Strategic customer behavior and optimal policies in a passenger-taxi double-ended queueing system with multiple access points and nonzero matching times

Hung Q. Nguyen[1] and Tuan Phung-Duc[2]

[1]Graduate School of Science and Technology, University of Tsukuba, Tennodai 1-1-1, Tsukuba, 305-8573, Ibaraki, Japan.
[2]Faculty of Engineering, Information and Systems, University of Tsukuba, Tennodai 1-1-1, Tsukuba, 305-8573, Ibaraki, Japan.

Contributing authors: s2030113@s.tsukuba.ac.jp; tuan@sk.tsukuba.ac.jp;

Abstract

This paper considers an observable double-ended queueing system of passengers and taxis, where matching times follow an exponential distribution. We assume that passengers are strategic and decide to join the queue only if their expected utility is nonnegative. We show that the strategy of passengers is represented by a unique vector of thresholds corresponding to different cases of the number of taxis observed in the system upon passenger arrival. Furthermore, we develop a heuristic algorithm to find an optimal range of fees to be levied on passengers to maximize social welfare or revenues.

Keywords: Double-ended queueing system, Strategic queueing, Matching queue

1 Introduction

1.1 Context and motivation

The topic of queues with strategic customers has attracted considerable attention. A great deal of attention has been devoted to individual optimality in observable queueing contexts. One of the interesting and notable extensions in the topic is double-ended queueing systems, which are practically seen in various social issues. In this study, we focus on one of the most typical examples of double-ended queueing systems: a passenger-taxi queue, which is usually seen at airports or railway stations.

In a double-ended queue, there usually exist two types of entities arriving at each side of the queue, forming a matching and leaving the system. In reality, the matching process usually takes time for several reasons, such as communication between passengers and taxi drivers or the fact that passengers may bring along bulky luggage; therefore, the boarding time will not be negligible. Furthermore, there are usually multiple access points for both taxis and passengers to match.

The lack of examination of such issues in the existing literature motivates us to consider a general system with nonzero matching times and multiple servers in this current paper.

1.2 Contributions

Compared to the queueing system with zero matching times, which has been widely studied in previous studies, the system in this research poses several challenges. In the system where there is no matching time, passengers and taxis are not simultaneously present in the queue; therefore, the system state can be represented by one variable, that is, the number of passengers or taxis in the queue and passengers base their decision on that variable only. When only passengers are strategic, due to the monotonic property of the expected waiting time with respect to the queue length, it is easily implied that passengers adopt a threshold strategy, indicating that they will balk the queue if observing a queue length that exceeds a certain threshold. Meanwhile, in the system with nonzero matching time, the queue may contain both passengers and taxis at the same time; thus, we need a two-dimensional Markov process to model this system. In this case, the strategic behavior of passengers depends on two variables, and the monotonic properties of the expected waiting time with respect to each variable are not obvious.

The main contributions of this study can be summarized as follows.

- We show that in an observable double-ended queueing system with multiple servers and nonzero matching times, passengers adopt a threshold strategy that is represented by a vector of $K+1$ thresholds corresponding to $K+1$ cases of the number of taxis in the system (K is the maximum number of taxis present in the system).

- We present a two-dimensional Markov model with an infinitesimal generator containing nonhomogeneous block matrices to calculate performance measures when passengers are strategic.
- We propose a heuristic algorithm that allows us to find an optimal range of toll fees levied on passengers at which social welfare revenues are maximized.

1.3 Literature review

The study of strategic queueing has attracted attention since the pioneering study of Naor [1], who considered the threshold-based behavior of customers arriving at an M/M/1 queueing system where system states can be observed. Later, Edelson and Hildebrand [2] created a counterpart framework to analyze customer behavior in case customers cannot observe system states. The two scenarios are usually referred to as the *observable case* and the *unobservable case*, respectively. The idea further led to various extensions with different settings, such as nonexponential service distributions [3, 4], reneging behavior [5, 6], feedback queues [7, 8], and preemptive queues [9, 10]. A comprehensive and detailed overview of the related literature and methodology is presented in [11, 12].

In observable queueing contexts, Altman and Shimkin [13] discussed the mechanism of convergence to a Nash equilibrium in a processor-sharing service system. Hassin [14] investigated the behavior of queueing customers with a first-come-last-served discipline. Legros [15] analyzed a service system with two types of customers classified by the level of emergency of the job. Most of the studies have shown that social welfare is not optimal when queueing participants make selfish strategic decisions. Ghosh and Hassin summarized measurements of this inefficiency in [16].

Double-ended queues have been used for modeling a variety of social systems, such as sharing economy [17, 18], disasters and repairs [19, 20], allocation of live organs [21, 22] and transportation [23]. There have been an increasing number of recent studies on customer behavior in transportation systems. Shi and Lian [24] analyzed the joining behaviors of passengers in a passenger-taxi queue with zero matching times. The model was extended by incorporating a gated policy [25]. In [26], the same system setting was considered in the context of customer loss aversion. In [27], different levels of information were considered. [28] considered a matching queue with two types of customers arriving at each end and differentiated by their priority. An observable taxi-passenger queue was considered in [29], where the authors attempted to optimize social welfare by adjusting the maximum buffer at both sides of the queue. [30] employed Markov modulated fluid flow to analyze a double-ended queue with abandonment. However, matching times between entities on two sides are ignored in all of the aforementioned papers.

To the best of our knowledge, the only two papers in which matching times were considered are [31, 32]. In the former paper, customer strategic behavior was not considered, while in the latter study, matching times were assumed to follow a simple two-mass point distribution, which seems to be unrealistic since matching times vary on a case-by-case basis. Additionally, the system was assumed to be a single server in both studies.

The rest of the paper is structured as follows. Section 2 describes our model and introduces some notation. Next, we analyze the strategic behavior of passengers in Section 3. In Section 4, we introduce an algorithm to find an optimal policy to maximize social welfare. In Section 5, we consider the scenario where the platform owner aims to maximize revenue. Section 6 presents a numerical example. In the last section, we make some concluding remarks and provide directions for future research.

2 Model description and notation

We consider an observable system with passengers arriving at a taxi stand containing S identical access points. The area (including S taxi access points) can accommodate at most K taxis at the same time ($K \geq S$). Passengers and taxis arrive at the taxi stand according to Poisson processes with arrival rates λ_p and λ_t, respectively. Matching times between passengers and taxi drivers follow an exponential distribution with mean $1/\mu$. To simplify the problem, we treat one passenger and a batch of passengers riding together as the same entity. When the parking area reaches its maximum capacity, the arrivals of taxis are blocked, and taxis leave immediately. We assume that the buffer capacity of passengers is infinite. If a passenger arrives when all access points are busy or there are no taxis available for matching, the passenger will wait in the queue under a FCFS (first-come-first-served) service discipline.

Let $I(t)$ and $J(t)$ denote the total number of passengers in the system and the number of taxis in the system at time t, respectively. Under this setting, the process $\{(I(t), J(t))|\ t \geq 0\}$ is a continuous-time Markov chain. When agents are not strategic (that is, there is no balking upon arrival), the transition diagram can be illustrated in Fig. 1.

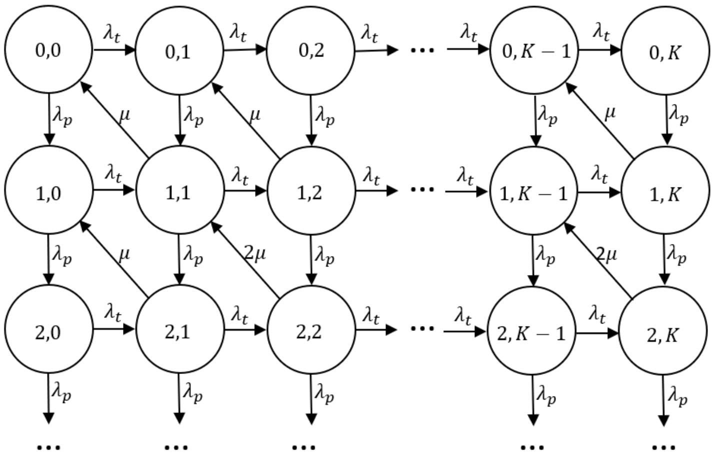


Fig. 1: Transition diagram of the nonstrategic queueing system

A single-server version of this model was thoroughly investigated in [31], in which the authors derived the stability condition and sojourn time distributions of both passengers and taxis. Furthermore, the stability condition of the multiserver system with the same setting was derived in [33] as follows.

$$\lambda_p < \left(\sum_{j=0}^{S-1} j\pi_j + S \sum_{j=S}^{K} \pi_j \right) \mu,$$

where π_j denotes the probability that there are j taxis in an M/M/S/K queue with arrival rate λ_t and service rate μ.

In the current research, we add economic parameters and study the system in equilibrium.

3 Self-optimization

In this section, we derive the strategic behavior of passengers, which is based on their utility. To maintain the concise and easy-to-follow structure of the paper, we put lengthy proofs in the Appendices.

Let U_p denote the individual utility of an arbitrary passenger who spends W time units in the system and receives a reward R_p after completing the service. Assume that each unit of time in the system costs C_p monetary units. Then, we have

$$U_p = R_p - C_p W.$$

Since sojourn time is stochastic, passengers base their behavior on the expected utility. Assume that passengers know their reward and waiting cost and estimate their expected utility as

$$\mathsf{E}(U_p) = R_p - C_p \mathsf{E}(W).$$

If $\mathsf{E}(U_p) \geq 0$, the passenger joins the queue; otherwise, that passenger balks.

It is reasonable to assume that

$$R_p \geq \frac{C_p}{\mu}.$$

This guarantees that passengers are willing to join an empty queue.

Note that the sojourn time can be decomposed into waiting time and service time, where the expected service time is constant at $\frac{1}{\mu}$, while the expected waiting time depends on the current system state. Denote by $T(p, j)$ the expected waiting time of a passenger observing a system state $s = (p, j)$ upon arrival, where j represents the current number of taxis in the system, and p represents the current "position" of the passenger. If $p = 0$ then the passenger is currently matching at an access point. If $p > 0$ then the passenger is p steps away from the taxi stand; in other words, $p - 1$ is the current queue length (which does not include passengers in matching) being observed. We need to calculate all the values of $T(p, j)$ to investigate the strategic behavior of customers. To this end, we provide a recursive procedure to calculate $T(p, j)$ based on the first step analysis. Specifically, $T(s)$ is recursively associated with the expected waiting time at any state s_i, which is reached by one step from state s, denoted by $T(s_i)$, that is,

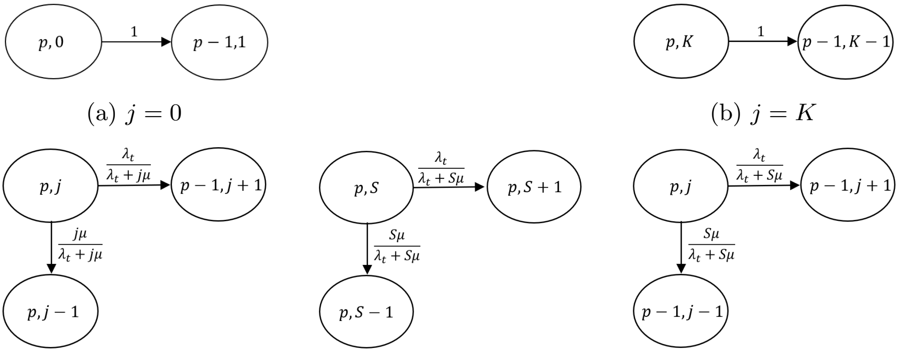
$$T(s) = \mathsf{T}_s + \sum_i \gamma_i T(s_i), \tag{1}$$

where γ_i is the probability that the passenger moves to the next state s_i, and T_s denotes the expected time needed to stay in state s. The one-step transitions from state (p, j) are illustrated in Fig. 2.

It immediately follows that $T(0, j) = 0$ for $j > 0$. When $p > 0$, using formula (1), we can derive $T(p, j)$ as follows.

$$T(p,j) = \begin{cases} \frac{1}{\lambda_t} + T(p-1, j+1) & \text{if } j = 0, \\ \frac{1}{\lambda_t + j\mu} + \frac{\lambda_t}{\lambda_t + j\mu} T(p-1, j+1) + \frac{j\mu}{\lambda_t + j\mu} T(p, j-1) & \text{if } 0 < j < S, \\ \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(p, j+1) + \frac{S\mu}{\lambda_t + S\mu} T(p, j-1) & \text{if } j = S, \\ \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(p, j+1) + \frac{S\mu}{\lambda_t + S\mu} T(p-1, j-1) & \text{if } S < j < K, \\ \frac{1}{S\mu} + T(p-1, j-1) & \text{if } j = K. \end{cases} \tag{2}$$

In what follows, we derive passengers' strategies in the form of a multi-threshold vector of maximum positions (at which they are willing to join)



(a) $j = 0$ (b) $j = K$

(c) $0 < j < S$ (d) $j = S$ (e) $S < j < K$

Fig. 2: Transition diagrams from state (p, j) $(p > 0)$ with transition probabilities

corresponding to a specific number of taxis present in the system in Theorem 1, which is obtained by Propositions 1 and 3. We present four propositions that show noticeable properties of expected waiting times and the derived threshold strategy. To be more specific, Proposition 1 shows the monotonicity property of $T(p, j)$ with respect to p. Proposition 2 shows the monotonicity property of $T(p, j)$ with respect to j, which is obtained by Lemmas 1–3. Proposition 3 shows that the expected waiting time reaches infinity as p reaches infinity. Proposition 4 shows the monotonicity property of thresholds.

Proposition 1 (Monotonicity property of expected waiting times with respect to the position)**.** $T(p, j) \leq T(p + 1, j)$ *for any fixed value of* j*.*

Proof See Appendix A. □

This result is intuitive in the sense that with the same number of taxis in the system, the farther passengers are away from the taxi stand, the longer they need to wait. However, looking at the recursive formulas, we can see that such a relationship between $T(p, j)$ and $T(p + 1, j)$ is not obvious.

Lemma 1

$$T(1, j) = \begin{cases} \frac{1}{\lambda_t} & \text{if } 0 \leq j \leq S - 1, \\ \frac{\lambda_t^2 + (S\mu)^2 + \lambda_t(S\mu)}{\lambda_t(S\mu)(\lambda_t + S\mu)} & \text{if } j = S, \\ \frac{1}{S\mu} & \text{if } S + 1 \leq j \leq K. \end{cases}$$

Proof See Appendix B. □

This result explicitly shows expected waiting times when passengers are one step away from the taxi stand. When there are fewer than S taxis, passengers are expected to wait for $\frac{1}{\lambda_t}$ units of time. When there are more than S taxis, passengers are expected to wait for $\frac{1}{S\mu}$ units of time. Meanwhile, the expected waiting time for those who observe exactly S taxis in the system is higher at $\frac{\lambda_t^2+(S\mu)^2+\lambda_t(S\mu)}{\lambda_t(S\mu)(\lambda_t+S\mu)}$ units of time.

We use this result to prove the following two lemmas.

Lemma 2 For any fixed value of p,

$$T(p, j) \leq \frac{1}{j\mu} + T(p, j-1),$$

for $1 \leq j \leq S-1$.

Lemma 3 For any fixed value of p,

$$T(p, j-1) \leq \frac{1}{\lambda_t} + T(p, j),$$

for $S+2 \leq j \leq K$.

Proof See Appendices C and D for the proofs of Lemma 2 and Lemma 3, respectively. □

We use these results to prove the following proposition.

Proposition 2 (Monotonicity property of expected waiting times with respect to the observed number of taxis). $T(p, j) \leq T(p, j+1)$ *for* $j = 0, 1, ..., S-1$, *and* $T(p, j) \geq T(p, j+1)$ *for* $S, S+1, ..., K-1$.

Proof See Appendix E. □

Intuitively, at the same arbitrary position, in case the current number of taxis in the system is greater than or equal to S, expected waiting times become shorter if there are more taxis in the system. Nevertheless, we see an opposite association between the number of taxis and waiting times when there are S or fewer than S taxis in the system. This is because when the number of taxis in the systems is smaller than or equal to the number of access points, the number of passengers being served at the taxi stand is equal to the number of taxis. In this case, more taxis indicates that more passengers occupy the taxi stand at the time of arrival. Regardless of the number of taxis in the system (which is less than or equal to S), the passenger at position p needs a fixed number of p more taxis for his turn to be served. If more passengers match at access points, it is more likely that the taxi stand becomes more “congested”, which may slow down the expected waiting time.

Proposition 3 (Infinity limit of waiting times)**.**

$$\lim_{p \to +\infty} T(p, j) = +\infty$$

for all $j = 0, 1, ..., K$.

Proof See Appendix F. □

This result indicates that the expected waiting time diverges to infinity if the current position is infinitely far from the taxi stand. We use this result to prove the following theorem.

Theorem 1 (Equilibrium strategy of passengers) *Passengers who arrive at the system adopt a threshold strategy represented by the vector* $\rho^s = (p_0^s, p_1^s, ..., p_K^s)$*, where* p_j^s *is the maximum position at which passengers are willing to join the system when they observe* j *taxis upon arrival.*

Proof This can be obtained from the monotonically nondecreasing property of $T(p, j)$ with regard to p obtained in **Proposition 1**. Since $R_p - \frac{C_p}{\mu} \geq 0$ by assumption and $R_p - C_p T(p, j) - \frac{C_p}{\mu} \to -\infty$ for all $j = 0, 1, ..., K$ when $p \to +\infty$ (due to the result in **Proposition 3**), for each fixed value of j, there must exist p_j^s such that $R_p - C_p T(p_j^s, j) - \frac{C_p}{\mu} \geq 0$ and $R_p - C_p T(p_j^s + 1, j) - \frac{C_p}{\mu} < 0$. p_j^s is the threshold strategy corresponding to each fixed value of the number of taxis observed upon arrival. □

Proposition 4 (Monotonicity property of thresholds)**.** $p_k^s \geq p_{k+1}^s$ *for* $k = 0, 1, ..., S-1$*, and* $p_k^s \leq p_{k+1}^s$ *for* $k = S, S+1, ..., K-1$.

Proof This can be proved by contradiction.

First, consider the case when $0 \leq j \leq S-1$. We will prove $p_j^s \geq p_{j+1}^s$ Assume that $p_j^s < p_{j+1}^s$, which implies $p_j^s + 1 \leq p_{j+1}^s$. Since p_j^s and p_{j+1}^s are both decision thresholds, we must have $T(p_j^s + 1, j) > \frac{R}{C} - \frac{1}{\mu}$ and $T(p_{j+1}^s, j+1) \leq \frac{R}{C} - \frac{1}{\mu}$, which imply $T(p_j^s + 1, j) > T(p_{j+1}^s, j+1)$. Additionally, by the monotonic properties of $T(p, j)$ on p and j (obtained in **Propositions 1** and **2**) and the earlier assumption, we have $T(p_{j+1}^s, j+1) \geq T(p_j^s + 1, j+1) \geq T(p_j^s + 1, j)$. This contradiction indicates $p_j^s \geq p_{j+1}^s$ for $j = 0, 1, ..., S-1$.

Second, consider the case when $S \leq j \leq K-1$. We will prove $p_j^s \leq p_{j+1}^s$. Assume that $p_j^s > p_{j+1}^s$, which implies $p_j^s \geq p_{j+1}^s + 1$. Since p_j^s and p_{j+1}^s are both decision thresholds, we must have $T(p_j^s, j) \leq \frac{R}{C} - \frac{1}{\mu}$ and $T(p_{j+1}^s + 1, j+1) > \frac{R}{C} - \frac{1}{\mu}$, which imply $T(p_{j+1}^s + 1, j+1) > T(p_j^s, j)$. Additionally, by the monotonic properties of $T(p, j)$ on p and j and the earlier assumption, we have $T(p_{j+1}^s + 1, j+1) \leq T(p_j^s, j+1) \leq T(p_j^s, j)$. This contradiction indicates $p_j^s \leq p_{j+1}^s$ for $j = S, S+1, ..., K-1$. □

This result indicates that the strategy threshold decreases as the number of taxis increases within the range from 0 to S. Passengers adopt a greater threshold when there are more taxis in the system and the number of taxis is larger than the number of access points.

4 Overall optimization

Let ξ_p and ξ_t denote the expected number of passengers and taxis being diverted from the service station per unit time, respectively. Let L_p and L_t denote the expected queue lengths of passengers and taxis. Additionally, denote by C_t the cost of staying in the system per unit time of taxi drivers and R_t the reward that taxi drivers receive after completing serving a passenger. Similarly, denote by C_p the cost of staying in the system per unit time of passengers and R_p the reward (service value) that passengers receive after being served.

Expected social welfare per unit time of all entities joining the system is then

$$U = (\lambda_p - \xi_p)R_p + (\lambda_t - \xi_t)R_t - C_pL_p - C_tL_t. \tag{3}$$

Since social welfare cannot be explicitly expressed in terms of threshold ρ, we need to use a brute-force search method to find the maximum value of social welfare. The traditional approach is to search for a socially optimal threshold strategy in the first place and then derive a corresponding optimal fee range that adjusts the self-optimal threshold to the socially optimal threshold. Such an approach is not feasible in this multidimensional case for several reasons. First, to perform an exhaustive search to find the socially optimal threshold, it is necessary that the number of cases being considered is finite, which requires an upper limit for the passenger buffer size, while we do not have this assumption. Second, even if we set a maximum buffer size of passengers at m, the number of cases to be considered is $(m+1)^{K+1}$, which becomes massively large when K and m are large. Finally, even if we manage to find a socially optimal threshold strategy, there is no guarantee that it can be shifted from the original self-optimal threshold strategy by implementing a fixed value for fee θ since the strategy being considered is a vector of $K+1$ values.

In this paper, we introduce a heuristic algorithm to find an optimal policy. Assume that the administrator of the system levies a toll fee θ on each passenger joining the system. We want to find an optimal range of θ that maximizes expected social welfare. With a toll fee of θ, the expected individual utility of passengers becomes

$$\mathsf{E}(U_p) = R_p - \theta - C_p\mathsf{E}(W). \tag{4}$$

Similar to the analysis in the previous section, when a toll fee is imposed, passengers still self-optimize and adopt a threshold strategy $\rho = (p_0, p_1, ..., p_K)$, which satisfies the property in **Proposition 4**. ρ remains unchanged as θ gradually increases within a certain fee range. When θ exceeds the upper bound of the range, some threshold element(s) p_j, which together with its corresponding number of taxis j yield the longest expected waiting time $T(p_j, j)$, decreases by

1. We then obtain a new ρ and repeat this procedure until all of the elements in ρ converge to $\mathbf{0}$, which is also the case where the taxi service becomes too expensive and passengers have no incentive to engage. Based on this property, we develop an algorithm to derive ranges of toll fees and strategy thresholds that passengers adopt accordingly.

Algorithm 1 Deriving fee ranges and threshold strategies

$\mathcal{T} \leftarrow \left\{T(p_j^s, j) \mid j = 0, 1, ..., K\right\}$
$\tau \leftarrow \max \mathcal{T}$
$\delta_1 \leftarrow 0$ ▷ lower bound of fee range (initial)
$\delta_2 \leftarrow R_p - C_p\tau - \frac{C_p}{\mu}$ ▷ upper bound of fee range (initial)
$\Delta \leftarrow [\delta_1, \delta_2]$ ▷ fee range (initial)
$\rho = (p_0, p_1, ..., p_K) \leftarrow (p_0^s, p_1^s, ..., p_K^s)$ ▷ threshold corresponding to fee range (initial)
$\mathcal{O} \leftarrow \{(\Delta, \rho)\}$ ▷ initial output: an initial pair of fee range and threshold
while $\rho \neq \mathbf{0}$ **do** ▷ $\mathbf{0}$: zero vector
  for $T(p, j)$ in $\mathcal{T}$ **do** ▷ updating ρ
    if $T(p, j) = \tau$ **then**
      $\mathcal{T} \leftarrow \mathcal{T} \setminus \{T(p, j)\} \cup \{T(p-1, j)\}$
      $p_j \leftarrow p_j - 1$
    end if
  end for
  $\tau \leftarrow \max \mathcal{T}$
  $\delta_1 \leftarrow \delta_2$
  $\delta_2 \leftarrow R_p - C_p\tau - \frac{C_p}{\mu}$
  $\Delta \leftarrow (\delta_1, \delta_2]$ ▷ updating Δ
  $\mathcal{O} \leftarrow \mathcal{O} \cup \{(\Delta, \rho)\}$ ▷ updating set of output
end while
return $\mathcal{O}$

For each pair of (Δ, ρ) obtained from **Algorithm 1**, we can correspondingly derive a value of social welfare by the following procedure.

In this study, we assume that passengers do not renege. Therefore, even when passengers adopt such a threshold strategy, system states may still exist at which the passenger queue length exceeds the threshold. Such states cannot be reached by the joining behavior of passengers but are stochastically attained in one of the following two scenarios.

- Successive completions of matching events: if a passenger arrives and decides to join at a certain state (i, j) where $i \leq \min\{j, S\} + p_j^s$, it may be the case that this passenger will reach a state $(i - q, j - q)$ where $q \leq i, q \leq j$ and $i - q > \min\{j - q, S\} + p_{j-q}^s$ if q pairs of passengers and taxis finish matching and leave the system in the next consecutive q events.

- Successive arrivals of taxis: if a passenger arrives and decides to join at a certain state (i, j) where $i \leq \min\{j, S\} + p_j^s$, it may be the case that this passenger will reach a state $(i+q, j)$ where $j > \min\{j+q, S\} + p_{j+q}^s$ if q taxis arrive in the next consecutive q events.

We then obtain the infinitesimal generator $\mathcal{Q}$ of the Markov chain modeling the system in equilibrium as below.

$$\mathcal{Q} = \begin{pmatrix} \mathcal{B}^{(0)} & \mathcal{C}^{(0)} & & & & \\ \mathcal{A}^{(1)} & \mathcal{B}^{(1)} & \mathcal{C}^{(1)} & & & \\ & \mathcal{A}^{(2)} & \mathcal{B}^{(2)} & \mathcal{C}^{(2)} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mathcal{A}^{(K-1)} & \mathcal{B}^{(K-1)} & \mathcal{C}^{(K-1)} \\ & & & & \mathcal{A}^{(K)} & \mathcal{B}^{(K)} \end{pmatrix},$$

where block matrices $\mathcal{A}^{(j)}, \mathcal{B}^{(j)}$ and $\mathcal{C}^{(j)}$ are dimensionally nonhomogeneous and defined as follows.

- $\mathcal{Q}$ has neither zero columns nor zero rows.
- $\mathcal{C}^{(j)}, \mathcal{B}^{(j+1)}, \mathcal{A}^{(j+2)}$ have the same number of columns for $j = 0, 1, ..., K-2$. $\mathcal{B}^{(0)}$ and $\mathcal{A}^{(1)}$ have the same number of columns. $\mathcal{C}^{(K-1)}$ and $\mathcal{B}^{(K)}$ have the same number of columns.
- $\mathcal{A}^{(j)}, \mathcal{B}^{(j)}, \mathcal{C}^{(j)}$ have the same number of rows for $j = 1, 2, ..., K-1$. $\mathcal{B}^{(0)}$ and $\mathcal{C}^{(0)}$ have the same number of rows. $\mathcal{A}^{(K)}$ and $\mathcal{B}^{(K)}$ have the same number of rows.
- $\mathcal{C}_{i,i}^{(j)} = \lambda_t$, for

$$i = 0, 1, ..., \max\left\{j + \max_{l \geq j}\left(\min\{l, S\} + p_l - l\right), \max_{l \leq j}\left(\min\{l, S\} + p_l\right)\right\}.$$

- $\mathcal{A}_{i,i-1}^{(j)} = \min(i\mu, j\mu, S\mu)$, for

$$i = 1, 2, ..., \max\left\{j + \max_{l \geq j}\left(\min\{l, S\} + p_l - l\right), \max_{l \leq j}\left(\min\{l, S\} + p_l\right)\right\}.$$

- $\mathcal{B}_{i,i+1}^{(j)} = \lambda_p$, for $i = 0, 1, ..., \min\{j, S\} + p_j - 1$.
- All other elements of $\mathcal{Q}$ which do not lie on the main diagonal are 0.
- $\mathcal{Q}_{i,i} = -\sum_{j \neq i} \mathcal{Q}_{i,j}$.

We can then derive the steady state probabilities defined as $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_K)$, where $\boldsymbol{\pi}_j = (\pi_{0,j}, \pi_{1,j}, ...)$ is a vector encoding all probabilities when there are j taxis in the system at the steady state by solving the following equations:

$$\begin{cases} \boldsymbol{\pi}\mathcal{Q} = \mathbf{0}, \\ \boldsymbol{\pi}\boldsymbol{e} = 1, \end{cases}$$

where $\mathbf{0}$ is a zero vector of appropriate dimension, and $\boldsymbol{e}$ is a unit vector of appropriate dimension.

The blocking probability of taxis is then given by $\boldsymbol{\pi}_K\boldsymbol{e}$, and the blocking probability of passengers is

$$\sum_{j=0}^{K}\sum_{i\geq p_j+\min\{j,S\}}\pi_{i,j}.$$

The numbers of passengers and taxis diverted from the system per unit time are, respectively given by

$$\xi_p=\lambda_p\sum_{j=0}^{K}\sum_{i\geq p_j+\min\{j,S\}}\pi_{i,j},$$

and

$$\xi_t=\lambda_t\boldsymbol{\pi}_K\boldsymbol{e}.$$

The mean lengths of the passenger queue and taxi queue are, respectively given by

$$L_p=\sum_{j=0}^{K}\sum_{i}i\pi_{i,j},$$

and

$$L_t=\sum_{j=0}^{K}\sum_{i}j\pi_{i,j}.$$

Substituting ξ_p, ξ_t, L_p, L_t into (3), we obtain the value of social welfare with respect to the fee range Δ and threshold strategy ρ of passengers. Comparing all obtained values of social welfare, we acquire the maximum social welfare together with the corresponding fee range and threshold vector, which yield that optimal value.

5 Revenue maximization

We examine the case in which the owner of the platform aims to maximize their revenue by imposing a toll fee of θ on each passenger. We consider the following two scenarios. In the first scenario, the platform owner collects a fixed fee for a seasonal toll pass from taxi companies. In this case, revenue maximization is equivalent to maximizing revenue from passengers. Thus, the objective function of the platform owner is given by

$$M_1=(\lambda_p-\xi_p)\theta. \tag{5}$$

In another scenario, the platform owner also levies a toll fee for each entrance of a taxi, denoted by θ_t. Assume that this amount is already fixed in advance. The objective function in this case is given by

$$M_2 = (\lambda_p - \xi_p)\theta + (\lambda_t - \xi_t)\theta_t. \tag{6}$$

In both scenarios, it is easily seen that the optimal value of the revenue is attained at one of the fee range upper bounds because all other parameters remain unchanged within the fee range. Since we already obtained all possible fee ranges and corresponding parameters in the previous section, it is possible to compare the revenue in all cases and find the maximum revenue similarly.

6 Numerical analysis

In this section, we illustrate the results with a specific numerical example. Set $\lambda_p = 7, \lambda_t = 6, \mu = 12, S = 4, K = 15, R_p = 20, R_t = 18, C_p = 5, C_t = 5, \theta_t = 10$. Calculated results show that

- When there is no intervention from the administrators, passengers adopt threshold strategy

$$\rho^s = (23, 23, 23, 23, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 33)$$

  corresponding to the number of taxis being observed upon arrival ranging from 0 to 15. This results in an expected social welfare of 136.221.
- Figure 3 shows that social welfare is discretely unimodal with respect to fee ranges, and peaks at 203.122 when a fee ranging in $(16.250, 16.854]$ is imposed on each entrance. Within this fee range, passengers adopt strategy

$$\rho^o = (3, 3, 3, 3, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14).$$

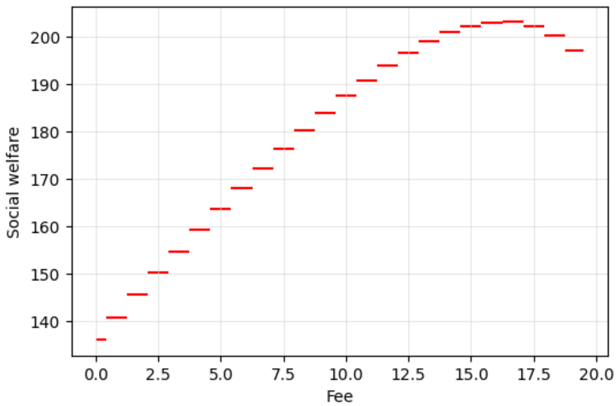


Fig. 3: Social welfare with respect to imposed fee

- The graph in Fig. 4, which is a dashed line (continuous within each range of toll fees), represents the relationship between the revenue from passengers and the toll fee levied on them. The maximum revenue from passengers is 114.893 when the platform charges a toll fee of 19.479 monetary units per passenger entrance.

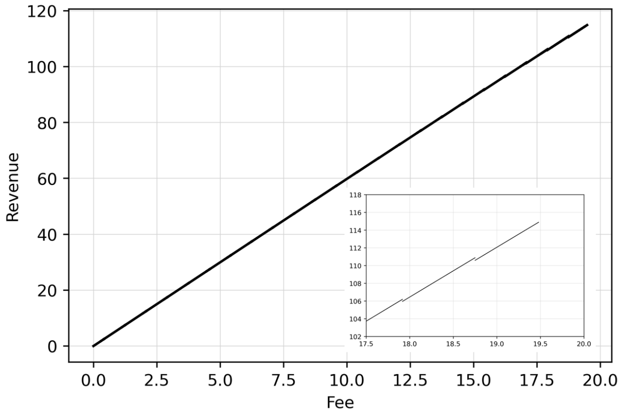


Fig. 4: Revenue from passengers with respect to imposed fee

- Figure 5 shows that the total revenue is maximum at 173.875 when the platform charges a toll fee of 19.479 monetary units per entrance of passengers.

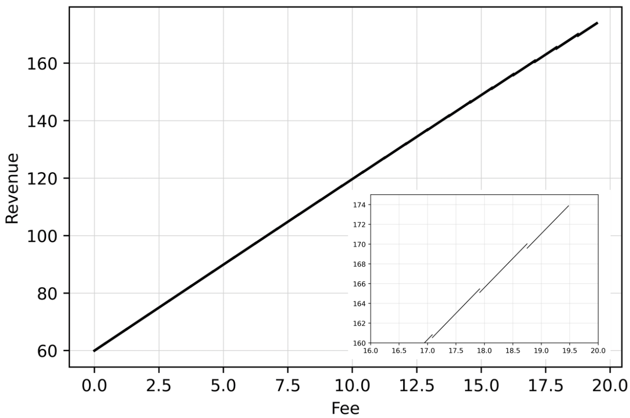


Fig. 5: Platform's revenue with respect to imposed fee

When the fee is larger than 19.479, passengers have no incentive to join the system since their expected utility becomes negative regardless of the number of taxis observed in the system.

- Finally, we present a sensitivity analysis of passengers' strategic behavior with respect to nonnegligible matching times. Except for the mean matching time, which is let vary, all parameters remain the same as in the previous experiments. The results are shown in Fig. 6.

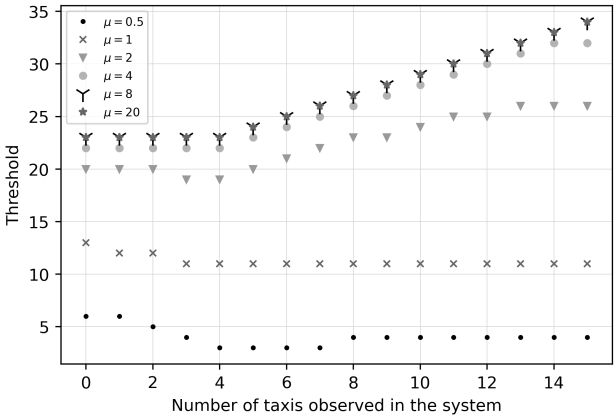


Fig. 6: Sensitivity of passengers' strategic behavior with respect to mean matching time

It can be observed that the thresholds adopted by passengers increase with an increased matching rate μ at first and then remain unchanged as μ becomes larger. Intuitively, as matching times becomes smaller, passengers' expected sojourn times also decrease, so they are willing to join a longer queue.

7 Concluding remarks

In this paper, we showed that passengers adopt a threshold-based strategy in a passenger-taxi queueing system with exponentially distributed matching times by proving the monotonically nondecreasing property of the expected waiting time with respect to the initial position upon arrival. We also derived passengers' strategies and corresponding optimal fee ranges that maximize social welfare or revenues. Theoretically, there might exist different threshold strategies that yield higher social welfare or revenues; however, a fixed toll fee cannot adjust passengers' behavior so that it coincides with such optimal behavior. This framework, although more computationally expensive, is a more general approach to the class of similar problems. For example, with the simple setting of an M/M/1 queueing system considered in the original work of Naor [1], instead of looking for a socially optimal threshold strategy from the beginning and deriving a corresponding fee range, a finite number of pairs of a fee range and a threshold strategy can be searched based on **Algorithm 1** and then the strategy that yields the maximal social welfare or revenue can be chosen. One limitation of this approach is that it does not cover the possibility that the global maximum of social welfare is attained when passengers adopt a higher threshold strategy and need a subsidy to do so.

In reality, not only passengers but also taxi drivers adopt strategic behavior. In this case, the strategy of one side (passengers or taxi drivers) may affect expected waiting times of the other side; therefore, decisions of entities not only depend on those of others of the same side but also depend on how the other side behaves, which finally leads to a Nash equilibrium of the system between two sides. This interesting but highly complex problem will be considered in the near future.

Acknowledgments. The research of Hung Q. Nguyen was supported by JST SPRING, Grant Number JPMJSP2124. The research of Tuan Phung-Duc was supported by JSPS KAKENHI Grant Numbers 18K18006 and 21K11765.

Appendix A Proof of Proposition 1

We will prove **Proposition 1** by induction on p. The statement is equivalent to

$$T(p, j) \leq T(p+1, j), \tag{A1}$$

for any fixed values of j.

Since $T(0, j) = 0$, it is obviously implied from the recursive formulas that $T(0, j) \leq T(1, j)$; thus, (A1) holds with $p = 0$. Assuming that (A1) holds with $p = q - 1$ for any integer $q \geq 1$, which indicates, for any fixed value of j,

$$T(q-1, j) \leq T(q, j). \tag{A2}$$

We show that it holds with $p = q$, which indicates that we need to prove that, for any fixed value of j,

$$T(q, j) \leq T(q+1, j),$$

by considering the following 5 cases.

- When $j = K$, from (2) we have

$$T(q, j) = \frac{1}{S\mu} + T(q-1, K-1), \tag{A3}$$

and

$$T(q+1, j) = \frac{1}{S\mu} + T(q, K-1). \tag{A4}$$

Since $T(q-1,K-1) \le T(q,K-1)$ by assumption (A2), from (A3) and (A4), we obtain

$$T(q,j) \le T(q+1,j) \text{ for } j = K. \tag{A5}$$

- When $S < j < K$, from (2) we have

$$T(q,j) = \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(q,j+1) + \frac{S\mu}{\lambda_t + S\mu} T(q-1,j-1),$$

and

$$T(q+1,j) = \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(q+1,j+1) + \frac{S\mu}{\lambda_t + S\mu} T(q,j-1).$$

Now, due to (A5), it is seen that the inequality $T(q,j) \le T(q+1,j)$ holds for $j = K-1$ because

$$\begin{aligned} T(q,K-1) &= \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(q,K) + \frac{S\mu}{\lambda_t + S\mu} T(q-1,K-2) \\ &\le \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(q+1,K) + \frac{S\mu}{\lambda_t + S\mu} T(q,K-2) \\ &= T(q+1,K-1). \end{aligned}$$

Then, it is easily obtained by induction on j, that

$$T(q,j) \le T(q+1,j) \text{ for } S < j < K. \tag{A6}$$

- When $j = 0$, from (1) we have

$$T(q,0) = \frac{1}{\lambda_t} + T(q-1,1), \tag{A7}$$

and

$$T(q+1,0) = \frac{1}{\lambda_t} + T(q,1). \tag{A8}$$

Since $T(q-1,1) \le T(q,1)$ because of the inductive assumption, from (A7) and (A8) we obtain

$$T(q,j) \le T(q+1,j) \text{ for } j = 0. \tag{A9}$$

- When $0 < j < S$, from (2) we have

$$T(q,j) = \frac{1}{\lambda_t + j\mu} + \frac{\lambda_t}{\lambda_t + j\mu} T(q-1,j+1) + \frac{j\mu}{\lambda_t + j\mu} T(q,j-1),$$

and

$$T(q+1,j) = \frac{1}{\lambda_t + j\mu} + \frac{\lambda_t}{\lambda_t + j\mu} T(q,j+1) + \frac{j\mu}{\lambda_t + j\mu} T(q+1,j-1),$$

Now, due to (A9), it is seen that the inequality $T(q,j) \le T(q+1,j)$ holds for $j=1$ because

$$\begin{aligned} T(q,1) &= \frac{1}{\lambda_t+\mu} + \frac{\lambda_t}{\lambda_t+\mu} T(q-1,2) + \frac{\mu}{\lambda_t+\mu} T(q,0) \\ &\le \frac{1}{\lambda_t+\mu} + \frac{\lambda_t}{\lambda_t+\mu} T(q,2) + \frac{\mu}{\lambda_t+\mu} T(q+1,0) \\ &= T(q+1,1). \end{aligned}$$

Then, it is easily obtained by induction on j, that

$$T(q,j) \le T(q+1,j) \text{ for } 0 < j < S. \tag{A10}$$

- Last, when $j=S$, from (2) we have

$$T(q,j) = \frac{1}{\lambda_t+S\mu} + \frac{\lambda_t}{\lambda_t+S\mu} T(q,S+1) + \frac{S\mu}{\lambda_t+S\mu} T(q,S-1), \tag{A11}$$

and

$$T(q+1,j) = \frac{1}{\lambda_t+S\mu} + \frac{\lambda_t}{\lambda_t+S\mu} T(q+1,S+1) + \frac{S\mu}{\lambda_t+S\mu} T(q+1,S-1). \tag{A12}$$

However, note that $T(q,S+1) \le T(q+1,S+1)$ and $T(q,S-1) \le T(2,S-1)$ (implied from results (A6) and (A10)). Therefore, from (A11) and (A12), we obtain

$$T(q,j) \le T(q+1,j) \text{ for } j = S. \tag{A13}$$

Equations (A5), (A6), (A9), (A10), (A13) complete our proof.

Appendix B Proof of Lemma 1

We prove **Lemma 1** by induction on j.

First, note that

$$T(1,0) = \frac{1}{\lambda_t} + \frac{\lambda_t}{\lambda_t+S\mu} T(0,1) = \frac{1}{\lambda_t},$$

and

$$T(1,K) = \frac{1}{S\mu} + \frac{S\mu}{\lambda_t+S\mu} T(0,K-1) = \frac{1}{S\mu}.$$

By induction on j, we have

$$\begin{aligned} T(1,j) &= \frac{1}{\lambda_t+j\mu} + \frac{\lambda_t}{\lambda_t+j\mu} T(0,j-1) + \frac{j\mu}{\lambda_t+j\mu} T(1,j-1) \\ &= \frac{1}{\lambda_t+j\mu} + \frac{j\mu}{\lambda_t+j\mu} . \frac{1}{\lambda_t} \end{aligned}$$

$$= \frac{1}{\lambda_t},$$

for $1 \leq\ j \leq S-1$; and

$$\begin{aligned}
T(1,j) &= \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(1,j+1) + \frac{S\mu}{\lambda_t + S\mu} T(0,j-1) \\
&= \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} . \frac{1}{S\mu} \\
&= \frac{1}{S\mu},
\end{aligned}$$

for $S+1 \leq\ j \leq K-1$.

Finally,

$$\begin{aligned}
T(1,S) &= \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(1,S+1) + \frac{S\mu}{\lambda_t + S\mu} T(1,S-1) \\
&= \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} . \frac{1}{S\mu} + \frac{S\mu}{\lambda_t + S\mu} . \frac{1}{\lambda_t} \\
&= \frac{\lambda_t^2 + (S\mu)^2 + \lambda_t (S\mu)}{\lambda_t (S\mu) (\lambda_t + S\mu)}.
\end{aligned}$$

It can also be noted that

$$\begin{aligned}
T(1,S) - T(1,S-1) &= \frac{\lambda_t^2 + (S\mu)^2 + \lambda_t (S\mu)}{\lambda_t (S\mu) (\lambda_t + S\mu)} - \frac{1}{\lambda_t} \\
&= \frac{S\mu}{\lambda_t(\lambda_t + S\mu)} > 0,
\end{aligned}$$

and

$$\begin{aligned}
T(1,S) - T(1,S+1) &= \frac{\lambda_t^2 + (S\mu)^2 + \lambda_t (S\mu)}{\lambda_t (S\mu) (\lambda_t + S\mu)} - \frac{1}{S\mu} \\
&= \frac{\lambda_t}{S\mu(\lambda_t + S\mu)} > 0.
\end{aligned}$$

Appendix C Proof of Lemma 2

We prove this lemma by induction on p. We can easily see that it holds with $p = 0$ and $p = 1$ due to **Lemma 1**. Assume that it holds with $p = q-1$ for any integer $q \geq 2$. Additionally, assume that when $p = q-1$, the inequality holds in the case of $j = S$ (we show that, under the same inductive assumptions, it also holds when $p = q$ and $j = S$ later in **Proposition 2**). Then, from assumptions we have

$$T(q-1,j) \leq \frac{1}{j\mu} + T(q-1,j-1), \text{ for } 1 \leq j \leq S. \tag{C14}$$

We show that the inequality holds with $p = q$, which indicates that we need to prove that

$$T(q,j) \leq \frac{1}{j\mu} + T(q,j-1), \text{ for } 1 \leq j \leq S-1. \tag{C15}$$

Assume there exists $1 \leq j \leq S-1$ such that

$$T(q,j) > \frac{1}{j\mu} + T(q,j-1). \tag{C16}$$

From (2) we have

$$\begin{aligned} T(q,j) &= \frac{1}{\lambda_t + j\mu} + \frac{\lambda_t}{\lambda_t + j\mu} T(q-1,j+1) + \frac{j\mu}{\lambda_t + j\mu} T(q,j-1) \\ &< \frac{1}{\lambda_t + j\mu} + \frac{\lambda_t}{\lambda_t + j\mu} T(q-1,j+1) + \frac{j\mu}{\lambda_t + j\mu} \left(T(q,j) - \frac{1}{j\mu} \right) \\ &= \frac{\lambda_t}{\lambda_t + j\mu} T(q-1,j+1) + \frac{j\mu}{\lambda_t + j\mu} T(q,j), \end{aligned}$$

which is equivalent to

$$T(q,j) < T(q-1,j+1). \tag{C17}$$

On the other hand, we also have

$$T(q-1,j+1) < \frac{1}{(j+1)\mu} + T(q-1,j), \tag{C18}$$

according to the inductive assumption. From (C16), (C17) and (C18), we obtain

$$\frac{1}{j\mu} + T(q,j-1) < \frac{1}{(j+1)\mu} + T(q-1,j+1),$$

which implies

$$T(q,j-1) < T(q-1,j). \tag{C19}$$

Additionally, from (2) we have

$$\begin{aligned} &T(q,j-1) \\ &= \frac{1}{\lambda_t + (j-1)\mu} + \frac{\lambda_t}{\lambda_t + (j-1)\mu} T(q-1,j) + \frac{(j-1)\mu}{\lambda_t + (j-1)\mu} T(q,j-2) \\ &> \frac{1}{\lambda_t + (j-1)\mu} + \frac{\lambda_t}{\lambda_t + (j-1)\mu} T(q,j-1) + \frac{(j-1)\mu}{\lambda_t + (j-1)\mu} T(q,j-2) \end{aligned}$$

(due to (C19)). This implies

$$T(q, j-1) > \frac{1}{(j-1)\mu} + T(q, j-2).$$

By induction on j, it finally implies

$$T(q, 1) > \frac{1}{\mu} + T(q, 0). \tag{C20}$$

However,

$$\begin{aligned}
T(q,1) - T(q,0) &= \left(\frac{1}{\lambda_t + \mu} + \frac{\lambda_t}{\lambda_t + \mu} T(q-1, 2) + \frac{\mu}{\lambda_t + \mu} T(q, 0)\right) - T(q, 0) \\
&= \frac{1}{\lambda_t + \mu} + \frac{\lambda_t}{\lambda_t + \mu} T(q-1, 2) - \frac{\lambda_t}{\lambda_t + \mu} T(q, 0) \\
&= \frac{1}{\lambda_t + \mu} + \frac{\lambda_t}{\lambda_t + \mu} T(q-1, 2) - \frac{\lambda_t}{\lambda_t + \mu} \left(\frac{1}{\lambda_t} + T(q-1, 1)\right) \\
&= \frac{\lambda_t}{\lambda_t + \mu} \left(T(q-1, 2) - T(q-1, 1)\right) \\
&\leq \frac{\lambda_t}{\lambda_t + \mu} \cdot \frac{1}{2\mu} \qquad \text{(due to the inductive assumption)} \\
&< \frac{1}{\mu},
\end{aligned}$$

which contradicts (C20). This indicates that (C15) holds and thus completes the proof.

Appendix D Proof of Lemma 3

We prove this lemma by induction on p. We can easily see that it holds with $p = 0$ and $p = 1$ due to **Lemma 1**. Assume that it holds with $p = q - 1$ for any integer $q \geq 2$. Additionally, assume that when $p = q - 1$, the inequality holds in the case of $j = S$ (we will show that, under the same inductive assumptions, it also holds when $p = q$ and $j = S$ later in **Proposition 2**). Then, from assumptions we have

$$T(q-1, j-1) \leq \frac{1}{\lambda_t} + T(q-1, j), \text{ for } S + 1 \leq j \leq K. \tag{D21}$$

We show that the inequality holds with $p = q$, which indicates that we need to prove that

$$T(q, j-1) \leq \frac{1}{\lambda_t} + T(q, j), \text{ for } S + 2 \leq j \leq K. \tag{D22}$$

First, notice that

$$
\begin{aligned}
&T(q,K-1)-T(q,K)\\
&=\left(\frac{1}{\lambda_t+S\mu}+\frac{\lambda_t}{\lambda_t+S\mu}T(q,K)+\frac{S\mu}{\lambda_t+S\mu}T(q-1,K-2)\right)-T(q,K)\\
&=\frac{1}{\lambda_t+S\mu}-\frac{S\mu}{\lambda_t+S\mu}T(q,K)+\frac{S\mu}{\lambda_t+S\mu}T(q-1,K-2)\\
&=\frac{1}{\lambda_t+S\mu}-\frac{S\mu}{\lambda_t+S\mu}\left(\frac{1}{S\mu}+T(q-1,K-1)\right)+\frac{S\mu}{\lambda_t+S\mu}T(q-1,K-2)\\
&=\frac{S\mu}{\lambda_t+S\mu}\left(T(q-1,K-2)-T(q-1,K-1)\right)\\
&\le\frac{S\mu}{\lambda_t+S\mu}.\frac{1}{\lambda_t}\text{(due to the inductive assumptions)}\\
&<\frac{1}{\lambda_t},
\end{aligned}
$$

which indicates that (D22) holds with $j = K$. Now, we make an inductive assumption on j; and for any $S+2 \le j \le K-1$, consider the following

$$
\begin{aligned}
&T(q,j-1)-T(q,j)\\
&=\left(\frac{1}{\lambda_t+S\mu}+\frac{\lambda_t}{\lambda_t+S\mu}T(q,j)+\frac{S\mu}{\lambda_t+S\mu}T(q-1,j-2)\right)\\
&-\left(\frac{1}{\lambda_t+S\mu}+\frac{\lambda_t}{\lambda_t+S\mu}T(q,j+1)+\frac{S\mu}{\lambda_t+S\mu}T(q-1,j-1)\right)\\
&=\frac{\lambda_t}{\lambda_t+S\mu}\left(T(q,j)-T(q,j+1)\right)+\frac{S\mu}{\lambda_t+S\mu}\left(T(q-1,j-2)-T(q-1,j-1)\right)\\
&\le\frac{\lambda_t}{\lambda_t+S\mu}.\frac{1}{\lambda_t}+\frac{S\mu}{\lambda_t+S\mu}.\frac{1}{\lambda_t}\text{ (due to the inductive assumptions)}\\
&=\frac{1}{\lambda_t}.
\end{aligned}
$$

Appendix E Proof of Proposition 2

We prove **Proposition 2** by induction on p. The statement is equivalent to the following inequalities.

$$T(p,j)\le T(p,j+1),\text{for } 0\le j\le S-1, \tag{E23}$$

and

$$T(p,j)\ge T(p,j+1),\text{for } S\le j\le K-1. \tag{E24}$$

We already showed that (E23) and (E24) hold with $p = 0$ and $p = 1$ in **Lemma 1**. Assuming that (E23) and (E24) hold with $p = q-1$ for any integer $q \ge 2$, which indicates that

$$T(q-1,j) \le T(q-1,j+1), \text{for } 0 \le j \le S-1,$$

and

$$T(q-1,j) \ge T(q-1,j+1), \text{for } S \le j \le K-1.$$

We show that it holds with $p = q$, which indicates that we need to prove that

$$T(q,j) \le T(q,j+1), \text{for } 0 \le j \le S-1,$$

and

$$T(q,j) \ge T(q,j+1), \text{for } S \le j \le K-1.$$

by considering the following 5 cases.

- When $j = 0$, consider the following

$$\begin{aligned} T(q,1) - T(q,0) &= \left(\frac{1}{\lambda_t+\mu} + \frac{\lambda_t}{\lambda_t+\mu}T(q-1,2) + \frac{\mu}{\lambda_t+\mu}T(q,0)\right) - T(q,0) \\ &= \frac{1}{\lambda_t+\mu} + \frac{\lambda_t}{\lambda_t+\mu}T(q-1,2) - \frac{\lambda_t}{\lambda_t+\mu}T(q,0) \\ &= \frac{1}{\lambda_t+\mu} + \frac{\lambda_t}{\lambda_t+\mu}T(q-1,2) - \frac{\lambda_t}{\lambda_t+\mu}\left(\frac{1}{\lambda_t} + T(q-1,1)\right) \\ &= \frac{\lambda_t}{\lambda_t+\mu}\left(T(q-1,2) - T(q-1,1)\right) \\ &\ge 0 \qquad \text{(due to the inductive assumption)}, \end{aligned}$$

  which indicates that

$$T(q,0) \le T(q,1). \tag{E25}$$

- When $1 \le j \le S-2$, from (2) we have

$$T(q,j) = \frac{1}{\lambda_t+j\mu} + \frac{\lambda_t}{\lambda_t+j\mu}T(q-1,j+1) + \frac{j\mu}{\lambda_t+j\mu}T(q,j-1), \tag{E26}$$

  and

$$\begin{aligned} T(q,j+1) = &\frac{1}{\lambda_t+(j+1)\mu} + \frac{\lambda_t}{\lambda_t+(j+1)\mu}T(q-1,j+2) \\ &+ \frac{(j+1)\mu}{\lambda_t+(j+1)\mu}T(q,j). \end{aligned} \tag{E27}$$

  We prove $T(q,j) \le T(q,j+1)$ by contradiction. Assuming $\exists j,\ T(q,j) > T(q,j+1)$, combining with (E27) we have

$$T(q,j) > \frac{1}{\lambda_t+(j+1)\mu} + \frac{\lambda_t}{\lambda_t+(j+1)\mu}T(q-1,j+2) + \frac{(j+1)\mu}{\lambda_t+(j+1)\mu}T(q,j),$$

which is equivalent to

$$T(q,j) > \frac{1}{\lambda_t} + T(q-1, j+2).$$

However, we also have $T(q-1, j+2) \geq T(q-1, j+1)$ (due to the inductive assumption), so

$$T(q,j) > \frac{1}{\lambda_t} + T(q-1, j+1). \tag{E28}$$

From (E26) and (E28) we obtain

$$T(q,j) < \frac{1}{\lambda_t + j\mu} + \frac{\lambda_t}{\lambda_t + j\mu}\left(T(q,j) - \frac{1}{\lambda_t}\right) + \frac{j\mu}{\lambda_t + j\mu} T(q, j-1),$$

which is equivalent to

$$T(q,j) < T(q, j-1).$$

By induction on j (by repeating the same procedure), it finally implies

$$T(q,1) < T(q,0),$$

which contradicts (E25) that is proved above. This contradiction indicates that

$$T(p,j) \geq T(p, j+1), \text{for } 1 \leq j \leq S-2. \tag{E29}$$

- When $j = K-1$, consider the following

$$\begin{aligned}
&T(q, K-1) - T(q,K) \\
&= \left(\frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(q,K) + \frac{S\mu}{\lambda_t + S\mu} T(q-1, K-2)\right) - T(q,K) \\
&= \frac{1}{\lambda_t + S\mu} + \frac{S\mu}{\lambda_t + S\mu} T(q-1, K-2) - \frac{S\mu}{\lambda_t + S\mu} T(q,K) \\
&= \frac{1}{\lambda_t + \mu} + \frac{S\mu}{\lambda_t + S\mu} T(q-1, K-2) - \frac{S\mu}{\lambda_t + \mu}\left(\frac{1}{S\mu} + T(q-1, K-1)\right) \\
&= \frac{S\mu}{\lambda_t + S\mu}\left(T(q-1, K-2) - T(q-1, K-1)\right) \\
&\geq 0 \qquad \text{(due to the inductive assumption)},
\end{aligned}$$

which indicates that

$$T(q, K-1) \geq T(q,K). \tag{E30}$$

- When $S+1 \leq j \leq K-2$, from (2) we have

$$T(q,j) = \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(q, j+1) + \frac{S\mu}{\lambda_t + S\mu} T(q-1, j-1), \tag{E31}$$

and

$$T(q, j+1) = \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(q, j+2) + \frac{S\mu}{\lambda_t + S\mu} T(q-1, j). \quad \text{(E32)}$$

Due to the inductive assumption, we have $T(q-1, j-1) \geq T(q-1, j)$; therefore, due to (E30), the inequality $T(q, j) \geq T(q, j+1)$ holds for $j = K-2$. By induction on j, we obtain

$$T(q, j) \geq T(q, j+1) \text{ for } S+1 \leq j \leq K-2.$$

Next, we prove that $T(q, S) \geq T(q, S+1)$. From (2) we have

$$T(q, S) = \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(q, S+1) + \frac{S\mu}{\lambda_t + S\mu} T(q, S-1), \quad \text{(E33)}$$

and

$$T(q, S+1) = \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(q, S+2) + \frac{S\mu}{\lambda_t + S\mu} T(q-1, S).$$

To prove the above inequality, we show that

$$T(q, S-1) \geq T(q-1, S). \quad \text{(E34)}$$

From (2) we have

$$\begin{aligned}
&T(q, S-1) \\
&= \frac{1}{\lambda_t + (S-1)\mu} + \frac{\lambda_t}{\lambda_t + (S-1)\mu} T(q-1, S) + \frac{(S-1)\mu}{\lambda_t + (S-1)\mu} T(q, S-2) \\
&\geq \frac{1}{\lambda_t + (S-1)\mu} + \frac{\lambda_t}{\lambda_t + (S-1)\mu} T(q-1, S) \\
&+ \frac{(S-1)\mu}{\lambda_t + (S-1)\mu} \left(T(q, S-1) - \frac{1}{(S-1)\mu} \right) \\
&\text{(due to } \textbf{Lemma 2}\text{)} \\
&= \frac{\lambda_t}{\lambda_t + (S-1)\mu} T(q-1, S) + \frac{(S-1)\mu}{\lambda_t + (S-1)\mu} T(q, S-1),
\end{aligned}$$

which implies that (E34) is true. Therefore,

$$T(q, S) \geq T(q, S+1). \quad \text{(E35)}$$

From (E33) and (E35), we have

$$T(q, S) \leq \frac{1}{\lambda_t + S\mu} + \frac{\lambda_t}{\lambda_t + S\mu} T(q, S) + \frac{S\mu}{\lambda_t + S\mu} T(q, S-1),$$

which implies

$$T(q,S-1)+\frac{1}{S\mu}\geq T(q,S),$$

and this also completes the proof of **Lemma 2**.

- Finally, we prove that $T(q,S)\geq T(q,S-1)$. To show the above equality, first note that

$$\begin{aligned}
&T(q,S-1)\\
&=\frac{1}{\lambda_t+(S-1)\mu}+\frac{\lambda_t}{\lambda_t+(S-1)\mu}T(q-1,S)+\frac{(S-1)\mu}{\lambda_t+(S-1)\mu}T(q,S-2)\\
&\leq\frac{1}{\lambda_t+(S-1)\mu}+\frac{\lambda_t}{\lambda_t+(S-1)\mu}T(q-1,S)+\frac{(S-1)\mu}{\lambda_t+(S-1)\mu}T(q,S-1),
\end{aligned}$$

due to (E29), and this implies

$$T(q,S-1)-T(q-1,S)\leq\frac{1}{\lambda_t}. \tag{E36}$$

Now, due to (E36), **Lemma 3** and the inductive assumptions, we have

$$\begin{aligned}
&T(q,S)-T(q,S+1)\\
&=\left(\frac{1}{\lambda_t+S\mu}+\frac{\lambda_t}{\lambda_t+S\mu}T(q,S+1)+\frac{S\mu}{\lambda_t+S\mu}T(q,S-1)\right)\\
&-\left(\frac{1}{\lambda_t+S\mu}+\frac{\lambda_t}{\lambda_t+S\mu}T(q,S+2)+\frac{S\mu}{\lambda_t+S\mu}T(q-1,S)\right)\\
&=\frac{\lambda_t}{\lambda_t+S\mu}\left(T(q,S+1)-T(q,S+2)\right)+\frac{S\mu}{\lambda_t+S\mu}\left(T(q,S-1)-T(q-1,S)\right)\\
&\leq\frac{\lambda_t}{\lambda_t+S\mu}.\frac{1}{\lambda_t}+\frac{S\mu}{\lambda_t+S\mu}.\frac{1}{\lambda_t}\\
&=\frac{1}{\lambda_t},
\end{aligned}$$

which indicates that

$$T(q,S)\leq\frac{1}{\lambda_t}+T(q,S+1). \tag{E37}$$

(Note that this conclusion also completes the proof of **Lemma 3**).
Now, due to (E37), we have

$$\begin{aligned}
T(q,S)&=\frac{1}{\lambda_t+S\mu}+\frac{\lambda_t}{\lambda_t+S\mu}T(q,S+1)+\frac{S\mu}{\lambda_t+S\mu}T(q,S-1)\\
&\geq\frac{1}{\lambda_t+S\mu}+\frac{\lambda_t}{\lambda_t+S\mu}\left(T(q,S)-\frac{1}{\lambda_t}\right)+\frac{S\mu}{\lambda_t+S\mu}T(q,S-1)
\end{aligned}$$

$$= \frac{\lambda_t}{\lambda_t + S\mu} T(q, S) + \frac{S\mu}{\lambda_t + S\mu} T(q, S-1),$$

which implies $T(q, S) \geq T(q, S-1)$.

Appendix F Proof of Proposition 3

First we prove that

$$T(p, 0) \geq \frac{p}{\lambda_t}, \tag{F38}$$

for all $p = 1, 2, \ldots$.

This inequality holds for $p = 1$ because $T(1, 0) = \frac{1}{\lambda_t}$. Assume that it also holds for $p = q \geq 1$, indicating that $T(q, 0) \geq \frac{q}{\lambda_t}$. We have

$$\begin{aligned} T(q+1, 0) &= \frac{1}{\lambda_t} + T(q, 1) \\ &\geq \frac{1}{\lambda_t} + T(q, 0) \qquad \text{(due to } \textbf{Proposition 2}\text{)} \\ &\geq \frac{q+1}{\lambda_t}. \end{aligned}$$

Therefore, by induction on p, we obtain that (F38) is true. $\lim_{p \to +\infty} \frac{p}{\lambda_t} = +\infty$, which implies

$$\lim_{p \to +\infty} T(p, 0) = +\infty.$$

By induction on j using formula (2), we can easily obtain

$$\lim_{p \to +\infty} T(p, j) = +\infty.$$

for all $j = 0, 1, 2, ..., K$.

References

[1] Naor, P.: The regulation of queue size by levying tolls. Econometrica **37**(1), 15–24 (1969)

[2] Edelson, N.M., Hildebrand, K.: Congestion tolls for Poisson queueing processes. Econometrica **43**(1), 81–92 (1975)

[3] Altman, E., Hassin, R.: Non-threshold equilibrium for customers joining an M/G/1 queue. In: Proceedings of 10th International Symposium on Dynamic Game and Applications (2002)

[4] Kerner, Y.: Equilibrium joining probabilities for an M/G/1 queue. Games. Econ. Behav. **71**(2), 521–526 (2011)

[5] Cripps, M.W., Thomas, C.D.: Strategic experimentation in queues. Theor. Econ. **14**(2), 647–708 (2019)

[6] Debo, L., Veeraraghavan, S.: Equilibrium in queues under unknown service times and service value. Oper. Res. **62**(1), 38–57 (2014)

[7] Brooms, A.C., Collins, E.J.: Stochastic order results and equilibrium joining rules for the Bernoulli feedback queue. Working Paper 1305, Department of Economics, Mathematics & Statistics, Birkbeck (2013)

[8] Fackrell, M., Taylor, P., Wang, J.: Strategic customer behavior in an M/M/1 feedback queue. Queueing Syst. **97**, 223–259 (2021)

[9] Altmann, J., Daanen, H., Oliver, H., Suarez, A.S.-B.: How to market-manage a QoS network. In: Proceedings Twenty-first Annual Joint Conference of the IEEE Computer and Communications Societies (2002)

[10] Wang, J., Cui, S., Wang, Z.: Equilibrium strategies in M/M/1 priority queues with balking. Prod. Oper. Manag. **28**(1), 43–62 (2018)

[11] Hassin, R.: Rational Queueing, 1st edn. Chapman and Hall/CRC, New York (2016)

[12] Hassin, R., Haviv, M.: To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems. Kluwer Academic Publishers, Boston (2003)

[13] Altman, E., Shimkin, N.: Individual equilibrium and learning in processor sharing systems. Oper. Res. **46**(6), 776–784 (1998)

[14] Hassin, R.: On the optimality of first come last served queues. Econometrica **53**(1), 201–202 (1985)

[15] Legros, B.: Agents' self-routing for blended operations to balance inbound and outbound services. Prod. Oper. Manag. (2021)

[16] Ghosh, S., Hassin, R.: Inefficiency in stochastic queueing systems with strategic customers. Eur. J. Oper. Res., **295**(1), 1–11 (2021)

[17] Cheng, M.: Sharing economy: A review and agenda for future research. Int. J. Hosp. Manag. **57**(1), 60–70 (2016)

[18] Jacob, J., Roet-Green, R.: Ride solo or pool: Designing pice-service menus for a ride-sharing platform. Eur. J. Oper. Res. Eur. J. Oper. Res., **295**(3), 1008–1024 (2021)

[19] Crescenzo, A.D., Giorno, V., Kumar, B.K., Nobile, A.G.: A double-ended queue with catastrophes and repairs, and a jump-diffusion approximation. Methodol. Comput. Appl. Probab. **14**, 937–954 (2012)

[20] Economou, A., Kanta, S.: Equilibrium balking strategies in the observable single-server queue with breakdowns and repairs. Oper. Res. Lett. **36**(6), 696–699 (2008)

[21] Boxma, O.J., David, I., Perry, D., Stadje, W.: A new look at organ transplantation models and double matching queues. Probab. Eng. Inform. Sci. **25**(2), 135–155 (2011)

[22] Elalouf, A., Perlman, Y., Yechiali, U.: A double-ended queueing model for dynamic allocation of live organs based on a best-fit criterion. Appl. Math. Model. **60**, 179–191 (2018)

[23] Wong, K.I., Wong, S.C., Bell, M.G.H., Yang, H.: Modeling the bilateral micro-searching behavior for urban taxi services using the absorbing Markov chain approach. J. Adv. Transp. **39**(1), 81–104 (2010)

[24] Shi, Y., Lian, Z.: Optimization and strategic behavior in a passenger-taxi service system. Eur. J. Oper. Res. **249**(3), 1024–1032 (2016)

[25] Wang, F., Wang, J., Zhang, Z.G.: Strategic behavior and social optimization in a double-ended queue with gated policy. Comput. Ind. Eng. **114**, 264–273 (2017)

[26] Jiang, T., Chai, X., Liua, L., Lva, J., Ammar, S.I.: Optimal pricing and service capacity management for a matching queue problem with loss-averse customers. Optim. **70**, 2169–2192 (2021)

[27] Wang, Y., Liu, Z.: Equilibrium and optimization in a double-ended queueing system with dynamic control. J. Adv. Transp. **2019** (2019)

[28] Diamant, A., Baron, O.: Double-sided matching queues: Priority and impatient customers. Oper. Res. Lett. **47**(3), 219–224 (2019)

[29] Shi, Y., Lian, Z.: Equilibrium strategies and optimal control for a double-ended queue. Asia-Pac. J. Oper. Res. **33**(3) (2016)

[30] Wu, H., He, Q.-M.: Double-sided queues with marked Markovian arrival processes and abandonment. Stochastic Models **37**(1), 23–58 (2020)

[31] Shi, Y., Lian, Z., Shang, W.: Study of a passenger-taxi queueing system with nonzero matching time. In: Proceedings of the 12th International Conference on Service Systems and Service Management (ICSSSM), pp. 1–5. IEEE, (2015).

[32] Wang, Z., Yang, C., Liu, L., Zhao, Y.Q.: Equilibrium and Socially optimal of a double-sided queueing system with two-mass point matching time (2021)

[33] Makino, Y., Phung-Duc, T.: Queueing analysis of taxi stand considering boarding time. In: Proceedings of The 2019 International Symposium on Nonlinear Theory and Its Applications (NOLTA 2019), pp. 32–35 (2019)