Queueing Theory Final Project

Chia-Hao Chang*

Dec 16, 2019

*UT EID:cc66887, email: chchangkh@utexas.edu

1 Model Construction and Definitions

We consider a system with two servers. One of the servers provides higher payoff v_+ than the other one, which provides payoff $v_- < v_+$. Customers are comprised of three genres. Each type of customers arrive according to a Poisson process, independent of one another. The first type of customers whom we call *informed* customers arrive at rate λ_i and is completely aware of whichever server provides a better payoff. In other words, informed customers always join the server with payoff v_+. The second type of customers whom we call *fake* customers arrive at rate λ_f and always join the the server with lower payoff, i.e., v_-. The third type of customers whom we call *regular* customers arrive at rate λ_r and is unaware as to which server provides better payoff, i.e., a regular customer does not know which server gives payoff v_+ and which gives v_-. More precisely, we assume that the customer does not know *a-priori* which server is better; mathematically, this is equivalent to saying that the regular customers have a uniform prior (i.e., a $(1/2, 1/2)$ prior) about which server is better.

As a result, a regular customer decides which queue to join based on his observance of the system state upon his arrival. Throughout this paper, we refer to system state as the (joint) queue length of the system. The low-quality server, i.e., the server providing v_-, however, could fake a prosperous queue in front his server by adjusting the arrival rate λ_f at a cost rate $c(\lambda_f)$ per unit time, where $c(\cdot)$ is assumed to be a strictly increasing functions. Realizing the fact that there are *informed*, *fake*, and *regular* customers, a regular customer, upon arriving at the system, decides which queue to join by maximizing his (subjective) expected payoff.

Notice that the described dynamics induces a game between regular customers and the low-quality server. We begin our discussion by a set of relevant definitions.

The state space is given by $S = \mathbb{Z}_+^2 = \{(x, y) \colon x, y \geq 0 \text{ and } x, y \in \mathbb{Z}\}$. We denote a generic state in S by (n_a, n_b), where n_a is the queue length of the server a and n_b is the queue length in front of the server b. Without loss generality, we assume throughout this paper the server a provides v_+ while server b provides v_-. Again be aware that the regular customer is indifferent between these two servers and all the information at hand is merely the system state (i.e. queue length.) In other words, an arrival of an informed customer and of a fake customer would increase n_a and n_b by 1, respectively. Note, however, a regular customer can not distinguish between the servers, and, as a result, he should treat states (x, y) and (y, x) indifferently. This motivates our following definition.

Definition 1 (Customer Strategy)**.** A *customer strategy* σ is a mapping $S \to \Delta(\{1, -1\})$ such that

$$\sigma((x, y))[a] = \sigma((y, x))[b], \text{ for all } x, y \in \mathbb{Z}_+,$$

where $\Delta(\{1, -1\})$ is the probability measure over $\{-1, 1\}$; $\sigma((x, y))[a]$ is the probability of joining server a.

In other words, $\sigma((n_a, n_b))[a]$ (resp. $\sigma((n_a, n_b))[b]$) is the probability that when a regular customer sees queue length distribution (n_a, n_b) he chooses to join the right queue (resp. the wrong queue.) Also, we assume that each customer makes his own decision, independent of any other people.

When a service is complete, the service provider receiver receives p dollars from the served customer. The (average) profit received by the low-quality server during $[0, T]$, when he chooses λ_f as the fake customer arrival rate, is therefore given by

$$\frac{1}{T}\sum_{k=1}^{n(T)} p\mathbb{I}\{\text{the } k\text{th customer goes to low-quality server}\} - \frac{1}{T}\int_0^T c(\lambda_f)dt,$$

where $n(T)$ is the total number of arrival of regular customers during $[0, T]$. We do not allow dynamic adjustment of the λ_f. Therefore, when $T \to \infty$, we have

$$\begin{aligned}
&\lim_{T\to\infty} \frac{1}{T}\sum_{k=1}^{n(T)} p\mathbb{I}\{\text{the } k\text{th customer goes to low-quality server}\} - \frac{1}{T}\int_0^T c(\lambda_f)dt \\
=&\lim_{T\to\infty} \frac{n(T)}{T}\frac{1}{n(T)}\sum_{k=1}^{n(T)} p\mathbb{I}\{\text{the } k\text{th customer goes to low-quality server}\} - c(\lambda_f) \\
=&\lambda_r p\mathbb{E}[R_r] - c(\lambda_f).
\end{aligned}$$

where R_r is the ratio of regular customers that joins the low-quality server.

Given a customer strategy $\sigma(\cdot)$, a Markov chain over $\mathbb{Z}_+^2$ is induced. As a result, $\mathbb{E}[R_r]$ can be obtained. To stress the role of customer strategy σ, we use the notation $\mathbb{E}_\sigma[R_r]$ when the customer strategy is given by σ.

We let $\pi_\sigma(\mathbf{n}) = \pi_\sigma(n_a, n_b)$ be the stationary distribution of the (two dimensional) Markov chain induced by customer strategy σ. We suppress λ_f for its effect is (implicitly) incorporated into the strategy σ.

1.1 Analysis of the perfect Bayesian Equilibrium: Customer Equilibrium

In this section, we analyze the equilibrium among the the customers when λ_i, λ_r, and λ_f are known to them.

To begin with, given a customer strategy σ, denote $\pi_{\sigma,a}(\mathbf{n})$ (resp. $\pi_{\sigma,b}(\mathbf{n})$) the stationary distribution of state $\mathbf{n}$ of the two-dimensional CTMC induced by the customer strategy σ, conditional on the first server is the good (resp. bad) server. Notice the symmetry in our definition of $\pi_{\sigma,a}(\mathbf{n})$

and $\pi_{\sigma,b}(\mathbf{n})$. It should not hard to find that due to our definition $\pi_{\sigma,a}(\mathbf{n})$ should satisfy

$$\pi_{\sigma,b}(\mathbf{n}) = \pi_{\sigma,a}(\mathbf{n}^\top),$$

where we have abused the notation $(\cdot)^\top$ by defining $\mathbf{n}^\top = (n_a, n_b)$ if $\mathbf{n} = (n_b, n_a)$. For example, the probability in state (2,3) conditional on the belief that the first server is good should be identical to the probability in state (3,2) conditional on the belief that the second server is good.

We assume all the customers are *ex ante* identical, risk-neutral, rational in the sense that they act to maximize their expected utility, and Bayesian. As a result, upon seeing a state $\mathbf{n}$, the customer updates his posterior belief by Bayesian rule, which gives

$$P_\sigma(\text{server } a \text{ is good}|\mathbf{n}) = \frac{\pi_{\sigma,a}(\mathbf{n})}{\pi_{\sigma,a}(\mathbf{n}) + \pi_{\sigma,b}(\mathbf{n})},\ P_\sigma(\text{server } b \text{ is good}|\mathbf{n}) = \frac{\pi_{\sigma,b}(\mathbf{n})}{\pi_{\sigma,a}(\mathbf{n}) + \pi_{\sigma,b}(\mathbf{n})},$$

where the subscript σ emphasizes the posterior probabilities are induced by σ.

Now, if the customer upon seeing $\mathbf{n}$ decides to choose "+" server with probability q then his (subjective) expected utility is given by

$$\mathbb{E}_\sigma[V|\mathbf{n}](q) = q\left(\frac{v_+\pi_{\sigma,a}(\mathbf{n})}{\pi_{\sigma,a}(\mathbf{n}) + \pi_{\sigma,b}(\mathbf{n})} + \frac{v_-\pi_{\sigma,b}(\mathbf{n})}{\pi_{\sigma,a}(\mathbf{n}) + \pi_{\sigma,b}(\mathbf{n})}\right) + (1-q)\left(\frac{v_-\pi_{\sigma,a}(\mathbf{n})}{\pi_{\sigma,a}(\mathbf{n}) + \pi_{\sigma,b}(\mathbf{n})} + \frac{v_+\pi_{\sigma,b}(\mathbf{n})}{\pi_{\sigma,a}(\mathbf{n}) + \pi_{\sigma,b}(\mathbf{n})}\right).$$

Definition 2 (Best Response)**.** The *best response* $\mathrm{BR}(\mathbf{n}|\sigma)$ at state $\mathbf{n}$ when the customer strategy is given by σ is the collection of strategies that maximizes the expected utility of the customer. Formally,

$$\mathrm{BR}(\mathbf{n}|\sigma) = \{q\colon \mathbf{E}_\sigma[V|\mathbf{n}](q) \geq \mathbf{E}_\sigma[V|\mathbf{n}](q'),\ \text{for all } q' \in [0,1]\}.$$

Clearly, when the customer decides to join server a (resp. join server b), $\mathrm{BR}(\mathbf{n}|\sigma)$ degenerates to 1 (resp. 0.)

This expression immediately motivates our following lemma,

Lemma 1. *The best response for customers in state* $\mathbf{n}$ *is to join server* a *if* $\pi_{\sigma,a}(\mathbf{n}) > \pi_{\sigma,b}(\mathbf{n})$*. The best response for customers in state* $\mathbf{n}$ *is to join server* b *if* $\pi_{\sigma,b}(\mathbf{n}) > \pi_{\sigma,a}(\mathbf{n})$*. When* $\pi_{\sigma,a}(\mathbf{n}) = \pi_{\sigma,b}(\mathbf{n})$*, the customer is indifferent between server* a *and* b*, and any randomized strategy is a best response.*

The proof is rather simple. One could notice that $\mathbb{E}_\sigma[V|\mathbf{n}]$ is maximized at $q = 1$ when $\pi_{\sigma,a}(\mathbf{n}) > \pi_{\sigma,b}(\mathbf{n})$, and is maximized at $q = 0$ when otherwise.

Now, we can finally define our notion of equilibrium already.

Definition 3 (Customer Equilibrium)**.** Given λ_i, λ_r, λ_f, a customer strategy σ is a *customer equilibrium* if it satisfies

$$\sigma(\mathbf{n}) \in \mathrm{BR}(\mathbf{n}|\sigma),$$

for all $\mathbf{n} \in \mathbb{Z}_+^2$.

That is, a customer strategy is equilibrium if it is a maximizer of the expected utility induced by itself.

1.2 Possible Customer Strategies

In this section, we discuss the possibilities of different customer strategies.

Let us start with something simple.

Conjecture 1. *Let λ_i, λ_f and λ_r be given. If $\lambda_i + \lambda_r < \lambda_f < \mu$, then the in customer equilibrium, the customer strategy is to always join the shorter queue.*

The assumption $\lambda_i + \lambda_r < \lambda_f < \mu$ captures the idea that merely a single server could serve all regular and informed customers. Further, under this assumption, even if all the regular customers are attracted to server a its effective arrival is still smaller than the fake server.

Under joining the shorter queue strategy, one could easily observe that the induced Markov chain is positive recurrent, which implies a stationary distribution exists and is unique. In the following context, we use $\pi(x, y)$ to denote $\pi_{\sigma,a}(x, y)$ for the customer strategy σ is clear from the context.

To show that joining the shorter queue is a costumer equilibrium, we have to show that the for all $x < y$

$$\pi(x, y) > \pi(y, x).$$

The global balance equation for (0,0) is given by

$$\lambda_T \pi(0, 0) = \mu\pi(1, 0) + \mu\pi(1, 0),$$

where $\lambda_T \equiv \lambda_r + \lambda_i + \lambda_f$. In particular for states $(1, 0)$ and $(0, 1)$ we have

$$(\mu + \lambda_T)\pi(0, 1) = \mu\pi(1, 1) + \mu\pi(0, 2) + (\lambda_f + \frac{1}{2}\lambda_r)\pi(0, 0),$$
$$(\mu + \lambda_T)\pi(1, 0) = \mu\pi(1, 1) + \mu\pi(2, 0) + (\lambda_i + \frac{1}{2}\lambda_r)\pi(0, 0).$$

For the rest of states on the x and y axis, i.e., states $(x,0)$ or $(0,x)$, $x \geq 2$, the global balance equations are

$$(\mu+\lambda_T)\pi(0,x) = \mu\pi(1,x) + \mu\pi(0,x+1) + (\lambda_f+\lambda_r)\pi(0,x-1),$$
$$(\mu+\lambda_T)\pi(x,0) = \mu\pi(x,1) + \mu\pi(x+1,0) + (\lambda_i+\lambda_r)\pi(x-1,0).$$

For general states (x,y), the global balance equations can be divided into:

(I) $x,y \geq 1, x-y \geq 2$

$$(2\mu+\lambda_T)\pi(x,y) = \mu\pi(x+1,y) + \mu\pi(x,y+1) + (\lambda_i+\lambda_r)\pi(x-1,y) + \lambda_f\pi(x,y-1);$$

(II) $x,y \geq 1$, $x-y=1$

$$(2\mu+\lambda_T)\pi(x,y) = \mu\pi(x+1,y) + \mu\pi(x,y+1) + (\lambda_i+\frac{1}{2}\lambda_r)\pi(x-1,y) + \lambda_f\pi(x,y-1);$$

(III) $x,y \geq 1, y-x \geq 2$

$$(2\mu+\lambda_T)\pi(x,y) = \mu\pi(x+1,y) + \mu\pi(x,y+1) + (\lambda_f+\lambda_r)\pi(x,y-1) + \lambda_i\pi(x-1,y);$$

(IV) $x,y \geq 1, y-x=1$

$$(2\mu+\lambda_T)\pi(x,y) = \mu\pi(x+1,y) + \mu\pi(x,y+1) + (\lambda_f+\frac{1}{2}\lambda_r)\pi(x,y-1) + \lambda_i\pi(x-1,y);$$

(V) $x=y\geq 1$

$$(2\mu+\lambda_T)\pi(x,y) = \mu\pi(x+1,y) + \mu\pi(x,y+1) + \lambda_f\pi(x,y-1) + \lambda_i\pi(x-1,y).$$

Another intuitive scenario is where the fake customers are absent, i.e., $\lambda_f = 0$.

Conjecture 2. *Let λ_i, $\lambda_r > 0$ and $\lambda_f = 0$. If $\lambda_i + \lambda_r < \mu$, then the customer strategy is to always join the longer queue when the queue lengths are not equal, and randomize with probability $\frac{1}{2}$ to join each queue.*

Under joining the longer queue strategy, the resulting stationary distribution is described by the global balance equations:

$$(\lambda_r+\lambda_i)\pi(0,0) = \mu\pi(1,0) + \mu\pi(0,1), \text{for state } (0,0),$$

whereas for states $(x,0)$ and $(0,x)$, the global balance equations are given by

$$(\lambda_T+\mu)\pi(x,0) = \lambda_T\pi(x-1,0) + \mu\pi(x+1,0) + \mu\pi(x,1),$$
$$(\lambda_T+\mu)\pi(0,x) = \lambda_r\pi(0,x-1) + \mu\pi(1,x) + \mu\pi(0,x+1), \text{ for } x \geq 2,$$

where $\lambda_T = \lambda_r + \lambda_i$ because $\lambda_f = 0$. For states $(1,0)$ and $(0,1)$ we have

$$(\lambda_T + \mu)\pi(1,0) = \mu\pi(2,0) + \mu\pi(1,1) + (\lambda_i + \frac{1}{2}\lambda_r)\pi(0,0),$$
$$(\lambda_T + \mu)\pi(0,1) = \mu\pi(0,2) + \mu\pi(1,1) + \frac{1}{2}\lambda_r\pi(0,0).$$

For $x - y \geq 2$, $x, y \geq 1$,

$$(\lambda_T + 2\mu)\pi(x,y) = \lambda_T\pi(x-1,y) + \mu\pi(x+1,y) + \mu\pi(x,y+1).$$

For $x - y = 1$, $x, y \geq 1$,

$$(\lambda_T + 2\mu)\pi(x,y) = (\lambda_i + \frac{1}{2}\lambda_r)\pi(x-1,y) + \mu\pi(x+1,y) + \mu\pi(x,y+1).$$

For $y - x \geq 2$, $x, y \geq 1$,

$$(\lambda_T + 2\mu)\pi(x,y) = \lambda_r\pi(x-1,y) + \lambda_i\pi(x,y-1) + \mu\pi(x+1,y) + \mu\pi(x,y+1).$$

For $y - x = 1$, $x, y \geq 1$,

$$(\lambda_T + 2\mu)\pi(x,y) = \frac{1}{2}\lambda_r\pi(x,y-1) + \lambda_i\pi(x-1,y) + \mu\pi(x+1,y) + \mu\pi(x,y+1).$$

The scenario described in conjecture 2 in fact admits a rather simple customer equilibrium which we describe as follows.

Proposition 1. *Let $\lambda_i, \lambda_r > 0$ and $\lambda_f = 0$. If $\lambda_i + \lambda_r < \mu$, then the customer strategy: "joining the longer queue when the queue lengths are not equal; join the first queue when the queue lengths are the same" is a customer equilibrium.*

Proof of Proposition 1. The proof is straightforward. We verify that the stationary distribution induced by this customer strategy admits this strategy, i.e., $\pi(x,y) > \pi(y,x)$ whenever $x > y$. To this end, notice that only states $(x,0)$ will be visited in equilibrium. This is because under this strategy, state $(0,0)$ always moves to $(1,0)$, which in turn admits only transitions from $(1,0)$ to $(2,0)$ (with rate λ_T) and from $(1,0)$ to $(0,0)$ (with rate μ). Following a similar induction process, we know that the only recurrent states are $\{(x,0)\colon x \geq 0, x \in \mathbb{Z}\}$. In fact, the stationary distribution is nothing more than that given by the M/M/1 queue formula: $\pi(x,0) = (1-\rho)\rho^x$, where $\rho = \frac{\lambda_T}{\mu}$; all other states have stationary probability 0. Hence, $\pi(x,0) > \pi(0,x) = 0$, which verifies that this strategy is indeed a customer equilibrium strategy. □

1.3 Related Literature on Join-the-Shortest-Queue

Our current methodology is similar to the famous join-the-shortest-queue (JSQ) problem in queueing theory. For the sake of exposition, we review and summarize a number of crucial works in the analysis of the JSQ systems. In particular, I surveyed a few papers which work on obtaining analytical solutions for the *asymptotic behavior* of JSQ systems.

1.3.1 The first paper: [2]

In [2], the authors consider a JSQ system where customers arrive according to Poisson process with rate λ and joins the shorter queue if the queue length is unequal and randomizes (with probability $\frac{1}{2}$) between two queues if the queue lengths are the same. The service rates of the servers are denoted by μ_1 and μ_2, where $\mu_1 + \mu_2 > \lambda$.

The authors first consider the systems with *symmetric servers* i.e., $\mu_1 = \mu_2$. Then they consider the more general scenario where $\mu_1 \neq \mu_2$. In each scenario, the authors implemented the *asymptotic analysis* of the JSQ system. In each scenario, the authors divide the analysis into three regimes: $n_1 = n_2 \to \infty$, $n_1 \to \infty$ while $n_2 = O(1)$, and $n_2 \to \infty$ while $n_1 = O(1)$. They obtained the asymptotic analytic solutions for each regime, for both symmetric and asymmetric scenarios.

The idea is that the stationary distributions can be described by a set of global balance equations, which are themselves homogeneous second-order (or higher) difference equations. Because the global balance equations are identical when the states (n_1, n_2) are bounded away from 0, one can then solve lesser balance equations-by neglecting the global balance equations that are around $(0, 0)$.

Finally, they also manifest their results by comparing with a M/M/2 system and a system with two independent M/M/1 queues, all with the same traffic intensity.

1.3.2 The second paper: [1]

In this paper, the authors considered a quite general JSQ system. Specifically, they considered a system with m servers. Denote $M \coloneqq \{1, 2, \ldots, m\}$ the collection of servers. They allow for each $A \subset M$ there is a stream of customers, arriving according to an independent Poisson process with rate $\lambda_A \geq 0$. For such customers, they join the server with the shortest queue length upon his arrival. For example, if $A = \{i\}$, the customer is a dedicated customer who only joins server i; if $A = M$, the customers join the shortest queue among all servers and randomizes uniformly among all equal-length servers. All arrival streams are assumed to be independent Poisson processes and all service times are assumed to be independent and memoryless.

The authors first obtained the stability conditions of this system. In particular, their results can be summarized as follows. Denote $\rho_A \coloneqq \frac{\sum_{B \subset A} \lambda_B}{\mu_A}$ where $\mu_A \coloneqq \sum_{i \in A} \mu_i$. $\rho_{\max} = \max_{A \subset M} \rho_A$; that is, ρ_A is the "average" load of the incoming stream whose servers are in A while $\rho_{\max}$ is the

the heaviest load among all possible incoming streams of customers. Essentially, the stability of the JSQ system can be categorized according to $\rho_{\max}$:

Theorem 1 (Stability Condition in [1]). *1. If $\rho_{\max} > 1$, the system is transient.*

2. If $\rho_{\max} = 1$, the system is either transient or null recurrent.

3. If $\rho_{\max} < 1$, the system positive recurrent. In addition, the average number of customers in the system is bounded by $-m + \mu_M/c$, for some constant c (described in the paper.)

The first two results are quite intuitive. The third result is somewhat surprising due to the generality of the current system: even if the system consists of heterogeneous customers (in the sense the available servers to them are different), join-the-shortest-queue is indeed a robust policy that stabilizes the system-provided that none of the incoming stream of customers have a load greater than 1.

The outline of the proof for statement 3 of theorem 1 is as follows. Essentially, the authors considered a similar system which only admits $\lambda'_{\{i\}} > 0, i \in M$; $\lambda'_A = 0$ if $A \subset M$ is not a singleton. They related $\lambda'_{\{i\}}$ in this fictitious system (which they call α-system) to the arrival rates λ_A in the original system (which they call λ-system).

Next, they consider a Markov decision process (MDP) where the decision maker can decide whether to switch from an α-system to a λ-system (or vice versa). The objective function for the MDP happened to coincide with the average number of customers in the system. Next, the authors show that in the MDP choosing an α-system is always worse than choosing a λ-system. Next, the authors further noted that if the objective function of the MDP is identified as a Lyapunov function, then by using the relation that an α-system is worse than a λ-system it can be shown that the Lyapunov functions satisfies the constraints for positive recurrence (i.e., negative drift on all but finite states.)

The next big results are the asymptotic analysis of the JSQ system. Most of the results focus on the expected value of *first-passage* time, in particular T_ℓ: the first-passage time that the system has ℓ customers, i.e., $||Q(t)||_1 = \ell$. They obtained the asymptotic analytic formula for $\mathbb{E}[T_\ell|\mathbf{Q}(0) = \mathbf{0}]$ when $\ell \to \infty$ and when the system has only two servers i.e., $m = 2$. $\mathbf{Q}(t) \in \mathbb{Z}_+^m$ is the system state at time t.

Specifically, they analyzed the system with only two servers. They found that the results can be categorized according to the relative magnitudes of ρ_1, ρ_2, and $\rho_M = \rho_{\{1,2\}}$ (hereafter we use $\rho \equiv \rho_M$ for shorthand.) They showed that the results can be divided into three regimes.

In the first regime which they call *strongly pooled servers* are such that $\rho > \max\{\rho_1, \rho_2\}$ and $\lambda_M > |\rho_M^2(\mu_2 - \mu_1) + \lambda_1 - \lambda_2|$. In this region, the expected first-passage time is given by

$$\mathbb{E}[T_\ell|\mathbf{Q}(0,0)] \sim g^{-1}\rho^{-1}(\lambda_1 + \lambda_2 + \lambda_M + \mu_1 + \mu_2)^{-1},$$

where g is a constant (defined in the paper.) As for asymptotic stationary distributions, for $k, \ell \in \mathbb{Z}_+$ we have

$$P\{Q_1 + Q_2 = \ell, Q_1 - Q_2 - k\} \sim 2\frac{f(0)}{d}\rho^\ell \varphi(k)\mathbb{I}\{k = \ell \text{mod } 2\},$$

where d is a constant defined in the paper and $f(\cdot)\colon \mathbb{R} \to \mathbb{R}$ is a functions also defined in the paper; $\varphi(\cdot)\colon \mathbb{Z}_+ \to \mathbb{R}$ is a function also defined in the paper. Note hereafter we use $\mathbf{Q} = (Q_1, Q_2)$ to denote the random vector distributed with stationary distribution of the underlying system.

In the second regime which they call *weakly pooled servers* are such that $\rho > \max\{\rho_1, \rho_2\}$ and $\lambda_M \leq |\rho_M^2(\mu_2 - \mu_1) + \lambda_1 - \lambda_2|$. In this region, the first passage time is given by

$$\mathbb{E}[T_\ell|\mathbf{Q}(0,0)] \sim g^{-1}\rho^{-1}(\lambda_1 + \lambda_2 + \lambda_M + \mu_1 + \mu_2)^{-1}.$$

The stationary distribution is given by

$$P\{Q_1 + Q_2 = \ell\} \sim \frac{f(0)}{d}\rho^\ell.$$

While the ratio between Q_1 and Q_2 when $Q_1 + Q_2 = \ell$ can be categorized by:

1. If $\rho^2(\mu_2 - \mu_1) + \lambda_1 - \lambda_2 > \lambda_M$, then

$$\frac{Q_1}{Q_2} = \frac{\lambda_1\rho^{-1} - \mu_1\rho}{(\lambda_2 + \lambda_M)\rho^{-1} - \rho\mu_2}, \quad Q_1 + Q_2 = \ell$$

2. If $\rho^2(\mu_2 - \mu_1) + \lambda_1 - \lambda_2 < -\lambda_M$, then

$$\frac{Q_1}{Q_2} = \frac{(\lambda_1 + \lambda_M)\rho^{-1} - \rho\mu_1}{\lambda_2\rho^{-1} - \mu_2\rho}, \quad Q_1 + Q_2 = \ell$$

3. If $|\rho^2(\mu_2 - \mu_1) + \lambda_1 - \lambda_2| = \lambda_M$, then $Q_1/Q_2 = 1, Q_1 + Q_2 = \ell$

In the last regime which they call *unpooled servers* is such that $\rho < \max\{\rho_1, \rho_2\}$. Assume without loss of generality $\rho_1 > \rho_2$ then in this region we have

$$\mathbb{E}[T_\ell|\mathbf{Q}(0) = (0,0)] = g^{-1}\rho_1^{-\ell}/(\lambda_1 + \lambda_2 + \lambda_M + \mu_1 + \mu_2),$$

where g is the same constant mentioned in the first regime. The stationary distributions are given by

$$P\{Q_1 + Q_2 = \ell, Q_2 = k\} \sim \frac{f}{\mu_1 - \lambda_1}\rho_1^{\ell-k}\left(1 - \frac{\lambda_2 + \lambda_M}{\mu_2}\right)\left(\frac{\lambda_2 + \lambda_M}{\mu_2}\right)^k \tag{1}$$

where f is a constant defined in paper.

We stop here by discussing the relation between [1] and my work. Notice that in conjecture 1, our system is equivalent to that in the third regime of the two-server-JSQ system in [1]. Indeed, notice that by relating server 1 (resp. server 2) to server b (resp. server a) in my model, we can immediately identify λ_i to be λ_2, λ_f to be λ_1, and λ_r to be λ_M. Moreover, we shall notice that $\frac{\lambda_f}{\mu} > \frac{\lambda_i+\lambda_f+\lambda_r}{\mu}$ in our setting, which verifies that we are in the third regime of the two-server-JSQ system in [1].

Now, we plug in the formula in (1) to check if conjecture 1 is true (or at least asymptotically correct.)

Remark 1 (Partial Result for conjecture 1)**.** *In conjecture 1, joining the shorter queue is an asymptotic customer equilibrium.*

Indeed, recall that to show joining the shorter queue is a customer equilibrium it suffices to show that $\pi(n_a, n_b) > \pi(n_b, n_a)$ for all $n_a < n_b$. With the help of (1), we have

$$\pi(n_a, n_b) = P\{Q_1 + Q_2 = n_b + n_a, Q_2 = n_a\} \sim \frac{f}{\mu - \lambda_f}\left(\frac{\lambda_f}{\mu}\right)^{n_b}\left(1 - \frac{\lambda_i + \lambda_r}{\mu}\right)\left(\frac{\lambda_i + \lambda_r}{\mu}\right)^{n_a},$$
$$\pi(n_b, n_a) = P\{Q_1 + Q_2 = n_b + n_a, Q_2 = n_b\} \sim \frac{f}{\mu - \lambda_f}\left(\frac{\lambda_f}{\mu}\right)^{n_b}\left(1 - \frac{\lambda_i + \lambda_r}{\mu}\right)\left(\frac{\lambda_i + \lambda_r}{\mu}\right)^{n_b}.$$

Note, however, by $(\lambda_i + \lambda_r)/\mu < 1$ by our assumption. Therefore we have $\pi(n_a, n_b) \gtrapprox \pi(n_b, n_a)$ for $n_a < n_b$, where we use $\gtrapprox$ to denote *asymptotically greater than*.

2 Approximation Schemes

Some approximation schemes arise in both economics and queueing when the underlying system is notoriously hard to solve. One the commonly used technique is to use *large economics approximation*. In large economics approximations, we use a continuum of players to approximate a system with a large number of player. Normally, such approximation yields a salient result only if the interaction among players is *anonymous* i.e., the payoff (or the structure) of the game is only related to the overall number of players who choose a particular action but not to the player who chooses it. One could be aware that large economics approximation is spiritually similar to fluid approximation in the sense that the effect of individual customer is "washed away" and only the aggregate effect is left.

However, this is reason why large economics approximation or fluid approximation is not appropriate for our current methodology. Notice that in the current model, the effect of *learning* is crucial in our analysis since it prompts the notion of *rationality* thereby *best response*. Due to the

Bayesian nature of our customers, the system state can not be discarded for it is central to updating the posterior belief. Hence, it would be inappropriate to use fluid model as a approximation scheme.

Similarly, in most of the current literature concerning large economics model they can only deal with *static* interactions among the players. However, the *system dynamics*, especially *Bayesian update*, is decisive in our methodology. Consequently, large economics approximation scheme is also improper.

3 Conclusions and Future Work

Although it is believed that the exact analytic solution for JSQ system can not be obtained. I think there is still hope for merely comparing two states in the system. A future direction of work might be analyzing the states around $(0, 0)$ and see if they satisfy conjecture 1 i.e., $\pi(x, y) > \pi(y, x)$ for all $x < y$.

References

[1] R. D. Foley and D. R. McDonald. Join the shortest queue: stability and exact asymptotics. *Ann. Appl. Probab.*, 11(3):569–607, 08 2001.

[2] C. Knessl, B. Matkowsky, Z. Schuss, and C. Tier. Two parallel queues with dynamic routing. *IEEE Transactions on Communications*, 34(12):1170–1175, December 1986.