Steady-State Convergence of the Continuous-Time JSQ System with General Distributions in Heavy Traffic

J. G. Dai
School of Operations Research and Information Engineering
Cornell University
jd694@cornell.edu

Jin Guang
School of Data Science
The Chinese University of Hong Kong, Shenzhen
jinguang@link.cuhk.edu.cn

Yaosheng Xu
Booth School of Business
University of Chicago
yaosheng.xu@chicagobooth.edu

ABSTRACT

This paper studies the continuous-time join-the-shortest-queue (JSQ) system with general interarrival and service distributions. Under a much weaker assumption than the one in the literature, we prove that each station's scaled steady-state queue length weakly converges to an identical exponential random variable in heavy traffic. Specifically, we establish our results by only assuming $2+\delta_0$ moment on the arrival and service distributions for some $\delta_0 > 0$. Our proof exploits the Palm version of the basic adjoint relationship (BAR) approach as a key technique.

1. INTRODUCTION

We consider a continuous-time queueing system with J parallel service stations, each with a single server and an infinite waiting queue. Jobs arrive at the system following a renewal process, and service times for each station are independent and identically distributed (i.i.d.) with general distributions. When a job arrives, it is routed to the station with the shortest queue length. This policy is known as the join-the-shortest-queue (JSQ) policy, and the system employing it is called the JSQ system. The JSQ policy is to equalize the queue lengths across stations, thereby reducing the average waiting time.

In this paper, we show that the scaled steady-state queue length for each station weakly converges to the same exponential random variable in heavy traffic. Specifically, we consider a sequence of JSQ systems indexed by $r \in (0,1)$. In heavy traffic with $r \to 0$ with fixed J, we prove that if interarrival and service times have finite $2+\delta_0$ moments for some $\delta_0 > 0$, then

$$\left(rZ_1^{(r)}, \ldots, rZ_J^{(r)}\right) \Rightarrow (Z^*, \ldots, Z^*),$$

where $\Rightarrow$ denotes convergence in distribution, $Z_j^{(r)}$ denotes the steady-state queue length at station j for the rth system, and Z^* is an exponential random variable. This result depends on the fact that the steady-state queue length vector collapses to the line where all queue lengths are equal, in the sense that the deviations from the line are uniformly bounded. This phenomenon of queueing systems in heavy traffic is called state-space collapse (SSC).

Based on a discrete-time framework, the literature studied similar results on SSC and heavy traffic limit for the JSQ system using drift method [6], transform method [8] and Stein's method [9]. These studies, however, assumed interarrival and service times with bounded supports, implying boundedness of all moments. In this work, we consider the continuous-time JSQ system and relax this assumption, demonstrating that only the $2+\delta_0$ moment is sufficient to ensure steady-state convergence in heavy traffic.

Our result is underpinned by a novel methodology called the basic adjoint relationship (BAR) approach. A significant benefit of the BAR approach is that it directly characterizes the stationary distribution of a queueing system, eliminating the need to address their transient dynamics. This approach has been successfully applied in recent studies [2, 3, 4, 7] to derive SSC or weak convergence for other various queueing systems.

2. MODEL SETTING

We consider a JSQ system with J parallel stations, indexed by $j \in \mathcal{J} \equiv \{1, \ldots, J\}$. For each station $j \in \mathcal{J}$, there is an i.i.d. sequence of random variables $\{T_{s,j}(i), i \in \mathbb{N}\}$ and a real number $\mu_j > 0$. For the arrival source, there are an i.i.d. sequence of random variables $\{T_e(i), i \in \mathbb{N}\}$ and a real number $\alpha > 0$. All of the above are defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We assume such $J+1$ sequences

$$\{T_e(i), i \in \mathbb{N}\}, \quad \{T_{s,j}(i), i \in \mathbb{N}\}_{j \in \mathcal{J}} \tag{1}$$

are independent and unitized, that is, $\mathbb{E}[T_e(1)] = 1$ and $\mathbb{E}[T_{s,j}(1)] = 1$ for all $j \in \mathcal{J}$. For each $i \in \mathbb{N}$, $T_e(i)/\alpha$ denotes the interarrival time between the ith and $(i+1)$th arriving jobs, and $T_{s,j}(i)/\mu_j$ stands for the service time of the ith job at station j. Accordingly, α is the arrival rate, and μ_j is the service rate at station j. We assume the following moment condition on interarrival and service time distributions.

Assumption 1. *We assume the interarrival and service times have finite $2+\delta_0$ moments for some $\delta_0 > 0$. Specifically, for $\delta_0 > 0$, we assume*

$$\mathbb{E}\left[T_e^{2+\delta_0}(1)\right] < \infty, \text{ and } \mathbb{E}\left[T_{s,j}^{2+\delta_0}(1)\right] < \infty \text{ for } j \in \mathcal{J}. \tag{2}$$

The routing decisions adopt the JSQ policy, which assigns the arriving job to the station with the shortest queue length. In the case of a tie, the job is assigned to the station with the smallest index. We use $u(z)$ to represent the routing decision when a job arrives and observes $z = (z_1, \ldots, z_J)$ jobs in the system, where z_j is the queue length at station j,

including possibly the one in service. Specifically, the job is routed to station j if $u(z) = e^{(j)}$, where $e^{(j)}$ denotes a J-dimensional unit vector where the jth element is 1 and all other elements are 0.

A JSQ system can be modeled as a Markov process as follows. For time $t \geq 0$, we denote by $Z_j(t)$ the queue length at station j. Let $R_e(t)$ be the residual time until the next arrival to the system, and $R_{s,j}(t)$ be the residual service time for the job being processed at station j if $Z_j(t) > 0$ or the service time of the next job to be processed at station j if $Z_j(t) = 0$. We write $Z(t)$ and $R_s(t)$ for J-dimensional random vectors whose jth element are $Z_j(t)$ and $R_{s,j}(t)$, respectively. For any $t \geq 0$, we set

$$X(t) = (Z(t), R_e(t), R_s(t)).$$

Then $\{X(t), t \geq 0\}$ is a Markov process with respect to the filtration $\mathbb{F}^X = \{\mathcal{F}_t^X, t \geq 0\}$ defined on the state space $\mathbb{S} = \mathbb{Z}_+^J \times \mathbb{R}_+ \times \mathbb{R}_+^J$, where $\mathcal{F}_t^X = \sigma(\{X(s), 0 \leq s \leq t\})$. We assume that each sample path of the process $\{X(t), t \geq 0\}$ is right-continuous and has left limits.

To carry out the heavy traffic analysis, we consider a sequence of JSQ systems indexed by $r \in (0, 1)$. To keep the presentation clean, we set the arrival rate as the only parameter dependent on r and denote by $\alpha^{(r)}$ the arrival rate for the rth system. All other parameters are assumed to be independent of r, including the service rates $\{\mu_j\}_{j\in\mathcal{J}}$, unitized interarrival and service times specified in (1). We parameterize r as

$$r = \sum_{j\in\mathcal{J}} \mu_j - \alpha^{(r)},$$

under which the traffic intensity $\rho^{(r)} \equiv \alpha^{(r)} / \sum_{j\in\mathcal{J}} \mu_j \to 1$ as $r \to 0$, that is, the system is in heavy traffic. We then denote by $\{X^{(r)}(t), t \geq 0\}$ the corresponding Markov process in the rth system. Our result is based on the steady-state behavior. This motivates us to make the following assumption. Under some mild distributional assumptions on interarrival times, the following assumption holds [1].

ASSUMPTION 2. *For each $r \in (0, 1)$, the Markov process $\{X^{(r)}(t), t \geq 0\}$ is positive Harris recurrent and has a unique stationary distribution $\pi^{(r)}$.*

For $r \in (0, 1)$, we denote by

$$X^{(r)} = (Z^{(r)}, R_e^{(r)}, R_s^{(r)})$$

the random vector that follows the stationary distribution. To simplify the notation, we use $\mathbb{E}_\pi[\cdot]$ (rather than $\mathbb{E}_{\pi^{(r)}}[\cdot]$) to denote expectation concerning the stationary distribution when the index r is clear from the context.

3. MAIN RESULTS

In this section, we demonstrate that the vector of the scaled steady-state queue length $rZ^{(r)}$ weakly converges to a vector whose elements are the same exponential random variable Z^* in heavy traffic.

THEOREM 1. *Suppose Assumptions 1 and 2 hold. As $r \to 0$, we have*

$$\left(rZ_1^{(r)}, \ldots, rZ_J^{(r)}\right) \Rightarrow (Z^*, \ldots, Z^*),$$

where Z^* is an exponential random variable with mean

$$m = \frac{1}{2J} \sum_{j\in\mathcal{J}} \mu_j \left(c_e^2 + c_{s,j}^2\right). \tag{3}$$

Here, c_e^2 is the squared coefficient of variation (SCV) of the interarrival time, and $c_{s,j}^2$ is the SCV of the service time at station j.

We recall that for a positive random variable U, its SCV, denoted as $c^2(U)$, is defined to be $c^2(U) = \mathrm{var}(U)/(\mathbb{E}[U])^2$.

To prove Theorem 1, we establish the SSC and weak convergence of the scaled average queue length as follows.

PROPOSITION 2 (STATE-SPACE COLLAPSE). *Suppose Assumptions 1 and 2 hold. The difference between the queue length and the average queue length is uniformly bounded in heavy traffic, i.e.,*

$$\sup_{r\in(0,\mu_{\min}/2)} \mathbb{E}\left[\max_{j\in\mathcal{J}} \left|Z_j^{(r)} - \bar{Z}^{(r)}\right|^{1+\delta_0/(1+\delta_0)}\right] < \infty, \tag{4}$$

where $\bar{Z}^{(r)} = \sum_{j\in\mathcal{J}} Z_j^{(r)}/J$ and $\mu_{\min} = \min_{j\in\mathcal{J}} \mu_j$.

REMARK 1. *Proposition 2 is enough to support Theorem 1. Furthermore, if the moment condition in Assumption 1 is strengthened to $M + \delta_0$, this SSC result can be similarly extended to $M + \delta_0/(M + \delta_0)$, as discussed in [5].*

PROPOSITION 3. *Suppose Assumptions 1 and 2 hold. As $r \to 0$, we have*

$$r\bar{Z}^{(r)} \Rightarrow Z^*,$$

where Z^* is an exponential random variable defined in (3).

Proposition 2 and Markov's inequality imply that for any station $j \in \mathcal{J}$, $rZ_j^{(r)} - r\bar{Z}^{(r)}$ converges to 0 in probability. Theorem 1 is, hence, a direct consequence of Proposition 3.

The proofs of Propositions 2 and 3 utilize the BAR, which will be introduced in Section 4. The proof sketch for designing and applying test functions to the BAR is outlined in Section 5, with a comprehensive version in [5].

4. BASIC ADJOINT RELATIONSHIP

In this section, we introduce the BAR of the JSQ system for our analysis, which enables us to characterize the stationary distribution of the JSQ system directly.

To characterize the jumps of states resulting from arrivals and service completions, we employ the Palm measure proposed in [3]. The Palm measure for external arrivals is represented by $\mathbb{P}_e$ and for service completions at station $j \in \mathcal{J}$ by $\mathbb{P}_{s,j}$. The following lemma characterizes the relationship between the pre-jump and post-jump states under the Palm measures, and its proof follows from Lemma 6.3 in [3].

LEMMA 4. *The pre-jump state X_- and the post-jump state X_+ have the following representation,*

$$X_+ = X_- + \sum_{j\in\mathcal{J}} \left(e^{(j)}, T_e/\alpha, 0\right) \mathbb{1}\left(u(Z_-) = e^{(j)}\right), \text{ under } \mathbb{P}_e,$$

$$X_+ = X_- + \left(-e^{(j)}, 0, e^{(j)} T_{s,j}/\mu_j\right), \quad \text{under } \mathbb{P}_{s,j}, j \in \mathcal{J},$$

where $T_e, T_{s,j}$ for $j \in \mathcal{J}$ are random variables defined on the measurable space $(\mathbb{S}^2, \mathscr{B}(\mathbb{S}^2))$, such that, under Palm distribution $\mathbb{P}_e$, T_e is independent of X_- and has the same distribution as that of $T_e(1)$ on $(\Omega, \mathscr{F}, \mathbb{P})$, and, under Palm distribution $\mathbb{P}_{s,j}$, $T_{s,j}$ is independent of X_- and has the same distribution as that of $T_{s,j}(1)$ on $(\Omega, \mathscr{F}, \mathbb{P})$.

Let $\mathcal{D}$ be the set of bounded function $f : \mathbb{S} \to \mathbb{R}$ satisfying the following conditions: for any $z \in \mathbb{Z}_+^J$, (a) the function $f(z, \cdot, \cdot) : \mathbb{R}_+ \times \mathbb{R}_+^J \to \mathbb{R}$ is continuously differentiable at all but finitely many points; (b) the derivatives of $f(z, \cdot, \cdot)$ in each dimension have a uniform bound over z.

For a JSQ system with a Markov process $\{X(t), t \geq 0\}$ and steady-state vector X defined in Section 2, we obtain the BAR as follows: for any $f \in \mathcal{D}$,

$$\begin{aligned}\mathbb{E}_\pi\left[\mathcal{A}f(X)\right] + \alpha\mathbb{E}_e\left[f(X_+) - f(X_-)\right] \\ + \sum_{j\in\mathcal{J}} \lambda_j \mathbb{E}_{s,j}\left[f(X_+) - f(X_-)\right] = 0, \end{aligned} \tag{5}$$

where $\lambda_j = \mu_j \mathbb{P}(Z_j > 0)$ is the departure rate at station j with the property $\sum_{j\in\mathcal{J}} \lambda_j = \alpha$ by conservation of flow, and the terms on the right-hand side of (5) correspond to state changes by jumps resulting from arrival and service completion, respectively; $\mathcal{A}$ is the "interior operator" defined as

$$\mathcal{A}f(x) = -\frac{\partial f(x)}{\partial r_e} - \sum_{j\in\mathcal{J}} \frac{\partial f(x)}{\partial r_{s,j}} \mathbb{1}(z_j > 0),\ x = (z, r_e, r_s) \in \mathbb{S},$$

which characterizes the system evolution between jumps. The derivation of the BAR (5) follows from Section 6 of [3].

5. SKETCH OF PROOF

In this section, we present the proof sketch for Propositions 2 and 3 using the BAR approach. The detailed proof is provided in [5]. To prove Proposition 2, we utilize the mathematical induction following the idea from [7] and the BAR in (5) with test functions inspired by [6, 7].

We first denote the components of the vector z parallel and perpendicular to $e \equiv (1, \ldots, 1)$ by

$$z_\parallel = \frac{\langle z, e\rangle}{\|e\|^2} e = \bar{z}e, \quad z_\perp = z - z_\parallel = (z_j - \bar{z})_{j\in\mathcal{J}},$$

where $\bar{z} = \sum_{j\in\mathcal{J}} z_j/J$ and the norm is Euclidean norm. To prove (4), it suffices to show that $\mathbb{E}_\pi[\|Z_\perp\|^{1+\delta_0/(1+\delta_0)}]$ is uniformly bounded for all $r \in (0, \mu_{\min}/2)$. Here, we present a prove sketch for the integer moment bound of $\|Z_\perp\|^M$ under the finite $(M+1)$th moment in Assumption 1 and then extend it to the non-integer case in [5].

Our statements include moment bounds for $\|Z_\perp^{(r)}\|$ and some auxiliary results. For each integer $n = 0, \ldots, M$, there exist positive and finite constants C_n, D_n, E_n, F_n that are independent of r such that the following statements hold for all $r \in (0, \mu_{\min}/2)$:

(S1) $\mathbb{E}_\pi[\|Z_\perp^{(r)}\|^n] \leq C_n$.

(S2) $\mathbb{E}_e[\|Z_{-,\perp}^{(r)}\|^n] + \sum_{\ell\in\mathcal{J}} \mathbb{E}_{s,\ell}[\|Z_{-,\perp}^{(r)}\|^n] \leq D_n$.

(S3) $\mathbb{E}_\pi[\|Z_\perp^{(r)}\|^n \psi_{M-n}(R_e^{(r)}, R_s^{(r)})] \leq E_n$.

(S4) $\mathbb{E}_e[\|Z_{-,\perp}^{(r)}\|^n \psi_{M-n}(R_{-,e}^{(r)}, R_{-,s}^{(r)})]$
$+ \sum_{\ell\in\mathcal{J}} \mathbb{E}_{s,\ell}[\|Z_{-,\perp}^{(r)}\|^n \psi_{M-n}(R_{-,e}^{(r)}, R_{-,s}^{(r)})] \leq F_n$,

where $\psi_n(r_e, r_s) = r_e^n + \sum_{j\in\mathcal{J}} r_{s,j}^n$.

The function ψ_{M-n} appearing in the auxiliary statements (S3)-(S4) depends on the moment condition of order $M+1$. This design of the auxiliary statements plays a crucial role in reducing the moment condition required for establishing the uniform bounds on $\mathbb{E}_\pi[\|Z_\perp^{(r)}\|^M]$.

For the induction step of the mathematical induction, we verify (S1)-(S4) for each given n, under the induction hypotheses that they are true for all $k = 0, \ldots, n-1$. To prove the above statements, we employ the BAR (5) with test functions as follows:

$$\begin{aligned} f_n(x) &= \frac{1}{n+1}\|z_\perp\|^{n+1} - z_\perp' u(z) \cdot \alpha^{(r)} r_e \cdot \|z_\perp\|^{n-1} \\ &\quad + z_\perp' (\mu \circ r_s) \cdot \|z_\perp\|^{n-1}, \\ f_{n,D}(x) &= \|z_\perp\|^n \psi_1(r_e, r_s), \\ f_{n,E}(x) &= \|z_\perp\|^n \psi_{M-n+1}(r_e, r_s), \\ f_{n,F}(x) &= \|z_\perp\|^n \psi_{M-n}(r_e, r_s)\psi_1(r_e, r_s), \end{aligned}$$

where $\circ$ is the element-wise product.

To prove Proposition 3, we utilize the following exponential test function in [3] to construct the moment generating function (MGF) of steady-state total queue length. For $\theta \leq 0$, we define

$$f_\theta(x) = \exp\left(\theta \sum_{j\in\mathcal{J}} z_j\right) \exp\left(-\alpha^{(r)}\eta(\theta) r_e - \sum_{j\in\mathcal{J}} \mu_j \xi_j(\theta) r_{s,j}\right),$$

where $\eta(\theta)$ and $\xi(\theta)$ satisfy some equations. With such a design, only the first term in BAR (5) regarding π will be kept, and the jump terms become 0. After setting θ to $r\theta$ and utilizing the Taylor expansions for $\eta(\theta)$ and $\xi_j(\theta)$, we prove the limit MGF of the scaled total queue length has the format of an exponential distribution with mean Jm:

$$\lim_{r\downarrow 0} \mathbb{E}_\pi\left[\exp\left(r\theta \sum_{j\in\mathcal{J}} Z_j^{(r)}\right)\right] = \frac{1}{1 - \theta Jm},$$

where m is defined in (3). The detailed proof is given in [5].

6. REFERENCES

[1] M. Bramson. Stability of join the shortest queue networks. *The Annals of Applied Probability*, 21(4):1568–1625, 2011.

[2] A. Braverman, J. G. Dai, and M. Miyazawa. Heavy traffic approximation for the stationary distribution of a generalized Jackson network: the BAR approach. *Stochastic Systems*, 7(1):143–196, May 2017.

[3] A. Braverman, J. G. Dai, and M. Miyazawa. The BAR approach for multiclass queueing networks with SBP service policies. *arXiv preprint arXiv:2302.05791*, 2023.

[4] J. G. Dai, P. Glynn, and Y. Xu. Asymptotic product-form steady-state for generalized Jackson networks in multi-scale heavy traffic. *arXiv preprint arXiv:2304.01499*, 2023.

[5] J. G. Dai, J. Guang, and Y. Xu. Steady-state convergence of the continuous-time JSQ system with general distributions in heavy traffic. *arXiv preprint arXiv:2405.10876*, 2024.

[6] A. Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72(3-4):311–359, 2012.

[7] J. Guang, X. Chen, and J. G. Dai. Uniform moment bounds for generalized Jackson networks in multi-scale heavy traffic. *arXiv preprint arXiv:2401.14647*, 2024.

[8] D. Hurtado-Lange and S. T. Maguluri. Transform methods for heavy-traffic analysis. *Stochastic Systems*, 10(4):275–309, 2020.

[9] X. Zhou and N. Shroff. A note on Stein's method for heavy-traffic analysis. *arXiv preprint arXiv:2003.06454*, 2020.