STABILITY ANALYSIS OF REGENERATIVE QUEUES

EVSEY MOROZOV[1] AND ROSARIO DELGADO[2]

1. Introduction

One of the topics which were intensively studied in the last decade is the stability of non-Markovian queueing systems. It is well-known that stability is one of the hard and actual problems and requires refinement and laborious mathematical technique especially outside the limits of Markovian queues. Stability analysis establishes the region of predefined parameters where the stability of the basic process holds. Various notions of stability are applied. We mention *weak* and *strong stability*, Chen [4], *global weak stability, global pathwise stability*, Dai and Vande Vate [10], and so on.

An effective and developed approach to stability analysis of a wide class queueing systems and networks is the *fluid approximation*. Among many works which treat this topic we mention Chen and Mandelbaum [6], Chen [4], Chen and Yao [5], Dai [7], Dai [8], Dai and Weiss [11], Dai and Vande Vate [10].

At the same time, the fluid approach is not direct in the sense that we study originally the stability/instability of the associated *fluid limit model* (and deal with deterministic fluid processes instead of original stochastic ones) to establish the similar property of the corresponding queueing process.

The most recent overview on stability analysis methods (with focus on networks) is the paper [12].

Unlike the mentioned above approaches, our approach to the stability is based on the *regeneration property* of the basic queueing process [1, 32].

We focus on the regenerative queues since they have numerous applications (for instance, [30, 31]). Also the regeneration of Harris recurrent Markov chains extends an area of this approach, [1]. The notable monograph [16] contains detailed description of stability analysis of Markov chains. For Markovian setting, this approach has paralellism with the one described in

[1] Supported by Russian Foundation for Basic Research, Grant 07-07-00088.
[2] Supported by Grant MEC-FEDER ref. MTM2006-06427.

this work. For instance, positive recurrence of renewal process of regenerations introduced below is similar to positive Harris recurrence of a Markov chain.

One more important feature of the approach is that we separate the predefined assumptions on the *negative drift* and *regeneration* assumptions. The latter guarantees the appearance of a *regeneration* with a positive probability regardless of the initial state in a compact set. (This assumption is weaker than Harris recurrence for which this positive probability must be one.) We note that negative drift typically does not allow to obtain stability directly and regeneration condition turns out to be very effective at the intermediate steps of the analysis, Morozov [26]. Also we mention an important role which the *tightness* of the stochastic processes plays in our stability analysis.

Another advantage is that our approach to stability is unified and allows to cover the multiserver queues where servers may be nonidentical. This case is especially difficult to be investigated because such queues are not monotone (unlike conventional queues). The reason of this difficulty is that the service times depend on server number in a node.

An important property of the approach is that in many cases it is extended to arbitrary initial state of the basic queueing process.

Finally, the proposed approach to stability analysis uses completely another technique based on a characterization of the renewal process of the regenerations and the asymptotic behavior of the forward renewal time. This work illustrates new possibilities of a unified approach to stability analysis of regenerative queueing processes developed earlier in [25]–[29].

Now we describe our approach to stability analysis.

Throughout this work (except in Section 5.2) we deal with a *renewal* input process that describes the arrival of customers to a system or queue, whose arrival instants are denoted by $(t_n)_{n\geq 0}$. We denote by $\tau_n = t_{n+1} - t_n > 0$, $n \geq 0$ the i.i.d. interarrival times. Let τ be a generic random variable with the distribution of the interarrival times. Assume that $\mathsf{E}\,\tau = \frac{1}{\lambda} \in (0, \infty)$; t_0 is the arrival instant of customer 0; if $t_0 = 0$ then the input process is *zero-delayed*. Otherwise, the delay $t_0 \geq 0$ may have another distribution diferent of τ, and we have a *delayed input*; this latest situation is only considered in Section 5.3. Consider the *forward renewal time* process for the input process, $\{\tau(t),\, t \geq 0\}$, defined by:

$$\tau(t) \stackrel{\text{def}}{=} \min\{\, t_k - t : t_k - t \geq 0 \,\}. \tag{1}$$

This process is tight (as proved in [26]), that is, for any fixed $\delta > 0$, a constant $C > 0$ exists such that

$$\inf_{t\geq 0} \mathsf{P}\left(\tau(t) \leq C\right) \geq 1-\delta\,.$$

We study the stability of some queueing systems. More specifically, in Sections 2 and 4 a single-server $GI/G/1$ queue is considered, while sections 3 and 5 deal with the multiple-server $GI/G/m$ queue with $m > 1$ servers, which are assumed to be identical in Section 3. We assume that a single queue or waiting line forms in front of the system (composed by 1 or m servers), and that the customer at the head of the queue will be handled by the first available server, by following a FIFS (first-in-first-served) service discipline, which moreover is assumed to be a no-idling (or work conserving) policy, that means that servers are never idle while there are customers waiting to be served at the waiting line.

Except for the Section 5.1, in all the work service times $\{S_n^{(k)},\, n \geq 0,\, k = 1, \ldots, m\}$ are assumed to be i.i.d. ($S_n^{(k)}$ denotes the service time for the nth customer that is handled by server k), with S a generic random variable whose distribution is that of any $S_n^{(k)}$, assumed to be > 0 w.p.1, and with expectation $\mathsf{E}\,S = \frac{1}{\mu} \in (0,\, \infty)$. In the single-server case, we drop the superscript in $S_n^{(1)}$ and just write S_n. In the multi-server case with identical servers, S_n is used to denote the service time (with the same distribution as S) of the nth customer arriving to the system, and this notation is coherent with that used in the single-server case. In this context (except for Sections 5.1 and 5.2), let we introduce the parameter

$$\rho \stackrel{\text{def}}{=} \frac{\mathsf{E}\,S}{\mathsf{E}\,\tau} = \frac{\lambda}{\mu}\,,$$

which can be interpreted as the *traffic intensity* through the $GI/G/1$ queue introduced before (but it also has a meaning for the multi-server queue).

In both cases, $GI/G/1$ and $GI/G/m$ (with $m > 1$) queues, as a measure of congestion of the system we can introduce the (*regenerative*) *queue size* process $\nu = \{\nu(t),\, t \geq 0\}$, where $\nu(t) \in \mathbb{N} \cup \{0\}$ is the number of customers in the system (at the waiting line or being served) at instant t. We denote by $\nu(0^-)$ the accumulated customers in the system at instant $t = 0$, and consider the *regenerative* embedded sequence $(\nu_n)_{n\geq 0}$ defined by

$$\nu_n \stackrel{\text{def}}{=} \nu(t_n^-)\left(= \lim_{t\to t_n^-} \nu(t) \text{ if } t_n > 0\right) \quad \text{if } n \geq 0\,.$$

By definition, ν_n is the number of customers in the system just when customer nth arrives. The embedded process $(\nu_n)_{n\geq 0}$ is also *regenerative*, and

its associated *renewal* process will be denoted by $\beta = (\beta_n)_{n\geq 0}$ (for any $n > 0$, the post-β_n−process $(\nu_k)_{k\geq\beta_n}$ is independent of the pre-history $(\nu_k)_{k<\beta_n}$, and its distribution does not depend upon n.) The renewal process β describes the times of occurrences of "regenerations" of process $(\nu_n)_{n\geq 0}$, that is, the end of a cycle and the beginning of the next one, and is defined by

$$\beta_{n+1} \stackrel{\text{def}}{=} \min\{\, k > \beta_n \,:\, \nu_k = 0\,\} \quad \text{for any } n \geq 0\,, \quad \beta_0 \stackrel{\text{def}}{=} 0\,. \tag{2}$$

The corresponding *renewal* process of regenerations for the *queue size* process ν is $(T_n)_{n\geq 0}$, with

$$T_{n+1} \stackrel{\text{def}}{=} t_{\beta_{n+1}} \Big(= \min\{\, t_k > T_n \,:\, \nu_k = 0\,\} \Big) \text{ for any } n \geq 0\,, \quad T_0 \stackrel{\text{def}}{=} t_0\,. \tag{3}$$

We have that $T_1 = t_0 + \tau_0 + \cdots + \tau_{\beta_1 - 1}$, and then $T_1 - T_0 = \tau_0 + \cdots + \tau_{\beta_1 - 1}$. If the input process is zero-delayed, $t_0 = 0$, then we have

$$T_0 = 0 \quad \text{and} \quad T_1 = \tau_0 + \cdots + \tau_{\beta_1 - 1}\,. \tag{4}$$

We refer to this situation as the "zero-delayed case" throughout the paper.

Jointly with the *queue size* process ν it is customary to consider another measure of congestion of the system, the *workload* process $W = \{W(t),\, t \geq 0\}$, with state space $(E = [0,\, \infty),\, \mathcal{E} = \mathcal{B}([0,\, \infty)))$ and whose paths are continuous on the right on $[0,\, \infty)$ and with limits on the left on $(0,\, \infty)$. Workload is defined in this way: for any $t \geq 0$, $W(t)$ is the amount of time needed for the system to complete service of all customers in queue at the waiting line or being served, at time t.

We denote by $W(0^-)$ the accumulated workload at instant $t = 0$, and consider the embedded sequence $(W_n)_{n\geq 0}$, where

$$W_n \stackrel{\text{def}}{=} W(t_n^-) \Big(= \lim_{t \to t_n^-} W(t) \text{ if } t_n > 0 \Big) \quad \text{if } n \geq 0\,.$$

Note that with this definition W_n is the waiting time in queue of customer n. Throughout the paper we will use a sample path relationship between waiting times W_n and W_{n+1} known as *Lindley's recursion* (see Example III.6.1 in [1]):

$$W_{n+1} = \Big(W_n - \big(\tau_n - S_n \big) \Big)^+ \tag{5}$$

where x^+ denotes $\max(x,\, 0)$.

Workload process W is also a *regenerative* process with the same *renewal* process of "regenerations" $(T_n)_{n\geq 0}$, and the *renewal* process of "regenerations" associated with the embedded sequence $(W_n)_{n\geq 0}$ is also process β, as for the queue size process (because $\nu(t) = 0 \Leftrightarrow W(t) = 0$).

We will use notation E_0 to denote expectation in the zero-delayed case, in which the first regeneration cycle lengths are β_1 for the discrete time embedded processes, and T_1 for the continuous time processes ν and W. As a rule, it is implicit from the context that $\mathsf{E} = \mathsf{E}_0$ but sometimes we use this notation in an explicit form. The renewal process β is called *positive recurrent* if $\beta_1 < \infty$ with probability 1 and moreover expectation of the first cycle length, which is β_1 in the zero-delayed case, is finite. That is, if

$$\boxed{\beta_1 < \infty \quad \text{w.p.1} \quad \text{and} \quad \mathsf{E}_0\, \beta_1 < \infty}$$

Analogously, the renewal processes of regenerations of ν and W, $(T_n)_{n\geq 0}$, is called *positive recurrent* if

$$\boxed{T_1 < \infty \quad \text{w.p.1} \quad \text{and} \quad \mathsf{E}_0\, T_1 < \infty}$$

We introduce the *forward renewal time* (or "unfinished time to renewal") for the discrete time processes:

$$\beta(n) \stackrel{\text{def}}{=} \min\{\, \beta_k - n : \beta_k - n > 0 \,\} \text{ for any } n \geq 0\,.$$

And the analogous for the continuous time:

$$T(t) \stackrel{\text{def}}{=} \min\{\, T_k - t : T_k - t > 0 \,\} \text{ for any } t \geq 0\,.$$

Note that $\beta(0) = \beta_1$ and $T(0) = T_1$ in the zero-delayed case.

We use the following dichotomy describing the asymptotic behavior of $\beta(n)$ and $T(t)$ (see [14]):

$$\mathsf{P} - \lim_{n\to\infty} \beta(n) = \infty \Leftrightarrow \mathsf{E}_0\, \beta_1 = \infty, \tag{6}$$

$$\mathsf{P} - \lim_{t\to\infty} T(t) = \infty \Leftrightarrow \mathsf{E}_0\, T_1 = \infty. \tag{7}$$

(Notation $\mathsf{P} - \lim$ means convergence in probability.) The main idea of our approach to stability analysis is to prove that convergence in the probability sense of $\beta(n)$ (or $T(t)$) to ∞ does not hold. Therefore, by (6) (or (7), respectively), we obtain the finiteness of the corresponding expectation $\mathsf{E}_0\beta_1 < \infty$ (or $\mathsf{E}_0 T_1 < \infty$). It is sufficient to obtain one of them for having the other, because by (4) and Wald's identity,

$$\mathsf{E}_0\, T_1 = \mathsf{E}\,\tau\ \mathsf{E}_0\, \beta_1 = \frac{1}{\lambda}\, \mathsf{E}_0\, \beta_1\,. \tag{8}$$

After that, we prove the finiteness with probability 1 of β_1 for having *positive recurrence* of $(\beta_n)_{n\geq 0}$, and we can apply Corollary VI.1.5 [1] if the distribution of the first regeneration cycle length β_1 is aperiodic, to

obtain the weak convergence of the discrete time processes to a limiting distribution, say π for $(\nu_n)_{n\geq 0}$, that implies in particular

$$\lim_{n\to\infty} \mathsf{P}(\nu_n \in A) = \frac{\mathsf{E}_0\Big(\sum\limits_{k=0}^{\beta_1-1} \mathbb{I}_{(\nu_k\in A)}\Big)}{\mathsf{E}_0\,\beta_1} \,(=\pi(A)\,) \quad \text{for any } A\subset \mathbb{N}\cup\{0\}\,, \tag{9}$$

where $\mathbb{I}_{(\cdot)}$ denotes the indicator function. Similarly for the continuous-time situation: once we have proved that process $(T_n)_{n\geq 0}$ is *positive recurrent*, we can apply Theorem VI.1.2 [1] if the distribution of the first regeneration cycle length T_1 is non-lattice, to obtain the weak convergence of the processes ν and W to a limiting distribution, say $\tilde{\pi}$ for ν, that in particular gives

$$\lim_{t\to\infty} \mathsf{P}(\nu(t) \in A) = \frac{\mathsf{E}_0\Big(\int\limits_{0}^{T_1} \mathbb{I}_{(\nu(t)\in A)}\Big)}{\mathsf{E}_0\,T_1} \,(=\tilde{\pi}(A)\,) \quad \text{for any } A\subset \mathbb{N}\cup\{0\}\,.$$

Note that all renewal process of regenerations for the discrete-time processes below are aperiodic, and thus *positive recurrence* implies the stability in the sense of convergence to the limit distribution (9). In consequence, our objective will be to prove *positive recurrence* of the renewal process β by checking that $\beta_1 < \infty$ w.p.1, and that P- $\lim\limits_{n\to\infty} \beta(n) \neq \infty$, which is equivalent to say that constants L, $\varepsilon > 0$ and non-random discrete instants $(n_i)_{i\geq 1}$ with $\lim\limits_{i\to\infty} n_i = \infty$ exist, such that

$$\inf_{i\geq 1} \mathsf{P}\left(\beta(n_i) \leq L\right) \geq \varepsilon\,. \tag{10}$$

Analogously for the renewal process $\{T(t),\, t\geq 0\}$: if we see that $T_1 < \infty$ w.p.1, and that constants b, $\varepsilon > 0$ and a sequence of non-random instants $(z_i)_{i\geq 1}$ with $\lim\limits_{i\to\infty} z_i = \infty$ exist, such that

$$\inf_{i\geq 1} \mathsf{P}\left(T(z_i) \leq b\right) \geq \varepsilon\,, \tag{11}$$

we will have proved its *positive recurrence*.

Throughout the work we treat different queueing models to which can be applied our techniques to find sufficient conditions for having stability in the sense explained before. The organization of the paper is as follows: in Section 2 we treat the standard single-server $GI/G/1$ queue, and in Section 3 the standard multi-server $GI/G/m$ queue, with $m > 1$.

Two extensions of the standard $GI/G/1$ queue are considered in Section 4: a queue with (partially) impatient customers, and the state-dependent case, from which the impatient customers situation is a particular case.

Finally, the extensions of the standard multi-server $GI/G/m$ queue considered in Section 5 are three: non-identical servers, regenerative input $R/G/m$, and the delayed-case. The first one only has sense in the $m > 1$ case, but the other two extensions are valid for $m \geq 1$, that is, include the single-server queue.

2. Stability analysis of a GI/G/1 queue

We start our stability study by considering a standard single-server $GI/G/1$ queue, and using notations already introduced in Section 1. Our main *negative drift assumption* is

$$\rho \left(= \frac{\lambda}{\mu} \right) < 1 \,, \tag{12}$$

which is equivalent to $\mathsf{E}(\tau - S) > 0$, and implies the *regeneration condition*

$$\mathsf{P}(\tau > S) > 0. \tag{13}$$

Theorem 1. *Under assumption (12), the renewal processes of regenerations $\beta = (\beta_n)_{n\geq 0}$ and $(T_n)_{n\geq 0}$ are positive recurrent, that is,*

$$\mathsf{E}_0\, \beta_1 < \infty \,, \quad \mathsf{E}_0\, T_1 < \infty \quad \textit{and} \tag{14}$$

$$\beta_1 < \infty \ \ w.p.1 \,, \quad T_1 < \infty \ \ w.p.1 \,, \tag{15}$$

regardless of the initial state $W0^-$).

Proof. Introduce the (non-negative) idle time for the server in the interval $[0,\, t]$

$$\mu(t) = \int_0^t \mathbb{I}_{(\nu(s)=0)}\, ds \quad \text{for any } t \geq 0 \,.$$

Let

$$N(t) = \#\{\, k = 0, 1, \ldots, : t_k \leq t \,\} = \min\{\, k \geq 0 \,:\, t_k > t \,\} \quad \text{for } t \geq 0 \,,$$

be the number of arrivals in interval $[0,\, t]$ (including that of customer 0.) In particular, $N(t_n) = n + 1$. In the zero-delayed case, $N(t) \geq 1$ for all $t \geq 0$ and $N(0) = 1$. Denote the total workload arriving in interval $[0, t]$ by

$$V(t) \stackrel{\text{def}}{=} \sum_{n=1}^{N(t)} S_{n-1} \,. \tag{16}$$

Obviously,

$$W(0^-) + V(t) = t - \mu(t) + W(t) \geq t - \mu(t)\,, \tag{17}$$

which implies

$$\mu(t) \geq t - V(t) - W(0^-) = t - V(t) + o(t) \quad \text{as} \quad t \to \infty\,.$$

Since

$$V(t)/N(t) \to \mathsf{E}\,S = \frac{1}{\mu},\ N(t)/t \to \frac{1}{\mathsf{E}\,\tau} = \lambda$$

by the Strong Law of Large Numbers (SLLN), then

$$\liminf_{t\to\infty} \frac{\mu(t)}{t} \geq 1 - \lim_{t\to\infty} \frac{V(t)}{N(t)} \frac{N(t)}{t} = 1 - \rho\,, \tag{18}$$

and by the negative drift assumption (12),

$$\liminf_{t\to\infty} \frac{\mu(t)}{t} > 0\,, \tag{19}$$

and consequently,

$$\mu(t) \to \infty \tag{20}$$

w.p.1. Because $\mu(t) \geq 0$, by Fatou's lemma we also have

$$\liminf_{t\to\infty} \frac{\mathsf{E}\,\mu(t)}{t} \Big(\geq \mathsf{E}\big(\liminf_{t\to\infty} \frac{\mu(t)}{t} \big) \Big) > 0\,. \tag{21}$$

Thus, by taking into account that $\mathsf{E}\,\mu(t) = \displaystyle\int_0^t \mathsf{P}(\nu(s) = 0)\,ds$, we have that

$$\liminf_{t\to\infty} \frac{1}{t} \int_0^t \mathsf{P}(\nu(s) = 0)\,ds > 0\,. \tag{22}$$

It is immediate then that $\mathsf{P}(\nu(t) = 0) \not\to 0$ as $t \to \infty$, and this means that a non-random sequence $z_i \to \infty$ and a constant $\delta > 0$ exist such that

$$\inf_{i\geq 1} \mathsf{P}(\nu(z_i) = 0) \geq \delta\,. \tag{23}$$

From the tightness property of the forward renewal time process for the input process $\{\tau(t),\, t \geq 0\}$ defined in (1) and from (23) we have that a constant $b > 0$ exists such that

$$\inf_{i\geq 1} \mathsf{P}(\nu(z_i) = 0,\, \tau(z_i) \leq b) \geq \frac{\delta}{2}. \tag{24}$$

Thus, for any i,

$$\mathsf{P}(T(z_i) \leq b) \geq \mathsf{P}(\nu(z_i) = 0,\, \tau(z_i) \leq b) \geq \frac{\delta}{2} = \varepsilon\,.$$

Hence, we have proved (11) and then $T(t)$ does not converge in probability to ∞, as $t \to \infty$. Therefore $\mathsf{E}_0\, T_1 < \infty$ and also $\mathsf{E}_0\, \beta_1 < \infty$ by (8), and (14) is proved.

Consider now first regeneration period T_1 in more detail. For the zero-delayed case, finiteness $T_1 < \infty$ w.p.1 follows from $\mathsf{E}_0\, T_1 < \infty$. Otherwise, by using that $\mu(t) \to \infty$ w.p.1 (see (20)), we can define a random instant $\hat{t} = \inf(t > 0 : \mu(t) > 0)$ which is finite w.p.1. Then (potential) initial busy period ends not later than $\hat{t}$, and a regeneration occurs not later next (after instant $\hat{t}$) arrival, that is, it occurs within interval $[0,\, \hat{t} + \tau(\hat{t}\,)]$, and thus

$$T_1 \le \hat{t} + \tau(\hat{t}\,) < \infty\,.$$

Moreover, discrete-time 1st regeneration period is also finite since

$$\beta_1 \le N(T_1) < \infty,$$

and with that we also have proved (15) and the proof of the theorem is finished. □

2.1. **Discrete-time approach to stability of GI/G/1 queue.** Now we present a discrete-time approach to stability for the same $GI/G/1$ queue and by using it we give an alternative proof for a part of Theorem 1, the one corresponding to (14), that is, we prove:

Under assumption (12),

$$\mathsf{E}_0\, \beta_1 < \infty \quad \textit{and} \quad \mathsf{E}_0\, T_1 < \infty \tag{25}$$

regardless of the initial state $W_0 = x$.

Remark 1. The discrete-time approach generally does not allow to prove finiteness of the 1st regeneration periods (15), that is, that

$$\beta_1 < \infty \ \text{ w.p.1} \quad \text{and} \qquad T_1 < \infty \ \text{ w.p.1}$$

unlike continuous-time approach used above.

Proof. As the basic process in this discrete-time approach, we consider the waiting time sequence $(W_n)_{n\ge 0}$, which is introduced in Section 1. Define $U_n = S_n - \tau_n$. Note that therefore $(U_n)_{n\ge 0}$ is a sequence of i.i.d. random variables and denote a generic variable for sequence U_n as U, with distribution function F_U. Then, by the Lindley's recursion (5) we have

$$W_{n+1} = (W_n + S_n - \tau_n)^+ = (W_n + U_n)^+,\ n \ge 0\,.$$

Consider increments $\Delta_n = W_{n+1} - W_n$. To estimate expectation $\mathsf{E}\Delta_n$ on the event $\{W_n \in dy\}$, we write down

$$\mathsf{E}(\Delta_n \,|\, W_n = y) = \mathsf{E}\left((y+U)^+ - y\right) = -y\,\mathsf{P}(U \le -y) + \int_{z>-y} z F_U(dz)\,. \tag{26}$$

Since by the negative drift assumption (12) we have $\mathsf{E}\,U = \mathsf{E}(S-\tau) \in (-\infty, 0)$, we see that

$$-y\,\mathsf{P}(U \le -y) \to 0 \quad \text{as} \quad y \to \infty\,,$$

and

$$\int_{z>-y} z\,F_U(dz) \downarrow \mathsf{E}\,U\,(<0) \quad \text{as} \quad y \to \infty\,.$$

Thus, by (26),

$$\lim_{y\to\infty} \mathsf{E}(\Delta_n \,|\, W_n = y) = \mathsf{E}\,U < 0\,. \tag{27}$$

Note that regardless of y,

$$(y+U)^+ - y = (y+S-\tau)^+ - y \le y + S - y = S\,,$$

and then, by (26) again,

$$\mathsf{E}(\Delta_n \,|\, W_n = y) \le \mathsf{E}\,S = \frac{1}{\mu} < \infty \quad \text{for any } \ y\,.$$

As a consequence, it is possible by using (27) to take $y_0 > 0$ (big enough) such that

$$\begin{aligned}\mathsf{E}\,\Delta_n =& \mathsf{E}(\Delta_n \,|\, W_n \le y_0)\,\mathsf{P}(W_n \le y_0) + \mathsf{E}(\Delta_n \,|\, W_n > y_0)\,\mathsf{P}(W_n > y_0) \le \\ \le& \mathsf{E}\,S\ \mathsf{P}(W_n \le y_0) + \frac{\mathsf{E}\,U}{2}\,\mathsf{P}(W_n > y_0)\,. \end{aligned} \tag{28}$$

Now we want to prove $\mathsf{P}-\lim\limits_{n\to\infty} W_n \neq \infty$, and for that we assume that $\mathsf{P}-\lim\limits_{n\to\infty} W_n = \infty$ and will arrive to a contradiction. Indeed, if this limit is ∞, we have that for any y_0 and any $\varepsilon > 0$, n_0 exists such that

$$\mathsf{P}\,(W_n \le y_0) \le \varepsilon \quad \text{for any } \ n \ge n_0\,.$$

Take $\varepsilon < \frac{\mu-\lambda}{\mu+\lambda}$ (that is possible by assumption (12)). Therefore, by (28) we conclude that

$$\mathsf{E}\,W_{n+1} - \mathsf{E}\,W_n = \mathsf{E}\,\Delta_n \le \frac{\varepsilon}{\mu} + \frac{1}{2}\,(\frac{1}{\mu} - \frac{1}{\lambda})\,(1-\varepsilon) < 0 \quad \text{for any } \ n \ge n_0\,,$$

and this implies that for any $n > n_0$,

$$\mathsf{E}\,W_n < \mathsf{E}\,W_{n_0} \le \mathsf{E}\left(\sum_{i=0}^{n_0-1} S_i\right) + x = \frac{n_0}{\mu} + x < \infty,$$

which contradicts the convergence W_n to infinity in probability.

By $\mathsf{P}-\lim\limits_{n\to\infty} W_n \neq \infty$ we have that constants $\delta > 0$ and $T < \infty$, and a (non-random) sequence $(n_i)_{i\ge1}$ with $n_i \to \infty$ exist, such that

$$\inf_{i\ge1} \mathsf{P}(W_{n_i} \le T) \ge \delta\,. \tag{29}$$

By (13) we can choose $\delta_0 > 0$ and $\delta_1 > 0$ such that

$$\mathsf{P}(\tau > S + \delta_0) \geq \delta_1 \,.$$

Denote $L = \lceil \frac{T}{\delta_0} \rceil$ (where, for a real number $x \geq 0$, $\lceil x \rceil$ denotes the minimum positive integer strictly greater than x), and fixing for a moment any $i \geq 1$, consider the event

$$A_i(L) \stackrel{\text{def}}{=} \bigcap_{k=0}^{L-1} \{\tau_{n_i+k} > S_{n_i+k} + \delta_0\}.$$

By the independence of components of the event,

$$\mathsf{P}(A_i(L)) \geq \delta_1^L \,. \tag{30}$$

By Lindley's recursion (5), we have that on $A_i(L)$,

$$\begin{aligned}
&0 \leq W_{n_i+1} = \Big(W_{n_i} - \big(\tau_{n_i} - S_{n_i}\big)\Big)^+ \leq \big(W_{n_i} - \delta_0\Big)^+ \\
&0 \leq W_{n_i+2} = \Big(W_{n_i+1} - \big(\tau_{n_i+1} - S_{n_i+1}\big)\Big)^+ \leq \\
&\qquad\qquad \leq \big(W_{n_i+1} - \delta_0\Big)^+ \leq \big(W_{n_i} - 2\,\delta_0\Big)^+ \\
&\vdots \\
&0 \leq W_{n_i+L} \leq \ldots \leq \Big(W_{n_i} - L\,\delta_0\Big)^+ \leq \Big(W_{n_i} - T\Big)^+ \big(= 0 \text{ in } \{W_{n_i} \leq T\}\big) \,.
\end{aligned}$$

Therefore, on the (independent) events $A_i(L) \cap \{W_{n_i} \leq T\}$, $W_{n_i+L} = 0$, that is, when customer $n_i + L$ arrives meets an empty system and regeneration occurs in interval $[n_i,\, n_i + L]$. In other words, the residual regeneration time at instant n_i must be not greater than L. (Recall that we use discrete-time scale counting arrivals.) As a consequence and by using (30), for any $i \geq 1$ we have

$$\mathsf{P}(\beta(n_i) \leq L) \geq \mathsf{P}(A_i(L) \cap \{W_{n_i} \leq T\}) \geq \delta\,\delta_1^L > 0 \quad \text{for any} \;\; i \geq 1 \,.$$

With $\varepsilon = \delta\,\delta_1^L > 0$, we obtain (10). Hence $\mathsf{P} - \lim\limits_{n\to\infty} \beta(n) \neq \infty$ and thus $\mathsf{E}_0\beta_1 < \infty$. Moreover, we also have finiteness of the expected continuous-time regeneration period T_1 by Wald's identity (8). With that we finish the proof of (25) (that is (14) in Theorem 1) by following the discrete-time approach. □

2.2. **One more approach to stability of GI/G/1 queue in continuous time.** In this subsection, we give an alternative proof of Theorem 1. More precisely, in the proof of Theorem 1, $\mathsf{E}_0\, T_1 < \infty$ and $\mathsf{E}_0\, \beta_1 < \infty$ follow from (21), and $T_1 < \infty$ and $\beta_1 < \infty$ w.p.1 follow from (20). In this Section, we give an alternative proof of (20) and (21). For that, let we define $b(t)$ as the number of customers served in interval $(0,\, t]$, and $b'(t)$ the number of renewals in $[0,\, t]$ in the zero-delayed renewal process generated by service times, that is,

$$b'(t) = \min\{\, k \geq 1 : S_0 + \cdots + S_{k-1} > t \,\} \text{ if } t > 0, \quad b'(0) = 1\,.$$

Note that $b'(\cdot)$ is a non-decreasing function. Thus, $b(t) \leq b'(t)$ and moreover, $b(t) \leq N(t) + \nu(0^-) = N(t) + o(t)$ as $t \to \infty$. By the elementary renewal theorem,

$$\lim_{t\to\infty} \frac{\mathsf{E}\, b'(t)}{t} = \frac{1}{\mathsf{E}\, S} = \mu\,, \tag{31}$$

and by the Strong Law of the Large Numbers (SLLN) for renewal processes,

$$\lim_{t\to\infty} \frac{b'(t)}{t} = \frac{1}{\mathsf{E}\, S} = \mu\,. \tag{32}$$

Since $\lambda < \mu$ by the negative drift assumption (12), and by the SLLN again,

$$\limsup_{t\to\infty} \frac{b(t)}{t} \leq \lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{\mathsf{E}\, \tau} = \lambda < \mu = \lim_{t\to\infty} \frac{b'(t)}{t}\,.$$

Denote

$$\delta(t) \stackrel{\text{def}}{=} \frac{b'(t) - b(t)}{t} \geq 0,$$

so we have seen that

$$\liminf_{t\to\infty} \delta(t) \geq \mu - \lambda > 0\,. \tag{33}$$

Since $\delta(t) \geq 0$, it then follows from Fatou's lemma that also

$$\liminf_{t\to\infty} \mathsf{E}\, \delta(t) \geq \mu - \lambda > 0\,. \tag{34}$$

Denote by F_t the distribution function of $\mu(t)$, that is, $F_t(x) = \mathsf{P}(\mu(t) \leq x)$. A key observation using coupling is that

$$b'(t) - b(t) \leq b'(t) - b'(t - \mu(t)) + 1 \leq_{st} b'(\mu(t)) + 1 \leq b'(\mu(t) + 1) + 1\,. \tag{35}$$

Then

$$\mathsf{E}\, \delta(t) = \frac{1}{t}\mathsf{E}(b'(t) - b(t)) \leq \frac{1}{t} \int_0^t \big(\mathsf{E}\, b'(x) + 1\big)\, F_t(dx). \tag{36}$$

Take any (fixed) $\varepsilon > 0$. By (31), $t_0 = t_0(\varepsilon) > 0$ exists such that

$$\frac{\mathsf{E}\, b'(t)}{t} \leq \mu + \varepsilon \quad \text{for any } t \geq t_0\,. \tag{37}$$

With this t_0 we can split the integral in (36) for any $t \geq t_0$ as

$$\int_0^t = \int_0^{t_0} + \int_{t_0}^t .$$

Obviously,

$$\int_0^{t_0} \mathsf{E}\, b'(x)\, F_t(dx) \leq \mathsf{E}\, b'(t_0) = o(t) \quad \text{as} \;\; t \to \infty .$$

Moreover, using (37) we obtain

$$\begin{aligned} \int_{t_0}^t \mathsf{E}\, b'(x)\, F_t(dx) &= \int_{t_0}^t \frac{\mathsf{E}\, b'(x)}{x}\, x\, F_t(dx) \leq \\ &\leq (\mu+\varepsilon) \int_{t_0}^t x\, F_t(dx) \leq (\mu+\varepsilon)\, \mathsf{E}\, \mu(t) . \end{aligned}$$

Then, by (34),

$$0 < \liminf_{t\to\infty} \mathsf{E}\, \delta(t) \leq (\mu+\varepsilon) \liminf_{t\to\infty} \frac{\mathsf{E}\, \mu(t)}{t},$$

and hence,

$$\liminf_{t\to\infty} \frac{\mathsf{E}\, \mu(t)}{t} > 0 ,$$

that is, we have proved (21).

Now we are going to prove (20). First of all, we notice that from (33) w.p.1,

$$\liminf_{t\to\infty} \delta(t)\, \sqrt{t} = \infty ,$$

and by (35),

$$\delta(t)\, \sqrt{t} = \frac{b'(t) - b(t)}{\sqrt{t}} \leq_{st} \frac{b'(\mu(t)+1)}{\mu(t)+1} \cdot \frac{\mu(t)+1}{\sqrt{t}} + \frac{1}{\sqrt{t}} .$$

Therefore,

$$\liminf_{t\to\infty} \frac{b'(\mu(t)+1)}{\mu(t)+1} \cdot \frac{\mu(t)+1}{\sqrt{t}} = \infty. \tag{38}$$

From (32) we have that for any fixed $\varepsilon > 0$, $t_1 = t_1(\varepsilon) > 0$ exists such that

$$\frac{b'(t)}{t} \leq \mu + \varepsilon \quad \text{for any} \;\; t \geq t_1 .$$

Furthermore we also have that

$$\frac{b'(t)}{t} \leq b'(t_1) \quad \text{for any} \;\; 1 \leq t < t_1 .$$

Therefore,

$$\frac{b'(t)}{t} \leq \mu + \varepsilon + b'(t_1) \quad \text{for any } t \geq 1 .$$

In particular, $\mu(t)+1 \geq 1$, so we have that

$$\frac{b'(\mu(t)+1)}{\mu(t)+1} \leq \mu + \varepsilon + b'(t_1) \quad \text{for any} \quad t \geq 0\,. \tag{39}$$

We can define a random variable finite w.p.1:

$$\xi \stackrel{\text{def}}{=} \mu + \varepsilon + b'(t_1)\,.$$

It now follows from (38) that

$$\liminf_{t\to\infty} \frac{\xi}{\sqrt{t}}\,\mu(t) = \infty\,, \tag{40}$$

while $\xi/\sqrt{t} \to 0$. Hence, $\lim_{t\to\infty}\mu(t) = \infty$ w.p.1, and (20) is proved. □

3. Stability analysis of a multi-server GI/G/m queue

In this Section, we consider a standard multi-server $GI/G/m$ queue with $m > 1$ identical servers fed by a single queue following a FIFS non idling service discipline. In this setting, the *negative drift assumption* (see (12) for the single-server case) takes the form:

$$\rho\left(=\frac{\lambda}{\mu}\right) < m\,, \tag{41}$$

although the *regeneration condition* is exactly the same that for the $GI/G/1$ queue (see (13)):

$$\mathsf{P}(\tau > S) > 0\,. \tag{42}$$

Regeneration condition is an extra assumption now because it cannot be deduced from the negative drift assumption (41) for $m > 1$ (that is an important difference by comparing with the single-server case $m = 1$, for which (13) was implied by (12)).

Remark 2. We prove now a result similar to Theorem 1 but we can only obtain the finiteness of the expectations (14) and do not finiteness of the first regeneration period (15), because unlike the single-server case, the convergence $\mu(t) \to \infty$ w.p.1 generally does not imply its finiteness. Nevertheless, we will establish this finiteness later, in Section 5.3 (see Theorem 6 there), by using another idea.

Theorem 2. *Under assumptions (41) and (42), we have (14), that is,*

$$\mathsf{E}_0\,\beta_1 < \infty \quad \textit{and} \quad \mathsf{E}_0\,T_1 < \infty\,,$$

regardless of the initial states.

Proof. For any server k, we denote by $\mu^{(k)}(t)$ the corresponding idle time in the interval $[0,\, t]$, and let $\mu(t)$ be the total amount of time in $[0,\, t]$ that the system is not completely full (along this time, a new customer that arrives to the system starts service immediately, without having to wait for it), that is,

$$\mu(t) = \int_0^t \mathbb{I}_{(\nu(s)<m)}\, ds\,.$$

Obviously,

$$\mu^{(k)}(t) \le \mu(t) \quad \text{for any } t \ge 0 \quad \text{and any server } k = 1, \ldots, m$$

(note that for the single-server queue, both time processes, $\mu^{(1)}(t)$ and $\mu(t)$, coincide).

Recall that $V(t) = \sum_{n=1}^{N(t)} S_{n-1}$ is the total workload arrived to the system in $[0,\, t]$ (see (16)). We have an inequality similar to (17):

$$W(0^-) + V(t) = tm - \sum_{k=1}^{m} \mu^{(k)}(t) + W(t) \ge mt - m\mu(t)\,.$$

Then, $\mu(t) \ge t - \frac{V(t)}{m} + o(t)$ as $t \to \infty$, and by (41),

$$\liminf_{t\to\infty} \frac{\mu(t)}{t} \ge 1 - \lim_{t\to\infty} \frac{V(t)}{N(t)}\, \frac{N(t)}{t}\, \frac{1}{m} = 1 - \frac{\rho}{m} > 0 \quad \text{by (41)} \tag{43}$$

(compare with (19)). This implies (20), and (21), which in turn can be rewritten as

$$\liminf_{t\to\infty} \frac{1}{t} \int_0^t \mathsf{P}(\nu(s) < m)\, ds > 0$$

instead of (22). Then, $\mathsf{P}(\nu(t) < m) \not\to 0$ as $t \to \infty$, that is, a non-random sequence $z_i \to \infty$ and a constant $\delta > 0$ exist such that

$$\inf_{i\ge 1} \mathsf{P}(\nu(z_i) < m) \ge \delta\,.$$

For each server k, we introduce the (right-continuous) *residual service time* process

$$\{S^{(k)}(t),\, t \ge 0\}\,,$$

where $S^{(k)}(t)$ is the amount of time needed by server k to finish the service of the customer that is handling at instant t, if any, with $S^{(k)}(t) = 0$ if the server is idle at time t. It has been proved in [26] that the residual service time process is tight. Note that the process $\{\sum_{k=1}^{m} S^{(k)}(t),\, t \ge 0\}$ is also tight

and moreover if $\nu(t) < m$ then $W(t) = \sum_{k=1}^{m} S^{(k)}(t)$ (because if there is at least one server free, it is not allowed to have customers waiting for service). Recall that process $\{\tau(t),\, t \geq 0\}$, which is the forward renewal time process for input introduced in (1), is also tight. Hence, by the tightness, a finite constant $D > 0$ exists such that

$$\inf_{i \geq 1} \mathsf{P}(\nu(z_i) < m,\, W(z_i) \leq D,\, \tau(z_i) \leq D) \geq \frac{\delta}{2}$$

(compare with (24) in the single-server case). It follows from the *regeneration condition* (42) and the fact that $\mathsf{E}\,\tau < \infty$, that finite and positive constants δ_0, δ_1 and R exist such that

$$\mathsf{P}(R \geq \tau > S + \delta_0) \geq \delta_1\,.$$

Denote $L = \lceil \frac{D}{\delta_0} \rceil$ and let $n(i) = \min(k : t_k \geq z_i)$ be the arrival number of the 1st arrival after z_i (that is, $t_{n(i)}$ is the first arrival instant after z_i). Introduce the event

$$A_i(L) \stackrel{\text{def}}{=} \bigcap_{j=0}^{L-1} \{R \geq \tau_{n(i)+j} > S_{n(i)+j} + \delta_0\},$$

and note that events $\{R \geq \tau_{n(i)+j} > S_{n(i)+j} + \delta_0\}$, $j = 0, \ldots, L-1$, are independent. Denote the event

$$E_i \stackrel{\text{def}}{=} \{\nu(z_i) < m,\, W(z_i) \leq D,\, \tau(z_i) \leq D\}\,, \tag{44}$$

which is independent of $A_i(L)$ for any i. By definition, on the event $E_i \cap A_i(L)$, a customer arrives which sees an empty system and thus a regeneration occurs. Moreover, it happens in interval $[z_i,\, z_i + D + L\,R]$ with constant length $D + L\,R$ and with a probability bigger or equal to

$$\mathsf{P}(E_i \cap A_i(L)) \geq \mathsf{P}(E_i)\,\mathsf{P}(A_i(L)) \geq \frac{\delta}{2}\,\delta_1^L,$$

because $\mathsf{P}(A_i(L)) \geq \delta_1^L$. Hence, the forward regeneration time $T(z_i)$ at instant z_i satisfies

$$\mathsf{P}(T(z_i) \leq b) \geq \mathsf{P}(E_i \cap A_i(L)) \geq \varepsilon \quad \text{for any} \quad i \geq 1\,,$$

with $b = D + L\,R$ and $\varepsilon = \frac{\delta}{2}\,\delta_1^L$, that is, we have (11).

Since instant z_i is arbitrary, then we obtain $\mathsf{P} - \lim_{t \to \infty} T(t) \neq \infty$, and by (7), $\mathsf{E}_0\, T_1 < \infty$, and also $\mathsf{E}_0\, \beta_1 < \infty$ by (8). □

Now we give an alternative proof of the fact that $\mu(t) \to \infty$ w.p.1. in the proof of Theorem 2, by following the same notations (adapted to the multi-server setting) of Section 2.2.

As the basic process, we consider the superposition $b'(t) = \sum_{k=1}^{m} b'_k(t)$ of m independent (stochastically equivalent) zero-delayed renewal processes

$$b'_k(t) = \min\{\, n \geq 1 : S_0^{(k)} + \cdots + S_n^{(k)} > t \,\} \quad \text{if} \;\; t > 0\,, \quad \text{and} \;\; b'_k(0) = 1\,,$$

generated by the (i.i.d.) service times $\{S_n^{(k)},\, n \geq 0\}$ of customers handled by server k, for any $k = 1, \ldots, m$. Note that $b'(0) = \sum_{k=1}^{m} b'_k(0) = m$. As in Section 2.2, number $b(t)$ of customers served in $(0,\, t]$ satisfies inequality $b(t) \leq N(t) + \nu(0^-)$, and moreover $b(t) \leq b'(t)$. So

$$b'(t) - b(t) \geq b'(t) - N(t) + o(t) \quad \text{as} \;\; t \to \infty\,.$$

Since w.p.1, by the SLLN for renewal processes, we have

$$\lim_{t\to\infty} \frac{b'(t)}{t} = m\,\mu \tag{45}$$

(compare with (32)), then

$$\liminf_{t\to\infty} \frac{b'(t) - b(t)}{t} \geq m\mu - \lambda = \mu(m - \rho) > 0 \tag{46}$$

by the *negative drift assumption* (41). Again (by using coupling) we obtain analogously to (35):

$$b'(t) - b(t) \leq b'(t) - \sum_{k=1}^{m} b'_k(t - \mu^{(k)}(t)) + m \leq b'(t) - \sum_{k=1}^{m} b'_k(t - \mu(t)) + m =$$

$$= b'(t) - b'(t - \mu(t)) + m \leq_{st} b'(\mu(t)) + m \leq b'(\mu(t) + 1) + m\,.$$

Moreover, it follows from (45) that for any $\varepsilon > 0$ an instant $t_1 = t_1(\varepsilon)$ exists such that

$$\frac{b'(\mu(t) + 1)}{\mu(t) + 1} \leq m\mu + \varepsilon + b'(t_1) \quad \text{for any} \;\; t \geq 0\,,$$

as in (39). Taking into account (46) we have that

$$\infty = \liminf_{t\to\infty} \frac{b'(t) - b(t)}{\sqrt{t}} \leq \liminf_{t\to\infty} \xi\, \frac{\mu(t) + 1}{\sqrt{t}}\,, \tag{47}$$

being $\xi = \mu\, m + \varepsilon + b'(t_1)$ a random variable finite w.p.1. This implies (40), and therefore $\mu(t) \to \infty$ w.p.1. □

4. Extensions of the GI/G/1 queue

In this Section, we study stability of some extensions of the standard single-server $GI/G/1$ queue considered in Section 2, in the *zero-delayed case*. For that, we use the discret-time approach of Section 2.1 (adapted to our setting) to prove the finiteness of expectations $\mathsf{E}_0\,\beta_1$ and $\mathsf{E}_0\,T_1$ regardless of the initial states. As it was commented in the Remark 1 there, this approach does not allow to prove finiteness of the first regeneration periods.

4.1. **The GI/G/1 queue with (partially) impatient customers.** In our first extension, we allow impatient customers in queue (that is, customers can leave the system without have been served). More specifically, let γ_n be the time that customer n may wait in the queue before to desist and leave the system. Recall that W_n denotes the waiting time of customer n in the queue line. Then customer n leaves the system with no service if $W_n > \gamma_n$. We assume $(\gamma_n)_{n\geq 0}$ to be i.i.d. and let γ be a generic variable with the distribution of any γ_n, that is assumed to be > 0 w.p.1. and independent of S.

We also introduce the *"persistent rate"*

$$p \stackrel{\text{def}}{=} \mathsf{P}(\gamma = \infty)\,, \quad \text{which is assumed to be in } [0,\,1)$$

because $p = 1$ would mean that any customer is patient, and then in fact we would be considering again the standard $GI/G/1$ queue, already studied in Section 2. On the opposite situation, $p = 0$, which is actually allowed, means that all customers are impatient, and $0 < p < 1$ corresponds to the partially impatient customers scenario.

Our main (*negative drift*) assumption takes now the following form,

$$\rho < \frac{1}{p}. \tag{48}$$

(Note that if $p = 0$, this negative drift condition always holds, if we put $\frac{1}{p} = \infty$, and that for the limit value $p = 1$ we obtain (12)). Note also that assumption (48) can be rewritten as

$$\mathsf{E}\,\tau - p\,\mathsf{E}\,S = \mathsf{E}(\tau - S\,\mathbb{I}_{(\gamma=\infty)}) > 0\,, \tag{49}$$

which in turn implies that $\delta_0 > 0$ and $\delta_1 > 0$ exist such that

$$\mathsf{P}(\tau > S\,\mathbb{I}_{(\gamma=\infty)} + \delta_0) \geq \delta_1\,. \tag{50}$$

Define

$$\varepsilon_0 \stackrel{\text{def}}{=} \mathsf{E}\,\tau - p\,\mathsf{E}\,S\,(> 0 \text{ by } (49))\,.$$

Assume also that the following *regeneration* assumption holds: for any finite constant $T \geq 0$, constants $\varepsilon(T) > 0$ and $\delta(T) > 0$ exist such that

$$\inf_{y \leq T} \mathsf{P}(\tau > S\,\mathbb{I}_{(\gamma > y)} + \delta(T)) \geq \varepsilon(T)\,. \tag{51}$$

Note that (51) is not implied by (50). Indeed, it is possible that $\delta(T) \to 0$ (and also $\varepsilon(T) \to 0$) as $T \to \infty$ while $\delta_0 > 0$ and $\delta_1 > 0$ are fixed. So in general we can not eliminate (51) from the analysis and we need to assume it.

As in the classical case, the workload process $W = \{W_n, n \geq 0\}$ is Markovian and satisfies a modified Lindley's recursion (see (5)):

$$W_{n+1} = \big(W_n + S_n\,\mathbb{I}_{(\gamma_n \geq W_n)} - \tau_n\big)^+, \quad n \geq 0\,. \tag{52}$$

If we define for any $y \in [0,\, \infty)$

$$X(y) \stackrel{\text{def}}{=} S\,\mathbb{I}_{(\gamma \geq y)} - \tau\,,$$

the modified Lindley's recursion implies that

$$W_{n+1} = \big(W_n + X(W_n)\big)^+, \quad n \geq 0\,,$$

and we have that with

$$X(\infty) \stackrel{\text{def}}{=} S\,\mathbb{I}_{(\gamma = \infty)} - \tau\,, \quad \mathsf{E}\,X(\infty) = -\varepsilon_0 < 0 \quad \text{by (49)}\,.$$

Theorem 3. *Under assumptions (48) and (51),*

$$\mathsf{E}_0\,\beta_1 < \infty \quad \textit{and} \quad \mathsf{E}_0\,T_1 < \infty\,.$$

Proof. The proof of this result is similar to that of (25) in Section 2.1 for the standard $GI/G/1$ queue, and then we do not write all details, but point out the differences: with $\Delta_n = W_{n+1} - W_n$ we have that for any y,

$$\begin{aligned} \mathsf{E}(\Delta_n \,|\, W_n = y) &= \mathsf{E}\,\big(\,(y + X(y))^+ - y\,\big) = \\ &= -y\,\mathsf{P}(X(y) \leq -y) + \int_{z > -y} z\,\mathsf{P}(X(y) \in dz)\,, \end{aligned} \tag{53}$$

analogously to (26) by substituting U by $X(y)$. Because $\mathsf{E}\,\tau < \infty$, then $y\,\mathsf{P}(X(y) \leq -y) \leq y\,\mathsf{P}(\tau \geq y) \to 0$ and hence, $-y\,\mathsf{P}(X(y) \leq -y) \to 0$, as $y \to \infty$. And moreover

$$\int_{z > -y} z\,\mathsf{P}(X(y) \in dz) \downarrow \mathsf{E}\,X(\infty) = -\varepsilon_0 < 0 \quad \text{as } y \to \infty\,,$$

because the family $\{X(y),\, y \geq 0\}$ is uniformly integrable and $\mathsf{E}\,X(y) \to \mathsf{E}\,X(\infty) < \infty$. Thus, by (53) we obtain that

$$\lim_{y \to \infty} \mathsf{E}(\Delta_n \,|\, W_n = y) = -\varepsilon_0 < 0 \tag{54}$$

(compare with (27)). We obtain the analogous formula to (28):

$$\mathsf{E}\Delta_n \le \mathsf{E}\, S\ \mathsf{P}(W_n \le y_0) - \frac{\varepsilon_0}{2}\,\mathsf{P}(W_n > y_0) \tag{55}$$

for any $y_0 > 0$ big enough. From that, the proof that $\mathsf{P} - \lim\limits_{n\to\infty} W_n \neq \infty$ follows as in Section 2.1 (by taking $\varepsilon < \dfrac{\varepsilon_0/2}{\varepsilon_0/2 + 1/\mu}$), and then we have that constants $\delta > 0$ and $T < \infty$, and a (non-random) sequence $(n_i)_{i\ge 1}$ with $n_i \to \infty$ exist, such that (29) holds, that is,

$$\inf_{i\ge 1} \mathsf{P}(W_{n_i} \le T) \ge \delta\,.$$

We can denote $L \stackrel{\text{def}}{=} \lceil \frac{T}{\delta(T)} \rceil$, where $\delta(T)$ is given by the *regeneration assumption* (51), and we can prove that a regeneration is attained within a finite interval with a positive probability. More precisely, we show that process W reaches zero state within interval $[n_i,\, n_i + L]$ from any point $W_{n_i} = x \in [0,\, T]$ with a probability which is uniformly lower bounded over the set by a positive constant. To do this, we first consider the sequences of events

$$A_i(L) = \bigcap_{k=0}^{L-1} \{\tau_{n_i+k} > S_{n_i+k}\, \mathbb{I}_{(\gamma_{n_i+k} \ge W_{n_i+k})} + \delta(T)\} \quad \text{and}$$

$$B_i = \{W_{n_i} \le T\}\,, \quad \text{for} \ \ i \ge 1\,. \tag{56}$$

By the modified Lindley's recursion (52), we have that on $A_i(L)$,

$$\begin{aligned}
&0 \le W_{n_i+1} = \Big(W_{n_i} + S_{n_i}\, \mathbb{I}_{(\gamma_{n_i} \ge W_{n_i})} - \tau_{n_i}\Big)^+ \le \Big(W_{n_i} - \delta(T)\Big)^+ \\
&0 \le W_{n_i+2} = \Big(W_{n_i+1} + S_{n_i+1}\, \mathbb{I}_{(\gamma_{n_i+1} \ge W_{n_i+1})} - \tau_{n_i+1}\Big)^+ \le \\
&\qquad\qquad \le \Big(W_{n_i+1} - \delta(T)\Big)^+ \le \Big(W_{n_i} - 2\,\delta(T)\Big)^+ \\
&\qquad\vdots \\
&0 \le W_{n_i+L} \le \ldots \le \Big(W_{n_i} - L\,\delta(T)\Big)^+ \le \Big(W_{n_i} - T\Big)^+ \ \big(= 0 \text{ on } \ B(i)\big)\,,
\end{aligned} \tag{57}$$

that is $W_{n_i+L} = 0$ on the events $A(i) \cap B(i)$. Therefore, when customer $n_i + L$ arrives he meets an empty system and regeneration occurs in interval $[n_i,\, n_i + L]$.

By the independence of the events whose intersection is $A_i(L)$, and by (51) we have that

$$\mathsf{P}\big(A_i(L)\,|\,B_i\big) = \prod_{k=0}^{L-1} \mathsf{P}(\tau_{n_i+k} > S_{n_i+k}\,\mathbb{I}_{(\gamma_{n_i+k}\geq W_{n_i+k})} + \delta(T)\,|\,W_{n_i} \leq T\big) \geq$$

$$\geq (\varepsilon(T))^L,$$

because by (57),

$$W_{n_i+k} \leq \Big(W_{n_i} - k\,\delta(T)\Big)^+ \leq W_{n_i} \quad \text{for any} \;\; k = 0, \ldots, L-1\,,$$

and then, if $W_{n_i} \leq T,$ we have that $W_{n_i+k} \leq T\,.$ As a consequence,

$$\mathsf{P}(A_i(L) \cap B_i) = \mathsf{P}(A_i(L)\,|\,B_i)\,\mathsf{P}(B_i) \geq \delta\,(\varepsilon(T))^L \quad \text{for any} \;\; i \geq 1\,,$$

and then,

$$\mathsf{P}(\beta(n_i) \leq L) \geq \mathsf{P}(A(i) \cap B(i)) \geq \delta\,(\varepsilon(T))^L > 0\,.$$

With $\varepsilon = \delta\,(\varepsilon(T))^L > 0$, we obtain (10) and from it, that $\mathsf{E}_0\,\beta_1 < \infty$, and also that $\mathsf{E}_0\,T_1 < \infty$ by Wald's identity (8). □

4.2. **A state-dependent G/G/1 queue.** Now we consider an extension of the single-server $GI/G/1$ queue with impatient customers considered in Section 4.1. We assume that on any event $\{W_n = y\}$, service time S_n and interarrival time τ_n are distributed as random variables $S(y)$ and $\tau(y)$, respectively, depending on y, with given (conditional) distributions. That is the *state-dependent* single-server $GI/G/1$ queue that we treat in this section.

By comparing with the impatient customers model of the previous section, there $\tau(y) = \tau$ did not depend in fact on y, and $S(y) = S\,\mathbb{I}_{(\gamma\geq y)}$, which depend on y through the random variable γ. Denote now for any $y \in [0,\,\infty)$,

$$X(y) \stackrel{\text{def}}{=} S(y) - \tau(y)$$

and assume that the following (non-degeneration) conditions hold:

$$\sup_{y\geq 0} \mathsf{E}\,S(y) < \infty\,, \tag{58}$$

$$\sup_{y\geq 0} \mathsf{E}\,\tau(y) < \infty\,. \tag{59}$$

We also assume that the *negative drift* condition holds. This condition now takes the form:

$$\limsup_{y\to\infty} \mathsf{E}\,X(y) < 0\,, \tag{60}$$

and also assume the *regeneration* condition: for any finite constant $T \geq 0$, constants $\varepsilon(T) > 0$ and $\delta(T) > 0$ exist such that

$$\inf_{y \leq T} \mathsf{P}(\tau(y) > S(y) + \delta(T)) \geq \varepsilon(T)\,. \tag{61}$$

Define $\varepsilon_0 \stackrel{\text{def}}{=} -\limsup_{y\to\infty} \mathsf{E}\, X(y)\,(> 0)\,.$

In this setting, analogously to Theorem 3 in the previous Section, we obtain the following statement.

Theorem 4. *Under assumptions (58)–(61),*

$$\mathsf{E}_0\, \beta_1 < \infty \quad \textit{and} \quad \mathsf{E}_0\, T_1 < \infty\,.$$

Proof. First of all we note that (58) and (59) imply

$$-\infty < \inf_{y\geq 0} \mathsf{E}\, X(y) \leq \sup_{y\geq 0} \mathsf{E}\, X(y) < \infty.$$

We obtain (54) as in the proof of Theorem 3, and instead of (55), we have:

$$\mathsf{E}\, \Delta_n \leq \sup_{y\geq 0} \mathsf{E}\, S(y)\ \mathsf{P}(W_n \leq y_0) - \frac{\varepsilon_0}{2}\, \mathsf{P}(W_n > y_0)$$

for any $y_0 > 0$ big enough. As in the proof of Theorem 3, we have that constants $\delta > 0$ and $T < \infty$, and a (non-random) sequence $(n_i)_{i\geq 1}$ with $n_i \to \infty$ exist, such that

$$\inf_{i\geq 1} \mathsf{P}(W_{n_i} \leq T) \geq \delta\,,$$

and with $L \stackrel{\text{def}}{=} \lceil \frac{T}{\delta(T)} \rceil$, where $\delta(T)$ is given by the regeneration assumption, it can be proved that a regeneration is attained within a finite interval with a positive probability. In fact, it can be shown similarly to Theorem 3 that process W reaches zero state in $[n_i,\, n_i + L]$ from any point $W_{n_i} = y \in [0,\, T]$ with a probability which is uniformly lower bounded over the set by a positive constant, by considering now (instead of (56)) the sequence of events

$$A_i(L) = \bigcap_{k=0}^{L-1} \{\tau_{n_i+k}(W_{n_i+k}) > S_{n_i+k}(W_{n_i+k}) + \delta(T)\} \quad \text{and}$$

$$B_i = \{W_{n_i} \leq T\}\,, \quad \text{as before, for any} \ \ i \geq 1\,,$$

and by using the modified Lindley's recursion

$$W_{n+1} = \big(W_n + S_n(W_n) - \tau_n(W_n)\big)^+\,, \quad n \geq 0\,.$$

From that we have (10) and then $\mathsf{E}_0\, \beta_1 < \infty$.

We cannot use Wald's identity (8) to obtain the finiteness of $\mathsf{E}_0\, T_1$ from that of $\mathsf{E}_0\, \beta_1$. Nevertheless, denote $\sup_y \mathsf{E}\, \tau(y) = d\,(< \infty$ by assumption

(59)), and note that β_1 is a stopping time (with respect to sequence $(\tau_n)_{n\geq 0}$) and that $T_1 = \tau_0 + \cdots + \tau_{\beta_1 - 1}$. Then we have

$$\mathsf{E}_0\, T_1 = \sum_{k=0}^{\infty} \mathsf{E}(\tau_k;\, \beta_1 > k) = \sum_{k=0}^{\infty} \mathsf{E}\tau_k \mathsf{P}(\beta_1 > k) \leq d\, \mathsf{E}_0\, \beta_1 < \infty\,. \quad \square \tag{62}$$

5. Extensions of the multi-server GI/G/m queue

5.1. **The multi-server GI/G/m queue with non-identical servers.** Consider a multi-server $GI/G/m$ queue with $m > 1$ servers as in Section 3, but now with non-identical servers. For any $i = 1, \ldots, m$, service times $\{S_n^{(i)},\, n \geq 0\}$ are assumed to be i.i.d, being $S^{(i)}$ a generic random variable whose distribution is that of any $S_n^{(i)}$, $n \geq 0$, assumed to be > 0 w.p.1, and with finite expectation

$$\mathsf{E}\, S_n^{(i)} = \frac{1}{\mu^{(i)}} \in (0,\, \infty)\,.$$

Random variables $S^{(i)}$ and $S^{(j)}$ are allowed to have different distributions for $i \neq j$, with $i,\, j \in \{1, \ldots, m\}$. Denote $\tilde{\mu} \stackrel{\text{def}}{=} \sum_{i=1}^{m} \mu^{(i)}$. Using the same other assumptions and notations as in Section 3, we assume *negative drift condition*, that takes the form:

$$\frac{\lambda}{\tilde{\mu}} < 1\,. \tag{63}$$

Note that in the identical-servers scenario of Section 3, $\mu^{(i)} = \mu^{(j)} = \mu$ for any $i,\, j$, so $\tilde{\mu} = m\,\mu$, and therefore (63) becomes (41). Moreover, we assume the *regeneration condition*:

$$\mathsf{P}(\tau > S^{(i)}) > 0 \quad \text{for all} \;\; i = 1, \ldots, m\,. \tag{64}$$

As in the standard (identical-servers) $GI/G/m$ queue (with $m > 1$) considered in Section 3, regeneration condition is an extra assumption that cannot be deduced from (63).

Analogously to Theorem 2 for the identical-servers setting, we can prove the next result:

Theorem 5. *Under assumptions (63) and (64), we have (14), that is,*

$$\mathsf{E}_0\, \beta_1 < \infty \quad \textit{and} \quad \mathsf{E}_0\, T_1 < \infty\,,$$

regardless of the initial states.

Proof. Denote by Q the original $GI/G/m$ queue, and consider also a modified queue Q^* with the same initial state, the same input, but in which an arriving customer goes to server i with probability $p^{(i)} = \mu^{(i)}/\tilde{\mu}$, for

any $i = 1, \ldots, m$. In the sequel, we will put index $*$ on the quantities corresponding to the system Q^*, and index (i) on that corresponding to any server i. Define

$$\lambda^{*(i)} \stackrel{\text{def}}{=} \lambda\, p^{(i)}, \;\; \rho^{*(i)} \stackrel{\text{def}}{=} \frac{\lambda^{*(i)}}{\mu^{(i)}}.$$

Therefore, (63) implies that

$$\rho^{*(i)} < 1 \quad \text{for all} \;\; i = 1, \ldots, m\,. \tag{65}$$

As a consequence, by Theorem 1, the renewal processes of regenerations in every single-server (which are standard $GI/G/1$ queues) of system Q^* are positive recurrent. By (14) and (15) we have that

$$\mathsf{E}_0\, T_1^{*(i)} < \infty \quad \text{and} \;\; T_1^{*(i)} < \infty \;\; w.p.1 \quad \text{for any } i = 1, \ldots, m\,. \tag{66}$$

Let $V^{(i)}(t)$ be the total workload arrived to server i (in queue Q) in the interval $[0, t]$, $W^{(i)}(t)$ be the residual workload for server i at instant t, and $\mu^{(i)}(t)$ be the idle time for server i in interval $[0, t]$. For queue Q^* the corresponding processes are denoted by $V^{*(i)}(t)$, $W^{*(i)}(t)$ and $\mu^{*(i)}(t)$, respectively.

First of all, note that for each $t \geq 0$, a server $i(t)$ exists such that $V^{(i(t))}(t) \leq V^{*(i(t))}(t)$ (because if $V^{(i)}(t) > V^{*(i)}(t)$ for any $i = 1, \ldots, m$, we will have that $N^*(t) > N(t)$, that is a contradiction with the fact that the total arrivals to both systems, Q and Q^*, are equal). With the notations used in the proof of Theorem 2 (Section 3), we have that

$$\mu^{(i)}(t) \leq \mu(t) = \int_0^t \mathbb{I}_{(\nu(s)<m)}\, ds \quad \text{for any} \;\; i = 1, \ldots, m \quad \text{and any} \;\; t \geq 0\,.$$

Since $W^{(i)}(0^-) = W^{*(i)}(0^-)$ for all $i = 1, \ldots, m$, and we have that

$$W^{(i)}(0^-) + V^{(i)}(t) = t - \mu^{(i)}(t) + W^{(i)}(t)$$

and

$$W^{*(i)}(0^-) + V^{*(i)}(t) = t - \mu^{*(i)}(t) + W^{*(i)}(t)\,,$$

as a consequence we obtain that for any $t \geq 0$, for server $i(t)$,

$$\mu^{*(i(t))}(t) - W^{*(i(t))}(t) \leq \mu^{(i(t))}(t) - W^{(i(t))}(t) \leq \mu^{(i(t))}(t)$$

and thus,

$$\mu(t) \geq \mu^{(i(t))}(t) \geq \min_{1 \leq i \leq m} \mu^{*(i)}(t) - \sum_{i=1}^{m} W^{*(i)}(t) \quad \text{for any} \;\; t \geq 0\,. \tag{67}$$

It now follows that in the collection of single-server $GI/G/1$ queues Q^*, all waiting time processes are tight, and moreover, positive recurrent, and in particular, $W^{*(i)}(t) = o(t)$, $t \to \infty$. The latter result one can also be

obtained from the following observation: let $T^{*(i)}(t)$ be the forward regeneration time for server i (in system Q^*) at instant t; then it follows from (66) that $T^{*(i)}(t) = o(t)$ as $t \to \infty$ w.p.1 (see [32]). Because $W^{*(i)}(t) \le T^{*(i)}(t)$, the desired result follows. Now as in (18) we obtain that w.p.1, as $t \to \infty$,

$$\liminf_{t\to\infty} \frac{\mu^{*(i)}(t)}{t} \ge 1 - \rho^{*(i)} > 0 \quad \text{for any} \quad i = 1, \ldots, m\,, \quad \text{by (65)}\,. \tag{68}$$

Hence, (67) implies

$$\liminf_{t\to\infty} \frac{\mu(t)}{t} \ge \min_{1\le i\le m} \liminf_{t\to\infty} \frac{\mu^{*(i)}(t)}{t} > 0\,.$$

It now easily follows, as in the proof of Theorem 2 from (43), that

$$\mathsf{E}_0\,\beta_1 < \infty \quad \text{and} \quad \mathsf{E}_0\,T_1 < \infty\,. \qquad \square$$

Remark 3. Note that we indeed need to impose *regeneration condition* (64) to finish the proof from (68), as in the proof of Theorem 2, unlike modified system Q^*, where such a kind of condition holds automatically for every server i from the negative drift assumption (65).

5.2. **A multi-server R/G/m queue with regenerative input.** In this section we consider a multi-server $R/G/m$ queue with $m \ge 1$ identical servers and a classical *regenerative (zero-delayed) input* with arrival instants $(t_n)_{n\ge 0}$ $(t_0 = 0)$, interarrival times $(\tau_n = t_{n+1} - t_n)_{n\ge 0}$, and regeneration points $(\alpha_n)_{n\ge 1}$, $1 \le \alpha_1 < \alpha_2 < \cdots$, which form an *imbedded renewal process*. That is, groups

$$(\tau_0, \ldots, \tau_{\alpha_1-1}), (\tau_{\alpha_1}, \ldots, \tau_{\alpha_2-1}), \ldots, (\tau_{\alpha_n}, \ldots, \tau_{\alpha_{n+1}-1}), \ldots$$

are i.i.d., and regeneration periods $\alpha_{n+1} - \alpha_n$, for $n \ge 1$, and α_1 (since we are in the zero-delayed case) are i.i.d. and for any $n \ge 1$, $\mathsf{E}_0\,(\alpha_{n+1} - \alpha_n) = \mathsf{E}_0\,\alpha_1 < \infty$.

Note that this model is a generalization of the multi-server $GI/G/m$ queue considered in Section 3, which corresponds to the particular case $\alpha_1 \equiv 1$ (in which, therefore, the interarrival times τ_n are moreover independent random variables).

The *renewal* process of regeneration instants for the input process is $(\Gamma_n)_{n\ge 0}$, with

$$\Gamma_n \stackrel{\text{def}}{=} t_{\alpha_n} \quad \text{for any} \quad n \ge 1\,, \quad \text{and} \quad \Gamma_0 = 0\,.$$

Assume that $\alpha = (\alpha_n)_{n\ge 1}$ and $(\Gamma_n)_{n\ge 0}$ are *positive recurrent*, that is,

$$\alpha_1 < \infty \ \text{ w.p.1}\,, \quad \mathsf{E}_0\,\alpha_1 < \infty; \tag{69}$$

$$\Gamma_1 < \infty \ \text{ w.p.1}\,, \quad \mathsf{E}_0\,\Gamma_1 < \infty. \tag{70}$$

The notations E_0 (as before) and P_0 (below) relate to the zero delayed case ($t_0 = 0$) in which the first input regeneration period α_1 is distributed as others.

Assume the next *regeneration condition*:

$$\mathsf{P}_0(\tau_0 > S,\, \alpha_1 = 1) > 0, \tag{71}$$

where S is a generic random variable with the distribution of any service time, assumed to be > 0 w.p.1, and with expectation $1/\mu \in (0,\, \infty)$.

The renewal processes of regenerations are now defined as:

$$\beta_{n+1} \stackrel{\text{def}}{=} \min\{\, \alpha_k > \beta_n \,:\, \nu_{\alpha_k} = 0 \,\} \quad \text{for any } n \geq 0\,, \quad \beta_0 \stackrel{\text{def}}{=} 0$$

(instead of (2)), and

$$T_{n+1} \stackrel{\text{def}}{=} t_{\beta_{n+1}} \Big(= \min\{\, \Gamma_k (= t_{\alpha_k}) > T_n : \nu_{\alpha_k} = 0\}\Big) \text{ for any } n \geq 0,$$

$$T_0 \stackrel{\text{def}}{=} t_0\,.$$

Note that these definitions are obtained by substituting $\{k \geq 1\}$ by $\{\alpha_k,\, k \geq 1\}$ in (2) and (3), respectively.

The input rate and traffic intensity are defined as

$$\lambda = \frac{\mathsf{E}_0\, \alpha_1}{\mathsf{E}_0\, \Gamma_1} \quad \text{and} \quad \rho = \frac{\lambda}{\mu}\,, \quad \text{respectively}\,.$$

The *negative drift condition* is, as for the standard $GI/G/m$ queue,

$$\rho < m\,. \tag{72}$$

Now we can establish the following result:

Theorem 6. *Suppose that assumptions (69), (70), (71) and (72) hold. Then,*

$$\mathsf{E}_0\, \beta_1 < \infty \quad \textit{and} \quad \mathsf{E}_0\, T_1 < \infty\,,$$

regardless of the initial states.

Proof. The proof is similar to that of Theorem 2 in Section 3. We only outline the differences: instead of process $\{\tau(t),\, t \geq 0\}$ defined in (1), we use here the *forward regeneration time process* for the input process, $\{\gamma(t),\, t \geq 0\}$, defined by

$$\gamma(t) \stackrel{\text{def}}{=} \min\{\, \Gamma_k - t \,:\, \Gamma_k - t \geq 0 \,\}\,,$$

and let for any $t \geq 0$,

$$k(t) \stackrel{\text{def}}{=} \min\{\, k \,:\, \Gamma_k \geq t \,\} \quad (\text{that is, } \gamma(t) = \Gamma_{k(t)} - t \;\; (\geq 0)\,)\,.$$

By the positive recurrence of the process of regenerations of input process (assumption (70)), the process $\{\gamma(t),\, t \geq 0\}$ is tight. Define the increments $V_n \stackrel{\text{def}}{=} V(\Gamma_{n+1}) - V(\Gamma_n)$, $n \geq 0$, of the cumulative process of arrived workload to the system $\{V(t), t \geq 0\}$ (see (16) for definition), and note that

$$V_0 = V(\Gamma_1) = V(t_{\alpha_1}) = \sum_{n=1}^{\alpha_1+1} S_{n-1}\,.$$

Therefore,

$$\mathsf{E}_0\, V_1 = \frac{\mathsf{E}_0\, \alpha_1 + 1}{\mu} < \infty\,.$$

It follows that the process $\{\Delta(t) \stackrel{\text{def}}{=} V(\Gamma_{k(t)}) - V(t),\, t \geq 0\}$ is also tight (see [32]), and we can introduce here similarly to (44), the events

$$E_i \stackrel{\text{def}}{=} \{\nu(z_i) < m,\, \Delta(z_i) \leq d,\, W(z_i) \leq D,\, \gamma(z_i) \leq D\}\,,$$

where D and d are some positive constants, and $(z_i)_{i\geq 1}$ is a non-random sequence such that $z_i \to \infty$, which verify that $\inf\limits_{i\geq 1} \mathsf{P}\big(E_i\big) \geq \dfrac{\delta}{2}$ for some $\delta > 0$.

By the *regeneration condition* (71), positive constants R, δ_1 and δ_0 exist such that

$$\mathsf{P}_0\Big(R \geq \tau_0 > S + \delta_0,\, \alpha_1 = 1\Big) \geq \delta_1\,. \tag{73}$$

We call a customer n *regenerative* if $n = \alpha_r$ for some $r \geq 1$. Let $r(i)$ be the next (after instant z_i) regenerative customer, in particular then

$$t_{r(i)} = z_i + \gamma(z_i)\,.$$

Thus, on the event E_i, regenerative customer $r(i)$ arrives to the system in the interval $[z_i,\, z_i + D]$, and the residual workload upon his arrival is $W(r(i)) \leq d + D$, and this occurs with probability $\geq \frac{\delta}{2}$, for any $i \geq 1$.

Define $L = \lceil (d+D)/\delta_0 \rceil$, and consider the event

$$A_i(L) \stackrel{\text{def}}{=} \bigcap_{j=0}^{L-1} \Big\{R \geq \tau_{r(i)+j} > S_{r(i)+j} + \delta_0,\, \alpha_{k(i)+j} = 1\Big\}. \tag{74}$$

Obviously, $\mathsf{P}_0\,(A_i(L)) \geq \delta_1^L$ and it is easy to see that on the event $E_i \cap A_i(L)$, a regenerative customer arrives in the interval $[z_i,\, z_i + D + L\,R]$ and finds an empty server, and this happens with probability $\geq \dfrac{\delta}{2}\,\delta_1^L$. Thus, for any $i \geq 1$,

$$\mathsf{P}_0\,(T(z_i) \leq b) \geq \varepsilon\,, \quad \text{with} \;\; b = D + L\,R \quad \text{and} \quad \varepsilon = \frac{\delta}{2}\,\delta_1^L\,.$$

Therefore, $\mathsf{E}_0\, T_1 < \infty$. Taking into account that $T_1 = t_{\beta_1}$ $(= t_{\alpha_\ell}$ for some $\ell)$, and that

$$t_{\alpha_\ell} = (\tau_0 + \cdots + \tau_{\alpha_1 - 1}) + (\tau_{\alpha_1} + \cdots + \tau_{\alpha_2 - 1}) + \cdots + (\tau_{\alpha_{\ell-1}} + \cdots + \tau_{\alpha_\ell - 1})\,,$$

by Wald's identity we obtain $\mathsf{E}T_1 = \mathsf{E}_0 \ell\, \mathsf{E}_0 \Gamma_1 < \infty$. Because

$$\beta_1 = \alpha_\ell = \alpha_1 + (\alpha_2 - \alpha_1) + \cdots + (\alpha_\ell - \alpha_{\ell-1}),$$

we have then $\mathsf{E}_0\, \beta_1 = \mathsf{E}_0\, \ell\, \mathsf{E}_0\, \alpha_1 < \infty$, and the proof is complete. □

5.3. **Stability analysis in the delayed case for the standard $GI/G/m$ queue and an application.** In this section, we extend stability analysis to the delayed case for a class of queueing models. More precisely, we show how to establish finiteness of the first regeneration periods β_1 and T_1 under the same negative drift and regeneration assumptions which have been used above to prove finiteness of the mean standard regeneration periods ($\mathsf{E}_0\, \beta_1$ and $\mathsf{E}_0\, T_1$).

As a basic model we consider the standard $GI/G/m$ queue with $m \geq 1$ servers, but this approach holds also for some other queues, as we will see in the application. First of all we recall that Theorem 2 (Section 3) says that under the *negative drift assumption* $\rho < m$ (41) and the *regeneration condition* $\mathsf{P}(\tau > S) > 0$ (42), $\mathsf{E}_0\, \beta_1 < \infty$ and $\mathsf{E}_0\, T_1 < \infty$. In its proof it is showed that w.p.1,

$$\liminf_{t\to\infty} \frac{\mu(t)}{t} \geq \delta_0\,, \tag{75}$$

for some $\delta_0 > 0$ ($\delta_0 = 1 - \frac{\rho}{m}$ in (43)) , with $\mu(t) = \displaystyle\int_0^t \mathbb{I}_{(\nu(s)<m)}\, ds$ the total amount of time in $[0,\, t]$ that there is at least one free server in the system. This implies that (see (20))

$$\lim_{t\to\infty} \mu(t) = \infty. \tag{76}$$

Remark 2 (in Section 3) attract our attention to the fact that from (76) it is not possible to deduce finiteness of the first regeneration periods (for the multi-server queue) by following the same reasoning that in the proof of Theorem 1 for the single-server queue. Now we use another approach to show this finiteness, whose key element is the finiteness of the number of instants the basic process hits any (fixed) compact set during a regeneration period.

Theorem 7. *Under assumptions (41) and (42), we have (15), that is,*

$$\beta_1 < \infty \;\; w.p.1 \quad \textit{and} \quad T_1 < \infty \;\; w.p.1\,,$$

regardless of the initial states.

Proof. For any $t \geq 0$, let $S(t) \stackrel{\text{def}}{=} (S^{(1)}(t), \ldots, S^{(m)}(t))$ (recall that in the proof of Theorem 2 we defined $S^{(k)}(t)$ as the residual service time for server k at instant t, that is, the remaining amount of time needed to finish the service of the customer that is handling at instant t). As the basic process for the queue we introduce the process $Z = \{Z_t, t \geq 0\}$, defined by $Z(t) \stackrel{\text{def}}{=} (\nu(t), S(t))$, and consider the embedded process $(Z_n)_{n\geq 0}$ with $Z_n \stackrel{\text{def}}{=} (\nu_n, S(t_n^-))$. (Recall that ν_n was defined as $\nu(t_n^-)$.)

We split the proof into two steps. Firstly we will show that the number of arrivals within the interval $[0, t]$ which see the basic process Z in a compact set B, increases to infinity as $t \to \infty$. In the second step, we will show that the number of visits to the set B by the process $(Z_n)_{n\geq 0}$ within the first regeneration period $[0, \beta_1)$ is finite w.p.1 for any initial state. From these facts, it is immediate to see that the total number of regeneration cycles cannot be less than two, and thus, $\beta_1 < \infty$ w.p.1. Moreover, recall that $T_1 = t_0 + \tau_0 + \cdots + \tau_{\beta_1 - 1}$; then, from the finiteness of β_1, it also follows that $T_1 < \infty$ w.p.1.

Step 1: It follows from [15] that fixed $\delta_0 > 0$ (the one given by (75)), $M = M(\delta_0) > 0$ exists such that if we define

$$B_M \stackrel{\text{def}}{=} [0, M] \times \cdots \times [0, M] \in \mathbb{R}_+^m \,,$$

we have that

$$\liminf_{t\to\infty} \frac{1}{t} \int_0^t \mathbb{I}_{(S(u)\in B_M)}\, du \geq 1 - \frac{\delta_0}{2} \,. \tag{77}$$

Define $B \stackrel{\text{def}}{=} \{0, \ldots, m-1\} \times B_M$. Then, from (75) and (77) it is easy to obtain that

$$\liminf_{t\to\infty} \frac{1}{t} \int_0^t \mathbb{I}_{(Z(u)\in B)}\, du \geq \frac{\delta_0}{2} \,. \tag{78}$$

Let $u(0) = 0$ and define for any $n \geq 1$,

$$u(n) \stackrel{\text{def}}{=} \min\{\, k > u(n-1) \,:\, Z_k \in B \,\} \,.$$

Then $(t_{u(n)})_{n\geq 1}$ are the arrival instants such that $Z_{u(n)} \in B$ (when the customer $u(n)$ meets the process Z in the set B).

Define the number $G_0(t)$ of arrivals within interval $(t_0, t]$, for $t > t_0$, which see the process Z in the set B, that is

$$G_0(t) = \#\{n \geq 1 \,:\, t_{u(n)} \leq t\} \,.$$

Then

$$\int_0^t \mathbb{I}_{(Z(u)\in B)}\, du \leq t_0 + \big(\, G_0(t) + 1 \,\big) \big(\max_{0\leq n\leq N(t)-1} \tau_n \big) \,, \tag{79}$$

where, recall, $N(t)$ denotes the number of arrivals in the interval $[0,\, t]$ (including t_0). Note that possible first period $[0,\, t_0]$ (provided $Z(0) \in B$) is included in the first right hand side term.

By assumption, $\mathsf{E}\tau < \infty$, and it is known (see for instance, [32]) that in this case

$$\max_{0 \le n \le N(t)-1} \tau_n = o(t)\ w.p.1 \ \text{ as } \ t \to \infty\,. \tag{80}$$

It now follows from (78)-(80) that $G_0(t) \to \infty$ as $t \to \infty$. Therefore, the number of arrivals within $[0,\, t]$ which see Z in B goes to infinity as $t \to \infty$.

Step 2: Fixed the compact set B introduced in the previous step, we know that constants $\varepsilon > 0$ and $L < \infty$ exist (both depending on B) such that

$$\inf_{z \in B} \mathsf{P}_z(\beta_1 \le L) \ge \varepsilon\,. \tag{81}$$

For *any* fixed initial state $z \in (\mathbb{N} \cup \{0\}) \times \mathbb{R}^m_+$, we obtain from (81) that

$$\begin{aligned} 1 &\ge \sum_{k \ge 0} \mathsf{P}_z(Z_{kL} \in B,\, k\,L < \beta_1 \le (k+1)\,L) = \\ &= \sum_{k \ge 0} \mathsf{P}_z(\beta_1 \le (k+1)\,L \,|\, Z_{k\,L} \in B,\, \beta_1 > k\,L)\, \mathsf{P}(Z_{k\,L} \in B,\, \beta_1 > k\,L) \ge \\ &\ge \varepsilon \sum_{k \ge 0} \mathsf{P}_z(Z_{k\,L} \in B,\, \beta_1 > k\,L)\,. \end{aligned}$$

By analogy, we have for $\ell = 0, \ldots, L-1$,

$$1 \ge \varepsilon \sum_{k \ge 0} \mathsf{P}_z(Z_{k\,L+\ell} \in B,\, \beta_1 > k\,L + \ell)\,.$$

Summing up all inequalities varying ℓ, we obtain the following upper bound (by taking $n = K\,L + \ell$):

$$\begin{aligned} \sum_{n \ge 0} \mathsf{P}_z(Z_n \in B,\, \beta_1 > n) &= \mathsf{E}_z \Big(\sum_{n=0}^{\infty} \mathbb{I}_{(\beta_1 > n,\, Z_n \in B)} \Big) = \\ &= \mathsf{E}_z \Big(\sum_{n=0}^{\beta_1 - 1} \mathbb{I}_{(Z_n \in B)} \Big) \le \frac{L}{\varepsilon} < \infty\,, \end{aligned} \tag{82}$$

which is independent of initial state z. Let $D \stackrel{\text{def}}{=} \sum_{n=0}^{\beta_1 - 1} \mathbb{I}_{(Z_n \in B)}$, which is the number of visits to the set B by the process $(Z_n)_{n \ge 0}$ within the first

regeneration period $[0,\, \beta_1)$. We have just seen that regardless of initial state, z, $\mathsf{E}_z\, D < \infty$ and then,

$$\mathsf{P}_z\,(D < \infty) = 1 \quad \text{for any initial state} \;\; z\,. \qquad \square$$

Application. Finally we show how this approach allows to simplify considerably some steps on the study of stability analysis of the well-known multi-server $GI/G/m/K$ queue with a finite buffer K. Assume the *negative drift assumption* $\rho < m$ (41) and the *regeneration condition* $\mathsf{P}(\tau > S) > 0$ (42). We show that the renewal process of regenerations for this queue (when arrivals see an empty system) is positive recurrent for any initial state. Obviously,

$$\max_n \nu_n \le K + m \quad \text{and} \quad W(t) \le_{st} \sum_{k=1}^{K} S_k + \sum_{i=1}^{m} S^{(i)}(t)\,,$$

and the workload process is tight by the tightness of residual service time (see [26]). So previous analysis allows us to conclude that mean regeneration period both for continuous- and discrete-time renewal processes are finite in the zero-delayed case, that is,

$$\mathsf{E}_0\, T_1 < \infty \quad \text{and} \quad \mathsf{E}_0\, \beta_1 < \infty\,. \tag{83}$$

As to the first regeneration period β_1, we introduce the (tight) process $(Z_n)_{n\ge 0}$ defined by $Z_n \stackrel{\text{def}}{=} (\nu_n,\, W_n)$. Analogously to (82) we have that for any compact set B,

$$\mathsf{E}_z \Big(\sum_{n=0}^{\infty} \mathbb{I}_{(\beta_1 > n,\, Z_n \in B)} \Big) < \infty\,,$$

for any initial state z. Hence, $\mathsf{P}_z\,(\beta_1 > n,\, Z_n \in B) \to 0$ as $n \to \infty$. But

$$\mathsf{P}_z\,(\beta_1 > n) = \mathsf{P}_z\,(\beta_1 > n,\, Z_n \in B) + \mathsf{P}_z\,(\beta_1 > n,\, Z_n \notin B)\,, \tag{84}$$

and it follows from the tightness of the process $(Z_n)_{n\ge 0}$ that the probability $\mathsf{P}_z\,(\beta_1 > n,\, Z_n \notin B)$ can be done arbitrarily small for the compact set $B = [0,\, K+m] \times [0,\, D]$, with a constant $D > 0$ large enough. Hence,

$$\mathsf{P}_z\,(\beta_1 > n) \to 0 \quad \text{as} \quad n \to \infty\,,$$

for any initial state z, that is, $\beta_1 < \infty$ w.p.1. (and hence also $T_1 < \infty$ w.p.1.)

Acknowledgements

Evsey Morozov thanks the staff of CRM for nice hospitality and remarkable conditions during his visit CRM in September 2007 when the main part of this joint work has been done.

References

[1] S. Asmussen, *Applied Probability and Queues* (Wiley, N.Y. 1987).

[2] F. Baccelli and S. Foss, Stability of Jackson-type queueing networks, *Queueing Systems*, **17** (1994) 5-72.

[3] C.-S. Chang, J.A. Thomas and S.-H. Kiang, On the stability of open networks: A unified approach by stochastic dominance, *Queueing Systems*, **15** (1994) 239-260.

[4] H. Chen, Fluid approximation and stability of multiclass queueing networks: work-conserving disciplines, *Annals of Applied Probabability*, **5** (1995) 637-665.

[5] H. Chen and D. Yao, Stable priority disciplines for multiclass networks, in: *Stochastic Networks: Stability and Rare Events*, eds. P. Glasserman, K. Sigman, D. Yao, Lect. Notes in Statistics (Springer-Verlag, 1996).

[6] H. Chen and A. Mandelbaum, Discrete flow networks: bottlenecks analysis and fluid approximations, *Mathematics of Operations Research*, **16** (1991) 408–446.

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

[8] J. Dai, A fluid limit model criterion for instability of multiclass queueing networks, *Annals of Applied Probabability*, **6** (1996) 751–757.

[9] J. Dai and T. Kurtz, A multiclass station with Markovian feedback in heavy traffic, *Mathematics of Operations Research*, **20** (1995) 721–742.

[10] J. Dai and J. Vande Vate, Global stability of two-station queuing networks, in: *Stochastic Networks: Stability and Rare Events*, eds. P. Glasserman, K. Sigman, D. Yao, Lect. Notes in Statistics (Springer-Verlag, 1996).

[11] J. Dai and G. Weiss, Stability and instability of fluid models for reentrant lines, *Mathematics of Operations Research*, **21** (1996) 115–134.

[12] S. Foss and T. Konstantopoulos, An overview on some stochastic stability methods, *Journal of the Operations Research Society of Japan*, **47(4)**, (2004) 275–303.

[13] D. Iglehart and W. Whitt, Multiple channel queues in heavy traffic I, *Advances in Applied Probabability*, **2** (1970) 150–170.

[14] W. Feller, *An Introduction to Probability Theory* vol. 2 (Wiley, New York, 1971).

[15] H. Kaspi, A. Mandelbaum, Regenerative closed queueing networks, *Stochastics and Stochastics Reports*, 39 (1992) 239-258.

[16] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[17] S. Meyn and D. Down, Stability of generalized Jackson networks, *Annals of Applied Probability*, **4** (1994) 124–148.

[18] E. Morozov, A comparison theorem for queueing system with non-identical channels, *Lecture Notes in Mathematics*, **1546** (1993) 130–133.

[19] E. Morozov, Regeneration of a closed queueing network, *Journal of Mathematical Sciences*, **69** (1994) 1186–1192.

[20] E. Morozov, Wide sense regenerative processes with applications to multi-channel queues and networks, *Acta Appl. Math.*, **34** (1994) 189–212.

[21] E. Morozov, The stability of non-homogeneous queueing system with regenerative input, *Journal of Mathematical Sciences*, **89** (1997) 407–421.

[22] E. Morozov, The tightness in the ergodic analysis of regenerative queueing processes, *Queueing Systems*, **27** (1997) 179–203.

[23] E. Morozov, A comparison theorem for queueing system with non-identical channels, *Lecture Notes in Mathematics, Stability Problems for Stochastic Models*, **1546** (1993) 129–133.

[24] E. Morozov. Wide sense regenerative processes with applications to multi-channel queues and networks, *Acta Applicandae Math.*, **34** (1994)189–212.
[25] E. Morozov, The stability of non-homogeneous queueing system with regenerative input, *Journal of Mathematical Sciences*, **89** (1997) 407–421.
[26] E. Morozov, The tightness in the ergodic analysis of regenerative queueing processes, *Queueing Systems*, **27** (1997) 179–203.
[27] E. Morozov, Queueing network stability: an unified approach via renewal technique, *Proceedings of the Intern. Conf. Probabilistic Analysis of Rare Events*, Riga (1999) 125–135.
[28] E. Morozov, Instability of nonhomogeneous queueing networks, *Journal of Mathematical Sciences*, **112** (2002) 4155–4167.
[29] E. Morozov, Weak regeneration in modeling of queueing processes, *Queueing Systems*, **46** (2004) 295–315.
[30] G. Shedler, *Regeneration and Networks of Queues* (Springer-Verlag, 1987).
[31] G. Shedler, *Stochastic Regenerative Simulation* (Academic Press Inc., 1993).
[32] W.L. Smith, Regenerative stochastic processes, *Proc. Royal Soc.*, Ser. A, **232** (1955) 6–31.
[33] K. Sigman, The stability of open queueing networks, *Stochastic Process and Their Applications*, **35** (1990) 11–25.

Evsey Morozov
Institute for Applied Mathematical Research
Karelian Research Centre Russian Academy of Sciences

Rosario Delgado
Department of Mathematics
University Autonomous of Barcelona
Bellaterra (Cerdanyola del Vallès), Barcelona, Spain