Monopoly Pricing When Customers Queue

Hong Chen
Faculty of Commerce and Business Administration
University of British Columbia, Vancouver, B.C. Canada
and
Murray Frank [1]
Faculty of Commerce and Business Administration
University of British Columbia, Vancouver, B.C. Canada
and
School of Business and Management
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

March 30, 1995

Abstract

It takes time to process purchases and as a result a queue of customers may form. The pricing and service rate decisions of a monopolist who must take this into account are characterized. We find that an increase in the average number of customers arriving in the market either has no effect on the monopoly price, or else causes the monopolist to *reduce* the price in the short run. In the long run the monopolist will increase the service rate and raise the price. When customer preferences are linear the equilibrium is socially efficient. When preferences are not linear equilibrium will not normally be socially efficient.

Journal of Economic Literature codes: L 12, L 15.
Key Words: Monopoly, Queue, Customer Information, Service Rate, Social Welfare.

[1]For research support, the authors would like to thank respectively: a Killam Faculty Research Fellowship and a grant from the NSERC (Canada), and a grant from the SSHRC (Canada). We also thank Vojislav Maksimovic for very helpful comments. Please address correspondence to M. Frank in Hong Kong; for e-mail: mfrank@usthk.ust.hk

1 Introduction

The median delay in delivery of a purchase is more than a month in many industrial markets such as airplane manufacturing, ship building, textile mill products, steel, fabricated metals, nonelectric machinery, and electric machinery. There are at least two interesting features of market clearing in such industries. First, there is commonly more variation in delivery lag than in posted price. This suggests that delay and queueing phenomena play a crucial role in clearing such markets. Second, the queue exists on the books of the firm and so is often not directly observable by the potential customer who is considering placing an order. Carlton and Perloff (1994) provide a valuable review of the evidence on the importance of time and delay in market clearing.

In this paper we study a monopoly which sets a price, and in the long run also chooses a service rate. The model differs from the standard theory of monopoly pricing because of the importance of delay. As in many of the industrial markets mentioned above, the queue of existing orders is not directly observable by the customer. We derive the customer demand function, the optimal price for the monopoly to set in the short run with a predetermined service rate, as well as the optimal monopoly price in the long run in which the firm also picks the service rate. The relationship between the market equilibrium and social welfare maximization is analyzed.

Fortunately we find that many of the comparative static effects derived in models without queues continue to hold. In some cases the magnitude of an expected effect is altered. However there are also instances in which an effect derived from the standard timeless model can be drastically altered. This fundamental point does show up in some ways in our analysis. When there is an increase in the number of customers coming to market, in the short run the monopolist will either leave the original price unchanged, or else will cut the price.

The literature on monopoly pricing with queues started with Naor (1969).[2] He demonstrated that when the customers can observe the queue prior to joining, the monopolist will charge a higher price than is socially efficient. Edelson and Hildebrand (1975) showed that when the customer preferences are linear, and they make their purchase decisions without observing the current state of the queue, the monopoly equilibrium price maximizes social welfare. Hassin (1986) showed that when the firm prefers to inform the customers of the queue length, then it is socially optimal to allow the firm to do so. But when the firm prefers not to inform the customers of the length of the queue, social optimality may or may not coincide with the firm's profit maximizing choices.

We allow for preferences that are more general than the linear preferences studied in these papers. This is not just a technical matter, it affects the economic interpretations of the results. Linear preferences are not consistent with customers who do discounting. We show that the equivalence of monopoly pricing and social welfare maximization discovered by Edelson

[2]Cooper (1990) reports that some years ago it was estimated that at that time more than 5,000 academic articles and books had been published relating to queues. Wolff (1989) provides a nice textbook treatment of the mathematics and operations research literature on queues. For helpful reviews of related optimization-based approaches to queueing theory see Stidham (1984) and Stidham and Webber (1993).

and Hildebrand (1975) continues to hold when the customers have linear preferences and the monopoly is able to choose the service rate as well as the price. However we also show that once one moves beyond the linear preference specification that they analyzed, this welfare equivalence no longer holds in general. Unlike the usual over-charging by a monopoly, here the equilibrium may involve either over-charging or under-charging relative to social welfare maximization.

DeVany (1976) is the only previous study of capacity, or service rate choice[3] in a monopoly queueing model. In his study the customers can observe the queue length. As we explain more fully in section 2.3, in comparison to the rest of the literature, there are some quite different features in his formulation of the customer's problem. DeVany (1976) makes some interesting observations about the monopoly equilibrium. First is his suggestion that the monopolist sets price equal to marginal cost. Second, while DeVany (1976) does not explicitly solve a social welfare maximization problem, he suggests that the monopolist chooses too little capacity for social efficiency. We will show that, at least when customers make the decision about joining the queue prior to observing its current length, the results are quite different.

There have been a number of other monopoly pricing and queueing models. Knudsen (1972) allowed for more than a single queue at the firm. Donaldson and Eaton (1981) showed that a monopolist may use a queue to separate out consumers who have different valuations of time. Mendelson and Whang (1990) analyzed the use of priority pricing for different classes of customers.[4] There have also been papers that analyze the formation of queues when the price is exogenously constrained to be below the market clearing level. An interesting example is the study by Deacon and Sonstelie (1985) of queues that arose when, by court order, a price ceiling was placed temporarily on the Chevron gas stations in California.

The rest of the paper is organized in the following manner. We start in section 2 by presenting the analysis of the example in which the customer preferences are linear. We present this first, both because the derivations are simpler in this case, and also because this is the problem that has attracted most attention in the previous literature. In section 3 we study the problem from the perspective of the customers. The demand curve is derived from the customer's optimization problem. In section 4 we analyze the short run situation in which the firm has a predetermined service facility, and the costs of operation are set to zero. Section 5 extends the firm problem to the long run in which the firm also has a service rate choice, and costly production. Social welfare properties of the monopoly equilibrium are presented in section 6. Finally the conclusions are set out in 7.

[3] What DeVany (1976) terms "capacity" we refer to as "service rate". We prefer the term "service rate" because it is more consistent with the time averaging approach adopted both by DeVany (1976) and by the current paper.

[4] Queueing has also been considered in economic settings other than that of a monopoly. Luski (1976), Levhari and Luski (1978), Kalai, Kamien and Rubinovich (1992), and Li and Lee (1994) have begun to integrate the analysis of queues with oligopoly considerations. DeVany and Saving (1983) and Davidson (1988) introduced queueing into competitive models. Mendelson (1985) studied queues that arise within the firm. Larson (1987) provided an intriguing discussion of some of the psychological aspects of queueing.

2 The Example of Customers With Linear Preferences

The logical structure of the problem is quite simple. The firm is a monopoly. The firm move first and selects the price to charge. It is not allowed to charge different customers different prices. The monopoly selects and commits to the price in advance, knowing how that will affect the behavior of the customers.

The customers see the price posted by the firm, and know the rate at which new potential customers arrive in the market, but cannot observe the current state of the queue at the moment that they are considering ordering. Accordingly they can infer and respond to the system average, but not to the actual value. Without that knowledge it is as if the customers are all playing a simultaneous move game. We look for a Nash equilibrium in the customers strategies. In such an equilibrium each customer will be able to calculate the rate at which customers join the queue at the firm. There is no discounting and we work with time averages.[5]

All customers are identical apart from their moment of arrival. They each will demand either nothing, or one unit of the good, from the firm. If the firm's offer is not sufficiently attractive then the customer will not place an order. A customer who does not place an order gets a return of v. Assume that the arrival times of potential customers are given by a Poisson process with rate Λ. The rate at which customers actually place orders is λ, which is to be determined. The information that is known to customer i is: R the reward from getting served by the firm, p the posted price, v the value of their alternative opportunity, and μ the exponential rate of service provided by the firm.

Let i refer to a customer and let s_i refer to his strategy choice. Customer i will select one of two feasible strategies: joining the queue, not joining the queue. We let $s_i = 0$ represent the decision not to join, and $s_i = 1$ represents the decision to join the queue. If customer i joins the queue the actual wait will be w_i, which is a random variable. The cost of waiting is denoted by $C(w_i)$. The customer's utility function $U(\cdot)$ satisfies $U' > 0$ and $U'' \leq 0$, and the customer's cost of delay function $C(\cdot)$ is nondecreasing with $C(0) = 0$. Customer i picks either $s_i = 0$, or $s_i = 1$, in order to maximize the expected utility, V_i where,

$$V_i = \begin{cases} U(v) & \text{if } s_i = 0 \\ U(R - C(w_i) - p) & \text{if } s_i = 1. \end{cases}$$

In some papers analysis is carried out in terms of a "full price" to the customer. The full price consists of the monetary price, plus the cost of waiting. Using our notation $P(\lambda) = p + C(w_i(\lambda))$ is the full price that customer i pays. In our model the firm selects a posted price. The behavior of the customers together with the rate at which orders are precessed, converts the posted price into a full price. The posted price is the firm's strategic variable. The full price is an equilibrium outcome.

At what rate will potential customers be placing orders? Obviously $0 \leq \lambda \leq \Lambda$. To go further requires consideration of the motivation for the customer's behavior. There are three

[5] There is a technical caveat that should be added to all of our derivations. We are supposing that the system has existed long enough that we can work with the stationary distribution.

cases to be considered: all customers pick $s_i = 0$, all customers pick $s_i = 1$, some fraction of the customers pick $s_i = 1$.

First, can it ever be an equilibrium for all of the customers to pick $s_i = 0$? Clearly this is possible, but not especially interesting. To rule this out we assume that at least when there are no other customers and the good is free, the customer will choose $s_i = 1$, i.e., $U(R - C(1/\mu)) > U(v)$, where $1/\mu$ is the mean service time.

Second, can it ever be an equilibrium for all of the customers to pick $s_i = 1$? This will depend on the policy of the firm. If the firm charges a low enough price, services the customers quickly enough, and the firm's product is enough more valuable than the customer's alternative opportunity, then all the potential customers become actual customers of the firm, and $\lambda = \Lambda$. In this case we know that the customer is making choices such that $\boldsymbol{E}U(R - P(\lambda)) \geq U(v)$, where as usual, $\boldsymbol{E}$ is the expectation operator. The expectation is taken with respect to the delay to be endured by the customer if making a purchase. If the customer buys from the firm, he accepts some risk since he does not know how long a wait he faces. When solving the firm's problem, we will find that the firm will never let the customer be strictly better off when buying from the firm. If that were the case then the firm could always raise its profits by raising its price very slightly. Accordingly in equilibrium the inequality will actually be an equality.

Third, suppose that if $\lambda = \Lambda$ then the customers would have a higher payoff from taking the outside opportunity than they expect to get by placing an order. Then at least some of the potential customers will not be joining the queue. The actual rate λ must be such as to cause the consumers to expect equal payoffs from placing an order with the monopolist, or from taking the outside opportunity. In other words, the demand function λ is found by solving

$$\boldsymbol{E}U(R - p - C(w_i(\lambda))) = U(v). \tag{1}$$

It is clear that (1) is a necessary condition for an equilibrium when some but not all customers will be making purchases from the monopolist. In this case we restrict attention to the symmetric equilibrium in which all of the customers randomize. An important feature of the randomized solution is that it preserves the Poisson form for the actual arrival process.

We now suppose that both U and C are linear functions. In this case equation (1) takes the simpler form $R - p - c\boldsymbol{E}w_i(\lambda) = v$. The next task is to determine at what rate customers will place orders with the firm. We are assuming that the arrival process of potential customers is Poisson, that a proportion of them joins the queue, and that there is a single firm that has an exponential service time with rate μ. Given these assumptions, solving for the expected waiting time is straight forward. It is a standard result from the queueing literature, derived for instance in Wolff (1989) section 5-5, that

$$\boldsymbol{E}w_i(\lambda) = \frac{1}{\mu - \lambda} \tag{2}$$

for $\lambda < \mu$, and $\boldsymbol{E}w_i(\lambda) = \infty$ for $\lambda \geq \mu$.

If the arrival rate exceeds the service rate, the expected wait becomes infinite because more and more customers keep on getting added faster than they are being sent away with their

good or service completed. When the service rate exceeds the arrival rate then equation (2) holds. The faster the service, and the fewer the expected number of customers, the shorter the expected wait.

In order for the problem to be of interest it must be the case that at least some potential customers wish to place orders. Accordingly we assume that $R - p - v - \frac{c}{\mu} \geq 0$. If this were false then no customers would place an order with the firm. Recall that $1/\mu$ is the mean service time, and so c/μ is the cost of waiting if the arriving customer does not have to wait for anyone else. When we turn to consideration of the monopoly price determination we will make the parallel assumption that it is feasible for the monopolist to set a price that attracts at least one customer. In other words we will assume that $R - v - \frac{c}{\mu} > 0$.

Consider the situation in which not all potential customers will actually wish to place orders with the firm. Combining (2) with the linear version of (1), we get

$$\lambda_0 = \begin{cases} \mu - \dfrac{c}{R-p-v} & \text{if } \ p < R - v - \dfrac{c}{\mu} \\ 0 & \text{if } \ p \geq R - v - \dfrac{c}{\mu}. \end{cases} \tag{3}$$

This would be the demand curve if the potential arrival rate is sufficiently large. Since the demand rate cannot exceed that potential rate we have

$$\lambda = \min\{\Lambda, \lambda_0\}.$$

It is clear that if $p \geq R - v - c/\mu$, then there is no incentive to place an order. For $p < R - v - c/\mu$, the interpretation of (3) is quite attractive. The faster the service offered by the firm, the more customers it will be able to sell to. The more impatient the customers, the fewer customers will buy from the firm. The more valuable the firm's product, the more customers will buy from the firm. The higher the firm's posted price, the fewer customers the firm will get. The higher the value of the alternative good to the customers, the fewer customers will buy from the firm.

For $p < R - v - c/\mu$, substituting λ into (2) yields

$$\boldsymbol{E}w_i = \begin{cases} \dfrac{R-p-v}{c} & \text{if } \ \Lambda \geq \mu - \dfrac{c}{R-p-v} \\ \dfrac{1}{\mu - \Lambda} & \text{if } \ \Lambda < \mu - \dfrac{c}{R-p-v}. \end{cases} \tag{4}$$

When there is a large pool of potential customers, given a price, the expected waiting time is independent of the service rate of the firm. This observation actually holds for general service time distributions. To see this fact note that when Λ is large, $R - p - c\boldsymbol{E}w_i(\lambda) = v$ from which we can also obtain (4).[6]

[6] Another simple example is an exponential utility function and a linear cost function, $U(x) = -e^{-\gamma x}$ and $C(w) = cw$, where $\gamma > 0$ satisfying $\gamma c < \mu$. Under the same distributional assumptions, we have $EU(R - P(\lambda)) = -\frac{\mu - \lambda}{\mu - \lambda - \gamma c} e^{-\gamma(R-p)}$. Now substituting the above into the customer's decision rule, we get a demand function

Recall that the queue is not directly observable by the customers. The monopolist could choose to inform the customers of their position in the queue. Would it be in the interests of the monopolist to do so?

The first issue is then whether the monopolist would wish to tell the customers the truth. If the monopolist is not constrained to tell the customers the truth, he would be tempted to always tell the customers that they are very close to the head of the queue. If the customers believed such claims they would have a relatively low expected cost of waiting. However, it is not apparent that such claims should be believed. If claims about position in the queue cannot be made credible then it is as if the monopolist is unable to make any claims at all.

To go further we simply suppose that the monopolist has some mechanism to make the claims credible, such as pledging his good name.[7] In that case will he wish to inform the customers of their place in the queue? It seems obvious that the answer is "no". If the customers know their place in the queue, then only those who get positive surplus, or at least zero surplus will place orders. If the customers do not know their place in the queue, then placing an order is like taking a risky gamble. As long as it is at least a fair bet, the customer will take it. After the fact, some of the customers realize positive surplus while others incur losses due to excessive waiting. Since those who incurred losses would not have joined had they known the true situation, it seems that there will be a greater level of demand when the customers are not informed of their position in the queue. This intuition is only part of the story however.

Proposition 2.1 *Let $\rho = \Lambda/\mu$. There exists a critical point ρ^* such that if $\rho < \rho^*$, there will be at least as many purchase orders placed when the customers cannot observe their position in the queue, as when they can observe their position. But this statement is reversed, if $\rho > \rho^*$.*

This proposition is proved in the appendix. The procedure is to derive the demand under each category of customer information, and then to compare the number of customers in the two cases. Neither case can be ruled out as being particularly implausible. Both $\rho > \rho^*$ and $\rho < \rho^*$ are conceivable situations.

What this proposition means is that, as long as the flow of potential customers is low relative to the speed of service being offered by the firm, the intuition suggested above carries

$\lambda = \min\{\Lambda, \mu - \frac{\gamma c}{1 - e^{-\gamma(R-p-v)}}\}$. The demand function has exactly the same economic interpretation as does the demand function with a linear utility function, though the rates at which λ changes in response to the change of R, p, v, c and μ are different. The expected waiting time is $Ew_i = \min\{\frac{1}{\mu-\Lambda}, \frac{1}{\gamma c}(1 - e^{-\gamma(R-p-v)})\}$. Again the economic interpretation is the same as in the linear case.

[7] The issues of credible versus incredible claims, and reputation building are very interesting. However to get into them here would take us well away from the main focus of this paper. In a dynamic setting we know from the supergame literature that one can construct equilibria in which a reputation for honesty can be sustained provided the horizon is infinite and the future is not discounted too much, for instance see the discussion in Tirole (1988). While we can readily construct such examples, we do not think that there would be any further insight to be derived, and so we adopt the simpler approach of Hassin (1986) and simply assume the existence of a commitment technology.

through. However the proposition also tells us that the intuition is incomplete.[8]

When the flow of customers is large relative to the service speed of the firm, the answer is reversed. Why is that? Suppose that the flow of customers is large relative to the service speed offered by the firm. In this case if the customer is not being told his position in the queue, the proportion of potential customers who actually place orders is fairly low. Suppose instead that the firm is correctly informing the customers as to their place in the queue. Now if the queue is currently long and the customer is told this, the customer walks away. But they were probably going to walk away anyhow. Suppose the queue is currently quite short. Now the customers becomes very likely to place orders that they would not otherwise have placed. Bad information (from the customer's perspective) causes little loss of orders, good information causes gains. The timing of these gains is precisely when they are most valuable, when the firm is facing the possibility of under utilizing its facility. If the flow of new customers into the market is high enough relative to the service speed of the firm, then this effect dominates.

The next question is to ask how ρ^* changes when the basic conditions change. The answer has a fairly simple intuition. Anything that makes it more likely that the customer will place an order reduces the need for the monopolist to say anything. This basic intuition is reflected in the following proposition that is proved in the appendix.

Proposition 2.2 *The critical point ρ^* has the following comparative statics. ρ^* increases as R and μ increase, and as c, p, and v decrease.*

Before closing this section we wish to emphasize that these propositions are propositions about the effect of information revelation on demand, for a given price. They are not propositions about profitability of information revelation. From Hassin (1986) we already know that a simple extension to a statement about profitability is not true. The reason for the difficulty in extending the result is that ρ^*, is a function of p.

Following Hassin (1986) we know that for the case of predetermined service rates and customers with linear preferences, if $(R-v)\mu/c \leq 2$, then the firm will always find it more profitable to reveal the queue length. If that condition is not satisfied then there will be a ρ^*, such that if $\rho < \rho^*$ then the firm will find it more profitable not to reveal the queue length, while if $\rho > \rho^*$ then it will be more profitable for the firm to reveal the queue length.

2.1 Short Run Monopoly Pricing

In this section we show how the firm faced with such customers will select a price to charge in the short run. This situation is short run in the sense that the service rate is predetermined and cannot be altered by the monopolist. The firm's problem is now

$$\max_{0 \leq p < R-v-c/\mu} \quad \pi = p \min\{\Lambda, \mu - \frac{c}{R-p-v}\}. \tag{5}$$

[8] A further qualification concerns customer risk aversion. Telling the customers their position in the queue reduces one source of uncertainty that they face. We leave such further complications for future study.

When Λ is great enough, the objective function is concave in $(0, R-v-c/\mu)$, and so from the standard first order condition, we find the optimal pricing for the firm

$$p_m = R - v - (c(R-v)/\mu)^{1/2}. \tag{6}$$

For this solution to make sense we need $p_m \geq 0$ and this happens when $\mu(R-v) \geq c$. We assume that this always holds. If this did not hold, then the customer would be unable to justify spending the necessary expected time in the queue.

When $\Lambda < \mu$, the demand function may take the form $\lambda = \Lambda$ or the form given by equation (3) depending on the price. The maximum price consistent with having all potential customers actually place an order ($\lambda = \Lambda$) is

$$p_\Lambda = R - v - \frac{c}{[\mu - \Lambda]^+}, \tag{7}$$

where $x^+ = \max\{x, 0\}$. The notation $[\mu - \Lambda]^+$ is used here so that the formula can be expressed without explicitly needing to restate the assumption that $\Lambda < \mu$. The price p_Λ is the lowest price that it could ever make sense for the firm to charge. If the firm chooses to set a higher price than that given by (7), then it will give up on serving the entire market. If the firm chooses to only serve some of the potential customers, the maximization is over $(p_\Lambda, R - v - c/\mu)$.

We will refer to p_m as "first order condition pricing", and p_Λ as "market capture pricing". The optimal price is determined by a simple comparison between $\pi(p_m)$ and $\pi(p_\Lambda)$. If the arrival rate of potential customers is low enough, the firm would like to attract even more customers than are coming to market. But that is not possible. In that case it might as well charge as high a price as possible to those customers who do come to market. The highest price that is "possible" is the price that leaves each of the customers just indifferent between placing an order and not doing so. This will be the lowest price that it ever makes sense for the monopolist to charge.[9] Accordingly the optimal price is given by $p^* = \max\{p_m, p_\Lambda\}$.

Substituting the optimal price into the demand function, we can obtain the actual demand in response to the optimal price:

$$\lambda = \min\{\Lambda, \mu - \left(\frac{c\mu}{R-v}\right)^{1/2}\}. \tag{8}$$

The next question is to ask how the price changes as c, μ, R, Λ, and v change. Direct calculation shows

Proposition 2.3 *When first order condition pricing is optimal for the monopolist, changes to the basic conditions in the market produce the following responses.*

$$\frac{\partial p}{\partial c} = -\frac{1}{2}\left(\frac{R-v}{c\mu}\right)^{1/2} < 0, \qquad \frac{\partial p}{\partial R} = 1 - \frac{1}{2}\left(\frac{c}{(R-v)\mu}\right)^{1/2} > 0,$$
$$\frac{\partial p}{\partial v} = -1 + \frac{1}{2}\left(\frac{c}{(R-v)\mu}\right)^{1/2} < 0, \qquad \frac{\partial p}{\partial \mu} = \frac{1}{2\mu}\left(\frac{c(R-v)}{\mu}\right)^{1/2} > 0, \qquad \frac{\partial p}{\partial \Lambda} = 0.$$

[9] Actually one could say that it is that price minus a vanishingly small amount so that all customers strictly prefer to place an order.

When the market capture pricing is optimal for the monopolist, changes to the basic market conditions produce the following responses.

$$\frac{\partial p}{\partial c} = -\frac{1}{\mu - \Lambda} < 0, \qquad \frac{\partial p}{\partial R} = 1 > 0,$$
$$\frac{\partial p}{\partial v} = -1 < 0, \qquad \frac{\partial p}{\partial \mu} = \frac{c}{(\mu - \Lambda)^2} > 0, \qquad \frac{\partial p}{\partial \Lambda} = -\frac{c}{(\mu - \Lambda)^2} < 0.$$

Most of the interpretations seem very natural. Increased cost of waiting causes the firm to cut the posted price. Increased value of the firm's good causes the firm to raise its price. Increased value of the outside opportunity for the customer causes the firm to cut its price. Increased speed of service by the firm allows it to raise the price that it charges. Finally, as long as there are enough customers arriving, the price that the firm charges is independent of the arrival rate. What limits the firm's ability to raise price is the rate at which it can process the customers, and the fact that excessive delay will induce potential customers to avoid becoming actual customers.

Perhaps the most curious of these results is given by $\frac{\partial p}{\partial \Lambda} < 0$. It has following interpretation. Suppose that it is strictly optimal for a firm to sell to all of the potential customers, so that it is still optimal for the firm to sell to all potential customers, if Λ increases by a small amount. The firm would be better off if there were more customers. If the number of potential customers does increase, then the firm will cut the posted price in order to continue to capture all available customers. To continue to get them all the firm must offset the increased cost of waiting that the extra customers impose on each other.

It is natural to ask how sensitive is the result to the details of our model. If the consumers have a quantity choice, as well as a decision of whether or not to join the queue, then the analysis is more complex. There are offsetting effects, and in general the sign of $\frac{\partial p}{\partial \Lambda}$ becomes ambiguous. There is one effect that reflects the attempt to capture more customers, and another effect that reflects the attempt to get more money from each customer. It is then an empirical question whether the increased demand translates into a price fall, no effect on price, or a price rise.

Sometimes Keynesian analysis is characterized as the analysis of situations in which price alone does not clear the market and there may be demand or supply shortages. It is often asserted that at less than "full capacity" any increase in demand translates into increase in production, while at or above full capacity any increase in demand would be translated into price increases. Here we found somewhat different effects. As we have just pointed out it is possible for increased demand to lead to no change in price, or even to induce a price cut. It should be emphasized that this analysis was all carried out for fixed service rate. In the long run we would expect adjustments to the service capabilities to come into play, as they do in section 5. The second caveat is that our analysis is all partial equilibrium. To get into the various implications of general equilibrium analysis would go well beyond the scope of this paper.

2.2 Choosing a Service Rate

In this subsection we turn to the long run problem facing the monopolist. The example is extended to allow the firm to choose the rate of service to offer. We let $q > 0$ be the marginal cost of increasing the speed of service, and $r \geq 0$ is the marginal cost of the actual production. For simplicity we do not have any fixed costs. The optimization problem takes the following form:

$$\begin{array}{ll} \max\limits_{p,\mu} & \pi(p,\mu) = (p-r)\lambda - q\mu \\ \text{s.t.} & \lambda = \min\{\Lambda, \mu - \frac{c}{R-p-v}\}, \\ & r < p < R - v - c/\mu, \\ & \mu \geq 0. \end{array} \tag{9}$$

The objective function (9) is not concave. If the firm is to be viable it is clear that it must cover the physical costs of production and so $p > r$. If the firm is to have any customers, it must be the case that $p < R - v - c/\mu$. Accordingly if the problem is to be of interest it must be the case that $r < R - v - c/\mu$.

The objective function is bounded by

$$\pi(p,\mu) \leq (p - r - q)\mu - \frac{c(p-r)}{R-p-v}. \tag{10}$$

Suppose that $R-v-c/\mu-r \leq q$. The second term on the right-hand-side of (10) is nonnegative. Recalling that $p < R - v - c/\mu$, and using (10) we see that

$$\pi(p,\mu) < ([R - v - c/\mu] - r - q)\mu.$$

If the coefficient on μ is less than or equal to 0, it is optimal to set $\mu = 0$. Operation would yield negative profit. For the rest of this subsection, we will consider more interesting case, $R - v - c/\mu - r > q$.

Whatever the customer arrival rate, the monopolist will select a processing rate to accommodate all of the potential customers. We can distinguish infinite and finite potential customer arrival rates. Suppose that $\Lambda = \infty$. In this case, if we take $p = R - v - c/\mu - \epsilon$, then for $\epsilon > 0$ small enough, $p - r - q > 0$. It is therefore clear that the optimal service rate is $\mu = \infty$.

Next suppose that $\Lambda < \infty$. If the firm chooses to operate and so $\mu > 0$, we claim that $\lambda = \Lambda = \mu - \frac{c}{R-p-v}$ must hold. Why is that? Suppose that this were not true, so that $\lambda = \mu - c/(R-p-v) < \Lambda$. The objective function is exactly the same as the right-hand-side of (10). The optimal p must be larger than $r + q$ as otherwise the objective function becomes negative. Hence it is desirable for μ to be as large as possible. Therefore, the optimal μ must be $\mu - c/(R-p-v) = \Lambda$ as claimed.

With constant returns to scale in the choice of service rate, the marginal cost of selling to a customer is constant. The marginal revenue is also constant up to the point when all the customers are being served, due to the identical unit demand assumption. If the marginal cost exceeds the marginal revenue, the firm is not viable. If the marginal revenue exceeds the

marginal cost, the monopolist would like to sell an infinite amount. If there are an infinite number of customers arriving at each moment, he does so. If there is only a finite arrival rate, then that determines the rate at which the monopolist will choose to sell.

This allows us to simplify the optimization problem to

$$\max_{r<p<R-v-c/\mu} \qquad (p-r)\Lambda - q(\Lambda + \frac{c}{R-p-v}).$$

The objective function of this problem is concave, so using the first order condition, we find the optimal solution. The solution to the monopolist's problem has

$$p^* = R - v - (cq/\Lambda)^{1/2}, \qquad \text{and} \qquad \mu^* = \Lambda + (c\Lambda/q)^{1/2}.$$

The firm will choose to operate if and only if

$$(R - v - q - r)\Lambda - 2(cq\Lambda)^{1/2} > 0. \tag{11}$$

This can be verified by substituting the values for (p^*, μ^*) back into the objective function and checking for nonnegative profits.

While the interpretations of p^* and μ^* seem quite sensible, two elements are worth comment. First, the physical costs of production do not affect the price that the monopolist charges in the long run. Neither does it affect the optimal speed of operation. It does affect whether the firm should choose to operate or not. This feature of the solution is really driven by the assumption that the customers demand either zero or one unit from the monopolist. The same sort of thing happens in monopoly pricing without queues, but with unit demand customers (see exercise 1.2 in Tirole (1988)).

The second aspect that is worth comment concerns the reaction when there is an increase in Λ the number of potential customers coming to market. It causes the firm to add capacity and to raise the price charged. Further, $\mu^* > \Lambda$. That is to say the optimal speed to process the customers is greater than the rate at which potential customers are arriving in the market. The reason for this is that if they were equal and all customers placed orders, the expected wait would grow to infinity. So not all customers would be willing to place order, and so some potential sales would be being lost.

By considering the partial derivatives of the left-hand-side of (11), we know that if a firm chooses to operate, it will still choose to operate, when R and Λ increase, and when v, c and q decrease. When a firm chooses to operate, the optimal service rate increases as Λ and c increase and as q decreases, and the optimal price increases as R and Λ increase and as v, c and q decrease.

2.3 Social Welfare When Customers have Linear Preferences

Social welfare is defined to be the sum of producer and consumer surplus. Given the presence of market power, one might be inclined to expect the equilibrium to be socially inefficient. However Edelson and Hildebrand (1975) derived the surprising result that the monopoly price

is also socially efficient when the customers cannot observe the length of the queue before placing their orders. They call this case the "no balking" case. When the customers can observe the queue length ("balking" permitted) social efficiency is not obtained by monopoly pricing.[10] We illustrate the welfare equivalence result for our model, and then extend it to show equivalence in the choice of service rate.

The consumer surplus per unit of time for a given arrival rate λ and price p is $CS = [\boldsymbol{E}U(R - p - C(w(\lambda))) - U(v)]\lambda$.The producer's surplus in this case is the same as the firm's revenue and so it is given by $PS = p\lambda$. The social welfare is $SW = CS + PS$, or equivalently,

$$SW = [\boldsymbol{E}U(R - p - C(w(\lambda))) - U(v) + p]\lambda \tag{12}$$

The social planner's problem is to maximize SW over p and λ subject to the constraints that $p \geq 0$ and $0 \leq \lambda \leq \Lambda$. To achieve this maximum, it is clear that $\lambda < \mu$ must hold because $C(\infty) = \infty$.

With linear preferences, it is clear from (12) that

$$SW = [R - c\boldsymbol{E}w(\lambda) - v]\lambda = [R - v - \frac{c}{\mu - \lambda}]\lambda,$$

for $\lambda < \mu$ and otherwise, $SW = -\infty$. As one might expect, in this case social welfare does not depend on the price, it only depends on the allocation. The price is a pure transfer. It can be directly verified that SW is a concave function of λ. Hence the social optimal demand rate λ^*, is given by the first-order condition

$$\lambda^* = \min\{\mu - \left(\frac{c\mu}{R - v}\right)^{1/2}, \Lambda\}.$$

The reason for the "min" in this expression is to reflect the presence of the constraint $\lambda \leq \Lambda$. The following result was derived by Edelson and Hildebrand (1975) in following up an earlier equivalence result of Edelson (1971).

Proposition 2.4 *(Edelson and Hildebrand (1975) Equivalence) The social welfare maximizing solution λ^*, is the same as the monopoly solution for λ in (8).*

[10] DeVany (1976) suggests that the monopolist supplies too little output for social efficiency. This seems to be driven by his assumptions concerning the customer's problem. There are some unusual features of the customer's decision making problem in DeVany (1976). The customer arrival rate depends on the posted price for reasons that are never explained. It is apparently not due to the customers' taste for the good, nor is it due to the customer's opportunity costs, since these features are explicitly accounted for elsewhere in the analysis. This extra role of the posted price appears to be double counting, and it does affect the conclusion to be reached about welfare. Perhaps more minor difficulties are: 1. If the arrival rate is to depend on a price, why the posted price rather than the expected full price? 2. When the customer finds an unusually long line at the monopolist, he is supposed to go to another firm with a stochastic queue. The length of the second queue is assumed to be uncorrelated with the first queue despite the fact that the postulated behavior of the customers creates just such a correlation.

Next consider the long run in which the firm and the social planner are both permitted to choose the service rate as well as the price. When the service rate is not predetermined, the firm's profit is $PS = (p - r)\lambda - q\mu$, where $r \geq 0$ is the marginal cost of the actual production and $q > 0$ is the marginal cost of increasing the speed of service. The social welfare in this case is

$$SW = (R - v - r - c/(\mu - \lambda))\lambda - q\mu, \tag{13}$$

which again does not depend on the price. The social planner's problem is to choose λ and μ subject to $0 \leq \lambda \leq \Lambda$ and $\lambda < \mu$, such that SW is maximized. We note that SW is concave in $\mu > \lambda$ although it is not concave jointly in λ and μ. Letting $\partial SW/\partial\mu = 0$ yields

$$\mu = \lambda + (c\lambda/q)^{1/2}. \tag{14}$$

Substituting (14) into (12) yields

$$SW = (R - v - r - q)\lambda - 2(cq\lambda)^{1/2}. \tag{15}$$

The social welfare function SW given by (15) is a convex function of λ and so maximization of (15) subject to $0 \leq \lambda \leq \Lambda$ must be achieved at either $\lambda = 0$ or $\lambda = \Lambda$. By comparing the values of SW at $\lambda = 0$ and $\lambda = \Lambda$, we find that the optimal $\lambda = \Lambda$ if $(R - v - r - q)\Lambda > 2(cq\Lambda)^{1/2}$, and $\lambda = 0$ otherwise. Substituting $\lambda = 0$ into (14) would yield $\mu = 0$, and substituting $\lambda = \Lambda$ would yield $\mu = \Lambda + (c\Lambda/q)^{1/2}$.

Proposition 2.5 *The social welfare maximizing solution is to set $\lambda = \mu = 0$ if $(R - v - r - q)\Lambda \leq 2(cq\Lambda)^{1/2}$, and to set $\lambda = \Lambda$ and $\mu = \Lambda + (c\Lambda/q)^{1/2}$ if $R - v - r - q > 2(cq\Lambda)^{1/2}$. This is the same as the monopoly solution in subsection 2.2.*

This result contrasts sharply with DeVany (1976), the only previous results on monopoly service rate choice that we know of in the literature. In DeVany (1976) the monopoly chooses too little of, what he terms capacity, and what we term service rate. This result shows that the Edelson and Hildebrand (1975) welfare equivalence does directly extend to the long run situation in which the firm has a choice of service rate.

3 The More General Customer Problem

Having worked out the example of linear customer preferences, it is natural to ask, to what extent do these results extend to more general forms of customer preferences? The rest of the paper is directed at answering this question. We will show that the comparative statics are mostly unaffected, but the social welfare results are sensitive to the linearity assumption.

Now consider the more general case of customer demand. We no longer restrict $U(x) = x$. We still assume that $U' > 0$, and, $U'' \leq 0$. It is clear that if $R - p - C(1/\mu) \leq v$, or equivalently, $p \geq R - C(1/\mu) - v$, there is no incentive for a customer to place an order and

hence the demand rate is zero. To avoid this trivial case, through this section we will only consider the case $p < R - C(1/\mu) - v$. Recall that the service time is exponentially distributed with a mean service rate μ. When the service rate μ is strictly greater than the arrival rate λ, the stationary waiting time is exponentially distributed with rate $\mu - \lambda$; otherwise, the expected waiting time is infinite. Since no one would like to wait for an infinite amount of time, it is clear that $\lambda < \mu$ must prevail for the customer's problem. When $\mu > \lambda$, we have

$$u \equiv \boldsymbol{E}U(R - p - C(w_i(\lambda))) = (\mu - \lambda) \int_0^\infty U(R - p - C(x))e^{-(\mu-\lambda)x}dx \tag{16}$$

$$= U(R - p) - \int_0^\infty U'(R - p - C(x))C'(x)e^{-(\mu-\lambda)x}dx, \tag{17}$$

where prime is used to denote a first derivative. Equation (17) is obtained through integration by parts. In the above, we assumed the expectation exists. In particular, we suppose $\lim_{x\to\infty} U(R - p - C(x))e^{-(\mu-\lambda)x} = 0$.

The demand function is given by $\lambda = \Lambda$ if and only if

$$(\mu - \Lambda) \int_0^\infty U(R - p - C(x))e^{-(\mu-\Lambda)x}dx \geq U(v). \tag{18}$$

Inequality (18) implies that $\mu > \Lambda$. When $\mu \leq \Lambda$, not all potential customers will be served in this market. The general form of the demand function is summarized as follows.

Proposition 3.1 *Let $\lambda_0 = \lambda_0(R, p, v, \mu)$ be the unique solution from equation (1) with $\boldsymbol{E}U(R - p - C(w_i(\lambda)))$ given by (16) and (17). Then the demand function is*

$$\lambda = \min\{\Lambda, \lambda_0(R, p, v, \mu)\}.$$

It should be pointed out that if we take $C(w) = cw$, then λ is also a function of c.

Proposition 3.2 *Case 1. Suppose that not all the customers will be choosing to join the queue. Then*

$$\frac{\partial\lambda}{\partial R} = \frac{\partial\lambda_0}{\partial R} > 0, \quad \frac{\partial\lambda}{\partial p} = \frac{\partial\lambda_0}{\partial p} < 0, \quad \frac{\partial\lambda}{\partial v} = \frac{\partial\lambda_0}{\partial v} < 0, \quad \frac{\partial\lambda}{\partial\mu} = \frac{\partial\lambda_0}{\partial\mu} = 1.$$

Case 2. Suppose that all the customers will be choosing to join the queue and that (18) holds with equality. Then increasing μ or R, and decreasing Λ, p, or v work as above.[11]
Case 3. Suppose that all the customers will be choosing to join the queue and that equation (18) is a strict inequality. Then

$$\frac{\partial\lambda}{\partial\Lambda} = 1, \quad \frac{\partial\lambda}{\partial R} = \frac{\partial\lambda}{\partial p} = \frac{\partial\lambda}{\partial v} = \frac{\partial\lambda}{\partial\mu} = 0.$$

[11] In a technical sense these are cases in which the left derivatives are not equal the right derivatives.

For the derivation of this proposition see the appendix. It can also be shown that, λ is increasing and concave in R, and is decreasing and concave in p. Unlike the linear problem we do not have an explicit dependence on c because we do not have an explicit functional form to work with. For many reasonable waiting cost functions, such as $C(x) = cx$, it can be verified in a manner paralleling the above analysis that the higher the waiting cost (the larger the value of c), the lower the demand rate.

The interpretation of this proposition is quite natural. In case 1 not all customers will be joining the queue. We find that the higher the value of service, the higher the demand rate. The higher the price charged for service, the lower the demand rate. The higher the value of the outside opportunity, the lower the demand rate. The higher the rate at which the firm can process orders, the higher the rate at which customers will place orders. Case 2 is just the boundary. Which way the results go depends on which of the other two cases the system moves to. Case 3 is a very favorable market for consumers. The firm could clearly improve profits by raising the price, and so we would not expect such a situation to arise.

4 Privately Optimal Pricing by the Monopolist

Having characterized the demand behavior of the customers, we are now in a position to analyze the decisions of the firm that would like to make a profit by selling to these customers. In this part of the paper we suppose that the firm has been exogenously endowed with a processing rate μ at no cost, and that there are constant costs of production which for simplicity are set equal to zero. This case can be interpreted as short run analysis in which the processing rate is predetermined. In section 5 we analyze the long run in which the monopolist chooses μ.

The firm's problem is

$$\begin{aligned} \max_{p} \quad & \pi(p) = p\lambda \\ \text{s.t.} \quad & p < R - v - C(1/\mu), \\ & p > 0. \end{aligned} \tag{19}$$

Depending on whether inequality (18) holds, either $\lambda = \Lambda$, or else λ is determined by equation (1). Once again we will have the distinction between first order condition pricing and market capture pricing.

First consider the case when inequality (18) does not hold so that λ is given by (1). The first order condition for optimality is

$$\lambda + p\frac{\partial \lambda}{\partial p} = 0. \tag{20}$$

The second order sufficient condition is

$$2\frac{\partial \lambda}{\partial p} + p\frac{\partial^2 \lambda}{\partial p^2} \leq 0. \tag{21}$$

In the appendix we show that the second order condition is satisfied.

Next consider the case when inequality (18) does hold for $p = R - v - C(1/\mu)$. In this case, the firm has the choice of selling to all potential customers or of charging a high enough price that only some fraction of the customers place purchase orders. If the firm chooses to take the whole market ($\lambda = \Lambda$) we again call this market capture pricing. In this case it is clear that the optimal pricing p_Λ is uniquely determined by choosing that value of p that sets condition (18) to be an equality.

If the firm decides not to sell to all potential customers, then λ is given by (1). As a result the optimal price can be determined as in the first case, by the first order condition (20). Let p_m denote this optimal price.

Overall, whether the firm should choose to take the whole market depends on whether $\pi(p_\Lambda) \geq \pi(p_m)$. Given that λ is from (1) the objective function π is concave. Accordingly it increases when $p \leq p_m$ and decreases when $p \geq p_m$. On the other hand, it is clear that the optimal price must be no lower than p_Λ. Recall that at p_Λ, the firm can take the whole market. Therefore, if $p_m \geq p_\Lambda$, then p_m is the overall optimal price, otherwise, p_Λ is optimal since $\pi(p)$ decreases for $p \geq p_\Lambda$. To summarize, the optimal price is $p = \max\{p_m, p_\Lambda\}$.

Having characterized the monopoly price, we now turn to consider how the monopoly price p varies with all other parameters. The comparative statics are broken into two cases depending on whether (18) holds or not.

Proposition 4.1 *When first order condition pricing is optimal for the monopolist, changes to the basic conditions in the market produce the following responses.*

$$\frac{\partial p}{\partial R} > 0, \quad \frac{\partial p}{\partial v} < 0, \quad \frac{\partial p}{\partial \mu} > 0, \quad \frac{\partial p}{\partial \Lambda} = 0.$$

When market capture pricing is optimal for the monopolist, changes to the basic conditions in the market produce the following responses.

$$\frac{\partial p}{\partial R} > 0, \quad \frac{\partial p}{\partial v} < 0, \quad \frac{\partial p}{\partial \mu} > 0, \quad \frac{\partial p}{\partial \Lambda} < 0.$$

Overall we see that the same basic effects arise here as were found in the linear case. When the waiting cost function is $C(x) = cx$, it can be verified similarly that the higher the waiting cost, the lower the price that the firm charges.[12]

5 The Firm's Choice of Service Rate

Having studied the situation in which the service rate is predetermined, we now turn to a consideration of the choice of service rate. There are three types of cost facing the firm. There is a constant marginal cost of speeding up the service rate, $q \geq 0$. There is a fixed cost $F \geq 0$.

[12] We have not presented the case when $p_m = p_\Lambda$. In this case, the left partial derivatives do not agree with the right partial derivatives; some are the same as the first case, while others are the same as the second. As a result it is similar to the discussion in section 3 when (18) holds with an equality.

The cost of production is a constant marginal cost $r \geq 0$. Some discussion of nonlinear cost of speeding up service is included at the end of this section.

Now the firm's problem is

$$\begin{aligned} \max_{p,\mu} \quad & \pi(p,\mu) = (p-r)\lambda - q\mu - F \\ \text{s.t.} \quad & r < p < R - v - C(1/\mu), \\ & \mu \geq 0, \end{aligned} \tag{22}$$

where $\lambda = \min\{\Lambda, \lambda_0\}$ is given by Proposition 3.1. First, note that $\pi(p,\mu) \leq \phi(p,\mu) := (p-r)\lambda_0 - q\mu - F$. Using Proposition 3.2 yields

$$\frac{\partial \phi}{\partial \mu} = p - r - q. \tag{23}$$

Recall the constraint $p < R - v - C(1/\mu)$ in (22). If $R - v - C(1/\mu) - r \leq q$, the partial derivative (23) is less than or equal to zero and so it is optimal to choose $\mu = 0$. Therefore, we will assume that $R - v - C(1/\mu) - r > q$. As in the linear case, it is easy to see that the optimal service rate is $\mu = \infty$, if $\Lambda = \infty$.

Proposition 5.1 *Suppose that $\Lambda < \infty$. If the optimal $\mu > 0$, then $\lambda = \Lambda = \lambda_0$. In other words, if the firm chooses to operate, then it is optimal for the firm to serve the whole market.*

Suppose to the contrary, the optimization problem (22) is equivalent to the one with an objective function ϕ defined above. At the optimum we must have $p - r - q > 0$; otherwise $\phi < 0$ since $\lambda_0 < \mu$ always holds. It follows from (23) that $\partial\phi/\partial\mu > 0$. In other words it is desirable to make μ as large as possible. While increasing μ also increases λ_0, the optimal μ must make $\lambda_0 = \Lambda$.

As a result of this proposition, the optimization problem simplifies to

$$\max_{r<p<R-v-C(1/\mu)} \quad (p-r)\Lambda - q\mu - F,$$

where $\mu = \mu(R,p,v,\Lambda)$ is a function of p; and it is found by solving $\lambda_0(R,p,v,\mu) = \Lambda$. It is shown in the appendix that

$$\frac{\partial^2 \mu}{\partial p^2} = -\frac{\partial^2 \lambda}{\partial p^2} \geq 0. \tag{24}$$

As a result we know that the objective function is concave, and so the first order condition gives the optimal price.

Proposition 5.2 *(1) If the firm chooses to operate, the optimal price p^* is the unique solution to*

$$q\frac{\partial \mu}{\partial p} = \Lambda, \tag{25}$$

and the optimal service rate is given by $\mu^* = \mu(R, p^*, v, \Lambda)$, *where* $p^* = p^*(R, v, \Lambda, q)$ *and* $\mu^* = \mu^*(R, v, \Lambda, q) := \mu(R, p^*(R, v, \Lambda, q), v, \Lambda)$. *The optimal price* p^* *increases as* R *and* Λ *increase and as* v *and* q *decrease. The optimal service rate increases as* Λ *and* v *increase and as* q *decreases, and remains unchanged as* R *varies.*

(2) The firm will choose to operate if and only if

$$(p^*(R, v, \Lambda) - r)\Lambda - q\mu^*(R, v, \Lambda) - F > 0.$$

Increasing R *and* Λ, *and decreasing* r, v, q *and* F *all make it more likely that the firm will choose to operate.*

The proof of the proposition in the appendix. This proposition shows that the major results found in the linear case are not exceptional. The interpretations are natural. In the long run, increases in the value of the good, and in the number of potential customers coming to market both trigger increases in the monopoly price. The more valuable the outside opportunity of the customers, the lower the monopoly price. The more expensive it is to speed up service, the lower the posted price. The more expensive it is to speed up service the slower the service will be. An increase in the rate of potential customers coming to market induces a more than proportionate increase in the service rate the monopolist will choose. This last observation, as with the others are direct generalizations of what was found for the linear customer case.

As in the case of customers with linear preference, the optimal level of the price and the service rate, do not depend on the production cost r. Only the decision whether or not to operate depends on r. We again caution that, this independence is driven by the unit demand customer assumption.

In this section of the paper we have included various costs of operating and have allowed for some generalizations on the demand side of the model. However, there are still further generalizations that are possible but not analyzed here. Three of these in particular can be mentioned: the cost of increasing the speed of processing orders could be nonlinear, the service rate distribution might not be exponential, customers might not have identical unit demands.

The cost of increasing the service rate could be nonlinear. Let $q(\mu)$ be the service rate cost function. If for all $\mu > 0$, $R - v - c/\mu \leq q'(\mu)$, then it is optimal not to operate. If $R - v - c/\mu > q'(\mu)$ for all $\mu > 0$, then the same result as the above holds. A sufficient condition for having this case is that q is concave and $q'(0) < R - v - c/\mu$. The case when $R - v - c/\mu > q'(\mu)$ holds only for some $\mu > 0$ is more complicated.

The next issue concerns the distributional assumption for the service rate. In our analysis we have assumed that the service rate μ is exponential. This is a common assumption to make, but it is not the only possibility. Some of the results in this subsection could be sensitive to this distributional assumption. The reason for this is that in general, we do not always have $\partial\lambda_0/\partial\mu = 1$. We do not pursue this technical issue any further here.

The final issue is the maintained hypothesis of identical unit demand customers. While this type of customer demand is widely used, as we have already pointed out, it is a somewhat special case. It has several advantages for our purposes. It makes the structure of the problem particularly similar to the standard queueing framework. It also allows direct comparison

with Naor (1969) and with Edelson and Hildebrand (1975). In our analysis unit demand customers are convenient, since they permit us to use the number of customers arriving in the market as a direct measure of the potential sales for the monopolist. Furthermore the unit demand assumption is quite a reasonable approximation for some cases, possible examples include purchases of airplanes, cars, pianos or other high value discrete, durable goods. More generally the number of units demanded will depend on the posted price, and perhaps even on the time spent in the queue in some cases.

An important related issue concerns customer heterogeneity. In much of the analysis, the results will carry over if we replace the identical customer, with the mean of an appropriate distribution. However, not all results will necessarily generalize so directly. We leave all these complications for future study.

6 Social Welfare

When analyzing social welfare in the case of customers with linear preferences, we found that monopoly pricing maximized social welfare. To go beyond the case of customers with linear preferences becomes analytically messy. The central question that needs to be answered by such an extension is, whether Edelson and Hildebrand (1975) equivalence continues to hold once we move beyond linear specifications? The answer to this question is, "no". To see this we directly illustrate the point using piecewise linear preference functions. Some expressions for more general preferences are set out in the appendix. They are messy and do not seem to offer further insight.

The central idea needed to understand the welfare properties of monopoly pricing in our context, concerns the presence or absence of consumer surplus in an equilibrium. In any equilibrium for the market the consumer's surplus is zero.

Why will an equilibrium have no consumer surplus? An equilibrium consists of a posted price by the monopoly, and a purchase-no purchase decision by each potential customer. Suppose that there was positive consumer surplus at a candidate equilibrium. If there were more potential customers, then at least one of them (or more rigorously a small positive proportion of them) would choose to join the queue expecting a positive surplus. If there were no more potential customers, then the monopolist could strictly increase his expected profits by raising the posted price by at least some small $\epsilon > 0$. Doing so will cost him no lost sales, and will give him extra revenue on each sale. So for a candidate solution to be an equilibrium there must be no consumer surplus.

While there will be zero consumer surplus in the market equilibrium under our assumptions, the same will not normally be true of the socially optimal solution. Hence in general there is no equivalence between social optimality, and monopoly pricing with customers who queue. We use the tractable case of piecewise linear preferences to make this point.

We now suppose that the customer's utility function takes the form

$$U(x) = \begin{cases} x & \text{if } \ x \leq A \\ A + a(x - A) & \text{if } \ x \geq A, \end{cases} \tag{26}$$

where both a and A are two positive constants, and we assume $a < 1$ and $A < R$. The argument will proceed in three steps. First we will show that social welfare maximization requires that the price be greater than or equal to $R - A$. We then show that the monopoly price must be less than or equal to $R - A$. Finally we show that at the socially optimal price and arrival rate, consumer surplus is negative, so it cannot be a market equilibrium.

Lemma 6.1 *The social welfare maximizing price p^* satisfies $p^* \geq R - A$.*

Why is this true? First of all suppose that $p \geq R - A$, then

$$\boldsymbol{E}U(R - p - cw_i(\lambda)) = R - p - c/(\mu - \lambda).$$

Whereas if $0 \leq p < R - A$, then

$$\boldsymbol{E}U(R - p - cw_i(\lambda)) = (1 - a)A + a(R - p) - \frac{ac}{\mu - \lambda} - \frac{(1 - a)c}{\mu - \lambda} e^{-\frac{(R - A - p)(\mu - \lambda)}{c}}. \tag{27}$$

Notice that $SW = [\boldsymbol{E}U(R - p - cw_i(\lambda)) - U(v) + p]\lambda$. It follows directly that for $0 \leq p \leq R - A$,

$$\frac{\partial SW}{\partial p} = \lambda(1 - a)[1 - e^{-\frac{(R - A - p)(\mu - \lambda)}{c}}] \geq 0.$$

Therefore the socially optimal p^* must be such that $p^* \geq R - A$. When $p \geq R - A$, $SW = [R - c/(\mu - \lambda) + U(v)]\lambda$ evidently does not depend on p. Accordingly the social welfare maximizing p^* can be any $p \geq R - A$.

Next consider the monopolists choice of price. While social welfare consists of both profit and consumer surplus, the monopolist is only concerned with maximizing $p\lambda$

Lemma 6.2 *If $\Lambda \geq \mu$ and*

$$U(v) + (c(R - U(v))/\mu)^{1/2} > A, \tag{28}$$

then the monopoly price p_m must satisfy $p_m \leq R - A$.

Start by noticing that for $p \geq R - A$,

$$\lambda = \mu - \frac{c}{R - U(v) - p}. \tag{29}$$

Since this is the same as the linear utility case, following the linear case, the optimal price for $p \geq R - A$ would be

$$p_0 = R - U(v) - (c(R - U(v))/\mu)^{1/2},$$

provided that $p_0 \geq R - A$. But inequality (28) implies that $p_0 < R - A$. In this case the objective value $p\lambda$ is decreasing for $p \geq R - A$, and accordingly the optimal price for the monopolist must be less than or equal to $R - A$, as claimed.

If the left-partial-derivative of $p\lambda$ with respect to p at $(R-A)$

$$\frac{\partial(p\lambda)}{\partial p}\Big|_{p=(R-A)-} < 0, \tag{30}$$

then the optimal monopolist's price must be strictly less than $R-A$, while the social optimal price is no less than $R-A$. Hence in this case the monopolist's optimal price is not socially optimal. The monopolist is charging a lower price than the social welfare maximizing price.

It can be shown in this case that at the social optimal price $R-A$, the consumer's surplus is negative. In contrast the privately optimal monopolist price is always such that the consumer surplus is zero. While the monopolist would be better off at the social optimal price, the monopolist cannot force the customers to take negative surplus. The basic point is that customers value of one dollar is less than the social value of one dollar when $R-p-cw_i(\lambda)$ is larger than A.

Lemma 6.3 *The inequality (30) holds if* $\Lambda \geq \mu$ *and*

$$\mu[A-U(v)]^2 - c[A-U(v)] + ac(R-A) < 0. \tag{31}$$

To see this, first note that for $p < R-A$ and $\Lambda \geq \mu$, by (1) and (27), λ is determined by the equation

$$(1-a)A + a(R-p) - \frac{ac}{\mu-\lambda} - \frac{(1-a)c}{\mu-\lambda}e^{-\frac{(R-A-p)(\mu-\lambda)}{c}} = U(v). \tag{32}$$

Multiply both sides of the above by $\mu-\lambda$; differentiate the both sides of the resulting equation with respect to p (note that λ is a function of p), and then setting $p = R-A$, we find

$$\frac{\partial\lambda}{\partial p}\Big|_{p=(R-A)-} = \frac{ac}{[A-U(v)]^2}.$$

Thus,

$$\frac{\partial(p\lambda)}{\partial p}\Big|_{p=(R-A)-} = \left[\lambda + p\frac{\partial\lambda}{\partial p}\right]\Big|_{p=(R-A)-} = \mu - \frac{c}{A-U(v)} + \frac{ac(R-A)}{[A-U(v)]^2}.$$

Hence as claimed the inequality (30) holds if and only if the inequality (31) holds.

There is no difficulty finding parameters such that (28), (31), $R > A$ and $\Lambda \geq \mu$ are all satisfied. One example is $R = 5$, $A = 1$, $v = 0$, $\mu = 0.25$, $\Lambda = 0.25$, $c = 0.125$ and $a = 0.5$. We can summarize the over all result in the following manner.

Proposition 6.4 *For nonlinear preference functions such as the piecewise linear example given in (26), the monopoly price choice is not in general equivalent to social welfare maximization.*

We have just worked out an example that has negative consumer surplus at the social optimum but not at the equilibrium. The situation can also be reversed. Consider the utility function of the form, $U(x) = \begin{cases} bx & \text{if } x \leq B \\ bB + (x - B) & \text{if } x \geq B, \end{cases}$ with $b > 1$. In this case it is possible to have consumer surplus that is positive at a social welfare maximum, but again equilibrium will entail zero consumer surplus.

Overall, the monopolist will not in general choose a price that results in social welfare maximization. We see that the monopolist may choose either too low or too high a price compared to the socially efficient choice. This contrasts sharply with the conventional analysis of the welfare properties of monopoly pricing.

7 Conclusions

There is considerable evidence that delay is an important element of market clearing in many markets. To what extent are standard comparative static and social welfare results altered by taking this into account? We have answered this question for the important case of a monopoly whose customers cannot observe the queue at the time that they must decide whether to place an order. We have extended the existing literature in several ways. Our major results are related to comparative statics, service rate selected by the firm, and social welfare of the monopoly equilibrium.

Concerning the comparative statics, we have presented a complete set of results. Perhaps unexpectedly, we found that when there is an increase in the number of customers coming to market, in the short run the monopolist will leave his price unchanged, or else will cut his price. The reason for the price cut is the need to offset the increased waiting time that the extra customers will be placing on the system when the monopolist hopes to capture all the extra potential customers. In the long run the monopolist will speed up the processing rate and raise the price. We have analyzed nonlinear specifications of preferences in order to show, among other things, that none of our comparative static results are sensitive to discounting by the customers.

The monopolist's choice of service rate at a given plant has been analyzed. We obtained the basic result that the monopolist will choose either not to operate, or else will choose to service the entire market. In the long run the monopolist will choose a service rate that strictly exceeds the average arrival rate of potential customers to the market.

Concerning social welfare, we have shown that the Edelson and Hildebrand (1975) result on the equivalence between social welfare maximization and monopoly pricing, depends crucially on their assumption of linear customer preferences. Even with piecewise linear preferences, the equivalence need not hold. Since linearity is crucial for establishing the efficiency of monopoly pricing, discounting may cause problems for the equivalence. However, for the widely analyzed special case of linear preferences, we show that their equivalence result can be extended to a monopoly's choice of the service rate.

We think it needs to be emphasized that there is a fundamental linkage between posted

price, and queueing time. Both impinge directly on the customer. This implies that neither price alone, nor queues alone, clear the market. To ignore their interactions can be quite misleading. We think the empirical evidence on the importance of delay, and the results of the examples presented in this paper, both suggest that further attention to the economics of queueing might be quite fruitful.

8 Appendix

8.1 Customer's Problem

Proof of Proposition 2.1

For simplicity, assume that $h := \frac{(R-p-v)\mu}{c}$ is an integer. Note that $h \geq 1$ must hold; otherwise, customers would never join the queue. Recall that $\rho = \Lambda/\mu$. The demand function is $\lambda_1 = \min\{\Lambda, \mu(1-\frac{1}{h})\}$, when customers cannot observe their position.

Now suppose that customers can observe their position. When an arriving customer finds q customers in the line, her expected waiting time, including her own service time, must be $(q+1)\times(1/\mu)$. Note that $1/\mu$ is the expected service time of one customer. Therefore, she will join the queue if and only if $R - p - \frac{c(q+1)}{\mu} \geq v$, or equivalently, $q \leq h-1$. So the maximum number of customers she can accept in front of her is $h-1$. In other words, if she finds more than $h-1$ customers in the system, she will leave without getting service and therefore get a reward v. What fraction of the customers will get service in this case? When every customer follows the strategy that she joins the queue only when there are at most $h-1$ customers already there (it will be h customers after she joins), this system is an $M/M/1$ queue with a queue limit h. This is discussed in many places including Subsection 5.7 (p.p. 252) of Wolff (1989). Then it is known that the actual rate of those who are served is (equation (69) on page 253 of Wolff (1989))

$$\lambda_2 = \Lambda \frac{1-\rho^h}{1-\rho^{h+1}},$$

where $\rho = \Lambda/\mu$. For $\rho = 1$, we can take limit as $\rho \to 1$ in the above, which gives $\lambda_2 = \Lambda \frac{h}{h+1}$.

If $\lambda_1 > \lambda_2$, then not letting customers observe their position would yield a higher actual demand rate, and if $\lambda_1 < \lambda_2$, we have the reversed case. What remains is to compare the magnitude between λ_1 and λ_2.

Case (i): $\Lambda \leq \mu(1-\frac{1}{h})$: in this case, $\lambda_1 = \Lambda$ and $\rho < 1$. Then

$$\lambda_2 - \lambda_1 = -\Lambda\rho^h \frac{1-\rho}{1-\rho^{h+1}} < 0.$$

Case (ii): $\Lambda > \mu(1-\frac{1}{h})$, or equivalently, $\rho > 1 - \frac{1}{h}$. Let

$$\begin{aligned} f(\rho) := \quad (\lambda_2 - \lambda_1)\tfrac{1}{\mu} = \quad & \frac{\rho - \rho^{h+1}}{1-\rho^{h+1}} - \left(1 - \frac{1}{h}\right) \\ = \quad & \frac{1}{h} - \frac{1-\rho}{1-\rho^{h+1}}. \end{aligned}$$

We show momentarily that f is an increasing function. As verified in Case (i), $f(1-\frac{1}{h}) < 0$ and

$$f(1) = \frac{1}{h} - \frac{1}{h+1} > 0.$$

Therefore, there exists a ρ^*, which is the solution to $f(\rho) = 0$, with

$$1 - \frac{1}{h} < \rho^* < 1 \tag{33}$$

such that $\lambda_2 < \lambda_1$ for $\rho < \rho^*$, $\lambda_2 = \lambda_1$ for $\rho = \rho^*$, and $\lambda_2 > \lambda_1$ for $\rho > \rho^*$.

To complete the analysis, we need to show that $f' \geq 0$. First,

$$f'(\rho) = \frac{1 - \rho^h + h\rho^h(\rho - 1)}{(1 - \rho^{h+1})^2}.$$

Hence, to show $f' \geq 0$, it is sufficient to show that

$$g(\rho) = 1 - \rho^h + h\rho^h(\rho - 1) \geq 0.$$

This is verified by observing that $g(1) = 0$, $g'(\rho) > 0$ for $\rho > 1$, and $g'(\rho) < 0$ for $\rho < 1$.

Finally, we took h to be an integer is just for the convenience of analysis. Otherwise, we take $\lfloor h \rfloor$ (the largest integer that is no greater than h) for the most of the analysis. The result still goes through.

Proof of Proposition 2.2

It follows from the previous proof that ρ^* is the unique solution of

$$1 - \rho^{h+1} = h(1 - \rho)$$

satisfying (33). For simplicity of notation, the superscript "$*$" on the ρ is omitted in this proof. Differentiating on both side of the above equality yields

$$\frac{d\rho}{dh} = \frac{1 - \rho + \rho^{h+1} \log \rho}{h - (h+1)\rho^h}. \tag{34}$$

We need to show that the above derivative is positive. First, we show that the denominator of (34) is positive, by making use of (33):

$$\begin{aligned} h - (h+1)\rho^h &> h - (h+1)\left(1 - \frac{1}{h}\right)^h \\ &= h\left[1 - \left(1 - \frac{1}{h^2}\right)\left(1 - \frac{1}{h}\right)^{h-1}\right] > 0. \end{aligned}$$

Next, it is clear that the inequality

$$\log \rho \geq 1 - \frac{1}{\rho}, \qquad \text{for } 0 < \rho < 1;$$

is sufficient to make the numerator of (34) positive. The inequality can be established by considering $f(\rho) := \log \rho + \frac{1}{\rho} - 1$, which satisfies $f'(\rho) < 0$ for $\rho \in (0, 1)$ and $f(1) = 0$. This shows that the derivative (34) is positive. The proposition is thus proved if we simply recall

that $h = [(R-p-v)\mu]/c$. Obviously h has positive partial derivatives respect to R and μ and has negative partial derivatives with respect to p, v and c.

Proof of Proposition 3.2

We now derive the partial derivatives of λ with respect to R, p, v and μ. Consider the case when $\lambda < \Lambda^*$, or equivalently when inequality (18) does not hold.

$$\frac{\partial u}{\partial \lambda} = -\int_0^\infty U'(R-p-C(x))C'(x)xe^{-(\mu-\lambda)x}dx < 0, \tag{35}$$

$$\frac{\partial u}{\partial R} = (\mu-\lambda)\int_0^\infty U'(R-p-C(x))e^{-(\mu-\lambda)x}dx > 0, \tag{36}$$

$$\frac{\partial u}{\partial p} = -(\mu-\lambda)\int_0^\infty U'(R-p-C(x))e^{-(\mu-\lambda)x}dx < 0, \tag{37}$$

$$\frac{\partial u}{\partial \mu} = \int_0^\infty U'(R-p-C(x))C'(x)xe^{-(\mu-\lambda)x}dx > 0. \tag{38}$$

We used the fact that both the utility function and the cost function are increasing function to obtain the inequality.

Then using the equality (1), we get

$$\frac{\partial \lambda}{\partial R} = -\frac{\partial u}{\partial R}\Big/\frac{\partial u}{\partial \lambda} > 0, \tag{39}$$

$$\frac{\partial \lambda}{\partial p} = -\frac{\partial u}{\partial p}\Big/\frac{\partial u}{\partial \lambda} = -\frac{\partial \lambda}{\partial R} < 0, \tag{40}$$

$$\frac{\partial \lambda}{\partial v} = U'(v)\Big/\frac{\partial u}{\partial \lambda} < 0, \tag{41}$$

$$\frac{\partial \lambda}{\partial \mu} = -\frac{\partial u}{\partial \mu}\Big/\frac{\partial u}{\partial \lambda} = 1. \tag{42}$$

This establishes the first part of proposition (2.3). The remaining parts are straight forward.

8.2 Firm's Problem

The firm's problem is given by (19). In view of (40), the first order condition for this problem is equivalent to

$$\lambda\int_0^\infty U'(R-p-C(x))C'(x)xe^{-(\mu-\lambda)x}dx = p(\mu-\lambda)\int_0^\infty U'(R-p-C(x))e^{-(\mu-\lambda)x}dx. \tag{43}$$

Note that λ is a function of R, p, v and μ. Though tedious, it is a matter of direct verification that there is a unique solution in $(0, R-v)$. The second order sufficient condition given in the body of the paper as (21). How do we know that the second order condition is satisfied? Because of (40), it suffices to show that

$$\frac{\partial^2 \lambda}{\partial p^2} \leq 0. \tag{44}$$

This in turn follows from

$$\frac{\partial^2 \lambda}{\partial p^2} = -\frac{\partial^2 u}{\partial p^2} \Big/ \frac{\partial^2 u}{\partial \lambda^2}$$

and

$$\frac{\partial^2 u}{\partial \lambda^2} = -\int_0^\infty U'(R-p-C(x))C'(x)x^2 e^{-(\mu-\lambda)x}dx < 0, \tag{45}$$

$$\frac{\partial^2 u}{\partial p^2} = (\mu-\lambda)\int_0^\infty U''(R-p-C(x))C'(x)e^{-(\mu-\lambda)}dx \leq 0. \tag{46}$$

We used the fact that U is concave and C is nondecreasing.

Next consider the comparative statics. To see how monopoly price p varies as R, v and μ changes, differentiate both sides of (20). This produces

$$\left(2\frac{\partial \lambda}{\partial p} + p\frac{\partial^2 \lambda}{\partial p^2}\right)dp + \left(\frac{\partial \lambda}{\partial R} + p\frac{\partial^2 \lambda}{\partial p \partial R}\right)dR$$
$$+\left(\frac{\partial \lambda}{\partial v} + p\frac{\partial^2 \lambda}{\partial p \partial v}\right)dv + \left(\frac{\partial \lambda}{\partial \mu} + p\frac{\partial^2 \lambda}{\partial p \partial \mu}\right)d\mu = 0. \tag{47}$$

Note that (45)-(46) and

$$\frac{\partial^2 u}{\partial p \partial R} = -(\mu-\lambda)\int_0^\infty U''(R-p-C(x))e^{-(\mu-\lambda)x}dx \geq 0,$$
$$\frac{\partial^2 u}{\partial p \partial \mu} = -\int_0^\infty U''(R-p-C(x))C'(x)e^{-(\mu-\lambda)x}dx \geq 0,$$

then using the equality (1) yields

$$\frac{\partial^2 \lambda}{\partial p \partial R} = -2\frac{\partial^2 u}{\partial p \partial R} \Big/ \frac{\partial^2 u}{\partial \lambda^2} \geq 0, \tag{48}$$

$$\frac{\partial^2 \lambda}{\partial p \partial \mu} = -2\frac{\partial^2 u}{\partial p \partial \mu} \Big/ \frac{\partial^2 u}{\partial \lambda^2} \geq 0, \tag{49}$$

and using (41) yields

$$\frac{\partial^2 \lambda}{\partial p \partial v} = -U'(v)\frac{\partial^2 u}{\partial \lambda^2}\frac{\partial \lambda}{\partial v} \Big/ \left(\frac{\partial u}{\partial \lambda}\right)^2 < 0. \tag{50}$$

In view of (47) and combining (39)-(42), the second order condition for the firm, and (48)-(50), we have

$$\frac{\partial p}{\partial R} = -\left(\frac{\partial \lambda}{\partial R} + p\frac{\partial^2 \lambda}{\partial p \partial R}\right) \Big/ \left(2\frac{\partial \lambda}{\partial p} + p\frac{\partial^2 \lambda}{\partial^2 p}\right) > 0, \tag{51}$$

$$\frac{\partial p}{\partial v} = -\left(\frac{\partial \lambda}{\partial v} + p\frac{\partial^2 \lambda}{\partial p \partial v}\right) \Big/ \left(2\frac{\partial \lambda}{\partial p} + p\frac{\partial^2 \lambda}{\partial^2 p}\right) < 0, \tag{52}$$

$$\frac{\partial p}{\partial \mu} = -\left(\frac{\partial \lambda}{\partial \mu} + p\frac{\partial^2 \lambda}{\partial p \partial \mu}\right) \Big/ \left(2\frac{\partial \lambda}{\partial p} + p\frac{\partial^2 \lambda}{\partial^2 p}\right) > 0, \tag{53}$$

$$\frac{\partial p}{\partial \Lambda} = 0. \tag{54}$$

The last equality follows from the fact that a small variation of Λ still preserves the violation of inequality (18), and hence, λ is still determined from equation (1) which does not depend on Λ. When the waiting cost function is $C(x) = cx$, it can be verified similarly that the higher the waiting cost, the lower the price that the firm charges.

Finally, consider the case when inequality (18) does hold for $p = R - v$. In this case, the firm has the choice of selling to all potential customers or of charging a high enough price that only some fraction of the customers place purchase orders. If the firm chooses to take the whole market ($\lambda = \Lambda$), then it is clear that the optimal pricing p_Λ is uniquely determined by

$$(\mu - \Lambda)\int_0^\infty U(R - p_\Lambda - C(x))e^{-(\mu-\Lambda)x}dx = U(v). \tag{55}$$

If the firm decides not to sell to all potential customers, then λ is given by (1). As a result the optimal price can be determined as in the first case, by the first order condition (20). Let p_m denote this optimal price.

Overall, whether the firm should choose to take the whole market depends on whether $\pi(p_\Lambda) \geq \pi(p_m)$. Given that λ is from (1) the objective function π is concave. Accordingly it increases when $p \leq p_m$ and decreases when $p \geq p_m$. On the other hand, it is clear that the optimal price must be no lower than p_Λ. Recall that at p_Λ, the firm can take the whole market. Therefore, if $p_m \geq p_\Lambda$, then p_m is the overall optimal price, otherwise, p_Λ is optimal since $\pi(p)$ decreases for $p \geq p_\Lambda$. In short, the optimal price is given by $p = \max\{p_m, p_\Lambda\}$.

Having characterized the monopoly price, we now turn to consider how the monopoly price p varies with all other parameters. If $p_m > p_\lambda$, then (51)-(54) prevails. If $p_m < p_\Lambda$, then the optimal $p = p_\Lambda$ is determined by (55). Hence,

$$\frac{\partial p}{\partial R} = -\frac{\partial u}{\partial R}\Big/\frac{\partial u}{\partial p} > 0, \tag{56}$$

$$\frac{\partial p}{\partial v} = -U'(v)\Big/\frac{\partial u}{\partial p} < 0, \tag{57}$$

$$\frac{\partial p}{\partial \mu} = -\frac{\partial u}{\partial \mu}\Big/\frac{\partial u}{\partial p} > 0, \tag{58}$$

$$\frac{\partial p}{\partial \Lambda} = -\frac{\partial u}{\partial \lambda}\Big/\frac{\partial u}{\partial p} < 0, \tag{59}$$

To obtain these results we use (35)-(38) and the fact that U is nondecreasing. All but the last have the same interpretation of the previous case.[13]

[13] We have not presented the case when $p_m = p_\Lambda$. In this case, the left partial derivatives do not agree with the right partial derivatives; some are the same as (51)-(54), while others are the same as (56)-(59). As a result it is similar to the discussion in section 3 when (18) holds with an equality.

8.3 Choosing a Service Rate

Here we prove claims and propositions in section 5. To this end, we first give a more explicit expression of $\mu = \mu(R, p, v, \Lambda)$. Using equation (16) and Proposition 3.1 we have

$$\lambda_0(R, p, v, \mu) = \mu - f(R, p, v),$$

where f is the unique solution to

$$f \int_0^\infty U(R - p - C(x))e^{-fx}dx = U(v).$$

It is clear that

$$\frac{\partial f}{\partial R} = -\frac{\partial \lambda_0}{\partial R}, \quad \frac{\partial f}{\partial p} = -\frac{\partial \lambda_0}{\partial p}, \quad \frac{\partial f}{\partial v} = -\frac{\partial \lambda_0}{\partial v}. \tag{60}$$

When $\lambda_0(R, p, v, \mu) = \Lambda$, we have $\lambda_0 = \lambda$ and

$$\mu = \mu(R, p, v, \Lambda) = f(R, p, v) + \Lambda. \tag{61}$$

With (60) and (61), it is immediate to verify the equality in (24). Then using inequality (44) gives the inequality in (24).

Next, we turn to the proof of the monotonicity properties in Proposition 5.2. First differentiating both sides of equality (25) yields

$$\frac{\partial \mu}{\partial p}dq + q\frac{\partial^2 \mu}{\partial p \partial R}dR + q\frac{\partial^2 \mu}{\partial p^2}dp + q\frac{\partial^2 \mu}{\partial p \partial v}dv = d\Lambda, \tag{62}$$

where $\partial^2\mu/\partial p\partial\Lambda = 0$ (due to (61)) was used. In view of (60) and (61), using (40), (50) and (44) we have

$$\frac{\partial \mu}{\partial p} = -\frac{\partial \lambda}{\partial p} > 0, \tag{63}$$

$$\frac{\partial^2 \mu}{\partial p \partial v} = -\frac{\partial^2 \lambda}{\partial p \partial v} > 0, \tag{64}$$

$$\frac{\partial^2 \mu}{\partial p \partial R} = -\frac{\partial^2 \lambda}{\partial p \partial R} = \frac{\partial^2 \lambda}{\partial p^2} \leq 0, \tag{65}$$

$$\frac{\partial^2 \mu}{\partial p^2} = -\frac{\partial^2 \lambda}{\partial p^2} \geq 0, \tag{66}$$

where we used (40) to obtain the second equality in (65). Noting (62) and using (63)-(66) yields

$$\frac{\partial p}{\partial R} = -\frac{\partial^2 \mu}{\partial p \partial R} \Big/ \frac{\partial^2 \mu}{\partial p^2} = 1, \tag{67}$$

$$\frac{\partial p}{\partial v} = -\frac{\partial^2 \mu}{\partial p \partial v} \Big/ \frac{\partial^2 \mu}{\partial p^2} < 0, \tag{68}$$

$$\frac{\partial p}{\partial \Lambda} = 1 \Big/ [q \frac{\partial^2 \mu}{\partial p^2}] > 0, \tag{69}$$

$$\frac{\partial p}{\partial q} = -\frac{\partial \mu}{\partial p} \Big/ \frac{\partial^2 \mu}{\partial p^2} < 0; \tag{70}$$

this verifies that the optimal price p increases as R and Λ increases and as v and q decreases.

Note that the optimal rate is determined by $\mu^* = \Lambda + f_0(R, p^*(R, v, \Lambda, q), v, \Lambda)$; hence, in view of (60), (40), and (67)-(70),

$$\frac{\partial \mu}{\partial R} = \frac{\partial f_0}{\partial R} + \frac{\partial f_0}{\partial p}\frac{\partial p}{\partial R} = 0, \tag{71}$$

$$\frac{\partial \mu}{\partial \Lambda} = 1 + \frac{\partial f_0}{\partial p}\frac{\partial p}{\partial \Lambda} > 1. \tag{72}$$

$$\frac{\partial \mu}{\partial q} = \frac{\partial f_0}{\partial p}\frac{\partial p}{\partial q} < 0; \tag{73}$$

this verifies that the optimal rate increases as Λ increases and as q decreases. Notice that (72) is not only positive, but is actually greater than one.

To see that the optimal rate increases as v increases, we only need to show that $\partial \mu / \partial v \geq 0$, which follows from

$$\begin{aligned}
\frac{\partial \mu}{\partial v} &= -[\frac{\partial \lambda}{\partial v} + \frac{\partial \lambda}{\partial p}\frac{\partial p}{\partial v}] \\
&= -[\frac{\partial^2 \lambda}{\partial p^2}\frac{\partial \lambda}{\partial v} - \frac{\partial^2 \lambda}{\partial p \partial v}\frac{\partial \lambda}{\partial p}] \Big/ \frac{\partial^2 \lambda}{\partial p^2} \\
&= \left(\frac{\partial \lambda}{\partial v}\right)^2 \frac{\partial^2 u}{\partial p^2} \Big/ [U'(v)\frac{\partial^2 \lambda}{\partial p^2}] \geq 0,
\end{aligned}$$

where we used (68) to obtain the second equality, used (45) and (44) to obtain the inequality, and used (40) and (41) to get the equality

$$\frac{\partial \lambda}{\partial p} \Big/ \frac{\partial \lambda}{\partial v} = -\frac{\partial u}{\partial p} \Big/ U'(v)$$

and then to differentiate on the both side of the above with respect to p to obtain the third equality. It should be noted that the derivative may be equal to zero, i.e., the optimal rate may remain unchanged as v varies, which is what happens in the linear customer model.

Finally, we consider the sufficient condition in the proposition as it responses to the change of parameters. It is sufficient to show that $g := (p - r)\Lambda - q\mu - F$ is non-decreasing as R and Λ increase and as r and q decreases. As proved above, p is increasing in R and μ remains unchanged as R varies; hence, g is increases in R. It is clear that g is decreasing in r and q

(for the latter, noticing that p is decreasing in q). As for Λ,

$$\begin{aligned}\frac{\partial g}{\partial \Lambda} &= \Lambda \frac{\partial p}{\partial \Lambda} + p - r - q\frac{\partial \mu}{\partial \Lambda} \\ &= p - r - q + [\Lambda - \frac{\partial \lambda}{\partial p}]\frac{\partial p}{\partial \Lambda} > 0,\end{aligned}$$

where we used (72) to obtain the second equality, and used (40), (69) and the fact that $p - r - q \geq 0$ to obtain the last inequality.

8.4 Social Welfare General Case

For general consumer preferences, the social welfare problem becomes analytically messy. The main insight to be derived is that the Edelson and Hildebrand (1975) equivalence does not hold. We illustrated this point in the text by means of a piecewise linear example. Here we record the general expressions merely for completeness.

To begin with, we rewrite the consumer surplus $CS = [\boldsymbol{E}U(R - p - C(w(\lambda))) - U(v)]\lambda$ in view of (16) and (17)

$$CS = \left[(\mu - \lambda)\int_0^\infty U(R - p - C(x))e^{-(\mu-\lambda)x}dx - U(v)\right]\lambda \tag{74}$$

$$= \left[U(R-p) - U(v) - \int_0^\infty U'(R - p - C(x))C'(x)e^{-(\mu-\lambda)x}dx\right]\lambda. \tag{75}$$

Hence, the social welfare (12) can be rewritten as

$$SW = \left[(\mu - \lambda)\int_0^\infty U(R - p - C(x))e^{-(\mu-\lambda)x}dx - U(v) + p\right]\lambda \tag{76}$$

$$= \left[U(R-p) - U(v) - \int_0^\infty U'(R - p - C(x))C'(x)e^{-(\mu-\lambda)x}dx + p\right]\lambda. \tag{77}$$

The social planner's problem is to maximize SW over p and λ subject to the constraints that $p \geq 0$ and $0 \leq \lambda \leq \Lambda$. To achieve this maximum, it is clear that $\lambda < \mu$ must hold because $C(\infty) = \infty$.

In this case, the first-order conditions are

$$\begin{aligned}\frac{\partial SW}{\partial \lambda} &= U(R-p) - U(v) - \int_0^\infty U'(R - p - C(x))C'(x)e^{-(\mu-\lambda)x}dx + p \\ &\quad - \int_0^\infty U'(R - p - C(x))C'(x)xe^{-(\mu-\lambda)x}dx = 0, \end{aligned} \tag{78}$$

$$\frac{\partial SW}{\partial p} = \left[1 - (\mu - \lambda)\int_0^\infty U'(R - p - C(x))e^{-(\mu-\lambda)x}dx\right]\lambda = 0, \tag{79}$$

where we used (77) for (78), and (76) for (79).

From equations (78) and (79), we can in principle solve for p and λ. If the second order conditions are satisfied, then the solution pair $p^* = p^*(R, v, \mu)$ and $\lambda^* = \lambda^*(R, v, \mu)$ would

be the optimal solution for the social welfare problem, provided that $p^* \geq 0$ and $\lambda^* \leq \Lambda$. Note that the optimal monopolist's price is determined by equation (20). It seems clear that equation (20) is in general different from equations (78) and (79); therefore, in general, the monopolist's optimal price is not socially optimal.

Finally, using (78) in (75), we can obtain the optimal consumer surplus:

$$CS^* = \lambda^* \left[\int_0^\infty U'(R - p^* - C(x))C'(x)xe^{-(\mu-\lambda^*)}dx - p^* \right]. \tag{80}$$

Thus, the optimal social welfare is

$$SW^* = \lambda^* \int_0^\infty U'(R - p^* - C(x))C'(x)xe^{-(\mu-\lambda^*)}dx. \tag{81}$$

References

[1] Carlton, D. and Perloff, J., 1994, *Modern Industrial Organization, Second Edition*, Harper Collins, New York.

[2] Cooper, R. B., 1990, "Queueing Theory", chapter 10 in D. P. Heyman and M. J. Sobel, eds., *Handbooks in OR and MS, Vol. 2*, North-Holland, Amsterdam.

[3] Davidson, C., 1988, "Equilibrium in Service Industries: An Economic Application of Queuing Theory," *Journal of Business* 61, 347-367.

[4] Deacon, R., and Sonstelie, J., 1985, "Rationing by Waiting and the Value of Time: Results from a Natural Experiment," *Journal of Political Economy*, 93, 627-647.

[5] DeVany, A., 1976, "Uncertainty, Waiting Time, and Capacity Utilization: A Stochastic Theory of Product Quality, " *Journal of Political Economy* 84, 523-541.

[6] DeVany, A., and Saving, T., 1983, "The Economics of Quality" *Journal of Political Economy* 91, 979-1000.

[7] Donaldson, D., and Eaton, B. C., 1981, "Patience More than its own Reward: A Note on Price Discrimination," *Canadian Journal of Economics*, 14, 93-105.

[8] Edelson, N. M., 1971, "Congestion Tolls Under Monopoly," *American Economic Review*, 59, 873-882.

[9] Edelson, N. M., and Hildebrand, D. K., 1975, "Congestion Tolls for Queuing Processes," *Econometrica*, 43, 81-92.

[10] Hassin, R., 1986, "Consumer Information in Markets with Random Product Quality: The Case of Queues and Balking," *Econometrica* 54, 1185-1195.

[11] Kalai, E., Kamien, M., and Rubinovich, M., 1992, "Optimal Service Speeds in a Competitive Environment" *Management Science* 38, 1154-1163.

[12] Knudsen, N. C., 1972, "Individual and Social Optimization in a Multiserver Queue with a General Cost-Benefit Structure," *Econometrica*, 40, 515-528.

[13] Larson, R. C., 1987, "Perspectives on Queues: Social Justice and the Psychology of Queueing", *Operations Research* **35**, 895-905.

[14] Li, L. and Lee, Y. S., 1994, "Pricing and Delivery-time Performance in a Competitive Environment," *Management Science* 40, 5, 633-646.

[15] Levhari, D., and Luski, I., 1978, "Duopoly Pricing and Waiting Lines," *European Economic Review*, 11, 17-35.

[16] Luski, I., 1976, "On Partial Equilibrium in a Queuing System with Two Servers," *Review of Economic Studies* 43, 519-525.

[17] Mendelson, H., 1985, "Pricing Computer Services: Queueing Effects," *Communications of the ACM* 28, 312-321.

[18] Mendelson, H., and Whang, S., 1990, "Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue," *Operations Research* 38, 870-883.

[19] Naor, P., 1969, "The Regulation of Queue Size by Levying Tolls," *Econometrica* 37, 15-24.

[20] Stidham, S., 1985, "Optimal Control of Admission to a Queueing System," *IEEE Transactions on Automatic Control*, AC-30, 8, 705-713.

[21] Stidham, S., and Weber, R., 1993, "A Survey of Markov Decision Models for Control of Networks of Queues," *Queueing Systems*, 13, 291-314.

[22] Tirole, J., 1988, *The Theory of Industrial Organization*, The MIT Press, Cambridge Massachusetts.

[23] Wolff, R. W., 1989, *Stochastic Modelling and the Theory of Queues*, Prentice-Hall, Engelwood, New Jersey.