Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimal Signaling Mechanisms in Unobservable Queues

David Lingenbrink, Krishnamurthy Iyer

informs
http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 67, No. 5, September–October 2019, pp. 1397–1416
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Optimal Signaling Mechanisms in Unobservable Queues

David Lingenbrink,[a] Krishnamurthy Iyer[a]

[a] School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853

Contact: dal299@cornell.edu, https://orcid.org/0000-0001-6214-9941 (DL); kriyer@cornell.edu, https://orcid.org/0000-0002-5538-1432 (KI)

Received: November 8, 2017
Revised: July 10, 2018
Accepted: September 4, 2018
Published Online in Articles in Advance: August 6, 2019

Subject Classifications: games/group decisions; queues; linear programming
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2018.1819

Abstract. We consider the problem of optimal information sharing in an unobservable single-server queue offering service at a fixed price to a Poisson arrival of delay-sensitive customers. The service provider observes the queue and may share state information with arriving customers. The customers, who are Bayesian and strategic, incorporate this information into their beliefs before deciding whether to join the queue. We pose the following question: Which signaling mechanism should the service provider adopt to maximize her expected revenue? We formulate this problem as an infinite linear program in the queue's steady-state distribution and establish that, in general, the optimal signaling mechanism requires the service provider to strategically conceal information in order to incentivize customers to join. In particular, we show that a binary signaling mechanism with a threshold structure is optimal. Finally, we prove that coupled with an optimal fixed price, the optimal signaling mechanism generates the same expected revenue as the optimal state-dependent pricing mechanism. This suggests that in settings where state-dependent pricing is infeasible, signaling can be effective in achieving the optimal revenue. Our work contributes to the literature on dynamic Bayesian persuasion and provides many interesting directions for extensions.

History: First Place Winner (Krishnamurthy Iyer), Junior Faculty Forum Paper Competition, 2017.
Funding: Financial support from the National Science Foundation Division of Civil, Mechanical and Manufacturing Innovation [Grants CMMI-1462592 and CMMI-1633920] is greatly appreciated.

Keywords: games: Bayesian persuasion • queues: strategic customers • programming: infinite linear program

1. Introduction

In many services systems, where resources to serve users are often costly and limited, the user experience depends on the state of the system, namely, the resource availability, wait times, the level of congestion, etc. As an example, in a call center, the wait time until service affects a caller's experience. Similarly, in a ride-hailing service, the availability of drivers in a ride-requester's neighborhood directly influences the time until the requester begins her ride, thereby affecting her utility. When resource availability is too low or the wait times are high, the users in the system might prefer to not avail the service, and perhaps instead choose an outside option. For example, if the time until the beginning of a ride is too long, a ride-requester may instead choose to use public transport.

However, as compared with the service providers, the users of such services typically have far less information about the system state. A call center may know the number and the nature of other requests currently on hold, whereas such information is not available to a caller. Similarly, in a ride-hailing service, the platform has access to the number of drivers and their location around a ride-requester's neighborhood, whereas the requester a priori does not, unless informed by the platform. Without the current state information, a user may choose to obtain service when the system is in a poor state and experience a low quality of service. One main goal of such systems is to minimize the occurrence of instances with poor quality of service, while maintaining revenue goals by providing service to a large number of users.

A common approach toward achieving this goal is to price the service based on the current system state. For example, in ride-hailing platforms, the price of obtaining rides often depends on the availability of the drivers in the platform. This price provides information to ride requesters, who may choose to not use the service if the price is too high. However, in some settings, practical considerations may render such state-dependent pricing infeasible or undesirable. This may be because there is no explicit price for the service being offered, or in other cases, the variability of prices may itself act as a source of user dissatisfaction.

When state-dependent pricing is infeasible or undesirable, a service provider may instead choose to share information about the system state directly to the users to help them decide whether to avail service. For example, a call center may choose to make anticipated delay announcements to incoming callers to help them decide whether to stay on the line (Armony and Maglaras 2004a). Similarly, a ride-hailing service may choose to provide information about wait times to help ride-requesters decide whether to hail rides. A natural question that arises then is how to effectively share information with users to reliably ensure they are satisfied with the service quality, while at the same time achieving revenue or profit goals. A secondary question is to quantify how the revenue so obtained compares with that under state-dependent pricing.

In this paper, we study this problem of information sharing in the context of a service system offering service at a fixed price. Customers arriving at the system must decide whether to leave without obtaining service or to possibly join a queue to obtain service. The queue length is observable to the service provider but unobservable to the customers. Each customer is strategic and incurs a cost of waiting until service completion. Furthermore, the customers are Bayesian and incorporate any information shared by the service provider into their beliefs prior to making their decision. We consider a service provider interested in maximizing her expected revenue. We pose the following question in this setting: How should the service provider share information about the queue to incentivize participation and maximize the expected revenue in the resulting customer equilibrium?

A central assumption in our model, as opposed to previous work (Allon et al. 2011), is that the service provider can *commit* to an information-sharing mechanism. Without this commitment power, the service provider will always prefer to share (possibly false) information that maximizes the likelihood a customer joins the queue, and, consequently, there cannot be any meaningful information transmission. On the other hand, as we show, by committing to a prespecified mechanism for information sharing, the service provider can credibly convey information about the state of the system.

Note that the set of all such mechanisms is quite complex. At one extreme, the service provider may choose to fully reveal the queue length to each arriving customers. At the other extreme, the service provider may choose not to disclose any information about the queue length to the customers. But, in between these two extremes, there exists a multitude of signaling mechanisms where the service provider sends a signal correlated with queue length to the customer. Moreover, each such choice of the signaling mechanism leads to the customers responding according to an equilibrium, and one must identify their equilibrium strategies in order to determine the resulting expected revenue. This task further exacerbates the complexity of identifying the optimal signaling mechanism.

The main contribution of this work is the rigorous formulation of the service providers' decision problem and identifying, for general waiting costs, the structure of the optimal signaling mechanism. In particular, we show that the service provider's decision problem can be formulated as an infinite linear program, whose variables correspond to the steady-state distribution of the queue under a feasible signaling mechanism. By analyzing the linear program, we show that for any given fixed price, there exists an optimal signaling mechanism that uses binary signals and has a threshold structure. This structure establishes that the optimal amount of information sharing requires the service provider to strategically provide ambiguous information about the queue, where the same signal is provided over a range of values of the queue length. In particular, the optimal signaling mechanism neither fully reveals nor fully conceals information about the system state.

We summarize our main results below.

1. *Linear programming formulation*: We begin in Section 2 by formulating the service provider's decision problem as an optimization problem, where the customers' behavior is constrained to be in an equilibrium. In Section 3, we first use a revelation principle-style argument (Fudenberg and Tirole 1991, Bergemann and Morris 2019) to show that it suffices to consider binary signaling mechanisms, where the signal the service provider sends is either "join" or "leave," and the customer equilibrium involves following the service provider's recommendation. Using this structural characterization of the set of signals, we show that the service provider's decision problem can be formulated as a linear program with a countable number of variables and constraints.

2. *Optimality of threshold mechanisms*: Next, by analyzing this linear program, we establish in Section 4 that the optimal signaling mechanism has a threshold structure, where the service provider sends the join signal if the queue length is below some threshold, and leave otherwise. (In addition, at the threshold, the service provider may randomize.) We establish this result through a perturbative analysis, where any feasible solution is perturbed to a solution with better objective in two steps. Furthermore, in Section 6.1, for the special case of linear waiting costs, we use the structural characterization of the optimal mechanism to obtain closed-form expressions for the optimal value of the threshold for any fixed-price.

3. *Comparison of signaling with optimal state-dependent prices*: Finally, in Section 5, we study the service provider's problem of setting the optimal fixed price in addition to subsequently choosing the optimal signaling mechanism. Interestingly, we find that with the optimal choice of the fixed price and using the corresponding optimal signaling mechanism, the service

provider can achieve the same revenue as with the optimal state-dependent pricing mechanism in an observable queue. (Hassin and Koshman 2017 obtain this result independently for the special case of linear waiting costs.) This suggests that in settings where state-dependent pricing is not feasible, the service provider can effectively use optimal signaling to achieve revenue comparable to that under state-dependent pricing.

Our paper provides a rigorous framework for analyzing the service provider's decision problem in a variety of related models that incorporate (exogenous) abandonments and customer heterogeneity, as we discuss in Sections 6.2 and 6.3. In particular, in these models, our framework leads to analogous linear programs whose solutions determine the optimal signaling mechanism. Our structural characterization of the optimal signaling mechanism continues to hold under abandonments. When customers are heterogeneous and their types are public, we show that the optimal mechanism may lack the threshold structure. However, we prove that the threshold structure of the optimal mechanism is restored if all customers types are charged the same price, or if the prices are set optimally.

1.1. Related Work

Our methods and results contribute to the emerging literature on Bayesian persuasion (Rayo and Segal 2010; Kamenica and Gentzkow 2011; Bergemann and Morris 2016a, b, 2019) that studies settings where an informed principal strategically chooses the amount of information to share with uninformed agents to incentivize them to act in a desired manner. In contrast to the literature on *cheap talk* (Crawford and Sobel 1982), the distinctive feature in Bayesian persuasion is the assumption that the principal can *commit* to sharing information in a prespecified manner. The main insight is that, in general, the principal's optimal signal must obfuscate information by carefully coalescing favorable and unfavorable states of the agents. Kolotilin et al. (2017) extends this basic model to settings with privately informed agents who must report their types to the principal before receiving information. For a general methodological approach to Bayesian persuasion and information design in finite settings, see Bergemann and Morris (2019) and Taneva (2019).

More specifically, our work fits in the framework of Bayesian persuasion in a *dynamic* setting. Several recent papers fit this description. Kremer et al. (2014) study a setting where a group of agents must sequentially choose an action from a set of actions with unknown, but deterministic, rewards; a principal observes the reward obtained by each agent and may share information about this to the next agent in sequence, with the goal being to maximize the expected average reward across all agents. The central tension in this setting is that agents prefer to exploit given their information, whereas the principal seeks to balance exploration and exploitation. Papanastasiou et al. (2017) extend this model to allow for stochastic rewards in an infinite-horizon, decentralized multiarmed bandit setting with discounted rewards and characterize the optimal disclosure policy as a solution to a linear program. Mansour et al. (2016) study a similar model (and other more general settings) and propose a bandit algorithm that achieves asymptotically optimal regret in maximizing social welfare. Our work differs from these papers in two aspects. First, because these papers study learning in a bandit setting, the focus is on a transient analysis starting with exogenously specified priors. In contrast, we perform a steady-state analysis, which leads to the customers' prior beliefs arising endogenously in equilibrium. Second, these papers focus on social welfare maximization, whereas we analyze a setting where the principal seeks to maximize her own revenue. Finally, Ely (2017) studies Bayesian persuasion in a dynamic setting where a principal provides information about a stochastically evolving state to a myopic agent. In contrast to our work, the state evolution here is independent of the agent's actions.

Our work also ties into the long line of work on strategic behavior in queues in both observable and unobservable settings. In the seminar paper, Naor (1969) studies revenue and welfare maximizing through static pricing in an *observable* M/M/1 queue (a single server queue with Poisson arrivals and exponential service times), where customers strategically choose to join or leave on arrival. Edelson and Hilderbrand (1975) study static pricing in an *unobservable* M/M/1 queue with strategic balking and observe that the revenue-maximizing static price equals welfare-maximizing static price. Chen and Frank (2001) study state-dependent pricing in an observable queue with homogeneous customers and prove that the revenue-optimal prices also maximize the social welfare. For more detailed discussions, see the book by Hassin and Haviv (2012) or the more recent extensive survey by Hassin (2016); in the following, we discuss a few papers closely related to our model and results.

A number of papers have analyzed service systems where strategic customers are partially informed about system parameters and state (Burnetas and Economou 2007; Economou and Kanta 2008a, b), and the service provider makes announcements about delay and service quality. Armony and Maglaras (2004a) analyze a customer contact center where arriving customers choose among joining a queue to obtain service, leaving (never to return), and putting a service request for a call-back. Customers receive a state-dependent anticipated delay information before making their decision. [Armony and Maglaras (2004b) study a similar setting without the anticipated delay announcements.] The authors analyze a many-server, heavy-traffic regime and propose an asymptotically consistent delay announcement policy and an asymptotically optimal routing rule. Yu et al. (2017a) perform an empirical study on how delay announcements impact customer behavior using call-center data and observe that delay announcements directly affect customers' waiting costs. Cui and Veeraraghavan (2016) consider a setting where customers in an observable queue do not know the service parameters, such as the service rate, and have arbitrary beliefs about them. The authors compare the effects of revealing these parameters and find situations where the announcement of service parameters hurts consumer welfare. Pender et al. (2017, 2018) consider a setting where customers choosing between two queues are provided delayed queue-length information (or a moving average of queue lengths over a time window). They find that such information can lead to oscillations in the two queues if the delay is beyond a critical value. Hassin and Roet-Green (2017) study an unobservable queue where customers can obtain the queue-length information by paying a cost of inspection. The authors prove the existence and uniqueness of the equilibrium and study its properties for a range of inspection costs.

Our work is closely related to that of Allon et al. (2011), who consider an unobservable, single-server queueing system where homogeneous customers with linear waiting costs choose to join or leave on arrival, after receiving a signal from the service provider. The authors assume that the service provider sends a deterministic signal at each queue state and focus on the setting of cheap talk, where the service provider cannot commit to the signaling mechanism. Essentially, the authors identify equilibria for the setting where the service provider and the customers choose their strategies simultaneously and study their properties for a range of settings differing in the alignment of the service provider's and the customers' incentives. Yu et al. (2017b) extend this model to include heterogeneous customers. In contrast to these papers, in our model, customers have general waiting costs, and the service provider can commit to the signaling mechanism. In other words, our model analyzes the Stackelberg setting where the service provider first selects (and commits to) a possibly randomized signaling mechanism, and the customers respond knowing the signaling mechanism. Finally, whereas they consider general objectives for the service provider, we focus on the setting where the service provider's goal is to maximize revenue.

Focusing on settings where the service provider has the power to commit, Hassin (1986) compares the social welfare between an observable queue and an unobservable queue, where customers have linear waiting costs and are charged revenue-maximizing static prices in each instance. The author notes that social welfare may be higher in the observable setting but not always. Guo and Zipkin (2007) study a similar setting where the service provider can commit to one of three specific signaling mechanisms (no information, total customers in the queue, or the exact total time needed to wait in the queue). Simhon et al. (2016) study a similar model under a specific class of signaling mechanisms, where the service provider reveals the queue length when it is below a threshold and reveals no information otherwise. They show that no such signaling mechanism can strictly increase the revenue over the full-information mechanism (in the overloaded regime) or the no-information mechanism (in the underloaded regime). In contrast, our model and methods do not a priori restrict the class of signaling mechanisms, and we show that typically the optimal signaling mechanism achieves strictly higher revenue than the full-information and the no-information mechanism. Furthermore, our analysis shows in fact that the service provider obtains higher revenue by revealing the queue length when it is large and concealing it when it is short.

Recently, Hassin and Koshman (2017) analyzed the case of linear waiting costs and observed that a threshold-signaling mechanism, together with an optimal choice of the fixed price, achieves the optimal revenue. We obtain the same result for a broader class of customer waiting costs. Furthermore, we characterize the optimal mechanism for any exogenously specified fixed price.

Finally, in our results for linear waiting costs, the expression for the threshold in the optimal signaling mechanism involves the Lambert-W function. Borgs et al. (2014) obtain similar expressions involving the Lambert-W function for determining the optimal threshold in an admission-control problem in observable queue.

2. Model

Our model consists of a service provider facing a sequence of potential customers who arrive according to a Poisson process with rate $\lambda > 0$. The service provider is capacity constrained, and, consequently, customers seeking to obtain service are put into a queue. Each customer upon arrival must decide whether to join the queue to obtain service at a fixed price $p > 0$ or to leave without obtaining service. We focus on the setting where the queue is unobservable—that is, the customers cannot directly see the state of the queue before deciding whether to join or leave. On the other hand, the service provider can observe the queue and may disclose information about its state to a customer upon her arrival.

We consider a setting where the queue is served by a single server. The service discipline in the queue is first-in-first-out (FIFO). Each customer's service requirement is distributed independently and identically as an exponential distribution and, without loss of generality, has unit mean. We restrict our attention to the setting where there is no abandonment: If the customer joins the queue upon arrival, they remain until service completion. We discuss different approaches to incorporate abandonment in our model in Section 6.2.

We make the assumption that the customers are homogeneous; we later discuss extensions to heterogeneous customers in Section 6.3. In particular, we represent the expected utility obtained by a customer upon joining the queue by a function $u(X)$, where X denotes the number of customers already in the queue upon arrival of the customer. The net expected payoff obtained by the customer upon choosing to join the queue is then given by $h(X,p) \triangleq u(X) - p$. We normalize the payoff of leaving without obtaining service to zero.

We require that the function u is nonincreasing in X, with $u(0) > 0$ and $\lim_{X\to\infty} u(X) < 0$. The first condition implies that customers incur a cost for waiting (longer) in queue. The latter two conditions are to avoid trivialities: The condition $u(0) > 0$ implies that a customer will prefer to join an empty queue if the price is low enough, whereas the final condition implies that for any $p \geq 0$, there exists an M such that $h(M,p) < 0$, making the customer prefer not joining the queue if she knows the queue length is larger than M. Given these assumptions, we restrict the values of p to the set $[0, u(0)]$, and for all $p \geq 0$, let M_p denote the smallest value of M for which we have $h(M,p) < 0$.

The arrival rate λ, the service requirement distribution, the customers' utility function $u(\cdot)$, and the fixed price p are common knowledge among the customers and the service provider.

2.1. Signaling Mechanism

The service provider seeks to maximize her expected revenue and has two controls to achieve this goal: (1) the fixed price p at which the service is provided and (2) the information shared with each arriving customer regarding the state of the queue. To formally describe the latter, we next introduce the notion of a signaling mechanism. A signaling mechanism $\Sigma = (\mathcal{S}, \sigma)$ is composed of a set $\mathcal{S}$ of possible signals together with a mapping[1] $\sigma : \mathbb{N}_0 \times \mathcal{S} \to [0,1]$, satisfying $\sum_{s\in\mathcal{S}} \sigma(n,s) = 1$ for each $n \in \mathbb{N}_0$. We interpret the mapping σ as follows: When a customer arrives to the system with X customers already in queue, the service provider sends a signal $s \in \mathcal{S}$ to the arriving customer with probability $\sigma(X,s)$.

To illustrate our definition, we briefly discuss two natural signaling mechanisms that have been analyzed in the literature and serve as extreme benchmarks for comparison:

1. *No-information mechanism*: At one extreme, we have the no-information mechanism, where the service provider reveals no information about the queue state to arriving customers. This setting can be represented by a signaling mechanism $\Sigma = (\mathcal{S}, \sigma)$, where $\mathcal{S} = \{\emptyset\}$ and $\sigma(n, \emptyset) = 1$ for each $n \geq 0$. Edelson and Hilderbrand (1975) consider revenue maximization in unobservable queues (without any possibility of signaling).

2. *Fully revealing mechanism*: At the other extreme, we consider the fully-revealing mechanism, where the state of the queue is completely revealed to arriving customers. This setting can be represented in our model by a signal set $\mathcal{S} = \mathbb{N}_0$, and $\sigma(n,s) = 1$ if $s = n$ and 0 otherwise. The seminal paper by Naor (1969) studies the problem of revenue maximization in observable queues with strategic customers.

We assume that the service provider can commit to the signaling mechanism publicly and that the signaling mechanism $\Sigma = (\mathcal{S}, \sigma)$ is common knowledge among the customers.

2.2. Customer Equilibrium

The customers are strategic and Bayesian and seek to maximize their total expected payoff given their beliefs. Given a signaling mechanism $\Sigma = (\mathcal{S}, \sigma)$, a pure strategy for a customer is a function $f : \mathcal{S} \to \mathcal{A} = \{0,1\}$, that specifies, for each possible signal $s \in \mathcal{S}$, an action $f(s) \in \{0,1\}$, where 1 denotes the action of joining the queue and 0 denotes the action of leaving without obtaining service. Similarly, a mixed strategy is specified by a function $f : \mathcal{S} \to [0,1]$, where $f(s) \in [0,1]$ denotes the probability that the customer will join the queue upon observing a signal $s \in \mathcal{S}$.

Recall that the service provider publicly commits to a signaling mechanism and seeks to maximize the expected revenue resulting from the customers' response. We model the customers' response as arising endogenously from an equilibrium. More precisely, we focus on the setting of a symmetric equilibrium where all customers follow the same (mixed) strategy. This is a mild assumption; in equilibrium, the customers' actions could possibly differ only at those signals under which they are indifferent between joining and leaving. To define the equilibrium notion, we consider a customer's decision problem when all other customers follow a given strategy.

Because customers are Bayesian, to describe a customer's decision problem, we must describe her beliefs. In particular, it is sufficient to describe the customer's *prior* belief about the state of the queue upon her arrival before receiving a signal from the service provider. (The customer's *posterior* belief after receiving a signal from the service provider is obtained via Bayes' rule.) Note that these prior beliefs are determined endogenously because the state of the queue upon a customer's arrival is dependent on the actions of all customers who arrived earlier. Because customer arrival is Poisson, using the PASTA property (Wolff 1982), we conclude that a customer upon arrival would see the queue in steady state. Consequently, in equilibrium, a customer's prior belief about the state of the queue must equal the queue's steady-state distribution.

Formally, given that all customers follow a strategy f and the service provider implements a signaling mechanism $\Sigma = (\mathcal{S}, \sigma)$, the queue evolves as a continuous time birth–death chain whose transition probabilities depend on f and σ. In particular, given that there are n customers already in the queue, a new customer enters the queue at rate $\lambda \sum_{s\in\mathcal{S}} \sigma(n,s)f(s)$, whereas a customer in service leaves the queue at rate 1. We restrict our attention to those customer strategies f for which the queue is stable.[2] Let $\pi_\infty(\Sigma, f)$ denote the steady-state distribution of the queue under the signaling mechanism Σ and customers' strategies f. For notational brevity, when the context is clear, we drop the explicit dependence on Σ and f to denote the steady-state distribution by π_∞, and we let X_∞ denote a random variable distributed independently as π_∞.

Upon arrival, a customer's prior belief about the state of the queue is given by π_∞. Thus, after observing a signal s, the customer's expected payoff is given by $\mathbf{E}^{\Sigma,f}[h(X_\infty, p)|s]$, where $\mathbf{E}^{\Sigma,f}[\cdot|s]$ denotes expectation with respect to the customers' posterior beliefs conditional on the signaling mechanism Σ, the strategy f, and the observed signal s. From this expression, we conclude that the customer's optimal action is to join the queue if $\mathbf{E}^{\Sigma,f}[h(X_\infty, p)|s] > 0$, to leave if $\mathbf{E}^{\Sigma,f}[h(X_\infty, p)|s] < 0$, and any mixed action if $\mathbf{E}^{\Sigma,f}[h(X_\infty, p)|s] = 0$. This leads to the following definition of a customer equilibrium:

Definition 1. Given a price p and a signaling mechanism Σ, a customer equilibrium is a strategy f satisfying for each $s \in \mathcal{S}$,

$$f(s) = \begin{cases} 1, & \text{if } \mathbf{E}^{\Sigma,f}[h(X_\infty, p)|s] > 0, \\ 0, & \text{if } \mathbf{E}^{\Sigma,f}[h(X_\infty, p)|s] < 0, \end{cases} \tag{1}$$

and $f(s) \in [0,1]$ otherwise.

To illustrate, consider the setting where the customers' utility is linear $u(X) = 1 - c(X+1)$ for some $c \in (0,1)$ and let $p \in [0, 1-c]$. Under the fully-revealing mechanism, the equilibrium strategy is trivially given by $f(s) = 1$ if $s < (1-c-p)/c$ and 0 if $s > (1-c-p)/c$. (If $(1-c-p)/c \in \mathbb{N}_0$, then $f((1-c-p)/c)$ can take any value between 0 and 1.) On the other hand, under the no-information mechanism, the customer equilibrium strategy f can be computed to be $f(\emptyset) = \min\{(1 - c/(1-p))/\lambda\}$. (See Appendix B for the details.)

2.3. Service Provider's Decision Problem

Having defined the customer equilibrium, we are now ready to formally specify the service provider's decision problem. For a choice of the fixed price p and the signaling mechanism Σ, consider a customer equilibrium f. In steady state, the queue throughput is given by

$$\begin{aligned}\mathsf{Th}(\Sigma, f) &\triangleq \mathbf{E}^{\Sigma,f}\left[\lambda \sum_{s\in\mathcal{S}} \sigma(X_\infty, s) f(s)\right] \\ &= \lambda \sum_{n=0}^{\infty} \pi_\infty(n) \sum_{s\in\mathcal{S}} \sigma(n,s) f(s).\end{aligned}$$

The preceding equation follows from the fact that customers are arriving according to a Poisson process with rate λ and, upon arrival, see the queue in steady state. In steady state, the number of customers already in the queue is n with probability $\pi_\infty(n)$, in which case the service provider sends a signal $s \in \mathcal{S}$ with probability $\sigma(n,s)$ and the customer joins the queue with probability $f(s)$. (Note that although the throughput $\mathsf{Th}(\Sigma, f)$ does not depend on the price p explicitly, there is an implicit dependence on p through the customer equilibrium f.) Thus, the service provider's expected revenue in equilibrium is given by

$$R(p, \Sigma, f) \triangleq p \cdot \mathsf{Th}(\Sigma, f).$$

The service provider's decision problem is then to choose a fixed price p and a signaling mechanism Σ in order to maximize her expected revenue in the resulting customer equilibrium[3] f:

$$\max_p \max_\Sigma R(p, \Sigma, f) \text{ subject to } f \text{ satisfying (1).} \tag{2}$$

Our main goal is to determine the optimal fixed price and to characterize the optimal signaling mechanism (if they exist) for the decision problem (2). As a first step in our analysis, we begin by studying the inner maximization problem, where the service provider seeks to choose an optimal signaling mechanism for a given (exogenously specified) fixed price p. For a given price, the service provider's problem can be equivalently cast as a throughput maximization problem, as specified below:

$$\max_\Sigma \mathsf{Th}(\Sigma, f) \text{ subject to } f \text{ satisfying (1).} \tag{3}$$

Subsequently, in Section 5, we address the problem of determining the optimal fixed price p.

3. Characterization of the Signal Space

There are two main difficulties in analyzing the decision problem (3). First, the space of possible signaling mechanisms is quite large. In particular, we have imposed no restrictions on the set $\mathcal{S}$. To make any progress, we must obtain some characterization of the set of possible signals that an optimal signaling mechanism might use. Second, given a particular signaling mechanism, one must characterize the customer equilibrium f^{σ}. This involves solving for a fixed point of an operator implicitly defined by (1), and for a general signaling mechanism, this could be a difficult problem. Hence, we first address these difficulties.

3.1. Equilibrium Characterization

Toward the goal of characterizing the set of signals in an optimal signaling mechanism, we start by defining the notion of *equivalence* between two mechanisms and the respective customer equilibria.

Definition 2. We say two signaling mechanisms $\Sigma_i = (\mathcal{S}_i, \sigma_i)$ and corresponding customer equilibria f_i, for $i = 1, 2$ are *equivalent* if they induce the same steady-state distribution—that is, if $\pi_\infty(\Sigma_1, f_1) = \pi_\infty(\Sigma_2, f_2)$.

We have the following lemma that states that it suffices to consider signaling mechanisms where the resulting customer equilibrium is pure. We provide the proof in Appendix A.

Lemma 1. *For any fixed price p, given a signaling mechanism $\Sigma = (\mathcal{S}, \sigma)$ and a customer equilibrium f, there exists a signaling mechanism $\Sigma_1 = (\mathcal{S}_1, \sigma_1)$ with customer equilibrium f_1 such that* (1) *(Σ_1, f_1) is equivalent to (Σ, f), and* (2) *f_1 is a pure strategy.*

Using the preceding lemma, we can further restrict the class of signaling mechanisms and customer equilibria to consider. We have the following lemma that states that it is enough for the service provider to consider mechanisms with binary signals with a specific customer equilibrium. The proof of the lemma uses a revelation-principle-style argument (Fudenberg and Tirole 1991, Bergemann and Morris 2019); we include the proof in Appendix A for the sake of completeness.

Lemma 2. *For any fixed price p, given a signaling mechanism $\Sigma = (\mathcal{S}, \sigma)$ and a customer equilibrium f, there exists an equivalent signaling mechanism $\Sigma_1 = (\mathcal{S}_1, \sigma_1)$ and customer equilibrium f_1, where $\mathcal{S}_1 = \{0, 1\}$ and $f_1(s) = s$ for $s \in \mathcal{S}_1$.*

Summing up the preceding two lemmas, we conclude that in order to determine an optimal signaling mechanism, it is sufficient to consider signaling mechanisms $\Sigma = (\mathcal{S}, \sigma)$ where $\mathcal{S} = \{0, 1\}$ and for which, the customer equilibrium is given by $f(s) = s$ for $s \in \{0, 1\}$. In other words, in the optimal signaling mechanism, the service provider sends a binary signal (join or leave) depending on the queue length, and in the resulting equilibrium, each customer finds it optimal to follow the recommendation. We refer to this customer strategy as the *obedient* strategy (Bergemann and Morris 2016a, 2019) and the resulting equilibrium to be the obedient equilibrium.

Given this reduction, the service provider's decision problem, for any fixed price p, simplifies to identifying a mapping $\sigma : \mathbb{N}_0 \times \{0, 1\} \to [0, 1]$ (with the restriction that $\sigma(n, 0) = 1 - \sigma(n, 1)$ for each $n \in \mathbb{N}_0$) that maximizes the throughput:

$$\begin{aligned} &\max_{\sigma} \ \mathbf{E}^{\sigma}[\lambda \sigma(X_\infty, 1)] \\ &\text{subject to } \mathbf{E}^{\sigma}[h(X_\infty, p) | s = 1] \geq 0, \\ &\qquad \mathbf{E}^{\sigma}[h(X_\infty, p) | s = 0] \leq 0. \end{aligned} \tag{4}$$

Here, the two inequalities impose the requirement that the customers find obedience to be optimal: When the signal $s = i$ is revealed to a customer, choosing action i is indeed an optimal action for her. We thus refer to the two constraints as *obedience constraints*. Note that, because we focus on the obedient equilibrium for the customers, and the signal space is fixed to be $\mathcal{S} = \{0, 1\}$, we simplify the notation and denote the expectation by $\mathbf{E}^{\sigma}$.

3.2. Linear Programming Formulation

Observe that the preceding optimization problem (4) is quite complex: In addition to having an infinite number of variables $\{\sigma(n, i) : n \geq 0, i = 0, 1\}$, the constraints are highly nonlinear. This nonlinearity implies that optimizing directly would be difficult. In this section, we provide a reformulation of (4) as a linear program. This reformulation paves the way for analyzing the service provider's decision problem and for characterizing the structure of the optimal mechanism.

The main insight behind the reformulation is that, instead of optimizing over the signaling mechanism σ, one can optimize directly over the resulting steady-state distribution π_∞^{σ}. By doing so, the preceding nonlinear optimization problem simplifies to the following linear program in $\{\pi_\infty^{\sigma}(n) : n \geq 0\}$, albeit with a countable number of variables and constraints:

$$\max_{\pi} \ \sum_{n=1}^{\infty} \pi_n$$

$$\text{subject to} \quad \sum_{n=1}^{\infty} \pi_n h(n-1, p) \geq 0, \tag{5a}$$

$$\sum_{n=0}^{\infty} h(n, p)(\lambda \pi_n - \pi_{n+1}) \leq 0, \tag{5b}$$

$$\lambda \pi_n - \pi_{n+1} \geq 0, \quad \text{for all } n \geq 0, \tag{5c}$$

$$\sum_{i=0}^{\infty} \pi_i = 1, \qquad \pi_n \geq 0, \quad \text{for all } n \geq 0.$$

To obtain this linear program, we first write the expectations in the obedience constraints of (4) as linear functions of the steady state distribution. The constraints (5c) are obtained from the detailed balance conditions $\pi_\infty^\sigma(n)\lambda\sigma(n,1) = \pi_\infty^\sigma(n+1)$ and using the fact that $\sigma(n,1) \in [0,1]$ for each $n \in \mathbb{N}_0$. We have the following lemma that relates the two optimization problems:

Lemma 3. *For every signaling mechanism $\sigma : \mathbb{N}_0 \times \{0,1\} \to [0,1]$ feasible for (4), there exists a feasible solution $\{\pi_n : n \geq 0\}$ to (5) with the same objective value. Conversely, let $\{\pi_n : n \geq 0\}$ be feasible for (5). Then the signaling mechanism $\sigma : \mathbb{N}_0 \times \{0,1\} \to [0,1]$, defined as $\sigma(n,1) = \pi_{n+1}/(\lambda\pi_n)$ if $\pi_n > 0$ and $\sigma(n,1) = 0$ otherwise, is feasible for (4) and has the same objective value.*

The preceding lemma not only allows us to optimize over the steady-state distribution $\{\pi_n : n \geq 0\}$, but also provides a rule to determine $\sigma(n,1)$ from the optimal solution and hence recover the signaling mechanism. The proof of this lemma is given in Appendix A. With this reformulation of the service provider's problem, we are now ready to identify an optimal mechanism.

4. Structure of Optimal Mechanism

Note that the signaling mechanism still has to determine which binary signal to send at each queue length. In the following, we show that this problem has a simple structure. Toward that end, we introduce below the class of *threshold* mechanisms.

Definition 3. We define a threshold mechanism $\sigma^x : \mathbb{N}_0 \times \{0,1\} \to [0,1]$ for $x \in \mathbb{R}_+ \cup \{\infty\}$ as follows: For $x \in \mathbb{R}_+$, we have

$$\sigma^x(n,1) \triangleq \begin{cases} 1, & \text{if } n < \lfloor x \rfloor, \\ x - \lfloor x \rfloor, & \text{if } n = \lfloor x \rfloor, \\ 0, & \text{otherwise.} \end{cases}$$

Also, we define $\sigma^\infty(n,1) = 1$ for all $n \geq 0$.

With this definition in place, we have our first main result:

Theorem 1. *For any fixed price p, there exists a threshold mechanism σ^x with $x \in \mathbb{R}_+ \cup \{\infty\}$ and $x \geq M_p$ that achieves the optimal revenue.*

Proof. The proof of the theorem involves three steps. First, analyzing the constraints of the linear program (5), we show that the optimal signaling mechanism would signal a customer to join the queue if she would have joined under full information. With this structure in place, we then show that any feasible solution that does not have a threshold structure can be perturbed to obtain another feasible solution corresponding to a threshold mechanism with equal or higher throughput. Finally, in Lemma A.1, we show that the set of feasible solutions corresponding to threshold mechanisms forms a compact set under the weak topology. Because the objective of the linear program (5) is continuous under the weak topology, we conclude that an optimal signaling mechanism with a threshold structure must exist.

Recall that $M_p \in \mathbb{N}$ is defined such that $h(M_p - 1, p) \geq 0$ and $h(M_p, p) < 0$. Consider any feasible solution $\{\pi_n : n \geq 0\}$ to the linear program (5). We first show that we can construct another feasible solution with weakly higher throughput by ensuring the (5c) constraints are tight for all $n \leq M_p$. Toward that end, we define $\{\hat{\pi}_n : n \geq 0\}$ by setting

$$\hat{\pi}_n = \begin{cases} \dfrac{1}{Z}\pi_0\lambda^n, & \text{for } n \leq M_p, \\ \dfrac{1}{Z}\pi_n, & \text{for } n > M_p, \end{cases}$$

where $Z \triangleq \pi_0 \sum_{i=0}^{M_p} \lambda^i + \sum_{i=M_p+1}^{\infty} \pi_i > 0$ is the normalizing constant to ensure $\sum_{n=0}^{\infty} \hat{\pi}_n = 1$.

We first show that $\hat{\pi}$ is feasible for (5). From the feasibility of $\{\pi_n : n \geq 0\}$, it is straightforward to show that the constraints (5c) continue to hold for $\{\hat{\pi}_n : n \geq 0\}$. Furthermore, we obtain that $\pi_n \leq \pi_0\lambda^n$ for all $n \geq 0$, and hence $\hat{\pi}_n \geq \pi_n/Z$ for all $n < M_p$. Because $h(n-1,p) \geq 0$ for all $n \leq M_p$, this implies that $\{\hat{\pi}_n : n \geq 0\}$ continues to satisfy the obedience constraint (5a). To show that $\hat{\pi}$ is a feasible solution, it remains to verify that (5b) holds. For this step, note that we have

$$\begin{aligned} \sum_{n=0}^{\infty} h(n,p)(\lambda\hat{\pi}_n - \hat{\pi}_{n+1}) &= \sum_{n=0}^{M_p-1} h(n,p)(\lambda\hat{\pi}_n - \hat{\pi}_{n+1}) \\ &\quad + \sum_{n=M_p}^{\infty} h(n,p)(\lambda\hat{\pi}_n - \hat{\pi}_{n+1}) \\ &= 0 + \sum_{n=M_p}^{\infty} h(n,p)(\lambda\hat{\pi}_n - \hat{\pi}_{n+1}) \\ &\leq 0. \end{aligned}$$

Here, the second equality follows from the definition of $\hat{\pi}$, and the inequality follows from the fact that $h(n,p) < 0$ for each $n \geq M_p$ and that $\hat{\pi}$ satisfies (5c). This proves the feasibility of $\hat{\pi}$.

The difference between the objective values for the two solutions is given by

$$\sum_{n=1}^{\infty} \hat{\pi}_n - \sum_{n=1}^{\infty} \pi_n = \pi_0 - \hat{\pi}_0 = \pi_0\left(1 - \frac{1}{Z}\right).$$

Now, because $\pi_n \leq \pi_0\lambda^n$ for all n, we obtain that $Z = \pi_0 \sum_{i=0}^{M_p} \lambda^i + \sum_{i=M_p+1}^{\infty} \pi_i \geq \sum_{n=0}^{\infty} \pi_n = 1$. This implies that the objective value of $\hat{\pi}$ is at least that of π. Furthermore, unless $\pi_n = \lambda^n\pi_0$ for all $n \leq M_p$, we obtain $Z > 1$, implying that the objective value of $\hat{\pi}$ is strictly greater than that of π. From this, we conclude that in any optimal solution π, we must have $\pi_i = \lambda^i\pi_0$

for $i \le M_p$. Henceforth, we restrict ourselves to feasible solutions π satisfying this property.

Consider now a feasible solution $\pi = \{\pi_n : n \ge 0\}$ such that there exists an $N > M_p$ with $\pi_n = \lambda^n \pi_0$ for all $n < N$, $0 < \pi_N < \lambda\pi_{N-1}$ and $\pi_{N+1} > 0$. We consider a perturbation of this feasible solution and show that it remains feasible and attains the same objective value as the original feasible solution. Toward this goal, define $\tilde{\pi} = \{\tilde{\pi}_n : n \ge 0\}$ as follows:

$$\tilde{\pi}_n = \begin{cases} \pi_n, & \text{for } n < N, \\ \pi_N + \beta \sum_{i>N} \pi_i, & \text{for } n = N, \\ (1-\beta)\pi_n, & \text{for } n > N, \end{cases}$$

for some $\beta \in (0, 1]$ to be chosen later. By construction, the linear programming objective for $\tilde{\pi}$ is same as that for π, and hence it suffices to show that $\tilde{\pi}$ is feasible. Note that $\tilde{\pi}_n \ge 0$ and $\sum_{n=0}^{\infty} \tilde{\pi}_n = 1$, so $\tilde{\pi}$ is a valid distribution. Thus, for $\tilde{\pi}$ to be feasible, we need (5a), (5b), and (5c) to hold.

First, observe that

$$\begin{aligned} &\sum_{n=1}^{\infty} \tilde{\pi}_n h(n-1,p) - \sum_{n=1}^{\infty} \pi_n h(n-1,p) \\ &\quad = \beta \sum_{n>N} \pi_n h(N-1,p) - \beta \sum_{n>N} \pi_n h(n-1,p) \\ &\quad = h(N-1,p)\beta \left(\sum_{n>N} \pi_n \left(1 - \frac{h(n-1,p)}{h(N-1,p)} \right) \right) \\ &\quad \ge 0. \end{aligned}$$

The inequality follows from the fact that because $N > M_p$, we have $h(N-1,p) \le h(M_p,p) < 0$, and the fact that because $h(n,p)$ is nonincreasing in n, we have $h(n-1,p)/h(N-1,p) \ge 1$ for $n > N$. Because π is feasible for the linear program, this implies that $\tilde{\pi}$ satisfies the constraint (5a) for all $\beta \in (0,1]$.

Next, note that because $\tilde{\pi}_{n+1} = \lambda\tilde{\pi}_n$ for all $n < M_p < N$, we obtain that $\sum_{n=0}^{\infty} h(n,p)(\lambda\tilde{\pi}_n - \tilde{\pi}_{n+1}) = \sum_{n=M_p}^{\infty} \cdot h(n,p)(\lambda\tilde{\pi}_n - \tilde{\pi}_{n+1})$. Because $h(n,p) < 0$ for all $n \ge M_p$, the latter expression is nonpositive, and (5b) holds, if $\lambda\tilde{\pi}_n - \tilde{\pi}_{n+1} \ge 0$ for all n—that is, if $\tilde{\pi}$ satisfies (5c). Finally, it is straightforward to verify that $\tilde{\pi}$ satisfies (5c) for all $n \ge 0$ if it is satisfied for $n = N-1$—that is, if $\lambda\tilde{\pi}_{N-1} - \tilde{\pi}_N \ge 0$. For this condition to hold, we need $\lambda\pi_{N-1} \ge \pi_N + \beta\sum_{i>N}\pi_i$, which holds for any $\beta \in (0,1]$ satisfying $0 < \beta \le (\lambda\pi_{N-1} - \pi_N)/\sum_{i>N}\pi_i$. So, for any such value of β, we obtain that $\tilde{\pi}$ is feasible for the linear program.

Note that if $(\lambda\pi_{N-1} - \pi_N)/\sum_{i>N}\pi_i \ge 1$, then choosing $\beta = 1$ yields $\tilde{\pi}_n = 0$ for all $n > N$. On the other hand, if $(\lambda\pi_{N-1} - \pi_N)/\sum_{i>N}\pi_i < 1$, then choosing $\beta = (\lambda\pi_{N-1} - \pi_N)/\sum_{i>N}\pi_i$ yields $\lambda\tilde{\pi}_{N-1} - \tilde{\pi}_N = 0$. Thus, we obtain that any $\{\pi_n : n \ge 0\}$, where $\pi_n = \lambda^n\pi_0$ for all $n < N$, $0 < \pi_N < \lambda\pi_{N-1}$ and $\pi_{N+1} > 0$ for some $N > M_p$, can be perturbed appropriately to obtain a feasible solution $\tilde{\pi}$ with equal objective and satisfying either (1) $\tilde{\pi}_n = \lambda^n\tilde{\pi}_0$ for all $n < N$, $0 < \tilde{\pi}_N \le \lambda\tilde{\pi}_{N-1}$, and $\tilde{\pi}_n = 0$ for all $n > N$ or (2) $\tilde{\pi}_n = \lambda^n\tilde{\pi}_0$ for all $n \le N$. In the latter case, if $0 < \tilde{\pi}_{N+1} < \lambda\tilde{\pi}_N$, one can perturb $\tilde{\pi}$ analogously. By induction, this implies that if the optimum to the linear program (5) is attained, then it is attained by a feasible solution $\{\pi_n : n \ge 0\}$ for which there exists an $N \ge M_p$ with $\pi_n = \lambda^n\pi_0$ for all $n < N$, $0 < \pi_N \le \lambda\pi_{N-1}$ and $\pi_n = 0$ for all $n > N$. (Here, N could be infinite.) Hence, we restrict our attention to feasible solutions of this form.

In Lemma A.1, we show that the set of all such feasible distributions is compact (under the weak topology). Because the objective is a continuous function of the steady-state distribution, we obtain that an optimal solution of this form exists.

Summarizing, there exists an optimal solution $\{\pi_n : n \ge 0\}$ to the linear program (5) for which there exists an $N \ge M_p$ (possibly infinity) such that $\pi_n = \lambda^n\pi_0$ for all $n < N$, $0 < \pi_N \le \lambda\pi_{N-1}$ and $\pi_n = 0$ for all $n > N$. Finally, by using Lemma 3, this implies that there exists an optimal signaling mechanism σ for which there exists an $N \ge M_p$ and a $q = \pi_N/(\lambda\pi_{N-1}) \in [0,1]$ such that $\sigma(n,1) = 1$ for all $n < N$, $\sigma(N,1) = q$ and $\sigma(n,1) = 0$ for all $n > N$. If $N = \infty$, we obtain that the mechanism $\sigma = \sigma^{\infty}$ is optimal. Otherwise, we obtain that σ^{N+q} is an optimal signaling mechanism. □

The preceding theorem has an important practical implication: The optimal signaling mechanism is easy to describe and implement. More precisely, our analysis assumes that the service provider can publicly announce and commit to a signaling mechanism. Given that signaling mechanisms are arbitrary mappings over the set of all nonnegative integers, this is a strong assumption for general signaling mechanisms. However, the structure of the optimal signaling mechanism renders this assumption innocuous. In particular, the service provider can easily implement the signaling mechanism σ^x by announcing a priori the value $N = \lfloor x \rfloor$ below which customers will be deterministically recommended to join the queue, and the probability $q = x - \lfloor x \rfloor$ with which they will be recommended to join when the queue length is exactly N.

We note that our proof implies that in any optimal signaling mechanism (not necessarily threshold), no customer would be told to leave if they would have joined under full information. This follows from the fact in any optimal solution π to (5), we have $\pi_{n+1} = \lambda\pi_n$ for all $n < M_p$. Notice that whenever a customer is told to leave, they know the length of the queue is more than M_p and joining will get them negative utility. This is similar to the results of Kamenica and Gentzkow (2011), where they showed that in an optimal persuasion mechanism, whenever an agent is told to take the principal's least-preferred action, the agent knows with certainty that it is in her best interest.

5. Optimal Pricing and Signaling

Having determined the structure of the optimal signaling mechanism for any fixed price p, we next investigate the service provider's decision problem of how to set p optimally in order to maximize her revenue.

In order to understand this problem, consider first as a detour the case of optimal state-dependent pricing in a fully-observable queue. More precisely, consider the setting where the arriving customers can observe the queue length, but the service provider is allowed to charge them a price dependent on the queue length. This setting of dynamic pricing serves as a natural benchmark against which we compare the revenue obtained under the optimal fixed-price and signaling mechanism. Surprisingly, we find that setting the fixed price optimally along with using an optimal signaling mechanism suffices to achieve the optimal revenue in the observable setting. We have the following main theorem.

Theorem 2. *With the optimal value of the fixed price p and the corresponding optimal signaling mechanism, the service provider obtains the same revenue as under optimal state-dependent prices in a fully observable queue.*

Before we state the proof, we note an important practical implication of this result. In many settings, state-dependent pricing is infeasible, either due to operational reasons, such as price stickiness arising out of menu costs associated with changing prices (Sheshinski and Weiss 1977), or due to exogenous reasons such as maintaining customer expectations about the price of service (Kalwani et al. 1990). In such settings, however, it may be feasible to make recommendations to customers based on the state of the system. Our result states that in such settings, as long as the fixed price is chosen optimally, the service provider can effectively use signaling to guarantee the same optimal revenue as with optimal state-dependent pricing.

Proof of Theorem 2. Let the optimal state-dependent pricing mechanism set a price $p(n)$ for service to an arriving customer when the number of customers already in the queue is n. Under our assumption of nonincreasing utilities, the prices $\{p(n) : n \geq 0\}$ can be shown to have the following form (Chen and Frank 2001): Up to a threshold of the queue length, the service provider sets prices that extract out all the surplus of the incoming customer, making those customers indifferent between joining and leaving; beyond this threshold, the service provider sets a large price, essentially denying entry to any incoming customers. Formally, the optimal prices satisfy $p(n) = u(n)$ for all $n < \kappa$ and $p(n) = \infty$ for all $n \geq \kappa$, for an appropriately chosen $\kappa > 0$. Let π_∞^κ denote the steady-state distribution of the queue under this pricing policy, and let X_∞^κ denote an (independent) random variable distributed as π_∞^κ. Note that under optimal state-dependent prices, the service provider's expected revenue is given by $\lambda \mathbf{E}\big[\mathbf{I}\{X_\infty^\kappa < \kappa\} p(X_\infty^\kappa)\big] = \lambda \mathbf{E}\big[\mathbf{I}\{X_\infty^\kappa < \kappa\} u(X_\infty^\kappa)\big]$.

Now, for the setting of an unobservable queue, consider the fixed price $\hat{p} \triangleq \mathbf{E}\big[u(X_\infty^\kappa) | X_\infty^\kappa < \kappa\big]$, and threshold signaling mechanism σ^κ—that is, the service provider sends signal 1 (or join) if the queue length is strictly less than κ, and 0 (or leave) otherwise. We claim that with this choice of fixed price $\hat{p}$ and the signaling mechanism σ^κ, the service provider achieves the same expected revenue in the obedient equilibrium as under the optimal state-dependent mechanism.

We start by showing that under the fixed price $\hat{p}$ and the threshold signaling mechanism σ^κ, the obedient strategy forms a customer equilibrium. To see this, observe that if all customers follow the recommendation, the steady-state distribution of the queue length is indeed given by π_∞^κ. By an abuse of notation, we let X_∞^κ denote the queue length upon a particular customer's arrival. Thus, the expected payoff to the customer for joining the queue upon receiving the signal $s = 1$ is given by

$$\begin{aligned}\mathbf{E}[h(X_\infty^\kappa, \hat{p}) | s = 1] &= \mathbf{E}[u(X_\infty^\kappa) | s = 1] - \hat{p} \\ &= \mathbf{E}[u(X_\infty^\kappa) | X_\infty^\kappa < \kappa] - \hat{p} = 0,\end{aligned}$$

where the second equality follows from the fact that the signaling mechanism σ^k sends signal $s = 1$ if and only if $X_\infty^\kappa < \kappa$. This implies that the resulting steady-state distribution satisfies the first obedience constraint (5a). Similarly, the expected payoff to the customer upon receiving the signal $s = 0$ is given by

$$\begin{aligned}\mathbf{E}[h(X_\infty^\kappa, \hat{p}) | s = 0] &= \mathbf{E}[u(X_\infty^\kappa) | s = 0] - \hat{p} = u(\kappa) - \hat{p} \\ &= u(\kappa) - \mathbf{E}[u(X_\infty^\kappa) | X_\infty^\kappa < \kappa] \leq 0,\end{aligned}$$

where the second equality follows from the fact that under the steady state, the signaling mechanism sends signal $s = 0$ if and only if $X_\infty^\kappa = \kappa$, the third equality from the definition of $\hat{p}$, and the inequality holds because u is nonincreasing. This implies that the resulting steady-state distribution satisfies the second obedience constraint (5b).

Next, observe that the service provider's expected revenue is given by

$$\begin{aligned}\lambda \mathbf{E}[\sigma(X_\infty^\kappa, 1)\hat{p}] &= \lambda \mathbf{E}\big[\mathbf{I}\{X_\infty^\kappa < \kappa\}\hat{p}\big] \\ &= \lambda \mathbf{E}\big[\mathbf{I}\{X_\infty^\kappa < \kappa\}\mathbf{E}\big[u(X_\infty^\kappa) | X_\infty^\kappa < \kappa\big]\big] \\ &= \lambda \mathbf{E}\big[\mathbf{I}\{X_\infty^\kappa < \kappa\} u(X_\infty^\kappa)\big],\end{aligned}$$

where the last equality follows from the tower property of conditional expectation. Thus, we observe that the service provider's expected revenue is the same as that of the optimal state-dependent pricing mechanism.

Finally, with homogeneous customers, the optimal state-dependent pricing mechanism is welfare-maximizing (Chen and Frank 2001) with zero customer surplus. Because the optimal fixed price $\hat{p}$ and the signaling mechanism σ^{κ} achieve this revenue, these values must be optimal. Thus, we obtain that with the optimal choice of the fixed price and the corresponding optimal signaling mechanism, the service provider obtains the same revenue as the optimal state-dependent prices. □

The preceding theorem leads to a natural question: Can the service provider increase her revenue in an unobservable queue through a combination of signaling and pricing? Under such a mechanism, customers not only receive information about the queue state from the signal but also from the price. Despite this generality, our result already implies that such mechanisms cannot improve the revenue. This follows from the fact that, in a fully observable queue, the optimal state-dependent pricing mechanism is welfare-maximizing and has zero customer surplus (Chen and Frank 2001). Because customer surplus in any mechanism must be nonnegative, there cannot be any combination of signaling and pricing that achieves strictly higher revenue than the optimal state-dependent pricing mechanism (or the optimal signaling mechanism with an optimal fixed price). However, there exist many mechanisms that achieve this optimal revenue through a combination of signaling and pricing. Specifically, given any partition of the set of queue lengths for which a customer joins under the state-dependent pricing mechanism, one can construct a combined signaling-and-pricing mechanism that achieves the optimal revenue: Such a mechanism would reveal to an arriving customer which set of the partition the queue-length lies in, and charge them the expected utility conditioned on the queue length being in that set. Furthermore, as long as the prices for different sets of the partition are different, the prices can themselves act as signals. This discussion suggests that, in an unobservable queue, the service provider has a flexibility in choosing the number of prices while optimizing her revenue.

6. Extensions

In this section, we discuss a few extensions to our results and our model. First, for the special case where the customers' utility is linear in time spent in queue, we obtain a closed-form expression for the threshold in the optimal signaling mechanism as a function of the fixed-price p. Subsequently, we discuss how our model can be extended to include abandonment and customer heterogeneity.

6.1. Linear Utility

A commonly studied model for customer utility is one where the customer receives a fixed value $V > 0$ from service and incurs a disutility that is proportional to the time spent while waiting until service completion. (Naor 1969, Allon et al. 2011, Borgs et al. 2014). Because we assume that the customers' service requirements are homogeneous and have unit mean, this assumption implies that the customer utility $u(\cdot)$ is given by $u(X) = V - c(X+1)$ for all $X \geq 0$, for some value of $c > 0$ that denotes the disutility per unit time of waiting. For this utility model, Theorem 1 implies that the linear program 5 can be analytically solved, resulting in a closed-form expression for the threshold in the optimal signaling policy.

To state our results, we assume, without loss of generality, that $V = 1$, and, to avoid trivialities, we let $c \in (0,1)$. Furthermore, let $W_0(\cdot)$ and $W_{-1}(\cdot)$ denote the two real branches of the Lambert-W function, defined as the set of functions that are the inverse of $f(X) = Xe^X$. (See Borgs et al. 2014 for a detailed description.) We have the following theorem.

Theorem 3. *Suppose $u(n) = 1 - c(n+1)$ with $c \in (0,1)$. Then, for each $p \in [0, 1-c]$, the threshold mechanism σ^x is optimal for $x = N + q$, where*

$$N = \begin{cases} \left\lfloor \dfrac{2(1-p)}{c} - 1 \right\rfloor, & \text{if } \lambda = 1, \\ \infty, & \text{if } \lambda \leq 1 - \dfrac{c}{1-p}, \\ \left\lfloor \dfrac{1}{\log(\lambda)} (W_i(-\kappa e^{-\kappa}) + \kappa) \right\rfloor, & \text{otherwise}, \end{cases}$$

with $\kappa = \left(\frac{1-p}{c} - \frac{1}{1-\lambda}\right)\log(\lambda)$ and where $i = 0$ when $\lambda > 1$ and $i = -1$ when $1 - c/(1-p) < \lambda < 1$. For all values of $\lambda < \infty$, we have

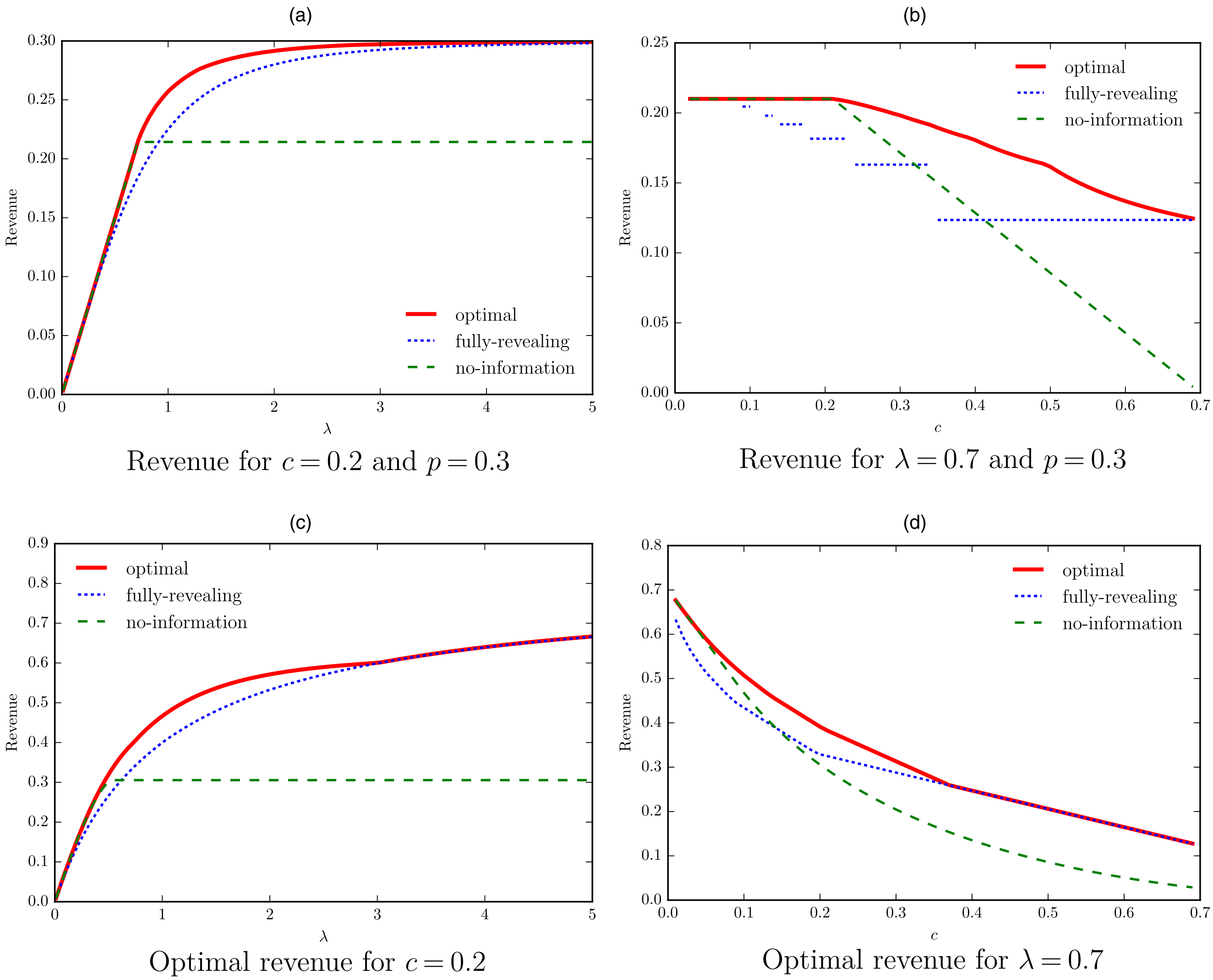
$$q = \frac{\sum_{k<N} \lambda^k (1 - p - c(k+1))}{\lambda^N (c(N+1) + p - 1)} \in [0,1].$$

The proof involves first showing that the throughput is increasing in the threshold as long as the obedient strategy is a customer equilibrium for the corresponding threshold mechanism. Then, using the equilibrium conditions for the obedient equilibrium, we obtain bounds on the optimal thresholds. We provide the full details in Appendix A.1.

Using this closed-form expression, we numerically compare the optimal signaling mechanism against those of fully-revealing and no-information mechanisms. In Figure 1(a), we plot the revenue of the optimal mechanism, along with those of the fully-revealing and the no-information mechanisms for a range of values of λ, when the customer utility is given by $u(X) = 1 - c(X+1)$ with $c = 0.2$ under a fixed price $p = 0.3$. As λ increases, the revenue of the fully-revealing mechanism and the optimal mechanism both converge to 0.3, the value when throughput is equal to 1; however, for any fixed λ, the optimal mechanism outperforms the others. Note that, for

Operations Research, 2019, vol. 67, no. 5, pp. 1397–1416, © 2019 INFORMS

Figure 1. (Color online) Comparison of the Optimal, the Fully-Revealing, and the No-Information Mechanisms



small arrival rates, the no-information mechanism outperforms the fully-revealing one; this is consistent with existing results (Simhon et al. 2016), as in this region, under the no-information mechanism, all customers join the queue, whereas under the fully-revealing mechanism, some customers will not join upon seeing a long queue.

Next, we consider the effect of changing c while fixing the values of the arrival rate ($\lambda = 0.7$) and the fixed-price ($p = 0.3$) in Figure 1(b). Notice the revenue for the fully-revealing mechanism is discontinuous: This is because under full information, customers join the queue only if the queue length is strictly less than $(1 - c - p)/c$, implying that the customer strategy is discontinuous in c. For low values of c, the revenue of the no-information and the optimal mechanism are both equal to $p\lambda = 0.21$, the maximal value, because all customers join the queue. As c increases, the revenue of the no-information mechanism goes to zero, whereas the revenue of fully-revealing and the optimal mechanism goes to 0.4, which arises when the customers only join the queue if it is empty when they arrive.

Finally, in Figure 1, (c) and (d), we compare the revenue under the optimal mechanism against the revenue of the no-information and the fully-revealing mechanisms, where under each setting we set the fixed-price optimally. In particular, in Figure 1(c), we fix $c = 0.2$ and vary λ, whereas in Figure 1(d), we fix $\lambda = 0.7$ and vary c. We see that the no-information outperforms the fully-revealing information for low values of the arrival rate λ, for a given c. As λ increases, we observe that the revenue of the optimal and the fully-revealing mechanisms converge, whereas the no-information mechanism's revenue is much lower. Similarly, for a fixed arrival rate, we see that optimal signaling is effective in increasing revenue over the no-information and the fully-revealing mechanisms for moderate values of c—that is, when customers incur moderate disutility for waiting.

6.2. Abandonment

Our model assumes that the customers who choose to join the queue stay in the queue until service completion. In many settings, this modeling assumption is unrealistic, and one must explicitly account for customer abandonment. A standard approach (Garnett et al. 2002) to incorporate customer abandonment is by modeling each arriving customer to have an independent and exogenously specified patience time τ and assuming that the customer will abandon the queue if she is still waiting to be served at time τ after joining the queue.

Consider the setting where customers' patience times are distributed independently and identically as an exponential distribution with rate γ. It is straightforward to show that our results continue to hold in this setting. Formally, as before, let $h(n,p)$ denote a customers' payoff upon joining the queue with n customers already in the queue, and the fixed price is p. (Note that, unlike in our original model, this payoff function now incorporates the fact that the customer may leave without obtaining service.) Then, by a similar argument as before, the service provider's decision problem can be written as the following linear program:

$$\begin{aligned}
\max_{\pi} \quad & \sum_{n=1}^{\infty} \pi_n \\
\text{subject to} \quad & \sum_{n=0}^{\infty} (1+\gamma n)\pi_{n+1} h(n,p) \geq 0, \\
& \sum_{n=0}^{\infty} (\lambda \pi_n - (1+\gamma n)\pi_{n+1}) h(n,p) \leq 0, \\
& \lambda \pi_n - (1+\gamma n)\pi_{n+1} \geq 0, \quad \text{for all } n \geq 0, \\
& \sum_{n=0}^{\infty} \pi_n = 1, \qquad \pi_n \geq 0, \quad \text{for all } n \geq 0.
\end{aligned}$$

(Here, we assume that once a customer is in service, she would not abandon the queue; without this assumption, one obtains a slightly modified linear program.) From each feasible solution π to this linear program, one can obtain the corresponding signaling mechanism σ as $\sigma(n,1) = (1+\gamma n)\pi_{n+1}/(\lambda \pi_n)$. For this setting, a similar analysis of the preceding linear program establishes Theorem 1 under same monotonicity conditions on the payoff function h.

Note, however, that the preceding model assumes that the customers only choose to abandon the queue when their patience runs out, and never before. When the queue is observable and the service time distributions are known, this is a fairly mild assumption, because a customer does not learn new information about her waiting time while she waits in the queue. However, in an unobservable queue, this assumption is strong and will in fact not be followed by a fully rational customer. In particular, a rational customer may find it optimal to abandon the queue before her patience runs out. This is because, be the time spent waiting in queue provides further information to a customer regarding the queue length, it is rational for a customer to abandon the queue if she believes that her waiting might be larger than her remaining patience. Modeling the abandonment decision endogenously is challenging, even in models without signaling (Ata and Peng 2017, Ata et al. 2017), and incorporating signaling in such models is an interesting direction for future work.

6.3. Heterogeneous Types

Another assumption of our model is that the customers are homogeneous. Our model can be naturally extended to allow for customer heterogeneity in the form of different utility for joining the queue. Formally, suppose there are K possible customer types $i \in \{1,\ldots,K\}$, where the arrival rate of customers of type i is given by λ_i. Suppose also that the service provider observes the type of a customer upon her arrival and charges a customer of type i a fixed-price p_i for obtaining service. Let $\sum_{i=1}^{K} \lambda_i = \Lambda$. Denote the expected payoff for a customer of type i upon joining the queue with X customers already in the queue as $h_i(X,p_i) \triangleq u_i(X) - p_i$, where each $u_i(\cdot)$ is nonincreasing with $u_i(0) > 0$ and $\lim_{X\to\infty} u_i(X) < 0$. In this setting, a signaling mechanism is specified by $\{\sigma(n,i,j) : n \geq 0, i = 1,\ldots,K; j = 0,1\}$, where $\sigma(n,i,1)$ denotes the probability with which the service provider tells a customer of type i to join the queue when there are n customers already in the queue and $\sigma(n,i,0)$ denotes the probability of telling them not to join. By letting $\phi_n^{i,j} = \lambda_i/\Lambda \pi_n \sigma(n,i,j)$, where $\pi = \{\pi_n : n \geq 0\}$ is the steady-state distribution of the queue, the service provider's decision problem can be reduced to the following linear program:

$$\begin{aligned}
\max_{\phi} \quad & \sum_{n=0}^{\infty} \sum_{i=1}^{K} p_i \cdot \phi_n^{i,1} \\
\text{subject to} \quad & \sum_{n=0}^{\infty} \phi_n^{i,1} h_i(n,p_i) \geq 0, \qquad \text{for } i = 1,2,\ldots,K, \\
& \sum_{n=0}^{\infty} \phi_n^{i,0} h_i(n,p_i) \leq 0, \qquad \text{for } i = 1,2,\ldots,K, \\
& \frac{1}{\lambda_1}(\phi_n^{1,0} + \phi_n^{1,1}) = \frac{1}{\lambda_i}(\phi_n^{i,0} + \phi_n^{i,1}), \\
& \qquad \text{for all } n \geq 0, i = 2,3\ldots,K, \\
& \frac{1}{\lambda_1}(\phi_n^{1,0} + \phi_n^{1,1}) = \sum_{i=1}^{K} \phi_{n-1}^{i,1}, \qquad \text{for all } n \geq 1, \\
& \sum_{n=0}^{\infty} \sum_{i=1}^{K} \phi_n^{i,1} + \phi_n^{i,0} = 1, \\
& \phi_n^{i,j} \geq 0, \qquad \text{for all } n \geq 0, i = 1,2,\ldots,K, \\
& \qquad \text{and } j = 0,1.
\end{aligned} \tag{6}$$

Here, the first two inequalities correspond to the obedience constraints for each type of customer. In particular, the first inequality requires that when a type i customer is told to join the queue, she finds it optimal to join, whereas the second inequality requires that the customer finds it optimal not to join the queue if she is told not to join. The remaining constraints in the linear program arise from the constraints on the steady-state distribution π. From a feasible solution ϕ, one can obtain the signaling mechanism as $\sigma(n,i,1) = \phi_n^{i,1}/(\phi_n^{i,1} + \phi_n^{i,0})$ and $\sigma(n,i,0) = \phi_n^{i,0}/(\phi_n^{i,1} + \phi_n^{i,0})$. Note that if $K = 1$, we are back to the case with homogeneous customers, and the preceding linear program reduces to the linear program (5), where $\phi_n^{1,1} = \pi_{n+1}/\lambda$, and $\phi_n^{1,0} = \pi_n - \pi_{n+1}/\lambda$ for all n.

For the heterogeneous customer type setting, our main result, Theorem 1, extends as follows:

Theorem 4. *Suppose all the customer types are charged the same fixed price; that is, $p_i = p$ for all $i = 1, \ldots, K$. Then, there exists an optimal signaling mechanism that signals each customer type by using a threshold mechanism: $\sigma(n,i,1) = 1$ for $n < N_i$ and $\sigma(n,i,1) = 0$ for $n > N_i$ for some N_i.*

This result is obtained by using a similar argument as to our main result: First, we show that under the optimal mechanism, each customer type is told to join the queue at all queue lengths for which they would have joined under full information; next, we show that any feasible mechanism satisfying this property but not of a threshold type can be perturbed appropriately without reducing the revenue. We omit the details for brevity.

On the other hand, if not all customer types are charged the same price, threshold mechanisms need not be revenue-optimal across all signaling mechanisms. We illustrate this using the following example: Consider a setting with two types of customers ($K = 2$), where each customer type has the following utility function:

$$u_1(n) = \begin{cases} 51, & n = 0, \\ 40, & n = 1, \\ -10{,}000, & n \geq 2, \end{cases} \qquad u_2(n) = \begin{cases} 2, & n \leq 1, \\ -8.5, & n \geq 2. \end{cases}$$

The type 1 customers are charged a price $p_1 = 50$ for service, whereas the type 2 customers are charged $p_2 = 1$. The arrival rates of the two types are $\lambda_1 = \lambda_2 = 1$. Solving the linear program (6), we obtain the optimal signaling mechanism to be

$$\sigma(n,1,1) = \begin{cases} 1, & n = 0, \\ \dfrac{1}{10}, & n = 1, \\ 0, & n \geq 2, \end{cases} \qquad \sigma(n,2,1) = \begin{cases} 0, & n = 0, \\ \dfrac{1}{10}, & n = 1, \\ 0, & n \geq 2. \end{cases}$$

Observe that this is not a threshold mechanism for customers of type 2. One can verify that no threshold mechanism achieves the same revenue as the preceding mechanism.

Nevertheless, the following theorem shows that if the (fixed) price for each type is set optimally, the revenue-optimal signaling mechanism is a threshold mechanism. Furthermore, an analogous result as in Theorem 2 holds: The optimal signaling mechanism (together with optimally set fixed type-dependent prices) achieves the same revenue as the optimal state-and-type-dependent pricing mechanism. We provide the proof in Appendix A.2.

Theorem 5. *For the optimal choice of fixed prices $p_i, i = 1, \ldots, K$, the optimal signaling mechanism has a threshold structure. Also, the revenue achieved by the service provider under this mechanism is same as that in the optimal state-and-type dependent pricing mechanism.*

Finally, a further extension of our model to heterogeneous customers involves the setting of private types, where the service provider cannot observe the types of the arriving customers. These settings in general require a combinatorial number of signals, where each signal corresponds to a subset of customer types who join the queue after receiving it. In the special case where all customer types are charged the same price p, and the types are ordered, meaning $h_i(n,p) \geq h_{i+1}(n,p)$ for all $i = 1, \ldots, K$ and $n \geq 0$, it suffices to consider mechanisms involving $K + 1$ signals, where the signal i corresponds to all customers with types less than or equal to i joining the queue. Although we can again formulate the service provider's decision problem as a linear program, we note once again that threshold mechanisms need not be optimal; there may exist signaling mechanisms that obtain higher revenue than threshold mechanisms.

7. Conclusion

We analyze the optimal information-sharing problem in the context of an unobservable queue with strategic customers. We establish that in the optimal signaling mechanism, the service provider does not fully reveal the queue state, nor completely conceals it. We show that the optimal signaling mechanism uses binary signals and has a threshold structure. We obtain analytical expressions for the optimal threshold in the case of linear customer utility. Furthermore, we show that with the optimal choice of the fixed price, the service provider can effectively use signaling to achieve the expected revenue achieved by the optimal state-dependent pricing mechanism.

Acknowledgments

The authors thank Siddhartha Banerjee, Robert Kleinberg, and the anonymous reviewers for comments and feedback.

A preliminary version of this work appeared as an extended abstract at the 18th ACM Conference in Economics and Computation.

Appendix A. Proofs

In this section, we provide the proofs of the results in the main body of the paper. We start with the proofs of the lemmas in Section 3.

Proof of Lemma 1. Given a signaling mechanism $\Sigma = (\mathcal{S}, \sigma)$ and customer equilibrium f, consider a new signaling mechanism $\Sigma_1 = (\mathcal{S}_1, \sigma_1)$, where $\mathcal{S}_1 = \mathcal{S} \times \{0,1\}$, and $\sigma_1 : \mathbb{N}_0 \times \mathcal{S}_1 \to [0,1]$ is given by

$$\sigma_1(n,s,1) = \sigma(n,s)f(s),$$
$$\sigma_1(n,s,0) = \sigma(n,s)(1-f(s)).$$

Now, consider the strategy f_1 under the signaling mechanism Σ_1, where $f_1(s,1)=1$ and $f_1(s,0)=0$. We begin by showing that the strategy f_1 constitutes a customer equilibrium under Σ_1. First, note that the steady-state distribution of the queue under (Σ_1, f_1) is same as that under (Σ, f). This follows from the fact that the queue has the same transition probabilities at each state under the two settings. Denote this common steady-state distribution by π_∞. From this, we obtain

$$\begin{aligned}\mathbf{E}^{\Sigma_1,f_1}[h(X_\infty,p)|(s,1)] &= \sum_{n=0}^{\infty}\pi_\infty(n|s,1)h(n,p)\\ &= \sum_{n=0}^{\infty}\pi_\infty(n|s)h(n,p)\\ &= \mathbf{E}^{\Sigma,f}[h(X_\infty,p)|s].\end{aligned}$$

Here, $\pi_\infty(n|s,1)$ and $\pi_\infty(n|s)$ denote the conditional probability that there are n customers in the queue upon seeing a signal $(s,1)$ and s respectively in the two signaling mechanisms. The second equality follows from the fact that the choice of the second component in σ_1 is independent of the number of customers in the queue.

Note that under Σ_1, a customer sees the signal $(s,1)$ only if $f(s)>0$, which implies, from the fact that f is a customer equilibrium under σ, that $\mathbf{E}^{\Sigma,f}[h(X_\infty,p)|s] \geq 0$. This implies that $\mathbf{E}^{\Sigma_1,f_1}[h(X_\infty,p)|(s,1)] \geq 0$, and indeed $f_1(s,1)=1$ is an optimal action. Similarly, we obtain that if the customer observes the signal $(s,0)$, then $f_1(s,0)=0$ is indeed an optimal action. Together, we obtain that the strategy f_1 is a customer equilibrium under Σ_1.

The proof then follows from the fact that f_1 is a pure strategy. □

Proof of Lemma 2. From Lemma 1, without loss of generality, assume that the customer equilibrium f is pure. Let $\mathcal{S}^i = \{s \in \mathcal{S} : f(s) = i\}$ for $i = 0,1$. Define $\sigma_1 : \mathbb{N} \times \{0,1\} \to [0,1]$ as follows

$$\sigma_1(n,i) = \sum_{s\in\mathcal{S}^i}\sigma(n,s), \quad \text{for } i = 0,1.$$

Now, consider the strategy f_1 under the signaling mechanism σ_1, where $f_1(i) = i$ for $i = 0,1$. By similar argument as in Lemma 1, it follows that the steady-state distribution under (Σ, f) is same as that under (Σ_1, f_1). Denote this steady-state distribution by π_∞. Thus, it follows that the two settings are equivalent, if we show that f_1 is indeed a customer equilibrium under Σ_1. This follows directly by observing that $\pi_\infty(n|i=1) = \pi_\infty(n|s \in \mathcal{S}^1)$, and hence

$$\begin{aligned}\mathbf{E}^{\Sigma_1,f_1}[h(X_\infty,p)|i=1] &= \mathbf{E}^{\Sigma,f}[h(X_\infty,p)|s\in\mathcal{S}^1]\\ &= \sum_{n=0}^{\infty}\pi_\infty(n|s\in\mathcal{S}^1)h(n,p)\\ &= \sum_{s\in\mathcal{S}^1}\frac{\pi_\infty(s)}{\pi_\infty(\mathcal{S}^1)}\sum_{n=0}^{\infty}\pi_\infty(n|s)h(n,p)\\ &= \frac{1}{\pi_\infty(\mathcal{S}^1)}\sum_{s\in\mathcal{S}^1}\pi_\infty(s)\mathbf{E}^{\Sigma,f}[h(X_\infty,p)|s]\\ &\geq 0,\end{aligned}$$

because $\mathbf{E}^{\Sigma,f}[h(X_\infty,p)|s] \geq 0$ for all $s \in \mathcal{S}^1$. Here, $\pi_\infty(s)$ denotes the probability in steady state of seeing signal s upon arrival, and $\pi_\infty(\mathcal{S}^1)$ denotes the probability of seeing a signal in $\mathcal{S}^1$. The third equation follows from the law of total probability. From this, we obtain that $f_1(1)=1$ is an optimal action on observing a signal 1 under σ_1. Similarly, we obtain that $f(0)=0$ is an optimal action on observing 0 under σ_1. Together, this implies that f_1 is a customer equilibrium under Σ_1 and the result follows. □

Proof of Lemma 3. We begin by showing that for any signaling mechanism $\sigma : \mathbb{N}_0 \times \mathcal{S} \to [0,1]$ feasible for (4), there exists a feasible solution $\pi = \{\pi_n : n \geq 0\}$ to the linear program (5) with the same objective value. Note that the steady-state distribution π_∞^σ of the queue under σ in the obedient equilibrium satisfies the following detailed balance equation,

$$\pi_\infty^\sigma(n)\lambda\sigma(n,1) = \pi_\infty^\sigma(n+1),$$

implying that

$$\pi_\infty^\sigma(n) = \lambda^n\left(\prod_{j=0}^{n-1}\sigma(j,1)\right)\pi_\infty^\sigma(0), \tag{A.1}$$

with $\pi_\infty^\sigma(0)$ given by

$$\pi_\infty^\sigma(0) = \left(\sum_{n=0}^{\infty}\lambda^n\left(\prod_{j=0}^{n-1}\sigma(j,1)\right)\right)^{-1}. \tag{A.2}$$

Define π as $\pi_n = \pi_\infty^\sigma(n)$ for all $n \geq 0$. Clearly $\pi_n \geq 0$ for all $n \geq 0$ and $\sum_{n=0}^{\infty}\pi_n = 1$. Similarly, using the detailed balance equation, we obtain for any $n \geq 0$,

$$\begin{aligned}\lambda\pi_n - \pi_{n+1} &= \lambda\pi_\infty^\sigma(n) - \pi_\infty^\sigma(n+1)\\ &\geq \lambda\pi_\infty^\sigma(n,1)\sigma(n,1) - \pi_\infty^\sigma(n+1) = 0.\end{aligned}$$

Thus, to show feasibility of π, we must verify that (5a) and (5b) hold. To see this, observe that

$$\begin{aligned}\mathbf{E}^{\sigma}[h(X_\infty,p)\mathbf{I}\{s=1\}] &= \sum_{n=0}^{\infty}\pi_\infty^{\sigma}(n,s=1)h(n,p)\\ &= \sum_{n=0}^{\infty}\pi_\infty^{\sigma}(n)\sigma(n,1)h(n,p)\\ &= \frac{1}{\lambda}\sum_{n=0}^{\infty}\pi_\infty^{\sigma}(n+1)h(n,p) \qquad (A.3)\\ &= \frac{1}{\lambda}\sum_{n=0}^{\infty}\pi_{n+1}h(n,p)\\ &= \frac{1}{\lambda}\sum_{n=1}^{\infty}\pi_n h(n-1,p).\end{aligned}$$

Here, the third equality follows from the detailed balance condition. Because σ is feasible for (4), we have $\mathbf{E}[h(X_\infty,p)|s=1]\geq 0$. From this, we conclude that $\sum_{n=1}^{\infty}\pi_n h(n-1,p)\geq 0$, and hence (5a) holds. Similarly, we have

$$\begin{aligned}\mathbf{E}^{\sigma}[h(X_\infty,p)\mathbf{I}\{s=0\}] &= \sum_{n=0}^{\infty}\pi_\infty^{\sigma}(n,s=0)h(n,p)\\ &= \sum_{n=0}^{\infty}\pi_\infty^{\sigma}(n)(1-\sigma(n,1))h(n,p)\\ &= \sum_{n=0}^{\infty}\left(\pi_\infty^{\sigma}(n)-\frac{1}{\lambda}\pi_\infty^{\sigma}(n+1)\right)h(n,p)\\ &= \frac{1}{\lambda}\sum_{n=0}^{\infty}(\lambda\pi_n-\pi_{n+1})h(n,p).\end{aligned} \qquad (A.4)$$

Again, we have used the detailed balance condition in the third equality. Because σ is feasible for (4), we have $\mathbf{E}^{\sigma}[h(X_\infty,p)|s=0]\leq 0$. From this and the preceding equalities, we conclude that (5b) holds. Finally, observe that

$$\mathbf{E}^{\sigma}[\lambda\sigma(X_\infty,1)] = \sum_{n=0}^{\infty}\lambda\pi_\infty^{\sigma}(n)\sigma(n,1) = \sum_{n=0}^{\infty}\pi_\infty^{\sigma}(n+1) = \sum_{n=1}^{\infty}\pi_n. \qquad (A.5)$$

Thus, we obtain that π is feasible for (5), with the same objective value as σ in (4).

Next, consider any feasible solution $\pi=\{\pi_n : n\geq 0\}$ for (5). We show that there exists a signaling mechanism σ feasible for (4) that attains the same objective value as π. Define $\sigma:\mathbb{N}_0\times\{0,1\}\to[0,1]$ as

$$\sigma(n,1)=\begin{cases}\dfrac{\pi_{n+1}}{\lambda\pi_n}, & \text{if } \pi_n>0,\\ 0, & \text{otherwise.}\end{cases}$$

In order to verify that the obedience constraints hold for σ, we first compute the steady-state distribution π_∞^{σ} when all customers follow the obedient strategy. Using (A.1) and (A.2), we get $\pi_\infty^{\sigma}(n)=\pi_n$ for all $n\geq 0$. Thus, from (A.3) and (A.4) and from the fact that π is feasible for (5), we obtain that $\mathbf{E}^{\sigma}[h(X_\infty,p)\mathbf{I}\{s=1\}]\geq 0$ and $\mathbf{E}^{\sigma}[h(X_\infty,p)\mathbf{I}\{s=0\}]\leq 0$. After conditioning on the appropriate event, we obtain that σ satisfies the obedience constraints. Finally, using (A.5), we conclude that σ achieves the same objective value in (4) as π in the linear program (5). □

The following lemma is used in the proof of Theorem 1 to show that the maximum in the linear program (5) is attained.

Lemma A.1. *Let $\mathcal{D}$ denote the set of all feasible solutions $\{\pi_n : n\geq 0\}$ to (5) of the following form: There exists an $N\geq M_p$, such that $\pi_n=\lambda\pi_{n-1}$ for all $n<N$, $0<\pi_N\leq\lambda\pi_{N-1}$ and $\pi_n=0$ for $n>N$. (Note N can equal ∞.) Then, the set $\mathcal{D}$ is compact under the weak topology.*

Proof. We will show that the set of distributions $\mathcal{D}$ is tight. The result then follows from Prokhorov's theorem. To show tightness, we prove that for any $\epsilon>0$, there exists an N such that for any $\pi\in\mathcal{D}$, we have $\sum_{n>N}\pi_n<\epsilon$.

Fix an $\epsilon>0$. First, note that if $\lambda<1$, then for any feasible solution π, we have $\pi_n\leq\lambda^n\pi_0\leq\lambda^n$. Hence, we obtain that for all large enough N, $\sum_{n>N}\pi_n\leq\sum_{n>N}\lambda^n<\lambda^N/(1-\lambda)<\epsilon$. Thus, for the rest of the proof, suppose $\lambda\geq 1$.

Let N be the first positive integer such that $\sum_{n=1}^{N-1}\lambda^n h(n-1,p)<0$ [because $h(n,p)\leq h(M_p,p)<0$ for all $n\geq M_p$, there must be such a value of N]. Note that for $\lambda\geq 1$ and for $\pi\in\mathcal{D}$, there must be an $m<\infty$ such that $\pi_m<\lambda\pi_{m-1}$. For a feasible $\pi\in\mathcal{D}$, let L be such that $\pi_n=\lambda^n\pi_0$ for all $n<L$, $0\leq\pi_L<\lambda\pi_{L-1}$ and $\pi_n=0$ for all $n>L$. If $L>N$, we have

$$\begin{aligned}\sum_{n=1}^{\infty}\pi_n h(n-1,p) &= \pi_0\sum_{n=1}^{L-1}\lambda^n h(n-1,p)+h(L-1,p)\pi_L\\ &\leq \pi_0\sum_{n=1}^{N-1}\lambda^n h(n-1,p)<0,\end{aligned}$$

contradicting the fact that $\pi\in\mathcal{D}$. Thus, for all $\pi\in\mathcal{D}$, we have $L\leq N$ and hence $\sum_{n>N}\pi_n=0<\epsilon$. Thus, by Prokhorov's theorem, the set $\mathcal{D}$ is compact. □

The following simple lemma states that the throughput is increasing in the threshold, and is used in the Proof of Theorem 3.

Lemma A.2. *The throughput of the threshold signaling mechanism σ^x is monotonically increasing in $x\in\mathbb{R}_+$.*

Proof. Note that for $x=N+q$ with $N\in\mathbb{N}_0$ and $q\in[0,1]$, we have $\pi_\infty^x(0)=(\sum_{i=0}^{N}\lambda^i+\lambda^{N+1}q)^{-1}$, where $\{\pi_\infty^x(n) : n\geq 0\}$ denotes the steady-state distribution of the queue under the signaling mechanism σ^x. This follows from the fact that under σ^x, there are at most $N+1$ customers in the queue, with a new customer joining the queue with probability 1 if the number of customers already in the queue is strictly less than N, and joining with probability q if the number of customers is equal to N, and balking otherwise. Thus, $\pi_\infty^x(0)$ is strictly decreasing in $x=N+q$. The result then follows from the fact that throughput satisfies $\mathsf{Th}(\sigma^x)=\lambda(1-\pi_\infty^x(0))$. □

A.1. Proof of Theorem 3

We now present the Proof of Theorem 3, obtaining analytical expression for the optimal threshold in the case of linear utility.

Proof of Theorem 3. Consider a threshold mechanism σ^x with $x\geq M_p$. We seek to find the largest value of x for which the obedient strategy is a customer equilibrium. Then, by Lemma A.2 and Theorem 1, we obtain the threshold mechanism σ^x is optimal.

To begin, note that because $x \geq M_p$, if a customer observes a signal $s = 0$, then the number of customers in the queue is at least $\lfloor x \rfloor = M_p$, and hence the expected payoff on joining the queue is at most $h(M_p, p) \leq 0$. Hence, leaving on seeing signal $s = 0$ is indeed optimal.

We have for $x = N + q \geq M_p$,

$$\pi_\infty^x(n|s=1) = \frac{\lambda^n \mathbf{I}\{n < N\} + q\lambda^N \mathbf{I}\{n = N\}}{\sum_{k<N} \lambda^k + q\lambda^N}.$$

This implies,

$$\mathbf{E}^{\sigma^x}[h(X_\infty, p)|s=1] = \frac{\sum_{k<N} \lambda^k(1-p-c(k+1)) + q\lambda^N(1-p-c(N+1))}{\sum_{k<N} \lambda^k + q\lambda^N}.$$

Thus, for joining the queue to be optimal for a customer on seeing a signal $s = 1$, we must have

$$\sum_{k<N} \lambda^k(1-p-c(k+1)) + q\lambda^N(1-p-c(N+1)) \geq 0. \quad \text{(A.6)}$$

We consider the following two cases separately:

Case 1. $\lambda = 1$**.** In this case, the Equation (A.6) becomes

$$(1-p-c)N - \frac{c}{2}N(N-1) + q(1-p-c(N+1)) \geq 0. \quad \text{(A.7)}$$

We first consider the case where $q = 0$ to find the largest N that satisfies this equation. The largest such value of N is

$$N^* = \left\lfloor \frac{2(1-p)}{c} - 1 \right\rfloor.$$

Unless the expression inside the floor operator on the right-hand side is an integer, we have $(1-p-c)N - cN(N-1)/2 > 0$, implying we can set $q > 0$ and not violate (A.7). The largest value of $q = q^*$ that can be set is for which (A.7) is an equality. (Note that q^* cannot be equal to 1, by definition of N^*.) From this, we obtain the following expression for q^*:

$$q^* = \frac{(1-p-c)N^* - \frac{c}{2}N^*(N^*-1)}{c(N^*+1)+p-1}.$$

Case 2. $\lambda \neq 1$**.** We, again, first consider the case where $q = 0$ and seek the largest value of N that satisfies (A.6). For any value N that satisfies (A.6), upon adding up the summations, we obtain

$$(1-p)\frac{1-\lambda^N}{1-\lambda} - c\left(\frac{1-\lambda^N}{1-\lambda} + \frac{\lambda - N\lambda^N + (N-1)\lambda^{N+1}}{(1-\lambda)^2}\right) \geq 0,$$

which on simplifying yields,

$$\frac{(1-p-c)(1-\lambda)}{\lambda c}(1-\lambda^N) \geq 1 - N\lambda^{N-1} + (N-1)\lambda^N.$$

Let $\alpha = (1-p-c)(1-\lambda)/(\lambda c)$ and $\beta = (1-\lambda)/\lambda$. Then, we obtain after some algebra,

$$(N\beta + 1 - \alpha)\lambda^N \geq 1 - \alpha. \quad \text{(A.8)}$$

Note that if $\alpha \geq 1$, implying that $\lambda \leq 1 - c/(1-p)$, then the right-hand side is nonnegative for all $N \geq 1$. Thus, all values of $N \geq M_p$ satisfy this equation, and hence we obtain $N^* = \infty$. In other words, the optimal signaling mechanism always signals the customer to join the queue.

Suppose now that $\alpha < 1$. Then, multiplying both sides of (A.8) by $\left(\lambda^{\frac{1}{\beta}}\right)^{1-\alpha} > 0$ gives us

$$(N\beta + 1 - \alpha)\left(\lambda^{\frac{1}{\beta}}\right)^{N\beta+1-\alpha} \geq (1-\alpha)\left(\lambda^{\frac{1}{\beta}}\right)^{1-\alpha}.$$

Let $\psi = N\beta + 1 - \alpha$ and $\gamma = \lambda^{1/\beta}$. Note that for all $\lambda \neq 1$, we have $\gamma < 1$. The preceding equation can be written as $(1-\alpha)\gamma^{1-\alpha} \leq \psi\gamma^\psi$. After multiplying both sides by $\log(1/\gamma) > 0$ and some algebra, we obtain

$$\psi \log\left(\frac{1}{\gamma}\right) \exp\left(-\psi \log\left(\frac{1}{\gamma}\right)\right) \geq (1-\alpha)\log\left(\frac{1}{\gamma}\right) \exp\left(-(1-\alpha)\log\left(\frac{1}{\gamma}\right)\right). \quad \text{(A.9)}$$

For $x > 0$, let $H(x)$ be the function defined implicitly by $H(x)\exp(-H(x)) = x\exp(-x)$ with $H(x) \neq x$ for $x \neq 1$. Observe that if $x > 1$, then $H(x) < 1$ and if $x < 1$, then $H(x) > 1$, with $H(1) = 1$.

Note that if $\lambda > 1$, then $\beta < 0$, $\alpha < 0$, which implies $1 - \alpha \geq 1$. Furthermore, we obtain that for $\lambda > 1$, $\gamma \leq e^{-1}$, which implies $\log(1/\gamma) \geq 1$. Hence, $z = (1-\alpha)\log(1/\gamma) \geq 1$. On the other hand, if $1 - c/(1-p) < \lambda < 1$, then $\beta > 0$, $1 - \alpha \in [0, 1]$, and furthermore $\log(1/\gamma) \leq 1$. Hence, $z = (1-\alpha)\log(1/\gamma) \leq 1$. Using these facts, and the definition of $H(\cdot)$, we obtain from (A.9),

$$H\left((1-\alpha)\log\left(\frac{1}{\gamma}\right)\right) \leq \psi\log\left(\frac{1}{\gamma}\right) \leq (1-\alpha)\log\left(\frac{1}{\gamma}\right), \quad \text{if } \lambda > 1;$$
$$(1-\alpha)\log\left(\frac{1}{\gamma}\right) \leq \psi\log\left(\frac{1}{\gamma}\right) \leq H\left((1-\alpha)\log\left(\frac{1}{\gamma}\right)\right), \quad \text{if } \lambda < 1.$$

Using the fact that $\psi = N\beta + 1 - \alpha$, and noting that $\beta < 0$ if $\lambda > 1$ and $\beta > 0$ if $\lambda < 1$, we get

$$N \leq \frac{1}{\beta\log\left(\frac{1}{\gamma}\right)}\left(H\left((1-\alpha)\log\left(\frac{1}{\gamma}\right)\right) - (1-\alpha)\log\left(\frac{1}{\gamma}\right)\right).$$

Because N^* is the largest such value of N, we have

$$N^* = \left\lfloor \frac{1}{\beta\log\left(\frac{1}{\gamma}\right)}\left(H\left((1-\alpha)\log\left(\frac{1}{\gamma}\right)\right) - (1-\alpha)\log\left(\frac{1}{\gamma}\right)\right)\right\rfloor.$$

Using the definition of the Lambert-W function and its two branches W_0 and W_{-1} [see Borgs et al. (2014)], it can be shown that for $x > 0$, we have $H(x) = -W_i(xe^{-x})$, where $i = 0$ if $x > 1$ and $i = -1$ if $x < 1$. Upon letting $\kappa = (1-\alpha)\log\left(\frac{1}{\gamma}\right) = \left(\frac{1-p}{c} - \frac{1}{1-\lambda}\right)\log(\lambda)$, we obtain

$$N^* = \left\lfloor \frac{1}{\log(\lambda)}(W_i(-\kappa e^{-\kappa}) + \kappa)\right\rfloor,$$

where $i = 0$ if $\lambda > 1$ and $i = -1$ if $1 - \frac{c}{1-p} < \lambda < 1$.

Finally, observe that unless the expression inside the floor operator in the expression for N^* is an integer, we have $\sum_{k<N^*} \lambda^k(1-c(k+1)) > 0$, and we can set $q > 0$ without violating (A.6). The largest value of $q = q^*$ that can be set is for which (A.6) is an equality. (Note that q^* cannot be equal to 1, by definition of N^*.) From this, we obtain

$$q^* = \frac{\sum_{k<N^*} \lambda^k (1-p-c(k+1))}{\lambda^{N^*}(c(N^*+1)+p-1)}.$$

This completes the proof. □

A.2. Proof of Theorem 5

In this subsection, we prove Theorem 5. The proof follows a similar structure to that of Theorem 2: We use the structure of the optimal state-and-type-dependent pricing mechanism in a fully observable queue to construct a threshold signaling mechanism (with fixed prices) that attains the same revenue. As the first step, we analyze the optimal state-and-type-dependent prices in a fully observable queue, where the service provider sets a price $p_i(n)$ for customer type i and queue length n. We have the following lemma.

Lemma A.3. *The optimal state-and-type-dependent pricing mechanism in a fully observable queue satisfies*

$$p_i(n) = \begin{cases} u_i(n), & \text{if } n < \kappa_i, \\ \infty, & \text{if } n \geq \kappa_i, \end{cases} \tag{A.10}$$

for some $\kappa = (\kappa_1, \ldots, \kappa_K) \in \mathbb{N}^{\mathbb{K}}$. Furthermore, this mechanism is welfare-optimal.

Proof of Lemma A.3. We begin by formulating the pricing problem in a fully observable queue as an infinite-horizon Markov decision process (MDP) with average rewards. We consider the embedded discrete time chain with states as follows. For each $n \geq 0$ and $i \in \{1, \ldots, K\}$, let (n, i) denote the state where there are n customers already in the queue and a customer of type i has arrived. Similarly, for $n \geq 0$, let $(n, 0)$ denote the state after a customer has departed, leaving n customers in the queue. Note that the service provider must choose a price $p_i(n) \geq 0$ at state (n, i) for $n \geq 0$ and $i \in \{1, \ldots, K\}$. (At state $(n, 0)$, the service provider chooses a dummy action.)

First, note that because $\lim_{X\to\infty} u_i(X) < 0$ for all i, there exists an N such that the $u_i(n) < 0$ for all i and $n > N$. Because $p_i(n) > u_i(n)$ for all $n > N$ and each i, no customer will join the queue when there are at least N customers already in the queue. Thus, it follows that the MDP is unichain (Puterman 1994), and the optimal prices can be found by solving the Bellman equation. Let V denote the average revenue under the optimal pricing mechanism, and let $g(n, i)$ denote the *bias* (Puterman 1994) of each state (n, i).

The Bellman equation for the pricing problem can then be written as follows: For each $n \geq 0$ and $i \in \{1, \ldots, K\}$, we have

$$\begin{aligned} &V + g(n,i) \\ &= \max_{p_i(n)\geq 0}\Bigg[\mathbf{I}\{p_i(n) \leq u_i(n)\}\left(p_i(n) + \sum_{j=1}^{K}\frac{\lambda_j}{1+\Lambda}g(n+1,j) + \frac{1}{1+\Lambda}g(n,0)\right) \\ &\qquad + \mathbf{I}\{p_i(n) > u_i(n)\}\left(\sum_{j=1}^{K}\frac{\lambda_j}{1_n+\Lambda}g(n,j) + \frac{1_n}{1_n+\Lambda}g(n-1,0)\right)\Bigg], \end{aligned} \tag{A.11}$$

and

$$V + g(n,0) = \sum_{j=1}^{K}\frac{\lambda_j}{1_n+\Lambda}g(n,j) + \frac{1_n}{1_n+\Lambda}g(n-1,0). \tag{A.12}$$

Here, we define $1_n \triangleq \mathbf{I}\{n > 0\}$ and recall that $\Lambda = \sum_{j=1}^{K}\lambda_j$. The first equation follows from the fact that if $p_i(n) \leq u_i(n)$, a customer of type i will join the queue at state (n, i), yielding an immediate revenue of $p_i(n)$. Subsequently, the queue state transitions to $(n+1, j)$ with probability $\lambda_j/(1+\Lambda)$ for $j \in \{1, \ldots, K\}$, and to state $(n, 0)$ with probability $1/(1+\Lambda)$. On the other hand, if $p_i(n) > u_i(n)$, then the customer does not join the queue at state (n, i), yielding no immediate revenue and similar subsequent transitions. The second equation follows from the fact that the service provider has a single dummy action at state $(n, 0)$ that yields no immediate revenue.

From the Bellman equation, it follows that one can always restrict to $p_i(n) \in \{u_i(n), \infty\}$: The price $p_i(n) = \infty$ performs equally as well as any $p_i(n) > u_i(n)$, and any $p_i(n) < u_i(n)$ is strictly dominated by $p_i(n) = u_i(n)$. Using this, we can write (A.11) as

$$\begin{aligned} V + g(n,i) = \max\Bigg[&u_i(n) + \sum_{j=1}^{K}\frac{\lambda_j}{1+\Lambda}g(n+1,j) + \frac{1}{1+\Lambda}g(n,0), \\ &\sum_{j=1}^{K}\frac{\lambda_j}{1_n+\Lambda}g(n,j) + \frac{1_n}{1_n+\Lambda}g(n-1,0)\Bigg], \end{aligned} \tag{A.13}$$

where the optimal price is $u_i(n)$ if the first term attains the maximum, and ∞ otherwise. Thus, it remains to show that the optimal pricing mechanism has a threshold structure—that is, $p_i(n) < \infty$ implies $p_i(m) < \infty$ for all $m < n$.

Substituting (A.12) into (A.13) and after simplifying, we obtain

$$\begin{aligned} g(n,i) &= \max\{u_i(n) + g(n+1,0), g(n,0)\} \\ &= g(n,0) + (u_i(n) - \eta(n))^+, \end{aligned}$$

where $\eta(n) \triangleq g(n,0) - g(n+1,0)$ for all n. Substituting this expression back into (A.12) and simplifying, we obtain the following equation that holds for all $n \geq 0$:

$$V = \sum_{i=1}^{K}\frac{\lambda_i}{1_n+\Lambda}(u_i(n) - \eta(n))^+ + \frac{1_n}{1_n+\Lambda}\eta(n-1). \tag{A.14}$$

Note that $p_i(n) < \infty$ if and only if $u_i(n) \geq \eta(n)$. Thus, to show that the optimal prices have a threshold structure, we must show that if $u_i(n) \geq \eta(n)$ for some n, then $u_i(m) \geq \eta(m)$ for all $m < n$. Because $u_i(n)$ is nonincreasing in n for each i, it suffices to show that $\eta(n)$ is nondecreasing in n. We prove this latter statement by induction.

First, note that because $\lim_{X\to\infty} u_i(X) < 0$ for all $i \in \{1, \ldots, K\}$, there exists an $N > 0$ such that $u_i(n) < 0$ for each i and $n > N$. By our earlier argument, this implies that optimal prices satisfy $p_i(n) = \infty$ for all i and $n > N$, which in turn implies $u_i(n) < \eta(n)$ for all i and $n > N$. Then, using (A.14), we obtain $\eta(n) = V(1+\Lambda)$ for all $n \geq N$. Hence, $\eta(n) \geq \eta(n-1)$ for all $n > N$.

Now suppose $\eta(n) \geq \eta(n-1)$ for some $n \geq 2$. Note that because $u_i(n) \leq u_i(n-1)$, this implies that $(u_i(n) - \eta(n))^+ \leq (u_i(n-1) - \eta(n-1))^+$. From (A.14) and using the fact that $n \geq 2$, we obtain

$$\begin{aligned} V &= \sum_{i=1}^{K} \frac{\lambda_i}{1+\Lambda}(u_i(n) - \eta(n))^+ + \frac{1}{1+\Lambda}\eta(n-1) \\ &\leq \sum_{i=1}^{K} \frac{\lambda_i}{1+\Lambda}(u_i(n-1) - \eta(n-1))^+ + \frac{1}{1+\Lambda}\eta(n-1) \\ &= V - \frac{1}{1+\Lambda}\eta(n-2) + \frac{1}{1+\Lambda}\eta(n-1). \end{aligned}$$

Thus, we have $\eta(n-1) \geq \eta(n-2)$. This completes the induction step, and we conclude that $\eta(n)$ is nondecreasing in n. Thus, if $u_i(n) \geq \eta(n)$ for some n, then $u_i(m) \geq \eta(m)$ for all $m < n$, and hence the optimal prices have a threshold structure.

Finally, it is straightforward to show that the problem of welfare optimization (where the service provider performs admission control to maximize social welfare) can be written as a dynamic program with the same Bellman equation, given by (A.12) and (A.13). This implies that the optimal state-and-type-dependent pricing mechanisms is also welfare optimal. □

We conclude with the proof of Theorem 5.

Proof of Theorem 5. From Lemma A.3, let $\kappa = (\kappa_1, \ldots, \kappa_K)$ denote the thresholds in the optimal state-and-type-dependent pricing mechanism. Let X_i^κ denote the steady-state distribution of the queue under this pricing mechanism.

For the unobservable queue, consider the signaling mechanism σ, where $\sigma(n, i, 1) = \mathbf{I}\{n < \kappa_i\}$ for each $i \in \{1, \ldots, K\}$, and the fixed (type-dependent) prices $p_i = \mathbf{E}[u(X_\infty^\kappa) | X_\infty^\kappa < \kappa_i]$. Using the same argument as in the proof of Theorem 2, it is straightforward to show that, for this setting, obedience is an equilibrium, and that under the obedient equilibrium, the service provider's revenue is the same as that of the optimal state-and-type-dependent pricing mechanism.

Finally, because the latter mechanism is welfare-optimal and has zero customer surplus, we conclude that the mechanism σ and the prices p_i together constitute the optimal fixed price and signaling mechanism. □

Appendix B. Comparison of the Fully-Revealing and the No-Information Mechanisms

In this section, we briefly compare the fully-revealing and no-information mechanisms in the case of linear expected utility $u(X) = 1 - c(X+1)$ and a fixed price p. Observe that for $p > 1 - c$, no customer will join the queue to obtain service, and hence the service provider's revenue is zero. We restrict our attention to $p \in [0, 1-c]$.

For the fully-revealing mechanism, we find the throughput to be

$$\mathsf{Th}^{\mathsf{full}} = \lambda\left(\frac{1-\lambda^{M_p}}{1-\lambda^{M_p+1}}\right),$$

where $M_p = \lceil (1-c-p)/c \rceil$. The optimal revenue for the full-information signal is given by

$$R^{\mathsf{full}} = \max_p \left(\frac{\lambda - \lambda^{M_p+1}}{1-\lambda^{M_p+1}}\right)p.$$

In the case of the no-information mechanism, a customer strategy is a probability q with which a customer joins the queue. We can view the queue as a thinned M/M/1 queue with arrival rate $q\lambda$. Recall that the stationary distribution for such a queue is $q\lambda/(1-q\lambda)$.

Note that $q = 0$ is not an equilibrium for $p < 1-c$: If so, joining the queue would have utility $1 - c - p > 0$. We see that $q = 1$ is an equilibrium if and only if the utility for joining the queue $(1 - p - c(\lambda/(1-\lambda)+1))$ is at least that of not (0), or, equivalently, if $\lambda \leq 1 - c/(1-p)$. Otherwise, if $\lambda > 1 - c/(1-p)$, we must have a mixed-strategy equilibrium $q \in (0,1)$. For this to be an equilibrium, the utility for joining the queue $[1 - p - c(q\lambda/(1-q\lambda+1)]$ must equal the utility for not joining the queue (0), so that a mixed strategy is optimal. This is equivalent to $q = (1 - c/(1-p))/\lambda$. Putting these cases together, we get that for any $p \in [0, 1-c]$, the customer equilibrium f is given by $q = \min\{(1 - c/(1-p))/\lambda, 1\}$, with the corresponding throughput given by $\mathsf{Th}^{\mathsf{no-info}} = \min\{1 - c/(1-p), \lambda\}$. Maximizing the revenue $p \cdot \mathsf{Th}^{\mathsf{no-info}}$ over values of $p \in [0, 1-c]$, we obtain the optimal price p^* to be

$$p^* = \begin{cases} 1 - \dfrac{c}{1-\lambda}, & \text{if } \lambda \leq 1 - \sqrt{c}, \\ 1 - \sqrt{c}, & \text{otherwise,} \end{cases}$$

with corresponding revenue given by

$$R^{\mathsf{no-info}} = \begin{cases} \lambda - \dfrac{c\lambda}{1-\lambda}, & \text{if } \lambda \leq 1 - \sqrt{c}, \\ (1-\sqrt{c})^2, & \text{otherwise.} \end{cases}$$

From the preceding discussion, we observe that for values where $\lambda < 1 - c/(1-p)$, we have $\mathsf{Th}^{\mathsf{full}} < \mathsf{Th}^{\mathsf{no-info}}$, implying that sharing no information about the queue with customers achieves higher throughput than revealing the number of customers in the queue. On the other hand, observe that when $p = 1 - c - \epsilon$ for small enough $\epsilon > 0$, we have $M_p = 1$, and hence,

$$\lim_{\lambda\to\infty} R^{\mathsf{full}} \geq \lim_{\lambda\to\infty} (1 - c - \epsilon)\left(\frac{\lambda - \lambda^2}{1-\lambda^2}\right) = 1 - c - \epsilon.$$

However, because $\sqrt{c} > c + \epsilon$ for small enough c, we have

$$\lim_{\lambda\to\infty} R^{\mathsf{no-info}} = (1-\sqrt{c})^2 = 1 - 2\sqrt{c} + c < 1 - c - \epsilon.$$

Thus, in this limiting regime, we have that revealing the number of customers in the queue obtains a higher revenue than not revealing, as seen for large values of λ in Figure 1(c).

Endnotes

[1] Here, and in the sequel, we let $\mathbb{N}_0$ denote the set of nonnegative integers.

[2] We note that in any equilibrium (as defined below), the queue will be stable.

[3] Note that in this formulation, we have ignored the possibility of the existence of multiple equilibria under a signaling mechanism. The right formulation would require that the service provider chooses, in addition to the signaling mechanism, a focal equilibrium f among all possible equilibria. Our results continue to hold under this formulation; we suppress the technical details for brevity and readability.

References

Allon G, Bassamboo A, Gurvich I (2011) We will be right with you: Managing customer expectations with vague promises and cheap talk. *Oper. Res.* 59(6):1382–1394.

Armony M, Maglaras C (2004a) Contact centers with a call-back option and real-time delay information. *Oper. Res.* 52(4):527–545.

Armony M, Maglaras C (2004b) On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* 52(2):271–292.

Ata B, Peng X (2017) An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. *Oper. Res.* 66(1):163–183.

Ata B, Glynn PW, Peng X (2017) An equilibrium analysis of a discrete-time Markovian queue with endogenous abandonments. *Queueing Systems* 86(1):141–212.

Bergemann D, Morris S (2016a) Bayes correlated equilibrium and the comparison of information structures in games. *Theoret. Econom.* 11(2):487–522.

Bergemann D, Morris S (2016b) Information design, Bayesian Persuasion, and Bayes correlated equilibrium. *Amer. Econom. Rev.* 106(5):586–591.

Bergemann D, Morris S (2019) Information design: a unified perspective. *J. Econom. Literature* 57(1):44–95.

Borgs C, Chayes JT, Doroudi S, Harchol-Balter M, Xu K (2014) The optimal admission threshold in observable queues with state dependent pricing. *Probab. Engrg. Inform. Sci.* 28(1):101–119.

Burnetas A, Economou A (2007) Equilibrium customer strategies in a single server markovian queue with setup times. *Queueing Systems* 56(3):213–228.

Chen H, Frank MZ (2001) State dependent pricing with a queue. *IIE Trans.* 33(10):847–860.

Crawford VP, Sobel J (1982) Strategic information transmission. *Econometrica* 50(6):1431–1451.

Cui S, Veeraraghavan S (2016) Blind queues: The impact of consumer beliefs on revenues and congestion. *Management Sci.* 62(12): 3656–3672.

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

Economou A, Kanta S (2008b) Optimal balking strategies and pricing for the single server markovian queue with compartmented waiting space. *Queueing Systems* 59(3/4):237–269.

Edelson NM, Hilderbrand DK (1975) Congestion tolls for Poisson queuing processes. *Econometrica* 43(1):81–92.

Ely JC (2017) Beeps. *Amer. Econom. Rev.* 107(1):31–53.

Fudenberg D, Tirole J (1991) *Game Theory* (MIT Press, Cambridge, MA).

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.

Guo P, Zipkin P (2007) Analysis and comparison of queues with different levels of delay information. *Management Sci.* 53(6):962–970.

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* 54(5):1185–1195.

Hassin R (2016) *Rational Queueing* (CRC Press, Boca Raton, FL).

Hassin R, Haviv M (2012) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, International Series in Operations Research & Management Science (Springer, New York).

Hassin R, Koshman A (2017) Profit maximization in the M/M/1 queue. *Oper. Res. Lett.* 45(5):436–441.

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.

Kalwani MU, Yim CK, Rinne HJ, Sugita Y (1990) A price expectations model of customer brand choice. *J. Marketing Res.* 27(3):251–262.

Kamenica E, Gentzkow M (2011) Bayesian persuasion. *Amer. Econom. Rev.* 101(6):2590–2615.

Kolotilin A, Mylovanov T, Zapechelnyuk A, Li M (2017) Persuasion of a privately informed receiver. *Econometrica* 85(6):1949–1964.

Kremer I, Mansour Y, Perry M (2014) Implementing the "wisdom of the crowd." *J. Political Econom.* 122(5):988–1012.

Mansour Y, Slivkins A, Syrgkanis V, Wu ZS (2016) Bayesian exploration: Incentivizing exploration in Bayesian games. *Proc. 2016 ACM Conf. Econom. Comput. (EC'16)* (ACM, New York), 661.

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

Papanastasiou Y, Bimpikis K, Savva N (2017) Crowdsourcing exploration. *Management Sci.* 64(4):1727–1746.

Pender J, Rand RH, Wesson E (2017) Queues with choice via delay differential equations. *Internat. J. Bifurcation Chaos* 27(4):1730016.

Pender J, Rand RH, Wesson E (2018) An analysis of queues with delayed information and time-varying arrival rates. *Nonlinear Dynam.* 91(4):2411–2427.

Puterman ML (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st ed. (John Wiley & Sons, New York).

Rayo L, Segal I (2010) Optimal information disclosure. *J. Political Econom.* 118(5):949–987.

Sheshinski E, Weiss Y (1977) Inflation and costs of price adjustment. *Rev. Econom. Stud.* 44(2):287–303.

Simhon E, Hayel Y, Starobinski D, Zhu Q (2016) Optimal information disclosure policies in strategic queueing games. *Oper. Res. Lett.* 44(1):109–113.

Taneva I (2019) Information design. *Amer. Econom. J. Microeconom.* Forthcoming.

Wolff RW (1982) Poisson arrivals see time averages. *Oper. Res.* 30(2): 223–231.

Yu Q, Allon G, Bassamboo A (2017a) How do delay announcements shape customer behavior? An empirical study. *Management Sci.* 63(1):1–20.

Yu Q, Allon G, Bassamboo A, Iravani S (2017b) Managing customer expectations and priorities in service systems. *Management Sci.* 64(8):3942–3970.

David Lingenbrink is a PhD candidate in the School of Operations Research and Information Engineering at Cornell University. His work focuses on optimal information sharing and persuasion.

Krishnamurthy Iyer is an assistant professor in the School of Operations Research and Information Engineering at Cornell University. His research interests include game theory, mechanism design, and applied probability, with applications in markets and service systems. His work was awarded first place in the 2017 INFORMS Junior Faculty Forum Paper Competition.