Priority Pricing in Queues with a Continuous Distribution of Customer Valuations

Sherwin Doroudi* **Mustafa Akan***
Mor Harchol-Balter
Jeremy Karp* **Christian Borgs†**
Jennifer T. Chayes†

May 2013
CMU-CS-13-109

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

*Tepper School of Business, Carnegie Mellon University
† Microsoft Research

This research was made possible by a Computational Thinking Grant from Microsoft Research. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution.

Keywords: Queues, priority, pricing, mechanism design, incentives

Abstract

We consider a service provider facing a *continuum* of delay-sensitive strategic customers. The service provider maximizes revenue by charging customers for the privilege of joining an M/G/1 queue and assigning them service priorities. Each customer has a valuation for the service, with a waiting cost per unit time that is proportional to their valuation; customer types are drawn from a continuous distribution and are unobservable to the service provider. We illustrate how to find revenue-maximizing incentive-compatible priority pricing menus, where the firm charges higher prices for higher queueing priority. We show that our proposed priority pricing scheme is optimal across all incentive-compatible pricing policies whenever the customer valuation distribution is regular. We compute the resulting price menus and priority allocations in *closed form* when customer valuations are drawn from Exponential, Uniform, or Pareto distributions. We find revenues in closed form for the special case of the M/M/1 queue, and compute revenues in the more general setting numerically. We compare our priority pricing scheme to the best fixed pricing scheme, as well as an idealized pricing scheme where customers always reveal their valuation. We observe the impact of service requirement variability on revenue and prices. We also illustrate how to create the optimal *discrete* priority pricing menu when the service provider is restricted to offering a *finite* number of priority classes.

1 Introduction

We consider a service provider facing an arrival stream of delay-sensitive, strategic customers. The service provider charges customers for the right to join an M/G/1 queue. Customer types are drawn from a continuous distribution where the "type" of a customer consists of her valuation (utility obtained from receiving service) and her delay sensitivity (waiting cost). While the arrival rate of customers and the distribution of customer types is known, we assume that a customer's type is her private information, and hence, *not* known to the service provider.

Our goal throughout is to maximize the revenue of the service provider. We do this by creating a **menu of prices**, where a customer pays a higher price to receive higher **queueing priority**, and hence lower delay. Based on a customer's delay-sensitivity, it may be better for the customer to choose the higher price option and incur less delay, or the lower price option and incur more delay. We assume that all customers behave strategically and selfishly, that is, they maximize their own utility (their value for service less delay costs and payments). Customers also have the option to forego joining the queue; if they opt not to join they pay nothing. The queue is unobservable, so customers do not know their delay ex-ante. It is thus important that customers make choices from the menu in a predictable fashion so that the service provider can report accurate "expected delays" as a function of customer priority (via queueing theory). Therefore, the service provider must ensure that the menu of prices is **incentive compatible**.

Motivation

The above scenario is representative of many existing applications. For example, today's airlines offer customers different prices, where higher prices allow the customer to board earlier or get her bags first (e.g., the Southwest Airlines *Business Select* ticket class). Many amusement parks (e.g., Disney World) offer customers a menu of prices, where customers paying a higher price for a higher priority class can jump ahead of lower priority classes in line. The goal of the service provider (the amusement park) is to maximize revenue. Note that while the queue length is unobservable, paying more yields lower expected delay due to having higher priority. The service provider would like to be able to provide customers with an expectation on what their delay will be at each priority level. To do this, the service provider must trust that customers behave in a predictable way when making menu selections. While delay is less of a problem in the internet today, there is apprehension that delay will become more of an issue in the future. There is already talk of providing differentiated services for shared databases (like orbitz.com) [34, 35, 44], as well as allowing Internet Service Providers (ISPs) to charge customers a higher price in order to give their packets higher priority at routers (a.k.a. DiffServ).

Continuous priority pricing

Our specific approach is to offer a priority class for each type of customer, resulting in a continuum of priorities. This approach allows for expressing the price menu as a continuous function over the space of customer types. In order to make analysis tractable, we make the natural simplifying assumption that a customer's delay sensitivity (waiting cost per unit time), c, is directly proportional to her valuation, v; that is, $c = \alpha v$ for some constant α, and so the greater a customer's value for service, the greater her delay sensitivity. By parametrizing customer types by their valuation, v, each type has its own priority class offered at price $p(v)$, where the continuous function $p(\cdot)$ is the price menu. With this simplifying assumption, our approach has the benefit of yielding the first *closed-form* incentive-compatible price menus in the setting with a continuum of customers. Among all possible incentive-compatible price menus, we select the one that maximizes revenue. In Section 7, we provide the explicit closed forms for the revenue-maximizing incentive-compatible price menu and resulting priority allocations in the cases where valuations are drawn from the Exponential, Uniform, or Pareto distributions. We also present the revenue generated by these pricing menus in closed form for the special case of M/M/1 queues; we compute revenues in the M/G/1 setting numerically. Throughout, we will refer to the above policy as **Priority Pricing** (PP).

Comparison with other pricing policies and service requirement variability

We compare PP to several other policies. The **Fixed Pricing** (FP) policy involves setting a single (optimally chosen) fixed price; all customers opting to join the queue pay this price and are served in first-come-first-serve order. The **Full Information** (FI) policy is the optimal Priority Pricing policy in the optimistic setting where customer types are visible to the firm (i.e., customer types are *not* private information). By definition, FI provides an upper bound on the revenue achievable under PP. Over a range of distributions and customer parameters, we find that revenue obtained from PP typically outperforms FP by 2–20%. We also prove that, when customer types are unobservable, PP is revenue-optimal across all incentive-compatible policies.

Finally, we consider the setting where the service provider is limited to offering a finite number of priority classes (e.g., gold, silver, and bronze priority). We call the revenue-maximizing incentive-compatible **Priority Pricing policy with n priority classes**, PP(n), and we find that even for small n, PP(n) performs nearly as well as PP. In particular, in a variety of settings, offering 5 priority classes is sufficient to capture over 99% of the revenue that can be obtained by offering a continuum of priority classes.

Organization of the paper

The remainder of this paper is organized as follows. In Section 2 reviews the relevant literature in the area of queueing with incentives. In Section 3, we formally model the customers' utilities

and the service provider's revenue maximization problem. We then proceed to characterize the incentive-compatible price functions in Section 4. Next, in Section 5 we approach our problem from the framework of Myerson's revenue-optimal auctions [37]. This framework allows for the classification of the set of distributions for which the Priority Pricing policy is revenue-optimal among across all pricing policies where customer priorities are static once they join the queue. We introduce the aforementioned alternate pricing schemes in Section 6. We then apply our findings from Sections 4 and 5 to specific distributions on customer valuations, namely, the Exponential, Uniform, and Pareto distributions in Section 7. A benefit of studying an M/G/1 model rather than the simpler M/M/1 model is that it allows us to witness the effect of service requirement variability on revenue and prices. In Section 8, we observe that as service requirement variability rises, revenue falls and the highest valuation customers pay more, while the lowest valuation customers either pay less or cease to join the queue. We also compare the performance of PP with FP and FI in Section 8, while Section 9 examines the impact of offering only finitely many priority classes (PP(n)).

2 Literature Review

2.1 Social welfare maximization

Most prior work in the area of queueing with incentives falls under maximizing social welfare, rather than maximizing the service provider's profits. Within social welfare maximization, much of the literature is concerned with **homogenous customers** (i.e., customers with identical preferences), as in [39, 47, 9, 16, 12]. These papers are primarily concerned with admission control, prices, and equilibrium behavior, as models with only a single class of customers do not necessitate priority queueing.

Other work in the area of social welfare maximization deals with **heterogenous customers**, who differ in either their valuation for the service or their delay sensitivities (i.e., waiting costs). In this setting, it is beneficial to make use of priority queues to give the best service to those customers who need it most. Kleinrock was the first to study priority pricing, although his work ignores customer incentives [30]. Later papers such as [33, 21, 22, 36, 19, 23] present a variety of models to address the incentive-compatible pricing of priority. In some of the related literature, as in the work of Kittsteiner and Moldovanu, priority is auctioned rather than sold directly using posted prices [28, 29].

2.2 Profit maximization

Contrastingly, this paper is focused on maximizing the service provider's profits. Once again, some of the research in this area examines models with **homogenous customers**. These papers are often concerned with **observable queues**, which allow customers to view the length of the

queue before making their entry decision. The observable queueing model in [39],[16], and [12] is also applicable to profit maximization. Although Chen and Frank give some results for multiple classes of customers, and [31] gives a formulation generalizing Naor's model to multiple customer classes, these findings are also restricted to observable queues. Çil, et.al. consider a rich framework where customers incur either a high or low holding cost for the service provider, with customer valuations distributed continuously within each group. Their results concern the structure of the optimal policy, but the optimal pricing policy is not obtained in closed-form. [17]. While these papers use bounded queue lengths to control customer arrivals, [8] approaches dynamic control from a different perspective where the service provider can adjust both the arrival and service rates. Although the aforementioned papers do not make use of priority queueing, observable priority queues can still be of use in the identical customer setting, as in [2] and [6].

Other work on profit maximization places greater focus on **heterogenous** customers in settings with **unobservable queues**. Within this space, customer heterogeneity is often captured via a **finite number** of different customer classes, as in [42] and [3]. Petersen and Rao assume customer waiting times rather than deriving them from queueing theory. Afèche considers a model with two classes of customers differing in their waiting costs and service requirements. In this work profits are maximized through a technique known as "damaging the goods," which amounts to strategic idling on the part of the server.

The small remainder of the literature, like our work, features models allowing for **infinitely many** customer classes. Plambeck considers customers who are either patient or impatient, where customer valuations are continuously distributed within each group [41]. However, unlike our work, the resulting priority price menus found in this work are not incentive-compatible, that is, they are not robust to strategic customers. Moreover, many of the results in this paper are valid only under heavy traffic assumptions. Abhishek, et al. consider a model where customers have one of two valuations for service, where customer waiting costs are continuously distributed within each group [1]. However, unlike our work, the focus of this work is not on deriving closed-form price menus, but rather on establishing some qualitative and theoretical economic results. The lack of focus on closed-form solutions allows the work to be more general, but at the expense of structural results, as the complexity of waiting time functions is effectively ignored. The work of Mandjes features customer heterogeneity in service requirements rather than in preferences [32], placing this work even further from the model investigated in our paper.

The work closest to ours appears in two working papers due to Katta and Sethuraman [26] and Afèche and Pavlin [4]. In both papers, customer valuations and waiting costs are distributed continuously and are related linearly. In addition to examining continuous pricing, Katta and Sethuraman study discrete price menus via simulation, while we use exact numerical analysis. The Aféche and Pavlin working paper was developed simultaneously with and independently from our work. We find the same implicit formula for the optimal incentive-compatible continuous price function given by Aféche and Pavlin. While these papers consider the M/M/1 setting, we consider the M/G/1 setting, allowing us to study the effect of service requirement variability. Most importantly, the results in these papers are *implicit*, whereas we consider particular distributions of customer valuations, yielding the first *explicit* closed-form price functions and allocations.

2.3 Analytic tools used in queueing with incentives and other directions

Much of the above work, including this paper, leverages priority queues in order to serve customers with greater valuations and greater delay sensitivity ahead of others. Both preemptive and non-preemptive priority queues are analyzed in the queueing theory literature in [18, 40, 43]. Moreover, much of the mechanism design research involving profit maximization invokes techniques from the celebrated paper by Myerson [37]. Our connection to Myerson is described at the end of Section 5.

Comprehensive surveys of the literature at the intersection of queueing and game theory are available in the books [24] and [45]. There is much other literature involving pricing and queueing omitted from the discussed above. For example, there is a long stream of literature on leadtime-dependent pricing, where the amount paid by customers depends on how soon they receive a good (their leadtime), see for example [27, 14, 15, 25, 13]. Much of this literature is far removed from our work, in that it does not involve incentives and strategic customers. The few papers that deal with leadtimes with incentives [5, 7] are still considerably different in their assumptions from our model those in the papers reviewed in Sections 2.1 and 2.2.

3 The Model

3.1 The queueing model and the customers

Consider a service provider serving customers in an unobservable M/G/1 queue. Customers arrive with average rate λ and have i.i.d. service requirements with mean $\mathbb{E}[S]$ and second moment $\mathbb{E}[S^2]$. We assume that load $\rho \equiv \lambda\mathbb{E}[S] < 1$ in order to ensure stability.[1] Each customer has a type (v, c), where v is her valuation for the service and c is her waiting cost per unit time. Customer valuations, v, are i.i.d. random variables drawn from a continuous distribution with support on an interval $\mathbf{X} \subseteq [0, \infty)$, cdf F, ccdf $\bar{F} \equiv 1 - F$, and pdf $f \equiv F'$. We assume that f is differentiable, and non-vanishing on the interior of $\mathbf{X}$. Moreover, $c = \alpha v$ for some fixed constant $\alpha \in (0, 1/\mathbb{E}[S])$. Hence, *waiting costs are proportional to valuations*, so a customer's *type* is parametrized by v alone. Henceforth, we will use the terms valuation and type interchangeably. A customer's type is her private information, while a customer's service requirement, S, is independent of the customer's type and is unknown to both the service provider and the customer.

Upon arriving, a customer chooses whether to join the queue. If the customer decides to join, she must report her type v to the firm, although she need not do so truthfully; the customer may *strategically* report that she has type x for any $x \in \mathbf{X}$. The customer *cannot* observe the current state of the queue when making the decision to join or when reporting her type. Upon entering, a

[1] In most cases, we can extend our results to allow for the case where $\rho > 1$ by ignoring all customers with valuations below some lower bound (so long as these customers would not join the queue in equilibrium), and thereby changing the support of the customer valuation distribution, so that the *effective* load falls under 1.

customer will pay a price $p(x)$, based on her reported type x, where p is a price function chosen and made public by the service provider. Customers receive *preemptive queueing priority* over all customers who report lower valuations (i.e., all customers who pay less).[2] The price function p may be viewed as a price menu corresponding to a continuum of priority classes each offered at their own price. Assuming truth-telling behavior on the part of all *other* customers, a customer of type v who reports type x will obtain a utility of

$$u_v(x) \equiv v - \alpha v \mathbb{E}[T(x)] - p(x) = v(1 - \alpha \mathbb{E}[T(x)]) - p(x) \tag{1}$$

for joining the queue, where $\mathbb{E}[T(x)]$ is the expected waiting time given reported type x, assuming all other customers report their valuations truthfully. Customers who decide not to join the queue obtain zero utility. We will prove that under appropriate choices of p, there exists a Bayesian Nash equilibrium where all customers will indeed report their true valuations.

3.2 The service provider's price function

The service provider seeks to set the price function p in order maximize its revenue, or more precisely, *the rate at which the firm earns revenue*, R, given by

$$R(p) = \lambda \int_{\mathbf{E}} p(v) f(v)\, dv, \tag{2}$$

where $\mathbf{E} \equiv \{v\colon u_v(v) > 0\}$ is the set of types willing to enter the system, given that all customers report their true valuations. Note that the utility functions u_v depend on p, and so the set $\mathbf{E}$ also depends on p. Since customers report their types in a self-interested manner, the assumption that customers are truth-telling is actually a constraint on the service provider's choice of the price function p. In particular, we require that every customer type is truth-telling whenever all other types are truth-telling,[3] that is

$$\forall v, x \in \mathbf{X}\colon u_v(x) \leq \max\{u_v(v), 0\}\}. \tag{3}$$

The constraint in (3) is called the *incentive-compatibility* constraint on p. This means that all customers are best off either reporting their true type or not joining the queue at all.

In most settings, it is irrelevant whether customers who receive zero utility from joining the queue decide to join the queue. For convenience, we assume that indifferent customers *do not* join the queue. That is, $\mathbf{E} = \{v\colon u_v(v) > 0\}$.

[2]We further assume for all the policies studied in this paper that a customer's priority may not change after they join the queue. That is, a customer may not be preempted by a lower priority customer on the basis of how long each customer has been in the system, how much service each customer has received, etc.

[3]To ensure consistent and stable system behavior, it would be sufficient to find any pricing mechanism where each customer of type v_0 would prefer to report some transformation of its type $\tau(v_0)$, given all customers of any type v were reporting $\tau(v)$. Consistent with the mechanism design literature [20], we restrict attention to incentive-compatible direct revelation mechanisms. This is without loss of generality, as the revelation principle tells us that for any general selling mechanism, any equilibrium of rational strategies for the service provider and the customers can be replicated by an equivalent incentive-compatible direct-revelation mechanism [38].

4 Characterizing incentive-compatible price functions

In order to find the revenue-maximizing incentive-compatible price function, we will characterize the class of all incentive-compatible price functions in terms of the given parameters α, λ, $\mathbb{E}[S]$, and $\mathbb{E}[S^2]$ as well as the type distribution captured by the cdf F.

4.1 The form of incentive-compatible price functions

Before expressing the *explicit* structure of this class of functions, we will prove some general results that will hold for incentive-compatible price functions in this setting. Our first observation will be that incentive-compatible price functions are non-decreasing.

Theorem 1. *If the price function p is incentive-compatible, then it is non-decreasing on* $\mathbf{E}$.

Proof. Proof. Assume by way of contradiction that there exists some incentive-compatible p that is not non-decreasing on $\mathbf{E}$. Then there exist $v_1, v_2 \in \mathbf{E}$ such that $v_1 < v_2$, but $p(v_1) > p(v_2)$. Since $v_1 \in \mathbf{E}$, we know that $u_{v_1}(v_1) > 0$, and so it follows from the incentive-compatibility of p, that $u_{v_1}(v_1) \geq u_{v_1}(v_2)$. However, since $\mathbb{E}[T(\cdot)]$ is nonincreasing on $\mathbf{X}$, as higher reported valuations never lead to slower service (i.e., longer wait times) in expectation, we see that the assumption that $p(v_1) > p(v_2)$ yields

$$u_{v_1}(v_1) = v_1(1 - \alpha\mathbb{E}[T(v_1)]) - p(v_1) < v_1(1 - \alpha\mathbb{E}[T(v_2)]) - p(v_2) = u_{v_1}(v_2),$$

contradicting the assumption that p is incentive-compatible. □

Next, we prove that nonnegative incentive-compatible price functions induce customers to join the queue according to a threshold strategy, that is, if a customer enters, all customers with higher valuation also enter the queue.

Theorem 2. *Let p be a nonnegative incentive-compatible price function. Then if type $v_0 \in \mathbf{X}$ joins the queue, then all types $v \geq v_0$ such that $v \in \mathbf{X}$ will also join the queue. That is, customers will join the queue according to a threshold strategy.*

Proof. Proof. We are given that $v_0 \in \mathbf{E}$, and hence $u_{v_0}(v_0) > 0$. It follows that $1 - \alpha\mathbb{E}[T(v_0)] > 0$ because $p(v_0) \geq 0$. Consequently, for all $v \geq v_0$

$$0 < u_{v_0}(v_0) = v_0(1 - \alpha \cdot \mathbb{E}[T(v_0)]) - p(v_0) \leq v(1 - \alpha \cdot \mathbb{E}[T(v_0)]) - p(v_0) \leq u_{v_0}(v)$$

and by incentive-compatibility $u_v(v) \geq u_v(v_0) > 0$, so $v \in \mathbf{E}$. □

A consequence of this theorem is that customers join the queue with *threshold-type* entry behavior: there exists some threshold $v_* \equiv \inf \mathbf{E}$, such that $\mathbf{E} = (v_*, \infty) \cap \mathbf{X}$. Hence, v_* is the greatest lower bound on customer types willing to join the queue. Consistent with our earlier assumption,

customers of type v_* will not join the queue, as prices will be set to ensure that they are indifferent. Note that threshold entry behavior also holds wherever the price function is nonnegative on $\mathbf{E}$ even if p is negative somewhere on $\mathbf{X}\backslash\mathbf{E}$.

The next theorem allows us to restrict the feasible set of incentive-compatible price functions in our optimization problem to those incentive-compatible price functions that are *nonnegative*. We prove that it is always in the interest of the firm to charge nonnegative prices, which guarantees that if an optimal (i.e., revenue-maximizing) incentive-compatible price function exists, then at least one such price function is nonnegative. This enables us to only restrict attention to threshold entry behavior.

Theorem 3. *For any incentive-compatible price function p, there exists a nonnegative incentive-compatible price function $\hat{p}$ such that $R(\hat{p}) \geq R(p)$. In particular, there exists a nonnegative revenue-maximizing incentive-compatible price function.*

Proof. Proof. Proof deferred to Appendix. □

From this point forward it will be convenient to define nonnegative price functions on just $\mathbf{E}$ (or for convenience on $\{v_*\} \cup \mathbf{E}$), rather than $\mathbf{X}$. That is, the service provider can elect not to offer any priority levels below that intended for customers of type v_*. Since customers with valuations less than v_* would be unwilling to join the queue with *any* report of their type, they would still be unwilling to join the queue if they were given fewer choices of potential types to report.

4.2 The expected response time in the preemptive priority queue

Characterizing the structure of incentive-compatible price functions explicitly will require computing the expected response time of a customer reporting type x in a preemptive priority queue. In particular, we want to compute $\mathbb{E}[T(x)]$ for all $x \in \mathbf{E}$, when entry behavior is threshold type.

The expected response time of a customer reporting type x is equal to the busy period[4] started by the preexisting work, W_x (at the time of entry) in the queue made up of only those customers with valuations (and hence reported valuations) of at least x, in addition to the work of the entering customer, S. That is,

$$\mathbb{E}[T(x)] = \frac{\mathbb{E}[W_{\geq x} + S]}{1 - \rho\bar{F}(x)}, \tag{4}$$

where $\rho = \lambda\mathbb{E}[S]$. To compute $\mathbb{E}[W_{\geq x} + S]$, observe that this quantity is the same as the expected response time of a customer in a FCFS queue made up of only those customers with valuations of at least x, which is given by the Pollaczek-Khinchin (P-K) formula

$$\mathbb{E}[W_{\geq x} + S] = \mathbb{E}[S] + \frac{\lambda\bar{F}(x)\mathbb{E}[S^2]}{2(1 - \rho\bar{F}(x))}. \tag{5}$$

[4]A busy period started by an amount of work, W, is the length of time until a system with initial work W becomes empty (this includes completing W work in addition to any work arriving during this process).

Substituting (5) into (4), we find that a customer reporting type x (given that all other customers report their types truthfully) experiences an expected response time of

$$\mathbb{E}[T(x)] = \frac{2\mathbb{E}[S] + \lambda\bar{F}(x)(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)}{2(1-\rho\bar{F}(x))^2}. \tag{6}$$

For any $x \notin \mathbf{X}\backslash\mathbf{E}$ (i.e., $x \leq v$), we must have $\mathbb{E}[T(x)] = \mathbb{E}[T(v_*)]$ as the customer reporting such a type is preempted by all entering customers, who have a collective arrival rate of $\lambda\bar{F}(v_*)$. However, in equilibrium no customer joining the queue will report a valuation $x \leq v_*$.

4.3 Explicit characterization of incentive-compatible price functions

Now that we have a closed form for $\mathbb{E}[T(x)]$, we proceed to find the incentive-compatible price menus in closed form. As stated in Section 4.1, we need define price functions only on $\mathbf{E}$ (or $\{v_*\} \cup \mathbf{E}$).

Theorem 4. *Let* $\mathbb{E}[T(\cdot)]$ *be as in* (6)*. For all* $v_* \in \mathbf{X}$*, such that* $\alpha\mathbb{E}[T(v_*)] \leq 1$*,*

$$p(x) = x(1 - \alpha\mathbb{E}[T(x)]) - \int_{v_*}^{x} (1 - \alpha\mathbb{E}[T(t)])\, dt$$

is a positive, incentive-compatible price menu on the entry set $\mathbf{E} = (v_*, \infty) \cap \mathbf{X}$*.*

Proof. Proof. We will first show that p is positive. Observe that for all $x \in \mathbf{E}$, $\mathbb{E}[T(x)]$ is decreasing in x and $x > v_*$, so $1 - \alpha\mathbb{E}[T(x)] > 1 - \alpha\mathbb{E}[T(v_*)] \geq 0$. It follows that

$$p(x) \geq x(1 - \alpha\mathbb{E}[T(x)]) - (x - v_*)(1 - \alpha\mathbb{E}[T(x)]) > v_*(1 - \alpha\mathbb{E}[T(x)]) \geq 0$$

establishing that p is positive (and hence, nonnegative), and confirming that entry behavior is threshold-type. It follows that (6) accurately gives the response time in the system for a customer reporting type $x \in \mathbf{E}$.

To verify incentive-compatibility, we check the first order condition:

$$\begin{aligned} u_v'(x) &= \frac{d}{dx}[v(1 - \alpha\mathbb{E}[T(x)]) - p(x)] \\ &= \frac{d}{dx}[(v - x)(1 - \alpha\mathbb{E}[T(x)])] - (1 - \alpha\mathbb{E}[T(x)]) \\ &= -\alpha(v - x)\left(\frac{d}{dx}[\mathbb{E}[T(x)]]\right) \\ &= 0, \end{aligned}$$

and since $\mathbb{E}[T(x)]$ is monotonic on $\mathbf{X}$, we may conclude that $x = v$ is the *unique* solution to $u_v'(x) = 0$ for $x \in \mathbf{X}$. Checking the second order condition, we have

$$\begin{aligned} u_v''(x)|_{x=v} &= -\alpha \left(\frac{d^2}{dx^2} \left[(v-x) \left(\frac{d}{dx} [\mathbb{E}[T(x)]] \right) \right] \right) \Bigg|_{x=v} \\ &= -\alpha(v-x) \left(\frac{d^2}{dx^2} [\mathbb{E}[T(x)]] \right) \Bigg|_{x=v} + \alpha \left. \frac{d}{dx} \mathbb{E}[T(x)]] \right|_{x=v} \\ &= \alpha \left(\frac{d}{dx} [\mathbb{E}[T(x)]] \right) \Bigg|_{x=v} \\ &< 0, \end{aligned}$$

as $\alpha > 0$ and $\mathbb{E}[T(x)]$ is decreasing in x on $\mathbf{X}$, establishing that p is incentive-compatible.

It remains to check that under equilibrium, we have the entry set $\mathbf{E} = (v_*, \infty) \cap \mathbf{X}$. Since we have established incentive-compatibility, $\mathbf{E} = \{v \in \mathbf{X} \colon u_v(v) \geq 0\}$. Now observe that the utility of a customer of type v,

$$u_v(v) = \int_{v_*}^{v} (1 - \alpha \mathbb{E}[T(t)])\, dt,$$

is positive if and only if $u_v(v)$ is $v > v_*$. □

The PP policy implements the price menu p with the revenue-maximizing choice of v_*. In particular, the optimal revenue must be given by[5]

$$R = \lambda \max_{v_*} \left\{ \int_{v_*}^{\infty} \left(v(1 - \alpha \mathbb{E}[T(v)]) - \int_{v_*}^{v} (1 - \alpha \mathbb{E}[T(t)])\, dt \right) f(v)\, dv \right\}. \tag{7}$$

We defer the discussion of solving this optimization problem until the next section.

5 The Myerson Auction Framework and the Optimality of Priority Pricing

In this section we provide an alternative framework for viewing the revenue-maximization problem posed in this paper. This approach is based on a generalization of Myerson's auction design framework. One advantage of adopting this framework is that it provides an easy way of classifying those distributions on customer valuations for which Priority Pricing is revenue-maximizing across all policies (see Theorem 5).

Myerson's celebrated paper [37] frames the problem in terms of a mechanism with a *finite* number of customers where the firm is auctioning *a good*, such that a customer's outcome is the

[5]Note that although the upper limit of the outer integral is taken to infinity, if $\mathbf{X}$ is bounded, f is vanishing beyond $\sup \mathbf{X}$, and so the upper limit of the integral could equivalently be written as $\sup \mathbf{X}$.

ex ante probability that the customer receives the good (or the fraction of a fungible good to be received). The sum of the probability over all customers must therefore be 1 (or between 0 and 1 if the good can be disposed). By contrast, in our setting the outcome a customer receives is dependent on the customer's delay, where delay is subject to a feasibility constraint that is considerably more complicated than a total probability constraint. Nonetheless, we can still use a generalization of this approach for our model.

We reformulate our model in a slightly different way. Recall that a customer of type v who joins the queue with reported type x obtains utility $u_v(x) = v(1 - \alpha\mathbb{E}[T(x)]) - p(x)$. Therefore, absent prices, the customer obtains $v(1-\alpha\mathbb{E}[T(x)])$. We call $y(x) \equiv v(1-\alpha\mathbb{E}[T(x)])$ the outcome experienced by a customer reporting type x (in expectation). Moreover, observe that a customer of type v values her outcome $y(x)$ linearly; that is, a customer with valuation v is willing to pay up to $vy(x)$ for the service when reporting x. Further assume that any customer opting not to join the queue receives a degenerate outcome equal to zero.

The firm is essentially selling outcomes to customers; an outcome $y(x)$ is priced at $p(x)$. The possible outcome allocations are subject to feasibility constraints reflecting the attainable response times in priority-based queueing systems. For instance, allocating all customers the minimal expected delay is not a feasible allocation of outcomes. The prices, $p(x)$, are subject to incentive-compatibility. Moreover, if $y(x) = 0$, then $p(x) = 0$, as this outcome is equivalent in value to not joining the queue.

5.1 Characterization of incentive-compatible price functions via the Meyerson framework

We now characterize the incentive compatible price functions. As before, we let $\mathbf{E}$ be the set of customers opting to join the queue. We have that for all $v \in \mathbf{E}$, $v = \arg\max_x vy(x) - p(x)$, which together with the first order condition, establishes that

$$\int_{v_*}^{v} p'(t)\,dt = \int_{v_*}^{v} vy'(v)dt$$

where v_* is the greatest lower bound on customer types willing to join the queue, so we set $p(v_*) = v_*y(v_*)$ in order to extract all value from customers of type v_* (should they join, although they ultimately will not); note that we again have $\mathbf{E} = (v_*, \infty) \cap \mathbf{X}$ (except for the case where $\mathbf{E} = \mathbf{X}$).

Applying the Fundamental Theorem of Calculus yields

$$p(v) = p(v_*) + \int_{v_*}^{v} ty'(t)\,dt = vy(v) - \int_{v_*}^{v} y(t)\,dt, \tag{8}$$

where the second equality follows from integration by parts.

5.2 The optimality of Priority Pricing

The revenue rate, R, can be found as follows

$$\begin{aligned} R &= \lambda \int_{v_*}^{\infty} p(v) f(v)\, dv \\ &= \lambda \int_{v_*}^{\infty} \left(v y(v) - \int_{v_*}^{v} y(t)\, dt \right) f(v)\, dv \\ &= \lambda \int_{v_*}^{\infty} v y(v) f(v)\, dv - \int_{v_*}^{\infty} \int_{v_*}^{v} y(t)\, dt f(v)\, dv, \\ &= \lambda \int_{v_*}^{\infty} v y(v) f(v)\, dv - \int_{v_*}^{\infty} \bar{F}(v) y(v)\, dv \\ &= \lambda \int_{v_*}^{\infty} y(v) \left(v - \frac{\bar{F}(v)}{f(v)} \right) f(v)\, dv \end{aligned}$$

where the last line is obtained by integration by parts.

Defining $\phi(v) \equiv v - \bar{F}(v)/f(v)$, we obtain

$$R = \lambda \int_{v_*}^{\infty} \phi(v) y(v) f(v)\, dv. \tag{9}$$

The function $\phi(v)$ is Myerson's *virtual valuation function* [37], so called because the firm maximizes its revenue when using the outcome allocation that maximizes the social welfare (subject only to feasibility constraints) when customer valuations are replaced by their corresponding virtual valuations. When $\phi(v)$ is increasing, we say that the customer valuation distribution is *regular*. This condition is important as it is sufficient to make the PP policy optimal across all pricing policies:

Theorem 5. *When the customer valuation distribution is regular (i.e., $\phi(v) = v - \bar{F}(v)/f(v)$ is increasing), the PP policy is the optimal pricing policy.*[6]

Proof. Proof. Consider two customers with types $v_1 > v_2$. It also holds that $\phi(v_1) > \phi(v_2)$ and the customers obtain outcomes $y(v_1) > y(v_2)$. If we were to exchange the queueing priorities of these customers, their response times (and hence, outcomes) would also be exchanged, so social welfare would decline by

$$\phi(v_1)y(v_1) + \phi(v_2)y(v_2) - [\phi(v_1)y(v_2) + \phi(v_2)y(v_2)] = [\phi(v_1) - \phi(v_2)][y(v_1) - y(v_2)],$$

so social welfare is maximized by serving higher valuation customers ahead of those with lower valuations whenever possible. Hence, the preemptive priority policy maximizes social welfare. □

We first note that the Exponential, Uniform, and Pareto distributions are all regular, so PP is optimal for these customer valuation distributions. We give closed-form expressions for the optimal price menus for these distributions in Section 7.

[6] We restrict attention to priority policies where a customer's priority level cannot change after it joins the queue; this assumption is not necessary for M/M/1 queues.

5.3 The optimal choice of v_*

Completing the interpretation of the PP policy in the Myerson framework, we have customers with valuation $v > v_*$ experience the outcome

$$y(v) = 1 - \frac{2\alpha\mathbb{E}[S] + \alpha(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)\lambda\bar{F}(v)}{2(1-\rho\bar{F}(v))^2}, \tag{10}$$

while all other customers obtain allocation $y = 0$. In order to choose the optimal price menu, it remains only to compute the optimal value of v_*. Applying the first order condition to (9), we have

$$R' = -\lambda f(v)\phi(v)y(v) = 0. \tag{11}$$

We find that (11) has at most two zeros in $\mathbf{X}$, as $-\lambda f(v)$ is nonvanishing on $\mathbf{X}$, while both $y(v)$ and $\phi(v)$ are increasing on $\mathbf{X}$. These zeros are z_ϕ and z_y, the zeros of ϕ and y, respectively (when they exist in $\mathbf{X}$).

Checking the second order condition, we find that

$$R''(v_*)|_{v_*=z_\phi} = -\lambda f(z_\phi)\phi'(z_\phi)y(z_\phi) \leq 0 \text{ if and only if } z_\phi \geq z_y,$$

as $-\lambda f(v) < 0$ and $\phi'(z) > 0$, while $y(z_\phi) \geq y(z_y) = 0$ if $z_\phi \geq z_y$ and $y(z_\phi) < y(z_y) = 0$ otherwise. Similarly,

$$R''(v_*)|_{v_*=z_y} = -\lambda f(z_y)\phi(z_y)y'(z_y) \leq 0 \text{ if and only if } z_y \geq z_\phi,$$

as $-\lambda f(v) < 0$ and $y'(z_y) > 0$, while $\phi(z_y) \geq \phi(z_\phi) = 0$ if $z_y \geq z_\phi$ and $\phi(z_y) < \phi(z_\phi) = 0$ otherwise.

Hence, the revenue-maximizing choice of v_* is $\max\{z_\phi, y_\phi\}$, should they both exist. If only one exists, then the optimal choice of v_* is given by whichever exists, and if neither exists, then the optimal choice of v_* is given by $\inf \mathbf{X}$ (i.e., in this case prices are set so that all customers opt to join the queue). The constraint $v_* \geq z_y$ ensures that $y(v_*) \geq 0$, and hence, all customers who join the queue receive a positive utility without needing to be subsidized (i.e., paying a negative cost).

It is worth noting that

$$z_y = y^{-1}(0) = \bar{F}^{-1}\left(\frac{1}{\rho} - \frac{\alpha}{2\lambda} + \frac{\alpha\mathbb{E}[S^2]}{4\lambda\mathbb{E}[S]^2} - \frac{\sqrt{8\alpha\mathbb{E}[S]\mathbb{E}[S^2] + \alpha^2(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)^2}}{4\lambda\mathbb{E}[S]^2}\right) \tag{12}$$

for any customer valuation cdf F, and consequently, $z_\phi \geq z_y$ if and only if

$$\alpha \leq \frac{2(1-\rho\bar{F}(z_\phi))^2}{2\mathbb{E}[S] + \lambda\bar{F}(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)}. \tag{13}$$

Furthermore, for the Exponential, Uniform, and Pareto distributions, we have z_ϕ is equal to the mean of the customer valuation distribution.

6 Alternative pricing policies

In this section, we introduce two alternatives to Priority Pricing (PP). Our motivation for introducing these policies is to provide a set of benchmarks by which we can subsequently evaluate the relative efficacy of the PP policy introduced in this paper.

6.1 The Fixed Pricing (FP) policy

FP assumes that the service provider sets a fixed price $p \geq 0$ and offers service in first-come-first-serve order (i.e., there are no priorities). Once again, we assume that the queue is unobservable. This structure again yields threshold entry behavior, with customers entering if and only if their type $v > v_*$. Hence, a customer's utility for service must be

$$v(1 - \alpha \cdot \mathbb{E}[T]) - p = v\left(1 - \alpha\mathbb{E}[S] - \frac{\alpha\lambda\bar{F}(v_*)\mathbb{E}[S^2]}{2(1-\rho\bar{F}(v_*))}\right) - p.$$

In particular since customers enter if and only if their utility is nonnegative, p and v_* are related as follows:

$$p = v_*\left(1 - \alpha\mathbb{E}[S] - \frac{\alpha\lambda\bar{F}(v_*)\mathbb{E}[S^2]}{2(1-\rho\bar{F}(v_*))}\right). \tag{14}$$

In this setting the revenue is given by $R_{\mathrm{FP}} = \lambda p\bar{F}(v_*)$ and hence, the optimal revenue is

$$R_{\mathrm{FP}} = \max_{v_*}\left\{\lambda v_*\bar{F}(v_*)\left(1 - \alpha\mathbb{E}[S] - \frac{\alpha\lambda\bar{F}(v_*)\mathbb{E}[S^2]}{2(1-\rho\bar{F}(v_*))}\right)\right\}, \tag{15}$$

with the optimal price (in terms of the optimal v_*) given in (14). Surprisingly this is a maximization problem that eludes closed form solutions for even simple distributions on customer valuations.

6.2 The Full Information (FI) policy

FI assumes that that the firm can observe customer types directly and extract full surplus from all customer types opting to join the queue. Like PP, this policy also prioritizes customers with higher valuations above those with lower valuations, as it can extract greater surplus from the former. In this setting the service provider charges customers with valuation v a price *arbitrarily close to, but less than*

$$p(v) = v\left(1 - \frac{2\alpha\mathbb{E}[S] + \alpha(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)\lambda\bar{F}(v)}{2(1-\rho\bar{F}(v))^2}\right). \tag{16}$$

If we assume that indifferent customers join the queue (in contrast to our earlier assumption), we can charge exactly $p(v)$. Note that it is beneficial to the firm to serve all customers willing to pay

a positive price, and hence, the optimal choice of $v_* = \max\{\inf \mathbf{X}, z_y\}$, with z_y as given in (12). The firm's revenue will be

$$R_{\mathrm{FI}} = \lambda \int_{v_*}^{\infty} v \left(1 - \frac{2\alpha\mathbb{E}[S] + \alpha(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)\lambda\bar{F}(v)}{2(1-\rho\bar{F}(v))^2}\right) f(v)\, dv. \tag{17}$$

Although the FI policy is not incentive-compatible, and therefore cannot be implemented in practice when customers are strategic, this policy provides an idealized upper bound on the feasible revenues from all incentive-compatible policies.

7 Closed form results for specific distributions

In this section we apply our implicit results from Sections 4 and 5 to provide closed form pricing menus, $p(v)$, (and the associated value of v_*) for the PP and FI policies for general α, λ, $\mathbb{E}[S]$, and $\mathbb{E}[S^2]$, where customer valuations v are drawn from the Exponential (Section 7.1), Uniform (Section 7.2), or Pareto (Secion 7.3) distributions. For the optimal revenue, R, closed forms do not always exist; we present R in closed form for the special case of M/M/1 queues (for all three valuation distributions), where we set $\mathbb{E}[S^2] = 2$ and $\mathbb{E}[S] = 1$. We compute R in the M/G/1 case numerically.

7.1 Customer valuations drawn from the Exponential distribution

We now assume that $v \sim \mathrm{Exponential}(\beta)$. Without loss of generality, we may set the scale parameter $\beta = 1$, and consequently $F(x) = 1 - e^{-x}$ and $\mathbb{E}[v] = 1$.

- Under PP, customers are served if and only if $v > v_*$, where

$$v_* = \begin{cases} 1 & \text{if } \alpha \le \alpha_0 \\ \log\left(\dfrac{4\rho\mathbb{E}[S]}{4\mathbb{E}[S] + \alpha(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2) - \sqrt{8\alpha\mathbb{E}[S]\mathbb{E}[S^2] + \alpha^2(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)^2}}\right) & \text{if } \alpha \ge \alpha_0 \end{cases}$$

  and

$$\alpha_0 = \frac{2(e-\rho)^2}{2e^2\mathbb{E}[S] - \lambda e(\mathbb{E}[S^2] - 2\mathbb{E}[S])}.$$

- The PP policy price menu (with v_* as given above) computed from (8) is

$$p(v) = v_* + \alpha\mathbb{E}[S]\log(e^v - \rho) + \frac{\alpha\lambda\mathbb{E}[S^2]}{2(e^{v_*} - \rho)} - \alpha\mathbb{E}[S]\log(e^{v_*} - \rho) + \frac{\alpha(\lambda^2\mathbb{E}[S]\mathbb{E}[S^2] - 2\mathbb{E}[S]ve^{2v} - \lambda e^v((\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)v + \mathbb{E}[S^2])}{2(e^v - \rho)^2}.$$

- Under PP, in the special case of the M/M/1 setting, where $\mathbb{E}[S] = 1$ and $\mathbb{E}[S^2] = 2$:

$$R^{\text{M/M/1}} = \begin{cases} \dfrac{\lambda}{e} + \alpha\left(\log(e-\lambda) - 1\right) & \text{if } \alpha \le (1-\lambda/e)^2 \\ \dfrac{\alpha\log(\alpha)}{2} + (1-\sqrt{\alpha})\left(\sqrt{\alpha} + (1-\sqrt{\alpha})\log\left(\dfrac{\lambda}{1-\sqrt{\alpha}}\right)\right) & \text{if } \alpha > (1-\lambda/e)^2 \end{cases}.$$

  We compute $R^{\mathrm{M/G/1}}$ numerically (not shown).

- Under FI, customers are served if and only if $v \ge v_*$, where

$$v_* = \begin{cases} 0 & \text{if } \alpha \le \alpha_0 \\ \log\left(\dfrac{4\rho\mathbb{E}[S]}{4\mathbb{E}[S] + \alpha(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2) - \sqrt{8\alpha\mathbb{E}[S]\mathbb{E}[S^2] + \alpha^2(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)^2}}\right) & \text{if } \alpha \ge \alpha_0 \end{cases}$$

  and

$$\alpha_0 = \frac{2(1-\rho)^2}{2\mathbb{E}[S] + \lambda(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)}.$$

- The FI policy price menu computed from (16) is

$$p(v) = v\left(1 - \frac{\alpha e^v\left(2\mathbb{E}[S]e^v + \lambda(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)\right)}{2(e^v - \rho)^2}\right).$$

- Under FI, in the special case of the M/M/1 setting, where $\mathbb{E}[S] = 1$ and $\mathbb{E}[S^2] = 2$:

$$R_{\text{FI}}^{\text{M/M/1}} = \begin{cases} \lambda + \alpha\log(1-\lambda) & \text{if } \alpha \le (1-\lambda)^2 \\ \dfrac{\alpha\log(\alpha)}{2} + (1-\sqrt{\alpha})\left(\sqrt{\alpha} + (1-\sqrt{\alpha})\log\left(\dfrac{\lambda}{1-\sqrt{\alpha}}\right)\right) + (1-\sqrt{\alpha})^2 & \text{if } \alpha > (1-\lambda)^2 \end{cases}$$

  It follows that $\lim_{\alpha\to 1}(R^{\mathrm{M/M/1}}/R_{\text{FI}}^{\text{M/M/1}}) = 1$, while $\lim_{\alpha\to 0}(R^{\mathrm{M/M/1}}/R_{\text{FI}}^{\text{M/M/1}}) = \bar{F}(1) = 1/e$.

7.2 Customer valuations drawn from the Uniform distribution

We now assume that $v \sim \mathrm{Uniform}(0, b)$. Without loss of generality, we may set $b = 2$, and consequently $F(x) = x/2$ for $0 \le x \le 2$ and $\mathbb{E}[v] = 1$.

- Under PP, customers are served if and only if $v > v_*$, where

$$v_* = \begin{cases} 1 & \text{if } \alpha \le \alpha_0 \\ \dfrac{2(\alpha+2\lambda)\mathbb{E}[S]^2 - 4\mathbb{E}[S] - \alpha\mathbb{E}[S^2] + \sqrt{\alpha 8\mathbb{E}[S]\mathbb{E}[S^2] + \alpha^2(\mathbb{E}[S^2] - \mathbb{E}[S]^2)^2}}{2\rho\mathbb{E}[S]} & \text{if } \alpha \ge \alpha_0 \end{cases}$$

  and

$$\alpha_0 = \frac{(2-\rho)^2}{4\mathbb{E}[S] - \lambda(\mathbb{E}[S^2] - 2\mathbb{E}[S])}.$$

- The PP policy price menu (with v_* as given above) computed from (8) is

$$p(v) = v_* + \frac{\alpha\mathbb{E}[S^2](v_*-2)}{(\rho(v_*-2)+2)\mathbb{E}[S]} - \frac{2\alpha(\mathbb{E}[S^2]-(\mathbb{E}[S^2]-\mathbb{E}[S]^2)v)}{(\rho(v-2)+2)\mathbb{E}[S]} - \frac{2\alpha v\mathbb{E}[S^2]}{(\rho(v-2)+2)^2\mathbb{E}[S]} + \frac{\alpha(\mathbb{E}[S^2]-2\mathbb{E}[S]^2)(\log(\rho(v_*-2)+2)-\log(\rho(v-2)+2))}{\rho\mathbb{E}[S]}$$

- Under PP, in the special case of the M/M/1 setting, where $\mathbb{E}[S]=1$ and $\mathbb{E}[S^2]=2$:

$$R^{\mathrm{M/M/1}} = \begin{cases} \dfrac{\lambda^2 - 4\alpha(\log(4)-\lambda) + 8\alpha\log(2-\lambda)}{2\lambda} & \text{if } \alpha \le (1-\lambda/2)^2 \\ \dfrac{2(2\sqrt{\alpha}(2-\lambda) - \alpha(3-\lambda) - (1-\lambda) + \alpha\log(\alpha))}{\lambda} & \text{if } \alpha > (1-\lambda/2)^2. \end{cases}$$

- Under FI, customers are served if and only if $v \ge v_*$, where

$$v_* = \begin{cases} 0 & \text{if } \alpha \le \alpha_0 \\ \dfrac{2(\alpha+2\lambda)\mathbb{E}[S]^2 - 4\mathbb{E}[S] - \alpha\mathbb{E}[S^2] + \sqrt{\alpha 8\mathbb{E}[S]\mathbb{E}[S^2] + \alpha^2(\mathbb{E}[S^2]-\mathbb{E}[S]^2)^2}}{2\rho\mathbb{E}[S]} & \text{if } \alpha \ge \alpha_0 \end{cases}$$

  and

$$\alpha_0 = \frac{2(1-\rho)^2}{2\mathbb{E}[S] + \lambda(\mathbb{E}[S^2]-2\mathbb{E}[S]^2)}.$$

- The FI policy price menu computed from (16) is

$$p(v) = v\left(1 - \frac{2\alpha\mathbb{E}[S] + \alpha\lambda(1-v/2)(\mathbb{E}[S^2]-2\mathbb{E}[S]^2)}{2(1-\rho(1-v/2))^2}\right).$$

- Under FI, in the special case of the M/M/1 setting, where $\mathbb{E}[S]=1$ and $\mathbb{E}[S^2]=2$:

$$R_{\mathrm{FI}}^{\mathrm{M/M/1}} = \begin{cases} 2\alpha + \lambda + \dfrac{2\alpha\log(1-\lambda)}{\lambda} & \text{if } \alpha \le (1-\lambda)^2 \\ \dfrac{4\sqrt{\alpha}(1-\lambda) + (2\lambda-1) - \alpha(3-2\lambda) + \alpha\log(\alpha)}{\lambda} & \text{if } \alpha > (1-\lambda)^2. \end{cases}$$

  It follows that $\lim_{\alpha\to 1}(R^{\mathrm{M/M/1}}/R_{\mathrm{FI}}^{\mathrm{M/M/1}}) = 1$, while $\lim_{\alpha\to 0}(R^{\mathrm{M/M/1}}/R_{\mathrm{FI}}^{\mathrm{M/M/1}}) = \bar{F}(1) = 1/2$.

7.3 Customer valuations drawn from the Pareto distribution

Finally, for our last specific distribution, we let v take a Pareto distribution with a scale parameter of 2, and (without loss of generality) $\mathbb{E}[v]=1$. That is, $F(x) = 1-(1+x)^{-2}$ for $x \ge 0$. Note that we have shifted the conventional Pareto distribution so that the support takes 0, rather than 1, as its lower bound, for consistency with the other distributions studied.

- Under PP, customers are served if and only if $v > v_*$, where

$$v_* = \begin{cases} 1 & \text{if } \alpha \le \alpha_0 \\ \left(\dfrac{4\rho\mathbb{E}[S]}{4\mathbb{E}[S] - 2\alpha\mathbb{E}[S]^2 + \alpha\mathbb{E}[S^2] - \sqrt{8\alpha\mathbb{E}[S]\mathbb{E}[S^2] + \alpha^2(\mathbb{E}[S^2] - 2\mathbb{E}[S])^2}} \right)^{1/2} - 1 & \text{if } \alpha \ge \alpha_0 \end{cases}$$

and

$$\alpha_0 = \frac{(4-\rho)^2}{16\mathbb{E}[S] - 2\lambda(\mathbb{E}[S^2] - 2\mathbb{E}[S])}.$$

- The PP policy price menu (with v_* as given above) computed from (8) is

$$\begin{aligned} p(v) = v_* &+ \alpha\mathbb{E}[S](v - v_*) + \frac{\alpha\mathbb{E}[S^2](v - v_*)}{4\mathbb{E}[S]} - \frac{\alpha\mathbb{E}[S^2]v(v+1)^4}{2\mathbb{E}[S]((v+1)^2 - \rho)^2} \\ &+ \frac{\alpha(\mathbb{E}[S^2](v-1) - 4v\mathbb{E}[S^2]}{4\mathbb{E}[S]((v+1)^2 - \rho)} + \frac{\alpha\mathbb{E}[S^2](v_*+1)^3}{\mathbb{E}[S]((v_*+1)^2 - \rho)} \\ &+ \frac{\alpha\sqrt{\lambda}(4\mathbb{E}[S]^2 + \mathbb{E}[S^2])}{2\sqrt{\mathbb{E}[S]}} \left(\log\left(\frac{1 - 2(v_*+1)}{1 + v_* - \sqrt{\rho}} \right) - \log\left(\frac{1 - 2(v+1)}{1 + v - \sqrt{\rho}} \right) \right) \end{aligned}$$

- Under PP, in the special case of the M/M/1 setting, where $\mathbb{E}[S] = 1$ and $\mathbb{E}[S^2] = 2$:

$$R^{\mathrm{M/M/1}} = \begin{cases} \dfrac{1}{4}\left(\lambda - 2\alpha\sqrt{\lambda}\coth^{-1}\left(\dfrac{2}{\sqrt{\lambda}} \right) \right) & \text{if } \alpha \le \alpha_0 \\ \dfrac{(2-\sqrt{\alpha})\sqrt{\lambda(1-\sqrt{\alpha})}}{2} - \left(1 - \sqrt{\alpha}\right)^2 - \dfrac{\alpha\sqrt{\lambda}\tanh^{-1}\left(\sqrt{1-\sqrt{\alpha}}\right)}{2} & \text{if } \alpha > \alpha_0 \end{cases}$$

- Under FI, customers are served if and only if $v \ge v_*$, where

$$v_* = \begin{cases} 0 & \text{if } \alpha \le \alpha_0 \\ \left(\dfrac{4\rho\mathbb{E}[S]}{4\mathbb{E}[S] - 2\alpha\mathbb{E}[S]^2 + \alpha\mathbb{E}[S^2] - \sqrt{8\alpha\mathbb{E}[S]\mathbb{E}[S^2] + \alpha^2(\mathbb{E}[S^2] - 2\mathbb{E}[S])^2}} \right)^{1/2} - 1 & \text{if } \alpha \ge \alpha_0 \end{cases}$$

and

$$\alpha_0 = \frac{2(1-\rho)^2}{2\mathbb{E}[S] + \lambda(\mathbb{E}[S^2] - 2\mathbb{E}[S]^2)}.$$

- The FI policy price menu computed from (16) is

$$p(v) = v - \frac{\alpha v\mathbb{E}[S]}{(1 - \rho/(1+v)^2)^2} + \frac{\alpha\lambda v(\mathbb{E}[S^2] - 2\mathbb{E}[S])}{2(1+v)^2(1 - \rho/(1+v)^2)^2}.$$

- Under FI, in the special case of the M/M/1 setting, where $\mathbb{E}[S] = 1$ and $\mathbb{E}[S^2] = 2$:

$$R_{\mathrm{FI}}^{\mathrm{M/M/1}} = \begin{cases} \lambda - \alpha\sqrt{\lambda}\tanh^{-1}\left(\sqrt{\lambda}\right) & \text{if } \alpha \leq (1-\lambda)^2 \\ \dfrac{2\lambda v_* + \lambda}{(v_*+1)^2} - \dfrac{\alpha\lambda v_*}{(v_*+1)^2 - \lambda} + \left(\dfrac{\alpha\sqrt{\lambda}}{2}\right)\log\left(\dfrac{2v_*+2}{v_*+1+\sqrt{\lambda}} - 1\right) & \text{if } \alpha > (1-\lambda)^2, \end{cases}$$

  where when $\alpha > (1-\lambda)^2$, we have

$$v_* = \sqrt{\frac{\lambda}{1-\sqrt{\alpha}}} - 1.$$

8 Comparison of policies and the impact of service requirement variability

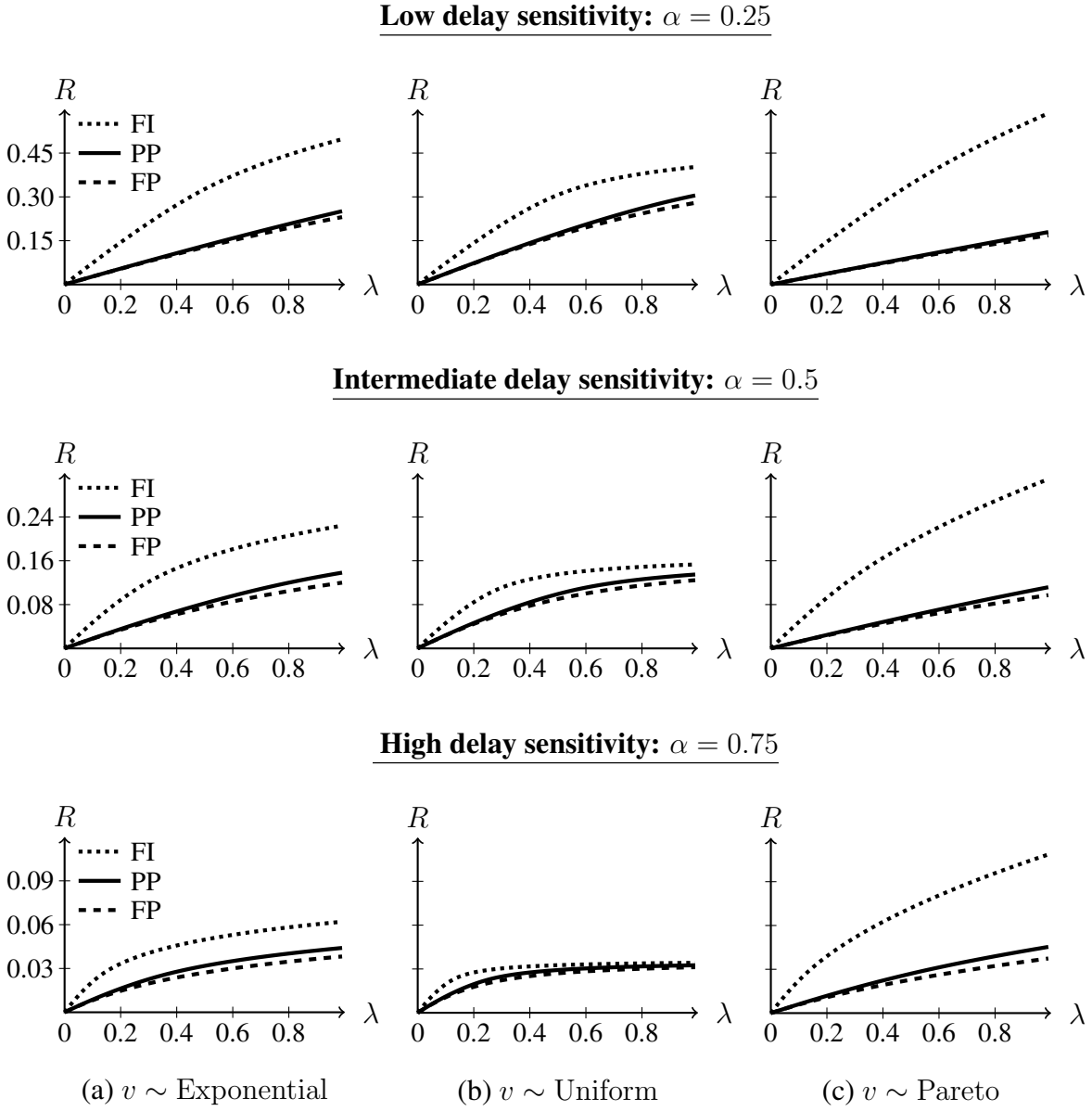
In this section, we conduct numerical experiments to test the efficacy (in terms of the revenue) of PP as compared to the two benchmark pricing policies—FP and FI—described in Section 6. For simplicity, we restrict ourselves to the M/M/1 setting where $\mathbb{E}[S] = 1$ and $\mathbb{E}[S^2] = 2$, so we can use our closed-form results from Section 7. In order to compute the the revenue under FP from (15), we must resort to numerical methods. Furthermore, we use numerical integration to study the impact of the variability of the customer service requirement distribution on the revenue obtained from PP introduced in M/G/1 queueing settings.

Figure 1 shows plots of the revenue obtained for PP, FP, and FI as a function of the arrival rate λ for the cases of $\alpha = 0.25$, 0.5, and 0.75, corresponding to low, medium, and high delay sensitivity, respectively.[7] As suggested by our theoretical results, as α approaches 1, the performance of PP approaches that of FI in the cases where customer valuations, v, are drawn from the Exponential and Uniform distributions.

Figure 2 shows the percentage improvement of PP over FP (i.e., $R/R_{FP} - 1$) as a function of λ for the cases of $\alpha = 0.25$, 0.5, and 0.75. We confirm that PP outperforms FP in all the cases we tested, as we have proven would hold for these distributions (see Theorem 5). Moreover, the improvement seems sharpest in the test cases where $\alpha = 0.75$ (that is, when customers are more delay sensitive). When customers are more delay sensitive (e.g., $\alpha = 0.5$ or $\alpha = 0.75$) and customer valuations are drawn from the Exponential or Uniform distributions, the improvement is most pronounced for intermediate values of λ in the range of 0.2–0.6, whereas in the remaining cases, where customer valuations are drawn from the Pareto distribution, or delay sensitivity is low, the improvement continues to increase as λ (and hence the load ρ) approaches 1.

Moving beyond the M/M/1 system, we study the impact of customer service requirement variability on the revenue obtained via the PP policy. That is, we keep $\mathbb{E}[S] = 1$ fixed and observe the

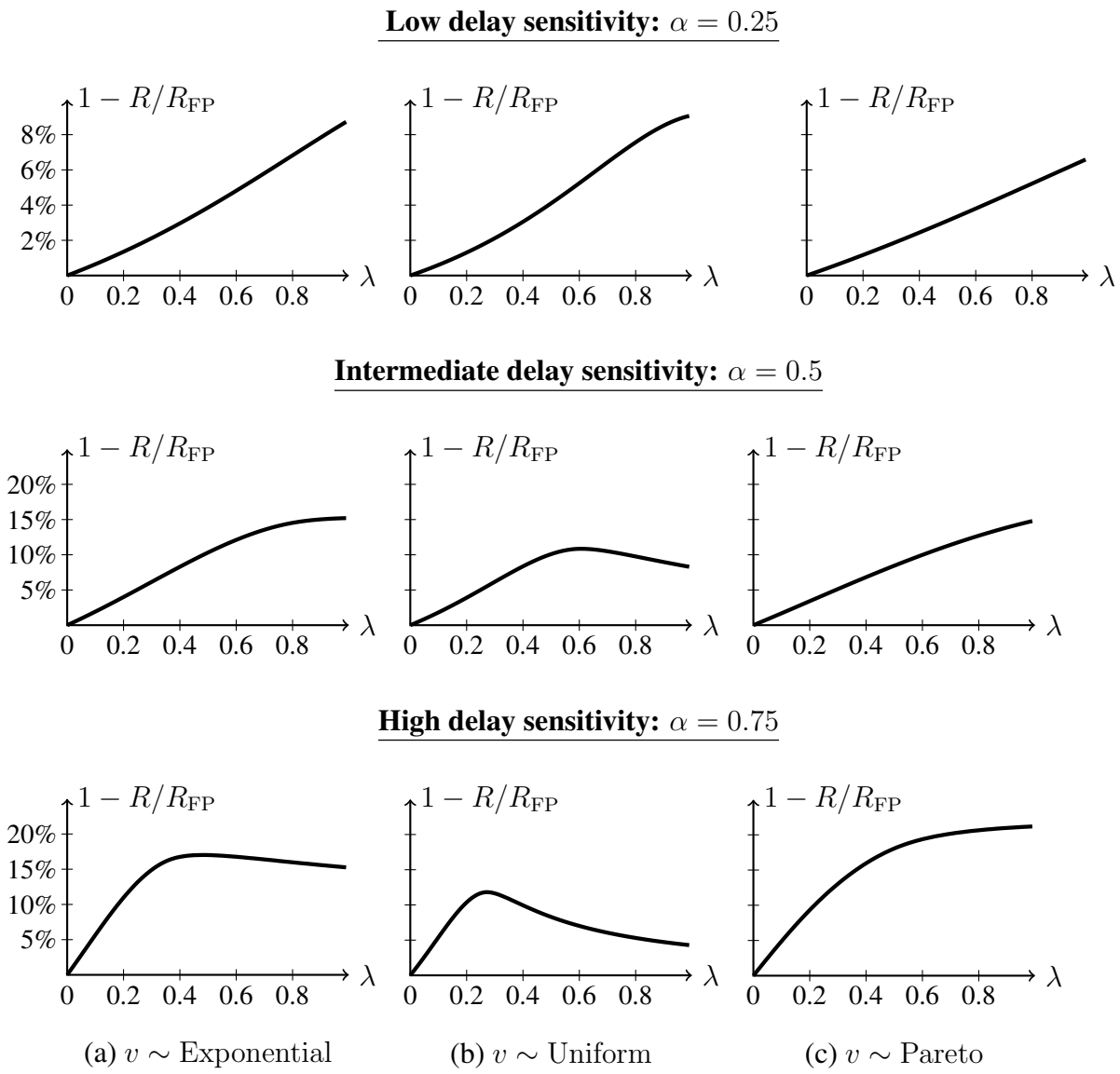
[7] Note that $\alpha \in (0, 1)$, since $\mathbb{E}[S] = 1$.



(a) $v \sim$ Exponential (b) $v \sim$ Uniform (c) $v \sim$ Pareto

Figure 1: Revenue versus arrival rate (λ) in three delay sensitivity regimes for PP (solid), FP (dahed), and FI (dotted). We assume an M/M/1 system where $\mathbb{E}[S] = 1$, $\mathbb{E}[S^2] = 2$, and customer valuations are drawn from the (a) Exponential, (b) Uniform, and (c) Pareto distributions from Section 7.

impact of $\mathbb{E}[S^2]$ on revenue. Note that in this case $\mathbb{E}[S^2] = \mathrm{Var}(S) + 1$. Figure 3 shows a plot of R vs. $\mathbb{E}[S^2]$ when $\alpha = \lambda = 0.5$. We observe that although revenue is always decreasing in variability, revenue does not decline significantly from $\mathbb{E}[S^2] = 100$ to $\mathbb{E}[S^2] = 200$ for any of the distributions studied. Revenue decreases when $\mathbb{E}[S^2]$ increases, as the increase in $\mathbb{E}[S^2]$ leads to



(a) $v \sim$ Exponential (b) $v \sim$ Uniform (c) $v \sim$ Pareto

Figure 2: Percentage increase in revenue generated by PP over the revenue generated by FP (i.e., $1 - R/R_{\text{FP}}$) as a function of the arrival rate, λ (in three delay sensitivity regimes). We assume an M/M/1 system where $\mathbb{E}[S] = 1$ and $\mathbb{E}[S^2] = 2$, and customer valuations are drawn from the (a) Exponential, (b) Uniform, and (c) Pareto distributions from Section 7.

an increase in response time, and therefore a decrease in surplus for all customers. Moreover, as $\mathbb{E}[S^2]$ rises, the set of customers admitted also begins to fall, that is, v_* begins to decline and fewer customers are paying for service.

Figure 4 demonstrates the impact of variability on the PP price menu. For all three distributions, we see that there exists some value $\tilde{v}$ such that $p(v)$ is increasing in $\mathbb{E}[S^2]$ for all $v > \tilde{v}$, while $p(v)$ is

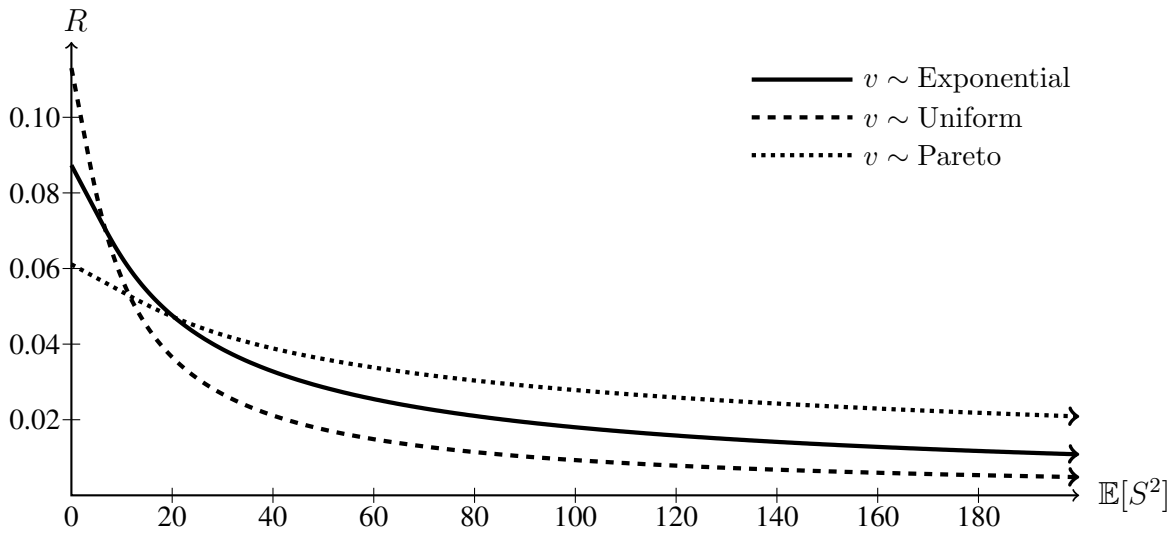


Figure 3: The revenue under the PP policy, R, as a function of $\mathbb{E}[S^2]$ when $\mathbb{E}[S] = 1$ (i.e., $\mathbb{E}[S^2] = \mathrm{Var}(S) + 1$), and $\alpha = \lambda = 0.5$. Customer valuations are drawn from the Exponential (solid), Uniform (dashed), and Pareto (dotted) distributions from Section 7.

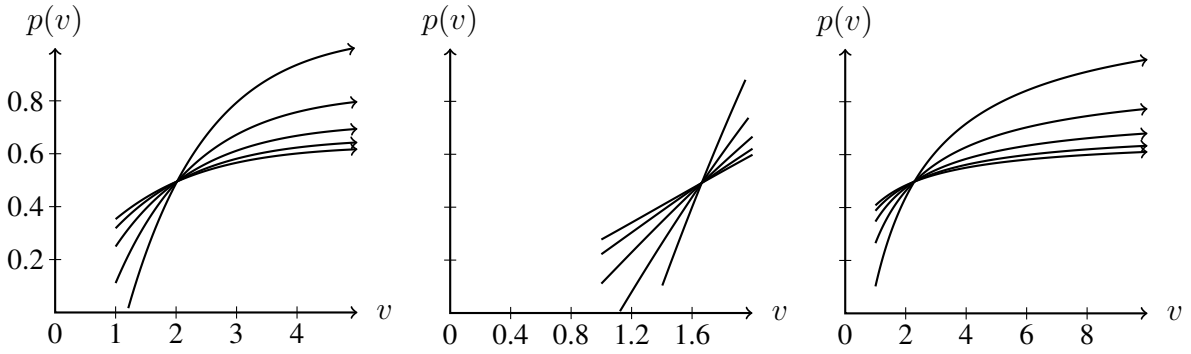
decreasing in $\mathbb{E}[S^2]$ for all $v < \tilde{v}$. That is, as variability increases, the highest valuation customers end up paying more, while the lowest valuation customers end up paying *less*, if they join the queue at all. We interpret this result as being due to the fact that higher valuation customers are impacted less by variability, as they have priority over most customers, while lower valuation customers are impacted the most, and hence, they are not able to pay as much for service. We were unable to derive $\tilde{v}$ in closed form.

Perhaps the most striking regarding observation regarding the impact of service requirement variability is as follows: while revenue generated under PP is greatest in this setting in the case of Uniformly distributed customer valuations and least for the case of Pareto distributed customer valuations when variability is low (as in the M/M/1 setting), this ordering is reversed under high variability. The Pareto case yields greater revenue that the Uniform case at around $\mathbb{E}[S^2] = 10$ and greater revenue than the Exponential case at around $\mathbb{E}[S^2] = 20$.

9 Discrete pricing

Thus far, we have proposed the use of a continuous pricing policy and found the revenue-optimal incentive-compatible price menu in terms of the continuous function $p(v)$. In this setting, customers can choose from a continuum of priority levels. Although we proved in Theorem 5 that the

The impact of service requirement variability on prices



(a) $v \sim$ Exponential (b) $v \sim$ Uniform (c) $v \sim$ Pareto

Figure 4: The PP price menu, $p(v)$, when $\mathbb{E}[S] = 1$ and $\mathbb{E}[S^2] = 1/2, 1, 2, 4, 8$ (from top to bottom on the left-hand side; reversed on the right-hand side), and $\alpha = \lambda = 0.5$. Customer valuations are drawn from the (a) Exponential, (b) Uniform, and (c) Pareto distributions from Section 7.

revenue-maximizing incentive-compatible price menu is a continuous pricing function when customer valuations are regularly distributed (e.g., Exponential, Uniform, or Pareto), in many settings it is not practical to offer a continuous menu of prices: the customers' decisions may be unnecessarily complicated and the service provider may have trouble representing this information to the customers. It may also be the case that it is too costly or otherwise impossible for a customer to communicate an arbitrarily precise valuation to the service provider, or it requires too much time for the service provider to sufficiently fine-tune a continuous price menu. Instead many services are offered according to *a finite number of discrete priority classes*. In this section we numerically study the effect of offering a discrete number of priority classes, and find that in many cases the profits that can be obtained from offering as few as three or four priority classes are comparable to those that can be obtained by offering a continuous menu of prices.

When the firm is restricted to offering only n preemptive priority levels at n prices $p_1 > p_2 > \cdots > p_n$, by a result analogous to Theorem 2, there exist thresholds $v_1 \geq v_2 \geq \cdots \geq v_n$, such that a customer of type $v \leq v_n$ opts not to receive service, a customer of type $v \in (v_i, v_{i-1})$ pays p_i for the i-th priority level, and a customer of type $v > v_1$ pays p_1 for the first priority level. Customers in the i-th priority level receive preemptive queueing priority over all customers in priority levels $i + 1$ and beyond. When implementing the optimal prices, we denote this policy by PP(n). We let R_n be the optimal revenue obtained under PP(n). Note that PP(1) is the same as Fixed Pricing (FP), while as $n \to \infty$ PP(n) approaches Priority Pricing (PP).

In order to guarantee incentive-compatibility it is sufficient that customers of type v_i be indifferent between being in priority level i and paying p_i and being in priority level $i + 1$ and paying p_{i+1} whenever $i < n$, while customers of type v_n are indifferent between paying p_n for the last

priority level and not joining the system at all. To see why this guarantees incentive-compatibility, let $\epsilon > 0$ and consider some customer $v_i - \epsilon$ who strictly prefers class i over class $i-1$. Since prices are nonnegative, however, this would imply customer v_i prefers class i to class $i-1$, by an even greater margin, but this contradicts the assumption that customer v_i is indifferent between these classes, so no such customer $v_i - \epsilon$ exists. A similar argument shows that no customer $v_i + \epsilon$ prefers class $i-1$ over class i.

We proceed to analyze these policies by calculating the delay experienced by the customers. We let τ_i be the mean response time of a customer with type $v \in (v_i, v_{i-1})$, assuming truthful reporting. From queuing theory, τ_1 is the expected response time in a FCFS M/G/1 queue with arrival rate $\lambda \bar{F}(v_1)$,

$$\tau_1 = \mathbb{E}[S] + \frac{\mathbb{E}[S^2]}{2(1 - \rho \bar{F}(v_1))},$$

while for all $i \geq 2$, τ_i is a busy period perpetuated by arrivals with valuation v_{i-1} and above started by the expected work in a FCFS system made up of customers with valuation v_i and above:

$$\forall i \in \{2, 3, \ldots, n\}: \qquad \tau_i = \frac{\mathbb{E}[S]}{1 - \rho \bar{F}(v_{i-1})} + \frac{\lambda \bar{F}(v_i) \mathbb{E}[S^2]}{2(1 - \rho \bar{F}(v_{i-1}))(1 - \rho \bar{F}(v_i))}.$$

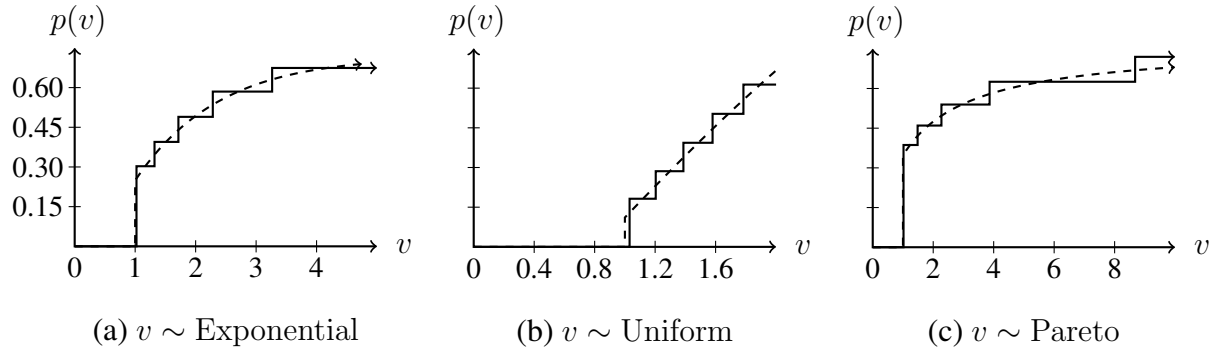
Therefore, the optimal choice of $\{p_i\}_{i=1}^n$ and the resulting $\{v_i\}_{i=1}^n$ are determined by solving the following nonlinear optimization program, with the objective value yielding R_n, the optimal revenue with n preemptive priority levels:

$$\begin{aligned} R_n = \max_{\{p_i, v_i\}_{i=1}^n} \quad & \lambda p_1 \bar{F}(v_1) + \lambda \sum_{i=2}^{n} \left\{ p_i (\bar{F}(v_{i-1}) - \bar{F}(v_i)) \right\} \\ \text{s.t.} \quad & v_i(1 - \alpha \tau_i) - p_i = v_i(1 - \alpha \tau_{i+1}) - p_{i+1} \qquad (\forall i \in \{1, 2, \ldots, n-1\}); \\ & v_n(1 - \alpha \tau_n) - p_n = 0; \\ & 0 < p_n \leq p_{n-1} \leq \cdots \leq p_1; \\ & 0 < v_n \leq v_{n-1} \leq \cdots \leq v_1; \\ & v_1, v_2, \ldots, v_n \in \mathbf{X}. \end{aligned} \tag{18}$$

We solve this nonlinear program numerically in the M/M/1 setting where $\mathbb{E}[S] = 1$ and $\mathbb{E}[S^2] = 2$. Our numerical findings suggest that the optimal $v_n \approx v_*$, the optimal threshold in the continuous case, even for small values of n.

Figure 5 shows the discrete pricing curves as a function of customer valuations (that is the price paid by a customer of a given type v) when the number of priority classes offered is $n = 5$. The lowest threshold v_5 is typically very close to the threshold v_* in the optimal (continuous) case. In virtually all cases, the thresholds $v_1, \ldots, v_5$ for Uniformly distributed v are nearly equally spaced, and hence, each priority level is purchased by roughly a $1/n$ fraction of the customers opting to join the queue. This is possibly due to the fact that the continuous pricing function is approximately linear in this case. For the other distributions, the pattern is not as simple.

As expected, our findings show that R_n is increasing in n (i.e., offering more priority classes yields greater revenue). We also wish to understand the number of classes that are sufficient for



(a) $v \sim$ Exponential (b) $v \sim$ Uniform (c) $v \sim$ Pareto

Figure 5: A juxtaposition of the PP(5) (solid) and PP (dotted) price menus where $\mathbb{E}[S] = 1$, $\mathbb{E}[S^2] = 2$, and $\alpha = \lambda = 0.5$, and customer valuations v are drawn from the (a) Exponential, (b) Uniform, and (c) Pareto distributions from Section 7.

capturing nearly all the available revenue. More formally, we study the loss in revenue as compared to the optimal revenue achievable via continuous pricing, R.

Figure 6(a) shows $R_1, R_2, \ldots, R_6$ (i.e., the optimal revenue obtained when the number of priority classes ranging from $n = 1$ to $n = 6$) in the case where $\lambda = \alpha = 0.5$, $\mathbb{E}[S] = 1$ and $\mathbb{E}[S^2] = 2$. For all the distributions studied there is little difference between R_3, R_4, R_5, R_6, and R (i.e., the optimal revenue under PP). In fact, a log-log plot of the proportional loss in revenue due to discretization, $1 - R_n/R$, versus the number of classes, n, for all the distributions studied suggests a linear decline with a slope of -2 (see Figure 6(b)). Using other values of λ and α yields plots similar to those in Figure 6. In particular the nearly linear decline with a slope of -2 (although with slightly different intercepts) was present in all cases examined. Therefore, it appears that the proportional loss in revenue, $1 - R_n/R \sim C/n^2$ as $n \to \infty$ for some constant C (i.e., $\lim_{n\to\infty} n^2(1 - R_n/R) \to C$). This is consistent with theoretical results on the efficiency of discrete price menus in other settings without queueing [46, 10, 11]. This suggests that a few priority classes are sufficient to capture nearly all the revenue available. We find that R_5 and R_6 are typically within 1% of optimal revenue, R.

10 Conclusion

We propose a pricing model for an *unobservable* M/G/1 queueing system that allows for infinitely many customer types. In particular, strategic customers are parametrized by their valuation, v, and their delay sensitivity, $c = \alpha v$, with constant α. We allow v to be a random variable drawn from a continuous distribution with cdf F. This extends upon the previous work in the area of revenue (profit) maximization for priority queues, where there are only finitely many customer types and

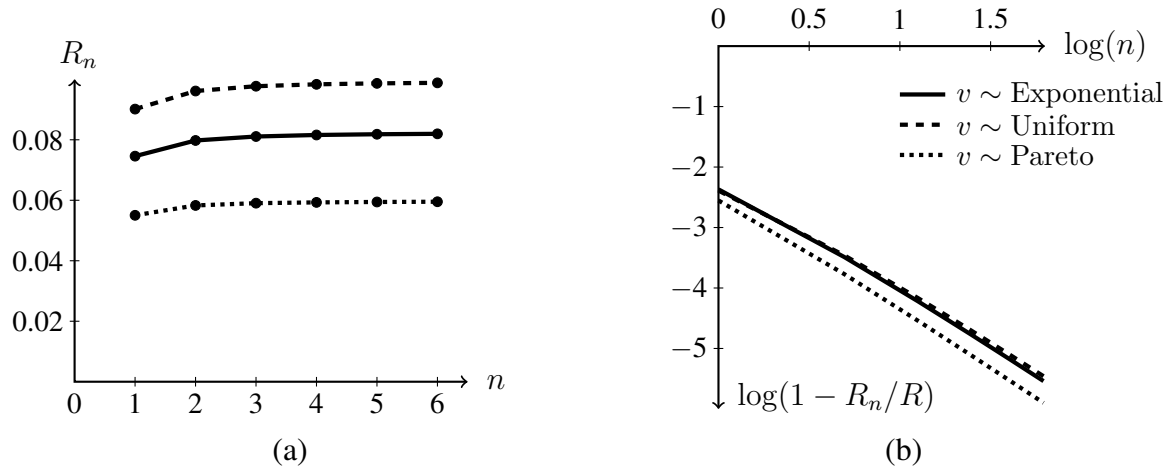


Figure 6: Plots of (a) the revenue R_n and (b) the fractional loss in revenue (compared to PP) as the number of discrete priority classes offered ranges from $n = 1$ to $n = 6$. In (a) we have a linear-linear plot and in (b) we have a log-log plot. The plots illustrate a representative case where $\mathbb{E}[S] = 1$, $\mathbb{E}[S^2] = 2$, and $\alpha = \lambda = 0.5$ for valuations drawn form the Exponential (solid), Uniform (dashed) and Pareto (dotted) distributions from Section 7.

priority classes, customers differ in only one way, or customers are non-strategic.

We derive the revenue-maximizing incentive-compatible priority price menus, $p(v)$, (where customers pay a greater price for greater priority) in closed form. Revenues, R, are computed numerically in the M/G/1 setting and in exact closed-form in the more restrictive M/M/1 setting. We solve this optimization problem in the cases where the customer valuation distribution is Exponential, Uniform, or Pareto and then compare our results with the Fixed Pricing (FP) and Full Information (FI) policies. Using the Myerson revenue-optimal auction framework, we are able to show that for these distributions on customer valuations, Priority Pricing (PP) is the best pricing policy, when priority levels must remain static; more generally, it is the optimal pricing policy whenever the distribution on customer valuations is regular. Gains of 2–20% over FP are typical for the PP policy. The M/G/1 setting also allows us to understand the significant role played by service requirement variability. Moreover, using only a discrete number of priority classes often does nearly as well as the the PP policy: offering five discrete classes, PP(5), typically yields revenues within 1% of the optimal (PP).

There are many avenues for further research in this area. For example, we could relax the assumption that $c = \alpha v$, and consider a joint-distribution of valuations and delay-sensitivities. Moreover, one could relax the assumption that customer types and service requirements are independent. Another direction for further research would be investigating policies where a customer's priority could change over time, based on, for example, how long she has been in the system or how much service she has received since entering the system.

References

[1] ABHISHEK, V., KASH, I., AND KEY, P. Fixed and market pricing for cloud services. In *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on* (2012), IEEE, pp. 157–162.

[2] ADIRI, I., AND YECHIALI, U. Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. *Operations Research 22*, 5 (1974), 1051–1066.

[3] AFÈCHE, P. Incentive-compatible revenue management in queueing systems: optimal strategic delay and capacity. *Working paper, University of Toronto* (2010).

[4] AFÈCHE, P., AND PAVLIN, M. Optimal price-lead time menus for queueing systems with customer choice: Priorities, pooling and strategic delay. *Working paper, Univeristy of Toronoto* (2011).

[5] AKAN, M., ATA, B., AND OLSEN, T. Congestion-based lead-time quotation for heterogenous customers with convex-concave delay costs: Optimality of a cost-balancing policy based on convex hull functions. *Operations Research 60*, 6 (2012), 1505–1519.

[6] ALPERSTEIN, H. Optimal pricing policy for the service facility offering a set of priority prices. *Management Science 34*, 5 (1988), 666–671.

[7] ATA, B., AND OLSEN, T. Congestion-based leadtime quotation and pricing for revenue maximization with heterogeneous customers. *Queueing Systems 73*, 1 (2013), 35–78.

[8] ATA, B., AND SHNEORSON, S. Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Science 52*, 11 (2006), 1778–1791.

[9] BALACHANDRAN, K. Purchasing priorities in queues. *Management Science 18*, 5, Part 1 (1972), 319–326.

[10] BERGEMANN, D., SHEN, J., XU, Y., AND YEH, E. Mechanism design with limited information: The case of nonlinear pricing.

[11] BERGEMANN, D., SHEN, J., XU, Y., AND YEH, E. Multi-dimensional mechanism design with limited information. In *Proceedings of the 13th ACM Conference on Electronic Commerce* (2012), ACM, pp. 162–178.

[12] BORGS, C., CHAYES, J., DOROUDI, S., HARCHOL-BALTER, M., AND XU, K. The optimal admission threshold in observable queues with state dependent pricing. Tech. rep., CMU-CS-12-145, School of Computer Science, Carnegie Mellon University, 2012.

[13] ÇELIK, S., AND MAGLARAS, C. Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. *Management Science 54*, 6 (2008), 1132–1146.

[14] CHARNSIRISAKSKUL, K., GRIFFIN, P., AND KESKINOCAK, P. Order selection and scheduling with leadtime flexibility. *IIE transactions 36*, 7 (2004), 697–707.

[15] CHARNSIRISAKSKUL, K., GRIFFIN, P. M., AND KESKINOCAK, P. Pricing and scheduling decisions with leadtime flexibility. *European Journal of Operational Research 171*, 1 (2006), 153–169.

[16] CHEN, H., AND FRANK, M. State dependent pricing with a queue. *IIE Transactions 33*, 10 (2001), 847–860.

[17] ÇIL, E. B., KARAESMEN, F., AND ÖRMECI, E. L. Dynamic pricing and scheduling in a multi-class single-server queueing system. *Queueing Systems 67*, 4 (2011), 305–331.

[18] COBHAM, A. Priority assignment in waiting line problems. *Operations Research 2*, 1 (1954), 70–76.

[19] DEWAN, S., AND MENDELSON, H. User delay costs and internal pricing for a service facility. *Management Science 36*, 12 (1990), 1502–1517.

[20] FUDENBERG, D., AND TIROLE, J. Game theory. 1991, 1991.

[21] GHANEM, S. Computing center optimization by a pricing-priority policy. *IBM Systems Journal 14*, 3 (1975), 272–291.

[22] GLAZER, A., AND HASSIN, R. Stable priority purchasing in queues. *Operations Research Letters 4*, 6 (1986), 285–288.

[23] HASSIN, R., AND HAVIV, M. Equilibrium threshold strategies: The case of queues with priorities. *Operations Research 45*, 6 (1997), 966–973.

[24] HASSIN, R., AND HAVIV, M. *To queue or not to queue: Equilibrium behavior in queueing systems*. Kluwer, 2003.

[25] KAPUSCINSKI, R., AND TAYUR, S. Reliable due-date setting in a capacitated mto system with two customer classes. *Operations research 55*, 1 (2007), 56–74.

[26] KATTA, A.-K., AND SETHURAMAN, J. Pricing strategies and service differentiation in queues — a profit maximization perspective. Tech. rep., CORC TR-2005-4, Columbia University, 2005.

[27] KESKINOCAK, P., RAVI, R., AND TAYUR, S. Scheduling and reliable lead-time quotation for orders with availability intervals and lead-time sensitive revenues. *Management Science 47*, 2 (2001), 264–279.

[28] KITTSTEINER, T., AND MOLDOVANU, B. Auction-based queue disciplines. *Working paper, University of Bonn* (2003).

[29] KITTSTEINER, T., AND MOLDOVANU, B. Priority auctions and queue disciplines that depend on processing time. *Management Science 51*, 2 (2005), 236–248.

[30] KLEINROCK, L. Optimum bribing for queue position. *Operations Research 15*, 2 (1967), 304–318.

[31] KOENIGSBERG, E. Queue systems with balking: a stochastic model of price discrimination. *RAIRO. Recherche opérationnelle 19*, 3 (1985), 209–219.

[32] MANDJES, M. Pricing strategies under heterogeneous service requirements. *Computer Networks 42*, 2 (2003), 231–249.

[33] MARCHAND, M. Priority pricing. *Management Science 20*, 7 (1974), 1131–1140.

[34] MCWHERTER, D., SCHROEDER, B., AILAMAKI, N., AND HARCHOL-BALTER, M. Priority mechanisms for OLTP and transactional web applications. In *the 20th* (Boston, MA, April 2004).

[35] MCWHERTER, D., SCHROEDER, B., AILAMAKI, N., AND HARCHOL-BALTER, M. Improving preemptive prioritization via statistical characterization of OLTP locking. In *the 21st* (San Francisco, CA, April 2005), pp. 446–457.

[36] MENDELSON, H., AND WHANG, S. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research 38*, 5 (1990), 870–883.

[37] MYERSON, R. B. Optimal auction design. *Mathematics of operations research 6*, 1 (1981), 58–73.

[38] MYERSON, R. B. Multistage games with communication. *Econometrica: Journal of the Econometric Society* (1986), 323–358.

[39] NAOR, P. The regulation of queue size by levying tolls. *Econometrica: Journal of the Econometric Society 37*, 1 (1969), 15–24.

[40] PHIPPS, T. Machine repair as a priority waiting-line problem. *Operations Research 4*, 1 (1956), 76–85.

[41] PLAMBECK, E. Optimal leadtime differentiation via diffusion approximations. *Operations Research 52*, 2 (2004), 213–228.

[42] RAO, S., AND PETERSEN, E. Optimal pricing of priority services. *Operations Research 46*, 1 (1998), 46–56.

[43] SCHRAGE, L., AND MILLER, L. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research 14*, 4 (1966), 670–684.

[44] SCHROEDER, B., HARCHOL-BALTER, M., IYENGAR, A., NAHUM, E., AND WIERMAN, A. How to determine a good multi-programming level for external scheduling. In *the 22nd* (Atlanta, GA, April 2006), pp. 60–70.

[45] STIDHAM JR., S. *Optimal design of queueing systems*. Chapman & Hall/CRC, 2009.

[46] WILSON, R. *Nonlinear Pricing: Published in association with the Electric Power Research Institute*. Oxford University Press, USA, 1997.

[47] YECHIALI, U. On optimal balking rules and toll charges in the GI/M/1 queuing process. *Operations Research 19*, 2 (1971), 349–370.

Appendix: Proof of Theorem 3

Before Proving Theorem 3, we will first prove the following Lemma.

Lemma 1. *For any incentive-compatible price function p, if there exist $z, z' \in \mathbf{E}$ such that $p(z) < 0 < p(z')$, then there exists some $y \in \mathbf{E}$ such that $p(y) = 0$.*

Proof. Proof of Lemma 1. Assume, by way of contradiction, that no such y exists. It immediately follows that either (i) p has discontinuity on the entry set $\mathbf{E}$, or (ii) $\mathbf{E} = \mathbf{E}_- \cup \mathbf{E}_+$, where p is negative on $\mathbf{E}_-$ and positive on $\mathbf{E}_+$ (and hence, the two sets are disjoint). We can rule out case (i), because such a discontinuity would violate the hypothesis that p is incentive-compatible (a customer on one side of the discontinuity would benefit from "jumping" to the other side). Similarly, in case (ii), it must be the case that $\sup \mathbf{E}_- < \inf \mathbf{E}_+$, (if the two were equal, say both are y', then the incentive compatibility constraint would not hold for y'). Therefore, by the continuity of waiting times, there exists a point in $\mathbf{E}_+$ which would increase its utility by reporting some type in $\mathbf{E}_-$, violating incentive compatibility. □

Proof. Proof of Theorem 3. We will prove the claim by construction. If all prices are positive on $\mathbf{E}$ (or if $\mathbf{E}$ is empty), then we are done (we can charge $\hat{p}(x) = +\infty$ for all $x \in \mathbf{X} \backslash \mathbf{E}$), and if all prices are negative on the entry set, then the price menu can be trivially improved by setting $\hat{p}(x) = +\infty$ for all x (i.e., not offer the service), so that $R(\hat{p}) = 0 \geq R(p)$. Otherwise, there exists $z, z' \in \mathbf{E}$ such that $p(z) < 0 < p(z')$, and so by Lemma 1 there exists $y \in \mathbf{E}$ such that $p(y) = 0$. Then $u_y(y) > 0$. Now let $w \equiv u_y(y)$ and

$$\hat{p}(x) \equiv \begin{cases} p(x) + w & \text{if } x \geq y, \\ w & \text{otherwise.} \end{cases}$$

We now show that $\hat{p}$ is incentive-compatible with the entry set $\mathbf{E}_y \equiv \mathbf{E} \cap (y, \infty)$. Let $\hat{u}$ denote utilities under $\hat{p}$ and note that $\hat{u}_y(y) = 0$.

For all $v \in \mathbf{E}_y$, understanding that the response time for these customers (and all reported types $x > y$) will be the same as it was under p and entry set $\mathbf{E}$, we have

$$\hat{u}_v(v) = u_v(v) - w \geq u_v(y) - w = \hat{u}_v(y) > \hat{u}_y(y) = 0$$

(since prices are nonnegative for all $x \geq y$) and so

$$\forall x \geq y \colon \hat{u}_v(v) = u_v(v) - w \geq u_v(x) - w = \hat{u}_v(x),$$

while

$$\forall x < y \colon \hat{u}_v(v) \geq \hat{u}_v(y) = \hat{u}_v(x),$$

as no customers are entering with valuations between x and y. Hence, the entering customers satisfy the incentive-compatibility constraint.

For all $v \notin \mathbf{E}$ with $v > y$, we have

$$\forall x \geq y \colon \hat{u}_v(x) = u_v(x) - w \leq -w \leq 0.$$

Meanwhile,

$$\forall x \leq y \colon \hat{u}_v(x) = \hat{u}_v(y) \leq 0,$$

so these customers also satisfy the incentive-compatibility constraints.

Finally, we consider $v < y$, and we see that

$$x \geq y \colon \hat{u}_v(x) \leq u_y(x) \leq u_y(y) = 0,$$

where $u_y(x) \leq u_y(y)$ follows from the argument that the customers with $v \geq y$, and in particular, $v = y$, satisfy the incentive-compatibility constraint for $\hat{p}$. Next, we see that

$$\forall x < y \colon \hat{u}_v(x) \leq \hat{u}_v(y) = 0.$$

Therefore, $\hat{p}$ is incentive compatible when the entry set is $\mathbf{E}_y$. Moreover, $R(\hat{p}) \geq R(p) + \lambda w \cdot \mathbb{P}(v \geq w) \geq R(p)$, proving the claim. □