J. Appl. Probab. 1–4 (2024)
*doi:*10.1017/jpr.2024.78

ON TRANSIENCE OF M/G/∞ QUEUES

SERGUEI POPOV ,* *University of Porto*

Abstract

We consider an M/G/∞ queue with infinite expected service time. We then provide the transience/recurrence classification of the states (the system is said to be at state n if there are n customers being served), observing also that here (unlike irreducible Markov chains, for example) it is possible for recurrent and transient states to coexist. We also prove a lower bound on the growth speed in the transient case.

Keywords: Recurrence; service time; heavy tails

2020 Mathematics Subject Classification: Primary 60K25
Secondary 60G55

In this note we consider a classical M/G/∞ queue (see e.g. [3]): the customers arrive according to a Poisson process with rate λ; upon arrival, a customer immediately enters the service, and the service times are i.i.d. (non-negative) random variables with some general distribution. For notational convenience, let S be a generic random variable with that distribution. We also assume that at time 0 there are no customers being served. Let Y_t denote the number of customers in the system at time t, which we also refer to as the *state of the system* at time t; note that, in general, Y is not a Markov process. Nevertheless, let us still call a state m *recurrent* if the set $\{t : Y_t = m\}$ is a.s. unbounded (i.e. m is 'visited infinitely many times') and *transient* if that set is a.s. bounded (i.e. the system escapes from m eventually). Let us also observe that $\mathbb{P}[S = \infty] > 0$ clearly implies that Y goes to infinity (indeed, it is straightforward to obtain that $\liminf_{t\to\infty} (Y_t/t) \geq \lambda\mathbb{P}[S = \infty]$), so from now on we assume that $S < \infty$ a.s.

We are mainly interested in the situation where the system is *unstable*, i.e. when $\mathbb{E}S = \infty$. In this situation, in principle, our (Markovian) intuition tells us that the system can be *transient* (in the sense $Y_t \to \infty$ a.s.) or recurrent (i.e. all states are visited infinitely many times a.s.). However, it turns out that for this model the complete picture is more complicated.

Theorem 1. *Define*

$$k_0 = \min\left\{k \in \mathbb{Z}_+ : \int_0^\infty (\mathbb{E}(S \wedge t))^k \exp(-\lambda\mathbb{E}(S \wedge t))\, \mathrm{d}t = \infty\right\} \tag{1}$$

(with the convention $\min\emptyset = +\infty$*). Then*

$$\liminf_{t\to\infty} Y_t = k_0 \quad a.s.$$

Received 17 July 2024; accepted 25 July 2024.
* Postal address: Centro de Matemática, University of Porto, Rua do Campo Alegre 687, 4169–007 Porto, Portugal. Email address: serguei.popov@fc.up.pt

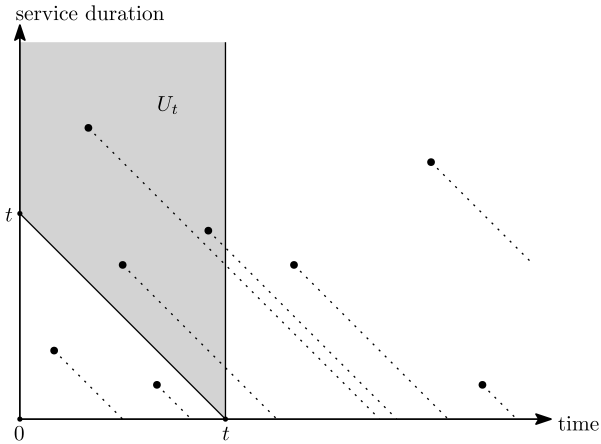

FIGURE 1. A Poisson representation of M/G/∞. In this example, there are *exactly* three customers at time t.

In particular, if

$$\int_0^\infty (\mathbb{E}(S \wedge t))^k \exp(-\lambda \mathbb{E}(S \wedge t))\, \mathrm{d}t < \infty \quad \textit{for all } k \geq 0, \tag{2}$$

then the system is transient; if

$$\int_0^\infty \exp(-\lambda \mathbb{E}(S \wedge t))\, \mathrm{d}t = \infty,$$

then the system is recurrent.

Before proving this result, we make the following remark. Let us define $M(t)$ to be the maximal remaining service time of the customers that are present at time t. This is a so-called *extremal shot noise process*; see [1] and references therein. It is not difficult to obtain that transience of $M(\cdot)$ is the same as transience of state 0 in M/G/∞; then Theorem 2.5 of [1] provides a criterion for the transience of $M(\cdot)$ (and therefore for the transience of state 0 in our situation).

Proof of Theorem 1. We start with a simple observation: for any $j \geq 0$, $\{\liminf Y_t = j\}$ is a tail event, so it has probability 0 or 1. This implies that $\liminf Y_t$ is a.s. a constant (which may be equal to $+\infty$).

We use the following representation of the process (see Figure 1): consider a Poisson process in $\mathbb{R}_+^2$, with the intensity measure $\lambda\, \mathrm{d}t \times \mathrm{d}F_S(u)$, where $F_S(u) = \mathbb{P}[S \leq u]$ is the distribution function of S. Then, a point (t, u) of this Poisson process is interpreted in the following way: a customer arrived at time t and the duration of its service will be u. Now, draw a (dotted) line in the southeast direction from each point, as shown in the picture; as long as this line stays in $\mathbb{R}_+^2$, the corresponding customer is present in the system. If we draw a vertical line from $(t, 0)$ in the upwards direction, then the number of dotted lines it intersects is equal to Y_t.

Next, for $k \in \mathbb{Z}_+$ let

$$T_k := \{t \geq 0 : \ Y_t = k\}$$

denote the set of times when the system has exactly k customers, and let

$$U_t = \big\{(s, u) \in \mathbb{R}_+^2 : \ s \in [0, t],\, u \geq t - s\big\}.$$

We note that Y_t equals the number of points in U_t, which has Poisson distribution with mean

$$\int_{U_t} \lambda \, \mathrm{d}t \, \mathrm{d}F_S(u) = \lambda \mathbb{E}(S \wedge t).$$

Therefore, by Fubini's theorem, we have (here $|A|$ stands for the Lebesgue measure of $A \subset \mathbb{R}$)

$$\mathbb{E}|T_k| = \mathbb{E} \int_0^\infty \mathbf{1}\{Y_t = k\} \, \mathrm{d}t = \frac{\lambda^k}{k!} \int_0^\infty (\mathbb{E}(S \wedge t))^k \exp(-\lambda \mathbb{E}(S \wedge t)) \, \mathrm{d}t. \tag{3}$$

Now, assume that $\mathbb{E}|T_k| < \infty$ for some $k \geq 0$; it automatically implies that $\mathbb{E}|T_\ell| < \infty$ for $0 \leq \ell \leq k$. This means that $|T_0|, \ldots, |T_k|$ are a.s. finite, and let us show that $T_0, \ldots, T_k$ have to be a.s. bounded (this is a small technical issue that we have to resolve because we are considering continuous time). Probably the cleanest way to see this is as follows. First notice that, in fact, T_0 is a union of intervals of random i.i.d. (with $\operatorname{Exp}(\lambda)$ distribution) lengths, because each time the system becomes empty it will remain so until the arrival of the next customer. Therefore, $|T_0| < \infty$ clearly means that $\sup T_0 \leq K_0$ for some (random) K_0. Now, *after* K_0 there are no longer any $1 \to 0$ transitions, so the remaining part of T_1 again becomes a union of such intervals, meaning that it should be bounded as well; we then repeat this reasoning a suitable number of times to finally obtain that T_k must be a.s. bounded. This implies that $\liminf_{t\to\infty} Y_t \geq k_0$ a.s.

Next, assume that $\{0, \ldots, k\}$ is a *transient set*, in the sense that $\liminf_{t\to\infty} Y_t \geq k+1$ a.s.; let us show that this implies that $\mathbb{E}|T_k| < \infty$. Indeed, first we can choose a sufficiently large $h > 0$ in such a way that

$$\mathbb{P}[Y_t \geq k+1 \text{ for all } t \geq h] \geq \frac{1}{2}.$$

Define a stopping time $\tau = \inf\{t \geq h : \; Y_t \leq k\}$ (again, with the convention $\inf \emptyset = +\infty$). Then, a crucial observation is that what one sees after τ is a superposition of two *independent* systems: one is formed by those customers (with their remaining lifetimes) present at τ, and the other is a copy of the original system. Then, a simple coin-tossing argument, together with the fact that an initially non-empty system (i.e. with some customers being served, with any assumptions on their remaining service times) dominates an initially empty system, shows that $|T_k|$ (in fact, $|T_0| + \cdots + |T_k|$) is dominated by $h \times \operatorname{Geom}_0(\frac{1}{2})$ random variable and therefore has a finite expectation. It means that we have $\liminf_{t\to\infty} Y_t \leq k_0$ a.s. (because otherwise, in the situation when $k_0 < \infty$, we would have $\mathbb{E}|T_{k_0}| < \infty$, which, by definition, is not the case). This concludes the proof of Theorem 1. □

Regarding this result, we may observe that in most situations one would have $k_0 = 0$ or $+\infty$; this is because convergence of such integrals is usually determined by what is in the exponent. Still, it is not difficult to construct 'strange examples' with $0 < k_0 < \infty$, i.e. where the process will visit $\{0, \ldots, k_0 - 1\}$ only finitely many times, but will hit every $k \geq k_0$ infinitely often a.s. (a behaviour one cannot have with irreducible Markov chains). For instance, let $\lambda = 1$ and fix $b > 0$; next, consider a service time distribution such that

$$1 - F_S(u) = \frac{1}{u} + \frac{b}{u \ln u}$$

for large enough u. Then it is elementary to obtain that $\mathbb{E}(S \wedge t) = \ln t + b \ln\ln t + \mathrm{O}(1)$ and the integrals in (1) diverge whenever $k \geq b - 1$, meaning that $k_0 = \lceil b \rceil - 1$.

https://doi.org/10.1017/jpr.2024.78 Published online by Cambridge University Press

Now, in the situation when (2) holds and Y is transient, it may also be useful to be able to say something about the speed of convergence of Y_t to infinity. We do not intend to enter deeply into this question here, but only prove a particular result needed for future reference. Namely, in [2] we work with a different model which in some sense dominates $\mathrm{M}/\mathrm{G}/\infty$. So, we will now give a lower bound on the growth of Y_t; more specifically, we will show that under certain conditions Y_t will eventually be at least a constant fraction of its expected value. For $q \in (0, 1)$, let us define $\gamma_q = 1 - q - q \ln q^{-1} > 0$.

Theorem 2. *Fix $q \in (0, 1)$ and assume that*

$$\int_0^\infty \exp(-\gamma_q \lambda \mathbb{E}(S \wedge t))\, \mathrm{d}t < \infty. \tag{4}$$

Then

$$\mathbb{P}[Y_t \geq q\lambda \mathbb{E}(S \wedge t) \textit{ for all large enough } t] = 1. \tag{5}$$

Proof. Let

$$H_q = \{t \geq 0 : \; Y_t < q\lambda \mathbb{E}(S \wedge t)\};$$

our goal is to show that H_q is a.s. bounded in the case when (4) holds. We recall a standard (Chernoff) tail bound: if X is Poisson(μ) and $q \in (0, 1)$, then

$$\mathbb{P}[X \leq q\mu] \leq \exp(-(q\mu \ln q + \mu - q\mu)) = \exp(-\gamma_q \mu). \tag{6}$$

Then, analogously to (3), we obtain from (6) that

$$\mathbb{E}|H_q| \leq \int_0^\infty \exp(-\gamma_q \lambda \mathbb{E}(S \wedge t))\, \mathrm{d}t; \tag{7}$$

so, by (4), we have $\mathbb{E}|H_q| < \infty$, meaning that $|H_q| < \infty$ a.s. To see that this has to imply that H_q is a.s. bounded, analogously to the proof of Theorem 1, one can reason in the following way. If $t \in H_q$, then $s \in H_q$ for all $s \in (t, A_t)$, where A_t is the first moment after t when a customer arrives in the system. This implies that the lengths of the intervals that constitute H_q dominate a sequence of i.i.d. random variables with Exp(λ) distribution; in its turn, this clearly implies that if $|H_q|$ is finite then it has to be bounded. □

Funding information

There are no funding bodies to thank relating to the creation of this article.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] FOUCART, C. AND LINGLONG Y. (2023). Extremal shot noise processes and random cutout sets. Available at arXiv:2302.03082 (to appear in *Bernoulli*).

[2] MENSHIKOV, M., POPOV, S. AND WADE, A. (2024). Semi-infinite particle systems with exclusion interaction and heterogeneous jump rates. Available at arXiv:2405.05246.

[3] NEWELL, G. F. (1966). The $\mathrm{M}/\mathrm{G}/\infty$ queue. *SIAM J. Appl. Math.* **14**, 86–88.

https://doi.org/10.1017/jpr.2024.78 Published online by Cambridge University Press