Self, Social and Monopoly Optimization in Observable Queues

Refael Hassin
Department of Statistics and Operations Research,
Tel Aviv University
Tel Aviv, Israel
hassin@post.tau.ac.il

Ran I. Snitkovsky
Department of Statistics and Operations Research,
Tel Aviv University
Tel Aviv, Israel
ransnit@gmail.com

ABSTRACT

Naor's [8] celebrated paper studies customer decisions in an observable M/M/1 queue where customers utility from joining the system is is a linear decreasing function of the joined position in queue. Naor derives the optimal threshold strategies for the individual, social planner and monopoly. The optimal threshold imposed by a monopoly is not greater than the socially optimal threshold, which is not greater than the individual's threshold. Studies show that this triangular relation holds in a more general setup where the utility function is not necessarily linear. Many of these extensions share common features. We point out conditions that imply the aforementioned result, and apply them to a new model motivated by order-driven markets. In the new model, customers choose between joining and balking when they might be forced to abandon the system before service completion, and the expected value of joining depends on the service completion probability, which is not linear in the observed queue size.

CCS CONCEPTS

• **Mathematics of computing** → **Queueing theory**; **Markov processes**; *Discrete optimization*; • **Applied computing** → *Multi-criterion optimization and decision-making*;

KEYWORDS

Rational Queueing, Observable Queues

ACM Reference Format:
Refael Hassin and Ran I. Snitkovsky. 2017. Self, Social and Monopoly Optimization in Observable Queues. In *VALUETOOLS 2017: 11th EAI International Conference on Performance Evaluation Methodologies and Tools, December 5–7, 2017, Venice, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3150928.3150949

Observable queues refer to systems in which *customers* (agents) arrive at a service station, observe its state (usually the queue size), and based on this information and common knowledge they act to maximize their own welfare. In most cases, the expected benefit of a customer joining the queue is a nonincreasing function of her joining position. Being able to choose join or balk upon arrival, risk neutral rational customers will join the queue as long as their expected total value from joining is nonnegative. Thus, when customers are homogeneous and the total value is monotone decreasing in the queue length, they apply a symmetric threshold joining strategy – join only in positions which are less than some predetermined threshold. It also implies that the joining of a customer induces negative externalities on other customers – by joining, in comparison to balking, one increases the expected queue length observed by future customers, hence, reducing their net utility.

Naor [8] studied the first model of risk neutral customers in an observable queue, considering customers choosing joining or balking upon arriving at an M/M/1 service station. Naor originally considered the total value as a function of waiting time, and not the joined position, what seems to be more natural when the loss is incurred directly by waiting. There, customers total value consists of some fixed reward, minus waiting expense, which is assumed to be linear in waiting time. Nevertheless, since a customer's decision is only made once, based solely on the observed queue length at the moment of arrival, it only relies on the expected total value given this queue length (or position).

Since in Naor's model the service duration of all waiting customers and the residual service of the one currently served are identically distributed, the expected total value of a joining customer is a decreasing linear function of her position. In some extensions of Naor's model the cost is not a linear function of the waiting time – examples we refer to later are given in [10] and [11] and a survey can be found in [3]§2.1. Indeed, in these cases the expected total value of customers as a function of position need not be linear. In this paper, specifically in Lemma 2.1 and Lemma 2.2 below, we discuss the relation between the value when expressed as a function of the waiting time, and the expected value when expressed as a function of position.

Naor defines three different threshold strategies: The first is the individually optimal (or *equilibrium*[1]) threshold, n_e, which is the threshold followed by customers when each of them joins if and only if her expected value from joining is nonnegative. The second is the *socially optimal* threshold, n_o, which is the threshold that maximizes aggregate social welfare per unit time. The third strategy is derived as follows – consider a toll-collecting profit-maximizing agency (or *monopoly*) which is completely divorced from the interest of customers. This agency seeks to impose a fixed toll which maximizes the rate of payments to the server. The value for a customer is now the reward minus waiting expense minus the toll, and it is assumed that customers behave strategically. Therefore,

[1] in dominating strategies

VALUETOOLS 2017, December 5–7, 2017, Venice, Italy

ACM ISBN 978-1-4503-6346-4/17/12…$15.00
https://doi.org/10.1145/3150928.3150949

the toll chosen by the agency uniquely determines a threshold joining strategy of customers. Moreover, the agency will choose the optimal toll such that for the threshold n_m induced, the customer joining in position $n_m - 1$ will be indifferent between joining and balking.

The problem of finding n_e in Naor's model is relatively simple and depends on the expected profit of a customer given her position in the queue. Solving for n_o and n_m, however, is not as immediate, because it also depends on the process of arrivals of future customers to the system. In his paper, Naor showed that $n_m \le n_o \le n_e$. Knudsen [5] generalizes [8], showing that the relation $n_m \le n_o \le n_e$ holds when the system is an M/M/s and the expected value of a joining customer is nonincreasing and concave in the joined position.

Naor [8] and Knudsen [5] assumed both the time between arrivals of successive customers and services are independent and memoryless. It follows that even when customers are allowed to renege their decision after joining, they will never use this option, also when they act to maximize social welfare. If customers are allowed to renege and inter-arrival times are not memoryless, then reneging may improve social welfare. Mendelson and Yechiali in [7] define the *conditional acceptence* strategy, which is a threshold strategy that allows the reneging of the last customer in the queue due to a preset rule. They show that this refinement may lead to a better admission policy in terms of social welfare than the standard threshold strategy. In the following papers, as well as in the present paper, it is assumed that reneging of any customer is prohibited: Yechiali in [12] shows how to compute n_o in a GI/M/1 queue and later in [13] generalizes the results for a GI/M/s. Simonovitz [9] generalizes [8] as well (but does not generalize [5]), showing that $n_m \le n_o \le n_e$ holds when the system is a GI/M/s and the net utility is nonincreasing and linear with respect to the queue position.

While the inequality $n_o \le n_e$ in Naor's model and in its extensions is fairly intuitive and follows an externalities-based argument (Proposition 1.1), the relation $n_m \le n_o$, when holds, cannot be easily justified. Following [8], [5] and [9], we introduce conditions implying the relation $n_m \le n_o$ (Proposition 1.3) that are strictly more general than both [5] and [9] (and therefore [8] as well), in two aspects: First, they apply for some models where the utility of a customer is not concave. Second, they do not require that the arrival processes will be a renewal processes. We demonstrate how to use the result in some concrete models, one of which is a model of customers arriving at an observable queue with the possibility that they will have to abandon before service is completed. This model is inspired by a non-strategic model suggested by Garman [2], where traders arrive at an order-driven market and place bidding orders lasting in the market for some stochastic 'lifetime'.

Some extensions of [8] deal with heterogeneous customers, thus, the joining strategy of customers need not be symmetric. Larsen [6] studies Naor's model when customers differ by service values, assuming the reward is uniformly distributed with an upper bound b. For the case where b is smaller than the expected cost of joining when there is a customer in service, Larsen shows that the profit maximizing fee is greater than the social-welfare maximizing fee. In contrast, Edelson and Hildebrand [1] show that this property does not necessarily hold if customers differ both by time and service values. A survey of extensions for Naor's model with heterogeneous customers is given by Hassin and Haviv [4]§2.5.

The contribution of the paper

The novelty of this paper is demonstrated in the following results:

- In §1 we present the general model, and derive necessary conditions for the relation $n_m \le n_o \le n_e$.
- In §2.1 we apply those conditions in a G/M/s system with concave utility. We prove that when the cost as a function of time is convex, then the expected utility as a function of queue length is concave, and use this result in analyzing the mentioned model.
- In §2.2 we introduce and analyze the Abandonment Model which is motivated by order-driven markets. Using the results obtained in §1 we show that in this model, $n_m \le n_o \le n_e$ holds.

Finally, in §3 we summarize our main results and discuss additional future research directions.

1 GENERAL UTILITY MODEL

Consider for a start a G/GI/s system, where s is possibly infinite. Later, in §2, we demonstrate how the results derived here for the general model apply for rather more concrete models. Let $u(k)$ be the utility for a customer who joins position k in the system (by *position* we refer to the number of customers in the system a moment after that customer has joined). We assume that customers are homogeneous (i.e., that they all share the same utility function), risk neutral, and that $u(k)$ is monotone nonincreasing over the domain of natural numbers, i.e.,

$$u(1) \ge u(2) \ge u(3) \ge \cdots,$$

and that $u(1) > 0$.

Upon arrival to a queue with $k-1$ customers, the customer either joins in position k or balks. We assume, without loss of generality, that balking customers receive zero utility. There exists an integer threshold $n \ge 1$ (possibly infinite) such that the threshold strategy n (that is "join if and only if the number of customers observed is $n-1$ or less") is the individually optimal strategy among customers. Consider the system in steady state and let the random variable $Q^{(n)}$ denote the number of customers in the system (the state) a moment before an arrival, when the entire population follow the threshold strategy n. Note that when customers follow threshold n, the only recurrent states are the set $\{0, 1, \ldots, n\}$, and therefore the system is stable for every n. The customer's utility, $S^{(n)}$, a random variable depending on $Q^{(n)}$, turns to be

$$S^{(n)} = u\left(Q^{(n)} + 1\right) \cdot \mathbf{1}_{\{Q^{(n)} < n\}}, \tag{1}$$

where the random variable $\mathbf{1}_{\{Q^{(n)} < n\}}$ is the indicator function of the event $\{Q^{(n)} < n\}$. When the service provider is a non-discriminating monopoly, its revenue per customer, $M^{(n)}$, is the toll levied when a customer joins and zero otherwise. Customers join as long as the toll is not larger than their expected revenue (reward minus time expense). As explained by [4], the optimal toll levied by the monopoly is either of the form $u(k)$ for some positive integer k, or $lim_{k\to\infty} u(k)$ (otherwise it can increase profit without changing the admission rate by increasing the toll). In other words,

the monopoly collects the same admission fee from every joining customer, which is equal to the utility of the customer who enters in position n. Therefore we define a random variable

$$M^{(n)} = u(n) \cdot \mathbf{1}_{\{Q^{(n)} < n\}}. \tag{2}$$

Clearly, since $u(n)$ is monotone nonincreasing, $S^{(n)} \geq M^{(n)}$ for every realization of $Q^{(n)}$.

The social objective function and the monopoly profit function, as functions of the threshold n, are $\lambda \mathrm{E}\left(S^{(n)}\right)$ and $\lambda \mathrm{E}\left(M^{(n)}\right)$, respectively. We denote

$$n_e = \max_{n\in\mathbb{N}}\{n \mid u(n) > 0\}, \quad n_o = \arg\max_{n\in\mathbb{N}} \mathrm{E}\left(S^{(n)}\right),$$
$$n_m = \arg\max_{n\in\mathbb{N}} \mathrm{E}\left(M^{(n)}\right), \tag{3}$$

when the maximum is attained, otherwise ∞. When finite, we assume n_o (n_m) is unique, otherwise we simply take the smallest n_o (similarly, n_m) such that the function $\mathrm{E}\left(S^{(n)}\right)$ (similarly, $\mathrm{E}\left(M^{(n)}\right)$) is maximal.

Proposition 1.1. $n_o \leq n_e$.

Proof. Suppose that a tagged customer observes n_e or more customers in the system upon arrival. If she joins the system, she will suffer negative utility. Moreover, it may only reduce the utility of future customers who arrive to the system. Customers who arrived before this tagged customer are not effected by her decision. Thus, overall, this customer's joining causes a strict decrease in social welfare. Therefore, when the state is n_e or more, the social planner should not let customers enter, meaning $n_o \leq n_e$. □

Proposition 1.2. $n_m \leq n_e$.

Proof. By the definition of n_e, $u(k) \leq 0$ for every $k > n_e$. This means that if the monopoly chooses a threshold $k > n_e$ it would suffer nonpositive expected revenue which clearly is not optimal. □

Theorem 1.1 (Knudsen [5]§6, Theorem 2). *If the system is an M/M/s and $\{u(k)\}_{k=1}^{\infty}$ is a concave sequence, then $n_m \leq n_o \leq n_e$.*

Theorem 1.2 (Simonovits [9]§5, Proposition 2). *If the system is a GI/M/s and $\{u(k)\}_{k=1}^{\infty}$ is a linear sequence, then $n_m \leq n_o \leq n_e$.*

Theorem 1.1 and Theorem 1.2 are neither more nor less general from each other. We later introduce a proposition that generalizes both theses theorems.

Denote $D^{(n)} = S^{(n)} - M^{(n)}$. An interpretation for $D^{(n)}$ is that it represents customers expected benefit in a system with threshold n and admission fee $u(n)$.

Lemma 1.3. *If $\mathrm{E}\left(D^{(n)}\right) \leq \mathrm{E}\left(D^{(n+1)}\right)$ for all $n \in [n_o, n_e - 1]$, then $n_m \leq n_o$.*

Proof. If $n_o = n_e$, it follows immediately by Proposition 1.2 that $n_m \leq n_o$. Suppose that $n_o \neq n_e$, then by Proposition 1.1, $n_o < n_e$, and assume that $\mathrm{E}\left(D^{(n)}\right) \leq \mathrm{E}\left(D^{(n+1)}\right)$ for all $n \in [n_o, n_e - 1]$. Thus, by definition of $D^{(n)}$,

$$\mathrm{E}\left(S^{(n)} - M^{(n)}\right) \leq \mathrm{E}\left(S^{(n+1)} - M^{(n+1)}\right), \quad \forall n \in [n_o, n_e - 1]. \tag{4}$$

Since n_o is the maximum point of $\mathrm{E}\left(S^{(n)}\right)$,

$$\mathrm{E}\left(S^{(n_o)}\right) \geq \mathrm{E}\left(S^{(n)}\right), \quad \forall n \in [n_o, n_e]. \tag{5}$$

Equation (4) implies that $\mathrm{E}\left(S^{(n)} - M^{(n)}\right)$ is a nondecreasing sequence in $n \in [n_o, n_e - 1]$, thus, we have

$$\mathrm{E}\left(S^{(n_o)}\right) - \mathrm{E}\left(M^{(n_o)}\right) \leq \mathrm{E}\left(S^{(n)}\right) - \mathrm{E}\left(M^{(n)}\right), \quad \forall n \in [n_o+1, n_e]. \tag{6}$$

Subtracting (6) from (5) we arrive at

$$\mathrm{E}\left(M^{(n_o)}\right) \geq \mathrm{E}\left(M^{(n)}\right), \quad \forall n \in [n_o + 1, n_e],$$

which means that $n_m \notin [n_o + 1, n_e]$. By Proposition 1.2, $n_m \leq n_e$, concluding that $n_m \leq n_o$. □

We next describe a technique that will allow us to compare the social and the monopoly profits between a system with threshold n and a system with threshold $n + 1$. For this aim, we shall make use of the following assumption: Suppose that for all $n \in \mathbb{N}$ there exists a coupling of the stationary states such that

$$Q^{(n+1)} = Q^{(n)} + \mathbf{1}_{A_n} \tag{7}$$

for some event A_n (see Figure 1). We shall mention in passing that such coupling can be constructed in many general settings, e.g. for the case of memoryless service, by assuming that the server regenerates the residual time of every customer in service with every new arrival to the system.

Note that by taking expected values of the variables in (7),

$$\Pr(A_n) = \mathrm{E}\left(Q^{(n+1)}\right) - \mathrm{E}\left(Q^{(n)}\right).$$

Moreover, $\{Q^{(n+1)} = n + 1\} \Leftrightarrow \{Q^{(n)} = n, A_n\}$, therefore

$$\begin{aligned}\Pr\left(Q^{(n)} < n, A_n\right) &= \Pr(A_n) - \Pr\left(Q^{(n)} = n, A_n\right) \\ &= \mathrm{E}\left(Q^{(n+1)}\right) - \mathrm{E}\left(Q^{(n)}\right) - \Pr\left(Q^{(n+1)} = n + 1\right).\end{aligned}$$

Lemma 1.4. *For $k = 0, 1, \ldots, n$,*

$$\Pr\left(Q^{(n)} = k, A_n\right) = \sum_{i=0}^{k} \left(\Pr\left(Q^{(n)} = i\right) - \Pr\left(Q^{(n+1)} = i\right)\right). \tag{8}$$

Proof. Basic probability implies, from (7),

$$\begin{aligned}&\Pr\left(Q^{(n)} = k, A_n^c\right) \\ &= \Pr\left(Q^{(n+1)} = k\right) - \Pr\left(Q^{(n)} = k - 1, A_n\right), \quad k = 1, \ldots, n,\end{aligned} \tag{9}$$

and clearly,

$$\begin{aligned}&\Pr\left(Q^{(n)} = k, A_n\right) \\ &= \Pr\left(Q^{(n)} = k\right) - \Pr\left(Q^{(n)} = k, A_n^c\right), \quad k = 0, \ldots, n.\end{aligned} \tag{10}$$

Since $\{Q^{(n+1)} = 0\} \Leftrightarrow \{Q^{(n)} = 0, A_n^c\}$, we have, from (10) for $k = 0$,

$$\Pr\left(Q^{(n)} = 0, A_n\right) = \Pr\left(Q^{(n)} = 0\right) - \Pr\left(Q^{(n+1)} = 0\right).$$

Substituting (9) in (10) for $k = 1, \ldots, n$ we get

$$\Pr\left(Q^{(n)} = k, A_n\right) = \sum_{i=0}^{k} \left(\Pr\left(Q^{(n)} = i\right) - \Pr\left(Q^{(n+1)} = i\right)\right).$$

□

Let $u'(k) = u(k+1) - u(k)$. Since $u(k)$ is nonincreasing we have that $u'(k)$ is nonpositive for all k. We have the following Lemma:

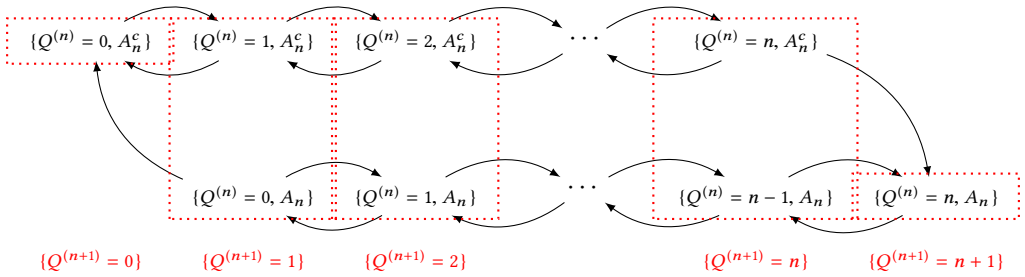

Figure 1: The transition between states in an n-threhold system with the event A_n.

Lemma 1.5.

$$\begin{aligned}&\mathrm{E}\left(D^{(n+1)} - D^{(n)}\right)\\ &= \mathrm{E}\left(u'\left(Q^{(n)}+1\right)\middle| Q^{(n)} < n, A_n\right)\cdot \Pr\left(Q^{(n)} < n, A_n\right)\\ &\quad - u'(n)\cdot \Pr\left(Q^{(n)} < n\right).\end{aligned} \tag{11}$$

Proof. For the ease of notation define $J^{(n)} = \mathbf{1}_{\{Q^{(n)}<n\}}$ and $B^{(n)} = \mathbf{1}_{\{Q^{(n)}=n\}} = 1 - J^{(n)}$. The random variables $J^{(n)}$ and $B^{(n)}$ are the indicators of the event that a new customer joins the queue and the event that she is blocked, respectively, when the threshold is n. Note that using (7) we have $J^{(n+1)} = J^{(n)} + B^{(n)}\cdot \mathbf{1}_{A_n^c} = J^{(n)}\cdot \mathbf{1}_{A_n} + \mathbf{1}_{A_n^c}$. Thus,

$$\begin{aligned}&\mathrm{E}\left(M^{(n+1)} - M^{(n)}\right)\\ &= \mathrm{E}\left(J^{(n+1)}\cdot u(n+1) - J^{(n)}\cdot u(n)\right)\\ &= \mathrm{E}\left(\left(J^{(n)} + B^{(n)}\cdot \mathbf{1}_{A_n^c}\right)u(n+1) - J^{(n)}\cdot u(n)\right)\\ &= u'(n)\cdot \mathrm{E}\left(J^{(n)}\right) + u(n+1)\cdot \mathrm{E}\left(B^{(n)}\cdot \mathbf{1}_{A_n^c}\right).\end{aligned} \tag{12}$$

In addition,

$$\begin{aligned}&\mathrm{E}\left(S^{(n+1)} - S^{(n)}\right)\\ &= \mathrm{E}\left(J^{(n+1)}\cdot u\left(Q^{(n+1)}+1\right) - J^{(n)}\cdot u\left(Q^{(n)}+1\right)\right)\\ &= \mathrm{E}\left(\left(J^{(n)}\cdot \mathbf{1}_{A_n} + \mathbf{1}_{A_n^c}\right)\cdot u\left(Q^{(n)} + \mathbf{1}_{A_n} + 1\right)\right)\\ &\quad - \mathrm{E}\left(\left(J^{(n)}\cdot \mathbf{1}_{A_n} + J^{(n)}\cdot \mathbf{1}_{A_n^c}\right)\cdot u\left(Q^{(n)}+1\right)\right)\\ &= \mathrm{E}\left(J^{(n)}\cdot \mathbf{1}_{A_n}\left(u\left(Q^{(n)}+1+1\right) - u\left(Q^{(n)}+1\right)\right)\right)\\ &\quad + \mathrm{E}\left(\mathbf{1}_{A_n^c}\left(u\left(Q^{(n)}+1\right) - J^{(n)}\cdot u\left(Q^{(n)}+1\right)\right)\right)\\ &= \mathrm{E}\left(J^{(n)}\cdot \mathbf{1}_{A_n}\cdot u'\left(Q^{(n)}+1\right)\right) + \mathrm{E}\left(B^{(n)}\cdot \mathbf{1}_{A_n^c}\cdot u\left(Q^{(n)}+1\right)\right)\\ &= \mathrm{E}\left(J^{(n)}\cdot \mathbf{1}_{A_n}\cdot u'\left(Q^{(n)}+1\right)\right) + u(n+1)\cdot \mathrm{E}\left(B^{(n)}\cdot \mathbf{1}_{A_n^c}\right).\end{aligned} \tag{13}$$

Subtracting (12) from (13) we achieve

$$\begin{aligned}&\mathrm{E}\left(D^{(n+1)} - D^{(n)}\right)\\ &= \mathrm{E}\left(J^{(n)}\cdot \mathbf{1}_{A_n}\cdot u'\left(Q^{(n)}+1\right)\right) - u'(n)\cdot \mathrm{E}\left(J^{(n)}\right)\\ &= \mathrm{E}\left(u'\left(Q^{(n)}+1\right)\middle| Q^{(n)} < n, A_n\right)\cdot \Pr\left(Q^{(n)} < n, A_n\right)\\ &\quad - u'(n)\cdot \Pr\left(Q^{(n)} < n\right).\end{aligned}$$

□

Equation (11) can be explained as follows: Suppose we have two coupled systems with different admission fees, one with threshold $n+1$ and one with threshold n, whose lengths can only differ when the queue in the former system is longer by 1 then in the latter. Then, if a customer is blocked in both systems she would receive zero total surplus in both systems and there will be no difference in her outcome. If she joins both systems and the queues differ by one, the difference in her total surplus would be the difference in utility, $u'\left(Q^{(n)}+1\right)$, minus the difference in fee, $u'(n)$. If she joins both systems and the queues' lengths are equal, the difference in her total utility would be only the difference in fee, $u'(n)$. In the case she joins only the $n+1$ system then she must have joined position $n+1$ in the queue, thus her utility equals the fee and her total surplus is 0, the same as balking from the n system. Other cases are impossible in the probability space that defines the coupling of the systems. Thus, overall, the expected difference in outcome sums up to the right hand side of (11).

Proposition 1.3. *If*

$$u'(n) \le \mathrm{E}\left(u'\left(Q^{(n)}+1\right)\middle| Q^{(n)} < n, A_n\right)\cdot \Pr\left(A_n\middle| Q^{(n)} < n\right), \tag{14}$$

for all $n \in [n_o, n_e]$, then $n_m \le n_o$.

Proof. Suppose (14) holds. This, with (11) implies that

$$\mathrm{E}\left(D^{(n+1)} - D^{(n)}\right) \ge 0, \quad \forall n \in [n_o, n_e]. \tag{15}$$

Thus, by Lemma 1.3, $n_m \le n_o$. □

To the end of the section we show how Proposition 1.3 can be utilized in proving $n_m < n_o$ when more properties of $u(k)$ are given in addition to monotonicity.

Proposition 1.4. *If $u(k)$ is concave, then $n_m \le n_o$.*

Proof. Since $u(k)$ is concave,

$$u'(n) = u(n+1) - u(n) \le u(k+1) - u(k) = u'(k), \quad \forall k \le n.$$

It follows that for all n,

$$u'(n) \le \mathrm{E}\left(u'\left(Q^{(n)}+1\right)\middle| Q^{(n)} < n, A_n\right), \tag{16}$$

because the right-hand side of (16) is a convex combination of terms $u'(k)$ such that $k \le n$. Since $u'(k)$ is also nonpositive we have, for all n,

$$u'(n) \le \mathrm{E}\left(u'\left(Q^{(n)}+1\right)\middle| Q^{(n)} < n, A_n\right)\cdot \Pr\left(A_n\middle| Q^{(n)} < n\right),$$

thus (14) holds, and by Proposition 1.3, $n_m \le n_o$. □

Proposition 1.4 emphasizes that the result established in Proposition 1.3 is more general than both the one presented in [5] and the one in [9].

Proposition 1.5. *If $u(k)$ is convex, and*

$$\frac{u'(n)}{\Pr\left(A_n \middle| Q^{(n)} < n\right)} \le u'(1), \quad \forall n \in [n_o, n_e],$$

then $n_m \le n_o$.

Proof. If $u(k)$ is convex, $u'(1) \le u'(k)$ for all $k \ge 1$. By the assumption,

$$\frac{u'(n)}{\Pr\left(A_n \middle| Q^{(n)} < n\right)} \le u'(1) \le \mathrm{E}\left(u'\left(Q^{(n)}+1\right) \middle| Q^{(n)} < n, A_n\right)$$

for all $n \in [n_o, n_e]$, where the second inequality follows since the right-hand side is a convex combination of $u'(k)$ such that $k \ge 1$. Thus (14) holds, and by Proposition 1.3, $n_m \le n_o$. □

2 EXAMPLES

This section presents examples of concrete models to which Proposition 1.3 applies.

2.1 G/M/s with Convex Waiting-Time Cost

Consider a G/M/s queue with service rates $\mu_1, \mu_2, \ldots, \mu_s$. We discuss both cases of homogeneous and heterogeneous servers. Assume that the utility of service after waiting t units of time in the system is given by $R - c(t)$, where $R > 0$ is some fixed reward and $c(t)$, the cost function, is convex. Customers observe the queue length upon arrival, then decide join or balk and reneging is not allowed.

Prior to analyzing the utility for a customer we shall introduce the following lemmas that we will later use in the example:

Lemma 2.1. *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d non negative random variables, and let $g(x)$ be a convex function. Define $S_n = \sum_{i=1}^{n} X_i$, then the sequence $\{\mathrm{E}(g(S_n))\}_{n=1}^{\infty}$ is convex.*

Proof. We show that $\{\mathrm{E}(g(S_n))\}_{n=1}^{\infty}$ is convex by showing that the sequence of its differences is nondecreasing. Let X_0 be a random variable independent of all $\{X_i\}_{i=1}^{\infty}$ and identically distributed, then, for all $n \in \mathbb{N}$,

$$g(S_{n+1}) - g(S_n) \sim g(X_0 + S_n) - g(S_n). \tag{17}$$

Since g is convex and $S_{n-1} \le S_n$ (with probability 1),

$$g(X_0 + S_{n-1}) - g(S_{n-1}) \le g(X_0 + S_n) - g(S_n), \quad n = 2, 3, \ldots \tag{18}$$

with probability 1. Taking expected values of both sides of (18) we arrive at

$$\begin{aligned}&\mathrm{E}(g(S_n)) - \mathrm{E}(g(S_{n-1}))\\ &= \mathrm{E}(g(X_0 + S_{n-1})) - \mathrm{E}(g(S_{n-1}))\\ &\le \mathrm{E}(g(X_0 + S_n) - \mathrm{E}(g(S_n))) = \mathrm{E}(g(S_{n+1})) - \mathrm{E}(g(S_n)),\end{aligned}$$

for $n = 2, 3, \ldots$, concluding that $\{\mathrm{E}(g(S_n))\}_{n=1}^{\infty}$ is convex. □

Lemma 2.2. *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d non negative random variables, let Y be a random variable independent of all $\{X_i\}_{i=1}^{\infty}$ and let $g(x)$ be a convex function. Define $S_0 = Y$ and $S_n = Y + \sum_{i=1}^{n} X_i$, $n = 1, 2, \ldots$, then the sequence $\{\mathrm{E}(g(S_n))\}_{n=0}^{\infty}$ is convex.*

The proof is similar to that of Lemma 2.1.

2.1.1 Homogeneous Servers. Suppose that $\mu_1 = \mu_2 = \cdots = \mu_s = \mu$, so the system is a standard G/M/s.

Proposition 2.1. *In the observable G/M/s (homogeneous servers) with fixed service reward R and convex time cost function, $n_m \le n_o \le n_e$.*

Proof. Consider a customer who joins the system in position k (i.e., upon arrival she observes $k-1$ customers in the system in total). If $k \le s$, her time spent in the system includes only the service time, which is exponentially distributed with rate μ. If $k > s$, her waiting time in the queue (excluding self service) is Erlang distributed with shape parameter $k - s$ and rate $s\mu$, and adds up with the time spent in service. Thus, conditioned on the queue length, the total time spent in the system, W, is a sum of $(k-s)^+$ i.i.d nonnegative random variables plus an independent random variable, therefore, by Lemma 2.2, $\mathrm{E}(c(W))$ is convex as a function of k. Denote by $u(k)$ the expected utility of that particular customer, then

$$u(k) = R - \mathrm{E}(c(W)). \tag{19}$$

It follows that $\{u(k)\}_{k=1}^{\infty}$ is a nonincreasing and concave sequence. From proposition 1.4, the thresholds n_e, n_o and n_m as defined in (3) sustain $n_m \le n_o \le n_e$. □

Special cases of this model were considered in the following examples:

- Naor [8], considers $s = 1$, Poisson arrivals and a linear cost function.
- Knudsen [5]§7, considers Poisson arrivals and a piecewise linear and convex cost function.
- Simonovits [9] considers general arrival with independent interarrival-times and a linear cost function.
- Sun and Li [10] consider $s = 1$, Poisson arrivals and $c(t) = C \cdot t^m$ for $m = 1, 2, 3$ and $C > 0$. In fact, Proposition 2.1 is valid for every real value $m \ge 1$.
- Wang, Zhang and Zhang [11] consider $s = 1$, Poisson arrivals and

  $$\mathrm{E}(c(W)) = C \cdot \mathrm{E}(W) + A \cdot C^2 \cdot \mathrm{Var}(W).$$

  For some nonnegative constants C, A. By linearity of expectation this is equivalent to

  $$c(t) = C \cdot t + A \cdot C^2 \cdot (t - \mathrm{E}(W))^2$$

  which is a convex function in t.

2.1.2 Heterogeneous Servers. Suppose now that service times are heterogeneous. We assume that waiting costs are incurred only when customers wait in the queue (but not in service).

Proposition 2.2. *In the observable G/M/s heterogeneous servers with fixed service reward R and convex time cost function, where customers pay only for their queueing time (but not for their service time), $n_m \le n_o \le n_e$.*

Proof. Consider a customer who joins the queue in position k (i.e., upon arrival she observes $s + k - 1$ customers in the system in total). Her waiting time in the queue (excluding self service), W_q, conditioned on the queue length, is Erlang distributed with shape

parameter k and rate $\sum_{i=1}^{s}\mu_i$. Therefore, by Lemma 2.1, $\mathrm{E}(c(W_q))$ is convex as a function of k. Denote by $u(k)$ the expected utility of that particular customer, then

$$u(k) = R - \mathrm{E}(c(W_q)).$$

It follows that $\{u(k)\}_{k=1}^{\infty}$ is a nonincreasing and concave sequence. From proposition 1.4, the thresholds n_e, n_o and n_m as defined in (3) sustain $n_m \le n_o \le n_e$. □

One can consider some extensions of the model incorporating service time expenses, although in general, if customers pay for their own service times then it is not necessarily true that the sequence $\{u(k)\}_{k=1}^{\infty}$ is nonincreasing. Depending on the service rates and the admission policy, it may hold that customers would favor waiting in the queue to joining a free server.

2.2 The Abandonment Model

This following model, motivated by applications of order-driven markets, is similar to a one presented by Garman [2].

Consider customers who arrive, following a Poisson process with rate λ, at a single-queue single-server system with exponential service rate μ. Each customer may abandon the system within some amount of time, unknown upon arrival, which is an independent exponential random variable with rate θ (the *abandonment rate*). A customer can leave the system either by service completion or by abandonment, whichever comes first. Each customer pays a fixed admission fee d upon joining, if she completes service before abandoning she receives R, and if she abandons she receives 0 (this is without loss of generality). For simplicity, assume that a customer may abandon the system at any time, even during her own service. We further assume that the service regime is FCFS and that $R/(\mu+\theta) \ge d$, i.e., customers' expected benefit from joining an empty server is positive. Upon arrival, each customer observes the queue length and chooses join or balk.

Consider a customer who joins the system in position k (i.e., this customer joins when there are $k-1$ customers in the system). The probability that she will not abandon before any of these $k-1$ customers leaves is $1-\theta/(\mu+k\theta)$. When there are $k-2$ such customers, this probability becomes $1-\theta/(\mu+(k-1)\theta)$ and so forth. Thus, the probability that this customer will eventually complete service is given by

$$\frac{\mu+(k-1)\theta}{\mu+k\theta}\cdot\frac{\mu+(k-2)\theta}{\mu+(k-1)\theta}\cdots\frac{\mu}{\mu+\theta}=\frac{\mu}{\mu+k\theta}, \tag{20}$$

Let $u(k)$ denote the total expected utility of that customer. Equation (20) implies

$$u(k) = \frac{\mu}{\mu+k\theta}R - d. \tag{21}$$

Note that $\{u(k)\}_{k=1}^{\infty}$ is monotone decreasing and strictly convex. Solving the inequality $u(k) \ge 0$ for k, we get by the definition of n_e in (3) that the pure threshold strategy n_e with

$$n_e = \left\lfloor \frac{\mu}{\theta}\left(\frac{R}{d}-1\right)\right\rfloor$$

is a dominant strategy for all customers and therefore induces equilibrium.

Proposition 2.3. *In the abandonment model described above, $n_m \le n_o \le n_e$.*

Proof. Recall $Q^{(n)}$, the state of the system when the threshold strategy is n. Define the following quantities

$$\beta_0 = 1; \quad \beta_k = \prod_{i=1}^{k}\frac{\lambda}{\mu+i\theta}, \quad k = 1,2,\ldots$$

Since the system is a standard birth-death process (M/M/1 + ∞), solving the steady state equations we have

$$\Pr\left(Q^{(n)}=k\right) = \frac{\beta_k}{\sum_{j=0}^{n}\beta_j}, \quad k=0,1,\ldots,n. \tag{22}$$

Define the event A_n such that $Q^{(n+1)} = Q^{(n)} + \mathbf{1}_{A_n}$. By (8) and (22),

$$\begin{aligned}\Pr\left(Q^{(n)}=k, A_n\right) &= \sum_{i=0}^{k}\left(\frac{\beta_i}{\sum_{j=0}^{n}\beta_j}-\frac{\beta_i}{\sum_{j=0}^{n+1}\beta_j}\right)\\ &= \frac{\beta_{n+1}\sum_{i=0}^{k}\beta_i}{\left(\sum_{j=0}^{n}\beta_j\right)\left(\sum_{j=0}^{n+1}\beta_j\right)}\end{aligned}$$

Let $\{u'(k)\}_{k=1}^{\infty}$ denote the sequence of differences of $\{u(k)\}_{k=1}^{\infty}$. By (21),

$$u'(k) = u(k+1)-u(k) = \frac{-R\mu\theta}{(\mu+k\theta)(\mu+(k+1)\theta)} = \frac{-R\mu\theta}{\lambda^2}\cdot\frac{\beta_{k+1}}{\beta_{k-1}}, \tag{23}$$

for $k = 1,2,\ldots$. We shall show that

$$\begin{aligned}&u'(n)\cdot\Pr\left(Q^{(n)}<n\right)\\ &\le \mathrm{E}\left(u'\left(Q^{(n)}+1\right)\middle| Q^{(n)}<n, A_n\right)\cdot\Pr\left(Q^{(n)}<n, A_n\right),\end{aligned} \tag{24}$$

and then, from Proposition 1.3 we will derive that $n_m \le n_s \le n_e$. First note that from (23) and (22),

$$\begin{aligned}u'(n)\cdot\Pr\left(Q^{(n)}<n\right) &= \frac{-R\mu\theta}{\lambda^2}\cdot\frac{\beta_{n+1}}{\beta_{n-1}}\cdot\frac{\sum_{j=0}^{n-1}\beta_j}{\sum_{j=0}^{n}\beta_j}\\ &= \frac{-R\mu\theta\beta_{n+1}}{\lambda^2\sum_{j=0}^{n}\beta_j}\cdot\frac{1}{\Pr\left(Q^{(n-1)}=n-1\right)}.\end{aligned}$$

Analyzing the right-hand side of (24) we get

$$\begin{aligned}&\mathrm{E}\left(u'\left(Q^{(n)}+1\right)\middle| Q^{(n)}<n, A\right)\cdot\Pr\left(Q^{(n)}<n, A\right)\\ &= \sum_{k=0}^{n-1}u'(k+1)\cdot\Pr\left(Q^{(n)}=k, A\right)\\ &= \sum_{k=0}^{n-1}\frac{-R\mu\theta}{\lambda^2}\cdot\frac{\beta_{k+2}}{\beta_k}\cdot\frac{\beta_{n+1}\sum_{i=0}^{k}\beta_i}{\left(\sum_{j=0}^{n}\beta_j\right)\left(\sum_{j=0}^{n+1}\beta_j\right)}\\ &= \frac{-R\mu\theta\beta_{n+1}}{\lambda^2\sum_{j=0}^{n}\beta_j}\cdot\sum_{k=0}^{n-1}\frac{\beta_{k+2}}{\beta_k}\cdot\frac{\sum_{i=0}^{k}\beta_i}{\sum_{j=0}^{n+1}\beta_j}\\ &= \frac{-R\mu\theta\beta_{n+1}}{\lambda^2\sum_{j=0}^{n}\beta_j}\cdot\sum_{k=0}^{n-1}\frac{\beta_{k+2}}{\sum_{j=0}^{n+1}\beta_j}\cdot\frac{1}{\Pr\left(Q^{(k)}=k\right)}.\end{aligned}$$

To show our aim (24) it suffices to show

$$\sum_{k=0}^{n-1}\frac{\beta_{k+2}}{\sum_{j=0}^{n+1}\beta_j}\cdot\frac{1}{\Pr\left(Q^{(k)}=k\right)} \le \frac{1}{\Pr\left(Q^{(n-1)}=n-1\right)}.$$

Note that $\Pr\left(Q^{(k)} = k\right)$ is positive and monotone decreasing in k, as it represents the blocking probability of a queue with threshold k. Thus, $1/\Pr\left(Q^{(k)} = k\right)$ is monotone increasing. Now,

$$\sum_{k=0}^{n-1} \frac{\beta_{k+2}}{\sum_{j=0}^{n+1} \beta_j} \cdot \frac{1}{\Pr\left(Q^{(k)} = k\right)}$$
$$\leq \sum_{k=0}^{n-1} \frac{\beta_{k+2}}{\sum_{j=0}^{n-1} \beta_{j+2}} \cdot \frac{1}{\Pr\left(Q^{(k)} = k\right)} \leq \frac{1}{\Pr\left(Q^{(n-1)} = n-1\right)},$$

where the first inequality evolves as we decrease the denominator and the second inequality follows as the second term is a convex combination of terms $1/\Pr\left(Q^{(k)} = k\right)$ such that $k \leq n-1$. □

3 CONCLUDING REMARKS

This work deals with the relation between the socially optimal threshold, n_o, and the monopoly threshold, n_m, in queues. The existence of the common relation in queues $n_m \leq n_o$ depends on the structure of customers' value of joining, and needs not necessarily hold in general. We establish sufficient conditions for $n_m \leq n_o$, based on coupling the system when customers' threshold is n with the same system when customers' threshold is $n+1$. When the service is exponential, this can be done by assuming that working servers regenerate their residual service time at the moment of each event (arrival or departure) in the system. A natural question to ask therefore is when the system is a G/M/s (or even an M/M/1), what properties form a necessary and sufficient condition for $n_m \leq n_o$.

The conditions for Proposition 1.3 demand that Equation (14) will hold for all $n \in [n_o, n_e]$. In fact, the interval's upper bound n_e can be substituted with any integer $\bar{n}$ such that $n_m \leq \bar{n}$. We conjecture that this observation can be utilized in showing that $n_o \leq n_m \leq n_e$ in the following *discount* model: Customers arrive at an M/M/1 service station choosing join or balk. A customer whose sojourn time is t receives $R \cdot e^{-\beta t} - d$, for some nonnegative parameters R, d and β. The sequence $\{u(k)\}_{k=1}^{\infty}$ can be then expressed using the moment-generating function of the Erlang(k, μ) distribution.

Another term that is mentioned in the literature and is not addressed in the paper is the monopoly threshold for *collectivistic customers* – the optimal threshold induced by the monopoly when customers cooperate and act as a collective that maximizes expected profit. This term was introduced by Yechiali [12], and studied by Simonovits [9], who shows that for an M/M/1 system with linear waiting costs, this threshold is not larger than the standard monopoly threshold, n_m. Simonovits conjectures that this result also holds in general for GI/M/s with linear waiting costs. We believe that the analytic tools presented in this paper can be exploited in proving (or disproving) that conjecture.

4 ACKNOWLEDGEMENT

The authors express their appreciation to Liron Ravner for his fruitful review of the paper. This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 355/15).

REFERENCES

[1] Noel M. Edelson and David K. Hildebrand. 1975. Congestion tolls for Poisson queueing processes. *Econometrica* 43 (1975), 81–92.
[2] Mark B. Garman. 1976. Market microstructure. *Journal of Financial Economics* 3 (1976), 257–275.
[3] Refael Hassin. 2016. *Rational Queueing*. CRC Press, Boca Raton.
[4] Refael Hassin and Moshe Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston.
[5] Niels Chr. Knudsen. 1972. Individual and Social Optimization in a Multiserver Queue with a General Cost-Benefit Structure. *Econometrica* 40 (1972), 515–528.
[6] Christian Larsen. 1998. Investigating sensitivity and the impact of information on pricing decisions in an M/M/1/∞ queueing model. *International Journal of Production Economics* 56-57 (1998), 365–377.
[7] Haim Mendelson and Uri Yechiali. 1981. Controlling the GI/M/1 queue by conditional acceptance of customers. *European Journal of Operational Research* 7 (1981), 77–85.
[8] Pinhas Naor. 1969. The Regulation of Queue Size by Levying Tolls. *Econometrica* 37 (1969), 15–24.
[9] András Simonovits. 1976. Self-And Social Optimization in Queues. *Studia Scientiarum Mathematicarum Hungarica* 11 (1976), 131–138.
[10] Wei Sun and Shiyong Li. 2012. Customer Threshold Strategies in Observable Queues with Partial Information of Service Time. *Information Computing and Applications* 307 (2012), 456–462.
[11] Jinting Wang, Zhe George Zhang, and Zhengwu Zhang. 2014. Performance analysis of a queue with strategic customers under qudratic utility criterion. *working paper* (2014).
[12] Uri Yechiali. 1971. On optimal balking rules and toll charges in the GI/M/1 queue. *Operations Research* 19 (1971), 349–370.
[13] Uri Yechiali. 1972. Customers' optimal joining rules for the GI/M/s queue. *Menegement Science* 18 (1972), 434–443.