On Nash Equilibrium of Self-Policy for an Observable Queue

Chia-Li Wang 王家禮
Dong Hwa University東華大學

July 16, 2018@ Chengdu

A Non-Cooperate Game

- In 1994, Chemical Bank in New York offered any customer who had spent 7 or more minutes in queue " a smile, a handshake and a crisp 5 dollar bill ",

the so-called 547 Program of Chemical Bank.

- Forced to jump the cliff together, Bank of America in California immediately announced a better 545 program

Features of This Game

- Arriving customers at the service center are different in service preference
- Each customer is an individual decision maker for self benefit
- Upon arrival, they observe the queue length, then decide to join for service or balk depending on expected personal gain

Corresponding Queueing Model I

- Based on different service preference, customers are classified into k types
- Customers of type i arrive at a rate λ_i and receive service that is independent with their types
- One way to characterize the preference:

  Joining type-i customers gain R_i at service completion, and incur waiting cost at rate C_i while in system; Net gain = $R_i - C_i w$, where w is expected waiting time

  Thus, join if and only if $R_i - C_i w \geq 0$, or, $w \leq R_i/C_i$
- Alternative way:

  Tolerance of waiting for type-i is $\theta_i = R_i/C_i$

Corresponding Queueing Model II

- Arriving customers first observe queue length, then decide to join for service or balk depending on net gain.
- Balking customers have no gain or loss
- Assumed asymmetric information, i.e., system administrator observes only { λ_i, θ_i }
- Thus, objectives of the administrator are to keep the system stable and maximize the throughput rate

General Assumptions on Service

- Customers' commencements of service are stochastically ordered in accordance with arrival times, such as FCFS, ROS, PS
- Departure rate μ_n, when n customers are in system, is non-decreasing and concave in n
- Once a service started, it will not be interrupted until completion, that is, no preemption is allowed nor will the service capacity allocated increases
- For all i, $\mu_1 \leq \theta_i$ to avoid triviality

Threshold-Type Decision Rule

- As a non-cooperate game, customers of different types are competing for some fixed service capacity so that the decision rules (joining or balking) for different types of customers are clearly dependent
- Decision rule of each type of customers is of threshold type, join if queue length is smaller than some n, balk otherwise
- Collection of each type's decision rules is called a policy. Denote a policy by $\mathbf{N} = (N_1, \ldots, N_k)$, where type-$i$ customer will join the system if the number of customers in the system, regardless their types, is strictly less than N_i for any $i = 1,2,\ldots k$

Self-Policy

- Let $w(\boldsymbol{N}, n)$ be expected waiting time of customer who finds n-1 customers present at the joining instant and future customers using policy $\boldsymbol{N}$
- $w(\boldsymbol{N}, n)$ is type invariant owing to type invariant service.
- Let $\boldsymbol{e}_i$ be the i th unit vector in $\boldsymbol{R}_k$. Policy $\boldsymbol{N}$ is said to be a self-threshold if

$$w(\boldsymbol{N}, N_i) \leq \theta_i < w(\boldsymbol{N}+\boldsymbol{e}_i, N_i+1)$$

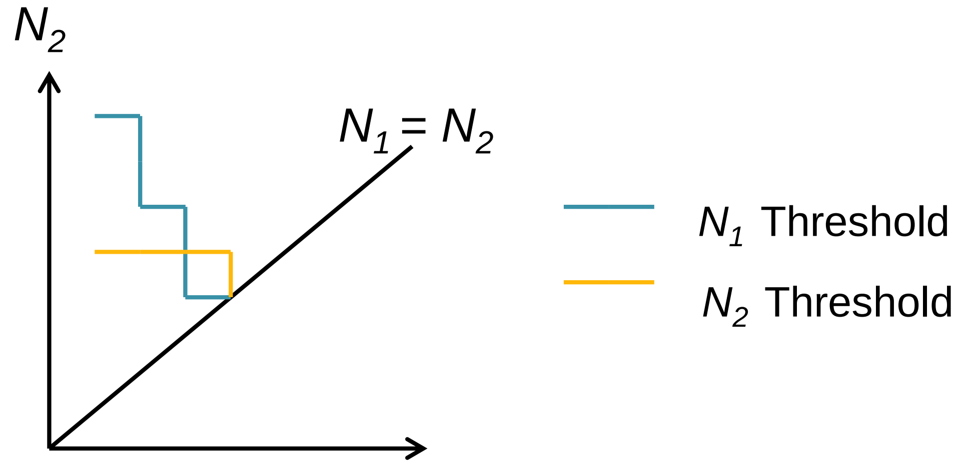
for all i

- We may consider that a self-threshold $\boldsymbol{N}$ is formed by negotiation among customer of all types and announced to every arriving customer

Self-Policy -- Existence

- For $\boldsymbol{N}$ to be self-threshold, order of $\{N_i\}$ must coincide with order of $\{\theta_i\}$. For simple exposition, we will make $\{\theta_i\}$, and, thus, $\{N_i\}$ in increasing order
- **Lemma** The queueing system with heterogeneous customers has at least one self-threshold.

Sketch of Proof : For $k = 2$



Example 1: M/M/1 PS Queue with 2-Types

- Assume $\lambda_i = \lambda_2 = 1$ and $\mu_n = 2$ for all n
- Conditioning on the next transition, we get, for example,

$$w((1, 3),1) = 1/(\lambda_2 + \mu_1) + \lambda_2 /(\lambda_2 + \mu_1)\ w((1, 3), 2)$$

- Plug the numbers in, solve the equations and show for

$\theta_1 \in (0.83, 0.95)$ and $\theta_2 \in (1.11, 1.14)$

$$w((1, 3), 1) = 0.64 \leq \theta_1 < 0.95 = w((2, 3), 2)$$

$$w((1, 3), 3) = 1.11 \leq \theta_2 < 1.43 = w((1, 4), 4)$$

and

$$w((2, 2), 2) = 0.83 \leq \theta_1 < 1.14 < w((3, 3), 3)$$

$$w((2, 2), 2) = 0.83 \leq \theta_2 < 1.14 = w((2, 3), 3)$$

- Thus, both (1, 3) and (2, 2) are self-thresholds

Self-Policy -- Multiplicity

- Example 1 demonstrates that for the multi-class queue, except for some special cases like the standard GI/M/c FCFS queue, there can be multiple self-policies for any given $\{\theta_i\}$
- It can be shown, under a mild condition, that all self-polices are connected in integral-lattice domain

Self-Policy -- Nash Equilibrium

- Decision rule of type-*i* customers is said to be optimal against policy ***N*** if using the rule yields largest expected utility for type-*i* while customers of all types use ***N***
- Self-policy ***N*** is called Nash equilibrium if N_i is optimal against ***N*** for all *i*
- Nash equilibrium is a stable policy. In other words, under a Nash equilibrium, no one has an incentive to deviate from the policy. Therefore, self-threshold policy ***N*** is equilibrium if

$$w(\boldsymbol{N}, N_i) \le \theta_i < w(\boldsymbol{N}, N_i + 1)$$

for all *i*

Example 2: M/M/1 PS Queue with 2-Types

- Does an equilibrium policy always exist?
- In Example 1, for $\theta_1 \in (0.83, 0.95)$ and $\theta_2 \in (1.11, 1.14)$ both (1, 3) and (2, 2) are self-thresholds
- As w((2, 2), 3) = 1.06 < θ_2, (2, 2) is not equilibrium;

  If $\theta_1 \in (0.83, 0.92)$, then w((1, 3), 2) = 0.92 > θ_1 and w((1, 3), 4) = 1.33 > θ_2, (1, 3) is equilibrium; if $\theta_1 \in (0.92, 0.95)$, it is not. So, no equilibrium policy
- The consequent question would be: Is it possible to have multiple equilibrium policies?

Existence and Uniqueness of Equilibrium

- For FCFS queue with multi-class customers, exists uniquely
- For PS queue with singe-class customers, exists none or one
- For M/M/1 queue with single-class and increasing μ_n, exists at least one
- **Theorem** The queue with multi-class customers and under general service mode and increasing μ_n has at most one equilibrium self-policy

What if there is no Equilibrium Self-Policy

- Any self-threshold without binding contract cannot be stable, and system's performance under decentralized decision fluctuate and never converge.
- A mathematical approach is to consider randomized threshold, i.e., real-valued thresholds. For example, Ben-Shahar, Orda and Shimkin [2000] shows PS queue with homogeneous customers always exists a unique equilibrium randomization self-threshold.
- With nice mathematical properties, but not practical

Collecting Toll to Equilibrate

- Common economical means to equilibrate the system is by imposing toll that modifies $\{\theta_i\}$: a type-*i* customer need to pay γ_i for joining, which will make

$$w(\boldsymbol{N}, N_i) \leq \theta_i - \gamma_i < w(\boldsymbol{N}, N_i+1)$$

  with an ``appropriate'' equilibrium self-threshold $\boldsymbol{N}$.

- However, there are incentive-compatibility and fairness issues that are unavoidable and difficult to resolve.
- Similar issues arise if equilibrating by regulating arrival rates $\{\lambda_i\}$

Our Approach to Achieving Equilibrium

- An operational means that is simpler than imposing tolls or regulating arrival rates is to modify service rate with $\{\theta_i\}$, $\{\lambda_i\}$ and service discipline unchanged
- For example, if a FCFS queue with c servers that has no equilibrium self-policy, we can let $\mu_n = \mu \min\{n, c\}$ to make it a standard GI/M/c queue under FCFS that has a unique equilibrium self-policy.
- Require to show for any given $\{\theta_i\}$ and $\{\lambda_i\}$, it always exists such a service rate modification. In other words, when $\boldsymbol{\mu} = \{\mu_n\}$ is at our disposal, the system can be led to equilibrium state

Equilibration

- Let $\boldsymbol{S}$ be feasible control space that contains all $\boldsymbol{\mu} = \{ \mu_n \}$ that μ_n is non-decreasing and concave in n
- We will construct a correspondence $R : \boldsymbol{S} \rightarrow \boldsymbol{S}$ with original $\boldsymbol{\mu} \in \boldsymbol{S}$, $R(\boldsymbol{\mu})$ outputs a set of service rates under which the system is equilibrium
- Key tool to show $R(\boldsymbol{\mu})$ is not empty is by Kakutani Fixed Point Theorem:

  A correspondence having a fixed point if it is

  defined on non-empty, compact and convex domain,

  non-empty, convex-valued, and having closed graph

Fixed Point Theorem

- **Lemma 1**. Feasible control set $\boldsymbol{S}$ is compact and convex.
- Next, construct point-to-set mapping

$$R_j = h_j \circ f_j \circ g : \boldsymbol{S} \rightarrow \boldsymbol{S}^{-j}$$

and $R : \boldsymbol{S} \rightarrow \boldsymbol{S}$ as Cartesian product of R_j, i.e.,

$$R = R_1 \times R_2 \times \cdots \times R_k \quad \text{where}$$

1. Function $g : \boldsymbol{S} \rightarrow \boldsymbol{I}^k$ as $g(\boldsymbol{\mu})=(N_1, N_2, \ldots, N_k)$, i.e., under $\boldsymbol{\mu}$, $\boldsymbol{N}$ is certain self-policy
2. Correspondence $f_j : \boldsymbol{I}^k \rightarrow \boldsymbol{S}$ as

$$f_j(\boldsymbol{N}) = \{ \boldsymbol{\mu} \in \boldsymbol{S}\text{: such that } N_j \text{ is equilibrium}\}$$

3. Function $h_j : \boldsymbol{S} \rightarrow \boldsymbol{S}^{-j}$ as

$$h_j(\mu_1, \mu_2, \ldots, \mu_k) = (\mu_{N_{j-1}+1}, \mu_{N_{j-1}+2}, \ldots, \mu_{N_j})$$

Existence of Feasible Rate Modification

- **Lemma 2**. Correspondence R_j is non-empty, convex-valued and has a closed graph for all j.
- With Lemma 1 and 2, R meets the conditions of Kakutani Fixed Point Theorem. So, we have shown that there exists some $\boldsymbol{\mu} = R(\boldsymbol{\mu})$.
- To conclude

Theorem. For the queue with arriving rate $\{\lambda_i\}$ and utility $\{\theta_i\}$, there exists service rate $\{\mu_n\}$ that under which an equilibrium self-policy is guaranteed

Incentive-Compatibility and Fairness

- The modification of service rate is based on number of customers in the system, not on the types, it is clearly incentive compatible
- While there could be infinitely many ways to modify the departure rate for equilibration, we define various criterion of fairness to find appropriate ones
- Denote μ_n as original rate and μ_n^m as modified rate

Minimal Adjustment

- A natural modification is to adjust $\{\mu_n\}$ as little as possible when converting non-equilibrium ***N*** to become equilibrium.
- It is appropriate if the cost of rate modification is of major concern and proportional to the amount of change
- For that goal, we obtain optimal modification from

$$\text{Min } \Sigma_{n=1}^{N_k} |\, \mu_n^m - \mu_n|$$

$$\text{s.t. } \boldsymbol{\mu}^m \in R(\boldsymbol{S})$$

Maximal Adjustment

- If system's concern is on the operation cost rate that is proportional to the service rate, then an appropriate adjustment is to **reduce** $\{\mu_n\}$ as much as possible when converting non-equilibrium ***N*** to become equilibrium.
- For that goal, we obtain optimal modification from

$$\text{Max } \Sigma_{n=1}^{N_k} \, | \, \mu_n^m - \mu_n^o | \, P_n$$

$$\text{s.t. } \boldsymbol{\mu}^m \in R(\boldsymbol{S})$$

Minimal Mean Waiting Time Increased

- To eliminate the arbitrage, the rate should be reduced so that the mean waiting times will increase accordingly. From customers' perspective, a fair adjustment, while leads to equilibrium, should increase their mean waiting time in the system as little as possible.
- For that goal, we obtain optimal modification from

$$\text{Min } \Sigma_{n=1}^{N_k} \mid w^m(\boldsymbol{N}, n) - w(\boldsymbol{N}, n) \mid$$

$$\text{s.t. } \boldsymbol{\mu}^m \in R(\boldsymbol{S})$$

Example 3: Optimal Rate Modifications

- In Example 2, $\theta_1 = 0.940$, $\theta_2 = 1.131$ and $\mu_n = 2$ for all n, self-policy (1, 3) is non-equilibrium due to N_1.
- Minimal adjustment, we get $\mu_1^m = 1.92$, $\mu_2^m = 1.95$, and $\mu_n^m = 2$ for $n \geq 3$. Amount of adjustment: (2-1.92)+(2-1.95) = 0.13, minimal.
- Maximal adjustment, we get $\mu_1^m = 1.75$ and $\mu_n^m = 2$ for $n \geq 2$. Amount of adjustment: 2-1.75 = 0.25, maximal.
- Fair adjustment, we get $\mu_1^m = 1.93$, $\mu_2^m = 1.93$ and $\mu_n^m = 2.04$ for $n \geq 3$. Total mean waiting time increased by 0.053, compared to 0.066 under minimal rate adjustment.

Notice that μ_3^m is increased from 2 to 2.04.

Thanks for Your Attention

Reference

- Adan and B. D'Auria, 2016, Sojourn time in a single-server queue with threshold service rate control, *SIAM J. Appl. Math*.
- K. R. Balachandran and S. Radhakrishnan, 1994, Extensions to type Dominance Characteristics, *Management Science*
- I. Ben-Shahar, A. Orda and N. Shimkin, 2000, Dynamic service sharing with heterogeneous preferences, *Queueing Systems*
- A. C. Brooms, 2005, On the Nash equilibrium for the FCFS queueing system with load-increasing service rate, *Adv. Appl. Prob*.
- R. Hassin and M. Haviv, 2003, *To queue or not to queue: Equilibrium behavior in queueing systems*, Kluwer Academic Publishers
- C. Wang, 2016, On Socially Optimal Queue Length, *Management Science*.