On the stability of some queuing models [1]

Attila Lovas[1,2], Miklós Rásonyi[1]

[1] Alfréd Rényi Institute of Mathematics
[2] BUTE, Department of Analysis

XXXVI International Seminar on Stability Problems for Stochastic Models

Petrozavodsk, Karelia, Russia, June 21–25, 2021

[1] Both authors benefited from the support of the "Lendület" grant LP 2015-6 of the Hungarian Academy of Sciences. The second author was also supported by the NKFIH (National Research, Development and Innovation

1 Motivation

2 The main results

3 Markov Chains in Random Environments

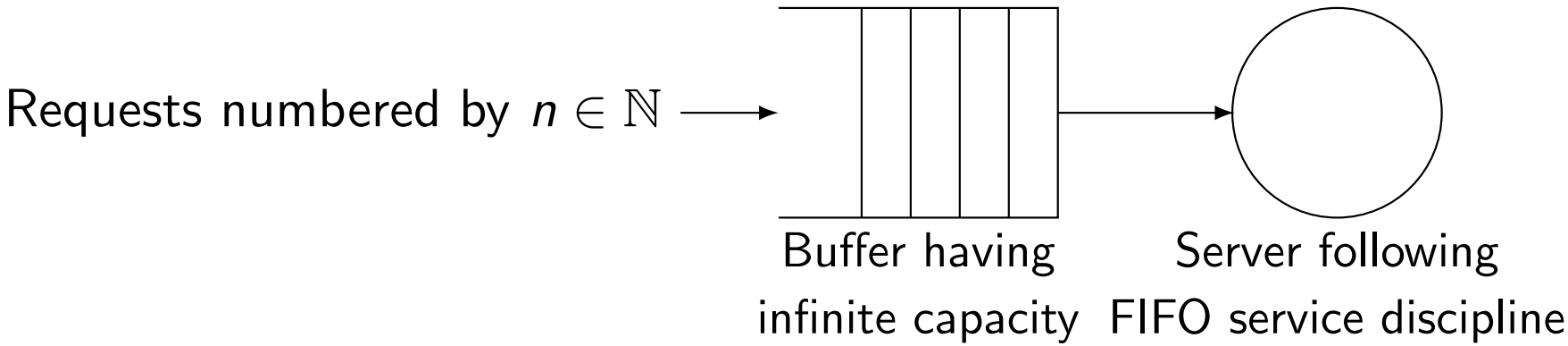
4 The sketch of the proof

Introduction

Queuing models have been widely applied in engineering and management sciences including but not limited to

- facility and service design,
- traffic control and scheduling,
- analysis of telecommunication networks.

We analyze single-server queuing systems with...



The inter-arrival and service times:

- $Z_n \in \mathbb{R}_+$ is the time between the arrival of n-th and $(n+1)$-th requests,
- $S_n \in \mathbb{R}_+$ denotes the service time of the n-th request.

Starting with an empty queue, the sequence of waiting times $(W_n)_{n\in\mathbb{N}}$ satisfies the Lindley recursion

$$W_n = (W_{n-1} + S_{n-1} - Z_n)_+, \ n > 0$$
$$W_0 = 0.$$

Further notations:

- $\mathcal{B}(\mathbb{R}_+)$ is the Borel σ-algebra of $\mathbb{R}_+$.
- For an $\mathbb{R}_+$-valued r.v. X, $\mathcal{L}(X)$ denotes its law on $\mathcal{B}(\mathbb{R}_+)$.
- Let $\mathcal{P}$ be the set of Borel probability measures on $\mathbb{R}_+$.
- We will use the total variation metric:

$$\mathrm{d}_{TV}(\mu_1, \mu_2) = |\mu_1 - \mu_2|(\mathbb{R}_+), \ \mu_1, \mu_2 \in \mathcal{P}.$$

Stability of waiting times

Definition

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.

Theorem (Loynes, 1962)

Assuming that the process $(S_n, Z_n)_{n\in\mathbb{Z}}$ is strongly stationary and ergodic, the following holds:

1. $E(S_0) < E(Z_0) \Rightarrow (W_n)_{n\in\mathbb{N}}$ *is stable (sub-critical queues)*
2. $E(S_0) > E(Z_0) \Rightarrow (W_n)_{n\in\mathbb{N}}$ *is unstable (super-critical queues)*
3. $E(S_0) = E(Z_0) \Rightarrow (W_n)_{n\in\mathbb{N}}$ *can be stable, properly substable, or unstable (critical queues)*

From now on, we will work in the sub-critical regime.

What we know about the speed of convergence?

If $(S_n)_{n\in\mathbb{Z}}$, $(Z_n)_{n\in\mathbb{Z}}$ are i.i.d. sequences, independent of each other, then $(W_n)_{n\in\mathbb{N}}$ is a Markov chain satisfying the Doeblin condition.

$\Rightarrow$ Geometric rate of convergence can be established.

However, if $(S_n, Z_n)_{n\in\mathbb{Z}}$ is merely stationary and ergodic, then the only bound which we have found is

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

where $\xi_n = S_n - Z_{n+1}$, $n \in \mathbb{Z}$.

1. This bound is not a concrete and explicit rate estimate.
2. It is inappropriate for statistical analysis.

Dependent queuing

The non-i.i.d. setting is interesting because...

1. dependence arises often in queuing networks where arrival processes are departure processes of other queues,
2. processing operations like batching or the presence of multiple, different classes of requests can also cause dependence.

What makes it challenging is that $(W_n)_{n\in\mathbb{N}}$ **fails to be Markovian**.

We assume that $(S_n)_{n\in\mathbb{Z}}$ and $(Z_n)_{n\in\mathbb{Z}}$ are strongly stationary, ergodic and independent of each other.

We consider two interesting cases:

I. $G/G_I/1/\infty$ queuing: $(Z_n)_{n\in\mathbb{Z}}$ is i.i.d.

II. $G_I/G/1/\infty$ queuing: $(S_n)_{n\in\mathbb{Z}}$ is i.i.d.

but $(S_n)_{n\in\mathbb{Z}}$ and $(Z_n)_{n\in\mathbb{Z}}$ are not required to be i.i.d. at the same time.

Definition

A sequence $(Y_n)_{n\in\mathbb{N}}$ of real-valued random variables satisfies a *Gärtner-Ellis-type condition* (G-Ec) if there is $\eta > 0$ such that the limit

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

exists for all $\alpha \in (-\eta, \eta)$ and Γ is differentiable on $(-\eta, \eta)$.

Assumption (for $G/G_I/1/\infty$ queuing)

The sequence of service times $(S_n)_{n\in\mathbb{N}}$ is uniformly bounded, satisfies the G-Ec, and $P(Z_0 > z) > 0$ for all $z > 0$.

Assumption (for $G_I/G/1/\infty$ queuing)

The sequence of inter-arrival times $(Z_n)_{n\in\mathbb{N}}$ is bounded, satisfies the G-Ec, and $\exists \beta_0 > 0$ $E[e^{\beta_0 S_0}] < \infty$ moreover $\mathcal{L}(S_0)$ has a density $s \mapsto f(s)$ (w.r.t. the Lebesgue measure) which is bounded away from 0 on compact subsets of $\mathbb{R}_+$.

Theorem

Under our standing assumptions, 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.

This result opens 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.

Let $Q : \mathbb{R}_+ \times \mathbb{R}_+ \times \mathcal{B}(\mathbb{R}_+) \to [0,1]$ be a mapping such that

I. for all $B \in \mathcal{B}(\mathbb{R}_+)$ the function $(x,y) \mapsto Q(x,y,B)$ is $\mathcal{B}(\mathbb{R}_+) \otimes \mathcal{B}(\mathbb{R}_+)$-measurable and

II. for all $(y,x) \in \mathbb{R}_+ \times \mathbb{R}_+$, $B \mapsto Q(y,x,B) \in \mathcal{P}$.

Let $Y : \mathbb{Z} \times \Omega \to \mathbb{R}_+$ be a strongly stationary stochastic process.

We will consider $\mathbb{R}_+$-valued process $(X_n)_{n\in\mathbb{N}}$ such that $X_0 = x_0 \in \mathbb{R}_+$ is fixed and for $n \in \mathbb{N}$,

$$\mathbb{P}(X_{n+1} \in B \mid \sigma(X_s,\, 0 \le s \le n;\; Y_s,\, s \in \mathbb{Z})) = Q(Y_n, X_n, B)\ \mathbb{P}-\text{a.s.}.$$

We interpret the process $(X_n)_{n\in\mathbb{N}}$ as a **Markov chain in a random environment (MCRE)** driven by the process $(Y_n)_{n\in\mathbb{Z}}$.

What makes $G/G_I/1/\infty$ and $G_I/G/1/\infty$ queuing amenable to analysis? $\Rightarrow$ **The key observation: Both of these processes is a MCRE.**

Lemma

For $G/G_I/1/\infty$ queuing, $(W_n)_{n\in\mathbb{N}}$ is a MCRE, where $(Z_n)_{n\in\mathbb{Z}}$ is the driving noise and $(S_n)_{n\in\mathbb{Z}}$ is the random environment, and vica versa, for $G_I/G/1/\infty$ queuing, the sequence of waiting times is also a MCRE with driving noise $(S_n)_{n\in\mathbb{Z}}$ and random environment $(Z_n)_{n\in\mathbb{Z}}$.

Proof.

We prove the latter case only. By the Lindley recursion and the independence of service times, we have

$$\begin{aligned}
&\mathbb{P}\left(W_{n+1}\in A \mid \sigma(W_i,\, 0\le i\le n;\, Z_m,\, m\in\mathbb{Z})\right) = \\
&= \mathbb{P}\left((W_n+S_n-Z_{n+1})_+\in A \mid \sigma(W_i,\, 0\le i\le n;\, Z_m,\, m\in\mathbb{Z})\right) \\
&= \int_{\mathbb{R}_+} \mathbb{1}_{(W_n+s-Z_{n+1})_+\in A}\, f(s)\,\mathrm{d}s = Q(Z_{n+1}, W_n, A),
\end{aligned}$$

where f is the density function of $\mathrm{Law}(S_0)$, and the parameterized kernel Q is given by $Q(z,w,A) := \mathbb{P}\left[(w+S_0-z)_+\in A\right]$. □

Ergodicity and stability of MCREs

Assumption (Foster–Lyapunov-type drift condition)

A) There exist measurable functions $V, K, \gamma : \mathbb{R}_+ \to \mathbb{R}_+$ *such that*

$$\int_0^\infty V(s)Q(y,x,\mathrm{d}s) \le \gamma(y)V(x) + K(y) \; x,y \in \mathbb{R}_+.$$

B) Furthermore, $\bar{\gamma} := \limsup_{n\to\infty} \mathbb{E}^{1/n}\left(K(Y_0)\prod_{k=1}^n \gamma(Y_k)\right) < 1.$

Assumption (Minorization condition)

A) For some $0 < \varepsilon < 1/\bar{\gamma}^{1/2} - 1$, *there is a measurable function* $\alpha : \mathbb{R}_+ \to (0,1)$ *and a probability kernel* $\kappa : \mathbb{R}_+ \times \mathcal{B}(\mathbb{R}_+) \to [0,1]$ *such that, for all* $y \in \mathbb{R}_+$, $A \in \mathcal{B}(\mathbb{R}_+)$, *and* $R(y) = \frac{2K(y)}{\varepsilon\gamma(y)}$

$$\inf_{x \in V^{-1}([0,R(y)])} Q(y,x,A) \ge (1-\alpha(y))\kappa(y,A).$$

B) Furthermore, $\lim_{n\to\infty} \mathbb{E}^{1/n^\theta}\left[\alpha(Y_0)^n\right] = 0$ *holds for some* $0 < \theta < 1$.

Ergodicity and stability of MCREs

Theorem

Under our assumptions, there exists a probability law $\mu_* \in \mathcal{P}$ such that

$$d_{TV}(Law(X_n), \mu_*) \leq C_0 \left(e^{-\nu_1 n^{2/3}} + \sum_{k=n}^{\infty} \mathbb{E}^{\frac{\nu_2}{k^{\kappa\theta}}} [\alpha^{k^\kappa - 1}(Y_0)] \right)$$

holds for all $n \in \mathbb{N}$, with some $C_0, \nu_1, \nu_2 > 0$, where $\kappa := \frac{1}{3(1-\theta)}$.

If Y is ergodic, then for any bounded and measurable $\Phi : \mathbb{R}_+ \to \mathbb{R}$

$$\frac{\Phi(X_1) + \ldots + \Phi(X_n)}{n} \to \int_{\mathcal{X}} \Phi(z)\, \mu_*(\mathrm{d}z),\ n \to \infty$$

holds in probability.

We prove the $G_I/G/1/\infty$ case.

The analogue statement for $G/G_I/1/\infty$ queuing can be proved along similar lines.

Let us recall the assumptions:

- The sequence of inter-arrival times $(Z_n)_{n\in\mathbb{Z}}$ is bounded, stationary, ergodic, and satisfies the G-Ec.
- The sequence of service times $(S_n)_{n\in\mathbb{Z}}$ is
  - i.i.d. and independent of $(Z_n)_{n\in\mathbb{Z}}$.
  - $\exists \beta_0 > 0$ $\mathsf{E}[e^{\beta_0 S_0}] < \infty$,
  - $\mathcal{L}(S_0)$ has a density $s \mapsto f(s)$ (w.r.t. the Lebesgue measure) which is bounded away from 0 on compact subsets of $\mathbb{R}_+$.

The following lemmas ensure that both the Foster-Lyapunov drift condition and the minorization condition holds.

Lemma (Drift condition, part A))

For $\bar{\beta} > 0$, we define $V(x) = e^{\bar{\beta}x} - 1$. Then

$$\int_0^\infty V(s)Q(y,x,\mathrm{d}s) \leq \gamma(y)V(x) + K(y)\ x,y \in \mathbb{R}_+.$$

holds with $K(z) := \gamma(z) := e^{-\bar{\beta}z}\mathbb{E}\left(e^{\bar{\beta}S_0}\right)$.

Proof.

$$\begin{aligned}\int_0^\infty V(s)Q(y,x,\mathrm{d}s) &= \mathbb{E}\left(e^{\hat{\beta}(x+S_0-z)_+} - 1\right) \leq \mathbb{E}\left(e^{\hat{\beta}(x+S_0-z)}\right) \\ &= e^{-\bar{\beta}z}\mathbb{E}\left(e^{\bar{\beta}S_0}\right) V(x) + e^{-\bar{\beta}z}\mathbb{E}\left(e^{\bar{\beta}S_0}\right)\end{aligned}$$

□

Lemma (Drift condition, part B))

There exists $\bar{\beta} > 0$ *such that*

$$\bar{\gamma} := \limsup_{n\to\infty} \mathbb{E}^{1/n}\left(K(Z_0)\prod_{k=1}^{n}\gamma(Z_k)\right) < 1.$$

Proof.

An equivalent formulation of the statement is

$$\exists \bar{\beta} > 0 \text{ s.t. } \limsup_{n\to\infty} \underbrace{\frac{1}{n}\log \mathbb{E}\left[e^{\bar{\beta}\sum_{j=1}^{n}(S_j - Z_j)}\right]}_{\text{we define } =:\lambda_n(\bar{\beta})} < 0,$$

where these λ_n are convex, moreover the point-wise limit $\lambda(\beta) := \lim_{n\to\infty}\lambda_n(\beta) = \Gamma(-\beta) + \log\mathbb{E}\left(e^{\beta S_0}\right)$ exists and finite for $|\beta| < \min(\eta, \beta_0/2)$.

Proof (Cont.)

For $0 < \beta < \min(\eta, \beta_0/2)$, we define

$$\psi_n(\beta) := \frac{1}{\beta}\mathbb{E}\left(e^{\beta\sum_{j=1}^{n}(S_j - Z_j)} - 1\right).$$

By the Lagrange mean value theorem and measurable selection, there exists a random variable $\xi_n(\beta) \in [0, \beta]$ such that

$$\psi_n(\beta) = \mathbb{E}\left[\left(\sum_{j=1}^{n}(S_j - Z_j)\right) e^{\xi_n(\beta)\sum_{j=1}^{n}(S_j - Z_j)}\right].$$

which is integrable. Hence reverse Fatou's lemma shows that

$$\limsup_{\beta\to 0+} \psi_n(\beta) \le \mathbb{E}\left[\sum_{j=1}^{n}(S_j - Z_j)\right] = n\mathbb{E}(S_0 - Z_0) < 0.$$

Proof (Cont.)

From this follows that, for all $n \geq 1$,

$$\lambda_n'(0) \leq \mathbb{E}(S_0 - Z_0) < 0.$$

By the Gärtner-Ellis-type property of $(Z_n)_{n\in\mathbb{N}}$,

$$\lambda_n(\beta) \to \lambda(\beta) \text{ for } 0 < \beta < \min(\eta, \beta_0/2).$$

Furthermore, the convexity of λ_ns implies that $\lambda_n'(0) \to \lambda'(0)$ hence $\lambda'(0) < 0$.

To sum up, we obtained

$$\lambda(0) = 0 \text{ and } \lambda'(0) < 0$$

which shows that there exists $\bar{\beta} > 0$ satisfying $\lambda(\bar{\beta}) < 0$ which proves our claim. □

Lemma (Minorization condition, part A))

Let $M > 0$ be such that $\mathrm{Ran}(Z) \subset [0, M]$, and $\varepsilon = \frac{1}{2}\left(\bar{\gamma}^{1/2} - 1\right)$. Then there exists $h > 0$ and $\alpha \in (0, 1)$ such that

$$\inf_{w \in V^{-1}([0,R(z)])} Q(z, w, A) \geq (1 - \alpha)\mathrm{Leb}(A \cap [h, h + 1])$$

Proof.

Observe that $R(z) = 2K(z)/(\varepsilon\gamma(z)) = 2/\varepsilon$ does not depend on z.

$$w \in V^{-1}([0, R(z)]) \Leftrightarrow 0 \leq w \leq \frac{1}{\bar{\beta}} \log\left(\frac{2}{\varepsilon} + 1\right)$$

Let $h := (1/\bar{\beta}) \log(2/\varepsilon + 1)$. So, we can write

$$\begin{aligned} \inf_{w \in [0,h]} Q(z, w, A) &= \inf_{w \in [0,h]} \mathbb{P}\left((w + S_0 - z)_+ \in A\right) \\ &\geq \inf_{w \in [0,h]} \mathbb{P}\left(w + S_0 - z \in A\right) \end{aligned}$$

Proof(Cont.)

$$\inf_{w\in[0,h]} \mathbb{P}\left(w + S_0 - z \in A\right) = \inf_{w\in[0,h]} \int_A f(s+z-w)\mathrm{d}s$$

$$\geq \inf_{w\in[0,h]} \int_{A\cap[h,h+1]} f(s+z-w)\mathrm{d}s$$

$$\geq \left[\inf_{v\in[0,M+h-1]} f(v)\right] \mathrm{Leb}(A\cap[h,h+1]),$$

where we used that $f : [0,\infty) \to (0,\infty)$ is bounded away from zero on $[0, M+h-1]$ which is a *fixed compact set* since M and h depends only on the distribution of S_0 and Z_0.

So, the minorization condition holds with
$\alpha = 1 - \inf_{v\in[0,M+h-1]} f(v)$ and $h := (1/\bar{\beta}) \log(2/\varepsilon + 1)$. □

Lemma (Minorization condition, Part B))

$\lim_{n\to\infty} \mathbb{E}^{1/n^\theta}\left[\alpha(Z_0)^n\right] = 0$ *for some* $0 < \theta < 1$.

Proof.

In our case $0 < \alpha < 1$ is a constant hence the claimed identity trivially holds for any $0 < \theta < 1$. □

Thank you for your attention!



FIGURE: Police dogs in China queuing up for their food.[2]

[2]The picture is taken from
https://www.awesomeinventions.com/queueing-police-dogs-in-china/

References

- A. A. Borovkov, *Stochastic Processes in Queuing Theory*, Nauka, Moscow, 1972.
- 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.
- A. Lovas and M. Rásonyi, Ergodic theorems for queuing systems with dependent inter-arrival times, *arXiv preprint:2004.01475* (2021).
- 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.
- S. P. Meyn and R. L. Tweedie., *Markov chains and stochastic stability*, Springer-Verlag, 1993.