Service Pricing with Loss-Averse Customers

Liu Yang

School of Economics and Management, Tsinghua University, China, yangliu@sem.tsinghua.edu.cn

Pengfei Guo, Yulan Wang

Faculty of Business, The Hong Kong Polytechnic University, Hong Kong, pengfei.guo@polyu.edu.hk, yulan.wang@polyu.edu.hk

We consider a service system in which customers are loss averse toward both price and delay attributes. That is, customers compare these two attributes to their rational expectations of outcomes, with losses being more painful than equal-sized gains are pleasant. We first study customers' equilibrium queueing strategies. We find that unlike the traditional case in which loss aversion is not considered, there may exist three equilibrium strategies, one of which is preferred in the sense that customers' utility is highest at this equilibrium. We then investigate the optimal pricing problem for a monopoly server and find that loss aversion polarizes queues, making long queues even longer and short queues even shorter. Furthermore, loss aversion toward the delay attribute drives the optimal price down, whereas loss aversion toward the price attribute drives it up. We also find that profit- and welfare-maximizing prices are not the same in a monopoly market. Finally, we consider pricing competition in a symmetric duopoly market and find that the conclusions depend on the size of the service capacity relative to the market size. For fast servers, there exists a unique symmetric price equilibrium. Under certain conditions, the effect of loss aversion on waiting time drives the price down, whereas that on the monetary term drives it up. For moderate-speed servers, there also exists a unique symmetric equilibrium. However, the effect of loss aversion on the two attributes works in reverse compared with that in the fast server case. For slow servers, we show that a symmetric equilibrium may not exist, and we numerically find that there may exist two asymmetric equilibria. Interestingly, with loss-averse customers, a firm can obtain a higher profit in a duopoly market than in a monopoly market.

Key words: service pricing, loss-averse customers, strategic queueing behavior, service competition

1. Introduction

Customers who seek service often assess two attributes of potential service systems: the net monetary reward (service reward−price) and the waiting time. A mental accounting of these two attributes determines whether the customer should join the service system or balk. In the classic literature (Hassin and Haviv 2003), customers are often assumed to be fully rational, in the sense that a simple utility function (the net monetary reward minus the cost of waiting) is used to represent the mental accounting process. However, this modeling approach may overlook features of customers' mental accounting activities in a queueing game.

According to the literature on service quality and customer satisfaction, a customer's satisfaction (utility) with a service is greatly affected by her expectations of the service attributes, with higher expectations often leading to lower utility and vice versa (Parasuraman et al. 1985, Anderson and Sullivan 1993, Lin et al. 2008). Such expectations are known as *reference points* and such customers are said to be *reference dependent*. This is also consistent with prospect theory (Kahneman and Tversky 1979), which demonstrates that people generally view a final outcome as a gain or a loss with respect to a certain reference point, which in our context is customer expectation. Nowadays, the abundance and ease of accessibility of information online allows customers to naturally form expectations that reflect real system performance before seeking a service. For example, the Ministry of Health of the province of Ontario posts emergency room, surgery and diagnostic imaging waiting time information on its government website (`http://www.health.gov.on.ca/en/public/programs/waittimes/`).

Loss aversion reflects customer behavior in the sense that compared with reference points on price and delay, losses are more painful than equal-sized gains are pleasant (Kahneman and Tversky 1979). There is a substantial body of anecdotal and empirical evidence of customer loss aversion on price and service quality attributes. For example, Lin et al. (2008) empirically study passengers' loss-aversion effects in a transportation system. They show that compared with passengers' service quality expectations, which generally include ticket fare, waiting time and travel time, service quality loss influences passengers' repurchase intentions more than service quality gain.

There is a difference between customers' mental accounting of money and time. According to Abdellaoui and Kemel (2014), there are three main differences between time and money. First, time is less fungible than money. Second, time cannot be easily saved or stored and hence its aggregation is difficult. Third, people are less capable of accounting for time than they are for money. These differences imply that people can have different degrees of loss aversion toward money and time attributes. Using a lab test, Abdellaoui and Kemel (2014) confirm that loss aversion exists in both money and time attributes, but to different degrees. Tereyagoglu et al. (2015) use ticket sales data from a theater and find that customers are reference dependent on both price and capacity sold (which indicates the congestion level of seating zones) in making their purchasing decisions. They show that customers are loss averse on both attributes and that the degree of loss aversion differs between these two attributes. Given these empirical findings, the monetary rewards and time attributes in a customer's utility must be considered separately, using different loss-aversion parameter values.

To explicitly consider the foregoing three issues, we adopt prospect theory by assuming that customers are reference dependent on their expectations regarding two attributes: price and waiting time. Under this assumption, we study customers' equilibrium queueing strategies and firms' pricing decisions to answer the following questions. How do reference price and waiting time affect a loss-averse customer's joining behavior? What is the optimal pricing strategy for a monopoly server when customers are loss averse toward both price and waiting time? Do the social planner and monopoly server price differently when customers are loss averse? When the servers are engaged in a duopoly price competition, how does customer loss aversion behavior affect their pricing decisions and system performance? We provide detailed answers to these questions in the remainder of the Introduction.

Following Koszegi and Rabin (2006), we posit that a customer's overall utility has two components: *intrinsic utility* and *gain-loss utility*. Intrinsic utility measures the direct effects of service attributes. Gain-loss utility is derived from a customer comparing the net monetary reward (service reward−price) and waiting time outcomes to her expectations of those outcomes and losses

being more painful than equal-sized gains are pleasant. A customer's reference points for the net reward and delay are fully endogenized as her *rational expectations* of the outcomes of these two attributes (Koszegi and Rabin 2006). Due to the stochastic property of waiting time, we further assume the expectations to be stochastic. This assumption is supported by the experimental study conducted by Rust et al. (1999). They show that a customer's service quality expectation is a probability density function that describes the relative likelihood of a particular quality outcome. In our context, this implies that the delay a customer expects to experience is stochastic and that a customer derives her gain-loss utility from comparing the real waiting time to her expectation.

Applying the concept of *personal equilibrium* proposed by Koszegi and Rabin (2006), we further define customer equilibrium in our queueing game as follows. Based on her reference points, a customer derives the gain-loss utility, which along with her intrinsic utility determines her optimal choice (whether to join or to balk). The customer's choice, along with the choices of all other customers, generates the outcomes of the waiting time and net reward, which are consistent with the customer's rational expectations (i.e., reference points).

We begin by conducting an equilibrium analysis of customers' queueing strategies. In sharp contrast to the uniqueness of the equilibrium in cases where loss aversion is not considered, we find that multiple equilibria are typical when customers are loss averse. Specifically, deviating from a customer's original plan probably causes a loss (see Koszegi and Rabin (2006)). If a customer had planned to join a service, she would feel a loss of service value if she did not join and her aversion to this loss would lead her to join the service even when the expected waiting time were long. However, if she had planned to balk, joining the service would bring her a loss in terms of waiting cost and her aversion to this loss would lead her not to join. We find that there are at most three equilibrium strategies, only one of which is the *preferred personal equilibrium*, which yields the largest expected utility for customers of all of the equilibria.

Next, we consider a service firm's optimal pricing decision in a monopoly market. Interestingly, we show that under optimal pricing, customers are either more or less likely to join the queue

than in the traditional case in which customer loss aversion is not considered. Specifically, if the traditional joining probability is high, meaning that the service reward is relatively high compared with the marginal cost of waiting, then loss aversion results in an even higher joining probability, as customers care more about the loss of the service reward than the gain of not having to wait. However, if the traditional joining probability is low, meaning that the marginal cost of waiting is relatively high compared with the service reward, then loss aversion results in an even lower joining probability, as customers try to avoid waiting. In short, loss aversion *polarizes* queues such that long queues become even longer and short queues even shorter.[1] We further show that both optimal price and joining probability decrease when customers are more loss averse toward waiting time and increase when they are more loss averse toward the net service reward.

We then study the welfare maximization problem and compare its optimal price to that of the profit maximization problem. It is well known that in an unobservable queue, without loss aversion the welfare- and profit-maximizing prices are equivalent when customers are identical (Edelson and Hildebrand (1975)). However, this conclusion no longer holds in our case, as the gain-loss utility is included in the welfare function and thus social welfare is always smaller than profit. We identify the sufficient conditions under which the profit-maximizing joining probability is larger than the socially desired joining probability.

Finally, we examine the price competition with loss-averse customers in a duopoly market. Compared with the monopoly case, customers in this setting have an expanded choice set from two options (balking and joining) to three options (balking, joining server 1 or joining server 2). We consider a symmetric setting in which the service rates of two firms are equal. We find that the equilibrium structure hinges on the size of the service capacity, specifically, whether the capacity is ample, moderate or scarce. We call servers in the corresponding cases the fast, moderate-speed and slow servers, respectively.

For fast servers, there exists a unique symmetric price equilibrium under which customers all join and equally split between two servers. Moreover, the customers obtain a higher utility from joining

than from balking. We then obtain two striking results. First, when the service reward is high, loss aversion on the monetary term drives the price down, whereas that toward waiting time drives the price up. The equilibrium price here can be higher than that when loss aversion is not considered, as we consider loss aversion in terms of two attributes: price and waiting time. When the degree of loss aversion on waiting time is high, a server tries to increase its price to reduce congestion. Unfortunately, the other server does the same and eventually congestion is not reduced, but the price is increased at equilibrium. Higher equilibrium prices along with an unchanged congestion level imply that loss aversion related to waiting time may harm customers, contradicting the results of many studies indicating that reference effects and loss aversion generally benefit customers (Heidhues and Koszegi 2008, Yang et al. 2014). The second striking result is that introducing a competitor does not necessarily harm the firm. This is in sharp contrast to the classic literature, which indicates that a firm's profit in a duopoly competition market cannot be higher than its monopoly profit. The reason is that, introducing an additional firm allows more customers to be served, which affects a customer's reference point and makes the balking option less attractive. The strengthened joining incentives from customers allow the two servers to increase their prices, resulting in a possibly higher profit for each server under competition than what they would obtain as monopolies.

For moderate-speed servers, there still exists a unique symmetric price equilibrium under which customers all join and equally split between two servers. However, the customers obtain the same utility from joining as they do from balking. Unexpectedly, the effect of loss aversion on the two attributes now works in reverse compared with the fast server case. That is, the effect of loss aversion on waiting time decreases the price, whereas that on the monetary term increases the price. Again, we find that competition can benefit a server compared with the monopoly case.

For slow servers, we show that there could exist a continuum of symmetric equilibria when the service rate is below a threshold. However, when the service rate is above this threshold, a symmetric equilibrium may not exist. The possible non-existence of symmetric equilibrium is

caused by a server's kinked profit function at the kink point where its price equals that of its competitor. When the profit function is kinked, a symmetric equilibrium requires the left-hand derivative of the server's profit function with respect to its price to be non-negative and the corresponding right-hand derivative to be non-positive at the kink point. However, with balking customers and loss-averse customer preferences, increasing the server's price in the region right above its competitor's price may be more beneficial than doing so in the region right below its competitor's price. In other words, the right-hand derivative may be larger than the left-hand derivative (and hence the requirements regarding the signs of the left- and right-hand derivatives at the kink point may not be satisfied simultaneously). Note that increasing a server's price in the region right above its competitor's price implies that the server is charging a price higher than that of its competitor. Hence, this server is positioned "closer" to the balking option than its competitor in terms of both the monetary reward and waiting time attributes. Consequently, this server together with the balking option have more influence on the customer's reference points than the competitor. Thus, customer loss aversion further mitigates the server's demand loss from the increased price. We note that Luski (1976) shows that the existence of balking customers and the heterogeneity of customer delay sensitivity drive the two servers to offer heterogeneous service products. Our result here further demonstrates that the existence of balking customers and loss aversion preferences can also drive the two servers to do so.

The remainder of this paper is structured as follows. In Section 2, we review the related literature. In Section 3, we consider a monopoly server case and study both customers' equilibrium queueing strategies and servers' optimal pricing decisions. In Section 4, we further examine the duopoly case and derive the equilibrium prices of the two servers. In Section 5, we provide concluding remarks. All the proofs are relegated to the online Appendix.

2. Literature Review

This work is closely related to three streams of literature: customers' loss aversion behavior in behavioral economics, reference effects in operations management (more specifically, service management) and customer queueing strategies and optimal firm decisions in queueing systems.

Customer loss aversion behavior was first modeled in prospect theory, as proposed by Kahneman and Tversky (1979). Koszegi and Rabin (2006) recently developed a model to consider the rational-expectation-based reference points and the effect of loss aversion on firm pricing strategies. We use this rational expectation based framework to consider customer loss aversion behavior in a service setting. Although we along with Koszegi and Rabin (2006) consider loss aversion in two dimensions, these two dimensions are different. In Koszegi and Rabin (2006), the two dimensions are the product and price. The two dimensions we consider are the monetary reward and delay. This is due to the different contexts. The stochastic outcome in Koszegi and Rabin (2006) is caused by the multiple prices a customer may face, whereas the stochastic outcome in our model comes from two sources: the probability choice over queueing options and the randomness of waiting time.

Since Koszegi and Rabin (2006), several studies have examined pricing issues with loss-averse customers. Heidhues and Koszegi (2008) examine price competition when customers base their references on their recent expectations of a product. They show the existence of a focal price equilibrium in the presence of loss aversion. Karle and Peitz (2012) also examine firm price competition, but only in the context of a proportion of customers being reference dependent. They show that a larger proportion of fully rational customers leads to a more competitive outcome. Baron et al. (2015) investigate a situation in which a newsvendor sells to strategic customers who have stochastic reference points toward both price and product availability. They find that expected price and firm profit may increase with the level of customer loss aversion. These studies consider pricing strategy in a traditional goods market rather than in a service operational setting. Lindsey (2011) studies congestion pricing given reference-dependent preferences in a transportation setting, yet we consider neither the congestion in a queueing framework nor the competitive pricing issue.

Our work contributes to the recent stream of research on reference effects in operations management (Popescu and Wu 2007, Nasiry and Popescu 2011, Roels and Su 2014, Chen et al. 2014, Baron et al. 2015). Popescu and Wu (2007) consider a dynamic pricing setting wherein the demand is a general reference-dependent function. They show that a firm's optimal pricing path either

decreases or increases in the presence of a reference effect. Nasiry and Popescu (2011) examine the same setting in which customers anchor to the lowest and most recent prices. They reveal that a range of constant pricing policies is optimal. Both of these works consider the reference effect in a traditional goods setting in which customers are reference dependent on a single price attribute and the reference price is a single point. In contrast, we consider a service operational setting wherein customers' reference points on the two attributes are stochastic, as information becomes more accessible and abundant. Roels and Su (2014) show that the reference point can be engineered by a social planner to achieve her objectives when individuals are behind averse and ahead seeking. The social planner can use the full reference distribution of outputs or an aggregate reference point.

To the best of our knowledge, the work of Yang et al. (2014) is alone in examining customer loss aversion in the context of service competition. They consider duopoly competition over waiting time when customers are reference dependent on the delay. Our work differs from theirs in three aspects. First, in their model, customers are reference dependent only on the delay attribute, whereas in our model they are reference dependent on both the price and delay attributes. Second, in their model, a reference point is determined by the waiting-time decisions of two firms, whereas we model a customer's reference point as her lagged rational expectations of the delay and price outcomes. Third, in their model, servers directly compete over delay by choosing their service rates, whereas in our model servers compete over prices.

Our work is also closely related to the literature that examines customers' queueing strategies and firms' optimal decisions in a service system; see Hassin and Haviv (2003) for a comprehensive review. One stream of literature focuses on fully rational customers (Cachon and Harker 2002, Chen and Wan 2003, Chen and Frank 2004, Ho and Zheng 2004, Allon and Federgruen 2007, Debo et al. 2012, Afèche et al. 2013). For instance, Afèche et al. (2013) study revenue-maximizing tariffs that depend on realized lead times when customers are risk averse regarding uncertain lead times. In contrast, we consider customer loss aversion behavior driven by waiting time uncertainty. Chen

and Frank (2004) consider firms' pricing decisions in a monopoly service setting and Chen and Wan (2003) examine price competition in a duopoly setting. In the absence of reference effects and loss aversion, our monopoly and duopoly models are reduced to those in the preceding two studies, respectively. By comparing our results with those obtained using traditional models, we show the effects of loss aversion on customers' equilibrium queueing strategies and firms' pricing decisions. Another emerging stream of literature considers boundedly rational customers, assuming that customers choose servers following a multinomial logit choice model (Shang and Liu 2011, Huang et al. 2013, Aksoy-Pierson et al. 2013, Li et al. 2016). Unlike the aforementioned work, we assume that customers are loss averse when they form preferences.

3. The Monopoly Case

In this section, we consider the monopoly case in which there is one server in the market. The customers, who are assumed to be identical, arrive to the system according to a Poisson process with rate λ, which is normalized to 1. The queue is unobservable and the customers decide to join or balk upon arrival based on the system's long-run statistics. The service times are i.i.d. and exponentially distributed with rate μ.

In the following subsections, we discuss the reference-dependent utility for loss-averse customers. We then consider customer equilibrium queueing strategies. Finally, we study profit- and welfare-maximizing prices.

3.1. Reference-dependent Utility

A customer who decides to join a service facility obtains a service reward, R, but must pay a price P and suffers a realized waiting cost. The customer's waiting cost is assumed to be proportional to her *sojourn* time, which is the sum of her waiting time in the queue and her service time. Hereafter, we refer to the sojourn time as the waiting time. A customer's expected waiting cost is θw, where w is the expected waiting time and θ is the unit-time waiting cost parameter. Hence, this customer's *intrinsic utility* is $R - P - \theta w$. Balking leads to an intrinsic utility of 0.

A customer's overall expected utility is the sum of her *intrinsic utility* and *gain-loss utility*, where the latter represents the additional utility the customer derives from comparing the actual outcomes to the reference points. We assume that service reward R and price P are deterministic. Hence, it is the waiting time uncertainty rather than the price uncertainty that a loss-averse customer faces in our system. In addition, as a customer always receives both the reward and price or neither depending on whether she joins the queue or balks, we can combine the service reward and price into a single attribute, $R-P$, referred to as the monetary reward. This is different from Koszegi and Rabin (2006) and Lindsey (2011), who consider random prices and treat price P as a different attribute than service reward R. In our setting, a customer's intrinsic utility consists of two attributes, net monetary reward $R-P$ and waiting time, which are on different psychological dimensions (Abdellaoui and Kemel 2014). We assume that the customer forms reference points for each of them.

Following Koszegi and Rabin (2006), we posit that a customer's reference points for the net reward and delay are fully endogenized as her rational expectations for the outcomes of these two attributes, which are consistent with real outcomes. The popularity of websites, such as Yelp.com, and social media platforms, such as Facebook, makes this assumption quite reasonable. In addition to her own past experiences, a customer can easily obtain other customers' experiences with a service system's waiting times online before joining the service. Hence, we assume that a customer knows the waiting time distribution of the server or, equivalently, the joining probability of other customers (denoted as δ).

Consider a tagged customer who plans to join a system with probability δ_p and expects to receive a net reward of $R-P$ and experience a random waiting time of W. Given that all other customers join the system with probability δ, the random waiting time W that the tagged customer expects to experience is an exponential random variable with a mean of $w=1/(\mu-\delta)$ (i.e., $W\sim\exp(\mu-\delta)$). The tagged customer plans to balk with probability $1-\delta_p$ and expects to receive 0 net reward and experience 0 waiting time. These net reward and delay outcome expectations are naturally treated

as her reference points. Thus, the reference point for the net monetary reward, denoted as $\hat{V}$, is a random variable defined as follows:

$$\hat{V} = \begin{cases} R - P, \text{ with probability } \delta_p; \\ 0, \qquad \text{with probability } 1 - \delta_p. \end{cases}$$

The reference point of the waiting time is also a random variable, denoted as $\hat{W}$, and is defined by the following expression:

$$\hat{W} = \begin{cases} W, \text{ with probability } \delta_p; \\ 0, \quad \text{with probability } 1 - \delta_p. \end{cases}$$

Let $F(t)$ be the cumulative distribution function (CDF) of W. The CDF of $\hat{W}$ can then be written as follows:

$$G_{\hat{W}}(r) = (1 - \delta_p) + \delta_p F(r), \quad r > 0. \tag{1}$$

Based on the net monetary reward and delay reference points, the customer then derives her gain-loss utility, which along with intrinsic utility further determines her optimal choice.

The overall expected utility for the customer to join is expressed as follows:

$$\begin{aligned} U(\text{Join}) = & \underbrace{R - P - \theta \mathrm{E}[W]}_{\text{Intrinsic utility}} + \underbrace{\mathrm{E}[(R - P - \hat{V})^+ + \alpha(R - P - \hat{V})^-]}_{\text{Gain and loss due to payment}} \\ & \underbrace{-\theta \mathrm{E}[\beta(W - \hat{W})^+ + (W - \hat{W})^-]}_{\text{Loss and gain due to waiting time}}, \end{aligned}$$

where $x^+ = \max\{0, x\}$ and $x^- = \min\{0, x\}$. The parameters $\alpha(\geq 1)$ and $\beta(\geq 1)$ measure the degree of customer loss aversion related to net reward and waiting, respectively. When they are greater than 1, customers are loss averse and care more about a loss than about an equal-sized gain. Empirical studies (Abdellaoui and Kemel 2014, Tereyagoglu et al. 2015) have shown that customers have different degrees of loss aversion toward the net reward and waiting time. We thus use different loss aversion parameters. Let t be the realized waiting time. Thus, the distribution of t is the same

as the distribution of the waiting time the customer plans to experience if she joins the queue. That is, $F(t) = F(r)$. Substituting (1) into the overall expected utility expression yields the following:

$$\begin{aligned} U(\text{Join}) &= R - P - \theta w + (1-\delta_p)(R-P) \\ &\quad + \theta \int_0^\infty \left(\beta(1-\delta_p)(-t) + \beta\delta_p \int_0^t (r-t)dF(r) + \delta_p \int_t^\infty (r-t)dF(r) \right) dF(t) \\ &= (2-\delta_p)\left(R - P - \theta w \frac{1+\beta}{2} \right). \end{aligned} \tag{2}$$

The detailed calculation of this utility function is listed in online Appendix B. The joining utility expression contains only parameter β. It does not contain parameter α, as the customer only experiences a loss due to waiting when she joins the service.

Although the customer obtains an intrinsic utility of 0 when she balks, her overall expected utility from balking is

$$U(\text{Balk}) = \underbrace{\mathrm{E}[\alpha(0-\hat{V})^-]}_{\text{Loss due to losing the net reward}} \quad \underbrace{-\theta \mathrm{E}[(0-\hat{W})^-]}_{\text{Gain due to avoiding waiting}},$$

where the first term measures the loss due to losing the net reward, $R-P$, and is thus multiplied by parameter α, and the second term measures the gain due to the avoidance of waiting and is thus not multiplied by parameter β. Again, by using (1), the balking utility is

$$U(\text{Balk}) = -\delta_p(\alpha(R-P) - \theta w). \tag{3}$$

In sharp contrast to traditional studies, the utility from balking can be a nonzero value, as balking generates mixed feelings for customers, such as losses from losing the service value and gains from avoiding waiting.

3.2. Customer Equilibrium

A customer makes her joining/balking decision by comparing her overall expected utility from joining the service with that from balking. Denote δ_o as the optimal joining probability, then $\delta_o = 1(0, \text{respectively})$ if $U(\text{Join})$ is larger (smaller, respectively) than $U(\text{Balk})$. Otherwise, when

$U(\text{Join})$ equals $U(\text{Balk})$, δ_o can be any value between 0 and 1. To derive the customer equilibrium, we consider the following difference:

$$U(\text{Join}) - U(\text{Balk}) = (2 - \delta_p + \delta_p\alpha)(R - P) - \theta w(1 + \beta - \delta_p(\beta - 1)/2), \tag{4}$$

where $w = 1/(\mu - \delta)$ (recall that total arrival rate $\lambda = 1$).

In the classical queueing game without considering the reference effect, the customer equilibrium requires a tagged customer's joining action to be the best response to the other customers' joining strategies. As customers are identical, the queueing game has indistinguishable infinitely many players. As is often the case in the queueing game literature, we consider only the symmetric equilibrium in which customers adopt the same joining strategy (Hassin and Haviv 2003). In our case, it is required that $\delta_o = \delta$. Moreover, after considering the reference effect, the concept of customer equilibrium has another layer of requirement. That is, a customer's reference points, $\hat{V}$ and $\hat{W}$, must be consistent with the net reward and the delay outcomes generated by her optimal strategy, δ_o. In other words, the customer's joining plan is consistent with her realized optimal choice; that is, $\delta_p = \delta_o$. This is defined as *personal equilibrium* in Koszegi and Rabin (2006). To conclude, customer equilibrium in our setting requires that $\delta_o = \delta_p = \delta$. Thus, in the following equilibrium analysis, we keep notation δ and get rid of the other two δs. The utility difference in (4) then becomes the following:

$$U(\text{Join}) - U(\text{Balk}) := g(\delta) = (2 - \delta + \delta\alpha)(R - P) - \theta w(1 + \beta - \delta(\beta - 1)/2). \tag{5}$$

There exist three types of equilibrium. If $g(0) \le 0$, then "all balk" is a pure-strategy equilibrium. If $g(1) \ge 0$, then "all join" is a pure-strategy equilibrium. If the customers adopt a mixed strategy (i.e., $0 < \delta < 1$), then the equilibrium joining probability must solve the equation $g(\delta) = 0$. Note that in this case, we have $U(\text{Join}) = U(\text{Balk}) < 0$. When customers are loss neutral in terms of the two attributes (i.e., $\alpha = \beta = 1$), the equation $g(\delta) = 0$ is reduced to $R - P - \theta w = 0$ (the equilibrium equation for the traditional case without considering reference effects; see §1.1 of Hassin and Haviv (2003)), suggesting that loss-neutral customers behave in the same way as reference-independent

customers, in which case the equilibrium is unique. However, in general, the uniqueness of the equilibrium no longer holds.

To find the number of equilibria for the general case, we first provide the following structural property of the function $g(\delta)$.

LEMMA 1. *If $\mu < 2(\beta+1)/(\beta-1)$, the function $g(\delta)$ is concave in δ. Otherwise, $g(\delta)$ is increasing in δ.*

Lemma 1 shows that the function $g(\delta) = 0$ has at most two solutions, implying that there are at most three equilibria for the entire game.; see online Appendix A for the details of the equilibria. Figure 1 illustrates such a three-equilibria case.

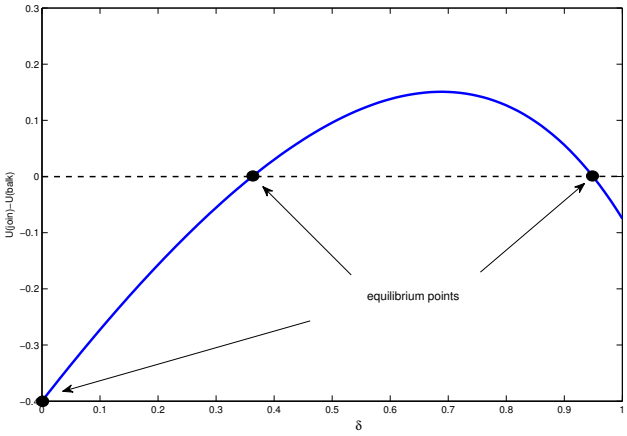

Figure 1 Illustration of a three-equilibria case: $R = 2, P = 0.45, \theta = 2, \mu = 2, \alpha = 2.5, \beta = 2.5$.

In the classic $M/M/1$ model without considering loss aversion behavior, as more customers intend to join the queue, the system becomes more congested, which reduces a tagged customer's tendency to join. Such avoid-the-crowd behavior assures that at most one equilibrium exists in the customer queueing game; see §1.6 of Hassin and Haviv (2003). Why do we have multiple equilibria here? The driving force for our multiple equilibria lies in the rational-expectation-based reference effect and customer loss aversion behavior. That is, deviating from a customer's original plan probably causes her a loss (Koszegi and Rabin (2006)). Here, customers treat the outcomes of their plans as

reference points. If a customer had planned to join the service, she would feel a loss of service value if she did not join and her aversion to this loss would lead her to join the service even when the expected waiting time were long. However, if she had planned to balk, joining the service would bring her a loss in terms of waiting cost, and her aversion to this loss would lead her not to join. Indeed, in the presence of loss aversion behavior (i.e., $\alpha, \beta > 1$), the utility difference between joining and balking increases in her planned joining probability, δ_p, as shown in (4). This implies that a customer's actual joining incentive becomes stronger as her planned joining probability increases. The existence of multiple equilibria is also found in other settings with rational-expectation-based customer loss aversion behavior (Koszegi and Rabin 2006, Heidhues and Koszegi 2008, Karle and Peitz 2012, Baron et al. 2015).

Remark 1. Consider the three-equilibrium case depicted in Figure 1. Between the two mixed-strategy equilibria represented by the two $g(\delta) = 0$ solutions, the smaller one must be locally unstable, as a slight decrease in δ would result in $U(\text{Join}) < U(\text{Balk})$ and thus more customers would balk, making δ even smaller. A slight increase in δ would result in $U(\text{Join}) > U(\text{Balk})$ and attract more customers to join. In contrast, the largest and smallest of the three equilibria are both locally stable. Between these two equilibria, we choose the one that generates a larger expected utility, known as the *preferred personal equilibrium* (PPE) (p. 1144 of Koszegi and Rabin (2006)). Among the three equilibria, the smallest yields zero utility for customers, whereas the medium-sized and largest both yield a negative utility. Thus, the one with a joining probability of 0 is the PPE in this example.

Although there may be multiple equilibria, there exists a unique PPE for the entire game, as stated in the following proposition. First, let

$$\overline{\mu} = \frac{\alpha+1}{\alpha - 2/(\beta+1)},\ \underline{P} = R - \theta\frac{\beta+3}{2(\alpha+1)(\mu-1)} \text{ and } \overline{P} = R - \theta\frac{\beta+1}{2\mu}.$$

Proposition 1. *When the server is a monopoly, there exists a unique PPE for the queueing game. Specifically, the PPE, denoted as* δ^{PPE}*, takes the following form:*

(1) When the service capacity is so small, such that $\mu \leq 1$*, then*

(a) if $P < \overline{P}$*, the PPE is a mixed-strategy equilibrium defined by the unique positive solution of* $g(\delta) = 0$*.*

(b) if $P \geq \overline{P}$*, "all balk" is the PPE (i.e.,* $\delta^{PPE} = 0$*).*

(2) When the service capacity is moderate, such that $1 < \mu < \overline{\mu}$*, then*

(a) if $P \leq \underline{P}$*, "all join" is the PPE (i.e.,* $\delta^{PPE} = 1$*).*

(b) if $\underline{P} < P < \overline{P}$*, the PPE is a mixed-strategy equilibrium defined by the unique positive solution of* $g(\delta) = 0$*.*

(c) if $P \geq \overline{P}$*, "all balk" is the PPE (i.e.,* $\delta^{PPE} = 0$*).*

(3) When the service capacity is ample, such that $\mu \geq \overline{\mu}$*, then*

(a) if $P \leq \overline{P}$*, "all join" is the PPE (i.e.,* $\delta^{PPE} = 1$*).*

(b) if $P > \overline{P}$*, "all balk" is the PPE (i.e.,* $\delta^{PPE} = 0$*).*

REMARK 2. In the limiting case with $\alpha = 1$ and $\beta = 1$, the two bounds for the price are $\underline{P} = R - \theta \frac{1}{\mu - 1}$ and $\overline{P} = R - \theta \frac{1}{\mu}$ and the threshold for the capacity is $\overline{\mu} = +\infty$. For this loss-neutral customer case, case (3) in Proposition 1 disappears and cases (1) and (2) define the customer equilibrium, consistent with the results in the literature (p. 46 of Hassin and Haviv (2003)). Case (3) is particularly interesting, as it shows that when the service capacity is larger than $\bar{\mu}$, the purchasing behavior of loss-averse customers is similar to that seen in a traditional goods market. That is, there exists a cut-off threshold regarding price P below which everyone buys and above which nobody buys. This is because that when the service capacity is ample, the effect of loss aversion on the net reward dominates its effect on the delay.

When the PPE is a mixed-strategy equilibrium (i.e., $\delta^{\text{PPE}} \in (0, 1)$), the joining probability δ^{PPE} decreases in price. To see this, note that the derivative of function $g(\delta)$ with respect to service price P is negative. Hence, as P increases, the function $g(\delta)$ decreases and its larger root becomes smaller. This suggests that δ^{PPE} is smaller, as it is essentially the larger root of the equation $g(\delta) = 0$, as shown in the proof of Proposition 1. Similarly, as α(respectively, β) increases, function $g(\delta)$ increases (respectively, decreases), leading to a larger (respectively, smaller) δ^{PPE}.

3.3. Profit and Welfare Maximization

In this section, we study the optimal pricing problem given that customers make their queueing decisions according to their reference-dependent utility. When there exist multiple equilibria in the customer queueing game, we consider the PPE. In the following, we first consider the decision maker to be a profit-maximizing service provider and then a welfare-maximizing social planner.

The profit-maximizing service provider chooses a price to maximize its expected profit, Π, as follows.

$$\max_{P} \quad \Pi = \delta^{\mathrm{PPE}} P.$$

According to Proposition 1, when $\overline{P} \leq 0$, the equilibrium joining strategy is $\delta^{\mathrm{PPE}} = 0$. Thus, we assume $\overline{P} > 0$ to avoid the trivial case in which the server receives zero demand. This condition can be rewritten as $R > \theta(\beta+1)/(2\mu)$. Note that customer loss aversion ($\beta > 1$) reduces the server's profitability.

As shown in Proposition 1, when $\mu \geq \overline{\mu}$, the profit-maximizing price approaches $\overline{P}$ from the left and the corresponding equilibrium joining probability $\delta^* = 1$. We thus focus on the case of $\mu < \overline{\mu}$. From Proposition 1, we know that for every price, P, there exists a unique positive δ^{PPE}. We begin by investigating the structural property of the profit function when $\mu < \overline{\mu}$ and $P < \overline{P}$, especially when a mixed-strategy equilibrium is involved, corresponding to cases 1(a) and 2(b) of Proposition 1. As a mixed-strategy equilibrium joining probability solves $g(\delta) = 0$, we can replace P in the profit function with the following function of δ:

$$P = R - \theta A(\delta) w,$$

where

$$A(\delta) = \frac{\beta + 1 - \delta(\beta-1)/2}{(\alpha-1)\delta + 2}$$

represents the adjusted marginal cost of waiting in the presence of reference effects. We can then consider δ as the decision variable instead of P.

Lemma 2. *The profit function,* Π*, is concave in* δ*, when* $\mu < \overline{\mu}$ *and* $P < \overline{P}$.

According to Lemma 2, the first-order condition (FOC) is sufficient to derive the optimal price and corresponding equilibrium joining probability. The equilibrium joining probability is determined by the following FOC:

$$\frac{d\Pi}{d\delta} = R - \theta w^2 \mu \left[A(\delta) + \delta \left(1 - \frac{\delta}{\mu} \right) \frac{dA(\delta)}{d\delta} \right] = 0. \tag{6}$$

Note that when customers are loss neutral, (6) is reduced to the traditional FOC without considering the reference effect; that is, $R - \theta w^2 \mu = 0$. We next investigate the effect of customer loss aversion on the optimal price and equilibrium joining probability.

PROPOSITION 2. *In a monopoly setting, both the optimal price, P^*, and the corresponding equilibrium joining probability, δ^*, increase with α and decrease with β.*

Recall that δ^{PPE} decreases with P, as stated in Remark 2. Hence, one may believe that the effect of the loss aversion parameter, α or β, on the optimal price, P^*, is simply the opposite of that on δ^*. However, Proposition 2 shows that this is not the case. When customers are more loss averse toward waiting time (i.e., when β increases), they are less likely to join the service. Thus, to obtain more profit, the server must lower its price to attract more customers. If customers are more loss averse toward the net monetary reward (i.e., if α increases), they are more likely to join the service to avoid losing the monetary reward. The server is then entitled to charge a higher price.

We now compare the optimal price, P^*, and the corresponding equilibrium joining probability, δ^*, to those in the traditional case without considering customer loss aversion. Note that in the traditional setting, if the market size is sufficiently large compared with the service capacity, μ, the equilibrium joining probability, δ^{trad}, solves the following equation:

$$d\Pi^{trad}/d\delta = R - \theta w^2 \mu = 0.$$

PROPOSITION 3. *Let $\hat{\delta} = \sup\{\delta : A(\delta) + \delta(1 - \delta/\mu) dA(\delta)/d\delta \geq 1\}$. The comparison result between δ^* and δ^{trad} depends on whether δ^{trad} is larger or smaller than $\hat{\delta}$. More specifically,*

- *if* $\delta^{trad} \le \hat{\delta}$, $\delta^* \le \delta^{trad}$;
- *if* $\delta^{trad} > \hat{\delta}$, $\delta^* > \delta^{trad}$.

In addition, the threshold, $\hat{\delta}$*, decreases with* α *but increases with* β*.*

If the equilibrium joining probability in the traditional case is high ($\delta^{trad} > \hat{\delta}$), then the net monetary reward is relatively high compared with the marginal cost of waiting. Loss aversion behavior results in an even higher joining probability. The mentality is, "Given that others are suffering a long wait, I will not bother to wait longer." This reference effect mitigates customers' feelings of loss, which helps them tolerate a longer waiting time in crowded environments, such as border-crossing systems. However, if the equilibrium joining probability in the traditional case is low ($\delta^{trad} \le \hat{\delta}$), the marginal cost of waiting is relatively high compared with the net service reward. Loss aversion behavior results in an even lower joining probability, as customers try to avoid loss from the delay. Consequently, customers' loss aversion behavior polarizes the queue, making short queues even shorter and long queues even longer. Polarized queue length has been widely observed in practice (e.g., see examples in Bikhchandani et al. (1992)). Managers should be aware of this polarizing effect, as it can otherwise cause misunderstandings in terms of customers' patience levels. For example, for a manager who wants to estimate the customer delay sensitivity parameter, θ, ignoring the polarizing effect can lead to underestimation of this parameter for heavy-traffic systems and overestimation for light-traffic systems.

We now consider the social-welfare-maximizing problem, which maximizes the sum of the aggregate reference-dependent utilities of all customers (including balking customers) and the server's profit. Note that customers' gain-loss utilities are included in the welfare function, which "respects the principle of consumer sovereignty" (Lindsey (2011)).

The social planner makes its pricing decision to maximize social welfare, SW, as follows:

$$\max_{P} \quad \mathrm{SW} = \delta(2-\delta)\left(R - P - \theta w \frac{1+\beta}{2}\right) + (1-\delta)[-\alpha\delta(R-P) + \delta\theta w] + P\delta.$$

The difference between social welfare, SW, and the server's profit, Π, is expressed as follows:

$$SW - \Pi = \frac{w\theta\delta\,(\delta-2)}{2}\frac{(\alpha(\beta+1)-2)}{(\alpha-1)\delta+2}.$$

If $\alpha = \beta = 1$, meaning that customers are loss neutral (loss and gain are weighted equally), the aforementioned difference is reduced to 0. Profit and social welfare maximization yield the same pricing decision. A similar result is identified by Edelson and Hildebrand (1975), who consider identical customers with intrinsic utilities. However, if either loss aversion parameter α or β is greater than 1, social welfare is always smaller than the servers' profit, as loss-averse customers place more weight on losses than on gains and servers fail to internalize customers' excess loss due to loss aversion. In this case, the profit-maximizing price differs from the welfare-maximizing price.

Denoting the welfare-maximizing joining probability as δ^s, we obtain the following result.

PROPOSITION 4. *In a monopoly setting, if either (1) $\mu \leq 2$ or (2) $\mu > 2$ and $\alpha < \frac{1}{1-2/\mu}$, the profit-maximizing joining probability is larger than the socially desired joining probability (i.e., $\delta^* > \delta^s$).*

As stated, with customer loss aversion, the server cannot absorb the entire customer surplus. Thus, when customers do not care much about losing the service reward (i.e., α is small), the social planner charges a higher price than the profit-maximizing server, which results in a lower joining probability. Doing this reduces not only the joining customers' congestion-induced utility loss, but also the unhappiness of the balking customers. If neither of the two conditions in Proposition 4 holds, the relationship between δ^* and δ^s cannot be determined analytically.

4. The Duopoly Case

In the foregoing section, we study the pricing problem in a monopoly setting in which customers have two options, to join a queue or to balk. We now further examine the pricing problem in a duopoly setting with two symmetric servers that have the same service rate, μ. No matter which server customers join, they receive the same service reward, R, and incur the same unit waiting cost, θ. We specifically consider the following two-stage game: in the first stage, server i determines its price, P_i, to maximize its revenue, $i = 1, 2$, and commits to it. In the second stage, arriving customers, who are assumed to be identical and to have full knowledge of the prices of the two servers, choose from the following three options: joining server 1, joining server 2 or balking. The subgame perfect Nash equilibrium for this aforementioned two-stage game can be solved

via backward induction. First, given a price pair, (P_1, P_2), we derive the corresponding customer equilibrium queueing strategy. Then, anticipating the second-stage customer equilibrium queueing strategy, we investigate the price competition game between the two servers.

4.1. Customer Equilibrium in Parallel Queues

In this subsection, we study the second-stage queueing game among customers given a price pair, (P_1, P_2). Similar to the monopoly setting, customers arrive according to a Poisson process with rate $\lambda = 1$. We posit that before taking any joining or balking actions, a customer knows the market shares of servers 1 and 2 (δ_1 and δ_2), that is, the joining probability of other customers. Hence, the customer knows the waiting time distributions of the two servers if she chooses to join one of them. In addition, we assume a customer plans to join server i with probability δ_{pi} and expects to receive a net reward and experience a delay according to her plan, which are naturally treated as her reference points.

We thus assume that a tagged customer expects to choose server i with probability δ_{pi}, pays price P_i, experiences random waiting time W_i and expects to balk with probability $1-\delta_{p1}-\delta_{p2}$. The reference point for the net monetary reward, $\hat{V}$, is thus a random variable defined as follows:

$$\hat{V} = \begin{cases} R - P_1, \text{ with probability } \delta_{p1}; \\ R - P_2, \text{ with probability } \delta_{p2}; \\ 0, \qquad \text{with probability } 1-\delta_{p1}-\delta_{p2}. \end{cases}$$

The reference point for the delay, denoted as $\hat{W}$, is a random variable defined as follows:

$$\hat{W} = \begin{cases} W_1, \text{ with probability } \delta_{p1}; \\ W_2, \text{ with probability } \delta_{p2}; \\ 0, \quad \text{with probability } 1-\delta_{p1}-\delta_{p2}, \end{cases}$$

where W_i is a random variable following an exponential distribution with a rate of $\mu_i - \delta_i$; that is, $W_i \sim \exp(\mu_i - \delta_i)$. Let $F_i(t)$ be the CDF of W_i. The CDF of $\hat{W}$, denoted as $H_{\hat{W}}(r)$, can thus be written as follows:

$$H_{\hat{W}}(r) = (1-\delta_{p1}-\delta_{p2}) + \delta_{p1}F_1(r) + \delta_{p2}F_2(r), \quad r > 0. \tag{7}$$

Based on the reference points for the net monetary reward and waiting time, the tagged customer derives her gain-loss utility, which along with her intrinsic utility further determines her optimal choice.

We first consider the case in which $P_1 \geq P_2$. Let w_i, $i = 1, 2$, denote the expected waiting time of server i. The tagged customer's utility from choosing server 1 can be derived as follows:

$$U_1 = \underbrace{(R - P_1 - \theta w_1)}_{\text{Intrinsic utility}} \quad \underbrace{-\alpha\delta_{p2}(P_1 - P_2)}_{\text{Loss from paying a higher price than } P_2} \quad \underbrace{+(R - P_1)(1 - \delta_{p1} - \delta_{p2})}_{\text{Gain from obtaining the service}}$$
$$+ \underbrace{\theta \int_0^\infty \left[\beta \int_0^t (r - t) dH_{\hat{W}}(r) + \int_t^\infty (r - t) dH_{\hat{W}}(r)\right] dF_1(t)}_{\text{Loss and gain due to delay: in the bracket, the first and second terms represent loss and gain, respectively}} .$$

In the preceding equation, the first term measures the intrinsic utility of joining server 1. The second term measures the customer's loss due to paying price P_1, higher than her plan of paying price P_2 with server 2. The third term measures the customer's gain from obtaining the net service reward from server 1 compared with her plan of balking. The fourth term measures the aggregate gain and loss due to the difference between the customer's realized and reference waiting times. The detailed calculation of the double integration is provided in online Appendix B. The utility function, U_1, can be further simplified as follows:

$$\begin{aligned} U_1 = {} & R - P_1 - \theta w_1 - \alpha\delta_{p2}(P_1 - P_2) + (R - P_1)(1 - \delta_{p1} - \delta_{p2}) \\ & + \theta \left[\beta\delta_{p2} w_2 - \beta(1 - \delta_{p1}) w_1 - (\beta - 1)\left(\delta_{p1}\frac{w_1}{2} + \delta_{p2}\frac{w_2^2}{w_1 + w_2}\right)\right]. \end{aligned} \tag{8}$$

Analogously, one can obtain the customer's utility from joining server 2 as follows:

$$U_2 = \underbrace{(R - P_2 - \theta w_2)}_{\text{Intrinsic utility}} \quad \underbrace{+\delta_{p1}(P_1 - P_2)}_{\text{Gain from paying a lower price than } P_1} \quad \underbrace{+(R - P_2)(1 - \delta_{p1} - \delta_{p2})}_{\text{Gain from obtaining the service}}$$
$$\underbrace{+\theta \left[\beta\delta_{p1} w_1 - \beta(1 - \delta_{p2}) w_2 - (\beta - 1)\left(\delta_{p2}\frac{w_2}{2} + \delta_{p1}\frac{w_1^2}{w_1 + w_2}\right)\right]}_{\text{Loss and gain due to the difference between experienced and benchmark delay}} .$$

Similar to the monopoly case, the utility of balking consists of two parts:

$$\begin{aligned} U_b = {} & \underbrace{\mathrm{E}[\alpha(0 - \hat{V})^-]}_{\text{Loss due to losing the net reward}} \quad \underbrace{-\theta\mathrm{E}[(0 - \hat{W})^-]}_{\text{Gain due to avoiding waiting}} \\ = {} & -\alpha\delta_{p1}(R - P_1) - \alpha\delta_{p2}(R - P_2) + \theta(w_1\delta_{p1} + w_2\delta_{p2}). \end{aligned} \tag{9}$$

The utilities from joining these two servers and from balking determine the customer's optimal choices, denoted as δ_{o1} and δ_{o2}. For instance, if $U_1 > U_2$ and $U_1 > U_b$, the optimal choices for the tagged customer are $\delta_{o1} = 1$ and $\delta_{o2} = 0$. Similar to the monopoly case, here we consider the symmetric equilibrium for customers' queueing games in which customers adopt the same joining strategy, which requires that $\delta_{o1} = \delta_1$ and $\delta_{o2} = \delta_2$. Moreover, following the requirement of *personal equilibrium* in Koszegi and Rabin (2006), a tagged customer's joining plan (based on which she forms her reference points) should be consistent with her realized optimal strategy; that is, $\delta_{p1} = \delta_{o1}$ and $\delta_{p2} = \delta_{o2}$. In short, the customer equilibrium in our setting requires that $\delta_{p1} = \delta_{o1} = \delta_1$ and $\delta_{p2} = \delta_{o2} = \delta_2$. Thus, in the following equilibrium analysis, we keep only notations δ_1 and δ_2.

Similarly, for the case in which $P_1 \leq P_2$, customers' utilities from joining server 1, joining server 2 and balking can be written, respectively, as follows:

$$\begin{aligned}
U_1 &= R - P_1 - \theta w_1 + \delta_2(P_2 - P_1) + (R - P_1)(1 - \delta_1 - \delta_2) \\
&\quad + \theta \left[\beta \delta_2 w_2 - \beta(1 - \delta_1) w_1 - (\beta - 1) \left(\delta_1 \frac{w_1}{2} + \delta_2 \frac{w_2^2}{w_1 + w_2} \right) \right], \\
U_2 &= R - P_2 - \theta w_2 - \alpha \delta_1(P_2 - P_1) + (R - P_2)(1 - \delta_1 - \delta_2) \\
&\quad + \theta \left[\beta \delta_1 w_1 - \beta(1 - \delta_2) w_2 - (\beta - 1) \left(\delta_2 \frac{w_2}{2} + \delta_1 \frac{w_1^2}{w_1 + w_2} \right) \right], \\
U_b &= -\alpha \delta_1(R - P_1) - \alpha \delta_2(R - P_2) + \theta(w_1 \delta_1 + w_2 \delta_2).
\end{aligned}$$

Due to the effect of loss aversion on price ($\alpha > 1$), the utilities from joining servers 1 and 2 ($U_1 and U_2$, respectively) when $P_1 \leq P_2$ are different from those when $P_1 \geq P_2$. In other words, utility functions U_1 and U_2 are not smooth at $P_1 = P_2$.

Given each set of prices (P_1 and P_2), customer equilibrium (δ_1 and δ_2) depends on the ordering of the three utilities, U_1, U_2 and U_b. Here, we provide three sets of conditions (stated in (10)) for solving the customer equilibrium depending on whether there exist balking customers. Set 1 represents the case in which joining is preferred over balking, and hence customers join either server 1 or 2 (i.e., $\delta_1 + \delta_2 = 1$). Set 2 represents the case in which customers are indifferent between joining and balking and there exist balking customers (i.e., $\delta_1 + \delta_2 < 1$). Set 3 provides a boundary case

between Sets 1 and 2, which represents the situation in which customers are indifferent between joining and balking and all customers choose joining. It is straightforward to show that any pair (δ_1, δ_2) satisfying one of the preceding three sets of conditions is a Nash equilibrium of the customer queueing game.

$$\text{(Set 1)} \begin{cases} U_1 = U_2 > U_b, \\ \delta_1 + \delta_2 = 1, \\ 0 \le \delta_i < \mu_i,\ i = 1, 2. \end{cases} \quad \text{(Set 2)} \begin{cases} U_1 = U_2 = U_b, \\ \delta_1 + \delta_2 < 1, \\ 0 \le \delta_i < \mu_i,\ i = 1, 2. \end{cases} \quad \text{(Set 3)} \begin{cases} U_1 = U_2 = U_b, \\ \delta_1 + \delta_2 = 1, \\ 0 \le \delta_i < \mu_i,\ i = 1, 2. \end{cases} \tag{10}$$

One can also specify other sets of sufficient conditions for the customer equilibrium where $U_1 \neq U_2$. However, without loss of generality, we can ignore such cases in studying servers' pricing games. The reason is that the server with a lower customer joining utility would have no customers joining it; hence, it would have an incentive to decrease its price until its customer joining utility matches that of the other server. Similarly, we can also ignore the case in which in equilibrium $U_1 < U_b$ and $U_2 < U_b$, as all customers would choose to balk. In other words, in considering the servers' pricing competition, the three sets of conditions stated in (10) are enough to characterize customer equilibria.

In general, to derive the closed-form solution for the customer equilibrium or to even discuss equilibrium uniqueness is highly challenging, as it involves mapping from the two-dimensional strategy set (δ_1 and δ_2) to the three-dimensional utility set and relies heavily on the topological properties of the set of strategies (Hassin and Roet-Green 2011). Fortunately, to conduct analysis of the pricing game between the two servers, we do not need the closed-form expressions of customer equilibrium joining probabilities. Instead, in deriving a server's best response function, we can list the sufficient conditions stated in (10) as the constraints in the server's revenue maximization problem, a strategy first adopted by Chen and Wan (2003). Furthermore, by utilizing the feature of the revenue maximization problem of each server, we can use a general set of looser conditions to replace all three sets of conditions stated in (10), as illustrated in the following subsection.

4.2. Two Servers' Pricing Competition Game

We are now ready to analyze the first-stage price competition game between two servers. We first derive the best response function of one server by assuming that the other server's price is fixed and then find the intersecting point of the two best response functions. As it is infeasible to derive the closed-form expression of the customer equilibrium joining strategy, we consider the corresponding equilibrium arrival rate as an implicit function of prices defined by (10).

To avoid trivial cases, we assume that each server is capable of attracting at least one customer by charging a price of 0 when there are no customers in the system, which implies that $R > \theta(\beta+1)/(2\mu)$. Define Π_i as server i's profit. We take server 1 as an example to illustrate how to identify its best response on pricing. Server 1's revenue maximization problem solves the following optimization problem:

$$\max_{0<P_1\le\overline{P},\delta_1,\delta_2} \quad \Pi_1 = P_1\cdot\delta_1, \tag{11}$$

$$\text{subject to} \quad U_1 = U_2, \tag{12}$$

$$U_1 \ge U_b, \tag{13}$$

$$\delta_1+\delta_2\le 1,\ 0\le\delta_1<\mu_1. \tag{14}$$

Similarly, we can find the best response function of server 2.

There are two points worth mentioning about the server's constrained optimization problem. First, as the customer equilibrium joining probability cannot be expressed as the closed-form expressions of the prices of the two servers, we use constraints (12)-(14) to represent the functional relationship between the customer equilibrium joining probabilities and the prices. Second, to avoid the cumbersome customer equilibrium conditions in (10), we can technically treat δ_1 and δ_2 as decision variables and use looser conditions in constraints (12)-(14) to capture customer equilibrium. To see this equivalence, note that δ_1 is in the objective function of server 1 (see (11)), and thus the maximization of server 1's revenue function forces δ_1 to be as large as possible. Consequently, either the constraint (13) is binding or in (14) $\delta_1+\delta_2=1$, which indeed forces the

constraints (12)-(14) to degenerate into one of the three sets of conditions stated in (10). (Note that the constraint $0 \leq \delta_2 < \mu$ is captured in server 2's optimization problem.) Hence, solving the preceding constrained optimization problem yields the best response function for server 1.

As it is difficult to conduct the general price equilibrium analysis, we shall instead focus on the symmetric Nash equilibrium for the pricing game, in which two servers adopt the same pricing strategy. If the symmetric Nash equilibrium does not exist, then we complement our analysis with numerical studies to explore what kind of equilibrium form it may take. Also note that the constraints (12)-(14) represent all of the customer equilibria given a general price pair (P_1, P_2); hence, the pricing game is studied over a very wide range. After we obtain the symmetric equilibrium price, we check whether the corresponding customer equilibrium is a PPE.[2] If it is a PPE, we can then claim that this symmetric price equilibrium is valid. It can be shown that if the symmetric price charged by the two servers is higher than $\overline{P}$, then the PPE is "balking." That is, no customer will join these two servers. Hence, to avoid the trivial cases of zero demand, price variables must be in the range of $[0, \overline{P}]$.

The price equilibrium structure depends on the relative magnitude of service rate μ compared with two thresholds, which are functions of $\overline{\mu}/2$, $\widetilde{\mu}$ and $\underline{\mu}$, where

$$\widetilde{\mu} = \frac{\sqrt{\left[(\beta+3)/(\alpha+1)\right]^2 + 8R/\theta \cdot (3\beta+5)/(\alpha+3)} + (\beta+3)/(\alpha+1)}{4R/\theta} + 0.5,$$

and $\underline{\mu}$ is the service rate that solves the equation (47) in online Appendix D.

Specifically, we consider the following three cases: fast servers with service rate $\mu > \min\{\widetilde{\mu}, \overline{\mu}/2\}$, moderate-speed servers with service rate $\min\{\underline{\mu}, \overline{\mu}/2\} \leq \mu \leq \min\{\widetilde{\mu}, \overline{\mu}/2\}$ and slow servers with service rate $\mu < \min\{\underline{\mu}, \overline{\mu}/2\}$. First, we briefly summarize the equilibrium results on the market coverage and customer utility in these three cases as follows. For the fast servers, all of the customers join the service, and their utility of joining is larger than that of balking. For the moderate-speed servers, all of the customers join the service, and their utility of joining equals that of balking. For the slow servers, some customers join the service and the others balk, and their utility of joining

equals that of balking. In the following, we only present the results on the equilibrium price and its comparison with the traditional counterpart, in addition to the sensitivity analysis with respect to the loss-aversion parameters. We also present the interesting conclusion that competition may be beneficial to a server when customers are loss averse. The detailed mathematical analysis can be found in online Appendix D.

4.2.1. Fast Server In this case, two over-capacitated servers compete for a relatively small population of potential customers. It turns out there exists a unique symmetric equilibrium, as stated in the following proposition.

PROPOSITION 5. *When $\mu > \min\{\widetilde{\mu}, \overline{\mu}/2\}$, there exists a unique symmetric Nash equilibrium between the two identical servers, denoted as $(P_1^e, P_2^e, \delta_1^e, \delta_2^e)$, with $\delta_1^e = \delta_2^e = 1/2$. The corresponding equilibrium prices, P_1^e and P_2^e, are*

$$P_1^e = P_2^e = \min\left\{\frac{(3\beta+5)}{2(\alpha+3)} \cdot \frac{\theta}{(\mu-0.5)^2},\ \overline{P}\right\}.$$

As shown in the proof of Proposition 5, if $\widetilde{\mu} < \overline{\mu}/2$, which requires the unit-time-cost-adjusted service reward (R/θ) to be above a certain threshold represented by some function of α and β, we have

$$P_1^e = P_2^e = \frac{(3\beta+5)}{2(\alpha+3)} \cdot \frac{\theta}{(\mu-0.5)^2} \tag{15}$$

for any $\mu > \widetilde{\mu}$. When $\alpha = \beta = 1$, the condition $\widetilde{\mu} < \overline{\mu}/2$ is always satisfied, as $\overline{\mu} = \infty$. Thus, when the service reward is relatively high, such that $\widetilde{\mu} < \overline{\mu}/2$, by (15), loss aversion toward waiting time drives the equilibrium price up, whereas loss aversion toward the net monetary reward drives the equilibrium price down. When customers care more about the loss resulting from a long waiting time, the server sets its price higher to decrease congestion, and the other server unfortunately thinks the same way. As a result, congestion is not reduced but price is increased. When customers are more loss averse toward the net reward, both servers want to undercut the other server's price, resulting in a lower equilibrium price. This result is similar to that obtained in the literature on price competition in the goods market, which concludes that loss aversion toward price lowers

equilibrium prices and benefits customers, whereas loss aversion toward product mismatch increases prices and softens competition (Zhou 2011, Heidhues and Koszegi 2008, Karle and Peitz 2012).

In a case where $\widetilde{\mu} < \overline{\mu}/2$, it would be interesting to further compare the equilibrium price in (15) to the traditional price where the reference effect is not considered. The traditional equilibrium prices in the fast server case (online Appendix E provides the detailed derivation) are

$$P_1^{\text{trad}} = P_2^{\text{trad}} = \frac{\theta}{(\mu - 0.5)^2},$$

which is very similar to (15), except that (15) has one extra loss-aversion-parameter-dependent term.

PROPOSITION 6. *When two servers are identical and $\widetilde{\mu} < \overline{\mu}/2$, the equilibrium price involving reference-dependent customers is higher (respectively, lower) than that in the traditional case when $3\beta > 2\alpha + 1$ (respectively, $3\beta < 2\alpha + 1$). Furthermore, the equilibrium price involving loss-neutral customers (i.e., $\alpha = \beta = 1$) is the same as that in the traditional case (i.e., $P_1^e = P_2^e = P_1^{trad} = P_2^{trad}$).*

When $\widetilde{\mu} < \overline{\mu}/2$, if α and β are equal and greater than 1, then condition $3\beta > 2\alpha + 1$ holds. Proposition 6 shows that the equilibrium price is higher than that in the traditional case. This implies that the effect of loss aversion toward waiting time dominates the effect of loss aversion toward net reward. This is due to the nature of waiting time uncertainty, which makes it more likely for customers to experience a loss in waiting time than in price.

4.2.2. Moderate-Speed Server When the capacity level decreases, waiting time increases and customers are more likely to balk. When the capacity level decreases to some critical point, the customers derive the same utility from joining as from balking, which implies that the competition between the two servers is softened. The following proposition summarizes the equilibrium results.

PROPOSITION 7. *When two servers are identical and $\min\{\underline{\mu}, \overline{\mu}/2\} \le \mu \le \min\{\widetilde{\mu}, \overline{\mu}/2\}$, the symmetric Nash equilibrium, denoted as $(P_1^e, P_2^e, \delta_1^e, \delta_2^e)$, is unique with $\delta_1^e = \delta_2^e = 1/2$ and*

$$P_1^e = P_2^e = R - \frac{\beta + 3}{2(\alpha + 1)} \cdot \frac{\theta}{\mu - 0.5}. \tag{16}$$

In equilibrium, customers are indifferent between joining the server and balking. Proposition 7 shows that the equilibrium price decreases with loss aversion toward waiting time (β) and increases with loss aversion toward net reward (α), a result that sharply contrasts that in the fast server case. Recall that when the capacity level is moderate, all of the customers join the service, although they are indifferent between joining and balking. When β increases, the customers are more loss averse toward waiting time and the servers decrease the price to compensate for customer loss in waiting time. However, when α increases, customers are more loss averse toward the net reward, which means they care more about losing the chance of receiving the service and thus the servers can increase the price.

Next, we compare the equilibrium price in (16) to the traditional price, where the reference effect is not considered. The traditional equilibrium prices in a moderate server case (online Appendix E provides the detailed derivation) are

$$P_1^{\text{trad}} = P_2^{\text{trad}} = R - \frac{\theta}{\mu - 0.5},$$

which is very similar to (16), except that (16) has one extra loss-aversion-parameter-dependent term.

PROPOSITION 8. *When two servers are identical and* $\min\{\underline{\mu}, \overline{\mu}/2\} \le \mu \le \min\{\widetilde{\mu}, \overline{\mu}/2\}$*, the equilibrium price involving reference-dependent customers is higher (respectively, lower) than that in the traditional case in which* $\beta < 2\alpha - 1$ *(respectively,* $\beta > 2\alpha - 1$*). Furthermore, the equilibrium price involving loss-neutral customers (i.e.,* $\alpha = \beta = 1$*) is the same as that in the traditional case (i.e.,* $P_1^e = P_2^e = P_1^{trad} = P_2^{trad}$*).*

When $\min\{\underline{\mu}, \overline{\mu}/2\} \le \mu \le \min\{\widetilde{\mu}, \overline{\mu}/2\}$, if α and β are equal and greater than 1, condition $\beta < 2\alpha - 1$ holds. Proposition 8 shows that the equilibrium price is higher than that in the traditional case. This implies that the effect of loss aversion on net reward dominates the effect of loss aversion on waiting time. Contrary to the fast server case, here customers are indifferent between joining the service and balking. The feeling of a loss in net reward when facing balking ($-\alpha(R - P_1)$) is much stronger than that when facing another competitor ($-\alpha(P_2 - P_1)$), making loss aversion related to net reward the dominant factor.

4.2.3. Slow Server When the capacity level decreases further, some customers choose to balk. In the traditional setting where the reference effect is not considered, Chen and Wan (2003) show that when the market is ample enough, such that balking customers exist, each of the two servers behaves like a monopoly and charges its monopoly price. However, this conclusion does not hold in our setting when the reference effect is considered. The change in one server's price not only affects its own arrival rate, but also indirectly affects the other server's arrival rate by changing customers' reference points. Thus, these two servers are still in competition when the reference effect is present. The equilibrium prices are the solutions of Karuch-Kuhn-Tucker conditions (48) to (57) shown in online Appendix D. The following analytical result can be derived with regard to this game.

PROPOSITION 9. *When two servers are identical and* $\theta(\beta+1)/(2R) < \mu < \min\{\underline{\mu}, \overline{\mu}/2\}$,

(1) If $\mu < 1/[\alpha - 1/(\beta+1)]$, *there exists a continuum of symmetric equilibria* $(P_1^e, P_2^e, \delta_1^e, \delta_2^e)$; *specifically, any joining probability in the interval*

$$\delta_1^e = \delta_2^e \in \left[\underline{\delta},\ \overline{\delta}\right]$$

constitutes an equilibrium, where $\underline{\delta}$ *and* $\overline{\delta}$ *are the solutions of equations (67) and (68) in online Appendix D, respectively.*

(2) If $\mu \geq 1/[\alpha - 1/(\beta+1)]$, *the symmetric equilibrium exists if and only if the following condition is satisfied:*

$$h(\alpha,\beta,\mu,R) := \left[\delta - \frac{\widehat{P}(\alpha-1)\delta}{X-2Y} - \frac{2+(\alpha-1)\delta}{2}\left(\frac{\widehat{P}}{X-2Y} + \frac{\widehat{P}}{X}\right)\right]\Bigg|_{\delta=\tilde{\delta}} < 0, \tag{17}$$

where the detailed expressions of $\widehat{P}$, X, Y *and* $\tilde{\delta}$ *are defined in equations (58), (59) and (65) in online Appendix D, respectively. If a symmetric equilibrium exists, then the equilibrium joining probability can be any value in the interval*

$$\delta_1^e = \delta_2^e \in \left[\underline{\delta},\ \overline{\delta}\right] \cap \left(\tilde{\delta}, \overline{\delta}\right].$$

When there exists a continuum of symmetric equilibria, the equilibrium prices are

$$P_1^e = P_2^e = R - \theta \frac{1}{\mu - \delta_1^e} \cdot \frac{\beta + 1 - (\beta - 1)\delta_1^e}{2 + 2(\alpha - 1)\delta_1^e},$$

and the equilibrium with the highest joining probability is the Pareto optimal equilibrium. In particular, when $\alpha = \beta = 1$, there exists a unique symmetric Nash equilibrium in which each server charges its monopoly price, the same as that under the traditional slow server case (stated in online Appendix E).

Although in the slow server case there exist balking customers, the two servers still compete due to the reference effect. Proposition 9 shows that the equilibrium structure in the slow server case is significantly different from the equilibrium structures in the moderate-speed and fast server cases in the following two aspects. First, there may be a continuum of symmetric equilibria when the service rate is below a certain level. Second, the symmetric equilibria may not exist when the service rate is above this level. The first difference can be explained by competition intensity: when the service rate is very low, most customers choose to balk. Hence, the reference points are heavily influenced by the net reward and waiting time associated with the balking option, allowing the competition between the two servers to be relatively weak and still making it possible for the two servers to charge the same price.

The key question is why a symmetric equilibrium may not exist when the service rate is relatively high. The answer lies in the server's profit function being kinked, with the kink point $P_1 = P_2$. This implies that the right- and left-hand derivatives of the profit function with respect to price are not the same. A symmetric equilibrium, if it exists, requires the left-hand derivative of a server's profit function with respect to its price to be non-negative and the corresponding right-hand derivative to be non-positive at the kink point, $P_1 = P_2$. This implies that the right-hand derivative cannot be larger than the left-hand derivative. However, this necessary condition may be violated in certain situations. Luski (1976) demonstrates such a case in which there exist balking customers and customers are heterogeneous in delay sensitivity, showing that the right-hand derivative is

larger than the left-hand derivative for a server's profit function at the point $P_1 = P_2$. However, Luski (1976) provides the mathematical proof for this statement only without any explanation. We provide our own understanding of the potential reasons for this result. In the setting of Luski (1976), when two servers set different prices, customers are segmented according to their own delay sensitivity. The highly delay-sensitive customers balk, the delay-insensitive customers join the low-price server and the moderately delay-sensitive customers join the high-price server. Due to the heterogeneity of the customer delay sensitivity attribute, slightly increasing a server's price in the region right above its competitor's price benefits the server more than doing so in the region right below its competitor's price, as in the former case reduced congestion allows the server (with a price now higher than that of its competitor) to attract the balking customers, which mitigates its demand loss. Indeed, Luski (1976) also shows that when balking customers do not exist, the left- and right-hand derivatives are equal and a symmetric equilibrium may exist.

Our setting is quite different from that of Luski (1976). However, a similar rationale can be applied to explain our equilibrium structure. Customer loss aversion (asymmetry between gains and losses) generates the kink in the servers' profit function at point $P_1 = P_2$. That is, at $P_1 = P_2$, the marginal profit change when a server (e.g., server 1) increases its price in a region right above P_2 does not equal that when the server does so in a region right below P_2. As in Luski (1976), here, the right-hand derivative may be larger than the left-hand derivative. To demonstrate this point, we can "position" the three options, joining the high-price server, joining the low-price server and balking along the two attributes: the net reward and waiting time. On each attribute, the option to join the high-price server is located between the other two and is hence "closer" to the balking option. Therefore, increasing a server's price in the region right above its competitor's price benefits it more than doing so in the region right below its competitor's price, as in the former case the server is located "closer" to the balking option, and hence this server together with the balking option have more influence on the customer's reference points than the competitor. Consequently, customer loss aversion further mitigates the server's demand loss from the increased price. Indeed,

when there is no balking customer, as in the fast- and moderate-speed server cases, symmetric equilibrium exists. Moreover, larger loss aversion parameter values of α and β lead to a larger mitigation effect on the demand loss and thus make symmetric equilibrium less possible, as shown by the following corollary.

COROLLARY 1. *The function $h(\alpha, \beta, \mu, R)$ increases in α and β.*

Corollary 1 and the decrease of threshold $1/[\alpha - 1/(\beta + 1)]$ in α and β imply that for a fixed capacity μ, with larger loss aversion parameters, α and β, symmetric equilibria are less likely to exist. Figure 2 depicts the existence and non-existence of symmetric equilibria when we vary α and β, which further illustrates Corollary 1.

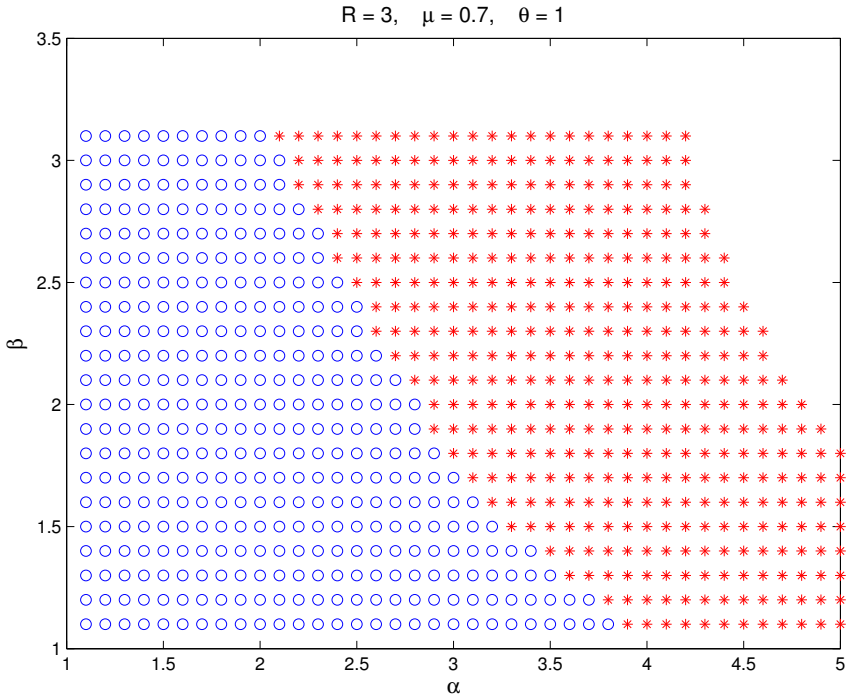

Figure 2 Illustration of the equilibrium map. The circles represent the region in which symmetric equilibria exist, and the stars represent the region in which symmetric equilibria do not exist. The two regions together represent the feasible region: $\theta(\beta+1)/(2R) < \mu < \min\{\underline{\mu}, \overline{\mu}/2\}$.

When symmetric equilibria do not exist, we conduct numerical experiments to study the equilibrium structure. Specifically, we set the service-related parameters to $R = 3$, $\theta = 1$ and $\mu = 0.7$.

We then vary the degree of loss aversion on waiting time, $\beta \in [2, 2.5, 3, 3.5, 4, 4.5]$, and the degree of loss aversion on net monetary reward, $\alpha \in [3.5, 4, 4.5, 5, 5.5, 6]$. The numerical results show that in all cases, there exist two asymmetric equilibria. Our conclusion here is very insightful: a large market size and customer loss aversion preference can drive two competing servers to provide heterogeneous service products.

4.3. Competition May be Beneficial for a Server

With equilibrium prices (P_1^e, P_2^e) and joining probabilities (δ_1^e, δ_2^e), it is easy to calculate the profit of each server under competition. A natural question arises: is it possible to introduce a competitor to benefit a monopoly? It is well established in the literature (e.g., Chen and Wan (2003)) that in a traditional service setting where loss aversion is not considered, it is always beneficial for a server to be a monopoly than to introduce a competitor. Surprisingly, in our setting with loss-averse customers, there are cases in which introducing a competitor actually benefits a monopolistic server. In the following, we first conduct a simple analysis of the changes of the customer joining and balking utilities when a competitor is introduced into the market to understand why competition may be beneficial. We then provide numerical examples to illustrate this point.

Consider that a server in a monopoly market charges price P and the equilibrium joining probability for customers is δ_p. Assuming that the service capacity is small enough, such that $\delta_p < 1/2$, we then have $U(\text{Join}) = U(\text{Balk}) < 0$. Now an identical server enters the market. Assume that both servers still charge price P and customers still join each server with probability δ_p. Substituting price P and customer joining probability δ_p into the customer joining utility expression stated in (8) and the baking utility expression stated in (9), we obtain

$$U_1 = U_2 = \frac{2(1-\delta_p)}{2-\delta_p} U(\text{Join}) \text{ and } U_b = 2U(\text{Balk}).$$

Based on this, we obtain

$$U_1 = U_2 = \frac{1-\delta_p}{2-\delta_p} U_b > U_b,$$

where the inequality holds because $1-\delta_p < 2-\delta_p$ and $U_b < 0$. Therefore, introducing a competitor can make the joining option more favorable than balking, and thus customers may join the original server with a higher probability. This implies that a server may earn more revenue in a competitive market than in a monopolistic market.

Next, we numerically show that competition can benefit a server. We set the degrees of loss aversion on waiting time and monetary reward to $\beta = 1.81$ and $\alpha = 2.25$, respectively. We also set $R = 5$ and $\theta = 1$. In the first example, we set $\mu = 1.03$. This falls into the fast server case ($\widetilde{\mu} < \mu < \overline{\mu}/2$). It turns out that the profit of the monopoly is 1.7065 with price $P^* = 2.7087$ and joining probability $\delta^* = 0.63$, whereas the profit of the duopolistic server is 1.7681 with an equilibrium price of $P^e = 3.5363$. In the second example, we set $\mu = 1.02$. This falls into the moderate-speed server case ($\underline{\mu} < \mu < \widetilde{\mu} < \overline{\mu}/2$). In this case, the profit of the monopoly is 1.6707 with price $P^* = 2.6947$ and joining probability $\delta^* = 0.62$, whereas the profit of the duopolistic server is 1.7885 with an equilibrium price of $P^e = 3.5769$. Both numerical examples illustrate that a server can benefit from competition when customers are loss averse.

Here, having more servers in the market increases the overall service capacity and allows more customers to consume the service. This changes customers' reference points, and balking without consumption of service is deemed to be a larger loss. The larger benefit of the joining option relative to the balking option allows servers to charge a price even higher than their monopolistic price and achieve a profit even larger than their monopolistic profit. This phenomenon is similar to the compromise effect in the marketing literature, which states that expanding the choice set for customers may increase the demand for a certain option (see Simonson (1989) for a discussion on this effect in context-dependent customer preferences).

5. Conclusions

In this study, we consider a customer's overall expected utility from a service to be the sum of her intrinsic and gain-loss utilities. Gain-loss utility measures the deviation of a customer's attribute values, the net monetary reward and delay based on her reference points. We consider

both monopoly and duopoly markets and examine customers' equilibrium queueing strategies in both. Based on the customer equilibrium analysis, we further analyze servers' pricing decisions.

Our conclusions are novel in at least three aspects. First, with reference effects and loss aversion, it is typical to have multiple equilibria, a finding that contradicts the unique equilibrium found by studies that do not consider loss aversion. Second, we show that the loss aversion effect polarizes the queues in a monopoly market, making long queues even longer and short queues even shorter. We also show that profit- and welfare-maximizing prices are not the same. Embedding the gain-loss utility into the welfare function makes social welfare smaller than server profit, as customers care more about losses than gains. Third, in a symmetric duopoly market, we examine the existence of the symmetric price equilibrium and conduct a sensitivity analysis of the equilibrium price depending on the magnitude of the service rate. In a large-capacity system (i.e., when the service speed is high), a unique symmetric price equilibrium exists. All customers are served and equally split between the two servers. Moreover, when the service reward is high, the equilibrium price decreases in the loss aversion parameter on the monetary attribute, but increases in the loss aversion parameter on the delay attribute. In a moderate-capacity system (i.e., when service speed is moderate), there also exists a unique symmetric price equilibrium. This symmetric price equilibrium, however, increases in the loss aversion parameter on the monetary attribute, but decreases in the loss aversion parameter on the delay attribute. In a small-capacity system (i.e., when service speed is low), we show that a symmetric equilibrium may not exist. We then numerically find that there may exist two asymmetric equilibria in this case. Our result from the slow server case demonstrates two driving forces for two identical servers to offer heterogeneous service products: the existence of balking customers and the customer's loss aversion preference. Strikingly, we show that a server's equilibrium profit in a duopoly market can be larger than its monopoly profit. As the customer choice set is expanded by introducing a competitive server, more incentives to join can be granted to customers, as reference-dependent utility is affected by available alternatives.

Our study appears to be the first to examine the reference effect on service systems when customers are loss averse toward both price and waiting time. We show that both customers'

equilibrium joining strategies and firms' pricing decisions are significantly different from the results obtained without considering the reference effect or only considering the reference effect in terms of waiting time. We hope our work stimulates empirical research testing our analytical results in different service systems, such as those in the healthcare field and at call centers. Future research may also consider analyzing loss-averse customers with observable queues.

Endnotes

1. The polarizing effect caused by reference effects and loss aversion may provide an explanation for herding behavior in queues, in addition to explanations from the viewpoints of information asymmetry and quality concern examined by Bikhchandani et al. (1992) and Veeraraghavan and Debo (2008).

2. Luckily, under a symmetric price equilibrium, it is possible to check whether the corresponding customer equilibrium is a PPE. Facing the same price, P, customers join two symmetric servers with the same probability (i.e., $\delta_1 = \delta_2$). This can be shown by substituting $P_1 = P_2 = P$ into $U_1 = U_2$ to obtain $\delta_1 = \delta_2$. The queueing game for customers is reduced to a choice between joining (one of the servers) and balking, and thus boils down to the queueing game under the monopoly setting. Consequently, similar to that stated in Proposition 1, there exist three different customer equilibrium structures corresponding to the cases of $\mu \leq 1/2$, $1/2 < \mu < \overline{\mu}/2$ and $\mu \geq \overline{\mu}/2$. As there are two servers in the duopoly market, here the capacity thresholds in the monopoly case, 1 and $\overline{\mu}$, shall be divided by 2.

References

Abdellaoui, M., E. Kemel. 2014. Eliciting prospect theory when consequences are measured in time units: "time is not money". *Management Science* **60**(7) 1844–1859.

Afèche, P., O. Baron, Y. Kerner. 2013. Pricing time-sensitive services based on realized performance. *Manufacturing & Service Operations Management* **15**(3) 492–506.

Aksoy-Pierson, M., G. Allon, A. Federgruen. 2013. Price competition under mixed multinomial logit demand functions. *Management Science* **59**(8) 1817–1835.

Allon, G., A. Federgruen. 2007. Competition in service industries. *Operations Research* **55**(1) 37–55.

Anderson, E. W., M. W. Sullivan. 1993. The antecedents and consequences of customer satisfaction for firms. *Marketing Science* **12**(2) 125–143.

Baron, O., M. Hu, S. Asadolahi, Q. Qian. 2015. Newsvendor selling to loss averse consumers with stochastic reference points. *Manufacturing & Service Operatios Management.* Forthcoming.

Bikhchandani, S., D. Hirshleifer, I. Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* **100**(5) 992–1026.

Cachon, G. P., P. T. Harker. 2002. Competition and outsourcing with scale economics. *Management Science* **48**(10) 1314–1333.

Chen, H., M. Frank. 2004. Monopoly pricing when customers queue. *IIE Transactions* **36** 569–581.

Chen, H., Y. W. Wan. 2003. Price competition of make-to-order firms. *IIE Transactions* **35**(9) 817–832.

Chen, X., P. Hu, S. Shum, Y. Zhang. 2014. Dynamic stochastic inventory management with reference price effects Working paper. University of Illinois at Urbana-Champaign, USA.

Debo, L., C. Parlur, U. Rajan. 2012. Signaling quality via queues. *Management Science* **58** 876–891.

Edelson, N. M., D. K. Hildebrand. 1975. Congestion tolls for poisson queuing processes. *Econometrica* **43**(1) 81–92.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Norwell, MA.

Hassin, R., R. Roet-Green. 2011. Equilibrium in a two dimensional queueing game: When inspecting the queue is costly Working paper. Tel Aviv, Israel.

Heidhues, P., B. Koszegi. 2008. Competition and price variation when consumers are loss averse. *American Economic Review* **98**(4) 1245–1268.

Ho, T. H., Y. S. Zheng. 2004. Setting customer expectation in service delivery: An integrated marketing-operations perspective. *Management Science* **50**(4) 479–488.

Huang, T., G. Allon, A. Bassamboo. 2013. Bounded rationality in service systems. *Manufacturing and Service Operations Management* **15**(2) 252–265.

Kahneman, D., A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* **47**(2) 263–292.

Karle, H., M. Peitz. 2012. Competition under consumer loss aversion. Working paper. Univ. of Mannheim.

Koszegi, B., M. Rabin. 2006. A model of reference-dependent preferences. *Quarterly Journal of Economics* **121**(4) 1133–1165.

Li, X., P. Guo, Z. Lian. 2016. Quality-speed competition in customer-intensive services with boundedly rational customers. *Production and Operations Management* Forthcoming.

Lin, J. H, T. R. Lee, W. Jen. 2008. Assessing asymmetric response effect of behavioral intention to service quality in an integrated psychological decision-making process model of intercity bus passengers: A case of taiwan. *Transportation* **35** 129–144.

Lindsey, R. 2011. State-dependent congestion pricing with reference-dependent preferences. *Transportation Research Part B* **45** 1501–1526.

Luski, I. 1976. On partial equilibrium in a queuing system with two servers. *The Review of Economic Studies* **43**(3) 519–525.

Nasiry, J., I. Popescu. 2011. Dynamic pricing with loss averse consumers and peak-end anchoring. *Operations Research* **59**(6) 1361–1368.

Parasuraman, A., V. A. Zeithaml, L. Berry. 1985. A conceptual model of service quality and implications for future research. *Journal of Marketing* **49**(4) 41–50.

Popescu, I., Y. Wu. 2007. Dynamic pricing strategies with reference effects. *Operations Research* **55**(3) 413–429.

Roels, G., X. Su. 2014. Optimal design of social comparison effects: setting reference groups and reference points. *Management Science* **60**(3) 606–627.

Rust, R. T., J. J. Inman, J. Jia, A. Zahorik. 1999. What you don't know about customer perceived quality: the role of customer expectation distributions. *Marketing Science* **18**(1) 77–92.

Shang, W., L. Liu. 2011. Promised delivery time and capacity games in time-based competition. *Management Science* **57**(3) 599–610.

Simonson, I. 1989. Choice based on reasons: The case of attraction and compromise effects. *The Journal of Consumer Research* **16**(2) 158–174.

Tereyagoglu, N., P. Fader, S. Veeraraghavan. 2015. Multi-attribute loss aversion and reference dependence: evidence from the performing arts industry Working paper. Georgia Institute of Technology, the USA.

Veeraraghavan, S., L. Debo. 2008. Joining longer queues: information externalities in queue choice. *Manufacturing and Service Operations Management* **11**(4) 543–562.

Yang, L., F. de Vericourt, P. Sun. 2014. Time-based competition with benchmark effects. *Manufacturing and Service Operations Management* **1**(16) 119–132.

Zhou, J. 2011. Reference dependence and market competition. *Journal of Economics and Management Strategy* **20**(4) 1073–1097.

Online Appendices

Service Pricing with Loss Averse Customers

L. Yang, P. Guo, Y. Wang

Appendix A: Detailed List of All Equilibria

To simplify the notions, define $\nu = \frac{R-P}{\theta/\mu}$, which refers to the maximal number of service periods a customer is willing to wait. The equation $g(\delta) = 0$ can then be rewritten as

$$-\delta^2(\alpha-1) + \delta\left(\frac{(\beta-1)\mu}{2\nu} - 2 + \mu(\alpha-1)\right) - \frac{(1+\beta)\mu}{\nu} + 2\mu = 0. \tag{18}$$

Let $\kappa(\delta)$ be the left-hand side of (18). It is easy to check that $\kappa(\delta)$ is a quadratic and concave function of δ. Equation (18) could have at most two roots. To characterize the number of roots, define

$$\Delta = \left(\frac{(\beta-1)\mu}{2\nu} - 2 + \mu(\alpha-1)\right)^2 - 4(\alpha-1)\left(\frac{(1+\beta)\mu}{\nu} - 2\mu\right).$$

Given that we are only interested in solutions within $[0,1]$, we shall consider the value of $\kappa(\delta)$ at the boundaries. It can be shown that

$$\kappa(0) = 2\mu - (1+\beta)\mu/\nu, \quad \kappa(1) = (\mu-1)(\alpha+1) - (\beta+3)\mu/(2\nu).$$

There may be at most three equilibria for the customer queueing game, as summarized in the following proposition.

PROPOSITION 10. *Depending on the sign of $\kappa(0), \kappa(1)$ and Δ, the equilibrium in the monopoly case can be characterized as follows.*

(1) If $\kappa(0) > 0$ and $\kappa(1) < 0$, there exists a unique mixed-strategy equilibrium $\delta^e \in (0,1)$, which is the solution of (18).

(2) If $\kappa(0) > 0$ and $\kappa(1) > 0$, it follows that $\kappa(\delta) \geq 0$ for all δ, and hence joining is a dominant strategy and $\delta^e = 1$.

(3) If $\kappa(0) < 0$ and $\kappa(1) > 0$. there exist three equilibria: a pure-strategy equilibrium $\delta^e = 0$, a mixed-strategy equilibrium $\delta^e \in (0,1)$, which is the solution of (18), and a pure strategy equilibrium $\delta^e = 1$.

(4) If $\kappa(0) < 0$ and $\kappa(1) < 0$, there may be three, two or one unique equilibria, depending on the value of Δ:

(i) if $\Delta > 0$, there exist three equilibria: two mixed-strategy equilibria, characterized by the two roots of (18), η_1 and η_2, and a pure strategy equilibrium (all balk) $\delta^e = 0$;

(ii) if $\Delta = 0$, the two roots decrease to one, defining a mixed-strategy equilibrium, and the other equilibrium is the pure strategy equilibrium (all balk) $\delta^e = 0$; and

(iii) if $\Delta < 0$, there is only a unique pure strategy equilibrium (all balk) $\delta^e = 0$.

The boundary cases with $\kappa(0) = 0$ and/or $\kappa(1) = 0$ can be easily derived and thus are ignored here.

Appendix B: Detailed Calculation of Reference-dependent Utilities

Detailed Calculation of Reference-dependent Utility in (2) is listed as follows:

We first show the following integration.

$$\begin{aligned}\int_0^t (r-t)dF(r) &= (r-t)F(r)|_0^t - \int_0^t F(r)dr \\ &= -(r+we^{-r/w})|_0^t \\ &= (w-t) - we^{-t/w}.\end{aligned}$$

We then obtain

$$\begin{aligned}\int_t^\infty (r-t)dF(r) &= \int_0^\infty (r-t)dF(r) - \int_0^t (r-t)dF(r) \\ &= (w-t) - [(w-t) - we^{-t/w}] \\ &= we^{-t/w}.\end{aligned}$$

Thus, the integration

$$\begin{aligned}&\int_0^\infty (-\beta t(1-\delta_p) + \beta\delta_p \int_0^t (r-t)dF(r) + \delta_p \int_t^\infty (r-t)dF(r))dF(t) \\ &= \int_0^\infty (-\beta t(1-\delta_p) + \beta\delta_p[(w-t) - we^{-t/w}] + \delta_p we^{-t/w})(1/w)e^{-t/w}dt \\ &= -\beta(1-\delta_p)w + \delta_p(1-\beta)\int_0^\infty e^{-2t/w}dt \\ &= -\beta(1-\delta_p)w + \delta_p(1-\beta)w/2.\end{aligned}$$

The utility can then be derived as

$$\begin{aligned}U(\text{Join}) &= R - P - \theta w + (1-\delta_p)(R-P) \\ &\quad + \theta(-\beta(1-\delta_p)w + \delta_p(1-\beta)w/2) \\ &= (2-\delta_p)(R-P) - \theta w(2-\delta_p)\frac{1+\beta}{2} \\ &= (2-\delta_p)\left(R - P - \theta w\frac{1+\beta}{2}\right).\end{aligned}$$

Detailed Calculation of Reference-dependent Utility in (8) is listed as follows:

Utilizing the expression of $H_{\hat{W}}(r)$ in (7), we can calculate the internal integration.

$$
\begin{aligned}
&\beta\int_0^t (r-t)dH_{\hat{W}}(r)+\int_t^{\infty}(r-t)dH_{\hat{W}}(r)\\
&=\beta(1-\delta_{p1}-\delta_{p2})(-t)+(\beta-1)\int_0^t (r-t)d\left(\delta_{p1}F_1(r)+\delta_{p2}F_2(r)\right)+\delta_{p1}(w_1-t)+\delta_{p2}(w_2-t)\\
&=(\beta-1)\left[(r-t)\left(\delta_{p1}F_1(r)+\delta_{p2}F_2(r)\right)\Big|_0^t-\int_0^t(\delta_{p1}F_1(r)+\delta_{p2}F_2(r))dr\right]\\
&\quad+\beta(1-\delta_{p1}-\delta_{p2})(-t)+\delta_{p1}(w_1-t)+\delta_{p2}(w_2-t)\\
&=(\beta-1)\left[0-\int_0^t\delta_{p1}\left(1-e^{-\frac{r}{w_1}}\right)+\delta_{p2}\left(1-e^{-\frac{r}{w_2}}\right)dr\right]+(\beta-1)(\delta_{p1}+\delta_{p2})t-\beta t+\delta_{p1}w_1+\delta_{p2}w_2\\
&=-(\beta-1)\left[\delta_{p1}\left(t+w_1e^{-\frac{t}{w_1}}-w_1\right)+\delta_{p2}\left(t+w_2e^{-\frac{t}{w_2}}-w_2\right)\right]+(\beta-1)(\delta_{p1}+\delta_{p2})t-\beta t+\delta_{p1}w_1+\delta_{p2}w_2\\
&=-\beta t+\beta(\delta_{p1}w_1+\delta_{p2}w_2)-(\beta-1)\left(\delta_{p1}w_1e^{-\frac{t}{w_1}}+\delta_{p2}w_2e^{-\frac{t}{w_2}}\right).
\end{aligned}
$$

We then calculate the entire gain-and-loss term as

$$
\begin{aligned}
&\int_0^{\infty}\left[-\beta t+\beta(\delta_{p1}w_1+\delta_{p2}w_2)-(\beta-1)\left(\delta_{p1}w_1e^{-\frac{t}{w_1}}+\delta_{p2}w_2e^{-\frac{t}{w_2}}\right)\right]\frac{1}{w_1}e^{-\frac{t}{w_1}}dt\\
&=\beta\delta_{p2}w_2-\beta(1-\delta_{p1})w_1-(\beta-1)\left(\delta_{p1}\frac{w_1}{2}+\delta_{p2}\frac{w_2^2}{w_1+w_2}\right).
\end{aligned}
$$

Appendix C: Proofs of Lemmas and Propositions in the Monopoly Case

Proof of Lemma 1: It can be shown that

$$g'(\delta)=(\alpha-1)(R-P)-\theta\frac{1+\beta-(\beta-1)\mu/2}{(\mu-\delta)^2}.$$

If $\mu<2(\beta+1)/(\beta-1)$, the term $\frac{1+\beta-(\beta-1)\mu/2}{(\mu-\delta)^2}$ is increasing in δ and hence $g'(\delta)$ is decreasing in δ. This implies that $g(\delta)$ is concave in δ. Otherwise, $g'(\delta)>0$, and $g(\delta)$ is increasing in δ.

Proof of Proposition 1: First, we examine the case of $\mu\le 1$. According to Lemma 1, $g(\delta)$ is concave in δ. The value of $g(\delta)$ at $\delta=0$ is $2(R-P)-\theta(\beta+1)/\mu$ and that at $\delta=\mu$ is $-\infty$. Depending on the value of price P, we are presented with the following two scenarios.

(1). If $P\ge\overline{P}$, $g(0)\le 0$. Hence, "all balk" ($\delta=0$) is an equilibrium because $U(\text{Balk})\ge U(\text{Join})$. In addition, there are at most two roots, η_1 and η_2, of the equation $g(\delta)=0$. However, the mixed-strategy equilibrium generates negative customer utility. We take η_2 as an example to illustrate it.

$$
\begin{aligned}
U(\text{Balk}|\delta=\eta_2)=U(\text{Join}|\delta=\eta_2)&=(2-\eta_2)\left(R-P-\theta\frac{\beta+1}{2(\mu-\eta_2)}\right)\\
&<(2-\eta_2)\left(R-P-\theta\frac{\beta+1}{2\mu}\right)\\
&<0=U(\text{Balk}|\delta=0),
\end{aligned}
$$

which means that $\delta = 0$ is a PPE and $\delta = \eta_2$ is not. One could similarly argue that the equilibrium defined by the smaller solution is not a PPE either.

(2). If $P < \overline{P}$, then $g(0) > 0$. Given that $g(\mu) = -\infty$ and the concavity of $g(\delta)$, the equation $g(\delta) = 0$ has a unique root and falls in the range of $(0, \mu)$. This solution defines a unique equilibrium that must be a PPE.

We then examine the case of $1 < \mu < \overline{\mu}$. In this case, $g(\delta)$ is also concave in δ. The value of $g(\delta)$ at $\delta = 1$ is $(\alpha+1)(R-P) - \theta(\beta+3)/2(\mu-1)$. Again, depending on the value of P, we are presented with the following three scenarios.

(1). If $P \geq \overline{P}$, given $\mu < \overline{\mu}$, we have $g(0) \leq 0$ and $g(1) < 0$. This is similar to the first scenario of $\mu \leq 1$. Thus, "all balk" ($\delta = 0$) is a PPE.

(2). If $\underline{P} < P < \overline{P}$, given $\mu < \overline{\mu}$, we have $g(0) > 0$ and $g(1) < 0$. This is the same as the second scenario of $\mu \leq 1$. Thus, we have a unique equilibrium $\delta^e \in (0, 1)$ that solves $g(\delta) = 0$ and is a PPE.

(3). If $P \leq \underline{P}$, given $\mu < \overline{\mu}$, we have $g(0) \geq 0$ and $g(1) > 0$. Therefore, $g(\delta) \geq 0$ and "all join" is a unique equilibrium, which must be a PPE.

Finally, we examine the case of $\mu \geq \overline{\mu}$.

(1). If $P \leq \overline{P}$, then $g(0) \geq 0$ and $g(1) > 0$. In this case, $g(\delta)$ can be either concave or increasing in δ, and in both cases $g(\delta) \geq 0$. Thus, 'all join' is a unique stable equilibrium and is a PPE.

(2). If $P > \overline{P}$, then $g(0) < 0$. Therefore, "all balk" ($\delta = 0$) is an equilibrium. Furthermore, depending on whether $g(1) > 0$, we are presented with the following two scenarios. First, if $g(1) > 0$, then $\delta = 1$ is an equilibrium. In addition, $g(\delta) = 0$ has one root, which, however, will generate a negative customer utility and hence cannot be a PPE. Furthermore, we have that

$$U(\text{Join}|\delta = 1) = R - P - \theta\frac{\beta+1}{2(\mu-1)} < R - P - \theta\frac{\beta+1}{2\mu} < 0 = U(\text{Balk}|\delta = 0),$$

which means that the equilibrium with $\delta = 0$ dominates that with $\delta = 1$ and is a PPE. Second, if $g(1) < 0$, then similar to the analysis of the first scenario of $\mu \leq 1$, the equilibrium with $\delta = 0$ is a PPE. This means that when $P \geq \overline{P}$, the PPE is $\delta = 0$.

Proof of Lemma 2: Given that $\Pi = [R - \theta A(\delta)w]\,\delta$, the first order derivative of the profit function Π with respect to δ is

$$\frac{\partial \Pi}{\partial \delta} = R - \theta A w - \theta\left(A'w + Aw^2\right)\delta = R - \theta A w^2 \mu - \theta A' w \delta.$$

The second equality follows from the equation $w + w^2\delta = w^2\mu$. Then, the second order derivative is

$$\frac{d^2\Pi}{d\delta^2} = -\theta\mu\left(A'w^2 + 2Aw^3\right) - \theta A'(w + w^2\delta) - \theta A'' w\delta = -2\theta w^2\left(A'\mu + A''\frac{\delta}{2w} + \mu w A\right).$$

Given that $A(\delta) = [\beta + 1 - \delta(\beta-1)/2] \,/\, [(\alpha-1)\delta + 2]$, we can derive

$$A' = -\frac{(\alpha-1)(\beta+1)+(\beta-1)}{[(\alpha-1)\delta+2]^2} \quad \text{and} \quad A'' = -\frac{2(\alpha-1)}{(\alpha-1)\delta+2}A'.$$

Plugging A'' into the above second order derivative expression, we then have

$$\begin{aligned}\frac{d^2\Pi}{d\delta^2} &= -2\theta w^2\left(\frac{2\mu+(\alpha-1)\delta^2}{2+(\alpha-1)\delta}A' + \mu w A\right)\\ &\leq -2\theta w^2\mu\left(A' + wA\right) = -2\theta w\mu(wA)',\end{aligned}$$

where the inequality holds because $A' < 0$ and $\mu \geq \delta$. Hence, to show this second-order derivative to be negative, it suffices to show that $(wA)' > 0$ for all δ^e.

Recall that when δ^e represents the mixed-strategy equilibrium (that is, its value is strictly between 0 and 1), it is the larger root of equation $P = R - \theta A w$, which is also stable. This stability means that function $R - \theta A w$ must down-cross line P as δ increases, implying that the derivative $(wA)' > 0$.

Proof of Proposition 2: We can rewrite the FOC (6) as

$$R - \theta w^2 \mu\left[A(\delta) + \delta\left(1 - \frac{\delta}{\mu}\right)A'(\delta)\right] = 0. \tag{19}$$

According to Lemma 2, the objective function is concave, implying that the left-hand-side (LHS) of (19) is decreasing in δ. Hence, to show the monotonicity of δ^* with respect to α or β, it suffices to demonstrate that the LHS of (19) is decreasing in β and increasing in α.

We can show that

$$\begin{aligned}\frac{\partial}{\partial\beta}\left[A + \delta\left(1-\frac{\delta}{\mu}\right)A'\right] &= \frac{2(1-\delta/2)^2 + \delta^2\alpha(1/\mu - 1/2)}{(2+\alpha\delta-\delta)^2}\\ &> \frac{2(1-\delta/2)^2 - \delta^2/2}{(2+\alpha\delta-\delta)^2} = \frac{2(1-\delta)}{(2+\alpha\delta-\delta)^2} \geq 0,\end{aligned}$$

where the first inequality follows from

$$\mu < \left(\bar{\mu} = \frac{\alpha+1}{\alpha - 2/(\beta+1)}\right) < \frac{2\alpha}{\alpha-1}.$$

Thus, the LHS of (19) is decreasing in β. Furthermore,

$$\begin{aligned}\frac{\partial}{\partial\alpha}\left[A + \delta\left(1-\frac{\delta}{\mu}\right)A'\right] &= \delta\frac{-\left[\beta+1-(\beta-1)\delta/2\right]\left[2+(\alpha-1)\delta\right] + (1-\delta/\mu)\left[(\beta+1)(\delta\alpha+\delta-2)-4\delta\right]}{(2-\delta+\delta\alpha)^3}\\ &\leq \delta\frac{-2(\beta+1)+\delta(\beta-1)-2(\beta+1)(1-\delta)(1-\delta/\mu)-4\delta(1-\delta/\mu)}{(2-\delta+\delta\alpha)^3}\\ &< 0,\end{aligned}$$

where the first inequality holds because the numerator decreases in α when

$$\mu \leq \left(\bar{\mu} = \frac{\alpha+1}{\alpha - 2/(\beta+1)}\right) < \frac{2(\beta+1)}{\beta-1}.$$

Thus, the LHS of (19) is increasing in α.

We then show that the optimal price P^* decreases with β and increases with α. Because the FOC must still be satisfied as β increases, we have

$$\frac{\partial}{\partial \delta}\left[w^2\left(A + \left(1 - \frac{\delta}{\mu}\right)\delta A'\right)\right]\frac{d\delta}{d\beta} + w^2\frac{\partial}{\partial \beta}\left[A + \left(1 - \frac{\delta}{\mu}\right)\delta A'\right] = 0.$$

Thus, we have

$$\frac{d\delta}{d\beta} = -\frac{\frac{\partial}{\partial \beta}\left[A + \left(1 - \frac{\delta}{\mu}\right)\delta A'\right]}{2A' + 2/(\mu - \delta)A + \delta(1 - \delta/\mu)A''}.$$

Using the preceding expression of $d\delta/d\beta$, we then obtain

$$\begin{aligned}
\frac{dP^*}{d\beta} &= -\theta\left[\frac{\partial(wA)}{\partial \delta}\frac{d\delta}{d\beta} + \frac{\partial(wA)}{\partial \beta}\right] \\
&= -\theta w\left\{-\frac{A' + 1/(\mu-\delta)A}{2(A' + 1/(\mu-\delta)A) + \delta(1-\delta/\mu)A''}\frac{\partial[A + (1-\delta/\mu)\delta A']}{\partial \beta} + \frac{\partial A}{\partial \beta}\right\} \\
&\leq -\theta w\left\{-\frac{1}{2}\frac{\partial[A + (1-\delta/\mu)\delta A']}{\partial \beta} + \frac{\partial A}{\partial \beta}\right\} \\
&= -\theta w\left\{\frac{1}{2}\frac{\partial[A - (1-\delta/\mu)\delta A']}{\partial \beta}\right\} \\
&< 0,
\end{aligned}$$

where the inequalities hold because $A' + 1/(\mu-\delta)A > 0$, $A'' > 0$, $\partial A/\partial \beta > 0$ and $\partial A'/\partial \beta < 0$.

Similarly, because the FOC must still be satisfied as α increases, we have

$$\frac{\partial}{\partial \delta}\left[w^2\left(A + \left(1 - \frac{\delta}{\mu}\right)\delta A'\right)\right]\frac{d\delta}{d\alpha} + w^2\frac{\partial}{\partial \alpha}\left[A + \left(1 - \frac{\delta}{\mu}\right)\delta A'\right] = 0.$$

Thus, we obtain

$$\frac{d\delta}{d\alpha} = -\frac{\frac{\partial}{\partial \alpha}\left[A + \left(1 - \frac{\delta}{\mu}\right)\delta A'\right]}{2A' + 2/(\mu - \delta)A + \delta(1 - \delta/\mu)A''}\ .$$

Using the preceding expression of $d\delta/d\alpha$, we have

$$\begin{aligned}
\frac{dP^*}{d\alpha} &= -\theta\left[\frac{\partial(wA)}{\partial \delta}\frac{d\delta}{d\alpha} + \frac{\partial(wA)}{\partial \alpha}\right] \\
&= -\theta w\left\{-\frac{A' + 1/(\mu-\delta)A}{2(A' + 1/(\mu-\delta)A) + \delta(1-\delta/\mu)A''}\frac{\partial[A + (1-\delta/\mu)\delta A']}{\partial \alpha} + \frac{\partial A}{\partial \alpha}\right\} \\
&> -\theta w\frac{1}{2}\frac{\partial[A - (1-\delta/\mu)\delta A']}{\partial \alpha} \\
&> 0,
\end{aligned}$$

where the inequalities hold because $A' + 1/(\mu-\delta)A > 0$, $A'' > 0$, $\frac{\partial[A + (1-\delta/\mu)\delta A']}{\partial \alpha} < 0$ and $\partial A'/\partial \alpha > 0$.

Proof of Proposition 3: A comparison of equation (6) and the FOC in the traditional setting in which reference effects are not considered, that is, $d\Pi^{trad}/d\delta = R - \theta w^2 \mu = 0$, shows that their only difference is the term

$$A(\delta) + \delta\left(1 - \frac{\delta}{\mu}\right)\frac{dA(\delta)}{d\delta} := f(\delta).$$

Note that $f(\delta)$ decreases in δ. Let $\hat{\delta} = \sup\{\delta : f(\delta) \geq 1\}$.

If $\delta^{trad} \leq \hat{\delta}$, we have $f(\delta^{trad}) \geq 1$, and then

$$\left.\frac{d\Pi}{d\delta}\right|_{\delta=\delta^{trad}} = R - \theta w^2(\delta^{trad})\mu f(\delta^{trad}) \leq R - \theta w^2(\delta^{trad})\mu = 0.$$

Because Π is concave in δ, we have the optimal joining probability $\delta^* \leq \delta^{trad}$. Morever

$$w^2(\delta^{trad}) = \frac{R}{\theta\mu} = w^2(\delta^*)f(\delta^*) < w^2(\delta^*)A(\delta^*),$$

implying that $w(\delta^{trad}) < w(\delta^*)A(\delta^*)$ given that $\delta^* \leq \delta^{trad}$. Thus,

$$P^* = R - \theta w(\delta^*)A(\delta^*) < R - \theta w(\delta^{trad}) = P^{trad}.$$

If $\delta^{trad} > \hat{\delta}$, we have $f(\delta^{trad}) < 1$ because $f(\delta)$ decreases in δ. Then,

$$\left.\frac{d\Pi}{d\delta}\right|_{\delta=\delta^{trad}} = R - \theta w^2(\delta^{trad})\mu f(\delta^{trad}) > R - \theta w^2(\delta^{trad})\mu = 0,$$

which implies that the optimal joining probability $\delta^* > \delta^{trad}$.

In addition, because $f(\delta)$ increases in β and decreases in α, according to the definition of $\hat{\delta}$, it also increases in β and decreases in α.

Proof of Proposition 4: To prove $\delta^* > \delta^s$, it suffices to prove that $d(SW - \Pi)/d\delta < 0$. We can easily show that

$$\begin{aligned}\frac{d(SW-\Pi)}{d\delta} &= \frac{d}{d\delta}\left(\frac{1}{2}\cdot\frac{\theta\delta\,(\delta-2)}{\mu-\delta}\cdot\frac{\alpha+\alpha\beta-2}{\alpha\delta-\delta+2}\right)\\ &= \frac{1}{2}\cdot\frac{\theta\left(\delta^2\alpha(\mu-2)-(\sqrt{\mu}(\delta-2))^2\right)}{(\mu-\delta)^2}\cdot\frac{\alpha+\alpha\beta-2}{(\alpha\delta-\delta+2)^2}.\end{aligned}$$

The sign of $d(SW - \Pi)/d\delta$ is fully determined by the sign of the term $\delta^2\alpha(\mu-2) - (\sqrt{\mu}(\delta - 2))^2$. Clearly, when $\mu \leq 2$, it is negative. Otherwise, we can rewrite it as $[\delta\sqrt{\alpha(\mu-2)} + \sqrt{\mu}(2 - \delta)][\delta\sqrt{\alpha(\mu-2)} - \sqrt{\mu}(2-\delta)]$. Thus, when $\mu > 2$, the sign depends on the range of δ: $\frac{d(SW-\Pi)}{d\delta} < 0$ when $\delta \in \left[0, \frac{2}{1+\sqrt{\alpha(1-2/\mu)}}\right]$ and $\frac{d(SW-\Pi)}{d\delta} > 0$ when $\delta \in \left[\frac{2}{1+\sqrt{\alpha(1-2/\mu)}}, 1\right]$. Clearly, as $\alpha < \frac{1}{1-2/\mu}$, $\frac{2}{1+\sqrt{\alpha(1-2/\mu)}} > 1$. Therefore, it is always true that $\frac{d(SW-\Pi)}{d\delta} < 0$.

Appendix D: Proofs of Lemmas and Propositions in the Duopoly Case

The organization of the proofs is as follows. We first establish the necessary conditions for the Nash equilibrium of the two servers' competition game. Then, we prove Proposition 5 for the fast server case.

To facilitate our analysis, we define a function

$$f(\delta_1, \delta_2) := \theta w_1 \left[\beta + 1 - (\beta - 1) \left(\frac{\delta_1}{2} + \delta_2 \frac{w_2}{w_1 + w_2} \right) \right].$$

Under the assumption of $P_1 \geq P_2$, the constraints (12) and (13) in server 1's optimization problem can be simplified as

$$\begin{aligned} U_1 = U_2 &\iff (P_1 - P_2)(2 - \delta_2 + \alpha\delta_2) = f(\delta_2, \delta_1) - f(\delta_1, \delta_2), \\ U_1 \geq U_b &\iff (R - P_1)(2 - \delta_1 - \delta_2 + \alpha\delta_1 + \alpha\delta_2) \geq f(\delta_1, \delta_2), \end{aligned}$$

which implies that

$$U_2 \geq U_b \iff (R - P_2)(2 - \delta_2 + \alpha\delta_2) + (R - P_1)(\alpha\delta_1 - \delta_1) \geq f(\delta_2, \delta_1).$$

One can simplify the constraints in a similar way for the case of $P_1 \leq P_2$.

Necessary Conditions for a Nash Equilibrium

First, assume $P_1 \geq P_2$. We start with the KKT conditions of the two servers' optimization problems. Based on server 1's optimization problem, (11) - (14), define the Lagrangian function of server 1 as

$$\begin{aligned} L_1(P_1, \delta_1, \delta_2) = \; & P_1\delta_1 + \eta_1 \left[(R - P_1)(2 - \delta_1 - \delta_2 + \alpha\delta_1 + \alpha\delta_2) - f(\delta_1, \delta_2) \right] \\ & + \eta_2 \left[(P_1 - P_2)(2 - \delta_2 + \alpha\delta_2) - f(\delta_2, \delta_1) + f(\delta_1, \delta_2) \right] \\ & - \eta_3(\delta_1 + \delta_2 - 1) - \eta_4(P_1 - \overline{P}) + \eta_5(P_1 - P_2), \end{aligned}$$

where $\eta_1, \eta_3, \eta_4, \eta_5 \geq 0$. By differentiating the Lagrangian function, the KKT conditions of server 1's optimization problem are

$$\begin{aligned} & \frac{\partial L_1}{\partial P_1} = \delta_1 - \eta_1(2 - \delta_1 - \delta_2 + \alpha\delta_1 + \alpha\delta_2) + \eta_2(2 - \delta_2 + \alpha\delta_2) - \eta_4 + \eta_5 = 0, \\ & \frac{\partial L_1}{\partial \delta_1} = P_1 + \eta_1 \left[(\alpha - 1)(R - P_1) - f_{\delta_1}(\delta_1, \delta_2) \right] + \eta_2 \left[f_{\delta_1}(\delta_1, \delta_2) - f_{\delta_1}(\delta_2, \delta_1) \right] - \eta_3 = 0, \\ & \frac{\partial L_1}{\partial \delta_2} = \eta_1 \left[(\alpha - 1)(R - P_1) - f_{\delta_2}(\delta_1, \delta_2) \right] + \eta_2 \left[(P_1 - P_2)(\alpha - 1) + f_{\delta_2}(\delta_1, \delta_2) - f_{\delta_2}(\delta_2, \delta_1) \right] - \eta_3 = 0, \\ & (P_1 - P_2)(2 - \delta_2 + \alpha\delta_2) = f(\delta_2, \delta_1) - f(\delta_1, \delta_2), \\ & (R - P_1)(2 - \delta_1 - \delta_2 + \alpha\delta_1 + \alpha\delta_2) \geq f(\delta_1, \delta_2), \end{aligned}$$

$$\delta_1+\delta_2\le 1,\ 0\le\delta_1<\mu_1, P_2<P_1\le\overline{P},$$
$$\eta_1\left[(R-P_1)(2-\delta_1-\delta_2+\alpha\delta_1+\alpha\delta_2)-f(\delta_1,\delta_2)\right]=0,$$
$$\eta_3(\delta_1+\delta_2-1)=0,$$
$$\eta_4(P_1-\overline{P})=0,\ \eta_5(P_1-P_2)=0,$$
$$\eta_1,\eta_3,\eta_4,\eta_5\ge 0.$$

Similarly, the Lagrangian function of server 2 can be written as

$$\begin{aligned}L_2(P_2,\delta_1,\delta_2)=\ & P_2\delta_2+\eta_1'\left[(R-P_2)(2-\delta_2+\alpha\delta_2)+(R-P_1)(\alpha\delta_1-\delta_1)-f(\delta_2,\delta_1)\right]\\ &+\eta_2'\left[(P_1-P_2)(2-\delta_2+\alpha\delta_2)-f(\delta_2,\delta_1)+f(\delta_1,\delta_2)\right]\\ &-\eta_3'(\delta_1+\delta_2-1)-\eta_4'(P_2-\overline{P})-\eta_5'(P_2-P_1).\end{aligned}$$

The corresponding KKT conditions are

$$\frac{\partial L_2}{\partial P_2}=\delta_2-\eta_1'(2-\delta_2+\alpha\delta_2)-\eta_2'(2-\delta_2+\alpha\delta_2)-\eta_4'-\eta_5'=0,$$
$$\frac{\partial L_2}{\partial \delta_1}=\eta_1'\left[(\alpha-1)(R-P_1)-f_{\delta_1}(\delta_2,\delta_1)\right]+\eta_2'\left[f_{\delta_1}(\delta_1,\delta_2)-f_{\delta_1}(\delta_2,\delta_1)\right]-\eta_3'=0,$$
$$\frac{\partial L_2}{\partial \delta_2}=P_2+\eta_1'\left[(\alpha-1)(R-P_2)-f_{\delta_2}(\delta_2,\delta_1)\right]+\eta_2'\left[(P_1-P_2)(\alpha-1)+f_{\delta_2}(\delta_1,\delta_2)-f_{\delta_2}(\delta_2,\delta_1)\right]-\eta_3'=0,$$
$$(P_1-P_2)(2-\delta_2+\alpha\delta_2)=f(\delta_2,\delta_1)-f(\delta_1,\delta_2),$$
$$(R-P_2)(2-\delta_2+\alpha\delta_2)+(R-P_1)(\alpha\delta_1-\delta_1)\ge f(\delta_2,\delta_1),$$
$$\delta_1+\delta_2\le 1,\ 0\le\delta_2<\mu_2, 0<P_2\le\min\{\overline{P},P_1\},$$
$$\eta_1'\left[(R-P_2)(2-\delta_2+\alpha\delta_2)+(R-P_1)(\alpha\delta_1-\delta_1)-f(\delta_2,\delta_1)\right]=0,$$
$$\eta_3'(\delta_1+\delta_2-1)=0,$$
$$\eta_4'(P_2-\overline{P})=0,\ \eta_5'(P_2-P_1)=0,$$
$$\eta_1',\eta_3',\eta_4',\eta_5'\ge 0.$$

Combining the two sets of KKT conditions gives us the necessary conditions for the Nash equilibrium when $P_1\ge P_2$. Analogously, one can write the necessary conditions for $P_1\le P_2$.

Proof of Proposition 5: In the fast server case, we study a highly competitive market, i.e., the two servers cover the entire market ($\delta_1+\delta_2=1$) and the customers enjoy a positive surplus compared with the balking option, i.e., $U_1=U_2>U_b$ (which also implies that $\eta_1=\eta_1'=0$). Assuming $P_1\ge P_2$, the necessary conditions can be simplified as

$$\delta_1+\eta_2(2-\delta_2+\alpha\delta_2)-\eta_4+\eta_5=0, \tag{20}$$
$$P_1+\eta_2\left[f_{\delta_1}(\delta_1,\delta_2)-f_{\delta_1}(\delta_2,\delta_1)\right]-\eta_3=0, \tag{21}$$

$$\eta_2\left[(P_1-P_2)(\alpha-1)+f_{\delta_2}(\delta_1,\delta_2)-f_{\delta_2}(\delta_2,\delta_1)\right]-\eta_3=0, \tag{22}$$

$$\delta_2-\eta_2'(2-\delta_2+\alpha\delta_2)-\eta_4'-\eta_5'=0, \tag{23}$$

$$\eta_2'\left[f_{\delta_1}(\delta_1,\delta_2)-f_{\delta_1}(\delta_2,\delta_1)\right]-\eta_3'=0, \tag{24}$$

$$P_2+\eta_2'\left[(P_1-P_2)(\alpha-1)+f_{\delta_2}(\delta_1,\delta_2)-f_{\delta_2}(\delta_2,\delta_1)\right]-\eta_3'=0, \tag{25}$$

$$(P_1-P_2)(2-\delta_2+\alpha\delta_2)=f(\delta_2,\delta_1)-f(\delta_1,\delta_2), \tag{26}$$

$$\delta_1+\delta_2-1=0, \tag{27}$$

$$\eta_4(P_1-\overline{P})=0,\ \eta_4'(P_2-\overline{P})=0, \tag{28}$$

$$\eta_5(P_1-P_2)=0,\ \eta_5'(P_2-P_1)=0, \tag{29}$$

where $\eta_3,\eta_4,\eta_5,\eta_3',\eta_4',\eta_5'\geq 0$.

For the symmetric equilibrium, we must have $P_1^e=P_2^e$, which, together with equations (26) and (27), implies that $\delta_1^e=\delta_2^e=1/2$. Conditions (21) and (22) imply that

$$P_1+\eta_2\left[f_{\delta_1}(\delta_1,\delta_2)-f_{\delta_1}(\delta_2,\delta_1)-f_{\delta_2}(\delta_1,\delta_2)+f_{\delta_2}(\delta_2,\delta_1)-(P_1-P_2)(\alpha-1)\right]=0, \tag{30}$$

where $\eta_2=-(\delta_1-\eta_4+\eta_5)/(2-\delta_2+\alpha\delta_2)$ according to condition (20). Similarly, conditions (24) and (25) imply that

$$P_2+\eta_2'\left[-f_{\delta_1}(\delta_1,\delta_2)+f_{\delta_1}(\delta_2,\delta_1)+f_{\delta_2}(\delta_1,\delta_2)-f_{\delta_2}(\delta_2,\delta_1)+(P_1-P_2)(\alpha-1)\right]=0, \tag{31}$$

where $\eta_2'=(\delta_2-\eta_4'-\eta_5')/(2-\delta_2+\alpha\delta_2)$ according to condition (23). From equations (30) and (31), we have $\eta_2=-\eta_2'$, which implies that $0\leq\eta_4-\eta_5=\eta_4'+\eta_5'<1/2$ and

$$P_1^e=P_2^e=\frac{3\beta+5}{2(\alpha+3)}\cdot\frac{\theta}{(\mu-0.5)^2}\cdot(1-2\eta_4+2\eta_5).$$

If $P_1^e=P_2^e<\overline{P}$, condition (28) gives us $\eta_4=\eta_4'=0$, which implies that $\eta_5=\eta_5'=0$. Thus, we have

$$P_1^e=P_2^e=\min\left\{\frac{3\beta+5}{2(\alpha+3)}\cdot\frac{\theta}{(\mu-0.5)^2},\ \overline{P}\right\}.$$

We then check whether the equilibrium solutions satisfy the following feasibility conditions:

$$U_1>U_b,\ \text{and}\ \eta_3,\eta_3'\geq 0.$$

The first condition is equivalent to $(R-P_1)(2-\delta_1-\delta_2+\alpha\delta_1+\alpha\delta_2)>f(\delta_1,\delta_2)$. To satisfy this condition, we need

$$P_1^e=P_2^e=\min\left\{\frac{3\beta+5}{2(\alpha+3)}\cdot\frac{\theta}{(\mu-0.5)^2},\ \overline{P}\right\}<R-\theta\frac{\beta+3}{2(\alpha+1)(\mu-0.5)}:=\widetilde{P}.$$

If $\mu > \overline{\mu}/2$, then $\overline{P} < \widetilde{P}$, which implies that $P_1^e < \widetilde{P}$; otherwise, if $\mu \le \overline{\mu}/2$, then $\overline{P} \ge \widetilde{P}$. Thus, $P_1^e < \widetilde{P}$ is satisfied if and only if $\mu > \widetilde{\mu}$. Therefore, the condition $U_1 > U_b$ is satisfied if and only if

$$\mu > \min\{\widetilde{\mu}, \overline{\mu}/2\} \ ,$$

where

$$\widetilde{\mu} = \frac{\sqrt{\left[(\beta+3)/(\alpha+1)\right]^2 + 8R/\theta \cdot (3\beta+5)/(\alpha+3)} + (\beta+3)/(\alpha+1)}{4R/\theta} + 0.5.$$

And then, from conditions (22) and (24), we have $\eta_3 = \eta_3' = (1-2\eta_4)(3\beta+5)/[4(\alpha+3)] > 0$.

It is worth noting that if $\widetilde{\mu} \le \overline{\mu}/2$, then for any $\mu > \widetilde{\mu}$, the equilibrium price is

$$P_1^e = P_2^e = \frac{3\beta+5}{2(\alpha+3)} \cdot \frac{\theta}{(\mu-0.5)^2}.$$

In summary, when $\mu > \min\{\widetilde{\mu}, \overline{\mu}/2\}$, we have found a unique symmetric solution that satisfies all the necessary conditions, specifically, when $U_1 > U_b$ and $\delta_1 + \delta_2 = 1$,

$$\delta_1^e = \delta_2^e = 1/2, \text{and } P_1^e = P_2^e = \min\left\{\frac{3\beta+5}{2(\alpha+3)} \cdot \frac{\theta}{(\mu-0.5)^2},\ \overline{P}\right\}.$$

To check whether this symmetric solution is the Nash equilibrium, we only need to check whether a server, say server 1, has the incentive to deviate from its price P_1^e, given that the other server sets its price at P_2^e. Note that server 1's problem has a unique solution satisfying the first order condition given server 2's price P_2^e. Furthermore, this solution leads to an objective value greater than the one at the boundary point ($\delta_1 = 0$ or $P_1 = 0$). Thus, it must be the best response of server 1. Hence, server 1 has no incentive to deviate from price P_1^e. The same reasoning applies to server 2. Therefore, the symmetric solution we find here is indeed a Nash equilibrium of the game.

We still need to check whether the corresponding customer equilibrium ($\delta_1^e = \delta_2^e = 1/2$) given the symmetric equilibrium price ($P_1^e = P_2^e$) is indeed a PPE. Recall that when the two servers charge the same price, the corresponding customer queueing game is reduced to be the one in the monopoly setting. The structure of the customer equilibrium is then similar to that stated in Proposition 1 except that the three thresholds in Proposition 1 $(\overline{\mu}, \underline{P}, \overline{P})$ are changed to be $(\overline{\mu}/2, \widetilde{P}, \overline{P})$ here. Hence, similar to Proposition 1, it is easy to show that a customer PPE is such that the two servers split the entire market, i.e., $\delta_1^{PPE} = \delta_2^{PPE} = 1/2$, if either of the following two conditions holds: 1) $1/2 < \mu < \overline{\mu}/2$ and $P \le \widetilde{P}$, 2) $\mu \ge \overline{\mu}/2$ and $P \le \overline{P}$. Therefore, we only need to check whether the current pricing equilibrium strategy satisfies these two sets of conditions. From the above equilibrium analysis, it is straightforward to see that

$$P_1^e = P_2^e < \widetilde{P},\ P_1^e = P_2^e \le \overline{P}, \text{ and } \mu > \min\{\widetilde{\mu}, \overline{\mu}/2\} > 1/2.$$

Therefore, the corresponding customer equilibrium is indeed a PPE.

So far we have found the necessary condition (i.e., $\mu > \min\{\widetilde{\mu}, \overline{\mu}/2\}$) for the existence of the symmetric equilibrium with $U_1 > U_b$ and $\delta_1 + \delta_2 = 1$. We still need to prove that it is also a sufficient condition, that is, when the condition $\mu > \min\{\widetilde{\mu}, \overline{\mu}/2\}$ holds, the symmetric equilibrium exists and must be the one with $U_1 > U_b$ and $\delta_1 + \delta_2 = 1$ rather than the one with $U_1 = U_b$ and $\delta_1 + \delta_2 = 1$ or the one with $U_1 = U_b$ and $\delta_1 + \delta_2 < 1$. The proof for the existence is easy because all conditions discussed above hold with the symmetric equilibrium solution. The only task left is to rule out the other two forms of symmetric equilibria. This has to be coupled with the proofs of Propositions 7 and 9. From the proof of Proposition 7, we know that the existence of a symmetric equilibrium with $U_1 = U_b$ and $\delta_1 + \delta_2 = 1$ requires the necessary condition $\min\{\underline{\mu}, \overline{\mu}/2\} \le \mu \le \min\{\widetilde{\mu}, \overline{\mu}/2\}$. From the proof of Proposition 9, we know that the symmetric equilibrium with $U_1 = U_b$ and $\delta_1 + \delta_2 < 1$ requires the necessary condition $\mu < \min\{\underline{\mu}, \overline{\mu}/2\}$. Consequently, the symmetric equilibrium with $U_1 > U_b$ and $\delta_1 + \delta_2 = 1$ is the only symmetric one in this case.

Proof of Proposition 7: In this case, we study a moderately competitive market, i.e., the two servers cover the entire market ($\delta_1 + \delta_2 = 1$) and the customers are indifferent between joining the service and balking, i.e., $U_1 = U_2 = U_b$. The necessary conditions under the assumption of $P_1 \ge P_2$ can then be simplified as

$$\delta_1 - \eta_1(\alpha+1) + \eta_2(2 - \delta_2 + \alpha\delta_2) - \eta_4 + \eta_5 = 0, \tag{32}$$

$$P_1 + \eta_1\left[(\alpha-1)(R-P_1) - f_{\delta_1}(\delta_1,\delta_2)\right] + \eta_2\left[f_{\delta_1}(\delta_1,\delta_2) - f_{\delta_1}(\delta_2,\delta_1)\right] - \eta_3 = 0, \tag{33}$$

$$\eta_1\left[(\alpha-1)(R-P_1) - f_{\delta_2}(\delta_1,\delta_2)\right] + \eta_2\left[(P_1-P_2)(\alpha-1) + f_{\delta_2}(\delta_1,\delta_2) - f_{\delta_2}(\delta_2,\delta_1)\right] - \eta_3 = 0, \tag{34}$$

$$\delta_2 - \eta_1'(2 - \delta_2 + \alpha\delta_2) - \eta_2'(2 - \delta_2 + \alpha\delta_2) - \eta_4' - \eta_5' = 0, \tag{35}$$

$$\eta_1'\left[(\alpha-1)(R-P_1) - f_{\delta_1}(\delta_2,\delta_1)\right] + \eta_2'\left[f_{\delta_1}(\delta_1,\delta_2) - f_{\delta_1}(\delta_2,\delta_1)\right] - \eta_3' = 0, \tag{36}$$

$$P_2 + \eta_1'\left[(\alpha-1)(R-P_2) - f_{\delta_2}(\delta_2,\delta_1)\right] + \eta_2'\left[(P_1-P_2)(\alpha-1) + f_{\delta_2}(\delta_1,\delta_2) - f_{\delta_2}(\delta_2,\delta_1)\right] - \eta_3' = 0, \tag{37}$$

$$(P_1 - P_2)(2 - \delta_2 + \alpha\delta_2) = f(\delta_2,\delta_1) - f(\delta_1,\delta_2), \tag{38}$$

$$(R - P_1)(\alpha+1) = f(\delta_1,\delta_2), \tag{39}$$

$$\delta_1 + \delta_2 = 1, \tag{40}$$

$$\eta_4(P_1 - \overline{P}) = 0,\ \eta_4'(P_2 - \overline{P}) = 0, \tag{41}$$

$$\eta_5(P_1 - P_2) = 0,\ \eta_5'(P_2 - P_1) = 0, \tag{42}$$

where $\eta_1, \eta_3, \eta_4, \eta_5, \eta_1', \eta_3', \eta_4', \eta_5' \ge 0$.

We focus on the symmetric solution, i.e. $P_1^e = P_2^e$. Then, conditions (38) and (40) imply that $\delta_1^e = \delta_2^e = 1/2$. Plugging this into condition (39) yields

$$P_1^e = P_2^e = R - \frac{\theta}{\mu - 0.5} \cdot \frac{\beta+3}{2(\alpha+1)} := \widetilde{P}. \tag{43}$$

To satisfy the feasibility condition of $0 < \mathrm{P}_1^e = P_2^e \le \overline{P}$, we need

$$\mu > \hat{\mu} := \frac{\theta}{R} \cdot \frac{\beta+3}{2(\alpha+1)} + 0.5 \text{ and } \mu \le \overline{\mu}/2.$$

It is easy to verify that $\hat{\mu} < \widetilde{\mu}$.

We need to further assure that the equilibrium solutions satisfy the feasibility conditions: $\eta_1, \eta_3, \eta_4, \eta_5, \eta_1', \eta_3', \eta_4', \eta_5' \ge 0$. We first examine the case with $\mu < \overline{\mu}/2$ and then check the case with $\mu = \overline{\mu}/2$.

When $\mu < \overline{\mu}/2$, we have $P_1^e = P_2^e < \overline{P}$. Thus, $\eta_4 = \eta_4' = 0$. From (36) and (37), we can solve η_2' as a function of η_1' as follows:

$$\eta_2' = -\frac{\eta_1'}{2} + \frac{2\widetilde{P}}{\theta w^2(3\beta+5)}.$$

Plugging the above equation into (35) and taking the feasibility condition $\eta_5' \ge 0$ into consideration, we get

$$\eta_1' \le \frac{2}{\alpha+3} - \frac{4\widetilde{P}}{\theta w^2(3\beta+5)}.$$

As $\eta_1' \ge 0$ is required, the RHS of the above inequality needs to be nonnegative, that is,

$$\frac{2}{\alpha+3} - \frac{4\widetilde{P}}{\theta w^2(3\beta+5)} \ge 0.$$

Plugging (43) into the above inequality yields

$$w^2\frac{3\beta+5}{\alpha+3} + w\frac{\beta+3}{\alpha+1} - \frac{2R}{\theta} \ge 0,$$

where $w = 1/(\mu - 0.5)$. It can be easily shown that the above condition is equivalent to $\mu \le \widetilde{\mu}$.

Next, we check the feasibility condition $\eta_3' \ge 0$. Adding (36) and (37) yields

$$\begin{aligned} 2\eta_3' &= \eta_1'\left[2(R-\widetilde{P})(\alpha-1) - f_{\delta_1}(\delta_2,\delta_1) - f_{\delta_2}(\delta_2,\delta_1)\right] + \widetilde{P} \\ &= \eta_1'\left[\theta w\frac{(\beta+3)(\alpha-1)}{\alpha+1} - \theta w^2\left(\beta+1-(\beta-1)\mu\right)\right] + \widetilde{P}. \end{aligned}$$

Since we can always set $\eta_1' = 0$, η_3' is nonnegative as the price $\widetilde{P}$ is nonnegative.

Similarly, from (33) and (34), we can solve η_2 as a function of η_1 as follows:

$$\eta_2 = \frac{\eta_1}{2} - \frac{2\widetilde{P}}{\theta w^2(3\beta+5)}.$$

Plugging this equation into (32) and taking into consideration the feasibility condition $\eta_5 \ge 0$, we get

$$\eta_1 \ge \frac{2}{3\alpha+1} - \frac{4\widetilde{P}(\alpha+3)}{\theta w^2(3\beta+5)(3\alpha+1)}. \tag{44}$$

It can be easily shown that if $\mu \le \widetilde{\mu}$, the RHS of (44) is always nonnegative.

We then check the feasibility condition $\eta_3 \geq 0$. Adding (33) and (34) yields

$$\begin{aligned} 2\eta_3 &= \eta_1 \left[2(R-\widetilde{P})(\alpha-1) - f_{\delta_1}(\delta_1,\delta_2) - f_{\delta_2}(\delta_1,\delta_2)\right] + \widetilde{P} \\ &= \eta_1 \left[\theta w \frac{(\beta+3)(\alpha-1)}{\alpha+1} - \theta w^2\left(\beta+1-(\beta-1)\mu\right)\right] + \widetilde{P}. \end{aligned}$$

Let

$$\tau(\mu) = (\beta+3)(\alpha-1)/(\alpha+1) - w\left(\beta+1-(\beta-1)\mu\right).$$

Thus, condition $\eta_3 \geq 0$ is equivalent to $\tau(\mu) \geq 0$, or $\tau(\mu) < 0$ and

$$\eta_1 \leq -\frac{\widetilde{P}}{\theta w \tau(\mu)}\ . \tag{45}$$

Together with condition (44), we need

$$\frac{\widetilde{P}}{\theta w^2}\left[\frac{4}{3\beta+5} - \frac{w}{\tau(\mu)}\cdot\frac{3\alpha+1}{\alpha+3}\right] \geq \frac{2}{\alpha+3}, \tag{46}$$

It can be shown that $\tau(\mu)$ increases in μ and $\tau(1/2) = -\infty$. Thus, the LHS of (46) increases in μ. Denote $\underline{\mu}$ as the service rate that solves

$$\frac{\widetilde{P}}{\theta w^2}\left[\frac{4}{3\beta+5} - \frac{w}{\tau(\mu)}\cdot\frac{3\alpha+1}{\alpha+3}\right] = \frac{2}{\alpha+3}, \tag{47}$$

then the condition $\eta_3 \geq 0$ is equivalent to $\mu \geq \underline{\mu}$. It is easy to verify that $\underline{\mu} > \hat{\mu}$, $\underline{\mu} < \widetilde{\mu}$ and $\underline{\mu} > 1/2$.

When $\mu = \overline{\mu}/2$, we need $\mu \leq \widetilde{\mu}$ to ensure that there exists a nonnegative η_1'. Then, we can find the corresponding $\eta_1, \eta_1', \eta_2, \eta_2', \eta_3, \eta_3', \eta_4, \eta_4', \eta_5, \eta_5'$ to satisfy the feasibility conditions $\eta_1, \eta_1', \eta_3, \eta_3', \eta_4, \eta_4', \eta_5, \eta_5' \geq 0$. For instance, let $\eta_1 = \eta_1' = 0,\ \eta_3 = \eta_3' = \widetilde{P}/2,\ \eta_5 = \eta_5' = 0$ and

$$\eta_2 = -\eta_2' = \frac{2\widetilde{P}}{\theta w^2(3\beta+5)},\ \eta_4 = \eta_4' = \frac{1}{2} - \frac{\widetilde{P}(3+\alpha)}{\theta w^2(3\beta+5)} \geq 0.$$

Then, they satisfy all the KKT conditions as well as the feasibility conditions. Therefore, if $\overline{\mu}/2 \leq \widetilde{\mu}$, all the feasibility conditions are satisfied.

In short, to satisfy all the feasibility conditions, we only need $\min\{\underline{\mu}, \overline{\mu}/2\} \leq \mu \leq \min\{\widetilde{\mu}, \overline{\mu}/2\}$.

To summarize, when $\min\{\underline{\mu}, \overline{\mu}/2\} \leq \mu \leq \min\{\widetilde{\mu}, \overline{\mu}/2\}$, we have found a unique symmetric solution that satisfies all the necessary conditions, specifically, when $U_1 = U_b$ and $\delta_1 + \delta_2 = 1$,

$$\delta_1^e = \delta_2^e = 1/2, \text{and } P_1^e = P_2^e = R - \frac{\theta}{\mu - 0.5}\cdot\frac{\beta+3}{2(\alpha+1)}.$$

It is easy to verify that this is also the unique symmetric solution when $P_1 \leq P_2$. Note that in this case, server 1's problem has a unique solution satisfying the first order condition for a given server 2's price P_2. Furthermore, this solution leads to an objective value greater than the one at the boundary point ($\delta_1 = 0$ or $\mathrm{P}_1 = 0$). Thus, it must be the optimal solution for server 1. The same

reasoning applies to server 2. Therefore, the symmetric solution we found here is indeed a Nash equilibrium of the game.

Again, we need to check whether the corresponding customer equilibrium is a PPE. Similar to that shown in the proof of Proposition 5, it requires that the equilibrium prices satisfy either of the following two conditions: 1) $1/2 < \mu < \overline{\mu}/2$ and $P \le \widetilde{P}$, 2) $\mu \ge \overline{\mu}/2$ and $P \le \overline{P}$. From the above equilibrium analysis, it is straightforward to see that

$$P_1^e = P_2^e = \widetilde{P} \le \overline{P} \text{ and } \mu \ge \min\{\underline{\mu}, \overline{\mu}/2\} > 1/2.$$

Hence, the corresponding customer equilibrium must be a PPE.

Using the same reasoning as that in the proof of Proposition 5, we can show that when $\min\{\underline{\mu}, \overline{\mu}/2\} \le \mu \le \min\{\widetilde{\mu}, \overline{\mu}/2\}$, the only symmetric Nash equilibrium is what we have found in this proof.

Proof of Proposition 9: In this case, we study a market with mild competition in which the two servers fail to cover the entire market ($\delta_1 + \delta_2 < 1$) (which implies that $\eta_3 = \eta_3' = 0$) and the customers are indifferent between joining the service and balking, i.e., $U_1 = U_2 = U_b$. The necessary conditions under the assumption of $P_1 \ge P_2$ can then be simplified as

$$\delta_1 - \eta_1(2 - \delta_1 - \delta_2 + \alpha\delta_1 + \alpha\delta_2) + \eta_2(2 - \delta_2 + \alpha\delta_2) - \eta_4 + \eta_5 = 0, \tag{48}$$

$$P_1 + \eta_1\left[(\alpha - 1)(R - P_1) - f_{\delta_1}(\delta_1, \delta_2)\right] + \eta_2\left[f_{\delta_1}(\delta_1, \delta_2) - f_{\delta_1}(\delta_2, \delta_1)\right] = 0, \tag{49}$$

$$\eta_1\left[(\alpha - 1)(R - P_1) - f_{\delta_2}(\delta_1, \delta_2)\right] + \eta_2\left[(P_1 - P_2)(\alpha - 1) + f_{\delta_2}(\delta_1, \delta_2) - f_{\delta_2}(\delta_2, \delta_1)\right] = 0, \tag{50}$$

$$\delta_2 - \eta_1'(2 - \delta_2 + \alpha\delta_2) - \eta_2'(2 - \delta_2 + \alpha\delta_2) - \eta_4' - \eta_5' = 0, \tag{51}$$

$$\eta_1'\left[(\alpha - 1)(R - P_1) - f_{\delta_1}(\delta_2, \delta_1)\right] + \eta_2'\left[f_{\delta_1}(\delta_1, \delta_2) - f_{\delta_1}(\delta_2, \delta_1)\right] = 0, \tag{52}$$

$$P_2 + \eta_1'\left[(\alpha - 1)(R - P_2) - f_{\delta_2}(\delta_2, \delta_1)\right] + \eta_2'\left[(P_1 - P_2)(\alpha - 1) + f_{\delta_2}(\delta_1, \delta_2) - f_{\delta_2}(\delta_2, \delta_1)\right] = 0, \tag{53}$$

$$(P_1 - P_2)(2 - \delta_2 + \alpha\delta_2) = f(\delta_2, \delta_1) - f(\delta_1, \delta_2), \tag{54}$$

$$(R - P_1)(2 - \delta_1 - \delta_2 + \alpha\delta_1 + \alpha\delta_2) = f(\delta_1, \delta_2), \tag{55}$$

$$\eta_4(P_1 - \overline{P}) = 0,\ \eta_4'(P_2 - \overline{P}) = 0, \tag{56}$$

$$\eta_5(P_1 - P_2) = 0,\ \eta_5'(P_2 - P_1) = 0, \tag{57}$$

where $\eta_1, \eta_1', \eta_4, \eta_4', \eta_5, \eta_5' \ge 0$.

We focus on the symmetric solution. Assume that $P_1^e = P_2^e = \widehat{P}$ and $\delta_1^e = \delta_2^e = \delta$. To ensure that $\delta < 1/2$ is a PPE under the price $\widehat{P} \le \overline{P}$, from Proposition 1, we need $\theta(\beta + 1)/(2R) < \mu < \overline{\mu}/2$.

We next solve the PPE δ from equations (48)-(57). From (55), we can write $\widehat{P}$ as a function of δ,

$$P_1^e = P_2^e = \widehat{P} = R - \theta w \frac{\beta + 1 - (\beta - 1)\delta}{2 + 2(\alpha - 1)\delta}. \tag{58}$$

From conditions (49) and (50), we can solve η_1 and η_2 as follows:

$$\eta_1 = \frac{\widehat{P}}{X-2Y} \text{ and } \eta_2 = \frac{Y}{X}\eta_1,$$

where

$$X = \theta w^2 \left(\beta + 1 - \frac{\beta-1}{2}\delta\right) > 0 \text{ and } Y = \theta w^2 \frac{\beta-1}{4}\delta + \theta w \frac{(\alpha-1)(\beta+1)+\beta-1}{2+2(\alpha-1)\delta} > 0. \tag{59}$$

Similarly, from conditions (52) and (53), we can solve η_1' and η_2'. It is easy to observe that

$$\eta_1' = \eta_1 \text{ and } \eta_2' = -\eta_2.$$

To satisfy the feasibility conditions on $\widehat{P}$, η_1 and η_1', we need

$$0 \le \widehat{P} \le \overline{P} \ , \ X - 2Y > 0.$$

We next show that $\widehat{P} \neq \overline{P}$, and we prove this by contradiction. If $\widehat{P} = \overline{P}$, together with $\mu < \overline{\mu}/2$, implies that $\delta = 0$ from Proposition 1. Plugging this into (51) yields

$$-2\eta_1 + 2\frac{Y}{X}\eta_1 - \eta_4' - \eta_5' = 0.$$

Since $\eta_1 > 0$ and $0 < Y/X < 1/2$, together with the feasibility conditions $\eta_4', \eta_5' \ge 0$, the above condition cannot be satisfied. Therefore, we have $\widehat{P} < \overline{P}$, which together with condition (56) imply that $\eta_4 = \eta_4' = 0$.

Since $\eta_4 = \eta_4' = 0$, and $\eta_5, \eta_5' \ge 0$, conditions (48) and (51) can be rewritten as

$$\delta - \frac{\widehat{P}}{X-2Y}\left\{[2+2(\alpha-1)\delta] - \frac{Y}{X}[2+(\alpha-1)\delta]\right\} \le 0 \ , \tag{60}$$

$$\delta - \frac{\widehat{P}}{X-2Y}\left(1 - \frac{Y}{X}\right)[2+(\alpha-1)\delta] \ge 0 \ . \tag{61}$$

Consequently, the symmetric equilibrium exists if and only if 1) the solution δ satisfies conditions (60), (61) and

$$0 \le \widehat{P} < \overline{P} \ , \ X - 2Y > 0, \tag{62}$$

and 2) the solution of the above three conditions ((60), (61) and (62)) satisfies $\delta < 1/2$.

Next, we simplify condition (62) by first showing that $\widehat{P}$ decreases in the PPE δ. Note that $w = 1/(\mu - \delta)$. The first order derivative of $\widehat{P}$ with respect to δ can be derived as

$$\frac{\partial \widehat{P}}{\partial \delta} = -\theta w^2 \frac{\beta+1-(\beta-1)\delta}{2[1+(\alpha-1)\delta]} + \theta w \frac{\alpha(\beta+1)-2}{2[1+(\alpha-1)\delta]^2} \tag{63}$$

$$= -\frac{\theta w^2}{2[1+(\alpha-1)\delta]^2} \underbrace{\{[\beta+1-(\beta-1)\delta][1+(\alpha-1)\delta] - (\mu-\delta)[\alpha(\beta+1)-2]\}}_{l(\delta)}. \tag{64}$$

Regarding $l(\delta)$, the term in the braces of (64), we can easily show that

$$\frac{\partial l(\delta)}{\partial \delta} = 2(\alpha-1)\left[\beta+1-(\beta-1)\delta\right] > 0,$$

which implies that $\widehat{P}$ is unimodal in δ. Furthermore, depending on the sign of $\partial\widehat{P}/\partial\delta$, we can have the following two situations.

Case 1: $\partial\widehat{P}/\partial\delta\Big|_{\delta=0} < 0$, which can be shown by simple algebra equivalent to

$$\mu < \frac{1}{\alpha - 2/(\beta+1)}.$$

Then, $\partial\widehat{P}/\partial\delta < 0$ for all $\delta > 0$. That is, $\widehat{P}$ decreases in δ. Note that $\widehat{P}\Big|_{\delta=0} = \overline{P}$. Thus, given a price $\widehat{P} < \overline{P}$, there exists a unique $\delta > 0$. Moreover, according to Proposition 1, this unique δ is the PPE.

Case 2: $\partial\widehat{P}/\partial\delta\Big|_{\delta=0} \geq 0$, which is equivalent to

$$\mu \geq \frac{1}{\alpha - 2/(\beta+1)}.$$

As $\partial l(\delta)/\partial\delta > 0$, $\widehat{P}$ first increases and then decreases in δ. Besides $\widehat{P}\Big|_{\delta=0} = \overline{P}$, there exists another value of δ, denoted by $\tilde{\delta}$ satisfying $\widehat{P}\Big|_{\delta=\tilde{\delta}} = \overline{P}$:

$$\tilde{\delta} = \frac{1}{\alpha-1}\left[\mu\left(\alpha - \frac{2}{\beta+1}\right) - 1\right]. \tag{65}$$

As $\mu < \overline{\mu}/2$, we have $\tilde{\delta} < 1/2$. Thus, $\widehat{P}$ is decreasing in δ when $\delta > \tilde{\delta}$. Given a price $\widehat{P} < \overline{P}$, there exists a unique $\delta > \tilde{\delta}$. Again, according to Proposition 1, this unique δ is the PPE.

Combining the above two cases, we have proved that $\widehat{P}$ decreases in the PPE δ, which implies that in condition (62), the requirement $0 \leq \widehat{P} < \overline{P}$ is equivalent to

$$\max\{0, \tilde{\delta}\} < \delta \leq \hat{\delta},$$

where $\hat{\delta}$ satisfies $\widehat{P}(\hat{\delta}) = 0$. Since $\widehat{P}$ decreases in the PPE δ, the requirement $X - 2Y > 0$ in condition (62) is automatically satisfied as

$$\begin{aligned} X - 2Y &= \theta w^2\left[\beta+1-(\beta-1)\delta\right] - \theta w\frac{\alpha(\beta+1)-2}{1+(\alpha-1)\delta} \\ &= -2\left[1+(\alpha-1)\delta\right]\frac{\partial\widehat{P}}{\partial\delta} > 0 \end{aligned}$$

for all $\max\{0, \tilde{\delta}\} < \delta \leq \hat{\delta}$. Therefore, condition (62) degenerates to

$$\max\{0, \tilde{\delta}\} < \delta \leq \hat{\delta}. \tag{66}$$

As to conditions (60) and (61), they can be rewritten as

$$\widehat{P}\left(\frac{1}{X-2Y}+\frac{1}{X}\right) \geq \frac{1-\frac{\widehat{P}}{X-2Y}(\alpha-1)}{1/\delta+(\alpha-1)/2},$$
$$\widehat{P}\left(\frac{1}{X-2Y}+\frac{1}{X}\right) \leq \frac{1}{1/\delta+(\alpha-1)/2}.$$

Because $\widehat{P}$ decreases in the PPE δ, the LHS term of the above two conditions decreases in the PPE δ while the RHS terms of the above two conditions increase in the PPE δ. Below, we prove that both $X-2Y$ and X increase in the PPE δ. Note that $w=1/(\mu-\delta)$. Thus,

$$\begin{aligned}\frac{\partial X}{\partial \delta} &= 2\theta w^3\left[\beta+1-\frac{\beta-1}{4}(\mu+\delta)\right] \\ &> 2\theta w^3\left[\beta+1-\frac{\beta-1}{4}\left(\frac{\beta+1}{\beta-1}+\frac{1}{2}\right)\right]>0,\end{aligned}$$

where the second inequality follows from $\mu<\overline{\mu}/2=\frac{\alpha+1}{2\alpha-4/(\beta+1)}\leq(\beta+1)/(\beta-1)$ and $\delta<1/2$. Recall that

$$X-2Y=\theta w^2\underbrace{\left[\beta+1-(\beta-1)\delta-\frac{1}{w}\cdot\frac{\alpha(\beta+1)-2}{1+(\alpha-1)\delta}\right]}_{k(\delta)}.$$

Regarding $k(\delta)$, the term in the square brackets, it can be shown that

$$\frac{\partial k(\delta)}{\partial \delta}=\frac{\alpha-1}{1+(\alpha-1)\delta}\left[\beta+1-(\beta-1)\delta+(\mu-\delta)\cdot\frac{\alpha(\beta+1)-2}{1+(\alpha-1)\delta}\right]>0.$$

This together with the fact that w increases in δ imply that $X-2Y$ increases in δ. Let $\underline{\delta}$ and $\overline{\delta}$ solve the following two equations, respectively:

$$\left[\widehat{P}\left(\frac{1}{X-2Y}+\frac{1}{X}\right)-\frac{1}{2/\delta+(\alpha-1)/2}\right]\Bigg|_{\delta=\underline{\delta}}=0, \tag{67}$$

$$\left[\widehat{P}\left(\frac{1}{X-2Y}+\frac{1}{X}\right)-\frac{1-\frac{\widehat{P}}{X-2Y}(\alpha-1)}{1/\delta+(\alpha-1)/2}\right]\Bigg|_{\delta=\overline{\delta}}=0. \tag{68}$$

Then, any $\delta\in[\underline{\delta},\overline{\delta}]$ satisfies conditions (60) and (61).

Based on the above analysis, we can conclude that the symmetric equilibrium exists if 1) the set $\delta\in\left[\underline{\delta},\overline{\delta}\right]\cap\left(\max\{0,\tilde{\delta}\},\hat{\delta}\right]$ is nonempty and 2) any $\delta\in\left[\underline{\delta},\overline{\delta}\right]\cap\left(\max\{0,\tilde{\delta}\},\hat{\delta}\right]$ satisfies $\delta<1/2$.

We now consider the aforementioned two cases:

$$\text{Case 1: } \mu<\frac{1}{\alpha-2/(\beta+1)} \text{ and Case 2: } \mu\geq\frac{1}{\alpha-2/(\beta+1)},$$

under which $\max\{0,\tilde{\delta}\}$ takes different values.

In the first case,

$$\max\{0,\tilde{\delta}\}=\max\left\{0,\frac{1}{\alpha-1}\left[\mu\left(\alpha-\frac{2}{\beta+1}\right)-1\right]\right\}=0.$$

Thus, the symmetric equilibrium exists if the following two conditions are satisfied: 1) the set $\delta \in [\underline{\delta}, \overline{\delta}] \cap (0, \hat{\delta}]$ is nonempty, and 2) any $\delta \in [\underline{\delta}, \overline{\delta}] \cap (0, \hat{\delta}]$ satisfies $\delta < 1/2$. We next prove that $\underline{\delta} > 0$ and $\overline{\delta} < \hat{\delta}$. It is easy to show that

$$\left[\widehat{P}\left(\frac{1}{X-2Y}+\frac{1}{X}\right)-\frac{1}{2/\delta+(\alpha-1)/2}\right]\Bigg|_{\delta=0} = \left[\widehat{P}\left(\frac{1}{X-2Y}+\frac{1}{X}\right)\right]\Bigg|_{\delta=0} > 0,$$
$$\left[\widehat{P}\left(\frac{1}{X-2Y}+\frac{1}{X}\right)-\frac{1-\frac{\widehat{P}}{X-2Y}(\alpha-1)}{1/\delta+(\alpha-1)/2}\right]\Bigg|_{\delta=\hat{\delta}} = -\frac{1}{1/\hat{\delta}+(\alpha-1)/2} < 0.$$

Comparing them with equations (67) and (68), we have $\underline{\delta} > 0$ and $\overline{\delta} < \hat{\delta}$. Thus, condition 1) is satisfied. To further satisfy condition 2), we need $\overline{\delta} < 1/2$, which is equivalent to

$$\left[\widehat{P}\left(\frac{1}{X-2Y}+\frac{1}{X}\right)-\frac{1-\frac{\widehat{P}}{X-2Y}(\alpha-1)}{1/\delta+(\alpha-1)/2}\right]\Bigg|_{\delta=\frac{1}{2}} < 0.$$

This can be rewritten as

$$\frac{\widehat{P}}{\theta w^2}\left[\frac{4}{3\beta+5}-\frac{w}{(\beta+3)(\alpha-1)/(\alpha+1)-w\left[\beta+1-(\beta-1)\mu\right]}\cdot\frac{3\alpha+1}{\alpha+3}\right] < \frac{2}{\alpha+3}.$$

Comparing it with (47), we can see that condition $\overline{\delta} < 1/2$ is equivalent to $\mu < \underline{\mu}$. Thus, in Case 1, if $\mu < \underline{\mu}$, the symmetric equilibrium exists and any δ^e that falls into the interval $\delta^e \in \left[\underline{\delta},\ \overline{\delta}\right]$ is an equilibrium.

Now we move to the second case:

$$\mu \geq \frac{1}{\alpha - 2/(\beta+1)},$$

under which, we have $\max\{0, \tilde{\delta}\} = \tilde{\delta}$. Thus, the symmetric equilibrium exists if the following two conditions are satisfied: 1) the set $\delta \in [\underline{\delta}, \overline{\delta}] \cap (\tilde{\delta}, \hat{\delta}]$ is nonempty, and 2) any $\delta \in [\underline{\delta}, \overline{\delta}] \cap (\tilde{\delta}, \hat{\delta}]$ satisfies $\delta < 1/2$. We have proved $\overline{\delta} < \hat{\delta}$ in the previous case, then the first condition requires $\overline{\delta} > \tilde{\delta}$, which is equivalent to

$$h(\alpha, \beta, \mu, R) := \left[\delta - \frac{\widehat{P}(\alpha-1)\delta}{X-2Y} - \frac{2+(\alpha-1)\delta}{2}\left(\frac{\widehat{P}}{X-2Y}+\frac{\widehat{P}}{X}\right)\right]\Bigg|_{\delta=\tilde{\delta}} < 0. \tag{69}$$

The second condition requires that $\overline{\delta} < 1/2$, which is equivalent to $\mu < \underline{\mu}$ as been shown in the previous case. To conclude, in Case 2, if $\mu < \underline{\mu}$ and the condition (69) is satisfied, then the symmetric equilibrium exists and lies in the interval

$$\delta^e \in \left[\underline{\delta},\ \overline{\delta}\right] \cap \left(\tilde{\delta}, \overline{\delta}\right].$$

It can be easily shown that when $\alpha = \beta = 1$, $\delta^e = \underline{\delta} = \overline{\delta} = \delta^{trad}$, and hence $P^e = P^{trad}$. Thus, there exists a unique symmetric equilibrium in which each server charges its monopoly price, same as that under the traditional case without considering the reference effect.

We next prove that when there exists a continuum of equilibria, the one with the highest joining probability is the Pareto optimal equilibrium. We prove this result by showing that the firm's profit $\pi = \widehat{P}\delta$ increases in δ within the interval $\delta \in \left[\underline{\delta},\ \overline{\delta}\right]$. Taking the first order derivative of π with respect to δ, we get

$$\begin{aligned}\frac{\partial \pi}{\partial \delta} &= \widehat{P} + \frac{\partial \widehat{P}}{\partial \delta}\delta \\ &= \widehat{P} - \frac{X-2Y}{2+2(\alpha-1)\delta}\delta \\ &= \frac{X-2Y}{2+2(\alpha-1)\delta}\left[\frac{\widehat{P}}{X-2Y}\left(2+2(\alpha-1)\delta\right)-\delta\right] \\ &> 0\end{aligned}$$

for all $\delta \in \left[\underline{\delta},\ \overline{\delta}\right]$, where the last inequality is due to that any equilibrium δ in this interval should satisfy condition (60).

So far, we have found the necessary conditions for the existence of the symmetric equilibrium with $U_1 = U_b$ and $\delta_1 + \delta_2 < 1$. We still need to prove that it is also a sufficient condition. Using the same reasoning as that in the proof of Proposition 5, we can show that when $\theta(\beta+1)/(2R) < \mu < \min\{\underline{\mu}, \overline{\mu}/2\}$, the only symmetric Nash equilibria are what we have found in this proof.

Proof of Corollary 1: Plugging $\delta = \tilde{\delta}$ into $\widehat{P}$, $X-2Y$ and X, we get

$$\begin{aligned}\widehat{P}(\delta=\tilde{\delta}) &= R - \frac{\theta(\beta+1)}{2\mu}, \\ (X-2Y)|_{\delta=\tilde{\delta}} &= \theta\frac{\tilde{\delta}}{\mu(\mu-\tilde{\delta})}(\beta+1)(\alpha-1), \\ X(\tilde{\delta}) &= \theta\frac{\beta+1}{2(\mu-\tilde{\delta})^2} + \theta\frac{(\beta+1)[1+(\alpha-1)\tilde{\delta}]}{2\mu(\mu-\tilde{\delta})}.\end{aligned}$$

It is easy to show that $\tilde{\delta}$ increases in α and β. Thus, from the above equations, we can easily observe that both $X-2Y$ and X increase in α and β while $\widehat{P}$ decreases in β and is constant in α.

Plugging the term $X-2Y$ into

$$h(\alpha,\beta,\mu,R) := \left[\delta - \frac{\widehat{P}(\alpha-1)\delta}{X-2Y} - \frac{2+(\alpha-1)\delta}{2}\left(\frac{\widehat{P}}{X-2Y}+\frac{\widehat{P}}{X}\right)\right]\Bigg|_{\delta=\tilde{\delta}},$$

we get

$$\frac{\widehat{P}(\alpha-1)\tilde{\delta}}{X-2Y} = \frac{\widehat{P}\mu(\mu-\tilde{\delta})}{\theta(\beta+1)} \text{ and } \frac{2+(\alpha-1)\tilde{\delta}}{2}\cdot\frac{\widehat{P}}{X-2Y} = \frac{\widehat{P}\mu(\mu-\tilde{\delta})}{\theta(\beta+1)}\cdot\frac{2+(\alpha-1)\tilde{\delta}}{2(\alpha-1)\tilde{\delta}}.$$

We can see that both of the above two terms decrease in α and β.

We next show that the term $[2+(\alpha-1)\delta]/(2X)$ decreases in α and β. Since

$$\frac{\partial X}{\partial \alpha} = \theta w^3\left[2(\beta+1) - \frac{\beta-1}{2}(\mu+\tilde{\delta})\right]\frac{\partial \tilde{\delta}}{\partial \alpha},$$

we have

$$\frac{\partial[2+(\alpha-1)\delta]/(2X)}{\partial\alpha} < \frac{\theta w^2\mu}{2X^2}\left[\frac{\beta-1}{2}\mu-(\beta+1)\right] < 0.$$

Since

$$\frac{\partial X}{\partial\beta} = \theta w^3\left[2(\beta+1)-\frac{\beta-1}{2}(\mu+\tilde{\delta})\right]\frac{\partial\tilde{\delta}}{\partial\beta}+\theta w^2\left(1-\frac{\tilde{\delta}}{2}\right),$$

we have

$$\frac{\partial[2+(\alpha-1)\delta]/(2X)}{\partial\beta} < \frac{\theta w^2\mu}{X^2(\beta+1)^2}\left[\frac{\beta-1}{2}\mu-(\beta+1)\right] < 0\ .$$

Therefore, the function $h(\alpha,\beta,\mu,R)$ increases in α and β.

Appendix E: Results under the Traditional Duopoly Price Competition

Using the same method as that in Appendix D, we can derive the equilibrium results for the duopoly price competition without considering the customer loss aversion preference. The procedure is similar to that in Chen and Wan (2003) except that in our case, the three regions corresponding to the three possible forms of equilibrium are defined in terms of the service capacity instead of the demand rate. Thus, we can simply make small changes over their findings to derive the results. We report our results as follows.

Assume the two servers are symmetric, i.e., $\mu_1=\mu_2=\mu$. Define the two threshold levels $\underline{\mu}^t$ and $\overline{\mu}^t$ as

$$\underline{\mu}^t = \frac{\sqrt{1+2R/\theta}+1}{2R/\theta}+0.5 \text{ and } \overline{\mu}^t = \frac{\sqrt{1+4R/\theta}+1}{2R/\theta}+0.5\ .$$

Then, under our setting, the fast server case corresponds to the case when $\mu>\overline{\mu}^t$, the moderate-speed server case corresponds to the case when $\underline{\mu}^t\le\mu\le\overline{\mu}^t$, and the slow server case corresponds to the case when $\mu<\underline{\mu}^t$. The corresponding equilibrium results are listed as follows.

- Fast server ($\mu>\overline{\mu}^t$)

There exists a unique symmetric equilibrium given by $P_1^{trad}=P_2^{trad}=\theta/(\mu-0.5)^2$, and the equilibrium demand rate $\delta_1^{trad}=\delta_2^{trad}=0.5$.

- Moderate-speed server ($\underline{\mu}^t\le\mu\le\overline{\mu}^t$)

There exists a unique symmetric equilibrium with the equilibrium price $P_1^{trad}=P_2^{trad}=R-\theta/(\mu-0.5)$ and the demand rate $\delta_1^{trad}=\delta_2^{trad}=0.5$.

- Slow server ($\mu<\underline{\mu}^t$)

There exists a unique equilibrium such that each server charges its own monopoly price $P_1^{trad}=P_2^{trad}=R-\sqrt{\theta R/\mu}$, and the equilibrium demand rate $\delta_1^{trad}=\delta_2^{trad}=\mu-\sqrt{\theta\mu/R}$.