\int_0^{v} (\lambda + h)(n_t, x_t + s, y_t + s)\, ds\Big)\lambda(n_t, x_t + v, y_t + v) \tag{2}$$

$$\times \exp\Big(-\int_0^{\Delta-v}(\lambda+h)(n_t+1,x_t+v+s',s')\,ds'\Big)\,dv;$$

the probabiity of exactly one service given X_t (of course, assuming $n_t>0$, otherwise no service is available and the probability in question equals zero) is

$$\int_0^{\Delta}\exp\big(-\int_0^{v}(\lambda+h)(n_t,x_t+s,y_t+s)\,ds\big)h(n_t,x_t+v,y_t+v) \tag{3}$$
$$\times \exp\Big(-\int_0^{\Delta-v}(\lambda+h)(n_t-1,s',y_t+v+s')\,ds'\Big)\,dv;$$

and so on, i.e., by induction a conditional probability of any finite number of events on this interval may be written as some multivariate integral, while the probability of infinitly many events on it equals zero. In particular, the (conditional given X_t) density of the moment of a new arrival *or* of the end of the current service after t at x_t+z, $z\geq 0$, equals,

$$(\lambda(n_t,x_t+z,y_t+z)+h(n_t,x_t+z,y_t+z))$$
$$\times\exp\big(-\int_0^{\infty}(\lambda+h)(n_t,x_t+s,y_t+s))\,ds\big).$$

This standard construction does not require any regularity of either intensity function, and even may allow some *un*bounded intensities; however, we do not touch this issue here and in the sequel both functions λ and h are assumed to be *bounded* and, of course, Borel measurable. In this case, for $\Delta>0$ small enough, the expression in (1) may be rewritten as

$$1-\int_0^{\Delta}(\lambda+h)(n_t,x_t+s,y_t+s)\,ds+O(\Delta^2),\qquad \Delta\to 0, \tag{4}$$

and this what is "usually" replaced by

$$1-(\lambda(X_t)+h(X_t))\Delta+O(\Delta^2).$$

However, in our situation, the latter replacement may be incorrect because of possible discontinuities of the functions λ and h. Emphasize that from time t and until the next jump, the evolution of the process X is *deterministic*, which makes the process *piecewise-linear Markov,* see, e.g., [10].

2.2 Main result

Let

$$\Lambda:=\sup_{n,x,y:\,n>0}\lambda(n,x,y)<\infty.$$

For establishing convergence rate to the stationary regime, we assume (cf. [18, 19]),

$$\inf_{n>0,y}h(n,x,y)\geq\frac{C_0}{1+x},\quad x\geq 0. \tag{5}$$

We also assume a new condition related to the function $\lambda_0(t):=\lambda(0,0,t)$:

$$0<\inf_{t\geq 0}\lambda_0(t)\leq\sup_{t\geq 0}\lambda_0(t)<\infty. \tag{6}$$

Recall that the process has no explosion with probability one due to the boundedness of both intensities, i.e., the trajectory may have only finitely many jumps on any finite interval of time.

Theorem 1 *Let the functions λ and h be Borel measurable and bounded and let the assumptions (5) and (6) be satisfied. Then, under the assumptions above, if C_0 is large enough, then there exists a unique stationary measure μ. Moreover, for any $k>0$ and any $m>k$, if C_0 is large enough, then there exists $C>0$ such that for any $t\geq 0$,*

$$\|\,\mu_t^{n,x,y}-\mu\,\|_{TV}\leq C\,\frac{(1+n+x+y)^m}{(1+t)^{k+1}}, \tag{7}$$

where $\mu_t^{n,x,y}$ is a marginal distribution of the process $(X_t,t\geq 0)$ with the initial data $X=(n,x,y)\in\mathcal{X}$. The constant C in (7) admits an effective bound.

In particular, this inequality holds true for the reliability characteristics introduced earlier. For any $m\geq 0$ denote

$$p_{\leq m}(t):=\mathbb{P}_x(X_t\in\{n(X_t)\leq m\}).$$

Then the following corollary holds true.

Corollary 1 *Under the assumptions of the Theorem 1, the probabilities $p_{\leq m}(t)$ converge to their limits, $p_{\leq m}(\infty)$, as $t\to\infty$, and for any $k>0$ and any $m>k$, if C_0 is large enough, then there exists $C>0$*

– the same as in (7) – such that the estimate is valid,

$$|p_{\le m}(t) - p_{\le m}(\infty)| \le C\,\frac{(1+n+x+y)^m}{(1+t)^{k+1}},$$

where x, y are the components of the initial state $X = (n, x, y) \in \mathcal{X}$.

Remark 1 *It is plausible that under the same set of conditions (5)–(6), the bound in (7) may be improved so that the right hand side does not depend on y. Moreover, we emphasize that given all other constants, the value C in (7) may be made "computable", with a rather involved but explicit dependence on the other constants. It is likely that the condition (6) may be replaced by a weaker one,*

$$\frac{C_0\prime}{1+t} \le \lambda_0(t) \le \sup_{t\ge 0}\lambda_0(t) < \infty, \tag{8}$$

along with the assumption that C'_0 is large enough; also, it is tempting to replace the condition (5) by a weaker one with some dependence of the bound in the right hand side of the variable y (under which new condition an improvement that was mentioned as a hypothesis in the first phrase of this remark becomes unlikely). However, all these issues require a bit more accuracy in the calculus and we do not pursue these goals here leaving them until further investigations.

The idea of the proof is based on constructing appropriate Lyapunov functions and yet on finding a new regeneration state instead of a "compromised" idle state of the system. Lyapunov functions guarantee that the distribution of the (independent) periods between regenerations admit some polynomial moments, which implies the desired statement. However, we first of all need some auxiliary results on a strong Markov property – which is essential in this approach – and on Dynkin's formula.

3 Lemmata

Recall [8] that the generator of a Markov process $(X_t, t \ge 0)$ is an operator $\mathcal{G}$, such that for a sufficiently large class of functions f

$$\operatorname*{suplim}_X \left\| \frac{\mathbb{E}_X f(X_t) - f(X)}{t} - \mathcal{G}f(X) \right\| = 0 \tag{9}$$

in the norm of the state space of the process; the notion of generator does depend on this norm. An operator $\mathcal{G}$ is called a *mild generalised generator* (another name is extended generator) if (9) is replaced by its corollary (10) below called *Dynkin's formula*, or *Dynkin's identity* [8, Ch. 1, 3],

$$\mathbb{E}_X f(X_t) - f(X) = \mathbb{E}_X \int_0^t \mathcal{G}f(X_s)\,ds, \tag{10}$$

also for a wide enough class of functions f. We will also use the non-homogeneous counterpart of Dynkin's formula,

$$\mathbb{E}_X \varphi(t, X_t) - \varphi(0, X) = \mathbb{E}_X \int_0^t \left(\frac{\partial}{\partial s}\varphi(s, X_s) + \mathcal{G}\varphi(s, X_s) \right) ds, \tag{11}$$

for appropriate functions of two variables $(\varphi(t, X))$. Both (10) and (11) play a very important role in analysis of Markov models and under our assumptions may be justified similarly to [19]. Here X is a (non-random) initial value of the process. Both formulae (10)–(11) hold true for a large class of functions f, φ with $\mathcal{G}$ given by the standard expression,

$$\mathcal{G}f(X) := \frac{\partial}{\partial x} f(X) 1(n(X) > 0) + \frac{\partial}{\partial y} f(X)$$

$$+\lambda(X)(f(X^+) - f(X)) + h(X)(f(X^-) - f(X)),$$

where for any $X = (n, x, y)$,

$$X^+ := (n+1, x, 0), \quad X^- := ((n-1) \vee 0, 0, y)$$

(here $a \vee b = \max(a, b)$). Under our minimal assumptions on regularity of intensities this may be justified similarly to [19].

Lemma 1 *If the functions λ and h are Borel measurable and bounded, then the formulae (10) and (11) hold true for any $t > 0$ for every $f \in C_b^1(\mathcal{X})$ and $\varphi \in C_b^1([0,\infty) \times \mathcal{X})$, respectively. Moreover, the process $(X_t, t \geq 0)$ is strong Markov with respect to the filtration $(\mathcal{F}_t^X, t \geq 0)$.*

Further, let

$$L_m(X) = (n+1+x+y)^m, \quad L_{k,m}(t,X) = (1+t)^k L_m(X). \tag{12}$$

The extensions of Dynkin's formulae for some unbounded functions hold true: we will need them for the Lyapunov functions in (12).

Corollary 2 *Under the assumptions of the Lemma 1,*

$$\begin{aligned} L_m(X_t) - L_m(X) = \int_0^t \lambda(X_s)\big[\quad \big(L_m(X_s^{(+)}) - L_m(X_s)\big) \\ +h(X_s)(L_m(X_s^-) - L_m(X_s)) + 1(n(X_s) > 0)\frac{\partial}{\partial x}L_m(X_s) + \frac{\partial}{\partial y}L_m(X_s)\big]\, ds + M_t, \end{aligned} \tag{13}$$

with some martingale M_t, and also

$$\begin{aligned} L_{k,m}(t,X_t) - L_{k,m}(0,X) = \int_0^t \big[\lambda(X_s)\big(L_{k,m}(s,X_s^{(+)}) - L_{k,m}(s,X_s)\big) \\ +h(X_s)\big(L_{k,m}(s,X_s^-) - L_{k,m}(s,X_s)\big) \\ +\Big(1(n(X_s) > 0)\frac{\partial}{\partial x} + \frac{\partial}{\partial y} + \frac{\partial}{\partial s}\Big)L_{k,m}(s,X_s)\big]\, ds + \widetilde{M}_t, \end{aligned} \tag{14}$$

with some martingale $\widetilde{M}_t$.

About a martingale approach in queueing models see, for example, [14]. The proof of the Lemma 1 is based on the next three Lemmata. The first of them is a rigorous statement concerning a well-known folklore property that probability of "one event" on a small nonrandom interval of length Δ is of the order $O(\Delta)$ and probability of "two or more events" on the same interval is of the order $O(\Delta^2)$. Of course, in queueing theory this is a common knowledge; moreover, the claims (15)–(18) follow immediatey from the definition of the process given earlier in (1)–(??). Yet for discontinuous intensities these properties have to be, at least, explicitly stated.

Lemma 2 *Under the assumptions of the Theorem 1, for any $t \geq 0$,*

$$\mathbb{P}_{X_t}(\,no\ \ jumps\ \ on\ \ (t,\ t+\Delta]\,) = \exp(-\textstyle\int_0^\Delta (\lambda+h)(X_t+s)\,ds) \quad (= 1 + O(\Delta)), \tag{15}$$

$$\mathbb{P}_{X_t}(\,at\ \ least\ \ one\ \ jump\ \ on\ \ (t,\ t+\Delta]\,) = O(\Delta), \tag{16}$$

$$\mathbb{P}_{X_t}(\,exactly\ \ one\ \ jump\ \ up\ \ \&\ \ no\ \ down\ \ on\ \ (t,\ t+\Delta]\,) = \textstyle\int_0^\Delta \lambda(X_t+s)\,ds + O(\Delta^2), \tag{17}$$

$$\mathbb{P}_{X_t}(\,exactly\ \ one\ \ jump\ \ down\ \ \&\ \ no\ \ up\ \ on\ \ (t,\ t+\Delta]\,) = \textstyle\int_0^\Delta h(X_t+s)\,ds + O(\Delta^2), \tag{18}$$

and

$$\mathbb{P}_{X_t}(\,at\ \ least\ \ two\ \ jumps\ \ on\ \ (t,\ t+\Delta]\,) = O(\Delta^2). \tag{19}$$

In all cases above, $O(\Delta)$ and $O(\Delta^2)$ are uniform with respect to X_t and only depend on the norm $\sup_X(\lambda(X) + h(X))$, that is, there exist $C > 0, \Delta_0 > 0$ such that for any X and any $\Delta < \Delta_0$,

$$\begin{aligned} \overline{\lim_{\Delta\to 0}}\{ \quad \Delta^{-1}\mathbb{P}_X(\,at\ \ least\ \ one\ \ jumps\ \ on\ \ (0,\Delta]\,) + \\ \Delta^{-2}\mathbb{P}_X(\,at\ \ least\ \ two\ \ jumps\ \ on\ \ (0,\ \Delta]\,) \\ +\Delta^{-2}\Big[\mathbb{P}_{X_t}(\,one\ \ jump\ \ up\ \ \&\ \ no\ \ down\ \ on\ \ (t,\ t+\Delta]\,) - \textstyle\int_0^\Delta \lambda(X_t+s)\,ds\Big] \\ +\Delta^{-2}\Big[\mathbb{P}_{X_t}(\,one\ \ jump\ \ down\ \ \&\ \ no\ \ up\ \ on\ \ (t,\ t+\Delta]\,) - \textstyle\int_0^\Delta h(X_t+s)\,ds\Big]\Big\} < C < \infty. \end{aligned} \tag{20}$$

The next two Lemmata are needed for the justification that the process with discontinuous intensities is, indeed, strong Markov.

Lemma 3 *Under the assumptions of the Theorem 1, the semigroup* $T_t f(X) = \mathbb{E}_X f(X_t)$ *is continuous in* t.

Lemma 4 *Under the assumptions of the Theorem 1 the process* $(X_t, t \geq 0)$ *is Feller, that is,* $T_t f(\cdot) \in C_b(\mathcal{X})$ *for any* $f \in C_b(\mathcal{X})$.

The proofs of all Lemmata 2–4 may be performed similarly to [19] where no regularity of the intensities was used, although the dependences were a bit less general. Further, according to [8, Theorem 3.3.10], any Feller process satisfying the claim of the Lemma 3 with right continuous trajectories is strong Markov, which guarantees the last assertion of the Lemma 1.

4 Outline of the Proof of the Theorem 1

1. The idea of the proof is to identify a regeneration state and to establish polynomial bounds for its hitting time. For the latter, we will use Lyapunov functions. The proof of convergence in total variation with rate of convergence basically repeats the calculus in [18] for the Lyapunov functions $L_m(X)$ and $L_{k,m}(t, X)$ from (12) with some changes, and on Dynkin's formulae (10) and (11) due to the Corollary 2. Without big changes, this calculus provides a polynomial moment bound

$$\mathbb{E}_X \tau_0^k \leq C L_m(X) \leq C(n + 1 + x + y)^m, \tag{21}$$

for certain values of k related to the exact value of the constat C_0, and for the hitting time

$$\tau_0 := \inf(t \geq 0 : X_t = (0,0,*).$$

However, it is not the set of idle states $\{(0,0,*)\}$ (i.e., the third component here is arbitrary non-negative) that will be a regeneration, but it is just an auxiliary one. Namely, once the process attains the set $\{(0,0,*)\}$, it may be then successfully coupled with another (stationary) version of the same process at their joint jump $\{n = 0\} \mapsto \{n = 1\}$. This is because, in particular, immediately after such a jump the state of each process reads as $(1,0,0)$. Clearly, *this state* $(1,0,0)$ *may be considered as a regeneration one*, and this is despite the fact that the process spends zero time in this state. The news in the calculus in comparison to [18] is that we have to tackle a wider class of intensities, which may be all variable (as well as discontinuous) rather than constant, including λ_0. However – beside a new regeneration state instead of a usual "zero" (idle) – this affects the calculus a little, once it is established that (10) and (11) hold true, because the major part of this calculus involves only time values $t < \tau_0$. Some change is also in the procedure of coupling, though, because at state $(1,0,0)$ the process can only spend zero time, which means that the process "cannot wait" at this state.

In turn, the inequality (21) provides a bound for the rate of convergence, for the justification of which rate there are various approaches such as versions of the coupling method as well as renewal theory. Convergence of probabilities in the definition of m-availability factors is a special case of a more general convergence in total variation. Although the changes in comparison to [18] are, in fact, minor, yet it would be not totally fair to say that a simple reference to this earlier paper may replace a full proof. So, as suggested by the Editors, and for the completeness of this paper, and for the convenience of the reader we outline the details of the proof of the Theorem here.

2. Let us inspect the properties of the functions $L_m(X)$ and $L_{k,m}(t, X)$. Assume that $m > 1$ and the value C_0 in the condition (5) is large enough, namely, C_0 satisifies

$$C_0 > \Lambda 2^{2(m+k)-1} + 2^{m+k}. \tag{22}$$

Recall that $\tau_0 := \inf(t \geq 0 : X_t = (0,0,*))$. Let $X_0 = X \equiv (n, x, y)$. Note that *it suffices to establish the*

estimate (21) for initial states with $n_0 + x > 0$. The reason for this is that in the case of $X_0 = (0,0,y)$, by virtue of the condition (6) irrespectively of the value of y, the time for the process to hit state (1,0,0) does satisfy this estimate – and even a better exponential bound holds true for such particular hitting time and initial state under our conditions – so that in this case we can start the estimate for τ_0 so to say from state (1,0,0). Hence, in the sequel we may and will assume $n_0 > 0$ without losing a generality.

3. Repeating the main steps in the calculus from [18] in our more involved but computationaly very similar situation, we obtain for $X_t \neq (0,0,*)$ (note that $n_t = 0$ is not excluded):

$$dL_m(X_t) = \lambda(X_t)\,((n+2+x_t+0)^m - (n+1+x_t+y_t)^m)\,dt +$$

$$+h(X_t)(n^m - (n+1+x_t+y_t)^m)\,dt +$$

$$+((n+1+x_t+y_t+dt)^m - (n+1+x_t+y_t)^m) + dM_t \equiv$$

$$\equiv (I_1 - I_2 + I_3)dt + dM_t,$$

where

$$I_1\,dt := \lambda(X_t)dt\,((n+2+x_t+y_t)^m - (n+1+x_t+0)^m),$$

$$I_2\,dt := h(X_t)((n+1+x_t+y_t)^m - (n+y_t)^m)\,dt,$$

$$I_3\,dt := ((n+1+x_t+y_t+dt)^m - (n+1+x_t+y_t)^m)$$
$$= m(n+1+x_t+y_t)^{m-1}\,dt,$$

and M_t is a local martingale (see, e.g., [14]). The following bound will be established:

$$(I_1 - I_2 + I_3) \le -CL_{m-1}(X_t), \quad t < \tau_0,$$

with some $C > 0$. The main purpose – beside the convenience of the reader – of the calculus which follows is, indeed, to make sure that there is no new difficulties due to the more involved model in comparison to [18]; in particular, some news is that unlike the paper [18], now to hit the state with $n = 0$ may not be enough, and because of this even the definition of the hitting time τ_0 here is different.

Later on, for the function $L_{m,k}(t,X) = (1+t)^k L_m(X)$ СЃ $k > 0$ under the appropriate condition on C_0 (see (22)) we will show that

$$\mathbb{E}_X L_{m,k}(\tau_0, X_{\tau_0}) \le L_{m'}(X) - c\mathbb{E}_X \tau_0^{k+1},$$

or, equivalently,

$$\mathbb{E}_X L_{m,k}(\tau_0, X_{\tau_0}) + c\mathbb{E}_X \tau_0^{k+1} \le L_{m'}(X),$$

with some $m' > m$ and $c > 0$, which would suffice for the desired result.

For any $m > 1$, $a \ge 1$ we have,

$$(a+1)^m - a^m = m\int_0^1 (a+s)^{m-1}\,ds \le m2^{m-1}a^{m-1}.$$

It follows that

$$I_1 \le m2^{m-1}\Lambda(n+1+x_t+y_t)^{m-1} \equiv m2^{m-1}\Lambda L_{m-1}(X_t).$$

Further,

$$I_2 \ge C_0(1+x_t)^{-1}((n+1+x_t+y_t)^m - (n+y_t)^m)$$

$$= C_0(1+x_t)^{-1}\int_0^1 m(n+y_t+s(1+x_t))^{m-1}(1+x_t)\,ds$$

$$\ge C_0 m\int_{1/2}^1 (n+y_t+\tfrac{1}{2}(1+x_t))^{m-1}\,ds$$

$$\ge C_0 m2^{-m}(n+1+x_t+y_t)^{m-1} \equiv C_0 m2^{-m}L_{m-1}(X_t).$$

Finally,

$$I_3 = m\,L_{m-1}(X_t).$$

So, we get,

$$I_1 - I_2 + I_3 \le m2^{m-1}\Lambda L_{m-1}(X_t) - C_0 m2^{-m} L_{m-1}(X_t) + m\, L_{m-1}(X_t).$$

Here if C_0 is large enough, namely, if

$$m2^{-m}C_0 > m2^{m-1}\Lambda + m \quad \sim \quad C_0 > 2^{2m-1}\Lambda + 2^m \tag{23}$$

(clearly, (23) is weaker than (22)), then the sum $I_1 - I_2 + I_3$ is strictly negative:

$$I_1 - I_2 + I_3 \le -m(2^{-m}C_0 - 2^{m-1}\Lambda - 1)\, L_{m-1}(X_t) < 0.$$

Note that (22) in full generality will be used in the sequel. Now, by virtue of Fatou's Lemma – if necessary with an appropriate localizing sequence – we obtain,

$$\mathbb{E}_X L_m(X_{t\wedge\tau_0}) + (m2^{-m}C_0 - m2^{m-1}\Lambda - m)\mathbb{E}_X \int_0^{t\wedge\tau_0} L_{m-1}(X_s)\, ds \le L_m(X),$$

and also

$$\mathbb{E}_X L_m(X_{\tau_0}) + (m2^{-m}C_0 - m2^{m-1}\Lambda - m)\mathbb{E}_X \int_0^{\tau_0} L_{m-1}(X_s)\, ds \le L_m(X). \tag{24}$$

From (24) it follows that, in particular, $\mathbb{E}_x\tau_0 < \infty$ (since $L_{m-1} \ge 1$), from which it may be concluded due to Harris–Khasminsky's principle that there exists a stationary measure (see [11]); in our case it is clearly unique (e.g., because of the convergence to *any* stationary measure, which follows from this proof); moreover,

$$\int_X L_{m-1}(X)\, \mu(dX) < \infty. \tag{25}$$

Also, by Hölder's inequality, for each $t \ge 0$,

$$\mathbb{E}_X L_{m'}(X_{t\wedge\tau_0}) \le L_{m'}(X), \quad \forall \quad m' \le m. \tag{26}$$

4. Note that the bound (26) has been established under the condition (23). Similarly, if it were known that C_0 satisfies

$$C_0 > 2^{2(m+\ell)}\Lambda + 2^{m+\ell}, \tag{27}$$

then we would be able to conclude that also

$$\mathbb{E}_X L_{m'}(X_{t\wedge\tau_0}) \le L_{m'}(X), \quad \forall \quad m' \le m + \ell. \tag{28}$$

for each $m' \le m + \ell$. In turn, if we need (27) and (28) for *some* ℓ greater than k – let arbitrarily close to k – then (22) suffices for this.

5. Now let us inspect the function $L_{m,k}(t,X) = (1+t)^k L_m(X)$ with $k > 0$ under the assumption (22). We have, similarly to the step 1,

$$dL_{m,k}(t,X_t) = (1+t)^k[I_1 - I_2 + I_3]\, dt + d\widetilde{M}_t$$

$$+k(1+t)^{k-1}\, L_m(X_t)\, dt$$

$$\le -(1+t)^k m(2^{-m}C_0 - \Lambda - 1)L_{m-1}(X_t)\, dt$$

$$+k(1+t)^{k-1}\, L_m(X_t)\, dt + d\widetilde{M}_t,$$

with some new local martingale $\widetilde{M}_t$. The second term $I_4 := k(1+t)^{k-1}\, L_m(X_t)$ may be split into two parts, $I_4 = I_5 + I_6$, where

$$I_5 := k(1+t)^{k-1}\, L_m(X_t)\, 1(k(1+t)^{k-1}\, L_m(X_t) \le \varepsilon(1+t)^k L_{m-1}(X_t)),$$

$$I_6 := k(1+t)^{k-1}\, L_m(X_t)\, 1(k(1+t)^{k-1}\, L_m(X_t) > \varepsilon(1+t)^k L_{m-1}(X_t)),$$

where $1(A)$ stands for the indicator of the event A. The term I_5 is clearly dominated by the main negative expression $-I_2$ in the sum $I_1 - I_2 + I_3$, if we put $\varepsilon < m(2^{-m}C_0 - 2^{m-1}\Lambda - 1)$. Let us now estimate the term I_6. For any $\ell > 0$ and $\varepsilon > 0$,

$$I_6 \le I_4\, \frac{(k\, L_m(X_t))^\ell}{(\varepsilon(1+t)L_{m-1}(X_t))^\ell} = I_4\, \frac{k^\ell}{(\varepsilon(1+t))^\ell} L_\ell(X_t).$$

So, I_6 does not exceed the value

$$k(1+t)^{k-1}\, \frac{k^\ell}{(\varepsilon(1+t))^\ell} L_{m+\ell}(X_t) \quad .$$

Let $\ell = k + \delta$ and assume that the value ℓ satisfies the condition (27). Recall that this is always possible if C_0 satisfies (22) and $\delta > 0$ is small enough. Then, due to (28) and, if necessary, by using a new auxiliary localizing sequence of stopping times with Fatou's Lemma we get,

$$\mathbb{E}_X L_{m,k}(t\wedge\tau_0, X_{t\wedge\tau_0})$$

$$+(m(2^{-m}C_0 - 2^{m-1}\Lambda - 1) - \varepsilon)\mathbb{E}_X \int_0^{t\wedge\tau_0} (1+s)^k L_{m-1}(X_s)\, ds \le$$

$$\le L_m(X) + C'\mathbb{E}_X \int_0^{\infty} \mathbb{E}_X 1(s \le t\wedge\tau_0)(1+s)^{k-1-\ell} L_{m+\ell}(X_s)\, ds \le$$

$$\le L_m(X) + C'\mathbb{E}_X \int_0^{\infty} (1+s)^{k-1-\ell}\mathbb{E}_X L_{m+\ell}(X_{s\wedge t\wedge\tau_0})\, ds \le$$

$$\le L_m(X) + C'' L_{m+\ell}(X) \le C''' L_{m+\ell}(X).$$

Again, by virtue of Fatou's Lemma this imples,

$$\mathbb{E}_X L_{m,k}(\tau_0, X_{\tau_0}) + C'\mathbb{E}_X \int_0^{\tau_0} (1+s)^k L_{m-1}(X_s)\, ds \le C''' L_{m+\ell}(X).$$

Since $L_{m-1}(X_s) \ge 1(s < \tau_0)$, we obtain,

$$\mathbb{E}_X \tau_0^{k+1} \le C L_{m+\ell}(X), \tag{29}$$

with some new constant $C > 0$, which does admits some effective bound similarly to all earlier constants.

6. The estimate (29) – along with the remark about an exponential moment for time to hit state (1,0,0) starting from state (0,0,∗) mentioned eralier – suffices for the desired inequality, and there are various ways to show it, including the coupling method (cf., e.g., [15, 17]), or renewal theory (see, e.g., [3]). Hence, for many readers a recommendation would be to stop reading here. However, for the convenience of the wider audience (as well as simply for the sake of completeness) we will now briefly recall the scheme of the coupling method mentioned earlier about how the proof may be completed without any big theory "by hand". Let (X_t) and $(\tilde{X}_t)$ be *two* independent copies of our Markov process where the first process starts at $X_0 = X$, while the second has a stationary initial distribution μ, which *existence* was mentioned earlier. (At the moment uniqueness is not proved, so we let *any* stationary distribution if there are more than one.) Denote $\bar{\tau}_0 := \inf(t \ge 0 : X_t = (0,0,*), \& \tilde{X}_t = (0,0,*))$ (the third components may be equal or different). Quite similarly to (29), the inequality

$$\mathbb{E}_X \bar{\tau}_0^{k+1} \le C L_{m+\ell}(X), \tag{30}$$

can be established, see, e.g., [18], with a new constant C, which also admits some effective bound. The proof follows from integration and from the fact that $\int L_{m+\ell}\, d\mu < \infty$ – which integral also allows some effective bound – due to (25), the latter guaranteed by the choice of large enough value of C_0, see (22).

7. Finally, by the *coupling inequality* ("c.i.") (see, for exampe, [16]),

$$|(\mu_t^X - \mu)(A)| = |\mathbb{E}_X(1(X_t \in A) - 1(\tilde{X}_t \in A))|1(t \ge \bar{\tau}_0)$$

$$+|\mathbb{E}_X(1(X_t \in A) - 1(\tilde{X}_t \in A))|1(t < \bar{\tau}_0)$$

$$\overset{c.i.}{\le} \mathbb{E}_X 1(t < \bar{\tau}_0) = \mathbb{P}_X(t < \bar{\tau}_0) \le \frac{\mathbb{E}_X \bar{\tau}_0^{k+1}}{t^{k+1}} \le \frac{C L_{m+\ell}(X)}{t^{k+1}}.$$

Note that, in particular, the uniqueness of the stationary distribution follows from this convergence. Finally, by the definition of the total variation $\| \mu_t^X - \mu \|_{TV} := 2\sup_A(\mu_t^X - \mu)(A)$, and hence, the obtained inequality (30) provides the claim of the Theorem.

References

[1] *Asmussen, S.*, Applied Probability and Queues, 2nd edition, Springer, Berlin et al. (2003).

[2] *Bambos, N., Walrand, J.*, On stability of state-dependent queues and acyclic queueing networks, Adv. Appl. Probab. 21(3) (1989), 681–701.

[3] *Borovkov, A. A.*, Asymptotic Methods in Queueing Theory, Chichester, NY: J. Wiley, 1984.

[4] *Borovkov, A. A., Boxma, O. J., Palmowski, Z.*, On the Integral of the Workload Process of the Single Server Queue, Journal of Applied Probability, 40(1) (2003), 200–225.

[5] *Bramson, M.*, Stability of Queueing Networks, École d'Été de Probabilités de Saint-Flour XXXVI-2006, Lecture Notes in Math., Vol. 1950 (2008).

[6] *Brémaud, P., Lasgouttes, J.-M.*, Stationary IPA estimates for nonsmooth $G/G/1/\infty$ functionals via palm inversion and level-crossing analysis Discrete Event Dynamic Systems, 3(4) (1993) 347-374.

[7] *Brémaud, P., Lasgouttes, J.-M.*, Stationary IPA estimates for nonsmooth $G/G/1/\infty$ functionals via palm inversion and level-crossing analysis. A preprint version 2012 of the paper [6], http://arxiv.org/pdf/1207.3241v1.pdf

[8] *Dynkin, E. B.*, Markov processes, V. I, Springer-Verlag, Berlin-Göttingen-Heidelberg (1965).

[9] *Fakinos, D.*, The Single-Server Queue with Service Depending on Queue Size and with the Preemptive-Resume Last-Come-First-Served Queue Discipline, Journal of Applied Probability, 24(3) (1987), 758–767.

[10] *Gnedenko, B. V., Kovalenko, I. N.*, Introduction to queueing theory. 2nd ed., rev. and suppl. Boston, MA et al., Birkhäuser (1991).

[11] *Hasminskii, R. Z.*,, Stochastic Stability of Differential Equations, Dordrecht, The Netherlands: Sijthoff & Noordhoff (1980).

[12] *Kim, B., Kim, J.* A note on the subexponential asymptotics of the stationary distribution of M/G/1 type Markov chains, European Journal of Operational Research, 220(1), (2012), 132-134.

[13] *Kimura, T., Masuyama, H., Takahashi, Y.* Subexponential Asymptotics of the Stationary Distributions of GI/G/1-Type Markov Chains, arXiv:1410.5554v3 (June 2016).

[14] *Liptser, R. Sh., Shiryaev, A. N.*, Stochastic calculus on filtered probability spaces, in: S. V Anulova, A. Yu. Veretennikov, N. V Krylov, et al., Stochastic calculus, Itogi Nauki i Tekhniki, Modern problems of fundamental math. directions, Moscow, VINITI (1989), 114–159 (in Russian); Engl. transl.: Probability Theory III, Stochastic Calculus, Yu. V. Prokhorov and A. N. Shiryaev Eds., Springer (1998), 111–157.

[15] *Thorisson, H.*, The queue $GI/G/1$: finite moments of the cycle variables and uniform rates of convergence, Stoch. Proc. Appl. 19(1) (1985), 85–99.

[16] *Thorisson, H.*, Coupling, Stationarity, and Regeneration, New York, NY: Springer, 2000.

[17] *Veretennikov, A. Yu.*, On polynomial mixing and convergence rate for stochastic difference and differential equations, Theory Probab. Appl. 45(1), 160–163 (2001).

[18] *Veretennikov, A. Yu.*, On the rate of convergence to the stationary distribution in the single-server queuing system, Autom. Remote Control 74(10), 1620-1629 (2013).

[19] *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, 479-504 (2014).