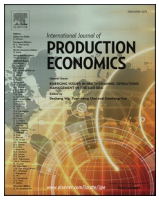
Int. J. Production Economics 228 (2020) 107696

Contents lists available at ScienceDirect

International Journal of Production Economics

journal homepage: http://www.elsevier.com/locate/ijpe



A service system with perishable products where customers are either fastidious or strategic

Gabi Hanukov [a,*], Tal Avinadav [a], Tatyana Chernonog [a], Uri Yechiali [b]

[a] Department of Management, Bar-Ilan University, Ramat Gan, 5290002, Israel
[b] Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv, 6997801, Israel

ARTICLE INFO

Keywords:
Perishable products
Inventory
Queueing
Strategic customers
Game theory

ABSTRACT

Many service systems in the fast food industry consist of two types of customers: fastidious and strategic. A fastidious customer will always join the queue and wait for a fresh product, while a strategic customer, depending on queue length and on stock availability, may either join the queue for fresh products, purchase an inventoried pre-prepared (perishable) product, or balk. When the system is empty, the server produces pre-prepared items up to a pre-determined capacity level. The server may assign different prices to pre-prepared items and to fresh products in order to maximize expected profit, while taking into account revenue from selling food, sojourn and balking costs, capacity costs, and costs associated with food deterioration. We formulate and analyze this stochastic queueing-inventory system and derive its steady-state probabilities using matrix geometric methods. Our economic analysis, which follows a Stackelberg game and Nash equilibrium, shows that the presence of strategic customers is always beneficial for non-strategic customers, and can also be beneficial for the operating server. Moreover, in some cases, the server benefits from charging a higher price for less-fresh (pre-prepared) products than for fresh items, and even when pre-prepared items are offered at a discount, the discount may have a stronger positive effect on non-strategic customers' utility than on strategic customers' utility. Notably, in some cases, the percentage increase in the customers' utility (as compared with the case in which pre-prepared food is not offered) may be even higher than the percentage increase in the server's expected profit, even though the server is the one who controls the decision variables.

1. Introduction

The revenue of the fast food industry amounts to hundreds of billions of dollars per year (National Restaurant Association, 2017). Even a small improvement in the efficiency of this industry has the potential to generate vast savings. The primary objective of a customer who enters a fast food restaurant is to obtain his/her order quickly. Yet, in many cases, faster service may imply lower-quality food. Customers may differ in the extent to which they prioritize rapid service over quality. In particular, many fast food restaurants stock pre-prepared food such as pizza, sandwiches or salad that can be purchased directly from the shelf. A customer who prioritizes rapid service might prefer to purchase pre-prepared items in order to spend as little time as possible waiting in the restaurant. In contrast, a customer who prioritizes quality might be willing to wait as long as required to get freshly prepared food.

In this study, we model and analyze a queueing-inventory system typical of the fast food industry. The system comprises a server and two different types of customers, distinguished according to whether they prioritize fresh food or a brief sojourn time. A customer who prioritizes sojourn time is referred to as "strategic", because we assume that such customers strategically decide whether to queue for fresh food, purchase a pre-prepared food item, or balk. Specifically, when a strategic customer enters a fast food dining establishment and observes a queue of waiting customers, he or she acts in accordance with a double threshold policy as follows: Up to a certain queue length the customer joins the queue and waits for a fresh product. When the queue length exceeds that threshold, the customer prefers to purchase a pre-prepared food item from stock (referred to as a PPS). If no PPSs are available, and the length of the queue does not exceed a second threshold, the customer joins the queue. Otherwise the customer balks. A customer who prioritizes fresh food, in turn, is referred to as "fastidious". Fastidious customers always join the queue to receive a freshly prepared food item, regardless of the

* Corresponding author.
E-mail addresses: german.kanukov@biu.ac.il (G. Hanukov), tal.avinadav@biu.ac.il (T. Avinadav), tatyana.chernonog@biu.ac.il (T. Chernonog), uriy@post.tau.ac.il (U. Yechiali).

https://doi.org/10.1016/j.ijpe.2020.107696
Received 11 October 2018; Received in revised form 23 January 2020; Accepted 25 February 2020
Available online 28 February 2020

queue's length.

Our model incorporates a special feature of the fast food industry: namely, the fact that food products can be prepared in advance and stored on the shelf. Accordingly, this work belongs to a special stream of literature dealing with queueing-inventory systems. In line with recent models developed in this vein (Hanukov et al., 2017, 2018a; b, 2019), we assume that the server prepares PPSs during its idle time. Finally, we assume that PPSs are perishable, meaning that they deteriorate in quality over the course of the time period between preparation and consumption. Thus, inventory is depleted either via demand or via spoilage of PPSs. Both depletion processes are stochastic.

We seek to identify the optimal capacity level of PPSs while considering customers' waiting time, holding costs for inventoried PPSs, and the fact that PPSs might spoil if they remain for too long on the shelf. Another goal is to identify an optimal approach to pricing PPSs, given that customers are sensitive to product freshness and waiting time. In this work, we consider a stationary pricing policy, in which customers do not know the age of the pre-prepared food products in stock (a reasonable assumption for the real-life scenario we consider), but are aware that they are not entirely fresh. We show that it may be economically beneficial for the server to apply a price discrimination policy based on one of the following two options: (i) giving a discount for a PPS in order to encourage strategic customers to purchase them. This option prevents food spoilage and reduces costs related to customers' sojourn time but at the same time reduces revenue; (ii) raising the price of PPSs. This option increases revenue but at the same time increases inventory costs and increases costs related to customers' sojourn time.

The contributions of this paper to the operations management literature can be summarized as follows:

- Being the first to provide a mathematical model that describes service systems with inventoried preliminary services and two types of customers: fastidious and strategic.
- Obtaining closed-form expressions for the rate matrix R, which then can be used to calculate quickly the steady-state probabilities and calculate the system performance measures.
- Finding a complete solution for the Nash equilibrium behavior of strategic customers facing multiple service options—join, purchase a PPS, or balk—and proving that it is a double threshold policy.
- Proving that the optimal price of a PPS is not necessarily cheaper than that of a fresh service and might even be more expensive.
- Finding that the production and deterioration rates of PPSs and the arrival rate of strategic customers do not affect the stability condition of the service system, which is dictated only by the arrival rate of fastidious customers and the server's service rate.

2. Literature review

This work is related to five streams of literature: (i) queueing-inventory systems; (ii) preparation and deterioration times of products; (iii) strategic customers and multi-service options in queueing systems; (iv) inventory management of perishable products; and (v) pricing of perishable products. In what follows we review the relevant literature in each of these streams.

2.1. Queueing-inventory systems

Systems that combine both queueing and inventory have been analytically investigated only in recent years. Works in this stream include studies by Zhao and Lin (2011) and Adacher and Cassandras (2014), who investigated systems in which each customer requires a unit from inventory when being served. Jeganathan et al. (2017) analyzed a perishable inventory system that uses a two-rate service policy within a finite queue under a continuous review (s,Q) ordering policy. Nair et al. (2015) considered a multi-server Markovian queueing model where the servers are considered as an inventory that is replenished according to the standard (s,S) policy. Krishnamoorthy et al. (2015) considered two control policies—(s,Q) and (s,S)—for an M/M/1 queueing system, in which an inventoried item is given to a customer upon service completion with a certain probability. Avşar and Zijm (2014) presented approximate queueing models for capacitated multi-stage inventory systems under base-stock control. Altendorfer and Minner (2015) modeled a production system as an M/M/1 queue with input rates depending on queue length and random customer-required lead time. Chebolu-Subramanian and Gaukler (2015) used queueing theory to analyze product contamination in a multi-stage food supply chain with inventory. Recent articles by Hanukov et al. (2017, 2018a,b, 2019) introduced models for utilizing servers' idle time to produce and store so-called "preliminary services". The current paper combines the use of preliminary services with two types of customers: fastidious and strategic.

2.2. Preparation and deterioration times of products

The system described herein incorporates assumptions with regard to the preparation and deterioration of perishable services. These processes are characterized by uncertainty. In general, there are three possible sources of uncertainty that affect preparation time of services: the server, the production process and raw materials (Hanukov et al., 2018a). To accommodate such uncertainty, prior studies of service systems have assumed that preparation time is exponentially distributed (see, e.g., Benjaafar et al., 2011; Flapper et al., 2012; Iravani et al., 2011; Hanukov et al., 2017). The exponential distribution is also commonly used in the literature to describe the lifetime of perishable products (see, e.g., Berman and Sapna, 2002; Chao et al., 2009; Kouki et al., 2016). According to Cancho et al. (2011), the exponential distribution assumption is considered to provide a simple, elegant and closed-form solution to many problems in lifetime testing and reliability studies. Kouki et al. (2018) claim that the exponential lifetime distribution is suitable for analyzing inventory systems in which product lifetimes are typically short and are rarely long. Examples of such products are food items with no printed expiration date, e.g., pizza, sandwiches or pre-prepared salad that are sold in fast-food restaurants. The current work provides a method for analyzing service systems that produce and sell such products. In line with prior research, we also assume exponential distributions for service preparation time and for product shelf life (see section 3).

2.3. Strategic customers and multi-service options in queueing systems

Models of strategic customers in queueing systems began with Naor's (1969) pioneering work, which investigated a customer's decision of whether to join an observable Markovian queue or balk. This decision took into account both the reward from getting the service and the expected sojourn time. Naor's work has since been extended by numerous others (see reviews by Hassin and Haviv, 2003; and recently by Hassin, 2016). For example, Kerner (2011) presented a full recursive algorithm for computing the mixed Nash equilibrium strategy among strategic customers deciding whether to join an observable queue or not. Boudali and Economou (2012) modeled customers' decisions regarding whether to join a queue or balk in a system with catastrophes, i.e., random events in which all customers are forced to leave the queue; the authors identified both individually-optimal and socially-optimal strategies. Shi and Lian (2016) analyzed the strategic behavior of taxi passengers faced with either observable or unobservable queue lengths. Haviv and Oz (2017) proposed a novel scheme to make customers adopt socially-optimal rather than individually-optimal queue-joining behavior. Bountali and Economou (2017) discussed the effects of two levels of information (observable versus unobservable queue) and service batch size on customers' strategic behavior and on the overall social welfare. For additional related works, we refer the reader to Yechiali

(1971, 1972), Manou et al. (2014), Ziani et al. (2015), Shone et al. (2016), Li et al. (2016a, 2016b), Wang et al. (2017), Hassin and Roet-Green (2017) and Wang et al. (2019).

Given that we consider two different categories of customers—i.e., strategic versus fastidious—studies of queueing/balking behavior among customers with heterogeneous preferences are particularly relevant to our context. Mandelbaum and Yechiali (1983), for example, studied a 'smart' customer's decisions to join, balk or 'wait and see' outside the system in an M/G/1 queue. Guo and Hassin (2012) investigated the join-or-balk decisions of customers with heterogeneous delay sensitivity in a queueing system with vacations, i.e., in which idle servers can leave the system to carry out ancillary work. Yu et al. (2016) leveraged empirical data from a medium-sized call center to analyze the abandonment behavior of customers queueing for call center service. They modeled customer heterogeneity by defining a ratio between the cost of waiting versus the reward of receiving service, which followed a folded normal distribution that varied across different classes of customers. The authors showed that delay announcements directly impact customers' waiting costs. Hu et al. (2017) modeled join-or-balk decisions in a service system with two streams of customers, one informed about real-time delay and the other uninformed, and investigated how the presence of a larger fraction of informed customers affects system throughput and social welfare.

Notably, most of the papers cited above assume that customers choose between two options: joining or balking. In our work, in contrast, we assume that customers have an additional option—to accept lower-quality service (a PPS) that does not entail a wait. Accordingly, our work relates to a stream of studies considering the decisions of queueing customers faced with multiple service options whose utility varies in accordance with customers' (heterogeneous) sensitivity to delay. These include a study by Akan (2012), who formulated a novel fluid model in which the system manager offers a menu of lead times and corresponding prices to arriving customers who differ in their delay sensitivity and their valuations of the service (or product). Afèche and Pavlin (2016) designed a price/lead-time menu and scheduling policy in a queueing system with multiple customer types who differ in their valuations of instant delivery and their delay costs.

2.4. Inventory of perishable products

In considering perishability, we draw from a vast stream of literature on inventory management of perishable items (see, e.g., Avinadav, 2020; Avinadav et al., 2013, 2014, 2017; Avinadav and Arponen, 2009; Berk and Gürler, 2008; Chao et al., 2015; Chen et al., 2014; Chernonog and Avinadav, 2019; Cooper, 2001; Herbon, 2017; Herbon and Khmelnitsky, 2017; Hu et al., 2015; Li et al., 2016a, 2016b; Zhang et al., 2016 and Chernonog, 2020). These studies deal with order quantity at fixed intervals or according to inventory position. For example, Avinadav and Arponen (2009) extended the classical EOQ model for products with a fixed expiry date and a declining demand rate due to a reduction in the quality of the product along time. They proved that the profit function is unimodal, and calculated the maximal average profit per unit time. Hu et al. (2015) formulated and analyzed a dynamic model of inventory and pricing decisions for perishable goods under uncertainty, where every period consists of two phases: clearance phase and regular-sales phase. They showed that a firm should either sell all the leftover units on discount, or dispose them, where the choice depends on the number of leftover units. Chao et al. (2015) developed approximation algorithms with worst-case performance guarantees for stochastic periodic-review perishable inventory systems for both backlogging and lost-sales models. The authors constructed two policies and demonstrated through an extensive numerical study that both policies perform consistently close to optimal. Chernonog (2020) investigated a two-echelon supply chain of a perishable product in which a manufacturer and a retailer interact via a wholesale price contract. In that model, product demand depends on the selling price, on advertising investment, and on the time a unit resides on the shelf before being sold. It is shown that under endogenous cycle length, there are cases in which Pareto improvement can be achieved by switching the roles of the leader and the follower in the supply chain. In our study, we consider the less explored scenario in which inventory is accumulated only when the server is idle (a stochastic process) and via production (also a stochastic process). Inventory is depleted either via demand or spoilage of pre-prepared food units (both are stochastic processes).

2.5. Pricing of perishable products

Pricing of perishable products is well studied in the marketing literature. Many studies in this vein consider dynamic pricing policies, in which products' prices change as they approach expiration. Rajan et al. (1992), for example, investigated dynamic pricing of a perishable product whose value to consumers drops as the product ages. The authors found that the optimal price may rise or fall according to changes in the product's accumulated cost and demand elasticity over time. Zhao and Zheng (2000) considered a dynamic pricing model for selling a perishable product to customers whose reservation price distribution changes over time. Levin et al. (2010) introduced a dynamic pricing model for a monopolistic company selling a perishable product, under the assumption that customers can act strategically—i.e., time their purchases so as to receive a discounted price. The authors found that a company that ignores strategic consumer behavior may receive much lower total revenues than one that uses a strategic equilibrium pricing policy. Herbon (2014) and Avinadav et al. (2017) found that a monopoly who faces deterministic demand that is dependent on the product's age, with non-strategic customers, should assign products a lower price at the early stages of their shelf-lives and then raise the price over time in order to compensate for the accumulated holding costs. Herbon and Khmelnitsky (2017) showed that a dynamic pricing policy may encourage customers to buy less-fresh products, potentially increasing revenue and eliminating waste. Feng et al. (2017) proposed an inventory model where the demand is a function of price, freshness, and displayed stock. The authors state that it may be profitable to always display fresh stock while selling less-fresh units at a discount. Li and Teng (2018) developed a joint pricing and lot-sizing model for retailers selling perishable products in which the demand depends not only on the selling price and reference price but also on product freshness. Sato (2019) investigated dynamic pricing decisions of a firm that sells perishable products in the presence of a firm offering a superior product. The author provided a probabilistic characterization of the optimal price trajectory under aggregate and multinomial logit demand models. Taken together, the diverse findings of studies in this vein suggest that the optimal pricing strategy for perishable products may vary in accordance with the business environment. For example, in some cases the optimal price of such products may increase over time, and in other cases it may decrease. Herein, we address considerations regarding pricing of fresh and pre-prepared food, taking into account customers' waiting time to be served (see sections 4 and 5).

3. The service system

For convenience, all notations used in the model are summarized in Appendix A.

3.1. Model description

We consider a service system in which, when the system is empty of customers, the server can prepare ready-to-purchase items (PPSs) and store them for future arrivals. Preparation time of a PPS is assumed to be exponentially distributed with parameter α. We assume that the maximum number of inventoried PPSs is limited to n (inventory capacity), where the value of n is determined according to economic considerations of the service provider (as elaborated below). When the

inventory level reaches n and no customers are present, the server stops producing PPSs and stays dormant. PPSs may spoil while being stored, and spoiled PPSs are disposed of by the server and do not reach the customers. We assume that the time to spoilage of a PPS is exponentially distributed with parameter θ.

PPSs offered to customers are assumed to meet a minimal standard of quality. However, customers perceive the quality of a PPS as being lower than that of a freshly-prepared product, since customers have no information on how long a PPS has been on the shelf and to what extent its quality has deteriorated. Therefore, a product prepared in front of a customer (fresh service, denoted FS) has a value of V_f, whereas a PPS has a value of V_p ($<V_f$).

There are two types of customers: fastidious and strategic. A fastidious customer is willing to wait for an FS as long as required, independent of whether PPSs are available or not. However, when a strategic customer arrives and a PPS is available, he/she acts as follows: if the number of customers in the system is lower than a given threshold m (common to all strategic customers), the customer will join the queue and wait for an FS; otherwise, the customer will take a PPS and leave. When PPSs are not available there are two possibilities: If the number of customers in the system is lower than a second threshold, $M(\geq m)$, the strategic customer will wait for an FS; otherwise, he/she will not join the queue and will leave the system without being served. In steady state, let L denote the number of customers present in the system. The possible system states and the actions taken upon arrival by a strategic customer are depicted in Fig. 1.

L = number of customers in the system; m = lower threshold; M ($>m$) upper threshold.

3.2. Steady-state analysis and system performance measures

Let the arrival rates of fastidious customers and of strategic customers follow Poisson distributions with parameters λ and η, respectively, and let the duration of time required to prepare an FS be exponentially distributed with parameter μ. We formulate the queueing-inventory system above as a quasi-birth-and-death (QBD) process where S denotes the number of PPSs in the system. The joint probability distribution function of the system states (L, S) is $p_{i,j} = P(L = i, S = j)$ where $i = 0, 1, 2, \ldots$ and $j = 0, 1, \ldots, n$. The system's states and a diagram of the transition rates are shown in Fig. 2. The balance equations for the probabilities $p_{i,j}$ are constructed with the aid of the so-called infinitesimal generator matrix Q given below, in line with standard practice in modeling multi-dimensional Markovian queues (see Neuts, 1981). This process entails arranging the system states in the following lexicographical order: $(\vec{\omega}_0, \vec{\omega}_1, \vec{\omega}_2, \ldots)$ where $\vec{\omega}_i \equiv ((i,0),(i,1),(i,2)\ldots(i,n))$, $i = 0, 1, 2, \ldots$, and Q provides the transition rates between those states. Thus,

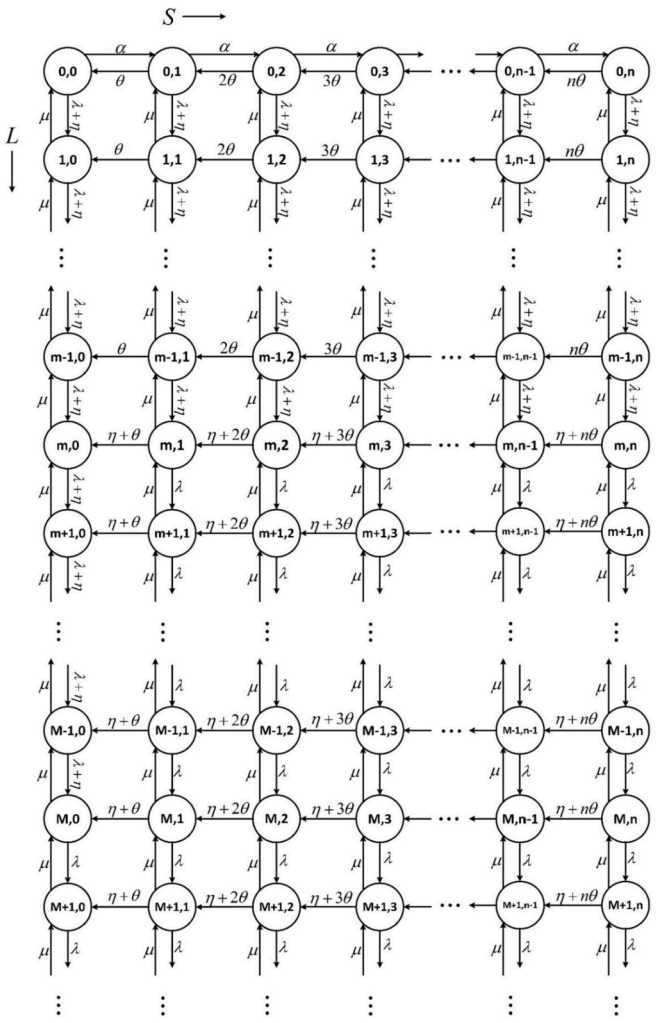


Fig. 2. System states and transition-rate diagram of the two-dimensional Markovian process.

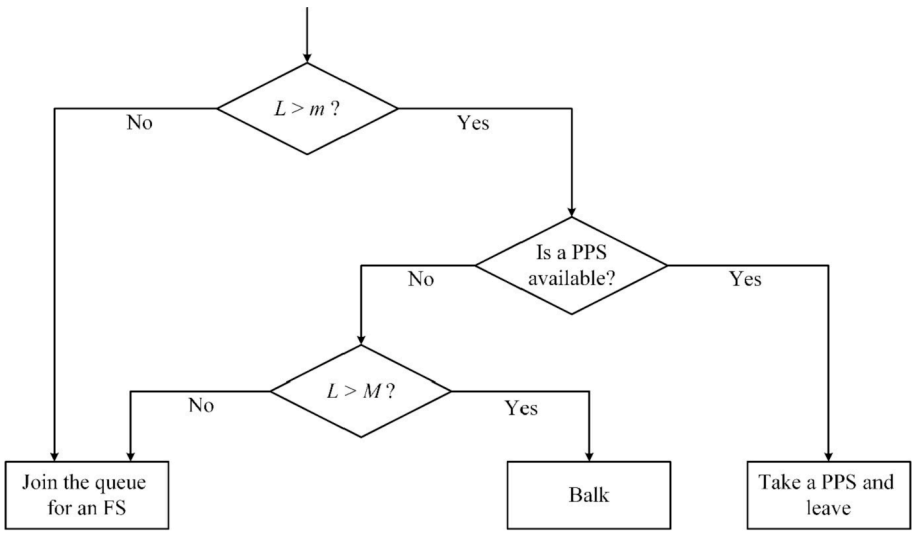


Fig. 1. Actions taken by a strategic customer.

$$Q = \begin{array}{c} \text{row} \backslash \text{block} \\ 0 \\ 1 \\ 2 \\ \vdots \\ m \\ m+1 \\ \vdots \\ M \\ M+1 \\ \vdots \end{array} \begin{pmatrix} B_0 & B_1 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ A_2 & B_2 & B_1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ 0 & A_2 & B_2 & B_1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ \vdots & & \ddots & \ddots & \ddots & & \vdots & \ddots & & & & & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & 0 & 0 & \cdots & A_2 & B_3 & B_4 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & A_2 & B_3 & B_4 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & & \ddots & \ddots & \ddots & & & \ddots & & \vdots & \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & A_2 & A_1 & A_0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & & \ddots & \ddots & \ddots \end{pmatrix}$$

where B_0, B_1, B_2, B_3, B_4, A_0, A_1 and A_2 are transition-rate matrices given in Appendix B. Note that the triplet (A_2, B_2, B_1) repeats itself in block-rows 1 to $m-1$ with a right shift along the block-rows; similarly, the triplet (A_2, B_3, B_4) repeats itself from block-row m to $M-1$, and the triplet (A_2, A_1, A_0) repeats itself from block-row M onwards, in both cases with a right shift along the block-rows. The matrix Q for each of the special cases where $(m=0, M=0)$, or $(m=0, M>0)$, or $(m>0, M=m)$ is presented in Appendix 3.

According to Neuts (1981, p. 83), the system's stability condition is $\vec{\pi} A_0 \vec{e} < \vec{\pi} A_2 \vec{e}$, where $\vec{e}$ is a column vector with all its entries equal to 1, and $\vec{\pi} = (\pi_0, \pi_1, ..., \pi_n)$ is the unique solution of the linear system $\vec{\pi}\,[A_0 + A_1 + A_2] = \vec{0}$ and $\vec{\pi} \cdot \vec{e} = 1$. In our case, $\vec{\pi} = (1, 0, ..., 0)$, and the stability condition translates into $\lambda < \mu$, which is identical to the stability condition of the classical M/M/1 queue. Hanukov and Yechiali (2019) showed that if each of the three matrices A_0, A_1 and A_2 is lower triangular, then the stability condition of the system is directly given by $a_0^{0,0} < a_2^{0,0}$ where $A_0 \equiv [a_0^{i,j}]$ and $A_2 \equiv [a_2^{i,j}]$(i.e., $\lambda < \mu$). So, the stability condition is unaffected by the production of PPSs.

Define $\vec{p} \equiv (\vec{p}_0, \vec{p}_1, \vec{p}_2, ...)$ as the vector of all the steady-state probabilities where $\vec{p}_i \equiv (p_{i,0}, p_{i,1}, p_{i,2}, ..., p_{i,n})$, $i = 0, 1, 2, ...$ Then, $\vec{p}$ is given by the solution of the system

$$\vec{p} Q = \vec{0} \text{ and } \vec{p} \cdot \vec{e} = 1, \tag{1}$$

where the first matrix-equation represents the balance equations, and the second is the normalization equation. In order to solve this system of equations, we use the matrix geometric method. Let the so-called rate matrix R be the positive minimal solution of the matrix quadratic equation

$$A_0 + RA_1 + R^2 A_2 = 0 \tag{2}$$

Then, $\vec{p}_i = \vec{p}_M R^{i-M}$, $i = M, M+1, M+2, ...$, where the initial conditions, represented by the vectors $\vec{p}_i$, $i = 0, 1, ..., M$, are obtained by solving part of Eq. (1) as follows (where matrix B is given in Appendix B):

(i) for $m = 0, M = 0$.

$$\begin{cases} \vec{p}_0 (B + RA_2) = \vec{0} \\ \vec{p}_0 [I - R]^{-1} \vec{e} = 1 \end{cases}$$

(ii) for $m = 0, M > 0$

$$\begin{cases} \vec{p}_0 B + \vec{p}_1 A_2 = \vec{0} \\ \vec{p}_{i-1} B_4 + \vec{p}_i B_3 + \vec{p}_{i+1} A_2 = \vec{0} \quad i = 1, 2, ..., M-1 \\ \vec{p}_{M-1} B_4 + \vec{p}_M (A_1 + RA_2) = \vec{0} \\ \sum_{k=0}^{M-1} (\vec{p}_k \cdot \vec{e}) + \vec{p}_M [I - R]^{-1} \vec{e} = 1 \end{cases}$$

(iii) for $m > 0, M = m$

$$\begin{cases} \vec{p}_0 B_0 + \vec{p}_1 A_2 = \vec{0} \\ \vec{p}_{i-1} B_1 + \vec{p}_i B_2 + \vec{p}_{i+1} A_2 = \vec{0} \quad i = 1, 2, ..., M-1 \\ \vec{p}_{M-1} B_1 + \vec{p}_M (A_1 + RA_2) = \vec{0} \\ \sum_{k=0}^{M-1} (\vec{p}_k \cdot \vec{e}) + \vec{p}_M [I - R]^{-1} \vec{e} = 1 \end{cases}$$

(iv) for $1 \le m < M$

$$\begin{cases} \vec{p}_0 B_0 + \vec{p}_1 A_2 = \vec{0} \\ \vec{p}_{i-1} B_1 + \vec{p}_i B_2 + \vec{p}_{i+1} A_2 = \vec{0} \quad i = 1, 2, ..., m-1 \\ \vec{p}_{m-1} B_1 + \vec{p}_m B_3 + \vec{p}_{m+1} A_2 = \vec{0} \\ \vec{p}_{i-1} B_4 + \vec{p}_i B_3 + \vec{p}_{i+1} A_2 = \vec{0} \quad i = m+1, m+2, ..., M-1 \\ \vec{p}_{M-1} B_4 + \vec{p}_M (A_1 + RA_2) = \vec{0} \\ \sum_{k=0}^{M-1} (\vec{p}_k \cdot \vec{e}) + \vec{p}_M [I - R]^{-1} \vec{e} = 1 \end{cases}$$

In most cases, the matrix R is calculated via successive substitutions (Neuts, 1981, p. 37). However, in our case, we are able to express *explicitly* the entries of $R \equiv [r_{i,j}]$, thus decreasing considerably the calculation effort. We claim:

Theorem 1. *The entries of R are given by*

$$r_{i,j} = \begin{cases} \lambda/\mu & i = j = 0 \\ 0 & 0 \le i < j \le n \\ \dfrac{\lambda + \mu + \eta + j\theta - \sqrt{(\lambda + \mu + \eta + j\theta)^2 - 4\lambda\mu}}{2\mu} & 1 \le i = j \le n \\ \dfrac{(\eta + (j+1)\theta) r_{i,j+1} + \mu \sum_{k=j+1}^{i-1} r_{i,k} r_{k,j}}{\lambda + \mu + \eta + j\theta - \mu (r_{i,i} + r_{j,j})} & 2 \le i \le n \ ; \ 1 \le j \le i-1 \\ \dfrac{(\eta + \theta) r_{i,1} + \mu \sum_{k=1}^{i-1} r_{i,k} r_{k,0}}{\mu (1 - r_{i,i})} & 1 \le i \le n \ ; \ j = 0 \end{cases}$$

Proof. The proof is presented in Appendix D.

Note that $r_{j,j}$ can be considered as the Laplace-Stieltjes transform, $\tilde{g}(s)$, of the duration of a busy period (i.e., a period in which customers are present in the system) in an M/M/1 queue with arrival rate μ and service rate λ, evaluated at $s = \eta + j\theta$ (Kleinrock, 1975, p. 215).

In what follows we will use the abbreviations "*fas*", "*str*" and "*uns*" to denote a fastidious, strategic, and unspecified customer (i.e., an arbitrary customer), respectively. For a given capacity of n PPSs, let $L^{[y]}$ and

$L_q^{[y]}$ denote the mean number of customers of type $y \in \{fas, str, uns\}$ in the system and in queue, respectively. Similarly, let $W^{[y]}$ and $W_q^{[y]}$ denote, respectively, the mean sojourn times of a type y customer in the system and in queue; let $\overline{S}$ denote the mean number of PPSs in the system; and let $\overline{T}$ denote the mean duration of time a PPS resides in the system. In addition, let η_{eff} denote the effective rate at which strategic customers join the queue; let η_{PPS} denote the rate at which strategic customers take PPSs; let Γ_{eff} denote the total effective rate at which customers join the queue; let α_{eff} denote the effective production rate of PPSs; and let θ_{eff} denote the effective deterioration rate of PPSs. In what follows, we provide the formulae to calculate each of these variables, based on the steady-state probabilities and model parameters. For simplicity of presentation, we define $p_{i,\cdot} \equiv \sum_{j=0}^{n} p_{i,j} = \vec{p}_i \cdot \vec{e}$, $i = 0, 1, 2, \ldots$, and $p_{\cdot,j} \equiv \sum_{i=0}^{\infty} p_{i,j}$, $j = 0, 1, \ldots, n$. Also let $\vec{v} \equiv (0, 1, 2, \ldots, n)^T$ and $\vec{u} \equiv (1, 0, 0, \ldots, 0)^T$.

Below are several relations, which are used in the economic analysis in the subsequent section:

(i) $\alpha_{eff} = \alpha(p_{0,\cdot} - p_{0,n})$

(ii) $\eta_{eff} = \eta(\sum_{i=0}^{m-1} p_{i,\cdot} + \sum_{i=m}^{M-1} p_{i,0})$

(iii) $\eta_{PPS} = \eta \sum_{i=m}^{\infty} \sum_{j=1}^{n} p_{i,j} = \eta \sum_{i=m}^{\infty} (\vec{p}_i \cdot \vec{e} - \vec{p}_i \cdot \vec{u}) = \eta(\sum_{i=m}^{M-1} \vec{p}_i + \vec{p}_M [I - R]^{-1}) \cdot (\vec{e} - \vec{u})$

(iv) $\theta_{eff} = \theta \sum_{j=1}^{n} (j p_{\cdot,j}) = \theta \sum_{i=0}^{\infty} (\vec{p}_i \cdot \vec{v}) = \theta(\sum_{i=0}^{M-1} (\vec{p}_i \cdot \vec{v}) + \vec{p}_M [I - R]^{-1} \vec{v})$

(v) $L^{[uns]} = \sum_{i=0}^{\infty} (i \vec{p}_i \cdot \vec{e}) = \sum_{i=0}^{M-1} i p_{i,\cdot} + \vec{p}_M [(M - 1)[I - R]^{-1} + [I - R]^{-2}] \vec{e}$

Additional relations are given in Appendix E, where some of the relations are obtained by substituting $\vec{p}_i = \vec{p}_M R^{i-M}$ and using the relations $\sum_{i=0}^{\infty} R^i = [I - R]^{-1}$ and $\sum_{i=0}^{\infty} (i + 1) R^i = [I - R]^{-2}$.

4. Behavioral and economic analysis

4.1. Short-term behavior of strategic customers

A newly-arrived strategic customer chooses whether to join the queue and receive a full service (FS), purchase a PPS (if available) without waiting, or balk. Each strategic customer makes the best choice that maximizes his/her individual utility. First, assume that an FS and a PPS are being sold for the same price b. To avoid triviality, we assume that $V_p > b$. Thus, the utility function of a strategic customer who sees i (unspecified) customers in the system is

$$U = \begin{cases} V_f - b - c(i+1)/\mu & \text{if the customer purchases an FS} \\ V_p - b & \text{if the customer purchases a PPS} \\ 0 & \text{if the customer balks} \end{cases}$$

where c is the sojourn time cost rate from the customer's perspective, and $(i+1)/\mu$ is the mean sojourn time of that strategic customer.

When PPSs are available, a strategic customer will join the queue if and only if his/her utility from joining is equal to or greater than that of taking a PPS, i.e., if and only if $V_f - c(i+1)/\mu \geq V_p$. Thus, the Nash equilibrium m (obtained by a pure strategy), indicating the minimal queue length for which the strategic customer refuses to stand in line and takes a PPS, is determined by

$$m = 1 + \max\left(i \middle| V_f - c(i+1)/\mu \geq V_p\right) = \left\lfloor (V_f - V_p)\mu/c \right\rfloor. \quad (3)$$

Note that a strategic customer follows "avoid the crowd" behavior; i.e., an increase in the propensity of others to join a queue tends to discourage the individual from joining, and he/she has to choose between two actions (see Fig. 1). Thus, according to Hassin and Haviv (2003, pp. 6–7), equilibrium m is unique.

When PPSs are not available, a strategic customer will join the queue if and only if his/her utility from joining is larger than the utility obtained from balking, i.e., $V_f - b - c(i+1)/\mu \geq 0$. Thus, the Nash equilibrium M (obtained by a pure strategy), indicating the minimal queue length for which the strategic customer balks, is determined by

$$M = 1 + \max\left(i \middle| V_f - b - c(i+1)/\mu \geq 0\right) = \left\lfloor (V_f - b)\mu/c \right\rfloor. \quad (4)$$

Equilibrium M is unique for the same reasons presented above with regard to equilibrium m. By Eq. (3), we conclude that the balking threshold, M, increases in the service rate, μ, and in the value of the fresh service, V_f, whereas it decreases in the selling price, b, and in the sojourn time cost, c.

4.2. Capacity planning and price discrimination

The fact that a given customer derives different utility levels from the two types of services (PPS and FS) suggests that the service provider may benefit from using a price discrimination policy. In particular, if the service provider charges less for a PPS than for an FS, then, on the one hand, income per PPS decreases (as compared with selling both services at the same price), but, on the other hand, the value of m decreases as well, resulting in a reduction of spoilage costs, holding costs and sojourn costs. Conversely, if the service provider charges more for a PPS than for an FS, then, on the one hand, income per PPS increases, but, on the other hand, the value of m increases, thereby increasing the associated costs. In what follows, we continue to use b to denote the price of the FS, whereas we denote the price of a PPS as $b - d$, where a positive value of d implies a price discount, and a negative value of d implies a price increase (as compared with the price of an FS).

The interaction between the server and the customers can be modeled via a Stackelberg game in which the server moves first by setting the value of d. In order to find the best response (the value of m) for a strategic customer, we adjust the strategic customer's utility function as follows:

$$U = \begin{cases} V_f - b - c(i+1)/\mu & \text{if the customer purchases an FS} \\ V_p - (b - d) & \text{if the customer purchases a PPS} \\ 0 & \text{if the customer balks} \end{cases} \quad (5)$$

which results in the following threshold value

$$m_d = \left\lfloor (V_f - V_p - d)\mu/c \right\rfloor. \quad (6)$$

Therefore, the service provider, who controls the capacity n of PPSs and the price discrimination value d, solves the following maximization problem:

$$\max_{n,d} \{ Z = (b - c_r)(\lambda + \eta_{eff}(n,d)) + (b - c_r - d)\eta_{PPS}(n,d) - [c_L L^{[uns]}(n,d) + c_n n + c_r \theta_{eff}(n,d) + \phi \eta_{balk}(n,d)] \} \quad (7)$$

$$S.t. \quad d \geq b - V_p, \quad (8)$$

where c_L is the sojourn time cost rate, from the service provider's perspective, per customer in the system; c_r is the cost of raw materials and labor for preparation of a single unit; c_n is the cost rate per unit of PPS capacity (a cost that includes, for example, costs resulting from storage equipment depreciation, maintenance and so forth); ϕ is the loss of reputation (in monetary units) and of future purchases associated with a customer who balks; and $\eta_{balk} = \eta - \eta_{eff} - \eta_{PPS}$ is the average balking rate of strategic customers. Note that if the optimal solution is $n^* = 0$, the price discrimination value d is irrelevant to the model.

5. Computational analysis

5.1. Baseline example

We were able to obtain a closed-form expression for the customer's

decision variable M. However, it is not possible to derive such expressions for the service provider's decision variables, n and d (and following (6) also for m). Therefore, in order to evaluate the effect of each parameter on these variables, we performed a numerical study. As a baseline, we used the following parameter values: For arrival and service rates we used $\lambda = 10\left[\frac{\text{fastidious customers}}{\text{hour}}\right]$, $\eta = 6\left[\frac{\text{impatient customers}}{\text{hour}}\right]$, $\mu = 20\left[\frac{\text{customers}}{\text{hour}}\right]$ and $\alpha = 20\left[\frac{\text{PPSs}}{\text{hour}}\right]$; these rates represent server utilization of approximately 80% and a service rate that is indifferent to whether customers are present or absent during food preparation. For the service provider's monetary parameters we used $b = 15\left[\frac{\$}{\text{unit}}\right]$ and $c_r = 5\left[\frac{\$}{\text{unit}}\right]$, which are close to the average values associated with hamburgers, Cobb salad and pizza (Priceconomics, 2017); $c_L = 30\left[\frac{\$}{\text{customer}\times\text{hour}}\right]$ (for example, Wang and Chang, 2016 use $200 per day); and $c_n = 0.1\left[\frac{\$}{\text{hour}}\right]$ per unit capacity. For the customer's monetary parameters, we used $V_f = 22\left[\frac{\$}{FS}\right]$, $V_p = 17\left[\frac{\$}{\text{PPS}}\right]$ and $c = 20\left[\frac{\$}{\text{hour}}\right]$, which imply $M = 7$ (see eq. (4)). Without price discrimination (i.e., $d = 0$), these values result in $m_0 = 5$, whereas a change of 1 dollar in the value of d for buying a PPS results in a change (in the opposite direction) of 1 unit in m_d (since $\mu/c = 1$). For the deterioration rate, we used $\theta = 0.3\left[\frac{\text{spoilt PPSs}}{\text{hour}}\right]$ (this assumption reflects the fact that it is difficult to sell a reheated pizza that has been prepared more than 3 h ago, for example), and $\phi = 20\left[\frac{\$}{\text{balking customer}}\right]$. Solving the maximization problem using an exhaustive search over the domain $n = 0,..,15$ with $d = -2,-1,0,..,5$ (according to the constraint in (8)) resulted in the following policy for the server: prepare capacity for up to $n^* = 9$PPSs, and offer a price discrimination (discount) value of $d^* = \$4$ per PPS. This set of values will result in an expected profit of $Z^* = \$90.93$ per hour. An illustration of the server's objective function is given in Fig. 3, and its values are given in Table 1.

Table 1
The service provider's expected profit as a function of n and d.

$n\backslash d$	5	4	3	2	1	0	−1	−2
0	61.59	61.59	61.59	61.59	61.59	61.59	61.59	61.59
1	68.90	72.16	71.89	70.60	68.99	67.35	65.77	64.26
2	74.42	79.08	78.51	76.32	73.63	70.88	68.25	66.70
3	78.22	83.65	82.84	80.03	76.59	73.07	69.70	68.16
4	80.75	86.67	85.69	82.44	78.47	74.40	70.50	68.83
5	82.36	88.63	87.53	83.96	79.61	75.14	70.85	68.94
6	83.32	89.85	88.67	84.88	80.24	75.47	70.89	68.66
7	83.82	90.55	89.31	85.36	80.50	75.51	70.72	68.10
8	83.98	90.88	89.60	85.52	80.51	75.34	70.38	67.36
9	83.90	**90.93**	89.62	85.46	80.32	75.02	69.93	66.48
10	83.65	90.78	89.46	85.23	80.00	74.60	69.40	65.51
11	83.26	90.49	89.16	84.88	79.58	74.10	68.82	64.48
12	82.79	90.09	88.77	84.44	79.09	73.55	68.20	63.41
13	82.26	89.62	88.30	83.95	78.55	72.96	67.57	62.31
14	81.69	89.10	87.79	83.42	77.99	72.36	66.93	61.20
15	81.10	88.56	87.25	82.87	77.42	71.76	66.30	60.08

5.2. Sensitivity analysis

In order to evaluate the effect of each parameter on the optimal solution, we carried out a sensitivity analysis with regard to the parameter values. In particular, we experimented with six additional values of α, θ, c_L, c_r, c_n and ϕ, which correspond to deviations of ±10%, ±25% and ±50% with respect to the values used in the baseline example presented above. Furthermore, we investigated 17 pairs of values of (λ,η), such that $\lambda + \eta = 16$, $\lambda = 0, 1, 2, \ldots, 16$. In total, we solved 54 variants of the problem, and the results are summarized in Tables 2 and 3.

5.2.1. Effect of parameters α, θ, c_L, c_r, c_n and ϕ

Table 2 shows that the sensitivity of the optimal price discrimination level to changes in each parameter value is relatively mild (changes in the value of d with respect to the value obtained in the baseline example range between −25% and 0%, but are mostly zero). The sensitivity of the optimal capacity of PPSs to changes in each parameter value is moderate (with variations with respect to the baseline solution ranging from −33.3% to +33.3%). The sensitivity of the expected profit can also be considered as moderate (with variations ranging from −47.4% to +48.8%). The influence of each parameter on the optimal capacity of PPSs is as expected. Specifically, when the spoilage rate θ increases, the inventory of PPSs is more exposed to loss, and thus the optimal capacity decreases (such that the service provider stores fewer items and avoids excessive spoilage costs). Moreover, it is observed that a higher production rate of PPSs, α, results in lower capacity of PPSs, which can be explained by the ability of the server to utilize its idle time more efficiently to produce PPSs. When customers' sojourn cost c_L increases, the server can increase the inventory of PPSs (via increasing the capacity n) to increase the likelihood that strategic customers will be able to make a purchase without incurring a high sojourn cost. As for the capacity cost

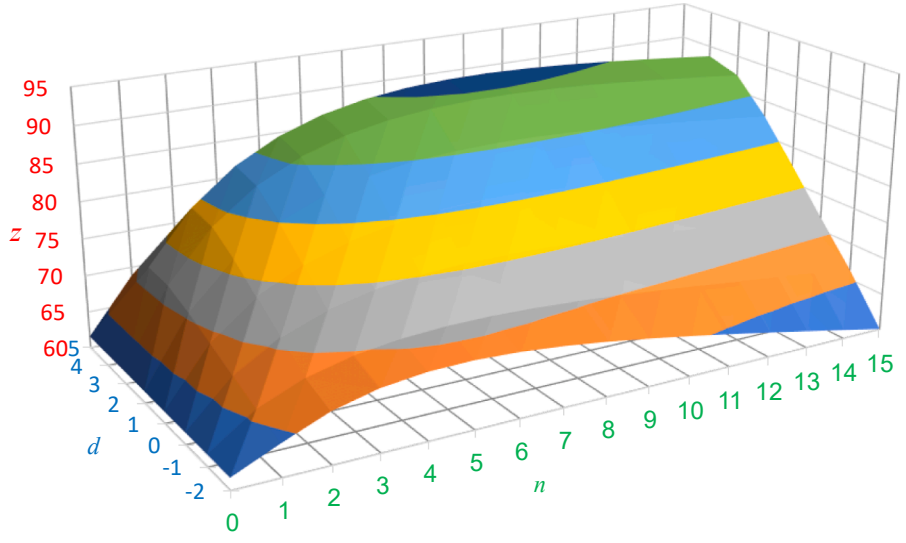


Fig. 3. The expected profit (Z) as a function of the PPS capacity (n) and the price discrimination value for a PPS (d) for the base example.

Table 2
Deviation in % from the optimal solution of the base example.($n^* = 9,\ d^* = 4,\ Z^* = 90.9,\ m_{d^*} = 1$)

Parameter	Influenced variable	Change in % in parameter value					
		−50	−25	−10	+10	+25	+50
θ	n^*	22.2	11.1	0.0	−11.1	−11.1	−22.2
	d^*	0.0	0.0	0.0	0.0	0.0	0.0
	Z^*	6.5	3.0	1.2	−1.1	−2.6	−4.9
	m_{d^*}	0.0	0.0	0.0	0.0	0.0	0.0
α	n^*	11.1	11.1	0.0	−11.1	−11.1	−11.1
	d^*	−25.0	0.0	0.0	0.0	0.0	0.0
	Z^*	−7.3	−2.4	−0.7	0.6	1.2	1.9
	m_{d^*}	100.0	0.0	0.0	0.0	0.0	0.0
c_L	n^*	−33.3	−11.1	−11.1	0.0	11.1	22.2
	d^*	−25.0	−25.0	0.0	0.0	0.0	0.0
	Z^*	27.3	12.8	4.9	−4.8	−12.0	−23.9
	m_{d^*}	100.0	100.0	0.0	0.0	0.0	0.0
c_n	n^*	0.0	0.0	0.0	0.0	0.0	0.0
	d^*	0.0	0.0	0.0	0.0	0.0	0.0
	Z^*	0.5	0.2	0.1	−0.1	−0.2	−0.5
	m_{d^*}	0.0	0.0	0.0	0.0	0.0	0.0
c_r	n^*	33.3	11.1	0.0	−11.1	−11.1	−22.2
	d^*	0.0	0.0	0.0	0.0	0.0	0.0
	Z^*	48.8	24.2	9.6	−9.6	−23.8	−47.4
	m_{d^*}	0.0	0.0	0.0	0.0	0.0	0.0
ϕ	n^*	−11.1	−11.1	0.0	0.0	0.0	0.0
	d^*	0.0	0.0	0.0	0.0	0.0	0.0
	Z^*	1.4	0.7	0.3	−0.3	−0.7	−1.3
	m_{d^*}	0.0	0.0	0.0	0.0	0.0	0.0

c_n, it has no effect on the optimal capacity of PPSs, which can be explained by the low baseline value of this cost as compared with the other cost components. As for the preparation cost c_r, clearly, when it is higher the optimal capacity of PPSs is lower.

An interesting result is that the discount in the selling price of a PPS is unaffected by deviations from the baseline values except in two cases: a decrease of 50% in the value of α and a decrease of 25% in the value of c_L. This result can be explained by the fact that the server has low motivation to encourage customers to buy a PPS when the customer's sojourn cost is low (i.e., lower c_L) or when PPS production efficiency is low (i.e., lower α). Similarly, the expected profit is robust to changes in the parameter values, except for c_L and c_r. This result is explained by the large share of these cost components in the total cost.

5.2.2. *Effect of parameters λ and η*

Table 3 provides a sensitivity analysis with respect to the ratio between the number of fastidious customers and the number of strategic customers, and also illustrates the improvement in profitability achieved by using the queueing-inventory system described herein (i.e., preparing and storing PPSs and selling them for a discounted price) compared with the regular (inventory-free) queueing model, in which all customers receive the same service ($n = 0, d = 0$). When the share of strategic customers is larger, given that the total arrival rate is fixed, it is clear that a larger capacity of PPSs should be prepared to address the higher demand. When demand for PPSs increases, their selling price can be raised by giving a smaller discount. Moreover, the server's expected profit is larger, and so is the improvement compared with a system without PPSs. Note that the expected profit increases in a concave manner in the share of strategic customers, which can be explained by the law of diminishing returns.

Table 3
The optimal solutions n^* and d^* for various values of λ and η ($\lambda + \eta = 16$), and profit improvement compared with a system without PPSs.

λ	η	n^*	d^*	Z^*	$Z(n=0)$	$\frac{Z^* - Z(n=0)}{Z(n=0)} \times 100$
0	16	10	3	102.4	64.2	59.6
1	15	10	3	101.9	64.1	58.9
2	14	10	3	101.2	64.0	58.1
3	13	10	3	100.5	63.9	57.2
4	12	10	3	99.7	63.8	56.2
5	11	10	3	98.7	63.6	55.0
6	10	10	4	97.5	63.4	53.8
7	9	10	4	96.5	63.1	52.8
8	8	9	4	95.1	62.8	51.5
9	7	9	4	93.3	62.3	49.8
10	6	9	4	90.9	61.6	47.6
11	5	8	4	87.8	60.7	44.8
12	4	8	4	83.6	59.3	40.9
13	3	7	4	77.9	57.4	35.8
14	2	5	4	69.8	54.3	28.6
15	1	3	4	57.7	49.2	17.4
16	0	0	–	40.0	40.0	0.0

5.3. *Example of a case in which it is appropriate to charge more for a PPS than for an FS*

In the above baseline example, the service provider offers a price discount for PPSs as a means of increasing the expected profit. The following example shows a scenario in which the opposite price discrimination policy is appropriate. This example uses the same parameter values used in the baseline example, except for $V_f = 26\ \left[\frac{\$}{FS}\right]$ and $V_p = 21\ \left[\frac{\$}{\text{PPS}}\right]$, and shows that it is optimal for the service provider to sell the PPSs for selling price of $16\ \left[\frac{\$}{\text{PPS}}\right]$, whereas the selling price of an FS is $15\ \left[\frac{\$}{FS}\right]$, i.e., $d^* = -\$1$. The service provider's expected profit for $n = 0, 1, 2, \ldots, 15$ with $d = -6, -5, \ldots, 5$ is given in Table 4. As we can see, the maximal expected profit of $Z^* = \$150.58$ per hour is obtained by using the capacity of PPSs $n^* = 1$ and $d^* = -\$1$.

5.4. *Customers' utility analysis*

We distinguish between the two types of customers when measuring customer utility. The utility of strategic customers has been defined in terms of monetary value in Eq. (5), while taking into account product freshness and sojourn time. In order to calculate the expected utility of a strategic customer over the various states of the system, we use the following formula:

Table 4
The service provider's expected profit as a function of n and d.

$n \setminus d$	5	4	3	2	1	0	−1	−2	−3	−4	−5	−6
0	150.19	150.19	150.19	150.19	150.19	150.19	150.19	150.19	150.19	150.19	150.19	150.19
1	142.23	147.31	149.24	150.08	150.45	150.57	**150.58**	150.52	150.43	150.33	150.25	150.22
2	137.57	145.26	148.33	149.72	150.33	150.55	150.57	150.47	150.33	150.17	150.03	150.38
3	134.26	143.57	147.41	149.16	149.96	150.25	150.28	150.17	149.99	149.79	149.61	150.26
4	131.66	142.06	146.44	148.48	149.41	149.76	149.81	149.69	149.49	149.27	149.05	149.86
5	129.47	140.66	145.45	147.69	148.73	149.14	149.21	149.09	148.88	148.63	148.39	149.25
6	127.57	139.34	144.43	146.84	147.97	148.42	148.51	148.40	148.18	147.92	147.67	148.49
7	125.85	138.08	143.40	145.95	147.15	147.65	147.75	147.64	147.42	147.16	146.89	147.63
8	124.29	136.87	142.38	145.03	146.29	146.83	146.95	146.85	146.63	146.37	146.09	146.69
9	122.84	135.70	141.36	144.10	145.42	145.98	146.13	146.04	145.82	145.55	145.28	145.72
10	121.49	134.57	140.36	143.17	144.54	145.13	145.29	145.22	145.01	144.74	144.46	144.71
11	120.22	133.48	139.38	142.25	143.66	144.28	144.46	144.39	144.19	143.92	143.64	143.70
12	119.03	132.44	138.42	141.35	142.80	143.44	143.64	143.58	143.38	143.12	142.84	142.69
13	117.90	131.44	137.50	140.48	141.95	142.62	142.83	142.78	142.60	142.34	142.06	141.68
14	116.84	130.48	136.61	139.63	141.13	141.82	142.05	142.01	141.83	141.58	141.30	140.69
15	115.84	129.57	135.75	138.81	140.34	141.05	141.29	141.26	141.09	140.84	140.57	139.72

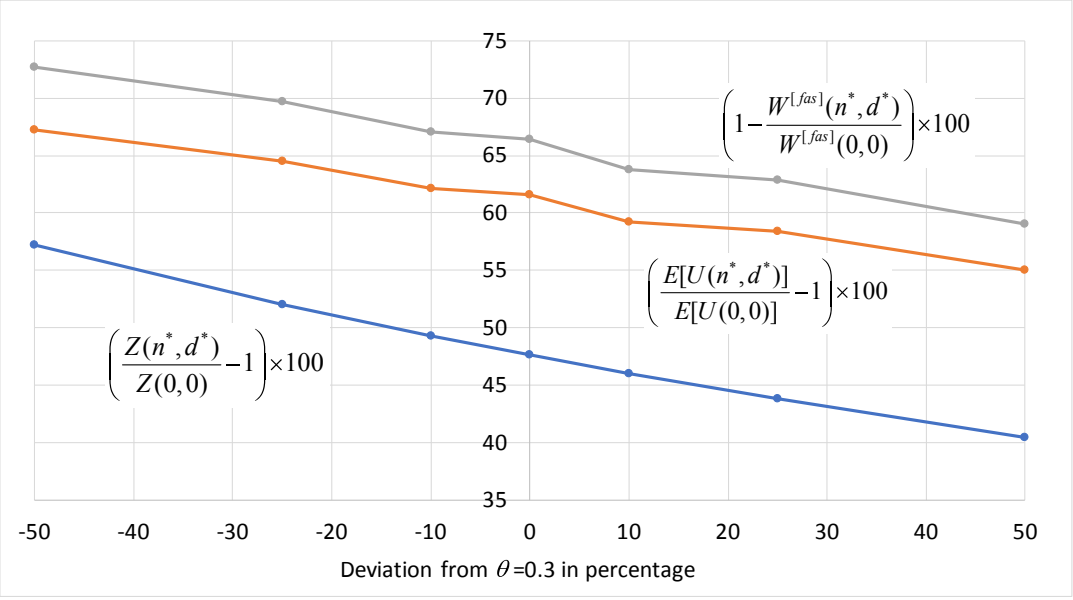


Fig. 4. Improvement in the server's expected profit, the expected utility of a strategic customer, and the expected sojourn time of a fastidious customer as a function of θ.

$$E[U(n,d)] = \sum_{i=0}^{m-1}(V_f - b - c(i+1)/\mu)p_{i,\cdot} + \sum_{i=m}^{M-1}(V_f - b - c(i+1)/\mu)p_{i,0} + (V_p - (b-d))\left(\sum_{i=m}^{M-1}\vec{p}_i + \vec{p}_M[I-R]^{-1}\right)\cdot(\vec{e}-\vec{u}) \quad (9)$$

The first element of $E[U(n,d)]$ refers to a strategic customer who joins the queue due to short anticipated sojourn time; the second element refers to a strategic customer who joins the queue due to a lack of PPSs; and the third element refers to a strategic customer who prefers to buy an available PPS. Figs. 4–9 present the improvement, as compared with the regular model ($n = 0$, $d = 0$), in the server's expected profit and in the expected utility of a strategic customer as a function of various parameters: θ, α, c_L, c_r, c_n and ϕ. As for fastidious customers, we measure their utility on the basis of their expected sojourn time.

For the analysis, we define three measures of percentage improvement: $\left(1-\frac{W^{[fas]}(n^*,d^*)}{W^{[fas]}(0,0)}\right)\times 100$, $\left(\frac{E[U(n^*,d^*)]}{E[U(0,0)]}-1\right)\times 100$ and $\left(\frac{Z(n^*,d^*)}{Z(0,0)}-1\right)\times 100$, corresponding to fastidious customers' utility (i.e., their sojourn time), strategic customers' utility, and the service provider's profit, respectively, where a positive value for a given measure implies that the proposed model yields better performance on that measure compared with the regular model ($n = 0, d = 0$). It is observed from Figs. 4–9 that the proposed system is better for all parties compared with the regular system. An interesting result that emerges from each of these figures is that the improvement for fastidious (non-strategic) customers is greater than the improvement for strategic customers. This result can be explained by the following intuition: For fastidious customers the proposed system is purely beneficial, given that they always decide to join the queue, and the availability of PPSs serves to shorten the queue; strategic customers, in contrast, benefit from shorter sojourn times on the one hand, but, on the other hand, may receive lower-quality service if they decide to purchase PPSs rather than to join the queue. Notably, in some cases, the improvement in customer utility is greater than the improvement in the server's profit, despite the fact that the server is the one who controls the decision variables.

Fig. 4 shows that higher values of θ result in lower values of the three measures above, implying that, as expected, all parties lose from a higher deterioration rate. Fig. 5 shows that all three measures are robust to changes in the capacity cost c_n. Fig. 6 shows that the general trend of the three measures is increasing in the production rate of PPSs, α. Fig. 7

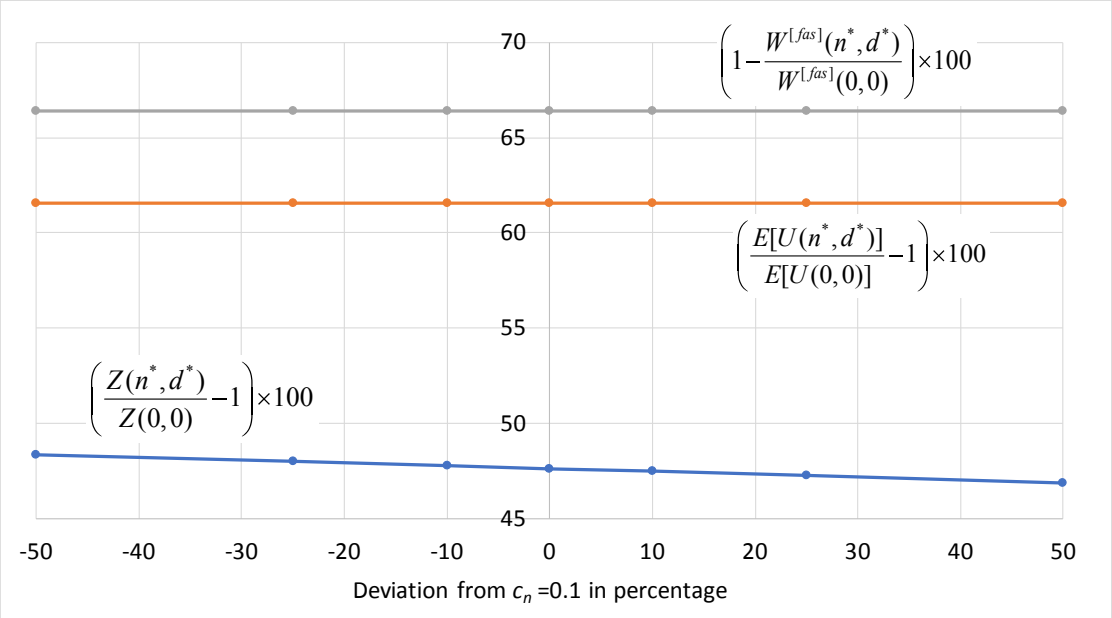


Fig. 5. Improvement in the server's expected profit, the expected utility of a strategic customer, and the expected sojourn time of a fastidious customer as a function of c_n.

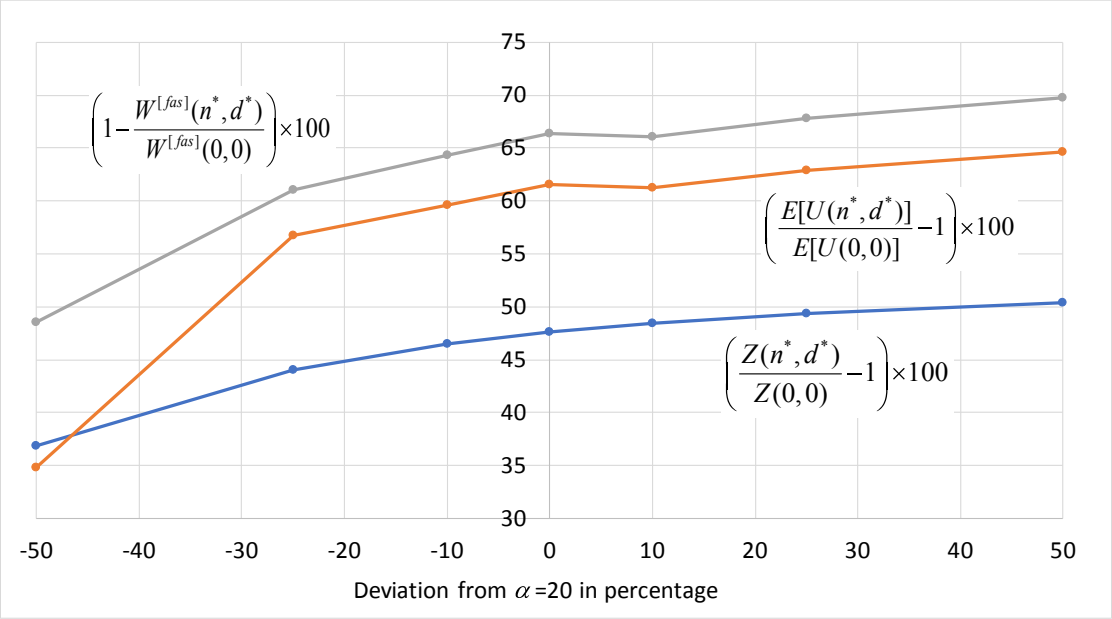


Fig. 6. Improvement in the server's expected profit, the expected utility of a strategic customer, and the expected sojourn time of a fastidious customer as a function of α.

shows that the improvement in the server's profits accelerates as the sojourn cost of customers increases, whereas the customers' measures improve slightly. Fig. 8 shows that the improvement in the service provider's profits accelerates as the per-unit cost c_r of labor and raw materials increases, whereas the customers' measures decrease slightly. Fig. 9 shows that the improvement in the server's profits is almost constant as the balking cost ϕ increases, whereas the customers' measures improve slightly.

In each of Figs. 4–6, we observe that the three measures of percentage improvement resemble one another in terms of their responses to variation in the focal parameter value (θ, c_n or α, respectively). In Figs. 7–9, in contrast, we observe that only two percentage improvement measures—namely, fastidious customers' and strategic customers' utility—show similar trends in response to variation in the focal parameter value (c_L, c_r and ϕ, respectively). The third measure—percentage improvement of the service provider's profit—shows a different trend, which is more sensitive to changes in c_L, c_r and ϕ. The intuition behind this surprising result is that the cost parameters have a stronger effect on the server's expected profit except for the capacity cost, whose contribution to the profit is relatively small.

5.5. Summary of results and implications

The main results obtained from our computational analysis can be summarized as follows:

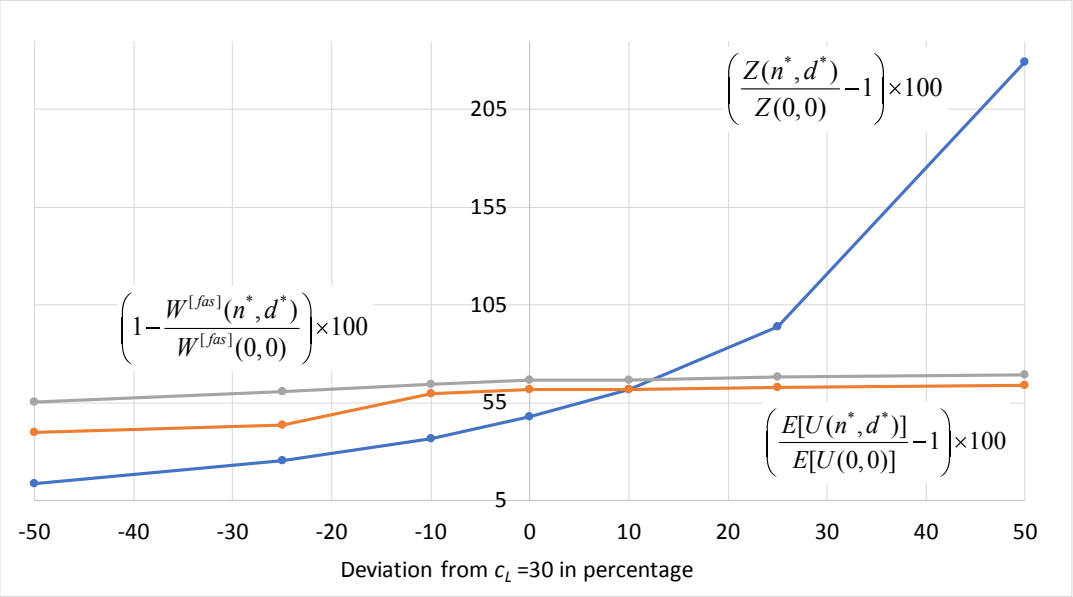


Fig. 7. Improvement in the server's expected profit, the expected utility of a strategic customer, and the expected sojourn time of a fastidious customer as a function of c_L.

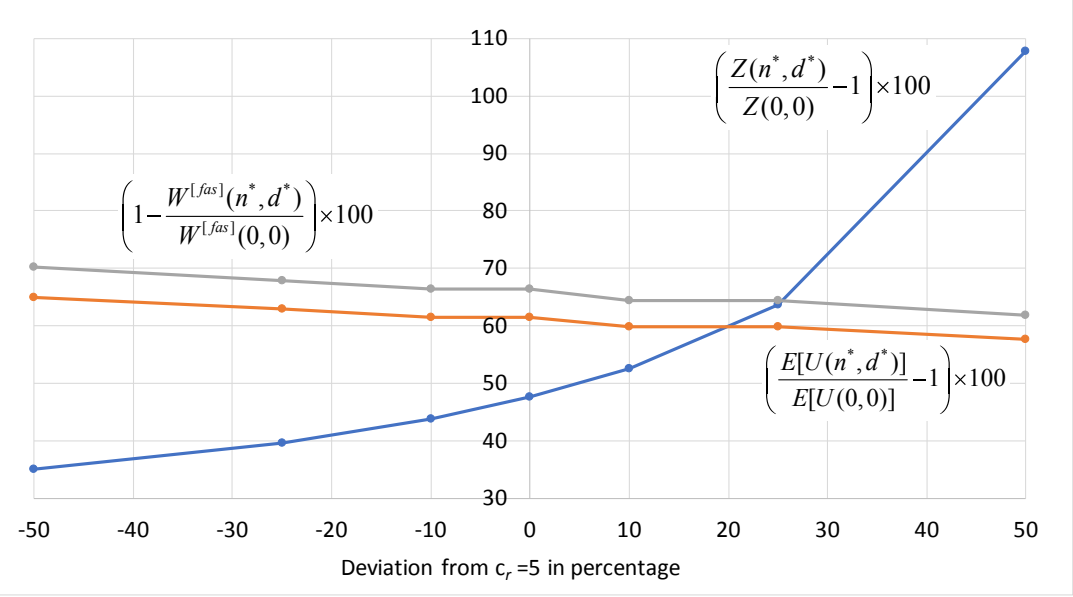


Fig. 8. Improvement in the server's expected profit, the expected utility of a strategic customer, and the expected sojourn time of a fastidious customer as a function of c_r.

- The server can benefit from limiting the capacity of inventoried PPSs, implying that it may not be beneficial for the server to utilize the entirety of its idle time for producing PPSs.
- The optimal pricing policy does not necessarily entail a discount on a less-fresh product (i.e., a PPS) as compared with a fresh product; rather, in some cases, it may be beneficial to sell a PPS for a higher price.
- The sensitivity of the value of the price discrimination level (d) to changes in the model parameter values (with the exception of the arrival rates of the two types of customers) is relatively mild. The sensitivity of the optimal inventory capacity to changes in each parameter value is moderate, as is the sensitivity of the server's expected profit.
- When the share of strategic customers increases, a larger capacity of PPSs should be prepared, and the selling price of PPSs should be raised. These steps enable the service provider to increase the expected profit, as well as to achieve greater improvement as compared with a traditional queueing system without PPSs.
- A counter-intuitive result is that the percentage improvement in customer utility that is achieved by utilizing the server's idle time to produce PPSs and offering a discount is larger for non-strategic customers than for strategic customers.
- Another counter-intuitive result is that, in some cases, the percentage improvement in customer utility (for customers of any type) achieved by the availability and discounting of PPSs is greater than the percentage improvement in the server's profit, despite the fact that the server is the one who controls the decision variables.

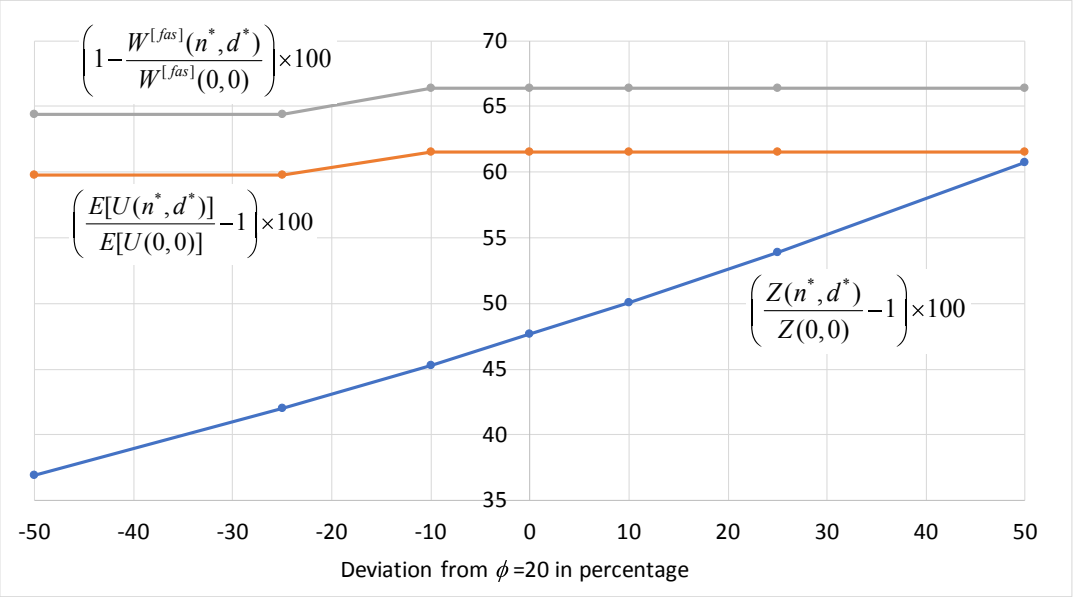


Fig. 9. Improvement in the server's expected profit, the expected utility of a strategic customer, and the expected sojourn time of a fastidious customer as a function of ϕ.

- Both the server and the customers (both types) lose from a higher deterioration rate.
- As the sojourn cost of customers increases, the improvement in the server's profits (when comparing the PPS system to a traditional queueing system) accelerates, whereas the customers' utility levels improve slightly.
- As the per-unit cost of labor and raw materials increases, improvement in the service provider's profits accelerates, whereas the customers' utility measures decrease slightly.
- The balking cost has a minor effect both on the server's profits and on customers' utility.

6. Conclusions

In this study we investigated a fast food service system with a strategic server who faces two types of customers: fastidious and strategic. Taking into account pricing, customer waiting time, and spoilage, while assuming that the server produces PPSs during his/her idle time (i.e., when no customers are present), we sought to determine: (i) the economically optimal capacity level of PPSs; and (ii) the optimal price discrimination level in the selling price that the server should offer to customers who are willing to purchase PPSs rather than wait in the queue for a freshly-prepared product. Although the mathematical model of this stochastic process is complex, we succeeded in solving it analytically using matrix geometric analysis. By providing closed-form expressions of all elements in the rate matrix R, our approach enables optimal values of PPS capacity and price discrimination level to be identified in a short time, even for large problems.

It is clear that non-strategic customers benefit when other customers are strategic, since the latter may decide not to join the queue, thereby reducing the sojourn time of the non-strategic customers (as compared with the case in which all customers are non-strategic). Moreover, our computational analysis shows that the presence of strategic customers who are willing to purchase PPSs can also be beneficial for the service provider. The reason is that by utilizing the server's idle time to increase its productivity, the service provider reduces the cost associated with customers' sojourn time without increasing its operational costs. At the same time, the ability to produce PPSs does not change the stability condition of the system (as compared with a regular system), despite the fact that it increases the server's productivity. In-depth analysis of the numerical example showed that the price discrimination level (i.e., the difference between the price of a PPS and the price of an FS) has a stronger positive effect (as reflected in percentage increase) on non-strategic customers' utility than on the utility of strategic customers. This result is explained by the fact that purchasing a PPS entails some loss of utility for strategic customers, as the quality of the product they receive is lower than that of a freshly-prepared product, whereas non-strategic customers, who do not purchase PPSs themselves, purely benefit when others do so. Notably, though the service provider is the one who controls the decision variables—and is focused primarily on his or her own self-interest (i.e., profit maximization)—there are some cases in which the percentage increase in the customers' utility is even higher than the percentage increase in server's expected profit.

We suggest several possible directions for further research in this domain. One option is to extend our two-stage operational model to a multi-server system or to a system with a limited customer queue capacity. Another interesting direction would be to investigate a system with durations that follow general distributions instead of Poisson/exponential distributions. Analysis of such a framework would require the use of simulations due to lack of mathematical tractability, where the Markovian model could be evaluated as an approximation for the general case.

Acknowledgements

This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 1448/17).

Appendix A. Notation

n	inventory capacity (the maximum number of PPSs the server is able to store; this is a decision variable of the service provider)
V_f	value of a freshly-prepared food item ("fresh service"; FS) from the customer's perspective

V_p	value of a PPS from the customer's perspective
m	number of customers in the system from which a strategic customer will prefer to take a PPS (joining-the-queue threshold, a decision variable of a customer)
M	number of customers in the system from which a strategic customer will balk (balking threshold, a decision variable of a customer)
λ	fastidious customers' arrival rate
η	strategic customers' arrival rate
μ	service rate
α	production rate of PPSs
θ	deterioration rate of an inventoried PPS
L	number of customers present in the system
S	number of PPSs in the system
$p_{i,j} = P(L = i, S = j)$	joint probability distribution function of the system states in steady state
Q	infinitesimal generator matrix
fas	denotes a fastidious customer
str	denotes a strategic customer
uns	denotes a customer of any type ("unspecified")
$L^{[y]}$	mean number of type $y \in \{fas, str, uns\}$ customers in the system
$L_q^{[y]}$	mean number of type $y \in \{fas, str, uns\}$ customers in the queue
$W^{[y]}$	mean sojourn time of a type $y \in \{fas, str, uns\}$ customer in the system
$W_q^{[y]}$	mean waiting time of a type $y \in \{fas, str, uns\}$ customer in the queue
$\overline{S}$	mean number of PPSs in the system
$\overline{T}$	mean duration of time that a PPS resides in the system
η_{eff}	effective rate at which strategic customers join the queue
η_{PPS}	rate at which strategic customers take PPSs
Γ_{eff}	total effective rate at which customers join the queue
α_{eff}	effective production rate of PPSs
θ_{eff}	effective deterioration rate of inventoried PPSs
b	selling price (the same for FS and PPS)
c	sojourn cost per unit of time from the customer's perspective
d	price discrimination value for a PPS in monetary units
m_d	joining the queue threshold as a function of d
c_L	sojourn cost per unit of time per customer, from the service provider's perspective
c_r	cost of raw materials and labor for preparation of a single unit
c_n	cost rate per unit of PPS capacity (e.g., storage equipment depreciation, maintenance)
ϕ	loss of reputation (in monetary units) and of future purchases associated with a customer who balks
η_{balk}	average balking rate of strategic customers

Appendix B. Matrices used to construct the matrix Q

$$B_0 = \begin{pmatrix} -(\alpha+\lambda+\eta) & \alpha & 0 & \cdots & 0 & 0 \\ \theta & -(\alpha+\lambda+\eta+\theta) & \alpha & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & (n-1)\theta & -(\alpha+\lambda+\eta+(n-1)\theta) & \alpha \\ 0 & 0 & 0 & \cdots & n\theta & -(\lambda+\eta+n\theta) \end{pmatrix}$$

$$B_1 = \begin{pmatrix} \lambda+\eta & 0 & \cdots & 0 & 0 \\ 0 & \lambda+\eta & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & & \lambda+\eta & 0 \\ 0 & 0 & \cdots & 0 & \lambda+\eta \end{pmatrix}$$

$$B_2 = \begin{pmatrix} -(\mu+\lambda+\eta) & 0 & 0 & \cdots & 0 & 0 \\ \theta & -(\mu+\lambda+\eta+\theta) & 0 & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & (n-1)\theta & -(\mu+\lambda+\eta+(n-1)\theta) & 0 \\ 0 & 0 & 0 & \cdots & n\theta & -(\mu+\lambda+\eta+n\theta) \end{pmatrix}$$

$$B_3 = \begin{pmatrix} -(\mu+\lambda+\eta) & 0 & 0 & \cdots & 0 & 0 \\ \eta+\theta & -(\mu+\lambda+\eta+\theta) & 0 & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \eta+(n-1)\theta & -(\mu+\lambda+\eta+(n-1)\theta) & 0 \\ 0 & 0 & 0 & \cdots & \eta+n\theta & -(\mu+\lambda+\eta+n\theta) \end{pmatrix}$$

$$B_4 = \begin{pmatrix} \lambda+\eta & 0 & \cdots & 0 & 0 \\ 0 & \lambda & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & & \lambda & 0 \\ 0 & 0 & \cdots & 0 & \lambda \end{pmatrix} \quad A_0 = \begin{pmatrix} \lambda & 0 & \cdots & 0 & 0 \\ 0 & \lambda & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & & \lambda & 0 \\ 0 & 0 & \cdots & 0 & \lambda \end{pmatrix} \quad A_2 = \begin{pmatrix} \mu & 0 & \cdots & 0 & 0 \\ 0 & \mu & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & & \mu & 0 \\ 0 & 0 & \cdots & 0 & \mu \end{pmatrix}$$

$$A_1 = \begin{pmatrix} -(\mu+\lambda) & 0 & 0 & \cdots & 0 & 0 \\ \eta+\theta & -(\mu+\lambda+\eta+\theta) & 0 & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \eta+(n-1)\theta & -(\mu+\lambda+\eta+(n-1)\theta) & 0 \\ 0 & 0 & 0 & \cdots & \eta+n\theta & -(\mu+\lambda+\eta+n\theta) \end{pmatrix}$$

$$B = \begin{pmatrix} -(\alpha+\lambda+\eta) & \alpha & 0 & \cdots & 0 & 0 \\ \eta+\theta & -(\alpha+\lambda+\eta+\theta) & \alpha & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \eta+(n-1)\theta & -(\alpha+\lambda+\eta+(n-1)\theta) & \alpha \\ 0 & 0 & 0 & \cdots & \eta+n\theta & -(\lambda+\eta+n\theta) \end{pmatrix}$$

Appendix C. Matrix Q for special cases

Matrix Q for $m = 0$, $M = 0$:

$$Q = \begin{pmatrix} B & A_0 & 0 & 0 & 0 & \cdots \\ A_2 & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ \vdots & & \ddots & \ddots & \ddots & \end{pmatrix}$$

Matrix Q for $m = 0$, $M > 0$:

$$Q = \begin{pmatrix} B & B_4 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ A_2 & B_3 & B_4 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ 0 & A_2 & B_3 & B_4 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ \vdots & & \ddots & \ddots & \ddots & 0 & 0 & 0 & 0 & \cdots \\ 0 & \cdots & 0 & 0 & \cdots & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & A_2 & A_1 & A_0 & \\ \vdots & & \vdots & \vdots & & \vdots & & \ddots & \ddots & \ddots \end{pmatrix}$$

Matrix Q for $m > 0$, $M = m$:

$$Q = \begin{pmatrix} B_0 & B_1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ A_2 & B_2 & B_1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ 0 & A_2 & B_2 & B_4 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ \vdots & & \ddots & \ddots & \ddots & 0 & 0 & 0 & 0 & \cdots \\ 0 & \cdots & 0 & 0 & \cdots & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & A_2 & A_1 & A_0 & \\ \vdots & & \vdots & \vdots & & \vdots & & \ddots & \ddots & \ddots \end{pmatrix}$$

Appendix D. Proof of Theorem 1

We start by obtaining the explicit expressions of $[m_{i,j}]_{(n+1)\times(n+1)} \equiv A_0 + RA_1 + R^2A_2$:

$$m_{i,0} = \begin{cases} \lambda - (\lambda+\mu)r_{0,0} + (\eta+\theta)r_{0,1} + \mu\sum_{k=0}^{n} r_{0,k}r_{k,0} & i = 0 \\ -(\lambda+\mu)r_{i,0} + (\eta+\theta)r_{i,1} + \mu\sum_{k=0}^{n} r_{i,k}r_{k,0} & i \neq 0 \end{cases} \tag{A1}$$

$$m_{i,j} = \begin{cases} \lambda - (\lambda+\mu+\eta+j\theta)r_{j,j} + (\eta+(j+1)\theta)r_{j,j+1} + \mu\sum_{k=0}^{n} r_{j,k}r_{k,j} & i = j \\ -(\lambda+\mu+\eta+j\theta)r_{i,j} + (\eta+(j+1)\theta)r_{i,j+1} + \mu\sum_{k=0}^{n} r_{i,k}r_{k,j} & i \neq j \end{cases}, j = 1, 2, \ldots, n-1 \tag{A2}$$

$$m_{i,n} = \begin{cases} \lambda - (\lambda+\mu+\eta+n\theta)r_{n,n} + \mu\sum_{k=0}^{n} r_{n,k}r_{k,n} & i = n \\ -(\lambda+\mu+\eta+n\theta)r_{i,n} + \mu\sum_{k=0}^{n} r_{i,k}r_{k,n} & i \neq n \end{cases} \tag{A3}$$

In what follows we show how to obtain an explicit solution to Eq. (1), which can be written as $[m_{ij}]_{(n+1)\times(n+1)} = 0$.

(i) We first prove that $r_{i,j} = 0, i < j$. According to Neuts (1981), starting from state (v, i) for any $v \geq M$, $r_{i,j}$ is equal to the expected number of visits in state $(v+1, j)$ before the process first re-enters level v. Since, for $i < j$, there is no feasible way to visit the state $(v+1, j)$ before re-entering level v (see Fig. 1), $r_{i,j} = 0$ for $i < j$.

(ii) Next we calculate $r_{j,j}$, $\forall j$. By Eq. (1) and (A1)-(A3) for $i = j$ we get

$$\lambda - (\lambda+\mu)r_{0,0} + (\eta+\theta)r_{0,1} + \mu\sum_{k=0}^{n} r_{0,k}r_{k,0} = 0, \tag{A4}$$

$$\lambda - (\lambda+\mu+\eta+j\theta)r_{j,j} + (\eta+(j+1)\theta)r_{j,j+1} + \mu\sum_{k=0}^{n} r_{j,k}r_{k,j} = 0, \tag{A5}$$

$$\lambda - (\lambda+\mu+\eta+n\theta)r_{n,n} + \mu\sum_{k=0}^{n} r_{n,k}r_{k,n} = 0. \tag{A6}$$

By (i) $r_{j,j+1} = 0$ for $\forall j$, $r_{j,k} = 0$ for $j < k$ and $r_{k,j} = 0$ for $k < j$, which implies $\sum_{k=0}^{n} r_{j,k}r_{k,j} = r_{j,j}^2$ for $\forall j$. Thus Eq. (A4)-(A6) reduce to

$$\lambda - (\lambda+\mu)r_{0,0} + \mu r_{0,0}^2 = 0, \tag{A7}$$

$$\lambda - (\lambda+\mu+\eta+j\theta)r_{j,j} + \mu r_{j,j}^2 = 0, j = 1, 2, \ldots, n \tag{A8}$$

Solving Eq. (A7)-(A8) leads to

$r_{0,0} = \lambda/\mu$ and $r_{j,j} = \frac{\lambda+\mu+\eta+j\theta-\sqrt{(\lambda+\mu+\eta+j\theta)^2-4\lambda\mu}}{2\mu}$, $j = 1, 2, \ldots, n$.

(iii) Finally, we calculate $r_{i,j}$, $i > j$. By Eq. (1) and (A1)-(A2) for $i \neq j$ we get

$$-(\lambda+\mu+\eta+j\theta)r_{i,j} + (\eta+(j+1)\theta)r_{i,j+1} + \mu\sum_{k=0}^{n} r_{i,k}r_{k,j} = 0, 2 \leq i \leq n, \quad 1 \leq j \leq i-1, \tag{A9}$$

$$-(\lambda+\mu)r_{i,0} + (\eta+\theta)r_{i,1} + \mu\sum_{k=0}^{n} r_{i,k}r_{k,0} = 0, 1 \leq i \leq n. \tag{A10}$$

By (i) $r_{i,k} = 0$ for $i < k$ and $r_{k,j} = 0$ for $k < j$, which implies $\sum_{k=0}^{n} r_{i,k}r_{k,j} = \sum_{k=j}^{i} r_{i,k}r_{k,j}$ for $\forall j$. Thus, Eq. (A9)-(A10) can be written as

$$-(\lambda+\mu+\eta+j\theta)r_{i,j} + (\eta+(j+1)\theta)r_{i,j+1} + \mu\left(r_{i,j}r_{j,j} + \sum_{k=j+1}^{i-1} r_{i,k}r_{k,j} + r_{i,i}r_{i,j}\right) = 0, 2 \leq i \leq n, \ 1 \leq j \leq i-1, \tag{A11}$$

$$-(\lambda+\mu)r_{i,0} + (\eta+\theta)r_{i,1} + \mu\left(r_{i,0}r_{0,0} + \sum_{k=0}^{i} r_{i,k}r_{k,0} + r_{i,i}r_{i,0}\right) = 0, 1 \leq i \leq n \tag{A12}$$

Solving Eq. (A11)-(A12) leads to.

$r_{i,j} = \frac{(\eta+(j+1)\theta)r_{i,j+1}+\mu\sum_{k=j+1}^{i-1} r_{i,k}r_{k,j}}{\lambda+\mu+\eta+j\theta-\mu(r_{i,i}+r_{j,j})}$, $2 \leq i \leq n$, $1 \leq j \leq i-1$, and $r_{i,0} = \frac{(\eta+\theta)r_{i,1}+\mu\sum_{k=1}^{i-1} r_{i,k}r_{k,0}}{\mu(1-r_{i,i})}$, $1 \leq i \leq n$. This completes the Proof.

Appendix E. Formulae for computing various performance measures

$$\Gamma_{eff} = \lambda + \eta_{eff}$$

$$W^{[uns]} = L^{[uns]} / \Gamma_{eff}$$

$$L_q^{[uns]} = \sum_{i=1}^{\infty}(i-1)p_{i,\cdot} = L^{[uns]} - (1 - p_{0,\bullet})$$

$$W_q^{[uns]} = L_q^{[uns]} \Big/ \Gamma_{eff}$$

$$W^{[fas]} = \sum_{i=0}^{\infty}\left(\frac{i+1}{\mu}\vec{p}_i \cdot \vec{e}\right) = \frac{1}{\mu}\left(\sum_{i=0}^{M-1}(i+1)p_{i,\cdot} + \vec{p}_M\left[[I-R]^{-2} + M[I-R]^{-1}\right]\vec{e}\right)$$

$$L^{[fas]} = \lambda W^{[fas]}$$

$$W_q^{[fas]} = \sum_{i=0}^{\infty}\left(\frac{i}{\mu}\vec{p}_i \cdot \vec{e}\right) = \frac{1}{\mu}\left(\sum_{i=0}^{M-1} i p_{i,\cdot} + \vec{p}_M\left[[I-R]^{-2} + (M-1)[I-R]^{-1}\right]\vec{e}\right)$$

$$L_q^{[fas]} = \lambda W_q^{[fas]}$$

$$W^{[str]} = \frac{1}{\mu}\left(\sum_{i=0}^{m-1}(i+1)p_{i,\cdot} + \sum_{i=m}^{M-1}(i+1)p_{i,0}\right)$$

$$L^{[str]} = \eta_{eff} W^{[str]}$$

$$W_q^{[str]} = \frac{1}{\mu}\left(\sum_{i=0}^{m-1} ip_{i,\cdot} + \sum_{i=m}^{M-1} ip_{i,0}\right)$$

$$L_q^{[str]} = \eta_{eff} W_q^{[str]}$$

$$\overline{S} = \sum_{i=0}^{\infty}(\overrightarrow{p}_i \cdot \overrightarrow{v}) = \sum_{i=0}^{M-1}(\overrightarrow{p}_i \cdot \overrightarrow{v}) + \overrightarrow{p}_M [I - R]^{-1}\overrightarrow{v}$$

$$\overline{T} = \overline{S}/\alpha_{eff}$$

References

Adacher, L., Cassandras, C.G., 2014. Lot size optimization in manufacturing systems: the surrogate method. Int. J. Prod. Econ. 155, 418–426.

Afèche, P., Pavlin, J.M., 2016. Optimal price/lead-time menus for queues with customer choice: segmentation, pooling, and strategic delay. Manag. Sci. 62 (8), 2412–2436.

Akan, M., Ata, B.Ş., Olsen, T., 2012. Congestion-based lead-time quotation for heterogenous customers with convex-concave delay costs: optimality of a cost-balancing policy based on convex hull functions. Oper. Res. 60 (6), 1505–1519.

Altendorfer, K., Minner, S., 2015. Influence of order acceptance policies on optimal capacity investment with stochastic customer required lead times. Eur. J. Oper. Res. 243 (2), 555–565.

Avinadav, T., 2020. The effect of decision rights allocation on a supply chain of perishable products under a revenue-sharing contract. Int. J. Prod. Econ. https://doi.org/10.1016/j.ijpe.2019.107587. In press.

Avinadav, T., Arponen, T., 2009. An EOQ model for items with a fixed shelf-life and a declining demand rate based on time-to-expiry. Asia Pac. J. Oper. Res. 26 (6), 759–767.

Avinadav, T., Chernonog, T., Lahav, Y., Spiegel, U., 2017. Dynamic pricing and promotion expenditures in an EOQ model of perishable items. Ann. Oper. Res. 248, 75–91.

Avinadav, T., Herbon, A., Spiegel, U., 2013. Optimal inventory policy for a perishable item with demand function sensitive to price and time. Int. J. Prod. Econ. 144, 497–506.

Avinadav, T., Herbon, A., Spiegel, U., 2014. Optimal ordering and pricing policy for demand functions that are separable into price and inventory age. Int. J. Prod. Econ. 155, 406–417.

Avşar, Z.M., Zijm, W.H., 2014. Approximate queueing models for capacitated multi-stage inventory systems under base-stock control. Eur. J. Oper. Res. 236 (1), 135–146.

Benjaafar, S., Cooper, W.L., Mardan, S., 2011. Production-inventory systems with imperfect advance demand information and updating. Nav. Res. Logist. 58 (2), 88–106.

Berk, E., Gürler, Ü., 2008. Analysis of the (Q,r) inventory model for perishables with positive lead times and lost sales. Oper. Res. 56 (5), 1238–1246.

Berman, O., Sapna, K.P., 2002. Optimal service rates of a service facility with perishable inventory items. Nav. Res. Logist. 49 (5), 464–482.

Boudali, O., Economou, A., 2012. Optimal and equilibrium balking strategies in the single server Markovian queue with catastrophes. Eur. J. Oper. Res. 218 (3), 708–715.

Bountali, O., Economou, A., 2017. Equilibrium joining strategies in batch service queueing systems. Eur. J. Oper. Res. 260 (3), 1142–1151.

Cancho, V.G., Louzada-Neto, F., Barriga, G.D., 2011. The Poisson-exponential lifetime distribution. Comput. Stat. Data Anal. 55 (1), 677–686.

Chao, G.H., Iravani, S.M., Savaskan, R.C., 2009. Quality improvement incentives and product recall cost sharing contracts. Manag. Sci. 55 (7), 1122–1138.

Chao, X., Gong, X., Shi, C., Zhang, H., 2015. Approximation algorithms for perishable inventory systems. Oper. Res. 63 (3), 585–601.

Chebolu-Subramanian, V., Gaukler, G.M., 2015. Product contamination in a multi-stage food supply chain. Eur. J. Oper. Res. 244 (1), 164–175.

Chen, X., Pang, Z., Pan, L., 2014. Coordinating inventory control and pricing strategies for perishable products. Oper. Res. 62 (2), 284–300.

Chernonog, T., 2020. Inventory and marketing policy in a supply chain of a perishable product. Int. J. Prod. Econ. 219, 259–274.

Chernonog, T., Avinadav, T., 2019. Pricing and advertising in a supply chain of perishable products under asymmetric information. Int. J. Prod. Econ. 209, 249–264.

Cooper, W.L., 2001. Pathwise properties and performance bounds for a perishable inventory system. Oper. Res. 49 (3), 455–466.

Feng, L., Chan, Y.L., Cárdenas-Barrón, L.E., 2017. Pricing and lot-sizing polices for perishable goods when the demand depends on selling price, displayed stocks, and expiration date. Int. J. Prod. Econ. 185, 11–20.

Flapper, S.D.P., Gayon, J.P., Vercraene, S., 2012. Control of a production–inventory system with returns under imperfect advance return information. Eur. J. Oper. Res. 218 (2), 392–400.

Guo, P., Hassin, R., 2012. Strategic behavior and social optimization in Markovian vacation queues: the case of heterogeneous customers. Eur. J. Oper. Res. 222 (2), 278–286.

Hanukov, G., Avinadav, T., Chernonog, T., Spiegel, U., Yechiali, U., 2017. A queueing system with decomposed service and inventoried preliminary services. Appl. Math. Model. 47, 276–293.

Hanukov, G., Avinadav, T., Chernonog, T., Spiegel, U., Yechiali, U., 2018a. Improving efficiency in service systems by performing and storing "preliminary services". Int. J. Prod. Econ. 197, 174–185.

Hanukov, G., Avinadav, T., Chernonog, T., Yechiali, U., 2018b. Performance improvement of a service system via stocking perishable preliminary services. Eur. J. Oper. Res. 274 (3), 1000–1011.

Hanukov, G., Avinadav, T., Chernonog, T., Yechiali, U., 2019. A multi-server queueing-inventory system with stock-dependent demand. IFAC-PapersOnLine 52 (13), 671–676.

Hanukov, G., Yechiali, U., 2019. Explicit solutions for continuous-time QBD processes by using relations between matrix geometric analysis and the probability generating functions method. Probab. Eng. Inf. Sci. https://doi.org/10.1017/S0269964819000470. Published online. DOI.

Hassin, R., 2016. Rational Queueing. Chapman and Hall/CRC..

Hassin, R., Haviv, M., 2003. To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems (Vol. 59). Springer Science & Business Media, New York.

Hassin, R., Roet-Green, R., 2017. The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. Oper. Res. 65 (3), 804–820.

Haviv, M., Oz, B., 2017. Self-regulation of an unobservable queue. Manag. Sci. 64 (5), 2380–2389.

Herbon, A., 2014. Dynamic pricing vs. acquiring information on consumers' heterogeneous sensitivity to product freshness. Int. J. Prod. Res. 52 (3), 918–933.

Herbon, A., 2017. A non-cooperative game model for managing a multiple-aged expiring inventory under consumers' heterogeneity to price and time. Appl. Math. Model. 51, 38–57.

Herbon, A., Khmelnitsky, E., 2017. Optimal dynamic pricing and ordering of a perishable product under additive effects of price and time on demand. Eur. J. Oper. Res. 260 (2), 546–556.

Hu, M., Li, Y., Wang, J., 2017. Efficient ignorance: information heterogeneity in a queue. Manag. Sci. 64 (6), 2650–2671.

Hu, P., Shum, S., Yu, M., 2015. Joint inventory and markdown management for perishable goods with strategic consumer behavior. Oper. Res. 64 (1), 118–134.

Iravani, S.M., Kolfal, B., Van Oyen, M.P., 2011. Capability flexibility: a decision support methodology for parallel service and manufacturing systems with flexible servers. IIE Trans. 43 (5), 363–382.

Jeganathan, K., Reiyas, M.A., Padmasekaran, S., Lakshmanan, K., 2017. An $M/E_K/1/N$ Queueing-inventory system with two service rates based on queue lengths. Int. J. Algorithm. Comput. Math. 3 (1), 357–386.

Kerner, Y., 2011. Equilibrium joining probabilities for an M/G/1 queue. Game. Econ. Behav. 71 (2), 521–526.

Kleinrock, L., 1975. Queueing systems. John Wiley & Sons 1 (New York).

Kouki, C., Babai, M.Z., Jemai, Z., Minner, S., 2016. A coordinated multi-item inventory system for perishables with random lifetime. Int. J. Prod. Econ. 181, 226–237.

Kouki, C., Babai, M.Z., Minner, S., 2018. On the benefits of emergency orders in perishable inventory systems. Int. J. Prod. Econ. 204, 1–17.

Krishnamoorthy, A., Manikandan, R., Lakshmy, B., 2015. A revisit to queueing-inventory system with positive service time. Ann. Oper. Res. 233 (1), 221–236.

Levin, Y., McGill, J., Nediak, M., 2010. Optimal dynamic pricing of perishable items by a monopolist facing strategic consumers. Prod. Oper. Manag. 19 (1), 40–60.

Li, Q., Guo, P., Li, C.L., Song, J.S., 2016a. Equilibrium joining strategies and optimal control of a make-to-stock queue. Prod. Oper. Manag. 25 (9), 1513–1527.

Li, Q., Yu, P., Wu, X., 2016b. Managing perishable inventories in retailing: replenishment, clearance sales, and segregation. Oper. Res. 64 (6), 1270–1284.

Li, R., Teng, J.T., 2018. Pricing and lot-sizing decisions for perishable goods when demand depends on selling price, reference price, product freshness, and displayed stocks. Eur. J. Oper. Res. 270 (3), 1099–1108.

Mandelbaum, A., Yechiali, U., 1983. Optimal entering rules for a customer with wait option at an M/G/1 queue. Manag. Sci. 29 (2), 174–187.

Manou, A., Economou, A., Karaesmen, F., 2014. Strategic customers in a transportation station: when is it optimal to wait? Oper. Res. 62 (4), 910–925.

Nair, A.N., Jacob, M.J., Krishnamoorthy, A., 2015. The multi server $M/M/(s, S)$ queueing inventory system. Ann. Oper. Res. 233 (1), 321–333.
Naor, P., 1969. The regulation of queue size by levying tolls. Econometrica: journal of the Econometric Society 15–24.
National Restaurant Association, 2017. Restaurant industry pocket factbook. National Restaurant Association, Washington, DC, p. 2017. Available at: https://www.restaurant.org/Downloads/PDFs/News-Research/Pocket_Factbook_FEB_2017-FINAL.pdf.
Neuts, M.F., 1981. Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach.
Priceconomics, 2017. How much do the ingredients cost in your favorite foods? Forbes. Available at: https://www.forbes.com/sites/priceonomics/2017/04/07/how-much-do-the-ingredients-cost-in-your-favorite-foods/#f665f4111eda.
Rajan, A., Rakesh, Steinberg, R., 1992. Dynamic pricing and ordering decisions by a monopolist. Manag. Sci. 38 (2), 240–262.
Sato, K., 2019. Price trends and dynamic pricing in perishable product market consisting of superior and inferior firms. Eur. J. Oper. Res. 274 (1), 214–226.
Shi, Y., Lian, Z., 2016. Optimization and strategic behavior in a passenger–taxi service system. Eur. J. Oper. Res. 249 (3), 1024–1032.
Shone, R., Knight, V.A., Harper, P.R., Williams, J.E., Minty, J., 2016. Containment of socially optimal policies in multiple-facility Markovian queueing systems. J. Oper. Res. Soc. 67 (4), 629–643.
Wang, T.Y., Chang, F.M., 2016. Dynamic control policy for the discrete-time queue. Int. J. Serv. Oper. Inf. 8 (2), 79–93.
Wang, J., Cui, S., Wang, Z., 2019. Equilibrium strategies in M/M/1 priority queues with balking. Prod. Oper. Manag. 28 (1), 43–62.
Wang, F., Wang, J., Zhang, Z.G., 2017. Strategic behavior and social optimization in a double-ended queue with gated policy. Comput. Ind. Eng. 114, 264–273.
Yechiali, U., 1971. On optimal balking rules and toll charges in the GI/M/1 queuing process. Oper. Res. 19 (2), 349–370.
Yechiali, U., 1972. Customers' optimal joining rules for the GI/M/s queue. Manag. Sci. 18 (7), 434–443.
Yu, Q., Allon, G., Bassamboo, A., 2016. How do delay announcements shape customer behavior? An empirical study. Manag. Sci. 63 (1), 1–20.
Zhang, H., Shi, C., Chao, X., 2016. Approximation algorithms for perishable inventory systems with setup costs. Oper. Res. 64 (2), 432–440.
Zhao, N., Lin, Z.T., 2011. A queueing-inventory system with two classes of customers. Int. J. Prod. Econ. 129, 225–231.
Zhao, W., Zheng, Y.S., 2000. Optimal dynamic pricing for perishable assets with nonhomogeneous demand. Manag. Sci. 46 (3), 375–388.
Ziani, S., Rahmoune, F., Radjef, M.S., 2015. Customers' strategic behavior in batch arrivals M2/M/1 queue. Eur. J. Oper. Res. 247 (3), 895–903.