On the stability of some queuing models

Attila Lovas [1,2]***, Miklós Rásonyi*** [1]

[1]Alfréd Rényi Institute of Mathematics, Budapest, Hungary, lovas@renyi.hu, rasonyi@renyi.hu
[2]Budapest University of Technology and Economics, Budapest, Hungary, lovas@math.bme.hu

Queuing models have many valuable applications in engineering and management sciences including scheduling, traffic control, facility design and telecommunication. In our contribution, we analyze single-server queuing systems with infinite buffer and requests numbered by $n \in \mathbb{N}$. Such models are described by a random iterative process driven by a two-dimensional sequence of non-negative random variables $(S_n, Z_n)_{n\in\mathbb{N}} \in \mathbb{R}_+^2$, where S_n denotes the service time of the n-th request and Z_n stands for the time between the arrival of n-th and $(n+1)$-th requests. For $n \in \mathbb{N}$, let W_n be the waiting time of the n-th request which is the time between the time of the request being received and the time when the request is being processed by the server. Starting with an empty queue ($W_0 = 0$), and assuming first-in, first-out (FIFO) service discipline, the sequence of waiting times $(W_n)_{n\in\mathbb{N}}$ satisfies the Lindley recursion

$$W_n = (W_{n-1} + S_{n-1} - Z_n)_+, \ n > 0.$$

Henceforth, $\mathcal{B}(\mathbb{R}_+)$ stands for the Borel σ-algebra of $\mathbb{R}_+$, and for an $\mathbb{R}_+$-valued random variable X we will denote by $\mathcal{L}(X)$ its law on $\mathcal{B}(\mathbb{R}_+)$. Let $\mathcal{P}$ be the set of Borel probability measures on $\mathbb{R}_+$. We consider the total variation metric on $\mathcal{P}$ which is given by $\mathrm{d}_{TV}(\mu_1, \mu_2) = |\mu_1 - \mu_2|(\mathbb{R}_+)$, $\mu_1, \mu_2 \in \mathcal{P}$, where $|\mu_1 - \mu_2|$ is the total variation of the signed measure $\mu_1 - \mu_2$. We say that the sequence of waiting times is stable if there exist a unique probability measure $\mu_* \in \mathcal{P}$ such that $\mathrm{d}_{TV}(\mathcal{L}(W_n), \mu_*) \to 0$ as $n \to \infty$, whatever the initialization W_0 is.

Loyens studied the stability of waiting times and introduced the terminology that a queue is *subcritical* if $\mathsf{E}(S_0) < \mathsf{E}(Z_0)$, *critical* if $\mathsf{E}(S_0) = \mathsf{E}(Z_0)$ and *supercritical* if $\mathsf{E}(S_0) > \mathsf{E}(Z_0)$. For mathematical convenience, from now on, we extend the process (S, Z) to the negative time axis, and start working with the index set $\mathbb{Z}$ instead of $\mathbb{N}$. Under the assumption that the process $(S_n, Z_n)_{n\in\mathbb{Z}}$ is (strong-sense) stationary and ergodic, Loynes showed that subcritical queues are stable, supercritical queues are unstable and critical queues can be stable, properly substable, or unstable Loynes [6]. However, in this general setting, it is just very little known about the limit distribution and the speed of convergence which would be necessary for the statistical analysis of such systems. The only quantitative result which we found in the literature regarding the

speed of the convergence is Theorem 4 on page 25 of Borovkov [1] which gives the following upper bound

$$\mathrm{d}_{TV}(\mathcal{L}(W_n), \mu_*) \leq \mathsf{P}\left(\min_{0<k<n} \sum_{j=1}^{k} \xi_j > \max\left(W_1, \xi_0 + \sup_{k\in\mathbb{N}} \sum_{j=1}^{k} \xi_{-j}\right)\right), \quad (1)$$

where $\xi_n = S_n - Z_{n+1}$, $n \in \mathbb{Z}$. Unfortunately, the expression standing on the right hand side of Eq. (1) does not provide a concrete and explicit rate estimate.

The ergodic theory of general state space Markov chains (see e.g. Meyn and Tweedie [5]) proved to be a powerful tool when $(S_n)_{n\in\mathbb{Z}}$, $(Z_n)_{n\in\mathbb{Z}}$ are i.i.d. sequences, independent of each other. However, this approach do not apply to the case when the process $(S_n, Z_n)_{n\in\mathbb{Z}}$ is merely stationary and ergodic as $(W_n)_{n\in\mathbb{N}}$ fails to be a Markovian process. We consider two special cases. If the inter-arrival times are i.i.d and the service times merely stationary *and* these two sequence are independent ($G/G_I/1/\infty$ queuing) then the process W is a Markov chain with driving noise Z in the random environment provided by S. And similarly, if the service times are i.i.d. and the (independent from service) inter-arrival times stationary ($G_I/G/1/\infty$ queuing) then W is a Markov chain driven by S in the random environment Z. Hence both these special cases of queuing systems fit into the scope of the framework we proposed in our recent paper (See Lovas and Rásonyi [2]).

We stipulate that the queue is sub-critical i.e. $\mathsf{E}[S_0] < \mathsf{E}[Z_0]$ (where the latter may be infinity). Furthermore, the sequences $(S_n)_{n\in\mathbb{N}}$ and $(Z_n)_{n\in\mathbb{N}}$ are assumed to be independent. We say that a sequence $(Y_n)_{n\in\mathbb{N}}$ of real-valued random variables satisfies a *Gärtner-Ellis-type condition* if there is $\eta > 0$ such that the limit

$$\Gamma(\alpha) := \lim_{n\to\infty} \frac{1}{n} \ln \mathsf{E}\left[e^{\alpha(Y_1+\ldots+Y_n)}\right] \quad (2)$$

exists for all $\alpha \in (-\eta, \eta)$ and Γ is differentiable on $(-\eta, \eta)$. Let us consider the following two set of conditions:

1. $(Z_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence, $(S_n)_{n\in\mathbb{N}}$ is uniformly bounded, ergodic, satisfying a Gärtner-Ellis type condition, and $\mathsf{P}(Z_0 > z) > 0$ for all $z > 0$.

2. $(S_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence, $(Z_n)_{n\in\mathbb{N}}$ is bounded, ergodic, satisfying a Gärtner-Ellis type condition. Furthermore, $\mathsf{E}[e^{\beta_0 S_0}] < \infty$ for some $\beta_0 > 0$ and $\mathcal{L}(S_0)$ has a density $s \mapsto f(s)$ (w.r.t. the Lebesgue measure) which is bounded away from 0 on compact subsets of $\mathbb{R}_+$.

Under our standing assumptions, we obtained that either family of conditions holds, there exists a probability μ_* on $\mathcal{P}$ such that

$$\mathrm{d}_{TV}(\mathcal{L}(W_n), \mu_*) \leq c_1 \exp\left(-c_2 n^{1/3}\right),$$

for some $c_1, c_2 > 0$. Furthermore, for an arbitrary measurable and bounded $\Phi : \mathbb{R}_+ \to \mathbb{R}$,

$$\frac{\Phi(W_1) + \ldots + \Phi(W_n)}{n} \to \int_{\mathbb{R}_+} \Phi(z)\mu_*(\mathrm{d}z),$$

in probability (See Lovas and Rásonyi [3]).

These results open door to the statistical analysis of $G/G_I/1/\infty$ and $G_I/G/1/\infty$ queuing models transcending the standard case with i.i.d arrival and service times.

Acknowledgements Both authors benefited from the support of the "Lendlet" grant LP 2015-6 of the Hungarian Academy of Sciences. The second author was also supported by the NKFIH (National Research, Development and Innovation Office, Hungary) grant KH 126505.

References

1. A. A. Borovkov, *Stochastic Processes in Queuing Theory*, Nauka, Moscow, 1972.

2. A. Lovas and M. Rásonyi, Markov chains in random environment with applications in queuing theory and machine learning, *Stochastic Processes and their Applications* **137** (2021) 294–326.

3. A. Lovas and M. Rásonyi, Ergodic theorems for queuing systems with dependent inter-arrival times, *arXiv preprint:2004.01475* (2021).

4. R. M. Loynes, The stability of a queue with non-independent inter-arrival and service times, *Mathematical Proceedings of the Cambridge Philosophical Society* **58** (1962) 497–520.

5. S. P. Meyn and R. L. Tweedie., *Markov chains and stochastic stability*, Springer-Verlag, 1993.