Customer differentiation with shipping as an ancillary service? free service, prioritization, and strategic delay

Sainathan, Arvind

2017

Sainathan, A. (2018). Customer differentiation with shipping as an ancillary service? free service, prioritization, and strategic delay. Decision Sciences, 49(4), 690-727. doi:10.1111/deci.12285

https://hdl.handle.net/10356/106422

https://doi.org/10.1111/deci.12285

Customer Differentiation with Shipping as an Ancillary Service? Free Service, Prioritization, and Strategic Delay

Arvind Sainathan

Nanyang Business School, Nanyang Technological University Singapore, ASAINATHAN@ntu.edu.sg

August 17, 2017 (*Forthcoming in Decision Sciences*)

A service provider/retailer offers *ancillary* service (e.g. shipping by an online retailer) to two types of customers, impatient and patient, who may be heterogeneous both in their delay sensitivities and service valuations. She can use *prioritization* and/or *strategic delay* to differentiate them by offering two service classes and charging different prices, potentially resulting in a *split* in which a single customer type selects both the classes. Her objective is to minimize cost while satisfying individual rationality and incentive compatibility conditions. We characterize the optimal solutions under both exogenous and endogenous capacities. We examine the conditions under which the following strategically important features of service delivery are optimal, and relate them to practical scenarios: (i) free service, (ii) single/differentiated service, (iii) split of customers, and (iv) strategic delay. We find that the presence of these features depends on (i) whether the retailer has *limited or sufficient* capacity and (ii) whether she sells *fashion goods or staple products*. A typical explanation for offering free service is that it increases demand from customers. We make an *operational case* for it by showing that even if demand does not change, free service is still optimal under some scenarios.

Keywords: service operations, heterogeneity, prioritization, e-Commerce, free-shipping

1. Introduction

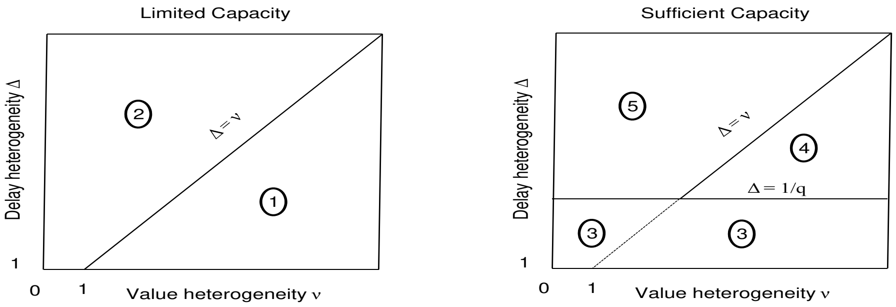
We consider a service provider (SP) who offers an *ancillary* service to her customers. We define a certain service to be an *ancillary service* if it satisfies the following two aspects: first, it is not a part of the primary services (or products) that the SP sells to its customers even though it provides value to these customers; second, the provision of this service is managed as a *cost center*, i.e., all the demand from customers needs to be satisfied while minimizing the cost. As we discuss below, shipping of products by online retailers is an example of such an ancillary service, and it is our main motivating example for the paper. Although ancillary services are costly, they are sometimes offered for free. We recognize that customers of such services may be heterogeneous both in how much they value them (*value heterogeneity*) and their *delay sensitivity*, the rate at which the valuation decreases over time (*delay heterogeneity*). Such heterogeneity is commonly observed in contexts involving service operations (e.g., see Afeche (2013), Allon and Federgruen (2009)). The SP can differentiate the customers using two simple mechanisms: (i) *prioritization* in which she prioritizes providing the service to some customers and (ii) *strategic delay* in which she deliberately delays it to some of them. Strategic delay has been considered in the past in the service operations management literature[1]. We investigate the impact of (value and delay) heterogeneity and SP's capacity on the optimality of free service and how she delivers the service and differentiates the customers. Next, we discuss about shipping by online retailers, how it can be considered as an ancillary service, and different shipping policies.

Shipping is a key business component for most online retailers. While economies of scale and the lack of maintenance and operating costs associated with "brick and mortar" retail stores help them, shipping costs them money. For instance, Amazon.com made a loss of about 3.5 billion US dollars due to shipping costs in 2013 (Forbes 2014). Shipping is an important factor not only for retailers but also for customers. An important reason why more customers are purchasing their products online is because they can get it shipped conveniently to the comfort of their homes; however, delays in shipping the orders can have adverse impacts on customer satisfaction. A vital question for such retailers involves the shipping policy that is used and how the shipping service is delivered. For instance, Zappos.com offers free-shipping to all customers (Zappos.com 2014). Amazon.com, its parent company, has a "standard" paid-shipping option and a "free" shipping option (for products/orders that qualify). Furthermore, standard shipping typically takes 3-5 business days while free shipping takes 5-8 business days (Amazon.com 2014), suggesting that customers who choose different options are heterogeneous in their delay sensitivity and how much they value shipping. Overstock.com, one of Amazon's competitors, used to charge a flat rate for standard shipping earlier but now employs a different strategy by charging it only for smaller orders and shipping other ones for free (Overstock.com 2014, Investor Press Release 2013).

We use the term *retailer* to denote the SP in the rest of the paper because the main application of our research involves online retailers and their shipping services. In the paper, we consider two types of customers, impatient and patient, with different delay sensitivities. We are primarily interested in answering the following questions concerning key service

[1] For instance, see Barnes and Mookherjee (2009), Afeche (2013), and Maglaras et al. (2013)

delivery features: (i) whether free service is provided, (ii) whether just a single service is provided or service is differentiated, (iii) whether there is any strategic delay, and (iv) whether there is a *split* in which a customer type selects both service classes[2]. We focus on these features due to their strategic importance to the retailer. Based on our analyses in the paper, we find that the answers depend on two crucial factors: (i) whether the retailer has *limited* or *sufficient* capacity and (ii) whether the retailer sells *fashion/branded goods* or she sells *staple/regular products.*



	Retailer Type (Capacity)	Product Type	Impatient Fraction q	Single Service	Free Service	Prioritized Service	Strategic Delay	Split
1	Limited	Fashion	No impact	No	Yes	Yes	No	Yes
2	Limited	Staple	No impact	Yes	Maybe	No	No	No
3	Sufficient	Fashion/ Staple	Low	No	No	Yes	No	No
4	Sufficient	Fashion	High	No	Yes	Yes	Yes	No
5	Sufficient	Staple	High	No	No	Yes	Yes	No

Figure 1: Optimal service delivery features under different cases

We classify a retailer as *limited* or *sufficient* based on her capacity (relative to her demand) for providing service. Retailer's capacity might be limited either because (i) her customer base is small so that she does not have economies of scale, or (ii) she has not yet made sufficient investment on the capacity. It results in a high waiting time for customers. On the other hand, a retailer with sufficient capacity can ensure that customers' waiting time is low. We refer to a product as a *fashion good* if its customers have a higher value heterogeneity than delay heterogeneity. Customers often have different tastes regarding such products (Jain and Paul 2001), which result in a higher variation of their product valuations than their delay sensitivities. Significant price changes for fashion goods (Pashigian 1988) also imply that its customers have a high value heterogeneity. A *staple product*, on

[2] We show that only impatient customers can possibly split, i.e., select both high-type and low-type service class, under optimality (see §6).

the other hand, is a product whose customers have a higher delay heterogeneity than value heterogeneity. There is little variation in how much different customers value such products; however, some customers can be quite delay-sensitive, and they expect their items to be delivered quickly. One reason for that can be because technology, with the resulting ubiquity of e-Commerce, is making some customers more impatient (Muther 2013).

Figure 1 provides a brief summary from the results of our analyses in §'s 6-8 and their implications. There are five main cases, as shown in the plots, that result in different service delivery features. In the first case, the retailer has limited capacity and she sells fashion goods ($\Delta < \nu$). She prioritizes her service but since the capacity is low, she has to provide free service, and impatient customers split. It is also the only case in which customers split. In the second case[3], the retailer still has limited capacity but she sells staple goods ($\Delta > \nu$). The retailer then does not prioritize, and she offers just a single service class. The other cases pertain to a retailer with sufficient capacity. The service delivery features, unlike for a limited capacity retailer, also depend on the fraction of impatient customers q. If this value is low (so that $q < 1/\Delta \Leftrightarrow \Delta < 1/q$), then the retailer prioritizes and does not provide free service, *regardless of whether she sells fashion goods/staple products.* If it is high, then she prioritizes and also uses strategic delay. Furthermore, she offers free service if she sells fashion goods, but she does not offer it if she sells staple products.

We make an important observation regarding free service (free-shipping) and how we model it in the paper. An oft-cited explanation for the prevalence of free-shipping is that it makes customers happy and increases demand either due to behavioral or economic/rational reasons involving customer behavior. *However, we show that even if this demand premium is ignored and there is no such increase in demand, free-shipping can still be optimal under some scenarios due to operational factors such as capacity and service delivery*[4]. We thereby make an *operational case* for why and when free-shipping can be optimal, which has not been done in the prior literature to the best of our knowledge.

2. Literature Review

This paper is related to three streams of literature: (i) queuing optimization which looks at pricing and other operational decisions for service classes in the context of customers who

[3] In this case, free service occurs only if $W(\mu) = v_1/\eta_1$ (see §3 for details regarding the notation).

[4] Incorporating this demand premium, which is already understood in the literature, will also lead to confounding and difficulty in separating the two types of benefits of free-shipping.

may be heterogeneous and self-select, (ii) free service provision (including free shipping) and why it may be beneficial, and (iii) vertically differentiated product variants with different quality.

Hassin and Haviv (2003) provide a comprehensive review of earlier literature on the first stream, we focus here on the more recent and relevant papers. Two papers from this stream come close to our research: Afeche (2013) and Katta and Sethuraman (2005). As in this paper, they consider (i) heterogeneity among customers in both their service valuations and delay sensitivities, and (ii) individual rationality (IR) and incentive compatibility (IC) conditions. They also analyze the optimization problem from the service provider's/retailer's view and maximize her revenue. However, they do not model ancillary services for which *the arrival rate/demand cannot be optimized and all the customers are satisfied at the minimum cost*. We only consider prioritization and strategic delay to differentiate customers because these policies are simple and easy to implement; however, we allow for the *splitting* of a customer type between two service classes. These aspects fundamentally change the formulation of the retailer's optimization problem and also lead to some key differences in the results (see §'s 4-6). Some research papers have considered pricing/service decisions that satisfy IC conditions in other contexts. Lederer and Li (1997) show that a competitive equilibrium results in incentive compatible pricing. Rao and Peterson (1998) study the optimal pricing of priority services in a static service facility with n customers that maximize their own profits. Van Mieghem (2000) and Hsu et al. (2009) consider socially optimal pricing and scheduling. Afeche and Mendelson (2004) consider pricing and priority auctions under a generalized delay cost. Zhang et al. (2007) model the pricing of communication services with delay guarantees in the presence of customers that are heterogeneous in their service valuations but homogeneous in their delay sensitivity. Zhao et al. (2012) analyze whether a firm should have uniform or differentiated price and lead time quotations in the presence of customers who are heterogeneous in their product valuations and delay cost rates. Li et al. (2012) consider competition between two service providers with *naive* customers who select them based on their prices and queue sizes. And Afeche et al. (2013) analyze how to price time-sensitive services based on realized lead times in the presence of customers who may be risk-averse. Other papers have considered similar decisions in other scenarios but without explicitly incorporating IC constraints (e.g., see So and Song (1998), Boyaci and Ray (2003), Maglaras and Zeevi (2003), Ray and Jewkes (2004), Allon and Federgruen (2009), Jayaswal et al. (2011), and Anand et al. (2011)).

There has been no prior research, to the best of our knowledge, in the queuing optimization literature, which (i) requires that all the demand from customers be satisfied (a key feature of ancillary services which we model in this paper), (ii) analyzes free service and when it can be beneficial operationally, and (iii) allows for splitting of a customer type between different service classes.

Because shipping is a key example for ancillary services in this paper, we next discuss about some articles and research papers that have considered free-shipping and its benefits. Customers attach a high level of importance to shipping charges when they shop with online retailers; e.g., in one survey, 93% of customers said they will buy more products online if shipping is free (eMC 2014) and in another, higher than expected shipping costs is a major reason for shipping cart abandonment (PayPal 2009). Some research from the marketing literature has analyzed whether the retailer would be better off by having the shipping free but instead inflating the base price of the product. They find that it is sometimes better to charge separately for shipping because customers perceive (and recall) a lower total price (Morwitz et al. 1998), but at other times, e.g., for books purchases, customers are much more sensitive to shipping prices than the price of product itself which makes it better to incorporate the shipping charge into the base price (Hamilton et al. 2010). For a review of research on price partitioning, we refer the reader to Morwitz et al. (2014). Gümüş et al. (2013) characterize the market equilibrium under competition among retailers who either charge shipping prices or offer free-shipping. Some papers (e.g., see Shampanier et al. (2007), Ariely (2008), and Kannan et al. (2009)) have reasoned the prevalence of free-shipping (and free provision of other services/products) by showing that zero price is a *special* price which can attract consumers and change their behavior. Leng and Parlar (2005) and Leng and Becerril-Arreola (2010) analyze contingent free-shipping (CFS), in which the retailer offers free -shipping only beyond a certain threshold order size, under B2B and B2C transactions respectively. The retailer benefits from CFS in both the papers because customers purchase more or the probability of repurchase increase. They do not model the time-sensitive nature of shipping services.

All the research above considers the benefit of free-shipping from changing customer behavior, either through purchase of more products/services, attraction to it due to the *special zero price* and willingness to pay a higher base price, or having a competitive advantage over retailers who do not offer it. In this regard, they analyze its benefits primarily from a *marketing perspective*. Although we recognize its key marketing implications, we take a different approach. We analyze its benefits primarily from an *operational perspective* by considering shipping as an ancillary service in which (i) the retailer is constrained by her capacity and minimizes the cost while satisfying all the demand, and (ii) the customers may be heterogeneous in their valuations and delay sensitivities.

This work is also related to research on product differentiation in which the product variants have different quality. Examples of such research can be found in Anderson et al. (1992), Moorthy (1984), and some references contained therein. The price(s) and quality measure(s) of the product variant(s) are akin to the price(s) and waiting time(s) for the service class(es) we model here[5]. However, there are two main differences. First, there is a negative externality which increases the waiting time as more customers purchase the service, while quality of a product is usually unaffected by its demand. Second, we consider an ancillary service in which the retailer has to provide adequate service to all her customers, while a monopolist selling differentiated product variants can partially satisfy the market.

3. Model

Customer arrivals for the ancillary service follow a Poisson process with rate λ. The retailer can provide at most two *classes* of service, h and l. She has two options to differentiate these service classes: *prioritization* and *strategic delay*. She may or may not prioritize, and she can strategically delay either service class. We focus our analysis on two service classes and these two differentiation options because they are relatively simple and easy to implement. Under prioritization, we refer to high priority and low priority service classes as *high-type* service (class h) and *low-type* service (class l) respectively. If there is no prioritization, strategic delay is the only way to differentiate, and we refer to the service class experiencing a lower (higher) delay as high-type (low-type) service. For each customer, the retailer charges prices of p_h and p_l ($p_h, p_l \geq 0$) for high- and low-type services respectively, and she incurs a marginal cost of $r > 0$. Strategic delay for high- and low-type services are denoted by d_h and d_l respectively ($d_h, d_l \geq 0$).

We do not consider the price of the primary product/service in our main analysis of the optimization problem involving ancillary service for the following reasons. First, firms typically set similar shipping policies for multiple ranges of products that may have very different (product) prices. It then becomes difficult to optimize each one of them in conjunction with

[5] The author thanks an anonymous reviewer for suggesting this connection.

the shipping price(s). Second, behavioral aspects become important and customers may not just look at the total price when they view the primary product/service and ancillary service together. Considerations involving pricing structure such as partitioned prices vs. single total price (Morwitz et al. 1998), which are not the focus of this paper, can then become pertinent. So we mainly focus on the pricing and service delivery decisions involving the ancillary service. In Appendix D, we show that our model and analysis can be easily extended to consider the impact of an *exogenous* primary product/service price; however, this price does not affect the key results so we omit it in the main paper for the sake of conciseness and easier exposition.

Customers are time-sensitive and they belong to two types. Type 1 and Type 2 customers have different delay sensitivities of η_1 and η_2 ($\eta_1 \neq \eta_2$) respectively. They value the ancillary service at v_1 and v_2 respectively. We assume wlog that $\eta_1 > \eta_2$, and we refer to Type 1 (Type 2) customers as impatient (patient) customers. The fraction of Type 1 customers is q ($0 < q < 1$). A Type i ($i = 1, 2$) customer purchases service class j ($j = h, l$) only if his net utility given by $v_i - p_j - \eta_i \cdot$ *total expected delay in service class* $j \geq 0$, in which the total expected delay is the sum of the expected waiting time and strategic delay. We measure *delay and value heterogeneity* among customers by $\Delta \equiv \eta_1/\eta_2$ and $\nu \equiv v_1/v_2$ respectively.

The retailer knows the values of v_1, v_2, η_1, η_2, and q; however, she does not know the customer type of an individual customer. Because the retailer provides an ancillary service, she minimizes the net cost of providing the service to customers, *while ensuring that all the customers use the service.* A customer's choice of a service class is strategic because the expected waiting time is not only affected by it but also by similar choices of other customers. The retailer anticipates the behavior by customers, and she sets the prices for service classes and decides how to provide them so that her net cost is minimized. We consider the retailer's cost minimization under two cases: (i) her capacity μ is exogenous and (ii) her capacity is optimized so that the total net cost including the capacity cost is minimized. Next, we discuss how the expected waiting time without strategic delay is related to different parameters.

When the retailer does not prioritize, all the customers are serviced in a first-come-first-serve (FCFS) manner. We denote the expected waiting time in this case by $W(\mu)$ in which μ is the retailer's capacity. We assume that $W(\mu)$ is a strictly decreasing and strictly convex function, i.e., $W'(\mu) < 0$ and $W''(\mu) > 0$. Note that even though this waiting time is the same across different customers, the retailer can still differentiate the two service classes by

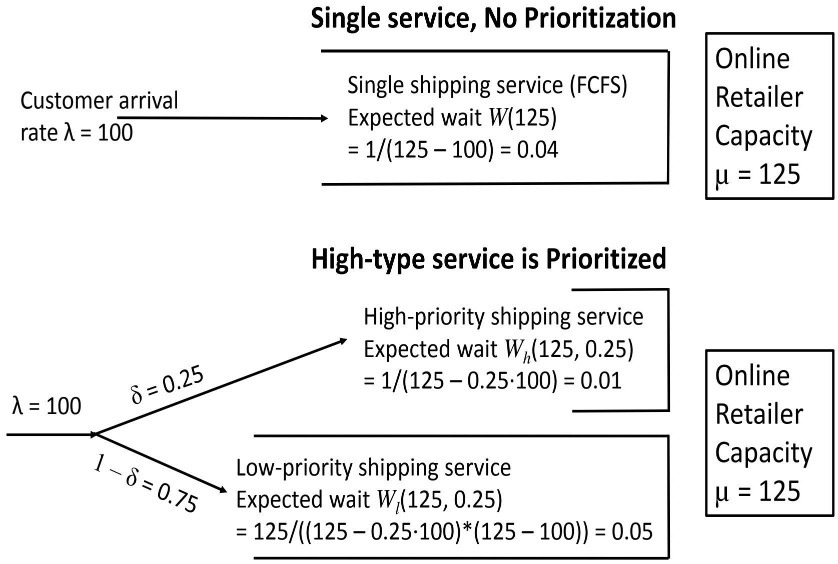

Single service, No Prioritization
Customer arrival rate λ = 100
Single shipping service (FCFS)
Expected wait W(125)
= 1/(125 – 100) = 0.04
Online Retailer Capacity μ = 125
High-type service is Prioritized
λ = 100
δ = 0.25
1 – δ = 0.75
High-priority shipping service
Expected wait W_h(125, 0.25)
= 1/(125 – 0.25·100) = 0.01
Low-priority shipping service
Expected wait W_l(125, 0.25)
= 125/((125 – 0.25·100)*(125 – 100)) = 0.05
Online Retailer Capacity μ = 125

Figure 2: Illustration of retailer's system and key model notation

using strategic delay. When the retailer prioritizes high-type service customers, there are two key aspects: (i) the expected waiting time will be the same for all customers from the same service class but a high-type service customer will have a different expected waiting time from that of a low-type service customer and (ii) the expected waiting times for high- and low-type services will depend not only on the retailer's capacity but also on δ ($0 \leq \delta \leq 1$), the fraction of customers who choose high-type service. We denote the expected waiting times for high- and low-type services by $W_h(\mu, \delta)$ and $W_l(\mu, \delta)$ respectively. However, for brevity, we also use just W_h and W_l to denote these waiting times. We assume that $\partial W_j/\partial\mu < 0$, $\partial^2 W_j/\partial\mu^2 > 0$, $\partial W_j/\partial\delta > 0$, $\partial^2 W_j/\partial\delta^2 > 0$, and $\partial W_h/\partial\delta < \partial W_l/\partial\delta$; $j = h, l$. These assumptions are intuitive (e.g., they are satisfied by $M/M/1$ queues with non-preemptive and preemptive priorities[6]); the expected waiting times decrease with a higher capacity at a diminishing rate but they increase with a higher load on the high-type service at an increasing rate; and an increase in the fraction of prioritized customers affects the expected waiting time of high-priority service less than that of low-priority service. Further, we assume that work conservation applies, i.e., $\delta W_h(\mu, \delta) + (1 - \delta) W_l(\mu, \delta) = W(\mu)\ \forall \mu, \delta$. Figure 2 illustrates the retailer's system[7] with $M/M/1$ queues and preemptive priority (Gross 2008), so that

[6] They also hold in $M/G/1$ queues with non-preemptive priority. Furthermore, the conditions on W are satisfied by the approximate waiting time for a general queue (Cox and Smith 1991), which is given by $\frac{1}{\mu} + \frac{1}{\lambda} \cdot \frac{\rho^{\sqrt{2(m+1)}}}{1-\rho} \cdot \frac{CV_a^2 + CV_s^2}{2}$ in which m is the number of servers, $\rho \equiv \frac{\lambda}{m\mu}$ is the utilization, and CV_a (CV_s) is the coefficient of variation in inter-arrival times (service times).

[7] While our analysis in §6 applies to more general queuing systems, we use $M/M/1$ queues and preemptive

$W(\mu) = 1/(\mu - \lambda)$, $W_h(\mu, \delta) = 1/(\mu - \lambda\delta)$, and $W_l(\mu, \delta) = \mu/\left((\mu - \lambda\delta) \cdot (\mu - \lambda)\right)$. Next, we formulate the retailer's optimization problem when there is no prioritization.

4. No Prioritization

The retailer does not prioritize, and she just uses strategic delay to differentiate high- and low-type services. We allow for the possibility that she can set the prices of high- and low-type services so that even the same type of customers *split*, i.e., some of them choose high-type service and others select low-type service (in §5 we show that splitting is *optimal* under certain cases). We let γ_i denote the fraction of Type i customers who choose the high-type service. The retailer's optimization problem can then be written as

$$(\mathcal{N}): \min_{p_h, p_l, d_h, d_l, \gamma_1, \gamma_2} \lambda r - \lambda\left(\gamma_1 q + \gamma_2(1-q)\right) p_h - \lambda\left((1-\gamma_1)q + (1-\gamma_2)(1-q)\right) p_l$$

$$\begin{aligned}
s.t.\quad & \gamma_i\left(p_h + \eta_i(W(\mu) + d_h)\right) \le \gamma_i v_i && i = 1,2 && (1)\\
& (1-\gamma_i)\left(p_l + \eta_i(W(\mu) + d_l)\right) \le (1-\gamma_i) v_i && i = 1,2 && (2)\\
& \gamma_i\left(p_h + \eta_i(W(\mu) + d_h)\right) \le \gamma_i\left(p_l + \eta_i(W(\mu) + d_l)\right) && i = 1,2 && (3)\\
& (1-\gamma_i)\left(p_l + \eta_i(W(\mu) + d_l)\right) \le (1-\gamma_i)\left(p_h + \eta_i(W(\mu) + d_h)\right) && i = 1,2 && (4)\\
& d_h \le d_l && && (5)\\
& 0 \le \gamma_i \le 1; d_h, p_j \ge 0 && i = 1,2; j = l,h && (6)
\end{aligned}$$

The retailer's objective is the net cost (per unit time), i.e., the difference of service cost and the total revenue obtained from charging prices p_h and p_l for the high- and low-type services. Constraints in (1) imply that if $\gamma_i > 0$ $(i = 1, 2)$, the net utility of Type i customers should be non-negative. They are *individual rationality* (IR) constraints of Type i customers for high-type service. If $\gamma_i = 0$, then no Type i customer selects high-type service and the corresponding IR constraint does not apply. Similarly, constraints in (2) are IR constraints of Type i customers for low-type service[8]. Constraints in (3) imply that if $\gamma_i > 0$, Type i customers obtain a higher net utility from high-type service than from low-type service. We refer to them as *incentive compatibility* (IC) constraints of Type i customers for high-type service. Similarly, constraints in (4) are IC constraints of Type i customers for low-type

priority, as illustrated by Figure 2, in §'s 7 and 8 for easier analysis.

[8] The IR constraints in (1) and (2) are akin to service-level agreements/guarantees and ensure that the retailer provides adequate service to the customers.

service. Constraint (5) follows from how high- and low-type services are defined (see §3). Constraints in (6) imply that the prices and strategic delay should be non-negative[9] and that the fraction of Type i customers choosing the high-type service should be between zero and one. We model γ_i's as retailer's decision variables purely for simplifying the formulation and making the exposition easier[10]. *However, we note that the optimal prices and strategic delays constitute an equilibrium among customers (with the corresponding optimal γ_i's) because they satisfy the requisite IR and IC conditions so that none of the customers would have an incentive to change their service classes.* We discuss about the implementation of scenarios in which Type i customers *split*[11], i.e., $0 < \gamma_i < 1$, in §6 following Lemma 5. We assume that the exogenous capacity μ is large enough with $W(\mu) \leq \min_{i=1,2} v_i/\eta_i$ so that $\mathcal{N}$ has a feasible solution. Next, we make some observations that simplify the optimization problem.

We note that $d_h = 0$ under optimality because otherwise the retailer can decrease d_h and d_l while increasing p_h and p_l thereby reducing the net cost. Also, if $d_l = 0$ under optimality, there is no service differentiation and the retailer provides just a single service class. Similarly, if $\gamma_1 = \gamma_2 = 0$ or $\gamma_1 = \gamma_2 = 1$, all the customers select a single service class. The minimum net cost from these scenarios in which there is no service differentiation is equal to that when the retailer offers just a single service class and there are no IC constraints, and it is given by $\max_{i=1,2} \lambda r - \lambda v_i + \lambda \eta_i W(\mu)$. Next, we consider what happens if there is service differentiation. Then $d_l > 0 = d_h$ under optimality. Further, Lemma 1 shows that the net cost is optimized only if $\gamma_1 = 1$ and $\gamma_2 = 0$.

Lemma 1 *If there is service differentiation under optimality, the retailer incurs the optimal net cost only if* $\gamma_1 = 1$ *and* $\gamma_2 = 0$.

All proofs are in Appendix A. Lemma 1 shows that the best way for the retailer to differentiate the customers is to have all the impatient (patient) customers purchase the high-type (low-type) service. She does *not split* customers who are of the same type. This result does not depend on the valuations v_1 and v_2. However, they do affect whether the retailer should differentiate the customers or just offer a single service class. Lemma 2 characterizes the optimal solution of $\mathcal{N}$.

[9] Prices are assumed to be non-negative because (i) charging negative prices might be impractical, and (ii) customers can perceive that the retailer is disingenuous and charges more for the primary product/service, which may result in a backlash from them.

[10] The retailer's pricing and strategic delay decisions *indirectly* affect γ_i's through customers' self-selection.

[11] As we show later in §'s 4-6, such a split can be optimal only under prioritization.

Lemma 2 *The optimal solution of* $\mathcal{N}$ *is given as follows:*
(i) if $\Delta \geq \nu$ *and* $W(\mu) \geq (v_1 - v_2)/(\eta_1 - \eta_2)$ *then* $p_h = p_l = v_1 - \eta_1 W(\mu)$, $d_h = d_l = 0$, *and the net cost is* $\lambda r - \lambda v_1 + \lambda \eta_1 W(\mu)$ *with no service differentiation;*
(ii) if $\Delta \geq \nu$, $W(\mu) < (v_1 - v_2)/(\eta_1 - \eta_2)$, *and* $\Delta \leq 1/q$ *then* $p_h = p_l = v_2 - \eta_2 W(\mu)$, $d_h = d_l = 0$, *and the net cost is* $\lambda r - \lambda v_2 + \lambda \eta_2 W(\mu)$ *with no service differentiation;*
(iii) if $\Delta \geq \nu$, $W(\mu) < (v_1 - v_2)/(\eta_1 - \eta_2)$, *and* $\Delta > 1/q$ *then* $p_h = v_1 - \eta_1 W(\mu)$, $p_l = (\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2)$, $d_h = 0$, $d_l = (v_1 - v_2)/(\eta_1 - \eta_2) - W(\mu)$, *and the net cost is* $\lambda r - \lambda \left(\frac{(\eta_1 q - \eta_2) v_1 + (1-q) \eta_1 v_2}{\eta_1 - \eta_2} \right) + \lambda \eta_1 q W(\mu)$ *with service differentiation;*
(iv) if $\Delta < \nu$ *and* $\Delta \leq 1/q$, $p_h = p_l = v_2 - \eta_2 W(\mu)$, $d_h = d_l = 0$, *and the net cost is* $\lambda r - \lambda v_2 + \lambda \eta_2 W(\mu)$ *with no service differentiation; and*
(v) if $\Delta < \nu$ *and* $\Delta > 1/q$, $p_h = \eta_1 (v_2/\eta_2 - W(\mu))$, $p_l = 0$, $d_h = 0$, $d_l = v_2/\eta_2 - W(\mu)$, *and the net cost is* $\lambda r - \lambda (\eta_1 q/\eta_2) v_2 + \lambda \eta_1 q W(\mu)$ *with service differentiation.*

Lemma 2 establishes the conditions under which the retailer differentiates her customers just through strategic delay. First, $\Delta > 1/q$, so that the patient and impatient customers have to be *sufficiently different* in their delay sensitivities. Further, if the fraction of impatient customers, q, is lower then a higher delay heterogeneity is necessary for service differentiation. Second, in addition to the value of Δ, whether the retailer differentiates also depends on how it compares with value heterogeneity ν. If $\Delta < \nu$, then the retailer differentiates for any feasible capacity. Also, she provides the low-type service *free of charge*. However, if $\Delta \geq \nu$, she also needs to have a *sufficiently high capacity* so that $W(\mu) < (v_1 - v_2)/(\eta_1 - \eta_2)$ for service differentiation to be optimal.

5. High-type Service is Prioritized

As we mentioned earlier in §3, without loss of generality, we refer to the service class which is given high (low) priority by the retailer as high-type (low-type) service. The notation for the problem parameters and decision variables remains the same as in §'s 3 and 4. As in §4, we assume that $W(\mu) \leq \min_{i=1,2} v_i/\eta_i$, i.e., the retailer has enough capacity to provide adequate service without prioritization and strategic delay. Also, we define $\tilde{\gamma}_1$ such that $W_l(\mu, \tilde{\gamma}_1 q) = \min_{i=1,2} v_i/\eta_i$ to characterize retailer's optimal solution later. It is the maximum fraction of impatient customers that choose the high-type service, so that the retailer can provide adequate low-type service to at least one customer type. There is a key aspect that makes the cost minimization problem here different from that in §4: the waiting times of high- and low-type services without delay, $W_h(\mu, \delta)$ and $W_l(\mu, \delta)$, not only depend on the capacity μ but also on the (total) fraction of customers who choose the high-type service δ which equals $\gamma_1 q + \gamma_2(1-q)$. This problem is formulated as

$$(\mathcal{P}): \min_{p_h, p_l, d_h, d_l, \gamma_1, \gamma_2} \lambda r - \lambda\left(\gamma_1 q + \gamma_2(1-q)\right) p_h - \lambda\left((1-\gamma_1)q + (1-\gamma_2)(1-q)\right) p_l$$

$$\begin{aligned}
s.t.\quad & \gamma_i\left(p_h + \eta_i(W_h(\mu, \gamma_1 q + \gamma_2(1-q)) + d_h)\right) \le \gamma_i v_i && i = 1,2 && (7)\\
& (1-\gamma_i)\left(p_l + \eta_i(W_l(\mu, \gamma_1 q + \gamma_2(1-q)) + d_l)\right) \le (1-\gamma_i) v_i && i = 1,2 && (8)\\
& \gamma_i\left(p_h + \eta_i(W_h(\mu, \gamma_1 q + \gamma_2(1-q)) + d_h)\right) \\
& \qquad \le \gamma_i\left(p_l + \eta_i(W_l(\mu, \gamma_1 q + \gamma_2(1-q)) + d_l)\right) && i = 1,2 && (9)
\end{aligned}$$

$$\begin{aligned}
& (1-\gamma_i)\left(p_l + \eta_i(W_l(\mu, \gamma_1 q + \gamma_2(1-q)) + d_l)\right) \\
& \qquad \le (1-\gamma_i)\left(p_h + \eta_i(W_h(\mu, \gamma_1 q + \gamma_2(1-q)) + d_h)\right)\ i = 1,2 && (10)\\
& 0 \le \gamma_i \le 1; d_j, p_j \ge 0 \qquad i = 1,2; j = l,h && (11)
\end{aligned}$$

Constraints (7) and (8) are the IR constraints of Type i customers for high- and low-type services respectively while constraints (9) and (10) represent the IC constraints of Type i customers for these two service classes. The logical constraints are represented by (11). Note that non-negativity of prices is a *non-trivial* constraint that has to be added in $\mathcal{P}$. This aspect is different from research works that consider *revenue/profit maximization* by service provider, which *automatically* results in non-negative prices (e.g., see Afeche (2013) and Katta and Sethuraman (2005)). We identify different cases in which the optimal price is zero and thereby intend to explain why free service provision for ancillary services (e.g., free-shipping) is prevalent.

The high-type service in $\mathcal{P}$, unlike in $\mathcal{N}$, can be delayed more or less than the low-type service because high- and low-type services are defined as high- and low-priority service classes respectively (see §3). Note that either $d_h = 0$ or $d_l = 0$ under optimality. Otherwise, the delay for both high- and low-type services can be decreased by the same amount $\epsilon > 0$ while increasing p_h and p_l by $\eta_2 \epsilon$, ensuring that the constraints in $\mathcal{P}$ are still satisfied and thereby reducing the net cost. Further, two conditions are necessary for service differentiation to occur: (i) $W_h + d_h \neq W_l + d_l$, i.e., the total expected delays are different, and (ii) neither $\gamma_1 = \gamma_2 = 0$ nor $\gamma_1 = \gamma_2 = 1$, i.e., there are some customers who select each service class. Next, we characterize how γ_i's are related if there is service differentiation.

Lemma 3 *Under service differentiation, the following relationships hold: (i) If* $0 < \gamma_1 < 1$, *either* $\gamma_2 = 0$ *or* $\gamma_2 = 1$, *and (ii) if* $0 < \gamma_2 < 1$, *either* $\gamma_1 = 0$ *or* $\gamma_1 = 1$.

Lemma 3 shows that both patient and impatient customers cannot be split between high- and low-type services. That is because they have different delay sensitivities and so the IC constraints of both types of customers cannot be satisfied for both service classes. Further, Lemma 3 implies that, under service differentiation, the problem $\mathcal{P}$ can be simplified into the following six problems: (i) $\mathcal{P}1$ with $0 < \gamma_1 < 1$ and $\gamma_2 = 0$; (ii) $\mathcal{P}2$ with $0 < \gamma_1 < 1$ and $\gamma_2 = 1$; (iii) $\mathcal{P}3$ with $\gamma_1 = 0$ and $0 < \gamma_2 < 1$; (iv) $\mathcal{P}4$ with $\gamma_1 = 1$ and $0 < \gamma_2 < 1$; (v) $\mathcal{P}5$ with $\gamma_1 = 1$ and $\gamma_2 = 0$; and (vi) $\mathcal{P}6$ with $\gamma_1 = 0$ and $\gamma_2 = 1$. Appendix B shows how problem $\mathcal{P}$ simplifies in each case, provides lemmas which characterize the optimal solutions for these problems, and describes the insights from these solutions.

Optimal Solution when High-Type Service is Prioritized

We use the results from Appendix B to find the optimal solution when the retailer prioritizes and thereby differentiates some of her customers. Lemma 4 characterizes this solution, and it enables us to compare prioritization vs. no-prioritization and find the overall solution in §6. We find that the optimal values depend on whether $\mathcal{P}5$, in which all impatient (patient) customers select high-type (low-type) service, is feasible.

Lemma 4 *The optimal solution for problem* $\mathcal{P}$ *is as follows:*
(i) If $W_l \leq v_2/\eta_2$ *and* $(\eta_1 - \eta_2)W_h + \eta_2 W_l \leq v_1$ *then* $\mathcal{P}5$ *is feasible,* $\gamma_1 = 1$, $\gamma_2 = 0$, *and the optimal prices, delay, and net cost are as given in Lemma B5;*
(ii) otherwise, $\mathcal{P}5$ *is infeasible,* $\gamma_1 = \tilde{\gamma}_1$, $\gamma_2 = p_l = d_h = d_l = 0$, *and the other optimal values are given by: (a) if* $\Delta \geq \nu$ *then* $p_h = v_1 - \eta_1 W_h(\mu, \tilde{\gamma}_1 q)$ *and the net cost is* $\lambda r - \lambda v_1 + \lambda \eta_1 W$, *(b) if* $\Delta < \nu$ *then* $p_h = \eta_1 \left(v_2/\eta_2 - W_h(\mu, \tilde{\gamma}_1 q)\right)$ *and the net cost is* $\lambda r - \lambda \eta_1 (v_2/\eta_2) + \lambda \eta_1 W$.

Lemma 4 shows that providing high-type service to any patient customer would be suboptimal to the retailer, *regardless of how much he values the service.* Further, if the retailer has sufficient capacity to provide adequate service to the customers even when all the impatient ones get prioritized, then $\gamma_1 = 1$ and $\gamma_2 = 0$, similar to when there was service differentiation without prioritization (see §4). If the capacity is insufficient then splitting impatient customers, in which some of them select high-type service while others select low-type service, is optimal under prioritization[12]. However, will it still be optimal even when the retailer can decide whether or not to prioritize? And does prioritization always perform strictly better? Next, we answer these questions by finding the retailer's overall optimal solution.

6. Prioritize or Not? Overall Optimal Solution

We consider the problem in which the retailer, in addition to the prices, delay and γ_i's, also decides whether or not to prioritize some of her customers. We assume that if their net costs are equal the retailer has the following preference relationship for different types of service delivery: single service class $\succ$ no-priority with strategic delay $\succ$ prioritization without strategic delay $\succ$ prioritization with strategic delay. The net cost has to strictly decrease for the retailer to implement a more complicated service delivery. Next, we characterize the retailer's optimal solution and provide conditions under which the following key features are optimal: (i) free service, (ii) single service, (iii) split of impatient customers, and (iv) strategic delay.

Theorem 1 *If $\Delta \geq \nu$, the retailer offers just a single service class with price $v_1 - \eta_1 W(\mu)$ and incurs a net cost of $\lambda r - \lambda v_1 + \lambda \eta_1 W$ when (i) $W_h(\mu, q) > (v_1 - v_2)/(\eta_1 - \eta_2)$ and $(\eta_1 - \eta_2) W_h(\mu, q) + \eta_2 W_l(\mu, q) > v_1$ or (ii) $W_h(\mu, q) \leq (v_1 - v_2)/(\eta_1 - \eta_2)$ and $W_l(\mu, q) > v_2/\eta_2$. Otherwise, or if $\Delta < \nu$, the retailer prioritizes the customers with the prices, delay, and net cost as in Lemma 4. Furthermore, strategic delay without prioritization is sub-optimal.*

Corollary 1 *The main features[13] in the optimal service delivery, when service is prioritized (differentiated), are characterized under different cases as follows:*

(i) If $\Delta \geq \nu$ and $W_l(\mu, q) = v_2/\eta_2$ then the low-priority service is free.

(ii) If $\Delta > \nu, 1/q$ and $W_l(\mu, q) < (v_1 - v_2)/(\eta_1 - \eta_2)$ then the low-priority service has a strategic delay.

(iii) If $\Delta = \nu$, $\Delta > 1/q$, and $W_l(\mu, q) < (v_1 - v_2)/(\eta_1 - \eta_2)$ then the low-priority service is free and it has a strategic delay.

[12] Note that the split is at an aggregate (customer type) level. Because the prices satisfy the IC constraints in (9) and (10), this split is an *equilibrium* among customers as none of them would have an incentive to select a different service class.

[13] For conciseness, the corollary mentions only about (a) features that are present and (b) cases in which at least one of them is present. E.g., it does not say "no strategic delay" when strategic delay is absent.

(iv) If $\Delta < \nu$ and $W_l(\mu, q) > v_2/\eta_2$ then the low-priority service is free and impatient customers split between high- and low-priority services with $\tilde{\gamma}_1 q$ customers choosing high-priority service, where $W_l(\mu, \tilde{\gamma}_1 q) = v_2/\eta_2$.

(v) If $\Delta < \nu$ and $W_l(\mu, q) = v_2/\eta_2$ then the low-priority service is free.

(vi) If $1/q < \Delta < \nu$ and $W_l(\mu, q) < v_2/\eta_2$ then the low-priority service is free and it has a strategic delay.

We first note that all the types of optimal service delivery in Theorem 1, including split of impatient customers, *result in equilibriums among customers because the IC and IR conditions are satisfied and none of them have any incentive to change their service class.* Next, we discuss about the different service delivery features from the results in Theorem 1 and its corollary.

A strategic delay without prioritization is sub-optimal because while a strategic delay makes low-type service worse, it cannot improve high-type service as it has to be non-negative, but prioritization not only makes the low-type service worse but it also makes the high-type service better, thereby enhancing the retailer's ability to differentiate her customers. Theorem 1 also shows that if strategic delay is optimal it is applied only to low-type (low-priority) service. So it is used in addition to prioritization in order to further differentiate patient and impatient customers. From Corollary 1, we find that delay heterogeneity Δ and value heterogeneity ν are key metrics that determine how service is delivered. If $\Delta > \nu$, i.e., the retailer sells *staple products* (as we discuss in §1) splitting customers is never optimal but offering a single service class is optimal under low capacity. Hence, in addition to the heterogeneity in delay sensitivities, *sufficient capacity* is required for the retailer to differentiate her customers. However, if $\Delta < \nu$, i.e., she sells *fashion goods*, then prioritization is always strictly better than single service and splitting impatient customers becomes optimal under low capacity. Furthermore, when there is a split, $\gamma_1 = \tilde{\gamma}_1$ which is reasoned as follows. The retailer prefers that the maximum fraction of impatient customers, which still ensures that adequate low-type service is provided (to patient customers), select high-type service. In summary, there are two ways in which the retailer can respond to low capacity: (i) single service with no prioritization and (ii) splitting impatient customers to reduce the need for high-priority service. The better way is determined based on whether staple products or fashion goods are sold. We note that neither single service nor split of customers is explicitly considered in some of the papers (Afeche 2013, Katta and Sethuraman 2005) that are closely related to this research.

Theorem 1 and its corollary also provide the conditions under which the retailer offers free service. Ignoring some *borderline* parametric values (e.g., μ s.t. $W_l(\mu, q) = v_2/\eta_2$ or v_i's, η_i's s.t. $\Delta = \nu$), we find that it happens when fashion goods are sold and (i) $W_l(\mu, q) \geq v_2/\eta_2$ or (ii) $W_l(\mu, q) < v_2/\eta_2$ and $\Delta > 1/q$. Strategic delay is zero in the first case, while it is positive in the second case. Also, there is another key distinction between the two cases: low capacity is responsible for free service in the first case (note that impatient customers are split and service is prioritized); however, free service in the second case is *driven by customer characteristics, i.e., their valuations, delay sensitivities, and the fraction of impatient customers.*

Finally, we note that there is some overlap in the results in Theorem 1 and Corollary 1 and those from Table 1 in §1. The single service case in Theorem 1 here corresponds to case 2 there, case (ii) corresponds to case 5, case (iv) corresponds to case 1, and case (vi) corresponds to case 4 in Table 1. Next, we prove a key result involving what happens when the optimal service delivery involves splitting of impatient customers.

Lemma 5 *If the optimal service delivery is to split impatient customers, then the corresponding prices and strategic delays (from Theorem 1) can only result in the following equilibrium values of γ_1 and γ_2: $\gamma_1 = \tilde{\gamma}_1$ and $\gamma_2 = 0$, or $\gamma_1 = \gamma_2 = 0$.*

Lemma 5 shows that $\gamma_1 = \tilde{\gamma}_1$ is not the unique equilibrium outcome when impatient customers are split. However, the retailer can still effect this outcome when *at least some customers select high-type service.* One way to do that is by exploiting customers' *risk perception*: the retailer provides information about the range of waiting times they would encounter when they select high- and low-type services[14]. Many online retailers already adopt this strategy (e.g., see Amazon.com (2014)). Although we do not model risk among customers and we assume that they make their decisions based on expected waiting time, *risk perception can still play a marginal role and ensure the optimal equilibrium for the retailer.* For this purpose, when impatient customers are split, we only consider $\gamma_1 = \tilde{\gamma}_1$.

[14] We assume that the retailer is truthful, customers anyway obtain the information in the long-run. The range has a lower bound and an upper bound on the waiting time (e.g., 1st and 99th percentiles respectively). Due to prioritization, low-type service would have a larger range than high-type service.

7. Optimal Service Delivery: Endogenous Capacity

For analytical simplicity, we assume that (i) the cost of capacity is linear with K per unit, and (ii) an $M/M/1$ queuing system with preemptive priority is used by the retailer under prioritization. She now minimizes her *total cost* given by the sum of net cost in §6 (see Theorem 1) and capacity cost of μK. As in §6, we only consider μ's such that $W(\mu) \leq \min_{i=1,2} v_i/\eta_i$. We can use this *sequential* approach to find the retailer's optimal decisions (pricing and strategic delivery decisions followed by capacity decision) because the best net cost, price(s), and strategic delay(s) are all *deterministic functions* of the capacity. Also, we can use the results from §6 by following this approach.

We first discuss how the optimal service delivery (for a given μ) evolves as μ increases. From Theorem 1, we find that it is straightforward when $\Delta < \nu$. For low μ's $W_l(\mu, q) > v_2/\eta_2$, the retailer splits impatient customers and offers two service classes, the low-type service is free and has no strategic delay. For higher values of μ, she still offers two service classes; if $\Delta \leq 1/q$ then there is no free service (unless $W_l(\mu, q) = v_2/\eta_2$) and no strategic delay, and if $\Delta > 1/q$ then the low-type service is free but has a strategic delay (unless $W_l(\mu, q) = v_2/\eta_2$). However, the change of service delivery becomes complicated when $\Delta \geq \nu$, mainly because of the complex range of (μ, q) values over which offering a single service class is optimal. We provide the details in Appendix C. Next, we characterize the retailer's optimal solution and her optimal service delivery when she has a large pool of customers.

Theorem 2 *The total cost is minimized by a unique optimal capacity μ^*. As $\lambda \to \infty$, single service, strategic delay, and split of impatient customers all become sub-optimal. Furthermore, if $1/q < \Delta < \nu$ ($\Delta > \nu$), then the low-priority service is free (charged).*

We explain the sub-optimality of (i) offering a single service class (or *split*) and (ii) having a strategic delay for a large system as follows. A single service class is sub-optimal because when $\lambda \to \infty$, due to economies of scale, service differentiation becomes relatively easier to implement since the retailer is not so constrained by her capacity. Also, when λ is high, raising the capacity beyond the threshold needed for the optimality of strategic delay, so that $W_l < \min\left(\frac{v_2}{\eta_2}, \frac{v_1 - v_2}{\eta_1 - \eta_2}\right)$, increases the capacity cost more than the increase in revenue from charging a higher price for high-type service (note that $p_l = 0$ or $p_l = (\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2)$ with strategic delay and it is independent of the capacity μ). So the retailer incurs less

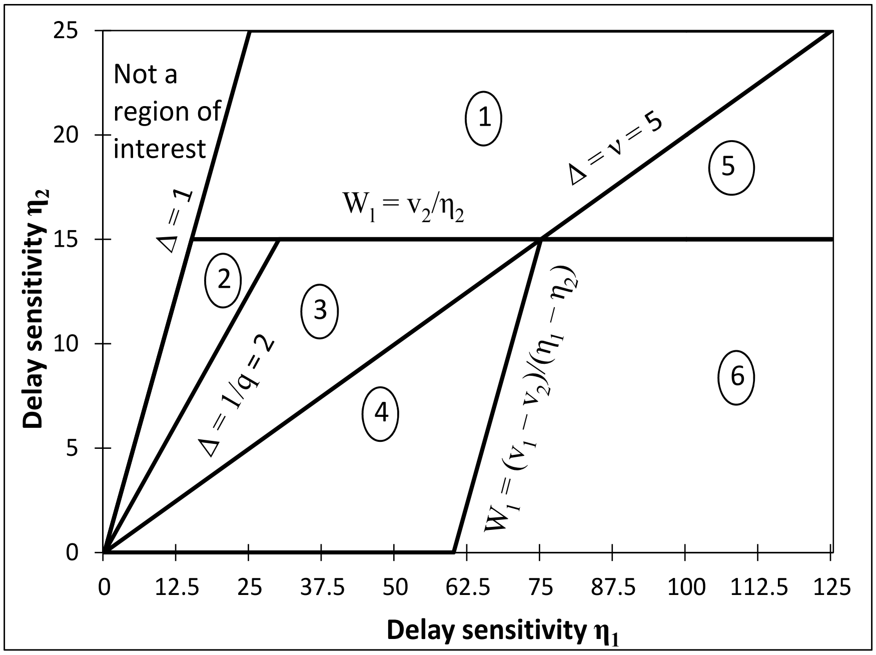


Figure 3: Variation of optimal service delivery with η_1 and η_2

Region	Single service	Free service	Prioritized service	Strategic delay	Split of impatient customers
1		✓	✓		✓
2			✓		
3		✓	✓	✓	
4			✓	✓	
5	✓				
6			✓		

Table 1: Features of optimal service delivery in different regions

total cost by having a lower capacity and no strategic delay instead of a high capacity with strategic delay.

Theorem 2 also shows that despite having a *large system* with $\lambda \to \infty$, free service is still provided when delay heterogeneity takes intermediate values (in between $1/q$ and value heterogeneity ν). Next, we analyze some numerical examples to better understand how retailer's service delivery changes with different parameters under exogenous and endogenous capacities.

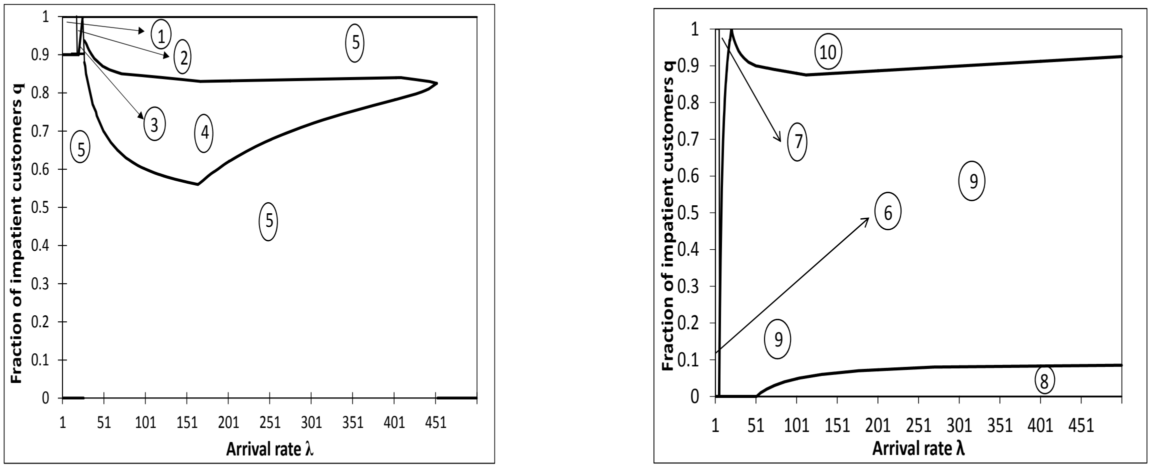
8. Numerical Analysis

This section comprises of two parts, in both of which we use $M/M/1$ preemptive priority under prioritization. In the first part (§8.1), we discuss how retailer's optimal service delivery changes with delay sensitivities η_1 and η_2, when the capacity is exogenous. We focus on η_1 and η_2 to better understand how the magnitude and heterogeneity of delay sensitivity affect retailer's decisions. In the second part (§8.2), we discuss how retailer's optimal service delivery changes with arrival rate λ and fraction of impatient customers q, when the capacity is endogenous. In the third part (§8.3), the capacity is still endogenous but we consider how optimal service delivery changes with unit capacity cost K and fraction q. When capacity is endogenous, we focus on the variation of λ, q, or K because these are the main parameters that typically affect capacity decisions (and optimal service delivery).

8.1 Exogenous Capacity: Impact of η_1 and η_2

Figure 3 and Table 1 show how the retailer's optimal service delivery changes when η_1 and η_2 vary. For this numerical example, we set the values of fixed parameters at $r = 2$, $\lambda = 100$, $\mu = 125$, $v_1 = 5$, $v_2 = 1$, and $q = 0.5$. Figure 3 shows that the (η_1, η_2) space, in which $\eta_1 > \eta_2$, gets divided into six regions. When η_2 is high so that $W_l(\mu, q) > v_2/\eta_2$, there are two possibilities. In region 1, when the delay heterogeneity $\Delta < \nu$, the value heterogeneity, and the retailer sells *fashion goods*[15], prioritization is preferred but some impatient customers need to be split so that adequate service gets provided for customers that select the low-priority service. In region 5, when $\Delta > \nu$ and the retailer sells *staple products*, she obtains optimal profits by just providing a single service class. A high value of delay heterogeneity Δ should favor prioritization but a high value of η_2 precludes the retailer from realizing this benefit. When η_2 is low and $W_l(\mu, q) < v_2/\eta_2$, there are four regions in Figure 3. The retailer uses prioritization in all these regions. However, other features of service delivery differ across these regions. In regions 2 and 6, there is no strategic delay or free service; in region 3, both of them are present, while region 4 has strategic delay but no free service. These differences are explained as follows. In region 2, since η_1 is low, the delay heterogeneity Δ is low so there is no strategic delay/free service. However, in region 6, η_1 is high so that $W_l(\mu, q) > (v_1 - v_2)/(\eta_1 - \eta_2)$ and the retailer does not have ample capacity that is needed for strategic delay/free service. In both regions 3 and 4, there is strategic delay since the values of η_1 and η_2 ensure that (i) there is enough delay heterogeneity and (ii) the retailer has ample capacity. Free service is determined by how Δ and ν compare. In region 3 (region 4), *fashion goods* (*staple products*) are sold, and free service (no free service) is offered.

[15] In §1, we briefly discuss how $\Delta < \nu$ and $\Delta > \nu$ can pertain to the retailer selling *fashion goods* and *staple products* respectively.



(a) $\Delta > \nu$ ($v_1 = 5$)

(b) $\Delta < \nu$ ($v_1 = 20$)

Figure 4: Variation of optimal service delivery with λ and q

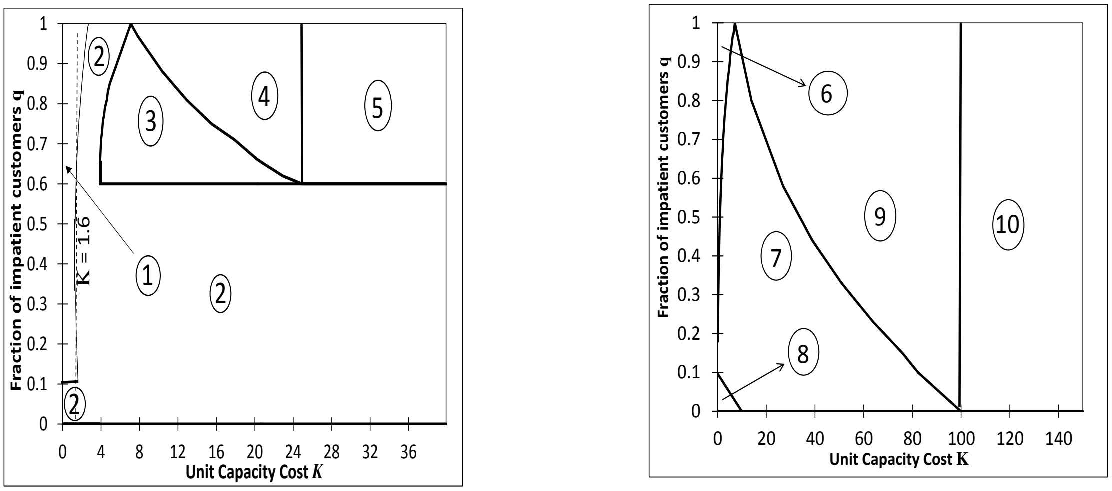
Region	Single service	Free service	Prioritized service	Strategic delay	Split of impatient customers
1	✓	✓			
2	✓				
3, 4, 9		✓	✓		
5, 8			✓		
6	✓	✓			
7		✓	✓		✓
10		✓	✓	✓	

Table 2: Features of optimal service delivery in different regions of (λ, q)

8.2 Endogenous Capacity: Impact of λ and q

We consider the variation of arrival rate λ and fraction of impatient customers q. These parameters, which fundamentally alter the composition of the customers, are likely to affect the retailer's capacity. So we consider the case of endogenous capacity (see §7). Based on different numerical examples, we find that the manner in which λ and q affect the retailer's optimal service delivery decision depends crucially on how Δ and ν compare with each other. We provide two demonstrative examples, one with $\Delta > \nu$ (for staple products) and the other with $\Delta < \nu$ (for fashion goods).

Figures 4a and 4b illustrate how the optimal service delivery changes with λ and q. The valuations of impatient customers are given by $v_1 = 5$ and $v_1 = 20$ in Figures 4a and 4b respectively. The values of other parameters are fixed at $v_2 = 1$, $\eta_1 = 100$, $\eta_2 = 10$, $r = 2$, and $K = 5$. Note that $\Delta = \eta_1/\eta_2 = 10 > \nu = v_1/v_2 = 5$ in Figure 4a, while $\Delta = 10 < \nu = 20$ in Figure 4b. Next, we analyze the optimal service delivery in different regions of these figures.



(a) $\Delta > \nu$ ($v_1 = 5$)

(b) $\Delta < \nu$ ($v_1 = 20$)

Figure 5: Variation of optimal service delivery with K and q

Region	Single service	Free service	Prioritized service	Strategic delay	Split of impatient customers
1			✓	✓	
2, 8			✓		
3, 7		✓	✓		
4	✓				
5, 10	✓	✓			
6		✓	✓	✓	
9		✓	✓		✓

Table 3: Features of optimal service delivery in different regions of (K, q)

Figure 4a is characterized by five regions. In regions 1, 2, and 3, λ is very low and q is high. Due to the low (optimal) capacity, the retailer offers only single service in regions 1 and 2; furthermore, it is free in region 1. In region 3, the retailer is able to prioritize her customers but she has to offer free service. The remaining two regions, 4 and 5, together comprise most of the (λ, q) combinations. In region 4, *with intermediate to high values of* q *and low to intermediate values of* λ (*limited capacity retailer*), free service is offered. However, in region 5, no free service is offered. The difference is explained as follows. In region 4, the retailer is constrained by her insufficient capacity, due to relatively higher values of q and lower values of λ, and has to therefore offer free service. In region 5. she does not face this constraint.

Figure 4b is also characterized by five regions. In regions 6 and 7, λ is very low which implies low capacity. The retailer offers single, free service in region 6. In region 7, she prioritizes the customers but the impatient ones get split due to low capacity. In region 9, which comprises most (λ, q) combinations, free service is offered. Again, the reason for free service is that the retailer's capacity is insufficient, i.e., $W_l(\mu^*, q) = v_2/\eta_2$. Free service is also offered, along with strategic delay, in region 10 in which the fraction of impatient customers q is very high. However, the reason for free service here is different, it is due to strategic delay. Due to a very high q, the retailer does have sufficient capacity μ^* so that $W_l(\mu^*, q) < v_2/\eta_2$. However, instead of charging the patient customers, she would rather provide them free service, strategically delay them, and then charge the impatient customers more to get a higher revenue. In region 8, there is no free service since q is low, the retailer has sufficient capacity, and there is no strategic delay. A key difference between Figures 4a and 4b is that free service is offered for *most* (λ, q) combinations in Figure 4b, unlike in Figure 4a. The main reason is due to how delay and value heterogeneity compare with each other, which depends on the nature of the product. In Figure 4b, $\Delta < \nu$ and the retailer sells *fashion goods*, while, in Figure 4a, $\Delta > \nu$ and the retailer sells *staple products.*

8.3 Endogenous Capacity: Impact of K and q

We consider the variation of unit capacity cost K and fraction of impatient customers q, which, respectively, affect the total cost of building capacity and customer composition, and are, therefore, also likely to affect the retailer's capacity. As in §8.2, we illustrate this variation by providing two demonstrative examples, one with $\Delta > \nu$ (for staple products) and the other with $\Delta < \nu$ (for fashion goods). We also use the same parametric values as those in §8.2; however, note that K now varies while we fix $\lambda = 100$.

Figure 5a shows how the optimal service delivery changes with K and q for staple products, and it is characterized by five regions. In region 1, because K is low and $q > 1/\Delta$, we have prioritization with strategic delay. However, in region 2, either q is too low or K is higher, so that strategic delay is no longer possible. Interestingly, we find that even when K becomes high, as long as q is not high ($q < 0.6$), the retailer differentiates her customers by offering prioritized service. The reasoning is as follows: because optimal capacity μ^* is such that $W(\mu^*) \leq \min_{i=1,2} v_i/\eta_i = v_1/\eta_1$ $(= 0.05)$ which is significantly less than v_2/η_2 $(= 0.1)$, *the retailer is able to prioritize, regardless of the value of K, as long as q is not high.* We make another key observation by looking at regions 1 and 2. We find that, for a small range of low values of K, as exemplified by the dashed line $K = 1.6$ in Figure 5a, the optimal service delivery exhibits the following complicated pattern as q increases from 0 to 1: *no strategic delay → strategic delay → no strategic delay → strategic delay.* Initially, strategic delay is futile because there are too many patient customers ($1 - q$ is high), and delaying all of them is counterproductive. Then it becomes optimal because the value of K ensures enough capacity for strategic delay. However, if q increases further, then the value of K in addition to the relatively high proportion of patient customers preclude strategic delay. Finally, if q is very high then there are too few patient customers and strategic delay again becomes optimal. Regions 3, 4, and 5 show that, when q is high, K has more impact on the optimal service delivery. *As K increases, the retailer first has to offer free service (even though she prioritizes her customers); she then offers just a single service but charges for it, and finally, for high values of K, she has to offer a single free service.*

Figure 5b shows how the optimal service delivery changes with K and q for fashion goods, and it is also characterized by five regions. Region 6, characterized by very low K and high q values, is the only region with strategic delay. Likewise, Region 8, characterized by very low K and low q values, is the only region in which the retailer does not offer free service. When K takes low (or intermediate) values, the optimal service delivery depends on whether q is below a certain threshold value which is decreasing in K: if that's the case, there is no split (Region 7); otherwise, impatient customers get split (Region 9). In either case, the retailer prioritizes and offers free service. Finally, if K is high then, regardless of the value of q, the retailer just offers a single, free service, as shown by Region 10.

9. Conclusion

We consider a retailer providing an ancillary service, a type of service that is characterized by its secondary nature in comparison to other products/services that the retailer sells and its management as a cost center, to two types of customers, patient and impatient, who may be heterogeneous both in their delay sensitivity, the rate at which their net utility decreases, and their service valuations. Our main example for such ancillary service is the shipping of products by online retailers to their customers. In this context, how should a retailer price her shipping service? Should she offer single service or two service classes? If she offers two service classes, should she prioritize some of her customers and/or strategically delay others? Should she provide free service? What are the factors which influence these decisions? We investigate these key questions in the paper, both when the retailer's capacity is exogenous and endogenous.

We find that (i) offering a single service class can be optimal even though customers are heterogeneous; (ii) the retailer prioritizes whenever two service classes are offered; and

(iii) *splitting* of impatient customers in which they select both high- and low-priority service classes can be optimal. There are two key factors that determine the optimal service delivery features: *her capacity and how customers' delay and value heterogeneity compare with each other.* If the delay heterogeneity is less than value heterogeneity then the retailer sells *fashion goods* for obtaining which the impatient customers are willing to pay *proportionately* more than patient ones. However, if the delay heterogeneity is higher then the retailer sells *staple products.* Then the impatient customers, even though they want the product more quickly, are not so brand-conscious and unwilling to pay proportionately more. Single service and splitting of impatient customers both happen when capacity is limited. Single service occurs when the retailer sells *staple products*, while splitting, in which the low-priority service is free, happens with *fashion goods.* Strategic delay is always used in conjunction with prioritization. It occurs when the *retailer has sufficient capacity* and the customers have a high delay heterogeneity. Furthermore, *strategic delay results in free service when the retailer sells fashion goods.*

Based on our numerical analysis in §8, we make some important observations about different service delivery features. First, we find that single service is offered when the capacity for satisfying customers is very low. That occurs when customers have high delay sensitivities η_1 and η_2 (Region 5 in Figure 3), arrival rate λ is too low (Regions 1 and 2 in Figure 4a and Region 6 in Figure 4b), or unit capacity cost K is high (Region 5 in Figure 5a and Region 10 in Figure 5b). On the other hand, strategic delay is used when the capacity is very high and the fraction of impatient customers q is sufficiently high (as seen in Regions 3 and 4 of Figure 3, Region 10 of Figure 4b, and Regions 1 and 6 in Figures 5a and 5b respectively). Single service and strategic delay can be optimal with both fashion goods and staple products. However, a *split in impatient customers* is only optimal when the retailer sells fashion goods; furthermore, she has insufficient capacity and/or a high fraction q (as seen in Region 1 of Figure 3, Region 7 of Figure 4b, and Region 9 of Figure 5b). If the cases above are excluded and we focus on scenarios in which the parameters $(\eta_1, \eta_2, \lambda, K, q)$ take intermediate (neither too high nor too low) values, the main question involves whether free service is offered. *We find that the answer crucially depends on whether the retailer sells fashion goods or staple products.* If she sells fashion goods, then she mostly *offers free service* (exemplified by Region 9 of Figure 4b and Region 7 of Figure 5b). However, if she sells staple products, she usually *does not offer free service* (shown by Region 5 of Figure 4a and Region 2 of Figure 5a). The only exception is Region 4 of Figure 4a in which λ and

q take intermediate-high values. The retailer then provides free service, in spite of selling staple products, because the optimal capacity is not enough to charge the patient customers (who select low-priority service).

Finally, we note that, although we examine the effect of primary product price on the pricing and delivery of ancillary service by using a simple extension of the model (see Appendix D), our analysis considers free service mainly from an operational perspective and ignores marketing elements pertaining to customer behavior. A future avenue of research might involve integration of operational and marketing aspects as well as the effect of competition to analyze the impact of free service provision.

Acknowledgment

The author is grateful to the anonymous reviewers, Associate Editor, and Editor for their valuable comments, suggestions, and assistance with the review process. The author is a Fellow of the Institute on Asian Consumer Insight (ACI) which focuses on Asian consumers and helps companies grow their businesses in Asia.

References

Afeche, P. 2013. Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing and Service Operations Management* **15**(3) 423–443.

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

Afeche, P., H. Mendelson. 2004. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science* **50**(7) 869–882.

Allon, G., A. Federgruen. 2009. Competition in service industries with segmented markets. *Management Science* **55**(4) 619–634.

Amazon.com. 2014. http://www.amazon.com/gp/help/customer/display.html.

Anand, K. S., M. F. Pac, S. Veeraraghavan. 2011. Quality-speed conundrum: trade-offs in customer-intensive services. *Management Science* **57**(1) 40–56.

Anderson, S. P., A. de Palma, J. F. Thisse. 1992. *Discrete choice theory of product differentiation*. MIT press.

Ariely, D. 2008. *Predictably irrational: The hidden forces that shape our decisions*. Harper Collins.

Barnes, D., V. Mookherjee. 2009. Customer delay in e-Commerce sites: Design and strategic implications. G. Adomavicius, A. Gupta, eds., *Business Computing*, chap. 5. Emerald Group Publishing Limited, 117–138.

Bazaraa, M. S., H. D. Sherali, C. M. Shetty. 2013. *Nonlinear programming: theory and algorithms*. 3rd ed. John Wiley & Sons.

Boyaci, T., S. Ray. 2003. Product differentiation and capacity cost interaction in time and price sensitive markets. *Manufacturing and Service Operations Management* **5**(1) 18–36.

Cox, D. R., W. L. Smith. 1991. *Queues*. CRC Press.

eMC. 2014. Shipping costs impact purchase decisions, customer satisfaction. Http://www.emarketingandcommerce.com/article/shipping-costs-impact-purchase-decisions-customer-satisfaction/1#.

Forbes. 2014. Amazon takes one step forward, two back. http://www.forbes.com/sites/stevebanker/2014/02/20/amazon-takes-one-step-forward-two-back/.

Gross, D. 2008. *Fundamentals of queueing theory*. John Wiley & Sons.

Gümüş, M., S. Li, W. Oh, S. Ray. 2013. Shipping fees or shipping free? A tale of two price partitioning strategies in online retailing. *Production and Operations Management* **22**(4) 758–776.

Hamilton, R. W., J. Srivastava, A. T. Abraham. 2010. When should you nickel-and-dime your customers? *MIT Sloan Management Review* **52**(1) 59–67.

Hassin, R., M. Haviv. 2003. *To queue or not to queue: Equilibrium behavior in queueing systems*, vol. 59. Kluwer Academic Publishers.

Hsu, V. N., S. H. Xu, B. Jukic. 2009. Optimal scheduling and incentive compatible pricing for a service system with quality of service guarantees. *Manufacturing and Service Operations Management* **11**(3) 375–396.

Investor Press Release. 2013. Overstock.com offers free shipping on orders over $50. http://investors.overstock.com/phoenix.zhtml?c=131091&p=irol-newsArticle_pf&ID=1771702&highlight=.

Jain, N., A. Paul. 2001. A generalized model of operations reversal for fashion goods. *Management Science* **47**(4) 595–600.

Jayaswal, S., E. Jewkes, S. Ray. 2011. Product differentiation and operations strategy in a capacitated environment. *European Journal of Operational Research* **210**(3) 716–728.

Kannan, P. K., B. K. Pope, S. Jain. 2009. Pricing digital content product lines: A model and application for the National Academies Press. *Marketing Science* **28**(4) 620–636.

Katta, A., J. Sethuraman. 2005. Pricing strategies and service differentiation in queues-A profit maximization perspective. Tech. Rep. TR-2005-04, CORC, Columbia University.

Lederer, P. J., L. Li. 1997. Pricing, production, scheduling, and delivery-time competition. *Operations Research* **45**(3) 407–420.

Leng, M., R. Becerril-Arreola. 2010. Joint pricing and contingent free-shipping decisions in B2C transactions. *Production and Operations Management* **19**(4) 390–405.

Leng, M., M. Parlar. 2005. Free shipping and purchasing decisions in B2B transactions: A game-theoretic analysis. *IIE Transactions* **37**(12) 1119–1128.

Li, L., L. Jiang, L. Liu. 2012. Service and price competition when customers are naive. *Production and Operations Management* **21**(4) 747–760.

Maglaras, C., J. Yao, A. Zeevi. 2013. Optimal price and delay differentiation in queueing systems. Columbia University working paper.

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science* **49**(8) 1018–1038.

Moorthy, K. S. 1984. Market segmentation, self-selection, and product line design. *Marketing Science* **3**(4) 288–307.

Morwitz, V., E. Greenleaf, E. Shalev E. J. Johnson. 2014. The price does not include additional taxes, fees, and surcharges: a review of research on partitioned pricing. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1350004.

Morwitz, V. G., E. A. Greenleaf, E. J. Johnson. 1998. Divide and Prosper: Consumers reactions to partitioned prices. *Journal of Marketing Research* **35** 453–463.

Muther, C. 2013. Instant gratification is making us perpetually impatient. Https://www.bostonglobe.com/lifestyle/style/2013/02/01/the-growing-culture-impatience-where-instant-gratification-makes-crave-more-instant-gratification/q8tWDNGeJB2mm45fQxtTQP/story.html.

Overstock.com. 2014. Shipping & Delivery. https://help.overstock.com/app/answers/detail/a_id/7/c/2.

Pashigian, B. P. 1988. Demand uncertainty and sales: A study of fashion and markdown pricing. *The American Economic Review* **78**(5) 936–953.

PayPal. 2009. New PayPal survey reveals why online shoppers abandon purchases. Https://www.paypal-media.com/press-releases/20090623005580.

Rao, S., E. R. Peterson. 1998. Optimal pricing of priority services. *Operations Research* **46**(1) 46–56.

Ray, S., E. M. Jewkes. 2004. Customer lead time management when both demand and price are lead time sensitive. *European Journal of Operational Research* **153**(3) 769–781.

Shampanier, K., N. Mazar, D. Ariely. 2007. Zero as a special price: The true value of free products. *Marketing Science* **26**(6) 742–757.

So, K. C., J. S. Song. 1998. Price, delivery time guarantees, and capacity selection. *European Journal of Operational Research* **111**(1) 28–49.

Van Mieghem, J. A. 2000. Price and service discrimination in queuing systems: Incentive compatibility of $Gc\mu$ scheduling. *Management Science* **46**(9) 1249–1267.

Zappos.com. 2014. Shipping and delivery questions. http://www.zappos.com/shipping-and-delivery-questions.

Zhang, Z., D. Dey, Y. Tan. 2007. Pricing communication services with delay guarantee. *INFORMS Journal on Computing* **19**(2) 248–260.

Zhao, X., K. E. Stecke, A. Prasad. 2012. Lead time and price quotation mode selection: Uniform or differentiated? *Production and Operations Management* **21**(1) 177–193.

Appendix A: Proofs

Proof of Lemma 1: We show that if service differentiation is optimal then $\gamma = 1$, $\gamma_2 = 0$ has to be the best solution. First, note that $d_l > 0 = d_h$ due to service differentiation, and $\gamma_1 = 0,\ \gamma_2 = 1$ is infeasible $\forall d_l > 0$. Either γ_1 or γ_2 is strictly in between zero and one because there is no service differentiation otherwise.

Suppose $0 < \gamma_1 < 1$, then the IC constraints (3) and (4) imply that $p_h - p_l = \eta_1 d_l > \eta_2 d_l$. Therefore, $\gamma_2 = 0$ because otherwise constraint (3) with $i = 2$ would be violated. Then the net cost becomes $\lambda r - \lambda\gamma_1 q p_h - \lambda(1 - \gamma_1 q)p_l$, which is decreasing in γ_1. Further, because $W(\mu)$ is independent of γ_1, the feasible region is independent of γ_1. So the optimal net cost is decreasing in γ_1. It further decreases when $\gamma_1 = 1$ because the constraint (4) no longer applies to Type 1 customers. Hence, the optimal net cost when $\gamma_1 = 1$ and $\gamma_2 = 0$ is less than that under any γ_1 s.t. $0 < \gamma_1 < 1$ and $\gamma_2 = 0$.

Similarly, suppose $0 < \gamma_2 < 1$ then $p_h - p_l = \eta_2 d_l < \eta_1 d_l$ and therefore, $\gamma_1 = 1$ (otherwise constraint (4) with $i = 1$ would be violated). And the net cost becomes $\lambda r - \lambda\left(1 - (1-\gamma_2)(1-q)\right)p_h - \lambda(1-\gamma_2)(1-q)p_l = \lambda r - \lambda q p_h - \lambda(1-q)p_l - \lambda\gamma_2(1-q)(p_h - p_l)$ which is decreasing in γ_2. So the optimal net cost is decreasing in γ_2. Hence, the optimal net cost when $\gamma_1 = 1$ and $\gamma_2 = 1$ is less than that under $\gamma_1 = 1$ and any γ_2 s.t. $0 < \gamma_2 < 1$. However, there is *no service differentiation* when $\gamma_1 = \gamma_2 = 1$. ∎

Proof of Lemma 2: Suppose there is service differentiation. Because of Lemma 1, $\mathcal{N}$ gets simplified to $\min_{p_h, p_l, d_l} \lambda r - \lambda q p_h - \lambda(1-q)p_l$ s.t. $p_h + \eta_1 W(\mu) \le v_1$, $p_l + \eta_2(W(\mu) + d_l) \le v_2$, $p_h - p_l \le \eta_1 d_l$, $p_h - p_l \ge \eta_2 d_l$, $d_l \ge 0$, and $p_l \ge 0$. Letting u_i denote the non-negative multiplier for the i^{th} constraint, KKT conditions (Bazaraa et al. 2013) imply that $u_1 + u_3 - u_4 = \lambda q$, $u_2 - u_3 + u_4 = \lambda(1-q) + u_6$, and $\eta_2(u_2 + u_4) = \eta_1 u_3 + u_5$. There is service differentiation only if $d_l > 0$ and hence $u_5 = 0$; otherwise the retailer's cost is $\max_{i=1,2} \lambda r - \lambda v_i + \lambda\eta_i W(\mu)$. Further, the second condition implies that $u_2 + u_4 > 0$ and hence $u_3 > 0$ from the third condition. So $p_h - p_l = \eta_1 d_l > \eta_2 d_l$ under optimality when there is service differentiation and hence $u_4 = 0$. Then solving for u_1, u_2, and u_3 from the three conditions, we get $u_1 = \frac{\lambda(\eta_1 q - \eta_2) - \eta_2 u_6}{(\eta_1 - \eta_2)}$, $u_2 = \frac{(\lambda(1-q) + u_6)\eta_1}{(\eta_1 - \eta_2)}$, and $u_3 = \frac{(\lambda(1-q) + u_6)\eta_2}{(\eta_1 - \eta_2)}$. We find that $u_1 \ge 0$ for some $u_6 \ge 0$ implies that $\Delta \ge 1/q$. There are then two possibilities for the optimal solution. First, $p_l > 0$ and hence $u_6 = 0$; then the first three constraints bind and $p_h = v_1 - \eta_1 W(\mu)$, $p_l = (\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2)$, and $d_l = (v_1 - v_2)/(\eta_1 - \eta_2) - W(\mu)$. Further for it to be feasible, $\Delta > \nu$ (since $p_l > 0$) and $W(\mu) < (v_1 - v_2)/(\eta_1 - \eta_2)$ (since $d_l > 0$). The corresponding cost is given by $\lambda r - \lambda\left(\frac{(\eta_1 q - \eta_2)v_1 + (1-q)\eta_1 v_2}{\eta_1 - \eta_2}\right) + \lambda\eta_1 q W(\mu)$. This case corresponds to case *(iii)* in the lemma. Second, $p_l = 0$; then the second and third constraints are binding because $u_2, u_3 > 0$, and $p_h = \eta_1\left(v_2/\eta_2 - W(\mu)\right)$ and $d_l = v_2/\eta_2 - W(\mu)$. Further, for it to be feasible $\Delta \le \nu$ (since $p_h + \eta_1 W(\mu) \le v_1$) and $W(\mu) \le v_2/\eta_2$ ($\because d_l \ge 0$). The corresponding cost is given by $\lambda r - \lambda\left(\eta_1 q/\eta_2\right)v_2 + \lambda\eta_1 q W(\mu)$. This case corresponds to case *(v)* in the lemma.

If conditions above for both cases *(iii)* and *(v)* are not satisfied, then there is no service differentiation. Then the retailer charges a single price $p_h = p_l = \min(v_1 - \eta_1 W(\mu), v_2 - \eta_2 W(\mu))$ to minimize the net cost while providing adequate service to everyone. If $\Delta < \mu$ then this price is $v_2 - \eta_2 W(\mu)$ since $W(\mu) \le v_2/\eta_2 < (v_1 - v_2)/(\eta_1 - \eta_2)$, which corresponds to case *(iv)* in the lemma. If $\Delta \ge \mu$ then the price depends on how $W(\mu)$ and $(v_1 - v_2)/(\eta_1 - \eta_2)$ compare with each other, which leads to cases *(i)* and *(ii)* in the lemma. ∎

Proof of Lemma 3: If $0 < \gamma_1 < 1$ then $p_h - p_l = \eta_1 \left(W_l + d_l - W_h - d_h\right)$. Due to service differentiation, either (i) $W_h + d_h < W_l + d_l$ which implies that $p_h - p_l > \eta_2 \left(W_l + d_l - W_h - d_h\right)$ and therefore $\gamma_2 = 0$ because otherwise the IC constraint of Type 2 customers for high-type service would be violated, or (ii) $W_h + d_h > W_l + d_l$ which implies that $p_h - p_l < \eta_2 \left(W_l + d_l - W_h - d_h\right)$ and therefore $\gamma_2 = 1$ because otherwise the IC constraint of Type 2 customers for low-type service would be violated. Similarly, we can show that if $0 < \gamma_2 < 1$, either $\gamma_1 = 0$ or $\gamma_1 = 1$. $\blacksquare$

Proof of Lemma 4: From the analysis of problem $\mathcal{P}5$ (see Appendix B), we find that it is feasible if and only if $W_l \leq v_2/\eta_2$ and $(\eta_1 - \eta_2)W_h + \eta_2 W_l \leq v_1$. We now show that (i) if $\mathcal{P}5$ is feasible then it dominates $\mathcal{P}6$ and (ii) if it is infeasible, then $\mathcal{P}1$ with $\gamma_1 = \tilde{\gamma}_1$ dominates $\mathcal{P}6$. This result and Lemmas B1-B4 imply that (a) $\gamma_2 = 0$ and (b) if $W_l(\mu, q) > \min_{i=1,2} v_i/\eta_i$ then $\gamma_1 = \tilde{\gamma}_1$ else $\gamma_1 = 1$.

Suppose $\mathcal{P}5$ is feasible. Then if $\Delta \geq \nu$, because $W_l(\mu, 1-q) \leq (v_1 - v_2)/(\eta_1 - \eta_2)$ and $\Delta > 1/q$ for service differentiation to be optimal in $\mathcal{P}6$ (from Lemma B6), $W_h(\mu, q) < W_l(\mu, 1-q)$ implies that only cases (ii) and (iv) of Lemma B5 apply. Comparing net costs in case (ii) and $\mathcal{P}6$, we have $\lambda r - \lambda q v_1 - \lambda(1-q)v_2 + \lambda\eta_1 q W_h(\mu, q) + \lambda\eta_2(1-q)W_l(\mu, q) < \lambda r - \lambda q v_1 - \lambda(1-q)(\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2) + \lambda\eta_1 q W_l(\mu, 1-q)$, $\Leftarrow \lambda\eta_2(1-q)\left(W_l(\mu, q) - \frac{v_1 - v_2}{\eta_1 - \eta_2}\right) < \lambda\eta_1 q \left(W_l(\mu, 1-q) - W_h(\mu, q)\right)$, $\Leftarrow \eta_2(1-q)\left(W_l(\mu, q) - \frac{v_1 - v_2}{\eta_1 - \eta_2}\right) < \eta_2 \left(W_l(\mu, 1-q) - W_h(\mu, q)\right)$ $(\because\ \eta_2 < \eta_1 q, W_l \geq W_h)$, $\Leftarrow W(\mu) + (1-q)\left(W_l(\mu, q) - \frac{v_1 - v_2}{\eta_1 - \eta_2}\right) < W_l(\mu, 1-q)$, which is true. Also, because $W_h(\mu, q) < W_l(\mu, 1-q)$, the optimal net cost in case (iv) is strictly less than that in $\mathcal{P}6$. If $\Delta < \nu$ then only case (vi) of Lemma B5 applies and $\lambda r + \lambda\eta_1 q \left(W_h(\mu, q) - v_2/\eta_2\right) < \lambda r + \lambda\eta_1 q \left(W_l(\mu, q) - v_2/\eta_2\right)$. Hence, if $\mathcal{P}5$ is feasible, it dominates $\mathcal{P}6$.

Suppose $\mathcal{P}5$ is infeasible. From Lemma B1, $\mathcal{P}1$ is feasible and $\tilde{\gamma}_1$ is the best γ_1 value. If $\Delta \geq \nu$, then $W_h(\mu, 1-q) < W_l(\mu, 1-q) \leq \frac{v_1 - v_2}{\eta_1 - \eta_2}$ (from Lemma B6). Also, $W_l(\mu, \tilde{\gamma}_1 q) = v_1/\eta_1$ so that the optimal values in $\mathcal{P}1$ are given by $p_l = d_l = d_h = 0$ (from case *(i)* of Lemma B1). Optimal net cost of $\mathcal{P}1$ is less than that of $\mathcal{P}6$ because $\lambda r - \lambda v_1 + \lambda\eta_1 W < \lambda r - \lambda q v_1 - \lambda(1-q)(\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2) + \lambda\eta_1 q W_l(\mu, 1-q)$ as $\eta_1 W - \eta_1 q W_l(\mu, 1-q) = \eta_1(1-q)W_h(\mu, 1-q) < (1-q)\left(v_1 - (\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2)\right) = \eta_1(1-q)\left(\frac{v_1 - v_2}{\eta_1 - \eta_2}\right)$.

Finally, if $\Delta < \nu$ note that $W_h(\mu, 1-q) < W_l(\mu, 1-q) \leq v_2/\eta_2$ (from Lemma B6). Also, $W_l(\mu, \tilde{\gamma}_1 q) = v_2/\eta_2$ so that the optimal values in $\mathcal{P}1$ are given by $p_l = d_l = d_h = 0$ (from cases *(iv)* and *(v)* of Lemma B1, both of which become equivalent). The optimal net cost of $\mathcal{P}1$ (from case *(iv)* of Lemma B1) is $\lambda r - \lambda\eta_1(v_2/\eta_2) + \lambda\eta_1 W(\mu) < \lambda r + \lambda\eta_1 q \left(W_l(\mu, 1-q) - v_2/\eta_2\right)$ because $\eta_1 W(\mu) - \eta_1 q W_l(\mu, 1-q) = \eta_1(1-q)W_h(\mu, 1-q) < \eta_1(1-q)v_2/\eta_2$. Thus if $\mathcal{P}5$ is infeasible, $\mathcal{P}1$ (with $\gamma_1 = \tilde{\gamma}_1$) dominates $\mathcal{P}6$. $\blacksquare$

Proof of Theorem 1 and Corollary 1: Comparing the net costs in Lemmas 2 and 4 gives the result. No prioritization does as well as prioritization if and only if $\Delta \geq \nu$ and $P5$ is infeasible. The results in the corollary are obtained from using Lemma 4. Furthermore, when prioritized service is optimal, we find that $\mathcal{P}5$ is infeasible if and only if $\Delta < \nu$ and $W_l(\mu, q) > v_2/\eta_2$. Then there is a split of impatient customers with $\tilde{\gamma}_1 q$ customers selecting high-priority service (as in Lemma 4). Otherwise, $\mathcal{P}5$ is feasible and results from Lemma B5 apply. $\blacksquare$

Proof of Lemma 5: We first note from Theorem 1 that when the split is optimal, $\Delta < \nu$

and the optimal values are given by $p_h = \eta_1 (v_2/\eta_2 - W_h(\mu, \tilde{\gamma}_1 q))$ and $p_l = d_h = d_l = 0$, with $W_l(\mu, \tilde{\gamma}_1 q) = v_2/\eta_2$. First, we find that $\gamma_1 \not> \tilde{\gamma}_1$ because otherwise $p_l + \eta_2 W_l(.) > \eta_2 W_l(\tilde{\gamma}_1 q) = v_2$ which violates the IR constraint for patient customers. Also, for such γ_1 we have $\gamma_2 = 0$ since $p_h + \eta_2 W_h(\mu, \gamma_1 q) > p_l + \eta_2 W_l(\mu, \gamma_1 q)$. The proof then follows from the property that $W_l(\mu, \gamma_1 q) - W_h(\mu, \gamma_1 q)$ is increasing in γ_1 (since $\partial W_h(\mu, \delta)/\partial \delta < \partial W_l(\mu, \delta)/\partial \delta$). Hence, we cannot have an equilibrium with $0 < \gamma_1 < \tilde{\gamma}_1$ because then $p_h + \eta_1 W_h(\mu, \gamma_1 q) > p_l + \eta_1 W_l(\mu, \gamma_1 q)$ and the IC constraint for impatient customers choosing high-type service is violated. However, we find that when $\gamma_1 = 0$, the optimal values above satisfy all the necessary IC and IR conditions, and so it is also an equilibrium. █

Proof of Lemma C1: Solving $W_h(\mu, q) = (v_1 - v_2)/(\eta_1 - \eta_2)$ and $W_l(\mu, q) = v_2/\eta_2$, we get a unique solution with $\tilde{\mu} = \frac{\lambda\left(\frac{v_2/\eta_2}{(v_1-v_2)/(\eta_1-\eta_2)}\right)}{\frac{v_2/\eta_2}{(v_1-v_2)/(\eta_1-\eta_2)}-1}$ and $\tilde{q} = \frac{1}{1-\frac{(v_1-v_2)/(\eta_1-\eta_2)}{v_2/\eta_2}}\left(1 - \frac{1}{\lambda}\left(\frac{\eta_1-\eta_2}{v_1-v_2} - \frac{\eta_2}{v_2}\right)\right)$. Note that $W_h(\mu, q) = (v_1 - v_2)/(\eta_1 - \eta_2)$ implies $\mu = \lambda q + (\eta_1 - \eta_2)/(v_1 - v_2)$ so that the slope of W_h curve is $1/\lambda$. Further, the slope of W_l curve is given by $\frac{-\partial W_l/\partial \mu}{\partial W_l/\partial q} = \frac{\frac{1}{(\mu-\lambda q)^2}+\frac{\lambda(\mu-\lambda+\mu-\lambda q)}{(\mu-\lambda q)^2(\mu-\lambda)^2}}{\frac{\mu\lambda}{(\mu-\lambda)(\mu-\lambda q)^2}} > \frac{\frac{1}{(\mu-\lambda q)^2}+\frac{\lambda(\mu-\lambda)}{(\mu-\lambda q)^2(\mu-\lambda)^2}}{\frac{\mu\lambda}{(\mu-\lambda)(\mu-\lambda q)^2}} = \frac{1}{\lambda}$. Then there are three possibilities: (i) $\tilde{\mu} < W^{-1}(v_1/\eta_1) = \lambda + \eta_1/v_1$ in which the W_h curve is always below the W_l curve, (ii) $\lambda + \eta_1/v_1 \leq \tilde{\mu}$ and $\tilde{q} \leq 1$ in which there is a unique intersection, and (iii) $\tilde{q} > 1$ in which the W_l curve is always below the W_h curve. █

Proof of Theorem 2: Suppose $\Delta \geq \nu$. From the results in Theorem 1 (and Lemmas 4 and B5) we find that there are two *general* ways in which optimal service delivery can evolve as capacity μ increases: (i) Single service → Differentiated service in which patient customers get surplus → Differentiated service in which no one gets surplus → Differentiated service with strategic delay (in which no one gets surplus) (ii) Single service → Differentiated service in which patient customers get surplus → Differentiated service in which no one gets surplus → Differentiated service (without strategic delay) in which impatient customers get surplus. Although there can be other patterns in which the optimal service delivery evolves, they are *subsequences* of at least one of the above patterns. We show that the optimal total cost is strictly convex with both patterns (so that it would be strictly convex with any subsequence pattern too). In the first pattern (for which $\Delta > 1/q$ is necessary), the four total profit functions (as capacity increases) are given by $\mu K + \lambda r - \lambda v_1 + \lambda \eta_1 W$, $\mu K + \lambda r - \lambda v_1 + \lambda \eta_2 W + \lambda(\eta_1 - \eta_2) W_h$, $\mu K + \lambda r - \lambda q v_1 - \lambda(1-q) v_2 + \lambda \eta_1 q W_h + \lambda \eta_2 (1-q) W_l$, and $\mu K + \lambda r - \lambda q v_1 - \lambda(1-q)\frac{\eta_1 v_2 - \eta_2 v_1}{\eta_1 - \eta_2} + \lambda \eta_1 q W_h$. The corresponding derivatives (with respect to μ) are related as follows: $K - \frac{\lambda\eta_1}{(\mu-\lambda)^2} < K - \frac{\lambda\eta_2}{(\mu-\lambda)^2} - \frac{\lambda(\eta_1-\eta_2)}{(\mu-\lambda q)^2} < K - \frac{\lambda\eta_2}{(\mu-\lambda)^2} - \frac{\lambda(\eta_1-\eta_2)q}{(\mu-\lambda q)^2} < K - \frac{\lambda\eta_1 q}{(\mu-\lambda q)^2}$. So the total cost is strictly convex in μ, and the optimal capacity is unique. In the second pattern (for which $\Delta \leq 1/q$ is necessary), the first three total cost functions are the same while the fourth one is given by $\mu K + \lambda r - \lambda v_2 + \lambda \eta_1 W - \lambda(\eta_1 - \eta_2) W_l$. The derivative of this function is greater than the derivative of the third total cost function if

$$\begin{aligned} & K - \frac{\lambda\eta_1}{(\mu-\lambda)^2} + \frac{\lambda(\eta_1-\eta_2)}{(1-q)(\mu-\lambda)^2} - \frac{\lambda q(\eta_1-\eta_2)}{(1-q)(\mu-\lambda q)^2} > K - \frac{\lambda\eta_2}{(\mu-\lambda)^2} - \frac{\lambda(\eta_1-\eta_2)q}{(\mu-\lambda q)^2} \\ \Leftrightarrow \quad & \frac{1}{(1-q)(\mu-\lambda)^2} - \frac{q}{(1-q)(\mu-\lambda q)^2} > \frac{1}{(\mu-\lambda)^2} - \frac{q}{(\mu-\lambda q)^2} \\ \Leftrightarrow \quad & \frac{1}{(\mu-\lambda)^2} > \frac{q}{(\mu-\lambda q)^2}, \end{aligned}$$

which is true. So the total cost is strictly convex and optimal capacity is unique, with the second pattern too.

Suppose $\Delta < \nu$. Then we find (from Theorem 1 and Lemmas 4 and B5) that the general pattens of optimal service delivery are (i) Differentiated service with split $\rightarrow$ Differentiated service with strategic delay and (ii) Differentiated service with split $\rightarrow$ Differentiated service without strategic delay. In the first pattern (for which $\Delta > 1/q$ is necessary), the total costs are $\mu K + \lambda r - \lambda\eta_1 \cdot v_2/\eta_2 + \lambda\eta_1 W$ and $\mu K + \lambda r + \lambda\eta_1 q\,(W_h - v_2/\eta_2)$. Their derivatives are such that $K - \frac{\lambda\eta_1}{(\mu-\lambda)^2} < K - \frac{\lambda\eta_1 q}{(\mu-\lambda q)^2}$. So the total cost is strictly convex and has a unique optimal capacity. In the second pattern (for which $\Delta \leq 1/q$ is necessary), the second total cost is $\mu K + \lambda r - \lambda v_2 + \lambda\eta_1 W - \lambda(\eta_1-\eta_2)W_l$. The derivative of this function is greater than the derivative of the first total cost function if $-\frac{\partial}{\partial\mu}\left(\lambda(\eta_1-\eta_2)W_l\right) > 0$ which is true since $\frac{\partial W_l}{\partial\mu} < 0$. So the total cost is strictly convex and the optimal capacity is unique, with the second pattern too.

We next consider what happens when $\lambda \rightarrow \infty$. We first show that single service, strategic delay, and split are all sub-optimal. When λ is high enough, case (iii) of Lemma C1 applies. So the W_l curve is always below the W_h curve (see Figure C1c for an illustration) and $W_l(\mu, q) = \frac{\mu}{(\mu-\lambda q)(\mu-\lambda)} > v_2/\eta_2$ for single service to be optimal. That means $1/(\mu-\lambda) > v_2/\eta_2(1-q)$ so that the derivative of total cost under single service given by $K - \frac{\lambda\eta_1}{(\mu-\lambda)^2} < K - \frac{\lambda\eta_1 v_2^2(1-q)^2}{\eta_2^2} < 0$ as $\lambda \rightarrow \infty$. So $W_l \leq v_2/\eta_2$. Strategic delay is sub-optimal since the derivative (both when $\Delta \geq \nu$ and $\Delta < \nu$) is given by $K - \lambda\eta_1 q/(\mu-\lambda q)^2 > K - \lambda\eta_1 q/(\lambda-\lambda q)^2 = K - \eta_1 q/\left(\lambda(1-q)^2\right) > 0$ as $\lambda \rightarrow \infty$. Since $W_l > v_2/\eta_2$ for split of impatient customers and the derivative of total cost is the same as that under single service, it is also sub-optimal.

Free service is determined by whether $W_l = v_2/\eta_2$. We denote the corresponding capacity by $\underline{\mu}$ so $1/(\underline{\mu}-\lambda) > v_2/\eta_2(1-q)$. If $\Delta > \nu$ then the right side derivative, which corresponds to the total cost function for differentiated service in which no one gets surplus, is given by $K - \frac{\lambda\eta_2}{(\underline{\mu}-\lambda)^2} - \frac{\lambda(\eta_1-\eta_2)q}{(\underline{\mu}-\lambda q)^2} < K - \frac{\lambda\eta_2}{(\underline{\mu}-\lambda)^2} < K - \frac{\lambda v_2^2(1-q)^2}{\eta_2} < 0$ as $\lambda \rightarrow \infty$. Hence the optimal capacity is higher than $\underline{\mu}$ and so low-priority service is charged. If $1/q < \Delta < \nu$ then the right side derivative (at $\underline{\mu}$) corresponds to that under strategic delay and it is positive (as shown above). So optimal capacity is $\underline{\mu}$ and low-priority service is free. $\blacksquare$

Appendix B: Optimal Solutions of Problems $\mathcal{P}1$-$\mathcal{P}6$

Problem $\mathcal{P}1$: $0 < \gamma_1 < 1$ and $\gamma_2 = 0$

The impatient customers split between high- and low-type services while all the patient customers select low-type service. Because $0 < \gamma_1 < 1$ and $\gamma_2 = 0$, the IC constraints imply that $p_h - p_l = \eta_1 (W_l + d_l - W_h - d_h) \geq \eta_2 (W_l + d_l - W_h - d_h)$. So $W_h + d_h \leq W_l + d_l$ and the high-type service is the *costlier* service class. Problem $\mathcal{P}1$ for any γ_1 can be written as $\min_{p_l,d_h,d_l} \lambda r - \lambda p_l - \lambda\gamma_1\eta_1 q (W_l + d_l - W_h - d_h)$ s.t. $p_l + \eta_i(W_l + d_l) \leq v_i$, $i = 1, 2$; $W_h + d_h \leq W_l + d_l$; $d_h, d_l, p_l \geq 0$. Lemma B1 characterizes the optimal solution of $\mathcal{P}1$. It shows that when $\gamma_2 = 0$, the net cost decreases because more impatient customers select the high-type service. Although the price of high-type service (low-type service) can decrease as in case (ii) (case (iv)) of Lemma B1, the effect of this price reduction on the net cost is more than compensated by the benefit from additional customers who choose the more expensive high-type service.

Lemma B1 *For problem $\mathcal{P}1$, $d_h = 0$ under optimality. Other optimal values are given by:*
(i) if $\Delta \geq \nu$ and $(v_1 - v_2)/(\eta_1 - \eta_2) \leq W_l \leq v_1/\eta_1$, then $p_h = v_1 - \eta_1 W_h$, $p_l = v_1 - \eta_1 W_l$, $d_l = 0$, and the net cost is $\lambda r - \lambda v_1 + \lambda\eta_1 W$;
(ii) if $\Delta \geq \nu$, $W_l < (v_1 - v_2)/(\eta_1 - \eta_2)$, and $\Delta \leq 1/(\gamma_1 q)$ then $p_h = v_2 - \eta_2 W_l + \eta_1(W_l - W_h)$, $p_l = v_2 - \eta_2 W_l$, $d_l = 0$, and the net cost is $\lambda r - \lambda v_2 + \lambda\eta_1 W - \lambda(\eta_1 - \eta_2)W_l$;
(iii) if $\Delta \geq \nu$, $W_l < (v_1 - v_2)/(\eta_1 - \eta_2)$, and $\Delta > 1/(\gamma_1 q)$ then $p_h = v_1 - \eta_1 W_h$, $p_l = (\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2)$, $d_l = (v_1 - v_2)/(\eta_1 - \eta_2) - W_l$, and the net cost is $\lambda r - \lambda v_1 + \lambda\eta_1 W + \lambda\eta_1(1 - \gamma_1 q)\left((v_1 - v_2)/(\eta_1 - \eta_2) - W_l\right)$;
(iv) if $\Delta < \nu$, $W_l \leq v_2/\eta_2$ and $\Delta \leq 1/(\gamma_1 q)$, $p_h = v_2 - \eta_2 W_l + \eta_1(W_l - W_h)$, $p_l = v_2 - \eta_2 W_l$, $d_l = 0$, and the net cost is $\lambda r - \lambda v_2 + \lambda\eta_1 W - \lambda(\eta_1 - \eta_2)W_l$; and
(v) if $\Delta < \nu$, $W_l \leq v_2/\eta_2$ and $\Delta > 1/(\gamma_1 q)$, $p_h = \eta_1 (v_2/\eta_2 - W_h)$, $p_l = 0$, $d_l = v_2/\eta_2 - W_l$, and the net cost is $\lambda r + \lambda\eta_1\gamma_1 q (W_h - v_2/\eta_2)$.
There is always service differentiation and the optimal net cost is decreasing in γ_1. So either (i) $W_l(\mu, q) > \min_{i=1,2} v_i/\eta_i$ and the best γ_1 is $\tilde{\gamma}_1$ (note that $W_l(\mu, \tilde{\gamma}_1 q) = \min_{i=1,2} v_i/\eta_i$), or (ii) $\mathcal{P}1$ is dominated by $\mathcal{P}5$.

Proof For $\mathcal{P}1$, $d_h = 0$ under optimality because otherwise just reducing it decreases the net cost. Then $\mathcal{P}1$ becomes $\min_{p_l,d_l} \lambda r - \lambda p_l - \lambda\gamma_1\eta_1 q (W_l + d_l - W_h)$ s.t. $p_l + \eta_i(W_l + d_l) \leq v_i$; $p_l, d_l \geq 0$. $\mathcal{P}1$ is similar to $\mathcal{N}$ (with $\gamma_1 = 1$ and $\gamma_2 = 0$): we have W_l instead of W and the IC constraint for high-type service (low-type service) is binding (non-binding) under optimality. Optimal values of p_l and d_l are found by using W_l instead of W in Lemma 2. Using $p_h = p_l + \eta_1 (W_l + d_l - W_h - d_h)$ and substituting the optimal values, we get the optimal net cost in Lemma B1. After some algebra, we find that it decreases with γ_1 in cases (i)-(iv). In case (v), its derivative wrt γ_1 is given by $\lambda\eta_1 q (W_h - v_2/\eta_2) + \lambda\eta_1\gamma_1 q^2 \partial W_h/\partial\delta$. Next, we show that it is negative. Work conservation implies that $\gamma_1 q W_h(\mu, \gamma_1 q) + (1 - \gamma_1 q)W_l(\mu, \gamma_1 q) = W(\mu)$. Differentiating wrt γ_1, we get $qW_h + \gamma_1 q^2 \partial W_h/\partial\delta - qW_l + q(1 - \gamma_1 q)\partial W_l/\partial\delta = 0$; $\Rightarrow$ $q(W_h - W_l) + \gamma_1 q^2 \partial W_h/\partial\delta < 0$ ($\because \partial W_l/\partial\delta > 0$); $\Rightarrow \lambda\eta_1 q (W_h - v_2/\eta_2) + \lambda\eta_1\gamma_1 q^2 \partial W_h/\partial\delta < 0$ ($\because W_l \leq v_2/\eta_2$ in case (v) and $\lambda\eta_1 > 0$). Because the optimal net cost is continuous, it is always decreasing in γ_1. Hence, the best γ_1 equals (i) $\tilde{\gamma}_1$ at which $\mathcal{P}1$ is just feasible or (ii) one. If $\mathcal{P}1$ is feasible with $\gamma_1 = 1$ then $\mathcal{P}5$, in which the IC constraint for high-type service does not have to bind, has a better net cost. $\blacksquare$

Problem $\mathcal{P}2$: $0 < \gamma_1 < 1$ and $\gamma_2 = 1$

The impatient customers split between high- and low-type services while the all the patient customers select high-type service. Because $0 < \gamma_1 < 1$ and $\gamma_2 = 1$, the IC constraints become $p_h - p_l = \eta_1 (W_l + d_l - W_h - d_h) \le \eta_2 (W_l + d_l - W_h - d_h)$. So $W_h + d_h \ge W_l + d_l$ and the high-type service is the *cheaper* service class. Problem $\mathcal{P}2$ for any γ_1 can be written as $\min_{p_h,d_h,d_l} \lambda r - \lambda p_h - \lambda\eta_1(1-\gamma_1)q\,(W_h + d_h - W_l - d_l)$ s.t. $p_h + \eta_i(W_h + d_h) \le v_i$, $i = 1, 2$; $W_l + d_l \le W_h + d_h$; $d_h, d_l, p_h \ge 0$. Lemma B2 shows that when $\gamma_2 = 1$, an increase in the fraction of impatient customers selecting the high-type service increases the net cost. This result has an intuitive explanation. The prioritized high-type service is the cheaper service class here. Hence, a higher γ_1 means more customers that are paying less. Further, giving priority to more customers increases the expected waiting time of low-type service, which reduces p_l. Both these factors increase the net cost.

Lemma B2 *For problem $\mathcal{P}2$, $d_l = 0$ under optimality. Service is differentiated in the following cases:*
(i) if $\Delta \ge \nu$, $W_l < (v_1 - v_2)/(\eta_1 - \eta_2)$, and $\Delta \ge \frac{1}{(1-\gamma_1)q}$ then $p_h = (\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2)$, $d_h = (v_1 - v_2)/(\eta_1 - \eta_2) - W_h$, $p_l = v_1 - \eta_1 W_l$, and the net cost is $\lambda r - \lambda v_1 + \lambda\eta_1(1 - q + \gamma_1 q)\,((v_1 - v_2)/(\eta_1 - \eta_2) + W_l)$; and
(ii) if $\Delta < \nu$, $W_l < v_2/\eta_2$ and $\Delta \ge \frac{1}{(1-\gamma_1)q}$ then $p_h = 0$, $p_l = \eta_1\,(v_2/\eta_2 - W_l)$, $d_h = v_2/\eta_2 - W_h$, and the net cost is $\lambda r - \lambda\eta_1(1-\gamma_1)q\,(v_2/\eta_2 - W_l)$.
Further, the net cost is increasing in γ_1 and hence $\mathcal{P}2$ is dominated by $\mathcal{P}6$.

Proof For $\mathcal{P}2$, $d_l = 0$ under optimality; otherwise $d_h, d_l > 0$ which is sub-optimal. Then $\mathcal{P}2$ becomes $\min_{p_h,d_h} \lambda r - \lambda p_h - \lambda\eta_1(1-\gamma_1)q\,(W_h + d_h - W_l)$ s.t. $p_h + \eta_1(W_h + d_h) \le v_1$, $p_h + \eta_2(W_h + d_h) \le v_2$, $W_l \le W_h + d_h$, and $p_h \ge 0$. Also, $u_3 = 0$ under service differentiation and KKT conditions give $u_1 + u_2 = \lambda + u_4$ and $\eta_1 u_1 + \eta_2 u_2 = \lambda\eta_1(1-\gamma_1)q$. Then $u_1 = \frac{\lambda(\eta_1(1-\gamma_1)q-\eta_2)-u_4\eta_2}{\eta_1-\eta_2}$ and $u_2 = \frac{(\lambda\gamma_1 q+u_4)\eta_1}{\eta_1-\eta_2}$. Because $u_4 \ge 0$, $\eta_2 \le \eta_1(1-\gamma_1)q$ so that $u_1 \ge 0$. If $p_h > 0$, then $u_4 = 0$ and the first and second constraints bind. The feasibility constraints then give case (i) in which the optimal net cost is increasing in γ_1 because both $1 - q + \gamma_1 q$ and $W_l(\mu, 1 - q + \gamma_1 q)$ are increasing in γ_1. If $p_h = 0$, the second constraint still binds and we get case (ii). The optimal net cost is again increasing in γ_1 because both $(1-\gamma_1)q$ and $(v_2/\eta_2 - W_l(\mu, 1 - q + \gamma_1 q))$ are non-negative and decreasing in γ_1. Hence the best γ_1 is zero and $\mathcal{P}6$, in which the IC constraint for low-type service does not have to bind, has a better net cost. $\blacksquare$

Problem $\mathcal{P}3$: $\gamma_1 = 0$ and $0 < \gamma_2 < 1$

All the impatient customers select low-type service while the patient customers split between high and low-type services. Because $\gamma_1 = 0$ and $0 < \gamma_2 < 1$, the IC constraints become $p_h - p_l = \eta_2 (W_l + d_l - W_h - d_h) \ge \eta_1 (W_l + d_l - W_h - d_h)$. So $W_h + d_h \ge W_l + d_l$ and the high-type service is the cheaper service class. Problem $\mathcal{P}2$ for any γ_2 can be written as

$\min_{p_l,d_h,d_l} \lambda r - \lambda p_l + \lambda\eta_2\gamma_2(1-q)\left(W_h + d_h - W_l - d_l\right)$ s.t. $p_l + \eta_i(W_l + d_l) \leq v_i$, $i = 1, 2$; $W_l + d_l \leq W_h + d_h$; $p_l \geq \eta_2\left(W_h + d_h - W_l - d_l\right)$ ($\because p_h \geq 0$); $d_h, d_l \geq 0$. Lemma B3 shows that when all the impatient customers select the costlier low-type service, offering a cheaper high-type service that splits the patient customers is sub-optimal. Two factors together explain the reason for this sub-optimality: (i) $p_h < p_l$, i.e., the high-type service class is priced lower and (ii) prioritizing some patient customers while the impatient customers purchase the low-type service misaligns the service classes with the customer types.

Lemma B3 *For problem $\mathcal{P}3$, there is no service differentiation under optimality.*

Proof Suppose $W_h + d_h > W_l + d_l$; decreasing d_h by a small $\epsilon > 0$ maintains the solution feasibility and reduces the net cost. Hence $W_h + d_h = W_l + d_l$ under optimality. ∎

Problem $\mathcal{P}4$: $\gamma_1 = 1$ and $0 < \gamma_2 < 1$

All the impatient customers select high-type service while the patient customers split between high and low-type services. Because $\gamma_1 = 1$ and $0 < \gamma_2 < 1$, the IC constraints become $p_h - p_l = \eta_2\left(W_l + d_l - W_h - d_h\right) \leq \eta_1\left(W_l + d_l - W_h - d_h\right)$. So $W_h + d_h \leq W_l + d_l$ and the high-type service is the costlier service class. Problem $\mathcal{P}4$ for any γ_2 can be written as $\min_{p_h,d_h,d_l} \lambda r - \lambda p_h + \lambda\eta_2(1-\gamma_2)(1-q)\left(W_l + d_l - W_h - d_h\right)$ s.t. $p_h + \eta_i(W_h + d_h) \leq v_i$, $i = 1, 2$; $W_h + d_h \leq W_l + d_l$; $p_h \geq \eta_2\left(W_l + d_l - W_h - d_h\right)$ ($\because p_l \geq 0$); $d_h, d_l \geq 0$. Lemma B4 shows that although increasing γ_2 means more customers select the costlier high-type service, the benefit from that is overshadowed by the negative effect from reduction in both p_h and p_l.

Lemma B4 *For problem $\mathcal{P}4$, $d_h = d_l = 0$ under optimality. Service is always differentiated and other optimal values are given as follows:*
(i) if $\Delta \geq \nu$, $W_h \geq (v_1 - v_2)/(\eta_1 - \eta_2)$ and $(\eta_1 - \eta_2)W_h + \eta_2 W_l \leq v_1$ then $p_h = v_1 - \eta_1 W_h$, $p_l = v_1 - (\eta_1 - \eta_2)W_h - \eta_2 W_l$, and the net cost is $\lambda r - \lambda v_1 + \lambda\eta_2 W + \lambda(\eta_1 - \eta_2)W_h$; and
(ii) if $\Delta \geq \nu$, $W_h < (v_1 - v_2)/(\eta_1 - \eta_2)$ and $W_l \leq v_2/\eta_2$; or if $\Delta < \nu$ and $W_l \leq v_2/\eta_2$ then $p_h = v_2 - \eta_2 W_h$, $p_l = v_2 - \eta_2 W_l$, and the net cost is $\lambda r - \lambda v_2 + \lambda\eta_2 W$.
Further, the net cost is increasing in γ_2 and hence $\mathcal{P}4$ is dominated by $\mathcal{P}5$.

Proof First, $d_l = 0$ under optimality because otherwise it can be decreased to reduce the net cost. Also, $d_h = 0$ because otherwise reducing it by $\epsilon > 0$ and increasing p_h by $\eta_2\epsilon$ maintains the feasibility and reduces the net cost by $\lambda\eta_2(1-\gamma_2)(1-q)\epsilon$. So $\mathcal{P}4$ becomes $\min_{p_h} \lambda r - \lambda p_h + \lambda\eta_2(1-\gamma_2)(1-q)\left(W_l + d_l - W_h - d_h\right)$ s.t. $p_h + \eta_1 W_h \leq v_1$, $p_h + \eta_2 W_h \leq v_2$, and $p_h \geq \eta_2\left(W_l - W_h\right)$. Either the first or second constraint binds because otherwise p_h can be increased. If $\Delta \geq \nu$ then the first condition binds if $W_h \geq (v_1 - v_2)/(\eta_1 - \eta_2)$ and the second one binds otherwise; however, if $\Delta < \nu$ then the second constraint always binds. Applying the feasibility of the third constraint then gives the two cases. In case (i), the net cost is increasing in γ_2 because $W_h\left(\mu, q + \gamma_2(1-q)\right)$ is increasing in γ_2, and in case (ii) it is independent of γ_2. Also, the optimal net cost is continuous in γ_2, and both $(\eta_1 - \eta_2)W_h + \eta_2 W_l$ and W_l are increasing in γ_2. Hence the best γ_2 is zero and so $\mathcal{P}4$ is dominated by $\mathcal{P}5$. ∎

Problem $\mathcal{P}5$: $\gamma_1 = 1$ and $\gamma_2 = 0$

All the impatient customers select high-type service while all the patient customers select low-type service. Note that there is no split. Because $\gamma_1 = 1$ and $\gamma_2 = 0$, the IC constraints become $\eta_2 (W_l + d_l - W_h - d_h) \le p_h - p_l \le \eta_1 (W_l + d_l - W_h - d_h)$. So $W_h + d_h \le W_l + d_l$ and the high-type service is the costlier service class. Problem $\mathcal{P}5$ can be written as $\min_{p_h,p_l,d_h,d_l} \lambda r - \lambda q p_h - \lambda(1-q)p_l$ s.t. $p_h + \eta_1(W_h + d_h) \le v_1$; $p_l + \eta_2(W_l + d_l) \le v_2$; $\eta_2 (W_l + d_l - W_h - d_h) \le p_h - p_l \le \eta_1 (W_l + d_l - W_h - d_h)$; $W_h + d_h \le W_l + d_l$; $p_l, d_h, d_l \ge 0$.

We first discuss when $\mathcal{P}5$ is feasible. It is feasible if and only if $W_l \le v_2/\eta_2$ and $(\eta_1 - \eta_2)W_h + \eta_2 W_l \le v_1$. Clearly, if these inequalities hold then $\mathcal{P}5$ is feasible since $p_h = p_l = d_h = d_l = 0$ satisfies all the necessary constraints. Next, we show that if $\mathcal{P}5$ is feasible then the two conditions should hold. Clearly, $W_l \le v_2/\eta_2$ because otherwise $p_l + \eta_2(W_l + d_l) > v_2$ for any $p_l, d_l \ge 0$ which violates a constraint of $\mathcal{P}5$. Also, since $p_h \ge p_l + \eta_2 (W_l + d_l - W_h - d_h)$ and $p_h + \eta_1(W_h + d_h) \le v_1$, we require that $p_l + (\eta_1 - \eta_2)(W_h + d_h) + \eta_2 (W_l + d_l) \le v_1$ which implies $(\eta_1 - \eta_2)W_h + \eta_2 W_l \le v_1$ since $p_l, d_h, d_l \ge 0$.

Lemma B5 characterizes the optimal solution. Some results here are similar to those in Lemma 2 in which there is no prioritization. If the capacity μ is large enough s.t. $W_l(\mu, q) < \min\left(\frac{v_1 - v_2}{\eta_1 - \eta_2}, \frac{v_2}{\eta_2}\right)$ (cases (iii)-(vi) of Lemma B5) then the expressions for optimal prices and delay are similar to those in Lemma 2 (cases (ii)-(v)) with W_l (W_h) replacing W for those of the low-type (high-type) service. Also, the condition for strategic delay to be optimal, $\Delta > 1/q$, is the same; regardless of whether the retailer prioritizes her customer or not, a strategic delay affects the net cost in the same way. However, if μ is lower (cases (i) and (ii)) then the optimal prices and delay have different expressions from those in Lemma 2.

Lemma B5 *For problem $\mathcal{P}5$, $d_h = 0$ under optimality. Service is always differentiated and other optimal values are given as follows:*
(i) if $\Delta \ge \nu$, $W_h \ge (v_1 - v_2)/(\eta_1 - \eta_2)$ and $(\eta_1 - \eta_2)W_h + \eta_2 W_l \le v_1$ then $p_h = v_1 - \eta_1 W_h$, $p_l = v_1 - (\eta_1 - \eta_2)W_h - \eta_2 W_l$, $d_l = 0$, and the net cost is $\lambda r - \lambda v_1 + \lambda \eta_2 W + \lambda(\eta_1 - \eta_2)W_h$;
(ii) if $\Delta \ge \nu$, $W_h < (v_1 - v_2)/(\eta_1 - \eta_2) \le W_l \le v_2/\eta_2$ then $p_h = v_1 - \eta_1 W_h$, $p_l = v_2 - \eta_2 W_l$, $d_l = 0$, and the net cost is $\lambda r - \lambda q v_1 - \lambda(1-q)v_2 + \lambda \eta_1 q W_h + \lambda \eta_2 (1-q) W_l$;
(iii) if $\Delta \ge \nu$, $W_l < (v_1 - v_2)/(\eta_1 - \eta_2)$ and $\Delta \le 1/q$ then $p_h = v_2 - \eta_1 W_h + (\eta_1 - \eta_2)W_l$, $p_l = v_2 - \eta_2 W_l$, $d_l = 0$, and the net cost is $\lambda r - \lambda v_2 + \lambda \eta_1 W - \lambda(\eta_1 - \eta_2)W_l$;
(iv) if $\Delta \ge \nu$, $W_l < (v_1 - v_2)/(\eta_1 - \eta_2)$ and $\Delta > 1/q$ then $p_h = v_1 - \eta_1 W_h$, $p_l = (\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2)$, $d_l = (v_1 - v_2)/(\eta_1 - \eta_2) - W_l$, and the net cost is $\lambda r - \lambda q v_1 - \lambda(1-q)(\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2) + \lambda \eta_1 q W_h$;
(v) if $\Delta < \nu$, $W_l \le v_2/\eta_2$ and $\Delta \le 1/q$ then $p_h = v_2 - \eta_1 W_h + (\eta_1 - \eta_2)W_l$, $p_l = v_2 - \eta_2 W_l$, $d_l = 0$, and the net cost is $\lambda r - \lambda v_2 + \lambda \eta_1 W - \lambda(\eta_1 - \eta_2)W_l$; and
(vi) if $\Delta < \nu$, $W_l \le v_2/\eta_2$ and $\Delta > 1/q$ then $p_h = \eta_1 (v_2/\eta_2 - W_h)$, $p_l = 0$, $d_l = v_2/\eta_2 - W_l$, and the net cost is $\lambda r + \lambda \eta_1 q (W_h - v_2/\eta_2)$.

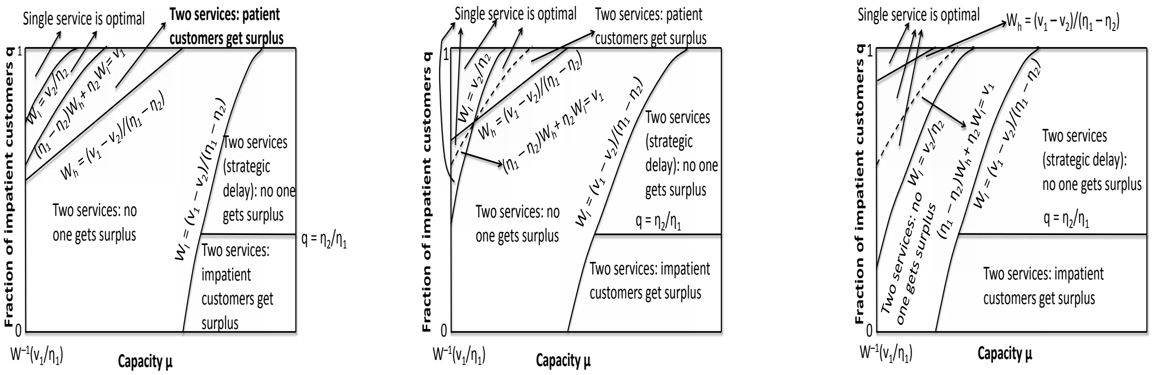
Proof The delay $d_h = 0$ under optimality; otherwise it can be reduced by $\epsilon > 0$ while increasing p_h by $\eta_2 \epsilon$ to reduce the net cost. Then $\mathcal{P}5$ becomes $\min_{p_h,p_l,d_l} \lambda r - \lambda q p_h - \lambda(1-q)p_l$ s.t. $p_h + \eta_1(W_h + d_h) \le v_1$, $p_l + \eta_2(W_l + d_l) \le v_2$, $p_h - p_l \le \eta_1 (W_l + d_l - W_h)$, $p_h - p_l \ge \eta_2 (W_l + d_l - W_h)$, $d_l \ge 0$, and $p_l \ge 0$. KKT conditions give $u_1 + u_3 - u_4 = \lambda q$, $u_2 - u_3 + u_4 = \lambda(1-q) + u_6$, and $(u_2 + u_4)\eta_2 = u_3 \eta_1 + u_5$. Because $W_h(\mu, q) < W_l(\mu, q)$, either $u_3 = 0$ or $u_4 = 0$. There are three possibilities. (1) Suppose $u_3 = 0$ and $u_4 > 0$ so that the fourth constraint binds; then $u_5 > 0$ and hence $d_l = 0$. Further $u_1 > 0$ and so the first constraint binds. So $p_h = v_1 - \eta_1 W_h$ and $p_l = p_h - \eta_2 (W_l - W_h)$. The feasibility conditions then give case (i). (2) Suppose $u_3 > 0$ and $u_4 = 0$ so that the third constraint binds. Then $u_2 > 0$ and so $p_l = v_2 - \eta_2(W_l + d_l)$. Also, $p_h = p_l + \eta_1 (W_l + d_l - W_h)$. (a) If $W_l < (v_1 - v_2)/(\eta_1 - \eta_2)$ and $d_l = 0$ then $p_h + \eta_1 W_h < v_1$ and so $u_1 = 0$. Hence $u_3 = \lambda q$, $u_2 = \lambda + u_6$, and $(\lambda + u_6)\eta_2 = \lambda\eta_1 q + u_5$. Further, $u_6 = 0$ when $W_l < v_2/\eta_2$ and hence $\Delta \leq 1/q$ for u_5 to be non-negative. That gives the results in cases (iii) and (v). (b) If $W_l < (v_1 - v_2)/(\eta_1 - \eta_2)$ and $d_l > 0$ then $u_5 = 0$ and $(\lambda(1-q) + u_3 + u_6)\,\eta_2 = u_3\eta_1$. Because $u_6 \geq 0$ and $u_3 \leq \lambda q$, we need $\lambda(1-q)\eta_2 \leq \lambda q(\eta_1 - \eta_2)$ or $\Delta \geq 1/q$. We get case (iv) or (vi), depending on whether the first constraint binds when $p_l = 0$. If $\Delta < \nu$ then $p_l = 0$ implies that $p_h + \eta_1 W_h = v_2\eta_1/\eta_2 < v_1$ so that $u_1 = 0$ and $u_3 = \lambda q$ which gives case (vi); otherwise the first constraint binds and we get case (iv). (3) Finally, if $u_3 = u_4 = 0$ then $u_1, u_2, u_5 > 0$. So the first, second, and fifth constraints bind. The feasibility conditions then give case (ii). ∎

Problem $\mathcal{P}6$: $\gamma_1 = 0$ and $\gamma_2 = 1$

All the impatient (patient) customers select low-type (high-type) service. Note that there is no split here. Because $\gamma_1 = 0$ and $\gamma_2 = 1$, the IC constraints become $\eta_1 (W_l + d_l - W_h - d_h) \leq p_h - p_l \leq \eta_2 (W_l + d_l - W_h - d_h)$. So $W_h + d_h \geq W_l + d_l$; the high-type service is the cheaper service class and it has a strategic delay. Problem $\mathcal{P}6$ can be written as $\min_{p_h,p_l,d_h,d_l} \lambda r - \lambda q p_h - \lambda(1-q)p_l$ s.t. $p_l + \eta_1(W_l + d_l) \leq v_1$; $p_h + \eta_2(W_h + d_h) \leq v_2$; $\eta_1 (W_l + d_l - W_h - d_h) \leq p_h - p_l \leq \eta_2 (W_l + d_l - W_h - d_h)$; $W_h + d_h \geq W_l + d_l$; $p_h, d_h, d_l \geq 0$. Lemma B6 shows that there are two conditions required for service differentiation to be optimal when $\gamma_1 = 0$ and $\gamma_2 = 1$, and high-type service is cheaper: (i) the capacity should be sufficiently high so that $W_l(\mu, q) < \min\left(\frac{v_1 - v_2}{\eta_1 - \eta_2}, \frac{v_2}{\eta_2}\right)$ and (ii) $\Delta > 1/q$. The first condition results from the fact that making the high-type service cheaper, through strategic delay, necessitates a higher capacity so that the retailer can provide adequate service. The second condition is the same as the condition for strategic delay to be optimal when $\gamma_1 = 1$ and $\gamma_2 = 0$ (see Lemma B5). It does not change because, even though which service class is costlier/cheaper changes, all the impatient customers select the costlier service class while the patient ones choose the cheaper service class and experience a strategic delay in both cases.

Lemma B6 *For problem $\mathcal{P}6$, $d_l = 0$ under optimality. There is service differentiation only under the following two cases:*
(i) If $\Delta \geq \nu$, $W_l < (v_1 - v_2)/(\eta_1 - \eta_2)$ and $\Delta > 1/q$ then $p_h = (\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2)$, $p_l = v_1 - \eta_1 W_l$, $d_h = (v_1 - v_2)/(\eta_1 - \eta_2)$ and the net cost is $\lambda r - \lambda q v_1 - \lambda(1-q)(\eta_1 v_2 - \eta_2 v_1)/(\eta_1 - \eta_2) + \lambda\eta_1 q W_l$; and
(ii) if $\Delta < \nu$, $W_l \leq v_2/\eta_2$ and $\Delta > 1/q$ then $p_h = 0$, $p_l = \eta_1(v_2/\eta_2 - W_l)$, $d_h = v_2/\eta_2 - W_h$ and the net cost is given by $\lambda r + \lambda\eta_1 q(W_l - v_2/\eta_2)$.

Proof First, $d_l = 0$ under optimality; otherwise $d_h, d_l > 0$ which is sub-optimal. Also, $W_h + d_h > W_l + d_l \Rightarrow d_h > 0$. Then $\mathcal{P}6$ becomes $\min_{p_h,p_l,d_h} \lambda r - \lambda q p_l - \lambda(1-q)p_h$ s.t. $p_l + \eta_1 W_l \leq v_1$, $p_h + \eta_2(W_h + d_h) \leq v_2$, $p_h - p_l \geq \eta_1 (W_l - W_h - d_h)$, $p_h - p_l \leq \eta_2 (W_l - W_h - d_h)$, and $p_h \geq 0$. KKT conditions give $u_1 + u_3 - u_4 = \lambda q$, $u_2 - u_3 + u_4 = \lambda(1-q) + u_6$ and $(u_2 + u_4)\eta_2 = u_3\eta_1$.


Single service is optimal
Two services: patient customers get surplus
W_l = v_2/η_2
(η_1 − η_2)W_h + η_2W_l = v_1
W_h = (v_1 − v_2)/(η_1 − η_2)
W_l = (v_1 − v_2)/(η_1 − η_2)
Two services: no one gets surplus
Two services (strategic delay): no one gets surplus
q = η_2/η_1
Two services: impatient customers get surplus
Fraction of impatient customers q
1
0
W^{-1}(v_1/η_1)
Capacity μ
Single service is optimal
Two services: patient customers get surplus
W_l = v_2/η_2
W_h = (v_1 − v_2)/(η_1 − η_2)
(η_1 − η_2)W_h + η_2W_l = v_1
W_l = (v_1 − v_2)/(η_1 − η_2)
Two services: no one gets surplus
Two services (strategic delay): no one gets surplus
q = η_2/η_1
Two services: impatient customers get surplus
Fraction of impatient customers q
1
0
W^{-1}(v_1/η_1)
Capacity μ
Single service is optimal
W_h = (v_1 − v_2)/(η_1 − η_2)
W_l = v_2/η_2
(η_1 − η_2)W_h + η_2W_l = v_1
W_l = (v_1 − v_2)/(η_1 − η_2)
Two services: no one gets surplus
Two services (strategic delay): no one gets surplus
q = η_2/η_1
Two services: impatient customers get surplus
Fraction of impatient customers q
1
0
W^{-1}(v_1/η_1)
Capacity μ

(a) Low λ (b) Medium λ (c) High λ

Figure C1: Variation of optimal service with capacity μ under different cases

So $u_3 > 0$ and the third constraint binds; otherwise $u_2 = u_3 = u_4 = 0$ which would violate the second condition. So $u_4 = 0$ and we get $u_1 = \frac{(\lambda(\eta_1 q - \eta_2) - u_6\eta_2)}{\eta_1 - \eta_2}$, $u_2 = \frac{(\lambda(1-q)+u_6)\eta_1}{\eta_1 - \eta_2}$, and $u_3 = \frac{(\lambda(1-q)+u_6)\eta_2}{\eta_1 - \eta_2}$. Hence $\Delta \geq 1/q$ so that $u_1 \geq 0$ for some $u_6 \geq 0$. Also, $u_2, u_3 > 0$ so $p_h = v_2 - \eta_2(W_h + d_h)$ and $p_l = p_h + \eta_1 (W_h + d_h - W_l)$. Cases (i) and (ii) occur based on whether the first constraint binds if $p_h = 0$. If $\Delta < \nu$ then $p_l + \eta_1 W_l = \eta_1 (W_h + d_h) = \eta_1 v_2/\eta_2 < v_1$ when $p_h = 0$. Then $u_1 = 0$, $u_6 > 0$, and the feasibility conditions give case (ii). If $\Delta \geq \nu$ then the first constraint binds and we get case (i). ■

Appendix C: Evolution of Optimal Service Delivery as μ Increases (when $\Delta \geq \nu$)

In order to understand this evolution, we first characterize the relationship between the curves in the (μ, q) plane given by $W_h(\mu, q) = (v_1 - v_2)/(\eta_1 - \eta_2)$ and $W_l(\mu, q) = v_2/\eta_2$, which is given by Lemma C1.

Lemma C1 *The curves given by $W_h(\mu, q) = (v_1 - v_2)/(\eta_1 - \eta_2)$ and $W_l(\mu, q) = v_2/\eta_2$ intersect at most once. Their relationship is given as follows:*
(i) if $\lambda < \frac{\eta_1}{v_1}\left(\frac{v_2/\eta_2}{(v_1-v_2)/(\eta_1-\eta_2)} - 1\right)$ then the W_h curve is always below the W_l curve and they do not intersect;
(ii) if $\frac{\eta_1}{v_1}\left(\frac{v_2/\eta_2}{(v_1-v_2)/(\eta_1-\eta_2)} - 1\right) \leq \lambda \leq \frac{(\eta_1-\eta_2)}{(v_1-v_2)}\left(\frac{v_2/\eta_2}{(v_1-v_2)/(\eta_1-\eta_2)} - 1\right)$ then there exists a unique (μ, q) such that $W_h(\mu, q) = (v_1 - v_2)/(\eta_1 - \eta_2)$ and $W_l(\mu, q) = v_2/\eta_2$; and
(iii) if $\lambda > \frac{(\eta_1-\eta_2)}{(v_1-v_2)}\left(\frac{v_2/\eta_2}{(v_1-v_2)/(\eta_1-\eta_2)} - 1\right)$ then the W_l curve is always below the W_h curve and they do not intersect.

Based on Lemma C1, we can find the optimal service delivery for different μ's. Figure C1 illustrates three cases corresponding to those in the lemma (note that the optimal service delivery features are obtained by just applying Theorem 1). In all of them, if μ is large enough so that $W_l(\mu, q) < (v_1 - v_2)/(\eta_1 - \eta_2)$, the optimal service delivery is the same; however, for smaller μ's, it varies between them. For instance, although offering two services in which only patient customers get surplus is optimal when λ is lower (Figures C1a and C1b) *that type of service delivery is always sub-optimal* if λ is high (Figure C1c). The intuition behind this result is as follows. If $\mu = \lambda + x$ in which $x > 0$ is a positive constant then

$W_h = 1/\left(\lambda(1-q)+x\right)$ is decreasing in λ while $W_l = (\lambda+x)/\left(x(\lambda(1-q)+x)\right)$ is increasing in λ. For a given additional capacity x, W_h and W_l diverge more as λ increases. For lower λ's, because W_h and W_l are not so different, the retailer has to ensure that patient customers do not select the high-type service by pricing the low-type service cheaper, which provides them surplus. However, for high λ's, W_h and W_l are quite different which automatically precludes them from choosing the high-type service and thereby the retailer extracting all their surplus.

Appendix D: Price of Primary Product

We consider what happens to the retailer's optimization problem when the price of primary product (or service)[16] is also taken into account. We denote the Type 1 and Type 2 customers' *total valuation* (of primary product and ancillary service) by $\tilde{v}_1$ and $\tilde{v}_2$ respectively. With a slight abuse of notation, we also let p_p to be the price of primary product. Then using $v_1 = \tilde{v}_1 - p_p$ and $v_2 = \tilde{v}_2 - p_p$, which are the *residual utilities* of Type 1 and Type 2 customers for the ancillary service[17], we can apply the analysis in §6 to find the optimal solution.

The figures and the accompanying table show the results from an example that demonstrates how p_p affects the optimal net cost (after including the price of primary product), and the optimal prices and service delivery for provision of the ancillary service. We find that the optimal net cost is constant initially in region 1. That is because in this region, the retailer offers prioritized service without any free service or strategic delay, and we also find that both p_h and p_l decrease at the same rate at which p_p increases. However, in region 2, the price of primary product is quite high which makes p_l zero. As p_p increases, p_h now decreases much more rapidly resulting in an increase in the net cost. Finally, in region 3, p_p is so high that $p_l = 0$ and impatient customers are split in order to provide adequate service to everyone. That further results in an increasing net cost. However, interestingly, because of the split, the rate of decrease of price p_h in region 3 is less than that in region 2.

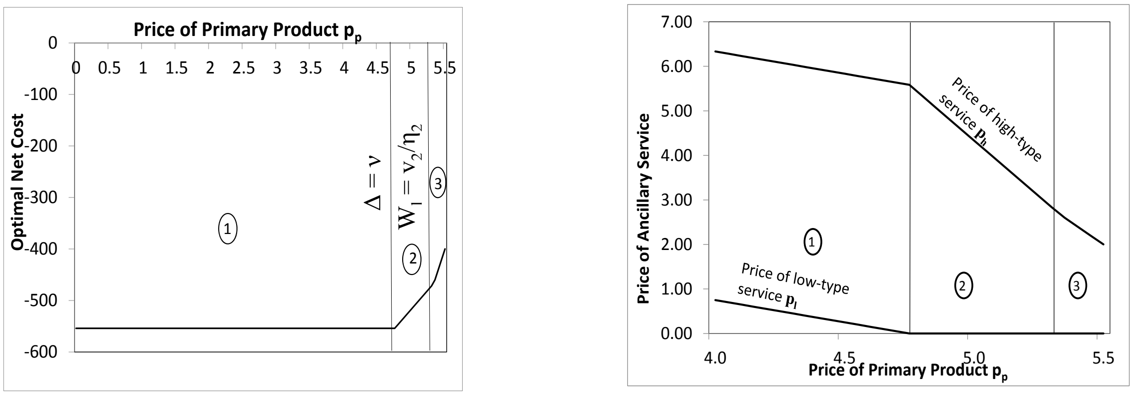
The discussion above considers an exogenous price p_p. Next, we analyze what happens if the retailer can decide how much p_p should be. We assume that there is a minimum price, $\underline{p}_p$, below which the retailer will not charge for the primary product. There can be multiple reasons for the presence of such a minimum price. The retailer might not want to *signal a "price war"* to its competitors by having the price too low, and there may be financial/accounting reasons for having a minimum threshold on how much the primary product should be charged. Also, customer perception is another key factor. The retailer would not want the customers to presume, from a very low price, that the product is inferior or she is *overcharging* for the ancillary service. Without any loss of generality, we normalize the minimum price to zero[18]. Lemma D1 identifies the optimal price for the primary product.

Lemma D1 *The retailer's optimal net cost is increasing in the price of primary product, and so the optimal price is zero.*

[16] For conciseness, we use *primary product* in the rest of the discussion; our analysis also applies when the retailer sells primary service.

[17] For simplicity, customers make decisions just based on the total valuation and total price(s).

[18] The valuations $\tilde{v}_1$ and $\tilde{v}_2$ would become $\tilde{v}_1 - \underline{p}_p$ and $\tilde{v}_2 - \underline{p}_p$ respectively, with other parameters unchanged. The optimal price of the primary product then would be the increment above $\underline{p}_p$.



(a) Variation of optimal net cost with the price of primary product

(b) Variation of prices p_h and p_l with the price of primary product

Figure D1: Optimal net cost and prices vs. p_p when $\tilde{v}_1 = 11$, $\tilde{v}_2 = 6$, $\eta_1 = 50$, $\eta_2 = 10$, $\lambda = 100$, $\mu = 125$, and $r = 2$.

Region	Single service	Free service	Prioritized service	Strategic delay	Split of impatient customers
1			✓		
2		✓	✓	✓	
3		✓	✓		✓

Table D1: Features of optimal service delivery in different regions

Proof We prove the inverse of the statement in the lemma: if the price of the primary product decreases then the optimal net cost also decreases, i.e., it cannot increase. It is observed by noting that, for any feasible solution to $\mathcal{N}$ or $\mathcal{P}$ with p_p as the price of the primary product, a new solution with all the decision variables remaining the same except for p_h and p_l which are both increased by ϵ is also feasible for the corresponding problem when the price of the primary product is $p_p - \epsilon$ $(0 \leq \epsilon \leq p_p)$. And furthermore, both solutions have the same net cost. ▮

The reasoning behind the result in Lemma D1 is as follows. When p_p is increased, an option is to decrease both p_h and p_l by the same amount. That is the case when no free service is offered in which the optimal net cost remains constant with increasing p_p (region 1). However, when free service is offered and p_p increases, only p_h decreases but it does so disproportionately. This decrease is necessary due to customer heterogeneity and the need to satisfy appropriate incentive compatibility (IC) conditions. It also increases the retailer's net cost (see regions 2 and 3).

The effect of primary product price on the pricing and delivery of ancillary service can be summarized as follows. If this price, p_p, is exogenous, i.e., the retailer is a *price taker*, then the valuations $v_1 = \tilde{v}_1 - p_p$ and $v_2 = \tilde{v}_2 - p_p$ can be used (in the analyses in §'s 6 and 7) to find how best to price and deliver the ancillary service. However, if this price is endogenous, the retailer sets the zero (minimum) price and the valuations $\tilde{v}_1$ and $\tilde{v}_2$ (after normalization) are used instead.