Faculty & Research Working Paper

Managing Retention in Service Relationships

Sam AFLAKI
Ioana POPESCU
2011/80/DS
Revised version of 2010/32/DS

Managing Retention in Service Relationships

Sam Aflaki*

Ioana Popescu**

Revised version of 2010/32/DS

.

* HEC Paris, Operations Management and Information Technology Area, 1 rue de la Libération, 78351 Jouy-en-Josas, France. Email : aflaki@hec.fr

* Booz & Company Professor in Strategic Revenue Management, Associate Professor of Decision Sciences at INSEAD, 1 Ayer Rajah avenue, 138676 Singapore.
Email: Ioana.popescu@insead.edu

Managing Retention in Service Relationships

Sam Aflaki
HEC Paris, Operations Management and Information Technology Area, 1 rue de la Liberation, 78351, Jouy en Josas, France. aflaki@hec.fr

Ioana Popescu
INSEAD, Decision Sciences Area, 1 Ayer Rajah Avenue, 138676 Singapore. ioana.popescu@insead.edu

Consider a firm that can actively manage and customize the service offered to customers in a repeat business context. What is the long-term value of such flexibility, and how should firms manage the service relationship over time? We propose a dynamic model of the firm–client relationship that relies on behavioral theories and empirical evidence to model the evolution of service quality expectations and their impact on customer retention and profitability. We find that firms can extract higher long-term value by managing service experiences and expectations over time. Varying service in the long run is not optimal, however. We characterize the optimal dynamic service policy and show that it converges to a steady-state service level. Loss aversion expands the range of constant optimal service policies, suggesting that behavioral asymmetries limit the value of responsive service. Sensitivity results characterize the effect of customer margin, loyalty, and memory on policies and profits.

Key words: service management, service quality, managing relationships, managing expectations

1. Introduction

The growth of the service sector has combined with increased availability of customer-level data to bring about a paradigm shift from managing transactions to managing customer relationships in both business-to-business (B2B) and business-to-consumer (B2C) sectors. As service providers focus on capturing long-term revenue streams from each customer, managing retention has emerged as a primary driver of profitability (Jones and Sasser 1995). Evidence suggests that a customer's assessment of the value of the relationship, as well as subsequent repatronage decisions, are critically influenced by the dynamics of service experiences with the firm (Bolton et al. 2006). The question for firms is then how to manage service experiences over time in order to capture the most value from each customer relationship.

Practitioners have recognized the importance of managing a series of service experiences in repeated interactions. "Every year companies have thousands, even millions of interactions with human beings, also known as customers. Their perceptions of an interaction are influenced by [...] the sequence of painful and pleasurable experiences. Companies care deeply about the quality of those interactions [...] Yet the application of behavioral science to service operations seems spotty at best", a recent McKinsey study reveals (DeVine and Gilson 2010).

We respond to this need by proposing a behavioral dynamic model for managing customized services over time in order to maximize expected long-run discounted profit from a firm's customer base. In each period, the provider decides what service level to offer to each customer. Improving service is costly, but it increases overall perception of service quality and hence the probability that a customer will renew the contract. We rely on empirical evidence and behavioral decision theories (adaptive expectations, prospect theory) to model realistic effects of service experiences on customer's service quality expectations, utility, and renewal decisions.

Our focus is on contractual settings in which the firm can use responsive service to manage retention. Here, the relevant defining feature of a contractual setting is that customer departure is observable, as explained in Fader and Hardie (2009); for example the customer cancels the contract or declines to renew a subscription or membership.[1] Examples include subscription and membership services, product service systems, insurance, banking or utilities contracts in B2C, as well as maintenance, support, or advertising contracts in B2B.

In this context, we define responsive service broadly as the noncontractible level of extra effort that the firm expends on retaining an individual customer, including sales-force effort (Liu et al. 2007; DeVine and Gilson 2010), number of contact hours (Bowman and Narayandas 2004), response time, value-added services, and so forth. For example, DeVine and Gilson (2010) report how an insurance company has implemented a schedule of customized phone calls from nurses to individual patients, in order to improve satisfaction and retention. Survival Chic, a lifestyle membership, manages retention through customized service in the form of weekly information and access to special deals and events. For TV stations or internet publishers selling advertising contracts, a controllable measure of service quality which affects future business is the percentage of target audience (or make-goods) delivered to the client over the year (Araman and Popescu 2010). A longitudinal study of high-technology markets shows that renewal of support service contracts is predicted by the sequence of service experiences as measured by additional resources (e.g. engineer work time) devoted to each encounter (Bolton et al. 2006). We prescribe how this sequence of service experiences should be optimized.

From a modeling perspective, our work fills a gap at the interface of the literatures on service operations and marketing by addressing a dual need to "incorporate findings from psychology and marketing into OM models of service management" (Bitran et al. 2008, p. 80) and to develop dynamic models that "determine fully personalized levels of marketing interventions ... over time,

[1] By contrast, in noncontractual settings (e.g., retail or airline travel) the firm does not know when customers become inactive, so the focus is on customer share as opposed to retention.

in such a way as to maximize CLV [customer lifetime value]" (Rust and Chung 2006, p. 575). Our contribution to these literatures is threefold: (i) we propose a behavioral dynamic model to maximize customer value by managing service at the individual level in contractual relationships while assuming that past service experiences affect retention; (ii) we characterize the structure of the firm's optimal service policy in the long run and also in a transient regime; and (iii) we explain how behavioral characteristics (e.g., customer loyalty and memory) affect the firm's policy and profits, and we identify the aspects of such behavior that should be assessed in this setting.

Our findings suggest that, even though firms can increase customer value by appropriately adjusting service and managing customer expectations, it is not optimal in the long run to vary service. In other words, the optimal service policy converges to an ideal long-run service level from which it is suboptimal to deviate. The optimal long-run service level is higher for customers who focus more on recent experiences, but it is not necessarily so for more loyal or higher margin customers. It is interesting that behavioral asymmetries drive the structure of our results and also limit the benefits of responsive service. Loss aversion—in other words, customers being more sensitive to perceived downgrades than to upgrades in service (Tversky and Kahneman 1991; Bolton et al. 2006)—leads to a range of optimal constant policies, which is confirmed by their practically observed prevalence. In contrast, if consumers are "gain seeking" (i.e., if service levels above expectations are more salient than those below expectations) then the optimal service policy oscillates.

Our results show how firms can improve long-term profitability by managing service over time and also show how behavioral characteristics—such as loyalty, memory, and loss aversion—affect policies and value. With the aim of providing analytical insights for this complex problem, we follow an evolutionary approach to model building. Focusing on retention, we develop a parsimonious model of service dynamics in response to realistic customer behavior in Section 3, characterize its optimal solution in Section 4, and report sensitivity results in Section 5. We then relax our assumptions and enrich the setup as follows. Section 6 allows for more general profit models that include volume effects, customer visits, and/or acquisition. Section 7 discusses general adaptation models, including such behavioral asymmetries as loss aversion. Section 8 incorporates parameter uncertainty, customer heterogeneity, and shared resources. Overall, as summarized in Section 9, these extensions reinforce our main insights and illustrate the versatility of our basic framework. Proofs for all formal results are given in the Appendix.

2. Related Literature

Our work bridges the literature on behavioral and service operations (Loch and Wu 2007; Bitran et al. 2008) and the marketing literature on customer and service relationship management, and customer lifetime value (Venkatesan and Kumar 2004; Rust and Chung 2006).

In the marketing stream, a few papers consider optimal investment decisions in terms of maximizing lifetime value. Ho et al. (2006) model customer purchases as a Poisson process whose rate depends on customer satisfaction, where the probability p that a customer is satisfied in a given period is controlled by the firm. Focusing on static policies, these authors find that customer value is increasing and convex in service quality. Ovchinnikov and Pfeifer (2011) show how to manage acquisition and retention spending between loyal and nonloyal customers when only a limited number can be served. They find that firms may prefer lower CLV customers and may spend more on retention when facing capacity constraints. Our focus is uncapacitated settings.

The papers cited so far do not capture dynamic effects of past firm policies on customer behavior, nor address the "lack of techniques for using individual-level data to create dynamic customer-marketing policies" (Lewis 2005, p. 986). This paper by Lewis is an exception in that it provides numeric results for dynamic relationship pricing without defection. We focus on managing service, but our analysis translates to managing any costly driver of retention—including prices—when defection is observable. For such contractual settings, Fader and Hardie (2009) provide a model of retention and lifetime value with heterogeneous customers, but they do not optimize it; we do so in Section 8.1.

In the service operations literature, Gans (2003) characterizes the firm's optimal stationary service policy in an oligopoly when customers learn about the firm's quality of service in a Bayesian fashion. By contrast, in our model the firm dynamically manages strategic, observable service levels and, in the process, may learn about the customer. In a duopoly setting, Hall and Porteus (2000) consider a finite-horizon dynamic model where demand is a function of "service failures", which are determined by the firm's investment in capacity. The customers in their setting are purely reactive in their switching behavior and have no memory of experiences prior to the current period. Olsen and Parker (2008) study the optimality of base-stock policies when customers who face stockouts defect with some probability but can be reacquired via advertising. Unlike these papers, we capture competition indirectly through customer choice and focus on modeling probabilistic retention as a function of the sequence of service experiences.

In that sense, our work belongs to a growing behavioral literature on dynamic models in which demand evolves adaptively based on the firm's past policies, including pricing (Popescu and Wu 2007), capacity (Liu and van Ryzin 2009), and quality (Caulkins et al. 2006). Liu et al. (2007) study the intertemporal allocation of an exhaustible resource (sales-force effort) over a customer's fixed lifetime when prior service experiences affect short-term profit but not retention. Gaur and Park (2007) derive steady-state results in an oligopoly where market shares evolve based on adaptive customer expectations about retailers' fill rates. In parallel work, Adelman and Mersereau (2010) use approximation techniques to dynamically manage capacity among clients whose stochastic demands depend adaptively on past fill rates. Contrary to our findings, they show that prioritizing clients by margin is optimal and that customer memory does not affect the steady state when demand is deterministic. Retention considerations and loyalty are absent from these models, as customers never defect.

Thus, a combination of features distinguishes our work from the preceding literature. First, we specifically model customer retention (or defection) as an observable and controllable probabilistic construct that is inherent to a contractual relationship (cf. Fader and Hardie 2009). In doing so, we draw on behavioral theories—such as prospect theory and adaptive expectations—to capture realistic effects of previous service experiences on customer behavior and retention in a dynamic service quality framework. Finally, we go beyond steady-state analysis to investigate the dynamics of customized service decisions over time and their sensitivity to behavioral factors such as customer loyalty, loss aversion, and memory.

3. The Model

A profit-maximizing firm decides in each period t what service level $x_t^i \in [0,1]$ to offer each active customer $i \in \{1, \ldots, N\}$.[2] This decision—as well as the customer's history of service experiences with the firm, $X_t^i = (x_0^i, \ldots, x_{t-1}^i)$—affect not only short-term profit from the customer, $\pi_i(x_t^i; X_t^i)$, but also repatronage behavior. Indeed, in each period the customer may decide to terminate the relationship (e.g., by canceling or not renewing the contract). This decision depends on the entire sequence of service experiences, including other customer or contract-specific factors and uncertain events, as described in Section 3.1. Defection is observed by the firm, and we assume that a customer who defects is lost forever; this assumption is typical in contractual settings, in contrast with non-contractual settings where "always a share" models are the norm (Fader and Hardie 2009).

By dynamically managing service levels for each customer, x_t^i, the firm's objective is to maximize the expected long-term value from its customer base. This value is given by

$$\lim_{T \to \infty} \sum_{t=0}^{T} \beta^t \sum_{i=1}^{N} \mathbb{P}(\text{customer } i \text{ is active by time } t | X_t^i) \pi_i(x_t^i; X_t^i), \tag{1}$$

where $\beta \in (0,1]$ is the firm's discount factor. In the absence of pooling effects, Problem (1) is separable at the customer level; hence we can omit the customer index i from notation and focus on managing individual relationships—that is, maximizing the CLV obtained from each customer. This assumption is common in the literature, as detailed in Ovchinnikov and Pfeifer (2011). We relax it in Section 8, where we consider customer heterogeneity in addition to shared costs and resources.

[2] Measuring service in percentage terms allows us to capture a variety of service decisions, including the binary (service, no service) models typically used in the literature; in this case, x represents the fraction of time that (good) service is offered within a given time period.

3.1. Customer Behavior

This section establishes a general, parsimonious model for the probability that a customer quits the relationship at time t as a function of the history of service experiences X_t. Renewal decisions are typically captured by a random utility model (e.g., Lewis 2005; Bolton et al. 2006); thus, at time t, the customer's utility from the relationship with the firm is given by

$$\tilde{U}_t = U_t(X_t; I) + \tilde{\varepsilon}_t. \tag{2}$$

Here U_t is a deterministic value function of the entire history of service experiences including an information vector of parameters, I, that captures individual characteristics (such as loyalty, memory, initial expectations) as well as factors that are contract, firm, or industry specific (e.g., price P, and switching costs). Additional random factors and events that affect the customer's decision are summarized in the random component $\tilde{\varepsilon}_t$, to be assumed i.i.d with a general distribution.

For tractability, we assume that the history X_t of service experiences can be represented by a one-dimensional recursive construct s_t that acts as a sufficient statistic to predict customer repatronage behavior—namely, $\tilde{U}_t = U(s_t; I) + \tilde{\varepsilon}_t$. So the probability that an active customer renews at time t is given by an increasing function $F(s_t) = F(s_t; I) = \mathbb{P}(\tilde{U}_t \geq 0) = \mathbb{P}(\tilde{\varepsilon} \geq -U(s_t; I))$. This model can be estimated from longitudinal data and can be extended to account for customer inertia (Su 2009) and choice among competing offers.[3] We refer to s_t as (the customer's perception of overall) *service quality* as determined recursively through an exponentially smoothed memory process:

$$s_{t+1} = \lambda s_t + (1 - \lambda) x_t. \tag{3}$$

The memory parameter $\lambda \in [0, 1)$ is the weight that the customer puts on prior experiences. Customers with lower λ focus more on more recent experiences; in the limit, if $\lambda = 0$ then customer decisions are determined solely by the most recent service experience.

The exponential smoothing model (3) captures the essence of theoretical models of belief formation (Hogarth and Einhorn 1992), service quality expectations (Cronin and Taylor 1992), and economic goodwill (Nerlove and Arrow 1962). Such models have been tested empirically in a service context (Boulding et al. 1993; Bolton et al. 2006) and used extensively to capture memory effects in operational models (e.g., Caulkins et al. 2006; Gaur and Park 2007; Liu et al. 2007; Popescu and Wu 2007; Adelman and Mersereau 2010). In addition to being widely used and empirically supported, model (3) is arguably the simplest that renders the main insights from our framework. Extensions capture smooth nonlinearities ($s_{t+1} = H(x_t, s_t)$) as well as asymmetries in perception, notably loss aversion, motivated by prospect theory (Tversky and Kahneman 1991); see Section 7.

[3] In this case $F(s_t; I) = \mathbb{P}(\tilde{U}_t \geq \tau + \tilde{V}_t) = \mathbb{P}(\tilde{\varepsilon} - \tilde{\epsilon} \geq V - U(s_t; I))$, where $V_t = V + \tilde{\epsilon}_t$ is the i.i.d. utility from the competitive offer and τ is a parameter reflecting consumer inertia.

In sum, our consumer behavior model makes no assumptions about the renewal probability $F(s)$ beyond that it is increasing. In other words, customers who have had overall better experiences with the firm are less likely to defect, as widely supported in the literature (see Zeithaml 2000). This assumption is consistent with the *adaptive expectations* framework (Cronin and Taylor 1992), under which previous service experiences create endogenous "will" expectations that evidence a positive effect on retention (Boulding et al. 1993).[4] Retention functions F commonly estimated in the literature include logit (Rust et al. 2004), exponential (Berger and Nasr 1998), and double-exponential functions (Bolton et al. 2006).

3.2. The Firm's Dynamic Optimization Problem

Based on the customer behavior model just described, the maximum expected long-term value from an active customer—given her history of service experiences as summarized by the state s_t—can be formulated recursively as the following stochastic dynamic program:

$$J(s_t) = \max_{\{x_t\}} \pi(x_t; s_t) + \sum_{l=1}^{\infty} \beta^l \left(\prod_{k=1}^{l} F(s_{t+k}) \right) \pi(x_{t+l}; s_{t+l}) = \max_{x_t \in [0,1]} \pi(x_t; s_t) + \beta F(s_{t+1}) J(s_{t+1}); \quad (4)$$

here, for all t, $s_{t+1} = \lambda s_t + (1-\lambda) x_t$ by (3). This is a stochastic shortest path problem with state- and decision-dependent transition probability $F(\lambda s_t + (1-\lambda) x_t)$ (Bertsekas 2007). The model assumes perfect screening and information regarding customer characteristics I; in Section 8.1 we extend this to incorporate customer heterogeneity and parameter uncertainty.

The main trade-off in (4) is between the short-term cost of providing high service and the long-term benefit of increasing retention and future revenue streams. To isolate the effect of past service experiences on retention, which is our main focus, for now we ignore volume effects; that is, we suppose $\pi(x_t; s_t) \equiv \pi(x_t)$. This assumption is consistent with need-based services (e.g. insurance, utilities) and all-inclusive subscriptions (e.g. lifestyle memberships, broadband access). In Section 6 we show that our insights remain valid even after we account for the positive effect of past service experiences on current customer spending as well as on visit frequency, timing of renewal decisions, and customer acquisition. In short, the following model provides the basic framework for our results:

$$J(s) = \max_{x \in [0,1]} \pi(x) + \beta F(\lambda s + (1-\lambda) x) J(\lambda s + (1-\lambda) x). \quad (5)$$

[4] These authors distinguish between *will* and *should* expectations, and their respective effects on retention; the former are based on the customer's previous experiences with the firm and typically dominate the latter, which are based on such exogenous factors as the firm's advertising, industry standards, and competitive offers. Our focus is on service quality expectations that are based on prior experiences with the firm (as summarized by s_t) and hence on *will* expectations.

As a static benchmark, consider a firm that ignores the effect of service quality expectations on retention and offers a constant service over time, $x_t = s_t \equiv s$. In this case the long-term profit is given by the well-known CLV formula (Rust et al. 2004):

$$\Pi(s) = \sum_{t=0}^{\infty} \beta^t F^t(s)\pi(s) = \frac{\pi(s)}{1-\beta F(s)} = \pi(s)L(s). \tag{6}$$

Given a constant service quality level s, the *expected customer lifetime value* $\Pi(s)$ is the product of the short-term profit $\pi(s)$ of each active customer and her *expected lifetime* $L(s) = 1/(1-\beta F(s))$.

We assume that the short-term profit π is strictly concave; its unique maximizer $\underline{s} \in [0,1]$ is the service level offered by a myopic firm (i.e. one that ignores future revenues). We further assume that the static objective Π has a unique interior maximizer $\bar{s} \in [0,1]$; in other words, we assume that Π is strictly quasi-concave.[5] Diminishing marginal returns to static service quality, and a non-monotone service–profitability relationship are supported empirically (Rust et al. 1995).

Our first formal result illustrates the advantage—relative to static and myopic benchmarks—of dynamically managing service in our framework. Let $x^*(s)$ denote the (largest) service that maximizes (5).

Lemma 1. *(a) The value function $J(\cdot)$ is increasing, and $\Pi(s) \le J(s) \le \dfrac{\pi(\underline{s})}{1-\beta}$ for all $s \in [0,1]$. (b) The optimal service level always exceeds the myopic one, $x^*(s) \ge \underline{s}$, for all s.*

Lemma 1 follows directly from the service quality–retention relationship F being increasing. This result shows that firm's dynamic management of service levels in response to customer expectations will result in higher than myopic service, and yield higher profits, than a static policy. Moreover, the firm can extract, on average, more value from customers who have had better overall experiences with the firm.

This setup raises several questions. Can the firm do better than offer $x_t = \bar{s}$ every period and, if so, how? Is it optimal to vary service on the long run, or is there an ideal service level for which the firm should aim? What about the short run? How do behavioral characteristics, such as memory and loyalty, affect the firm's service policy and customer lifetime value? We provide formal answers to these questions in what follows.

4. Optimal Policy

This section describes the optimal policy of the firm. We first characterize an ideal long-run service level from which it is suboptimal to deviate. We then investigate the firm's optimal transient policy—that is, the optimal sequence of service experiences and its convergence properties.

[5] To guarantee uniqueness of optimal solutions, it is sufficient to assume that $\pi'(s)/L'(s)$ is strictly decreasing in s. This condition is satisfied for common parametric models used in the literature, such as logistic $F(x) = 1/(1+\exp(-\alpha x))$ or exponential $F(x) = 1-\exp(-\alpha x)$ retention, and for power cost $c(x) = x^{1+\theta}$ with $\theta \ge \alpha$.

4.1. Steady State

By definition, s^{**} is a *steady state* if the firm has no incentive to move away from it (i.e., it is a fixed point of the optimal service policy: $x^*(s^{**}) = s^{**}$). In other words, if customers are accustomed to a service level s^{**}, then it is optimal for the firm to offer $x \equiv s^{**}$ in each period.

Our next result establishes necessary conditions for a steady state to exist. Existence and global stability are subsequently confirmed in Proposition 2. In order to ensure uniqueness, we assume throughout that the following function is strictly quasi-concave:

$$W(s) = \lambda\pi(s) + (1-\lambda)\Pi(s). \tag{7}$$

PROPOSITION 1. *If Problem* (5) *admits a steady state* s^{**}*, then this is the unique interior maximizer of* $W(s)$ *over* $s \in [0,1]$*. Furthermore, in this case* $\underline{s} \le s^{**} \le \bar{s}$ *and* s^{**} *is increasing in* β*.*

The result shows that, in steady state, the firm balances short-term profit π and long-term customer value Π weighed by customer memory λ, by solving $W'(s) = \lambda\pi'(s) + (1-\lambda)\Pi'(s) = 0$. In particular, if customer renewals are insensitive to past service experiences (i.e., if F is constant), this reduces to $\pi' = 0$ and so a myopic policy $s^{**} = \underline{s}$ is optimal.

Proposition 1 is useful for relating the firm's long-run service policy to its strategic outlook and to customer- and contract-specific factors (further discussed in Section 5). A firm with a short-term outlook puts less weight β on future cash flows (e.g., $\beta = 0$ for a fully myopic firm), and it provides less long-run service and lower service quality, because it focuses on (short-term) cost savings. In particular, the ideal long-run service level exceeds the myopic benchmark: $s^{**} \ge \underline{s}$.

On the other hand, $s^{**} \le \bar{s}$; that is, the steady-state service level is lower than the level offered by a strategic firm that is oblivious to customer expectations and their impact on retention. This result is robust, and preserved under volume effects and nonlinear adaptation models, as shown in Section 6. Indeed, by appropriately managing and responding to customers' endogenous expectations, a strategic firm achieves systematic cost savings in the long run.

4.2. Transient Policy and Global Stability

The previous section characterized a steady-state service policy but did not indicate if and how the firm might reach it. The following result establishes existence and global stability of the unique interior steady state determined by Proposition 1, and it also characterizes the structure of the transient policy. Define the optimal service quality (or state) policy as $s^*(s) = \lambda s + (1-\lambda)x^*(s)$.

PROPOSITION 2. *The optimal service quality policy* $s^*(\cdot)$ *is increasing. Moreover, all optimal paths* $\{s_t^*\}$ *converge monotonically to the unique steady state* s^{**} *characterized by Proposition 1. The optimal service path* $\{x_t^*\}$ *also converges to the steady state* s^{**}*.*

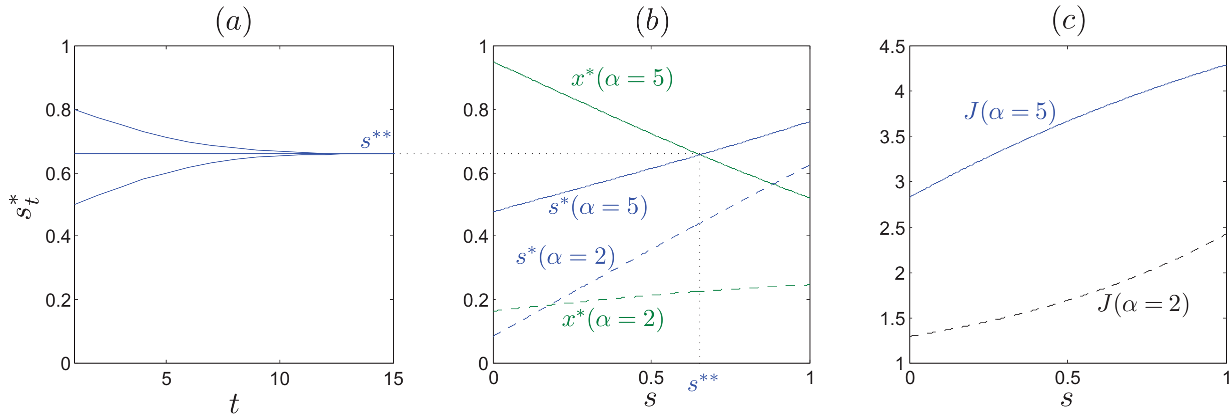

Figure 1 **Optimal service quality paths (a), policies (b) and customer value (c).** $\pi(x)=1-x^2$, $F(s)=1/(1+\exp(-\alpha s))$, $\lambda=0.5$, $\beta=0.94$.

Proposition 2 demonstrates the existence of an ideal long-run service level s^{**} that does not depend on customers' prior expectations s_0. In general, the firm benefits from adjusting service levels, and the optimal way of doing so induces a monotone service quality path that converges to s^{**}; this is illustrated in Figure 1(a).

If customers have relatively low initial perceptions of service quality (based on prior experiences with firm), then our model prescribes gradually increasing service quality to s^{**}; the opposite holds if customers have high initial perceptions. To illustrate these situations we consider, for example, a firm that has been ignoring future revenue streams and thus offering (in each period) the myopic service level $\underline{s}$. When customer perceptions are anchored at $s_0=\underline{s}\le s^{**}$, this firm can maximize customer lifetime value by gradually improving service quality perceptions up to s^{**}. In contrast, the opposite prescription applies for a firm that has been ignoring customer adaptation processes, and thus offering the constant service level $s_t=\bar{s}\ge s^{**}$.

Our results in this section indicate that, in a transient regime, the firm can benefit from managing customers' service quality perceptions/expectations. Initially, these may not be perfectly known to the firm. For a customer who has a history with the firm, s_0 summarizes the service quality delivered so far, and can be estimated from longitudinal data (as in, e.g. Bolton et al. 2006); for new customers, s_0 represents an initial stock of goodwill, as from word of mouth or other 'will' expectations. The assumption that s_0 is perfectly known is relaxed in Section 8.1.

Proposition 2 shows that all service paths x_t^* converge to the same steady state, but it makes no statement regarding monotonicity of the service policy. Indeed, it is possible that the optimal service policy is not increasing, as illustrated in Figure 1(b) and explained in the next section.

4.3. Marginal Returns to Service Quality and Service Policy

Here we investigate both the structure of the optimal service policy and the nature of marginal returns to service quality as captured by the shape of the value function J. We argue that both depend on the shape of the retention function F (see Section 3.1); this can be convex if customers have low switching costs, as evidenced in more competitive industries (Jones and Sasser 1995).

PROPOSITION 3. *If F is convex, then the optimal service policy $x^*(s)$ is increasing and the value function $J(s)$ is convex. These relationships are ambiguous if F is concave; see Figure 1(b,c).*

A convex value function J goes against the conventional wisdom of diminishing marginal returns to service quality (Rust et al. 1995). Ho et al. (2006) find that customer value can be convex if costs are not too steep; our result is not driven by cost but rather by the shape of the retention function, which does not figure in their paper. Jones and Sasser (1995), suggest that the effect should be determined by whether the retention rate is convex or concave. Although it confirms their intuition for F convex, Proposition 3 implies that firms may experience increasing marginal returns to service quality even when the latter has a diminishing marginal effect on retention (i.e., when F is concave). Indeed, panel (c) of Figure 1 shows that J can be either convex or concave for F concave. Intuition suggests that the shape of the value function J depends on the degree of concavity of F; in particular, J is concave for sufficiently concave F. This intuition is confirmed in numerical experiments for parametric logit, exponential, and power specifications of F. However, characterizing the relationship analytically is complicated by the multiplicative interaction effects in the value-to-go (5), which make preservation of concavity difficult.[6]

5. Sensitivity to Price and Behavioral Characteristics

As stated in the Introduction, part of our goal is to understand how long-run service and customer lifetime value are affected by behavioral characteristics (such as customer loyalty and memory) and by prices and switching costs.

5.1. Customer Memory

The firm provides higher long-run service to customers who are more forgetful—that is, those who anchor on more recent experiences, as captured by a lower λ (this is a consequence of Proposition 1). In other words, customers who adapt faster should receive better service in the long run, ceteris paribus, and so are more costly to maintain. But are these customers more profitable? The next result states that the value function is locally monotonic in the adaptation parameter λ, but the direction of monotonicity depends on customer initial expectations.

PROPOSITION 4. *(a) The steady-state service level $s^{**}(\lambda)$ is decreasing in λ. (b) For any given λ, the marginal effect of memory on the value function $\frac{d}{d\lambda}J(s;\lambda)$ is positive for $s > s^{**}(\lambda)$ and negative for $s < s^{**}(\lambda)$.*

[6] A sufficient condition for $J(s)$ to be concave (resp. convex) is the concavity (convexity) of $(FJ)(s)$; see the proof of Proposition 3. Yet this property is not preserved in general because FJ need not be concave even when both F and J are (increasing and) concave.

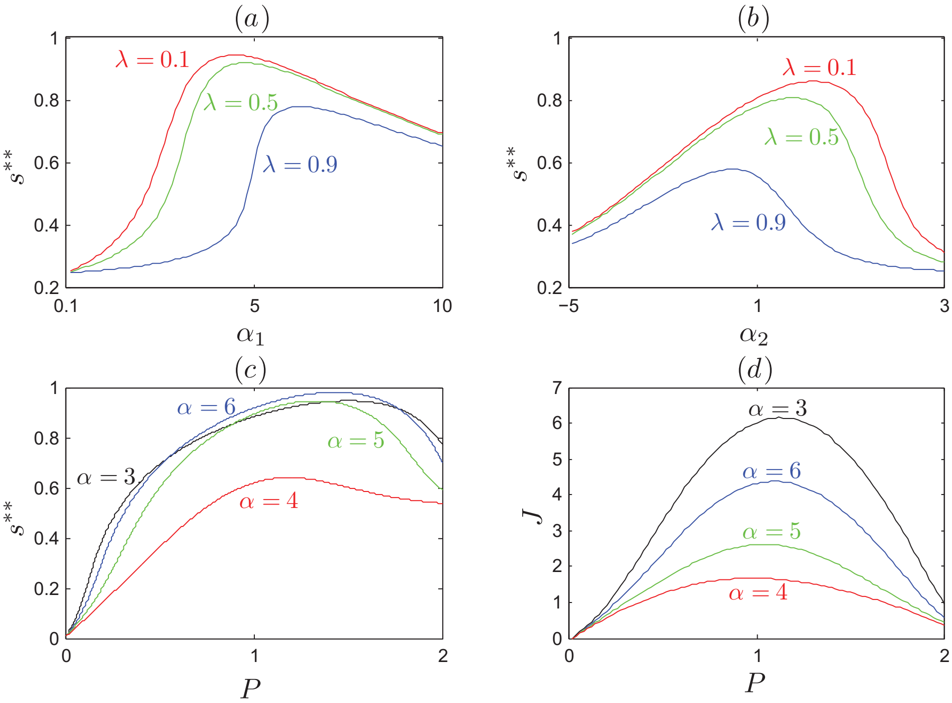

Figure 2 **Sensitivity to loyalty (α) and price (P);** $\pi(x,P) = Pd(x,P) - c(x)$, $c(x) = x^2, d(x,P) = 0.5x - P + 2$, $F(s) = 1/(1 + exp(-\alpha_1 s + \alpha_2 + P))$, $\beta = 0.94$. **(a)** $\alpha_2 = 1$; **(b)**$\alpha_1 = 3$; **(c, d)** $\alpha_2 = 1$, $\lambda = 0.5$.

Among customers who had relatively good prior experiences with the firm, those who adapt more slowly provide a better long-term return on investment because they yield higher value J and demand lower long-run service s^{**}. Among dissatisfied customers, those with shorter memory are more profitable but do require higher maintenance cost in the long run.

5.2. Loyalty and Switching Costs

The steady-state service level is affected by factors that increase retention, such as customer inertia, loyalty, and switching costs. We capture these factors using a generic parameter α such that $F(s;\alpha)$ is increasing. This implies $J(s;\alpha)$ is increasing in α, confirming the intuitive result that, all else equal, loyal customers are more valuable. But do they also receive better service?

As illustrated in Figure 2, the steady state $s^{**}(\alpha)$ is typically nonmonotonic: it increases up to a certain point but then decreases. This suggests that there is an "ideal" loyalty level beyond which the firm will treat customers as a captured audience and therefore reduce its costly investment in retention. In the additive specification, α can be viewed a proxy for customer inertia, or switching costs, whereas in the multiplicative case it reflects customer's sensitivity to prior service experiences.

By moderating F's degree of concavity, α affects marginal returns to service quality and the nature of the transient service policy (cf. panels (b,c) in Figure 1). This parameter also links to the level of competition in the market environment (Jones and Sasser 1995), suggesting that—from the customers' point of view—there is an optimal level of market competition that results in the highest service level. This is partially consistent with Hall and Porteus (2000), who find that loyalty decreases long-run service levels in an oligopoly. Our model captures competitive effects implicitly, through customer choice; strategic interaction remains an interesting topic for future research.

5.3. Price

There are multiple arguments for managing service levels, as opposed to prices, in long-term relationships; see for example Bolton and Drew (1991), Liu et al. (2007), and Ovchinnikov and Pfeifer (2011). We assume for simplicity that prices are exogenously fixed throughout the customer's lifetime (e.g. based on strategic, competitive, or brand image considerations); our framework can be adapted to manage discounts rather than service levels.[7] In this context, we investigate the effect of price P on long-term service and customer profitability. Higher prices may increase margin, but they also diminish customer surplus, and thereby reduce the probability of retention; in other words, $F(s;P)$ is decreasing in P (because U is so, cf. (2)). Proposition 1 shows that the effect of price on long run service $s^{**}(P)$ is determined by their interaction in $W(s;P) = \lambda\pi(s;P) + (1-\lambda)\Pi(s;P)$.

In general, we find that the unilateral effect of margin on long-run service is nonmonotonic, if short-term profit π is unimodal, in particular increasing in P as with subscription services (i.e., $\pi(x) = P - c(x)$). The unimodal relationship illustrated in Figure 2(c) suggests that higher-margin customers command better service—up to a certain point, where the pattern is reversed and the relationship becomes more transactional. Prices that are too high make it easier for customers to defect, and better service cannot indefinitely make up in retention what is lost by extracting such high rents. This trade-off is reflected in the unimodal relationship between price and customer lifetime value, illustrated in Figure 2(d).

Figures 1 and 2 illustrate the trade-offs among customer characteristics (margin, adaptation, and loyalty), and how they affect value and long-run service. The insights, which are summarized in Section 9 (see Table 2), remain robust when customers are heterogeneous (see Section 8) as well as under more general profit and adaptation models discussed next.

6. General Profit Models

This section shows that our insights remain valid when we account for more general profit models, including volume effects. Unlike need-based services (e.g., insurance, utilities), certain hedonic and utilitarian services are such that past service experiences can have a positive impact on customers' volume and/or frequency of purchase. For example, a more satisfied client may purchase additional

[7] In this case x_t captures the discount from the reference price, and customers form adaptive price expectations that affect demand and, in our case, renewal; this is unlike the treatment in Lewis (2005) and Popescu and Wu (2007).

insurance services or use a certain loan service more often. Bolton et al. (2006) show that prior service experiences affect the purchase of additional support contracts. So far, our stylized model has focused on customer retention, and ignored the positive effect of past service experiences on customer spending, purchase frequency, and acquisition. This section demonstrates that incorporating these effects does not change the nature of our insights. For simplicity we discuss each extension in isolation, but combining them will not change the results. We conclude that, despite its simplicity, the stylized model (5) captures essential features of the problem at hand.

6.1. Volume Effects

Suppose that better past experiences increase purchase volume, so instant profit $\pi(x,s)$ is an increasing function of s. For example, an advertiser may spend a larger share of its budget with a publisher who delivers more audience. Redefine $W(s) = \lambda\pi(s,s) + (1-\lambda)\Pi(s)$, where $\Pi(s) = \pi(s,s)L(s)$, and denote partial derivatives with corresponding subscripts, in particular $\pi_2(s,s) = \frac{\partial}{\partial s}\pi(x,s)|x = s$. Then we have the following result.

PROPOSITION 5. *The statements in Lemma 1 and Propositions 1–4 extend under volume effects provided that short-term profit $\pi(x,s)$ is supermodular, concave in x, and increasing and convex in s and that $\pi(s,s)$ is concave in s. In particular, the steady state $s^{**} \in [\underline{s}, \bar{s}]$ solves $W'(s) = (1-\lambda)\pi_2(s,s)$.*

Our main insights are robust to volume effects. The steady-state service level is lower than its static counterpart, and this remains true even if defection decisions are affected only by the most recent service encounter ($\lambda = 0$) because $s^{**}(0) \leq \bar{s}$. In the absence of volume effects ($\pi_2 \equiv 0$) we recover Proposition 1, and in particular $s^{**}(\lambda = 0) = \bar{s}$. With volume effects, $\pi_2(s,s)$ is positive; hence, ceteris paribus, the firm needs to invest less effort over the long run in customers who can be easily enticed to spend more (e.g., through cross- or up-selling). These customers are also more profitable, providing an overall higher return on investment. This finding underscores the benefit of value-added services that stimulate customer spending and reduce retention costs.[8]

Existing models of adaptive demand have focused on the effect of past policies on volume but not on the probability of defection (see Section 2). Without defection (i.e., with F constant), Proposition 5 shows that customer value J is convex and the optimal service policy x^* is increasing; this extends Proposition 3. In contrast with Adelman and Mersereau (2010), who find no memory effect under deterministic demand, the effect persists in our model ($s^{**}(\lambda)$ decreasing in λ).

We next provide intuition for the assumptions underlying Proposition 5. Increasing marginal effects of service quality on demand (i.e., assuming that $\pi(x,s)$ is convex in s) is consistent, for

[8] We have not even included the increase in customer switching costs (and hence retention) that results from capturing a larger share of the customer wallet.

example, with Gans (2003) and the concept of consumer "delight" (Bowman and Narayandas 2004). Supermodularity means that customers who had better experiences with the firm are more sensitive to changes in service, or that the marginal cost of serving them is lower; this assumption, together with the concavity of π in x, ensures global stability. The last condition, which is used to show that s^{**} is unique and decreasing in λ, requires that service have diminishing marginal effects on immediate profit for a firm that keeps service quality constant. For example, these conditions hold when $\pi(x,s) = r(s) - c(x)$ if both customer spending r and costs c are increasing and convex in their respective arguments and if the marginal cost of service grows faster than the marginal effects of past service quality on demand.

6.2. Customer Visits

Our stylized model (5) assumes that customers periodically visit the firm and decide whether or not to renew. However, our main insights are not changed by uncertainty in visits and renewal timing decisions nor by the positive effects of service quality on the frequency of these events.

6.2.1. Endogenous visit frequency. Suppose that, in each period, a customer who has received better quality of service s has a higher probability $v(s)$ of visiting the firm. We then obtain $\bar{J}_v(s) = v(s)\{\max_x \pi(x,s) + \beta F(\lambda s + (1-\lambda)x)\bar{J}_v(\lambda s + (1-\lambda)x)\} + \beta(1-v(s))F(s)\bar{J}_v(s)$. Denoting $F_v(s) = \frac{v(s)}{1-\beta F(s)(1-v(s))}$, we obtain that $J_v(s) = \bar{J}_v(s)/F_v(s)$ solves model (5) with F replaced by $F \cdot F_v$, which is increasing in s. All our results so far extend in this case.

6.2.2. Viscous demand. In contractual settings such as insurance, utilities, and subscriptions, the customer visits the firm regularly but only occasionally—that is, with probability $p(s)$—considers whether or not to defect. A boundedly rational attention budget can explain such demand viscosity: "the consumer rethinks such decisions from time to time, regularly or at some random intervals, perhaps triggered by some events" (Radner 2003, p. 190). The resulting model $J_p(s) = \max_x \pi(x,s) + \beta[p(S)F(S)J_p(S) + (1-p(S))J_p(S)]$, where $S = \lambda s + (1-\lambda)x$, has the same structure as model (5) but with F replaced by the increasing function $1 - p(s)(1-F(s)) \geq F(s)$.

6.3. Customer Acquisition

Although the main focus in this paper is on maximizing profit from an existing customer base (i.e. models for customer base analysis, e.g. Fader and Hardie 2009), our framework can be adjusted to incorporate the positive effects of service experiences on customer acquisition through referrals. Suppose that, in each period t, an existing customer may bring in a new customer with the same service quality expectations s_t; this reflects "word of mouth", which occurs with probability $F_a(s_t)$. In this viral marketing type of acquisition model, growth is proportional to the customer base and so the problem remains separable at the customer level. All our results extend by replacing F with $F + F_a$ if this sum does not exceed $1/\beta$ (in order to ensure a bounded value function; see the Appendix).

7. General and Asymmetric Adaptation Models

This section extends our model, based on exponential smoothing (3), to allow for smooth non-linearities, as well as non-smooth behavioral asymmetries in memory and adaptation processes.

7.1. Nonlinear Adaptation Models

In order to account for smooth nonlinear effects in the customers' perception and adaptation process, we set $s_{t+1} = H(x_t, s_t)$. Our results extend under this general adaptation process by replacing λ with $\lambda(s) = H_2(s, s)$, which is the derivative of $H(x, s)$ with respect to s evaluated at $x = s$; for exponential smoothing (3), we recover precisely $\lambda(s) = \lambda$.

Proposition 6. *Suppose that the adaptation process satisfies the following assumptions: (a) $H(x, s)$ is increasing in x and s, and $H(s, s) = s$; (b) $H_{11}(x, s) \leq 0$; (c) $H_{12}(x, s) \geq 0$; (d) $H_{22}(s, s) \geq 0$. Then the results in Lemma 1 and Propositions 1–4 extend to this model, and the corresponding steady state solves $\lambda(s)\pi'(s) + (1 - \lambda(s))\Pi'(s) = 0$.*

It is interesting that, when service is aligned with expectations, the steady state depends on the adaptation process only via $\lambda(s)$, the marginal effect of previous experiences on service quality.We next discuss the assumptions in Proposition 6. Part (a) states that, all else equal, both current and previous service levels have a positive effect on service quality; this is consistent with empirical evidence (Boulding et al. 1993). Diminishing marginal sensitivity to service, part (b), is a natural assumption. Parts (c) and (d) are technical conditions stating that perception of service quality is more sensitive to a change in (current or past) service for customers who had better past experiences with the firm. To our knowledge, these latter two hypotheses have not been tested.

The assumptions in Proposition 6 are satisfied, for example, by exponential smoothing models in which the weight attached to the current experience depends on the current service level: $H(x, s) = \lambda(x)s + (1 - \lambda(x))x$, provided that $\lambda(x)$ is increasing with bounded curvature, $\lambda'(x) \geq |\lambda''(x)|/2$ (as with, e.g., $\lambda(x) = x$, e^{x-1}, and $1 - e^{-x}$). In this model, lower service is more salient in memory—that is, the lesser the current experience, the more it weighs on service quality. Section 7.2 describes a model in which experiences below expectations are more salient than those above expectations.

7.2. Loss Aversion

In this section we investigate, along lines that are consistent with prospect theory and empirical evidence, the effect of behavioral asymmetries on adaptation and decision processes. We find that these asymmetries—in particular, loss aversion—have important effects on the firm's policy: (i) they make constant service policies more prevalent, leading to a range of steady states; and (ii) optimal service policies oscillate if the asymmetries are reversed.[9]

[9] These insights are preserved in an alternative retention model, driven by disconfirmation–i.e., the gap between experience and expectation, $F(x - s)$, where customers react more to *changes* in service quality than to absolute levels.

Prospect theory postulates that decision makers code new information as gains or losses relative to a status quo, a principle that applies to both decision and experience utility (Kahneman et al. 1997). "Experienced utility" is the decision maker's hedonic value at the moment of experience, a value that in our model is captured by the concept of service quality. Moreover, a negative change from the status quo has a larger effect on value than a positive change of the same magnitude. Generally known as "loss aversion", this phenomenon has received vast empirical support in the service quality literature—not only in B2C but also in B2B markets (Bolton et al. 2006).

Consider the following asymmetric (kinked) service quality updating process:

$$s^K_{t+1} = \begin{cases} s_t + (1-\lambda_G)(x_t - s_t) & \text{if} \quad x_t \geq s_t, \\ s_t + (1-\lambda_L)(x_t - s_t) & \text{if} \quad x_t < s_t, \end{cases} \tag{8}$$

where $1-\lambda_L > 1-\lambda_G$ expresses loss aversion. Such a kinked learning model is used by Gaur and Park (2007) for consumers who form expectations about product availability.

The Bellman equation for loss-averse adaptation can be written as

$$J^K(s_t) = \max_{x_t \in [0,1]} \pi(x_t) + \beta F(s^K_{t+1}) J^K(s^K_{t+1}), \tag{9}$$

where s^K_{t+1} follows the transition dynamics (8). Let J^L and J^G denote the value functions of the smooth problems (5) corresponding to λ_L and λ_G, respectively. Loss aversion implies that, by Proposition 1, the corresponding steady states satisfy $s^{**}_L = s^{**}(\lambda_L) > s^{**}(\lambda_G) = s^{**}_G$. We show in the Appendix that $J^K(s) \leq \min\{J^L(s), J^G(s)\}$. This suggests that loss aversion—in other words, the asymmetric effect of disappointing experiences relative to pleasurable ones—has a negative effect on profitability, which is consistent with the findings in Gaur and Park (2007).

Proposition 7. *Assume that customers are loss averse, $\lambda_L < \lambda_G$. Then Problem* (9) *admits a range of steady states $[s^{**}_L, s^{**}_G]$; that is, starting from any $s_0 \in [s^{**}_L, s^{**}_G]$, a constant service path $x^*_t = s^*_t = s_0$ is optimal. For $s_0 > s^{**}_L$, the service quality path $\{s^*_t\}$ decreases to s^{**}_L and $J^K(s_0) = J^L(s_0)$. For $s_0 < s^{**}_G$, $\{s^*_t\}$ increases to s^{**}_G and $J^K(s_0) = J^G(s_0)$. The optimal service path $\{x^*_t\}$ converges to the same steady state as the corresponding service quality path.*

Adding loss aversion to the model does not affect the general structure of the transient policy, but it does enlarge the set of steady states and thereby increases the prevalence of constant service policies (see Figure 3(a)). Intuitively, the firm has less leverage to improve perception when customers anchor more strongly on negative experiences. Technically, this is due to the kink in service quality updating.

For completeness, we briefly consider the case where customers are "gain seeking" (i.e., where $\lambda_G < \lambda_L$). This means that service experiences above expectations (positive disconfirmation) are more salient than those below expectations (negative disconfirmation). Bolton et al. (2000) find

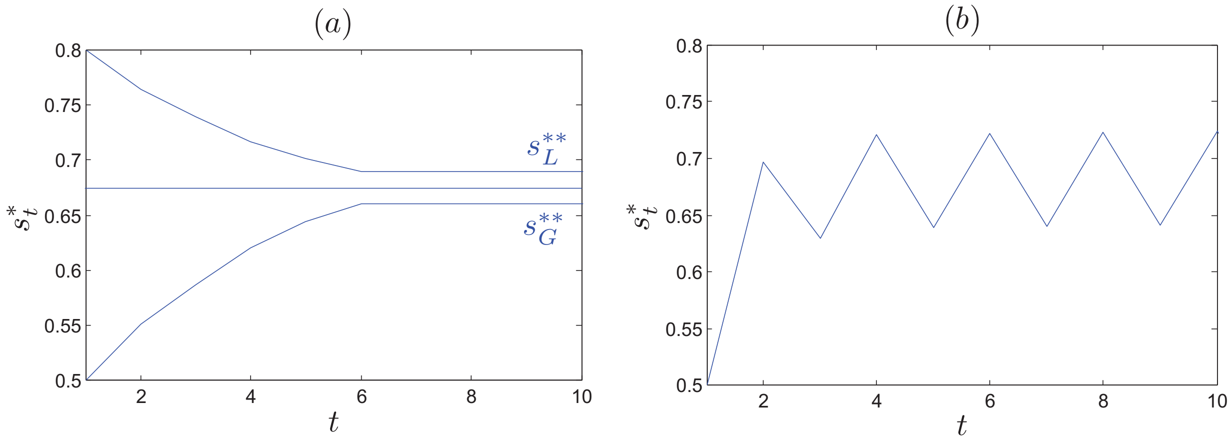

Figure 3 **Optimal service quality paths for (a), loss averse and (b) gain seeking customers.** $\pi(x) = 1 - x^2$, $F(s) = 1/(1 + \exp(-5s))$, $\lambda = 0.5$, $\beta = 0.94$. **(a)** $\lambda_L = 0.4$ **and** $\lambda_G = 0.6$; **(b)** $\lambda_L = 0.6$ **and** $\lambda_G = 0.4$.

evidence that members of loyalty rewards programs tend to discount or overlook negative service experiences. In this case, we find that no interior steady state exists. Under a high–low policy (cf. Figure 3(b)), the firm benefits in the long run by manipulating customer expectations. Here the benefits to the firm are attributable to the positive net effect of first increasing service and then decreasing it.

PROPOSITION 8. *If $\lambda_L > \lambda_G$, then problem* (9) *admits no interior steady state; that is to say, any optimal service path oscillates.*

8. Extensions: Heterogeneous Customers and Shared Resources

This section extends our setup to incorporate such "pooling" effects as customer heterogeneity, shared resources, and shared costs. In these cases, Problem (1) is no longer separable and so the firm must make trade-offs between customers.

8.1. Customer Heterogeneity and Parameter Uncertainty

So far we have assumed that the firm is fully informed about customer characteristics, I, and that it can perfectly customize service. In this section, we relax these assumptions to acknowledge the firm's uncertainty about customer types—in particular, regarding initial expectations s_0 and retention drivers (loyalty or switching costs) α. This model reflects the case in which a firm offers the same service levels to a heterogeneous pool of customers that cannot be targeted individually.

Suppose the firm knows the distribution $\tilde{\theta}$ of customer types, and $F(\cdot;\theta)$ is strictly increasing in θ; that is, suppose higher types are more likely than lower types to renew, ceteris paribus. Here $\tilde{\theta}$ may capture the uncertainty about loyalty $\tilde{\alpha}$ or about initial expectations $\tilde{s}_0$ (or both).[10] The firm's objective, given $\tilde{\theta}$, can be written as

[10] Technically, for $\tilde{\theta} = \tilde{s}_0$, we keep uncertainty separate from the state s_t by defining the latter as "delivered" service quality, i.e., setting $s_0 = 0$ in the recursive definition of s_t and $F_t(s_t;\tilde{s}_0) = F(s_t + \lambda^t \tilde{s}_0)$, a minor abuse of notation.

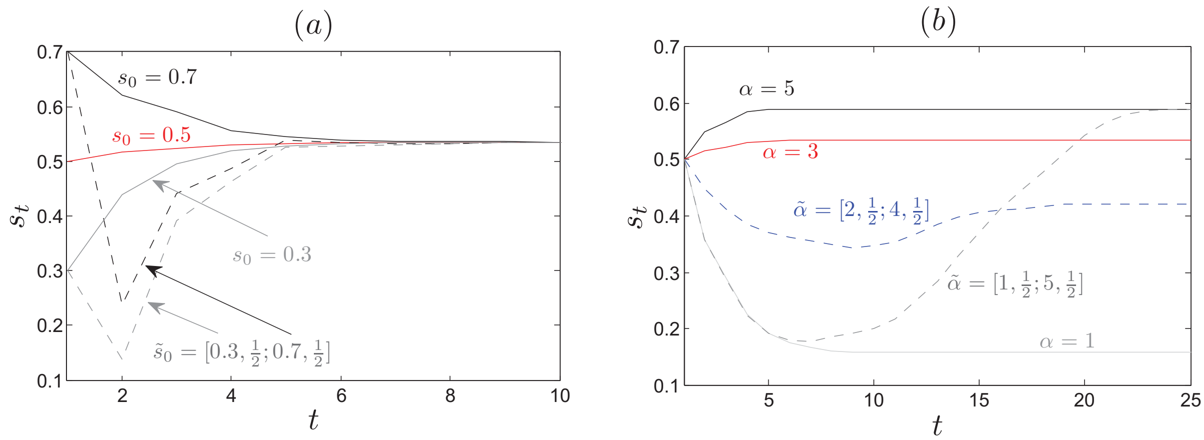

Figure 4 **Optimal state paths (dashed lines $\tilde{s}_t$) when customers are heterogeneous in (a) initial expectations $\tilde{s}_0$ (for $\alpha = 3$) and (b) loyalty $\tilde{\alpha}$ (for $s_0 = 0.5$). Solid lines represent state paths s_t for homogeneous customers; $\pi(x) = 1 - x^2$, $F(x) = 1/(1 + \exp(\alpha x))$, $\lambda = 0.5$, $\beta = 0.94$.**

$$\tilde{J}(\tilde{\theta}) = \max_{X=\{x_t\}} \mathbb{E}_{\tilde{\theta}}[J(X; \tilde{\theta})]. \tag{10}$$

A recursive expression of this imperfect observation model requires, for a given policy X, the probability that the consumer renews at stage t given that she has renewed in all previous stages:

$$\mathbb{P}\left(\tilde{U}_{t+1}(s_{t+1}; \tilde{\theta}) \geq 0 | \tilde{U}_i(s_i; \tilde{\theta}) \geq 0, i = 1, \ldots, t\right) = \frac{\mathbb{E}_{\tilde{\theta}}\left[\prod_{i=1}^{t+1} F(s_i; \tilde{\theta})\right]}{\mathbb{E}_{\tilde{\theta}}\left[\prod_{i=1}^{t} F(s_i; \tilde{\theta})\right]} = \mathbb{E}_{\tilde{\theta}_t}[F(s_{t+1}; \tilde{\theta}_t)].$$

The distribution $\tilde{\theta}_t$ of types who are alive at time t is updated recursively each period in a Bayesian fashion: $\tilde{\theta}_{t+1} \sim (\tilde{\theta}_t | \text{customer renews at time} \, t)$ with $\tilde{\theta}_0 = \tilde{\theta}$. The following Bellman equation holds for the finite T-horizon version of (10):

$$\tilde{J}_t^T(s_t, \tilde{\theta}_t) = \max \pi(x_t) + \beta \mathbb{E}_{\tilde{\theta}_t}[F(s_{t+1}; \tilde{\theta}_t)] \tilde{J}_{t+1}^T(s_{t+1}, \tilde{\theta}_{t+1}), \quad \tilde{J}_T^T(s_T, \tilde{\theta}_T) = 0, \tag{11}$$

where $s_{t+1} = \lambda x_t + (1 - \lambda) s_t$ and

$$\mathbb{P}(\tilde{\theta}_{t+1} \geq k) = \frac{\mathbb{E}[F(s_{t+1}; \tilde{\theta}_t) | \tilde{\theta}_t \geq k] \mathbb{P}(\tilde{\theta}_t \geq k)}{\mathbb{E}[F(s_{t+1}; \tilde{\theta}_t)]}, \quad \forall k \in [\theta_L, \theta_H].$$

In particular and by the same argument as in Proposition 2, concavity of π implies that, all else equal, the optimal transient policy $\tilde{s}_{t+1}^*(s_t, \tilde{\theta}_t)$ is increasing in s_t for any given $\tilde{\theta}_t$. However, this term $\tilde{\theta}_t$ also affects the optimal state paths, which need not be monotonic (see Figure 4).

Proposition 9. *For any distribution $\tilde{\theta}$ with support $[\theta_L, \theta_H]$, Problem* (10) *admits the unique steady state $\tilde{s}^{**}(\tilde{\theta}) = s^{**}(\theta_H)$.*

Proposition 9 characterizes the optimal long-run policy of the firm under parameter uncertainty. It shows that, among a pool of heterogeneous customers, the firm targets the highest types in the long run because they are more likely to be active. Since lower types are more likely to defect, the distribution of surviving types, $\tilde{\theta}_t$, increases every period (in the sense of first-order dominance) in a Bayesian fashion. This sorting argument can explain the empirical evidence of increasing retention rates, as illustrated in Fader and Hardie (2009) for the special case of $F(s_t;\theta) \equiv \theta$. Our result extends their model and insights to capture the effect of service dynamics on retention and to explore possible implications for the optimization problem of the firm.

Problem (11) is generally intractable owing to the high dimensionality of the state space. For two-point distributions $\tilde{\theta} = [\theta_H, p; \theta_L, 1-p]$, (11) is amenable to a two-dimensional state dynamic program with $\tilde{\theta}_t$ replaced by p_t and updated as $p_{t+1} = \dfrac{p_t F(s_{t+1};\theta_H)}{p_t F(s_{t+1};\theta_H) + (1-p_t) F_t(s_{t+1};\theta_L)}$. We shall use this formulation in numerical experiments to illustrate the effect of customer heterogeneity in initial expectations and loyalty (or switching costs) on the firm's policy and profit.

8.1.1. Heterogeneity in initial expectations. A consequence of Proposition 9 is that—regardless of the distribution of customers' initial expectations, $\tilde{\theta} = \tilde{s}_0$—the optimal service paths converge to the same steady-state s^{**} given by Proposition 2, because $\tilde{s}^{**}(\tilde{s}_0) = s^{**}(s_0^H) \equiv s^{**}$. In other words, facing heterogeneous customers, the firm will provide the same long-run service level as if assuming "average", or incorrect initial expectations. Figure 4(a) illustrates this effect as well as the transient policy of the firm. Consumers with different initial expectations maintain different perceptions of service quality (grey versus black dotted lines), but these converge relatively fast to the same steady state as the memory effect of initial expectations fades. In the short run, unlike the homogeneous case (Proposition 2), heterogeneous customers may each experience nonmonotonic service quality as the firm optimally juggles between extracting rents from low-type customers (who defect earlier) and maintaining its long-term relationships with high types.

8.1.2. Heterogeneity in retention drivers. A similar transient oscillation effect is triggered by customer heterogeneity in loyalty or switching costs, $\tilde{\theta} = \tilde{\alpha}$. As illustrated in Figure 4(b), the firm first targets transactional customers and then converges to the optimal policy for loyal customers in the long run. Unlike the case with initial expectations, ignoring heterogeneity in loyalty (i.e., assuming $\alpha = \mathbb{E}\tilde{\alpha}$) may lead the firm to offer service levels that are either too high or too low, in the long run, because of the unimodal relationship between long-run service and loyalty (cf. Section 5.2). Indeed, Figure 4(b) shows that $\tilde{s}^{**}(\tilde{\alpha} = [1, \frac{1}{2}; 5, \frac{1}{2}]) = s^{**}(5) > s^{**}(3) > s^{**}(4) = \tilde{s}^{**}(\tilde{\alpha} = [2, \frac{1}{2}; 4, \frac{1}{2}])$.

8.1.3. Value of information. Table 1 reports, along three dimensions, the value of information regarding customer types: (1) the effect of parameter uncertainty (variance) in $\tilde{\theta}$ on customer value (i.e., the percentage gap from certainty equivalence); (2) the cost of assuming that customers are homogeneous of average type $\theta = \mathbb{E}\tilde{\theta}$ and delivering the optimal service policy $X^*(\mathbb{E}\tilde{\theta}) = \arg\max J(\mathbb{E}\tilde{\theta}))$ corresponding to the homogeneous case; and (3) the value of full information and

Table 1 **Value of information about s_0 and α with parameters as in Figure 4;** $\tilde{s}_0 = [0.3, \frac{1}{2}; 0.7, \frac{1}{2}]$, $\tilde{\alpha} = [3, \frac{1}{2}; 7, \frac{1}{2}]$.

Uncertainty Measure		Initial Expectations $\tilde{s}_0$	Loyalty $\tilde{\alpha}$
Price of uncertainty	$\frac{J(\mathbb{E}\tilde{\theta})-\tilde{J}}{\tilde{J}(\mathbb{E}\tilde{\theta})}$	6%	45%
Cost of ignoring uncertainty	$\frac{\tilde{J}(X^*(\mathbb{E}\tilde{\theta});\tilde{\theta})-\tilde{J}}{\tilde{J}}$	4%	30%
Value of screening	$\frac{\mathbb{E}[J(\tilde{\theta})]-\tilde{J}}{\tilde{J}}$	5%	70%

targeting of individual customer types. These measures all suggest that information on customer loyalty is more valuable than initial expectations, and so is targeting service on that dimension.

At the cost of higher computational complexity, our framework in this section extends to simultaneously optimizing service and learning about customer parameters by updating a *prior* on $\tilde{\theta}$ in each period. Numerical results (not reported here for conciseness) evidence a similar policy structure with relatively limited additional profit gains (less than 3% for $\tilde{\alpha}$ and even less for $\tilde{s}_0$).

In sum, our results in this section show that the service trajectory for heterogeneous customers converges to a steady state $s^{**}(\theta_H)$ that corresponds to the most loyal types but is independent of the distribution of initial expectations. Convergence need not be monotonic, and it occurs faster for $\tilde{s}_0$ than $\tilde{\alpha}$ because of the decaying effect on service quality perceptions, and hence on retention. Overall, our results support the importance of measuring retention drivers such as loyalty and switching costs—more so than initial expectations—and of targeting service along these dimensions.

8.2. Shared Resources and Costs

In this section, we briefly discuss how our framework can extend to incorporate shared resources and cost-pooling effects across customers. For simplicity, we assume that the firm serves two customers, A and B, with corresponding retention functions F_A and F_B and memory parameters λ_A and λ_B. The total short-term profit from offering service levels x^A and x^B to these two customers is denoted $\pi(x^A, x^B)$; we assume it to be concave in each dimension and submodular. We further consider the possibility of resource constraints, $x_A + x_B \le Q$, and without loss of generality we normalize $Q = 1$. For example, $\pi(x^A, x^B) = \pi^A(x^A) + \pi^B(x^B) + \pi^C(x^A + x^B)$, where the first two terms capture the individual short-run profit (or revenue) from each customer and the latter the pooled cost of total resources deployed, or the salvage value of unutilized resources (e.g. spot market profit in the case of advertising contracts). Concavity of π^C ensures that π is submodular.

The Bellman equation to account for capacity constraints can be written as

$$J(s_t^A, s_t^B) = \max_{x_t^A + x_t^B \le 1} R(x_t^A, s_t^A, x_t^B, s_t^B) + \beta F_A(s_{t+1}^A) F_B(s_{t+1}^B) J(s_{t+1}^A, s_{t+1}^B), \tag{12}$$

where $s_{t+1}^j = \lambda_j s_t^j + (1-\lambda_j) x_t^j$, for $j \in \{A, B\}$ and

$$R(x_t^A, s_t^A, x_t^B, s_t^B) = \pi(x_t^A, x_t^B) + \beta F_A(s_{t+1}^A) \bar{F}_B(s_{t+1}^B) J_A(s_{t+1}^A) + \beta \bar{F}_A(s_{t+1}^A) F_B(s_{t+1}^B) J_B(s_{t+1}^B)$$

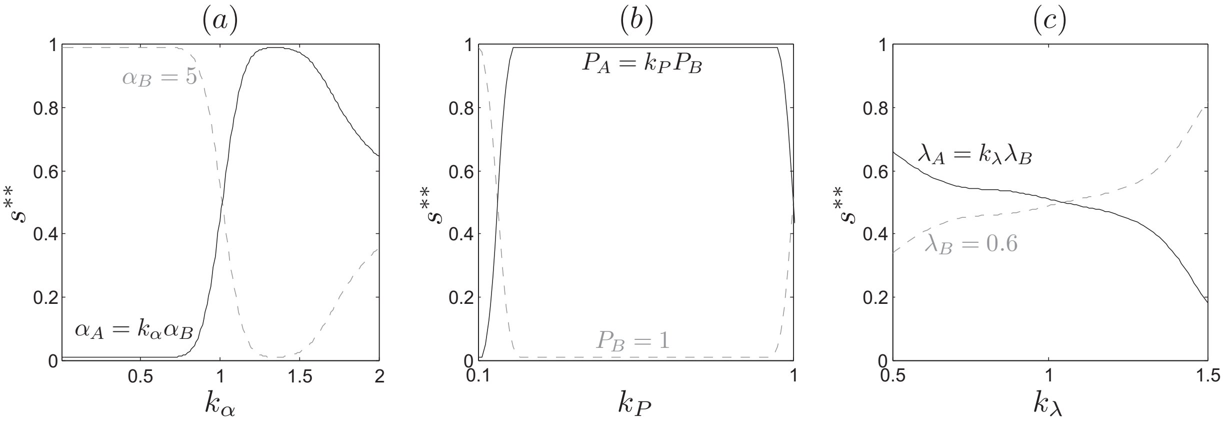

Figure 5 **Steady states vs. loyalty (a), price (b) and memory (c), for two customers with shared resources.** $\pi(x^A, x^B) = P_1 + P_2$, $\beta = 0.94$, $F_i(x) = 1/(1 + exp(-\alpha_i x + \rho_i P + 1))$, $s_0^i = 0.5, i \in \{A, B\}$. **(a)** $\lambda_i = 0.6, P_i = \rho_i = 1$; **(b)** $\lambda_i = 0.6, \alpha_i = 10, \rho_i = 5$; **(c)** $\alpha_i = 5, P_i = \rho_i = 1$.

for J_A and J_B the expected long-term profit in the corresponding single-customer models. Define $\Pi(x^A, x^B) = \frac{R(x^A, x^A, x^B, x^B)}{1 - \beta F_A(x^A) F_B(x^B)}$, the profit that corresponds to offering a constant policy (x^A, x^B) until the first customer dies and thereafter using the optimal single-customer policy for the remaining customer. Define also $W^j(x^A, x^B) = \lambda_j \pi(x^A, x^B) + (1 - \lambda_j)\Pi(x^A, x^B)$ for $j \in \{A, B\}$.

PROPOSITION 10. *(a) The steady states for Problem (12), (s_A^{**}, s_B^{**}), satisfy one of the following conditions: (i) $W_1^A(s_A, s_B) = 0, W_2^B(s_A, s_B) = 0$ and $s_A + s_B \leq 1$; (ii) $W_1^A(s_A, s_B) = W_2^B(s_A, s_B)$ and $s_B = 1 - s_A$; (iii) s_A is a steady state of J_A and $s_B = 0$; (iv) s_B is a steady state of J_B and $s_A = 0$. (b) The optimal service quality policy for customer A, $s_A^*(s^A, s^B)$ is increasing in s^A and decreasing in s^B, and conversely for customer B.*

Part (a) of the proposition characterizes steady-state service levels for the two customers under capacity constraints, extending Proposition 1. In the first two cases, unlike the latter two, both customers are served in the long run. In particular, if capacity is binding (case (ii)) then the firm balances in steady state the marginal long-term profit from each customer. In this case, Figure 5 illustrates the robustness of our sensitivity results in Section 5 when $\pi(x^A, x^B) = P_1 + P_2$ and $x^A + x^B \leq 1$. Specifically, the steady state s^{**} remains unimodal with respect to changes in loyalty α and price P, and it also remains decreasing in memory λ. This is contrary to Adelman and Mersereau (2010), who find that higher-margin customers always get better service and that memory does not affect the steady state when demand is deterministic; however neither retention nor loyalty is captured in their average reward model.

Part (b) of Proposition 10 extends Proposition 2 in the capacitated case. Our numerical results suggest that the optimal service quality paths need not be monotonic (since there may be strategic substitution between the two customers) but in the long run they converge to a steady state.

9. Conclusions

We relied on behavioral theories to develop a dynamic programming model of using responsive service strategies to manage retention over time in a contractual relationship. In this context, we showed that firms can extract more value in the long run by gradually managing service experiences and expectations over time toward an "ideal" steady-state service level from which it is suboptimal to deviate. This steady state does not depend on customers' service quality expectations, and it is lower than the optimal constant service offered by firms that ignore the effect of past experiences on retention. We discover that behavioral asymmetries (such as loss aversion) drive the structure of optimal long-run policies by ensuring convergence, and increasing the prevalence of constant service policies. Indeed, the more customers are averse to service downgrades, the wider is the range of optimal constant service policies offered by the firm and the lower are its profits from those policies.

Higher-margin customers are not always more valuable, and they need not receive better service because prices above a threshold make the relationship more transactional. Customers who are more loyal are more valuable, but they may not receive better service if they are inherently too sticky. Our model predicts a unimodal relationship between loyalty and service. From the customers' perspective, this means that there is an ideal, intermediate level of loyalty (driven, e.g., by switching costs or the level of market competition) that fetches the best service. The firm offers less service in the long run to less adaptive customers (i.e., those who anchor more on past experiences); these are more profitable than customers who focus on recent experiences, but only if they have received better treatment in the past. The relationships between various consumer characteristics and the firm's policy and profits are summarized in Table 2.

These findings are robust when customers are heterogeneous and when accounting for the impact of prior service experiences on purchase volumes, visit frequency, and referrals. It is interesting that the insights reported here are also independent of the shape of the retention function. However, that function does influence the initial sequence of service experiences as well as the marginal returns to service quality. Indeed, in a transient regime, we find that customers who had better experiences may receive higher or lower service depending on the marginal effect of service quality on retention. In heterogeneous markets, the firm capitalizes first on customers who are more transactional but, over the long run targets those who are most likely to renew (e.g., because they are more loyal, have lower switching costs, or had better past experiences). Ignoring heterogeneity in loyalty or switching costs is more costly than ignoring differences in service quality expectations, and the value of information and screening on these dimensions is significantly higher.

Table 2 Effects of Margin and Consumer Behavior on Policy and Value

Characteristics (I)	**Long-run Service** (steady state s^{**})	**Customer Value** (J)
Price (P)	Inverse U-shape	Inverse U-shape
Loyalty (α)	Inverse U-shape	Increasing
Adaptation (λ)	Increasing	Increasing for high s_0; decreasing for low s_0
Loss aversion ($\rho = \lambda_G/\lambda_L > 1$)	Expands range (oscillates if $\rho < 1$)	Decreasing
Initial expectations (s_0)	No effect	Increasing

This paper is a first step toward capturing behavioral effects of service dynamics on customer retention and profitability in a business relationship. We have therefore strived for parsimony in developing the simplest stylized model capable of transmitting the main insights from this framework. Ample opportunities exist for future research to extend this model and address its limitations from an operational, marketing, or economic perspective—for example, by incorporating strategic interactions, acquisition spending, or richer operational structures. For expository purposes, we cast our model and results in the context of managing service relationships. However, our setup can be expanded to broader contexts, including dynamic pricing, employee retention and effort and quality management.

References

Adelman, D., A.J. Mersereau. 2010. Dynamic capacity allocation to customers who remember past service. Tech. rep., Working Paper.

Araman, V.F., I. Popescu. 2010. Media revenue management with audience uncertainty: Balancing upfront and spot market sales. *Manufacturing & Service Operations Management* **12**(2) 190–212.

Berger, P. D., N. I. Nasr. 1998. Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing* **12**(1) 17–30.

Bertsekas, D.P. 2007. *Dynamic programming and optimal control*. Athena Scientific.

Bitran, G. R., J. C. Ferrer, P. Rocha Oliveira. 2008. Managing customer experiences: Perspectives on the temporal aspects of service encounters. *Manufacturing & Service Operations Management* **10**(1) 61.

Bolton, R. N., J. H. Drew. 1991. A multistage model of customers' assessments of service quality and value. *Journal of Consumer Research* **17**(4) 375–384.

Bolton, R. N., P.K. Kannan, M.D. Bramlett. 2000. Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the Academy of Marketing Science* **28**(1) 95–108.

Bolton, R. N., K. N. Lemon, M. D. Bramlett. 2006. The effect of service experiences over time on a supplier's retention of business customers. *Management Science* **52**(12) 1811–1823.

Boulding, W., A. Kalra, R. Staelin, V. A. Zeithaml. 1993. A dynamic process model of service quality: From expectations to behavioral intentions. *Journal of Marketing Research* **30**(1) 7–27.

Bowman, D., D. Narayandas. 2004. Linking customer management effort to customer profitability in business markets. *Journal of Marketing Research* **41**(4) 433–447.

Caulkins, J.P., G. Feichtinger, J. Haunschmied, G. Tragler. 2006. Quality cycles and the strategic manipulation of value. *Operations Research* **54**(4) 666–677.

Cronin, J. J., S. A. Taylor. 1992. Measuring service quality: a reexamination and extension. *Journal of Marketing* **56**(3) 55–68.

DeVine, J., K. Gilson. 2010. Using behavioral science to improve the customer experience. *McKinsey Quarterly* **February**.

Fader, P.S., B.G.S. Hardie. 2009. Probability models for customer-base analysis. *Journal of Interactive Marketing* **23**(1) 61–69.

Gans, Noah. 2003. Customer loyalty and supplier quality competition. *Management Science* **48**(2) 207–221.

Gaur, V., Y. Park. 2007. Asymmetric consumer learning and inventory competition. *Management Science* **53**(2) 227.

Hall, J., E. Porteus. 2000. Customer service competition in capacitated systems. *Manufacturing & Service Operations Management* **2**(2) 144–165.

Ho, T., Y. H. Park, Y. P. Zhou. 2006. Incorporating Satisfaction into Customer Value Analysis: Optimal Investment in Lifetime Value. *Marketing Science* **25**(3) 260 – 277.

Hogarth, R. M., H. J. Einhorn. 1992. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology* **24**(1) 1–55.

Jones, T. O., W. E. Sasser. 1995. Why satisfied customers defect. *Harvard Business Review* **73**(6) 88–91.

Kahneman, D., P. P. Wakker, R. Sarin. 1997. Back to bentham? explorations of experienced utility. *Quarterly Journal of Economics* **112**(2) 375.

Lewis, M. 2005. Research note: a dynamic programming approach to customer relationship pricing. *Management Science* 986–994.

Liu, B. S. C., N. C. Petruzzi, D. Sudharshan. 2007. A service effort allocation model for assessing customer lifetime value in service marketing. *Journal of Services Marketing* **21** 24–45.

Liu, Q., G. van Ryzin. 2009. Strategic capacity rationing when customers learn. *Working paper, Columbia University* .

Loch, Ch., Y. Wu. 2007. Behavioral operations management. *Foundations and Trends in Technology, Information and Operations Management* **1**(3).

Nerlove, M., K.J. Arrow. 1962. Optimal advertising policy under dynamic conditions. *Economica* 129–142.

Olsen, T.L., R.P. Parker. 2008. Inventory management under market size dynamics. *Management Science* **54**(10) 1805–1821.

Ovchinnikov, A., Ph.E. Pfeifer. 2011. Marketing spending and customer lifetime value for firms with limited capacity. Darden School of Business Working Paper Series.

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

Primas, H. 1999. Basic elements and problems of probability theory. *Journal of Scientific Exploration* **13**(4) 579–613.

Radner, R. 2003. Viscous demand. *Journal of Economic Theory* **112**(2) 189–231.

Rust, R. T., T. S. Chung. 2006. Marketing models of service and relationships. *Marketing Science* **25**(6) 560–580.

Rust, R. T., K. N. Lemon, V. A. Zeithaml. 2004. Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing* **68**(1) 109–127.

Rust, R. T., A. J. Zahorik, T. L. Keiningham. 1995. Return on quality (roq): Making service quality financially accountable. *The Journal of Marketing* **59**(2) 58–70.

Su, X. 2009. A model of consumer inertia with applications to dynamic pricing. *Production and Operations Management* **18**(4) 365–380.

Topkis, D. M. 1998. *Supermodularity and complementarity*. Princeton University Press.

Tversky, A., D. Kahneman. 1991. Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics* **106**(4) 1039–1061.

Venkatesan, R., V. Kumar. 2004. A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing* 106–125.

Zeithaml, V. A. 2000. Service quality, profitability, and the economic worth of customers: What we know and what we need to learn. *Journal of the Academy of Marketing Science* **28**(1) 67.

Appendix A: Proofs

Proof of Lemma 1. (a) Monotonicity of the value function holds because $F(\lambda s + (1-\lambda)x)$ is increasing in s. Monotonicity is preserved by induction for the corresponding finite-horizon model and then at the limit for our infinite-horizon formulation. The firm can extract at least $\Pi(s)$ from the customer by maintaining the service quality at current expectations (i.e., by offering $\{x_t \equiv s\}$), so $\Pi(s) \leq J(s)$. An upper bound on customer value is obtained if the customer never defects and the firm offers the short-term profit maximizing service; in this case $J(s) \leq \sum_{t=0}^{\infty} \beta^t \max_s \pi(s) = \frac{\pi(\underline{s})}{1-\beta}$.

(b) Part (a) and F increasing imply that the profit-to-go in the Bellman equation is increasing. Because π is concave, its maximizer $\underline{s} \leq x^*(s)$ for all s; in particular, $\pi'(x^*(s)) \leq 0$.

Proof of Proposition 1. Define $\bar{\pi}(s,S)=\pi(\frac{S-\lambda s}{1-\lambda})$. Then, in terms of the variable $s_{t+1}=\lambda s_t+(1-\lambda)x_t$, Problem (5) becomes

$$J(s_t)=\max_{s_{t+1}\in\mathbf{s}(s_t)}\bar{\pi}(s_t,s_{t+1})+\beta F(s_{t+1})J(s_{t+1}). \tag{13}$$

Here $\mathbf{s}(s_t)=[\lambda s_t,\lambda s_t+(1-\lambda)]$ is the feasible set of next-period service quality s_{t+1} associated with the constraint $x_t\in[0,1]$. We will use this alternative formulation for technical convenience and because it facilitates extensions.

First, boundary steady states can be ruled out because π and Π have interior maximizers. Indeed, by Lemma 1(b), $s^{**}\geq\underline{s}>0$ and so 0 cannot be a steady state. Also, $s=1$ cannot be a steady state because $J(1)\geq J(\bar{s})\geq\Pi(\bar{s})>\Pi(1)$; hence the constant path $x_t\equiv 1$ cannot be optimal.

We can therefore focus on interior steady states, which are given by the following Euler equation:

$$\frac{\partial}{\partial s_{t+1}}\left\{\bar{\pi}(s_t,s_{t+1})+\beta F(s_{t+1})\Big(\bar{\pi}(s_{t+1},s_{t+2})+\beta F(s_{t+2})\Pi(s_{t+2})\Big)\right\}_{s_t=s_{t+1}=s_{t+2}=s}=0.$$

This can be rewritten as

$$\bar{\pi}_2(s,s)+\beta F'(s)\Pi(s)+\beta F(s)\bar{\pi}_1(s,s)=0. \tag{14}$$

Differentiating $\Pi(s)=\frac{\bar{\pi}(s,s)}{1-\beta F(s)}$ then gives $(1-\beta F(s))\Pi'(s)=\bar{\pi}_1(s,s)+\bar{\pi}_2(s,s)+\beta F'(s)\Pi(s)$, which allows to write the Euler Equation (14) as

$$\Pi'(s)=\bar{\pi}_1(s,s). \tag{15}$$

This is equivalent to $W'(s)=\lambda\pi'(s)+(1-\lambda)\Pi'(s)=0$ because $\bar{\pi}_1(s,s)=-\frac{\lambda}{1-\lambda}\pi'(s)$. Thus a steady state maximizes W, and the strict quasi-concavity of W ensures uniqueness. Furthermore, $s^{**}\leq\bar{s}$ follows because Π is unimodal and $\Pi'(s^{**})=-\frac{\lambda}{1-\lambda}\pi'(s^{**})\geq 0$ by Lemma 1 (b).

To show that $s^{**}(\lambda)$ is decreasing in λ, we use the envelope theorem to derive $W_{s\lambda}(s^{**})=\pi'(s^{**})-\Pi'(s^{**})=\frac{1}{1-\lambda}\pi'(s^{**})\leq 0$, as argued previously.

To show that $s^{**}(\beta)$ is increasing, it is sufficient to show that $\Pi(s;\beta)$ and hence $W(s;\beta)$ are supermodular. In fact, it suffices to show that $\log(\Pi(s;\beta))=\log(\pi(s))-\log(1-\beta F(s))$ is supermodular in (s,β). The derivative of this expression with respect to β is $F(s)L(s)$, so indeed it is increasing in s.

Proof of Proposition 2. Since $\pi(x)$ is concave, it follows that $\bar{\pi}(s_t,s_{t+1})$ is supermodular and so is the term on the right-hand side of the Bellman equation (13). Moreover, the feasible sets $\mathbf{s}(s_t)$ are ascending in s_t; that is, for any $s_t\leq s_t'$, $r\in\mathbf{s}(s_t)$, and $r'\in\mathbf{r}(s_t')$ we have $\min(r,r')\in\mathbf{s}(s_t)$ and $\max(r,r')\in\mathbf{s}(s_t')$. Therefore, by Topkis's theorem (Topkis 1998, Thm. 2.8.2), the policy function $s^*(\cdot)$ is increasing on $[0,1]$. It follows that $s^*(\cdot)$ has a fixed point, $s^{**}=s^*(s^{**})$, that is a steady state of Problem (13).

Finally, monotonicity of $s^*(\cdot)$ implies that the state path $\{s_t^*\}$ is monotonic (by induction); because the feasible set $\mathbf{s}(\cdot)$ is compact, $\{s_t^*\}$ must converge to a steady state. Hence a steady state exists, and our assumptions ensure that it is unique and interior. Note that, in steady state we have $s^{**} = \lambda s^{**} + (1-\lambda)x^{**}$, so $s^{**} = x^{**}$. In particular, because the service quality paths converge monotonically to s^{**} (which maximizes the unimodal function W), it follows that W is a Lyapounov function for our problem.

Proof of Proposition 3. The result holds because convexity is preserved by maximization and limits. Indeed, if $F(\cdot)$ is convex then it follows by induction that the corresponding finite-horizon value function is also convex. By value iteration, we can take limits to obtain that the infinite-horizon value function $J(s)$ is convex; therefore, $V(s) = F(s)J(s)$ is convex.

We further show that V convex (resp. concave) is sufficient for the service policy to be increasing (decreasing) and the value function J to be convex (concave), confirming the statement at the end of Section 4.3. Now, V convex (concave) is equivalent to the argument on the right-hand side of the Bellman equation, $Q(x,s) = \pi(x) + \beta V(\lambda s + (1-\lambda)x)$, being supermodular (submodular). Monotonicity of the optimal policy then follows by Topkis's theorem. Moreover, V convex implies that $Q(x,s)$ is convex in s, so J is convex. On the other hand, V and π concave implies that $Q(x,s)$ is jointly concave, so J must be concave.

Proof of Proposition 4. For a given s_0, consider the optimal service path $s_{t+1} = s^*(s_t)$ for all t. By the envelope theorem, for all t we have

$$\frac{\partial}{\partial\lambda} J(s_t;\lambda) = \frac{s_{t+1} - s_t}{(1-\lambda)^2}\pi'\Big(\frac{s_{t+1} - \lambda s_t}{1-\lambda}\Big) + \beta F(s_{t+1})\frac{\partial}{\partial\lambda} J(s_{t+1};\lambda). \tag{16}$$

In particular, at the steady state $s_t = s_{t+1} = s^{**}(\lambda)$, this gives $\frac{\partial}{\partial\lambda} J(s;\lambda)|_{s=s^{**}} = \beta F(s)\frac{\partial}{\partial\lambda} J(s;\lambda)|_{s=s^{**}}$ or $\frac{\partial}{\partial\lambda} J(s;\lambda)|_{s=s^{**}} = 0$. For $s_0 \geq s^{**}(\lambda)$, by Proposition 2 the optimal service quality path is decreasing $s_{t+1} \leq s_t$ for all t. Because $\pi'(x_t) \leq 0$ on an optimal path by Lemma 1(b), the first term on the right-hand side of (16) is positive. Thus, for any $t > 0$, we have

$$\begin{aligned}
\frac{\partial}{\partial\lambda} J(s_0;\lambda) &\geq \beta F(s_1)\frac{\partial}{\partial\lambda} J(s_1;\lambda) && (17)\\
&\geq \beta^2 F(s_1)F(s_2)\frac{\partial}{\partial\lambda} J(s_2;\lambda) \geq \cdots && (18)\\
&\geq \lim_{t\to\infty} \beta^t \left(\prod_{i=1}^{t} F(s_i)\right)\frac{\partial}{\partial\lambda} J(s_t;\lambda) = 0. && (19)
\end{aligned}$$

The last derivative is bounded as $t \to \infty$ because $s_t \to s^{**}(\lambda)$ (Proposition 2). The case $s_0 \leq s^{**}(\lambda)$ is proved similarly.

Proof of Proposition 5. It is easy to see that Lemma 1 extends—in particular, $\pi_1(x^*(s), s) \leq 0$. Here we follow the same steps used to obtain the Euler Equation (15) in the proof of Proposition 1.

In this case, $\Pi'(s) = \bar{\pi}_1(s,s) = \pi_2(s,s) - \frac{\lambda}{1-\lambda}\pi_1(s,s)$ or equivalently $W'(s) = (1-\lambda)\pi_2(s,s)$, which gives the desired results. The right-hand side of the first equation is positive at an optimal solution, via Lemma 1(b), which implies $s^{**} \le \bar{s} = \arg\max \Pi(s)$. That s^{**} is unique and decreasing in λ follows if $\bar{\pi}_1(s,s)$ is increasing in s, which in turn is implied by the assumptions in the proposition. Indeed,

$$\frac{\partial}{\partial s}\bar{\pi}_1(s,s) = \frac{1}{1-\lambda}\frac{\partial}{\partial s}\left(\pi_2(s,s) - \lambda\frac{\partial}{\partial s}\pi(s,s)\right) = \frac{1}{1-\lambda}\left[\frac{\partial}{\partial s}\pi_2(s,s) - \lambda\frac{\partial^2}{\partial s^2}\pi(s,s)\right].$$

The first term is positive because $\pi(x,s)$ is supermodular and convex in s, and the second because $\pi(s,s)$ is concave in s.

For Proposition 2 to extend, all we need is supermodularity of $\bar{\pi}(s,S) = \pi(\frac{S-\lambda s}{1-\lambda}, s)$ in (s,S); this ensures that the state path is monotonic. The condition obtains because $\pi(x,s)$ is supermodular in (x,s) and concave in x. Convexity of $\pi(x,s)$ in s ensures that Proposition 3 extends. It is also easy to see that Proposition 4 extends, via Lemma 1(b).

Proof of the Results in Section 6.3. Let $J(N,s_t)$ denote the expected profit from N consumers in state s_t. With the referral acquisition process, the problem remains separable at the customer level; that is, $J(N,s_t) = NJ(1,s_t)$. Denote by $\bar{F} = 1-F$ and $\bar{F}_a = 1-F_a$ the defection probabilities. The Bellman equation for the latter is

$$\begin{aligned}
J(1,s_t) &= \max_x \pi(x) + \beta F(s_{t+1})F_a(s_{t+1})J(2,s_{t+1}) + \beta[F(s_{t+1})\bar{F}_a(s_{t+1}) + \bar{F}(s_{t+1})F_a(s_{t+1})]J(1,s_{t+1})\\
&= \max_x \pi(x) + \beta(2F(s_{t+1})F_a(s_{t+1}) + F(s_{t+1})\bar{F}_a(s_{t+1}) + \bar{F}(s_{t+1})F_a(s_{t+1}))J(1,s_{t+1})\\
&= \max_x \pi(x) + \beta(F(s_{t+1}) + F_a(s_{t+1}))J(1,s_{t+1}).
\end{aligned}$$

We can recover our original model by using $F + F_a$ instead of F. To explain the first equation we remark that, for every customer who is alive at time t, we have at time $t+1$: (a) two identical customers in state s_{t+1} with probability $F(s_{t+1})F_a(s_{t+1})$; (b) one (old or new) customer in state s_{t+1} with probability $F(s_{t+1})\bar{F}_a(s_{t+1}) + F_a(s_{t+1})\bar{F}(s_{t+1})$; or (c) no customer. The second equation follows from the separability property and the last by rearranging terms.

Proof of Proposition 6. Let h denote the inverse function of $H(x,s)$; thus, $h(H(x,s),s) = x$. The Bellman equation can be rewritten with respect to the variable s_{t+1} as

$$J(s_t) = \max_{s_{t+1}\in\mathbf{s}(s_t)} Q(s_t,s_{t+1}) = \pi(h(s_{t+1},s_t)) + \beta F(s_{t+1})J(s_{t+1}), \tag{20}$$

where $\mathbf{s}(\cdot)$ represents the corresponding feasible set. Monotonicity of H ensures that the results in Lemma 1 are preserved—in particular, $\pi'(x_t) \le 0$ at an optimal solution—so we can focus on the interval where π is decreasing. Much as in the proof of Proposition 2, supermodularity of Q implies monotonicity of the policy function $s^*(\cdot)$. The reason is that $H(x_t, s_t)$ is increasing in s_t and so the feasible sets $\mathbf{s}(s_t)$ are ascending. We have

$$Q_{12}(s_t, s_{t+1}) = \pi''(h(s_{t+1}, s_t))h_1(s_{t+1}, s_t)h_2(s_{t+1}, s_t) + \pi'(h(s_{t+1}, s_t))h_{12}(s_{t+1}, s_t), \tag{21}$$

where, as before, derivatives are denoted by corresponding subscripts. Differentiating

$$H(h(s_{t+1}, s_t), s_t) = s_{t+1} \tag{22}$$

with respect to s_{t+1} and s_t yields (respectively)

$$h_1(s_{t+1}, s_t) = \frac{1}{H_1(x_t, s_t)} > 0 \quad \text{and} \quad h_2(s_{t+1}, s_t) = -\frac{H_2(x_t, s_t)}{H_1(x_t, s_t)} \leq 0 \tag{23}$$

by the assumptions in the proposition. Finally, we take the cross partial derivative of (22) with respect to (s_{t+1}, s_t) and substitute $h_1(s_{t+1}, s_t)$ and $h_2(s_{t+1}, s_t)$ from (23); the result is

$$h_{12}(s_{t+1}, s_t) = \frac{H_2(x_t, s_t)H_{11}(x_t, s_t) - H_{12}(x_t, s_t)H_1(x_t, s_t)}{[H_2(x_t, s_t)]^3} \leq 0. \tag{24}$$

Since π is decreasing and concave on the relevant domain, it follows from (23) and (24) that $Q_{12} \geq 0$. The rest of the proof is similar to that of Proposition 1.

Proof of Proposition 7. We transform Problem (9) to obtain

$$J^K(s_t) = \max_{s^K_{t+1} \in \mathbf{s}(s_t)} \bar{\pi}(s_t, s^K_{t+1}) + \beta F(s^K_{t+1}) J^K(s^K_{t+1}), \tag{25}$$

where the state transition s^K satisfies (8). This can also be written as $s^K_{t+1} = \min\{s^G_{t+1}, s^L_{t+1}\}$, where $s^j_{t+1} = s_t + (1 - \lambda_j)(x_t - s_t), j \in \{G, L\}$. Indeed, when $x_t \geq s_t$ we have $s^K_{t+1} = s^G_{t+1} \leq s^L_{t+1}$ (because $1 - \lambda_L > 1 - \lambda_G$) and when $x_t \leq s_t$ we have $s^K_{t+1} = s^L_{t+1} \leq s^G_{t+1}$.

We shall use (P_κ) to denote the smooth Problem (5) with $\lambda = \lambda_\kappa = \kappa\lambda_G + (1 - \kappa)\lambda_L$, so $s^\kappa_{t+1} = \kappa s^G_{t+1} + (1 - \kappa)s^L_{t+1}$. We also define (P_G) and (P_L) as (P_κ) when $\kappa = 0$ and $\kappa = 1$, respectively. Let $J^i(s_t)$ denote the value function and s^{**}_i the steady state of the corresponding problems (P_i), $i \in \{G, L, \kappa\}$. In particular, $s^{**}_{\kappa=0} = s^{**}_L$ and $s^{**}_{\kappa=1} = s^{**}_G$.

Lemma 2. *(a) $J^K(s) \leq J^\kappa(s)$ for all s. (b) If s^{**}_κ is a steady state for (P_κ), then it is also a steady state for Problem* (25)*.*

Proof. (a) The claim follows from $s^\kappa_{t+1} = \kappa s^G_{t+1} + (1 - \kappa)s^L_{t+1} \geq \min\{s^G_{t+1}, s^L_{t+1}\} = s^K_{t+1}$ if we first use induction on the finite-horizon versions of the corresponding problems and then employ value iteration. (b) Starting from s^{**}_κ, a constant service quality path is optimal for (P_κ). This path is feasible for our kinked Problem (25) and achieves the same value $J^K(s^{**}_\kappa) = J^\kappa(s^{**}_\kappa)$. Therefore, by Lemma 2(a), the constant path s^{**}_κ must be optimal for Problem (25) and so s^{**}_κ is also a steady state for this problem. $\square$

By Proposition 2, if we start from $s_0 > s^{**}_{\kappa=0} = s^{**}_L$ then the optimal service quality path in (P_L) decreases to s^{**}_L. This path is feasible for Problem (25) and gives the same value in both problems. Because $J^L(s) \geq J^K(s)$ for all s, the optimal path for (P_L) is also optimal for Problem (25). The case $s_0 < s^{**}_{\kappa=1} = s^{**}_G$ is analogous.

By Proposition 1, s^{**}_κ solves $W'(s;\lambda_\kappa) = 0$ when $\lambda_\kappa = \kappa\lambda_G + (1-\kappa)\lambda_L$. For $\kappa = 0$ (resp. $\kappa = 1$), the solution to $W'(s;\lambda_\kappa) = 0$ is s^{**}_L (resp. s^{**}_G). Continuity of $W'(s;\lambda)$ ensures that, for all $s \in [s^{**}_G, s^{**}_L]$, there exists a $\kappa \in [0,1]$ such that $W'(s;\lambda_\kappa) = 0$; in other words, s is a steady state for (P_κ). By Lemma 2, it is also a steady state for Problem (25) and hence for Problem (9).

Proof of Proposition 8. Suppose that the interior steady state exists and is equal to x. It follows that, starting from $s_0 = \mathrm{x}$, any deviation from the constant service path $\{x_t \equiv \mathrm{x}, \forall t\}$ is not profitable. Consider the following two deviations:

$$\wp_+(e) = \{x_1 = \mathrm{x} + e,\, x_2 = \mathrm{x} - d,\, x_t = \mathrm{x},\, t \geq 3\}, \tag{26}$$

$$\wp_-(e) = \{x_1 = \mathrm{x} - e,\, x_2 = \mathrm{x} + f,\, x_t = \mathrm{x},\, t \geq 3\}; \tag{27}$$

here $d = \frac{\lambda_L(1-\lambda_G)}{1-\lambda_L} e$ and $f = \frac{\lambda_G(1-\lambda_L)}{1-\lambda_G} e$. It follows that $s_2 = \mathrm{x}$ for both paths. The profit associated with each path is

$$J^{\wp_+(e)}(\mathrm{x}) = \pi(\mathrm{x}+e) + \beta F(\lambda_G \mathrm{x} + (1-\lambda_G)(\mathrm{x}+e))\Big(\pi(\mathrm{x}-d) + \beta F(\mathrm{x})\Pi(\mathrm{x})\Big), \tag{28}$$

$$J^{\wp_-(e)}(\mathrm{x}) = \pi(\mathrm{x}-e) + \beta F(\lambda_L \mathrm{x} + (1-\lambda_L)(\mathrm{x}-e))\Big(\pi(\mathrm{x}+f) + \beta F(\mathrm{x})\Pi(\mathrm{x})\Big). \tag{29}$$

Suboptimality of any deviation from the constant path $\{\mathrm{x}\}$ implies that the derivative of $J^{\wp_+(e)}(\mathrm{x})$ and $J^{\wp_-(e)}(\mathrm{x})$ with respect to e, evaluated at $e = 0$, should be negative. Since $\pi'(\mathrm{x}) \leq 0$ at a steady state by Lemma 1(b), it follows that

$$\frac{d}{de} J^{\wp_+(e)}(\mathrm{x}) + \frac{d}{de} J^{\wp_-(e)}(\mathrm{x})\Big|_{e=0} = \beta(\lambda_L - \lambda_G)\left(F'(\mathrm{x})\Pi(\mathrm{x}) - \frac{1-\lambda_L\lambda_G}{(1-\lambda_G)(1-\lambda_L)} F(\mathrm{x})\pi'(\mathrm{x})\right) \leq 0. \tag{30}$$

This expression contradicts $\lambda_L > \lambda_G$, so an interior steady state does not exist. Boundary steady states are ruled out in the same way as in Proposition 1.

Proof of Proposition 9. Monotonicity of the policy function follows as usual from the supermodularity of short-term profit $\bar{\pi}$. For notational convenience, we omit the horizon-length superscript T. Because F is increasing in θ we have $\mathbb{E}[F(s;\tilde{\theta}_t)|\tilde{\theta}_t \geq k] \geq \mathbb{E}[F(s;\tilde{\theta}_t)]$ for all t and $k \in [\theta_L, \theta_H]$, which implies that

$$\mathbb{P}(\tilde{\theta}_{t+1} \geq k) = \frac{\mathbb{E}[F(s_{t+1};\tilde{\theta}_t)|\tilde{\theta}_t \geq k]\mathbb{P}(\tilde{\theta}_t \geq k)}{\mathbb{E}[F(s_{t+1};\tilde{\theta}_t)]} \geq \mathbb{P}(\tilde{\theta}_t \geq k). \tag{31}$$

Hence $\tilde{\theta}_{t+1} \succeq_{FSD} \tilde{\theta}_t$ and so, by the monotone convergence theorem (Primas 1999), the distributions must converge to that of $\tilde{\theta}^* \succeq_{FSD} \tilde{\theta}$. Rearranging terms and taking limits in (31) yields that, for any $k \in [\theta_L, \theta_H]$, either $0 = \mathbb{P}(\tilde{\theta}^* \geq k) \geq \mathbb{P}(\tilde{\theta} \geq k)$ (implying $k = \theta_L$) or $0 = \lim_t \frac{\mathbb{E}[F(s_t;\tilde{\theta}^*)|\tilde{\theta}^* < k]}{\mathbb{E}[F(s_t;\tilde{\theta}^*)]}$. Because F is strictly increasing in θ, the latter cannot hold for all $k \in [\theta_L, \theta_H]$ unless $\tilde{\theta}^*$ is a Dirac at θ_H.

Proof of Proposition 10. Define $\bar{\pi}(s^A, S^A, s^B, S^B) = \pi(x^A, x^B)$ and similarly $\bar{R}(s^A, S^A, s^B, S^B) = R(x^A, s^A, x^B, s^B)$, where $x^A = \frac{S^A - \lambda_A s^A}{1-\lambda_A}$ and $x^B = \frac{S^B - \lambda_B s^B}{1-\lambda_B}$.

(a) The system of Euler equations obtains by equating to zero the partial derivatives with respect to s^A_{t+1}, respectively s^B_{t+1}, of

$$R(s^A_{t+1}, s^A_t, s^B_{t+1}, s^B_t) + \beta F_A(s^A_{t+1}) F_B(s^B_{t+1}) \left(R(s^A_{t+2}, s^A_{t+1}, s^B_{t+2}, s^B_{t+1}) + \beta F_A(s^A_{t+2}) \bar{F}_B(s^B_{t+2}) \Pi(s^A_{t+2}, s^B_{t+2}) \right),$$

evaluated at $s^i_t = s^i_{t+1} = s^i_{t+2}, i \in \{A, B\}$. These equations, with arguments omitted for convenience, can be written as

$$\bar{R}_2 + \beta F_A F_B \bar{R}_1 + \beta F'_A F_B \Pi = \bar{R}_4 + \beta F_A F_B \bar{R}_3 + \beta F_A F'_B \Pi = 0.$$

Differentiating Π with respect to each component we obtain

$$(1 - \beta F_A F_B)\Pi_1 = \bar{R}_1 + \bar{R}_2 + \beta F'_A F_B \Pi \text{ and } (1 - \beta F_A F_B)\Pi_2 = \bar{R}_3 + \bar{R}_4 + \beta F_A F'_B \Pi.$$

Pairing these expressions up with the Euler equations, the latter can be written as $\bar{R}_1 = \Pi_1$ and $\bar{R}_3 = \Pi_2$. Moreover, from the definition of $\bar{R}$ it is easy to see that $\bar{R}_1 = \bar{\pi}_1$ and $\bar{R}_3 = \bar{\pi}_3$. Finally, we use the definition of $\bar{\pi}$ to obtain precisely the equations in case (i).

The system just described might not admit a solution that meets the capacity constraint. In that case, either the firm serves only one customer (case (iii) or (iv)) or the resource is fully utilized in steady state; that is, $s^{**}_A + s^{**}_B = 1$. Plugging this into the Euler equation gives the desired result.

(b) This part follows from the Topkis theorem (Topkis 1998, Thm. 2.8.2) because π submodular and concave implies $\bar{\pi}$ supermodular in (s^j, S^j) and submodular in (s^i, S^j) and in (S^i, S^j), $i \neq j$.

Europe Campus
Boulevard de Constance
77305 Fontainebleau Cedex, France
Tel: +33 (0)1 60 72 40 00
Fax: +33 (0)1 60 74 55 00/01

Asia Campus
1 Ayer Rajah Avenue, Singapore 138676
Tel: +65 67 99 53 88
Fax: +65 67 99 53 99

Abu Dhabi Campus
Muroor Road - Street No 4
P.O. Box 48049
Abu Dhabi, United Arab Emirates
Tel: +971 2 651 5200
Fax: +971 2 443 9461

www.insead.edu

INSEAD

The Business School
for the World®

Printed by INSEAD