Economic Theory
DOI 10.1007/s00199-006-0173-8

RESEARCH ARTICLE

Pascal Courty · Mario Pagliero

Responsive pricing

Received: 21 February 2006 / Revised: 9 October 2006

Abstract We study the efficiency property of responsive pricing, a scheme that proposes to increase prices as a function of the level of capacity utilization in environments where traditional allocation schemes (e.g. competitive markets, non-linear pricing) cannot be implemented in practice. We show that although responsive pricing implements allocations that are arbitrarily close to full capacity utilization (no wasted capacity and no excess demand), these allocations are not always efficient. We identify conditions under which efficiency occurs and discuss implications for the use of responsive pricing.

Keywords Responsive pricing · Congestion pricing

JEL Classification Numbers D60 · L50 · R48

We would like to thank seminar participants at the LSE, Venezia, Toulouse, and Copenhagen as well as Piero Gottardi, Karel Mertens, Marco Ottaviani, Markus Poschke, Karl Schlag, and Sanne Zwart for useful comments.

P. Courty (✉)
Department of Economics, European University Institute,
Villa San Paolo Via della Piazzuola 43, 50133 Firenze, Italy
E-mail: pcourty@iue.it
Tel.: +39-055-4685925
Fax: +39-055-4685902

M. Pagliero
Department of Economics and Finance and Collegio Carlo Alberto,
University of Turin, 10134 Turin, Italy
E-mail: pagliero@econ.unito.it

1 Introduction

Economists have long recognized the necessity to vary prices to allocate congestible resources efficiently when demand changes over time. In this paper, we investigate the extent to which responsive pricing, a pricing scheme introduced by Vickrey in 1971 that proposes to vary prices in real time as a function of the level of capacity utilization, can increase efficiency when demand changes cannot be anticipated.[1] The class of applications that are relevant include

- Telephone use: This was the original application used by Vickrey to motivate responsive pricing. Vickrey proposed to quote each new user a charge that would vary as a function of the level of network congestion. Other economists have proposed to vary price in real time in electricity markets (Borenstein 2001) and Internet pricing (MacKie-Mason and Varian 1995).[2]
- Road pricing: The San Diego's Regional Planning Agency has used responsive pricing to allocate fast track lanes in highways. Cars that want to use the fast track lanes have to pay a fee that varies in real time as a function of congestion. Consumers face a trade-off between the amount of time they want to save and the fees they are willing to pay (http://argo.sandag.org/fastrak/).
- Ski resorts: Prices could vary in real time to give an incentive to ski less during high demand periods thus reducing lines, and to ski more when demand is low thus achieving a more efficient use of capacity. The same principle could be applied to price access to other sport facilities and theme parks.[3]

Other examples can easily be found. In these applications, traditional allocation schemes, such competitive resale markets, auctions, or even advance screening contracts, would be difficult to implement in practice. Responsive pricing is much simpler. It only requires to measure congestion (i.e. utilization rate) in real time and to be able to communicate congestion-contingent prices to consumers. Responsive pricing proposes to increase access prices as utilization rates increase, that is, as the level of capacity utilization gets closer to congestion.

To understand why prices have to respond to demand shocks, consider what happens under unresponsive pricing. If prices are set according to the expected level of demand at a given time the very nature of the randomness of the arrival process implies that there are times when the number of new arrivals exceeds or falls short of available capacity. If prices do not vary as a function of realized demand, some potential buyers are denied access when there is a sudden arrival of consumers and capacity is wasted when there is a low demand realization.

The set of applications where responsive pricing could be used have the characteristics that although demand variations, due to changes in the number of consumers requesting access, are to some extent impossible to predict, it may be possible

[1] Vickrey's main message was to "call attention to the possibilities that arise if one attempts seriously to promote efficiency through causing prices to fluctuate so as to clear the market [. . .] even in response to those fluctuations that can not be fully predicted in advance."

[2] To illustrate, easyEverything, the largest chain of Internet café in the world, followed Vickrey's proposal and gives discounts that are a function of the number of vacant terminals (http://www.easyeverything.com/; Courty and Pagliero 2003).

[3] To deal with waiting on popular rides, some theme parks sell fast track passes that enables holders to bypass queues (http://www.sixflags.com/parks/wyandotlake/parkinfo/fastlane.asp) while others offer reservation systems which replace waits with virtual lines assigning ride times (http://www.themeparksonline.org/).

to influence the length of time consumers use the service. When this is the case, one can seriously think of using prices to achieve more efficient allocations of the congestible resource between users. The welfare gains from using responsive pricing are potentially great since congestion and/or unused capacity otherwise prevail. For example, lines in ski resorts and unused telephone capacity are common.

There are two basic elements to responsive pricing. First, responsive pricing charges consumers in real time, as consumption takes place. If ω denotes an arrival realization and t time, responsive pricing computes and announces the price for consumption in interval $t + \mathrm{d}t$, $p_t(\omega)$, only at time t. This rules out, for example, advance bookings. Second, the instantaneous price depends on a single state variable: the level of capacity utilization. If capacity utilization is $q_t(\omega)$, then the instantaneous price is set according to $p_t(\omega) = r(q_t(\omega))$ where $r()$ is a given non-decreasing function. Once the function $r()$ is set, consumers play a game of incomplete information. They try to guess future prices to make their consumption decisions. In turn, their consumption decisions determine future prices in equilibrium.

This work is a first step towards understanding the efficiency properties of responsive pricing. We consider a social planner who sets the responsive pricing function $r()$ to maximize social welfare. Can the social planner achieve, or at least approach, the efficient allocation with responsive pricing? Stated formally, does there exist a function $r()$ such that the allocation that results from the game that consumers subsequently play be arbitrarily close to the efficient allocation?

We model the dynamic allocation problem as follows. At every point in time a random arrival flow of consumers requests access. Consumers consume one unit of service per unit of time and value each additional unit less than the previous one. We focus the analysis on the consumers' incentives to terminate consumption. Additional decisions that could also matter include the decisions to request access (endogeneous arrival flow) and to delay consumption, but for tractability concerns, and given the early stage of the research as well, it is sensible to narrow down the problem. The decision to terminate consumption is arguably central to most, if not all, of the congestion problems mentioned earlier while these other issues aren't so.

The analysis proceeds in three steps. We first derive the efficient allocation. Second, we compute the equilibrium under responsive pricing and show that there is no function $r()$ that implements the efficient allocation. Finally, we investigate whether it is possible to construct a sequence of responsive pricing functions $r()$ that approach the efficient allocation.

Our analysis establishes several results. We show that responsive pricing achieves full capacity utilization in the limit – when the price is extremely responsive to changes in the level of capacity utilization. We also show that the limit outcome is efficient under a simple condition on consumer demands, called the no-crossing condition. When this condition holds, equilibrium consumption strategies are very simple. Consumer terminate consumption when their marginal willingness to pay is equal to the instantaneous price. The efficiency result, however, does not generalize to the case where the no-crossing condition does not hold. In fact, we present an example where consumer demands may cross and where no responsive pricing function can approximate the efficient allocation.

This work stresses the distinction between the concepts of full capacity utilization and efficiency. These two concepts are equivalent in the standard textbook model of supply and demand. In our application, they are not always equivalent. Although responsive pricing achieves outcomes that are arbitrarily close to full capacity utilization, these outcomes are not always efficient.

The closest work to our analysis is Vickrey (1971). Vickrey introduced the concept of responsive pricing and speculated that it may achieve efficiency. Vickrey's intuition has been applied, for example, to pricing in electricity markets: Joskow and Tirole (2004) argue that "the case of price-sensitive consumers who react efficiently to real time prices is the textbook representation of consumer demand." Our analysis qualifies this conjecture and shows that efficiency is not always warranted under responsive pricing. Our analysis builds on a concern already identified by Vickrey (in the context of an application to telephone pricing) in his original proposal: " one significant imperfection would remain with such a system: a user upon being informed of the current rate may still be unclear as to whether he should let the call go through at the current rate or defer the call until later, since he has no assurance of what the rate would be at the later time." [4] Our model formalizes Vickrey's conjecture that consumer forward looking behavior may impede efficiency. In addition, we identify a condition under which the efficient outcome is always achieved.

Our work is also related to the large literature that applies non-linear pricing techniques to congestion problems (Wilson 1993). This literarture has considered many pricing schemes (e.g. peak load pricing and priority pricing among others[5]) but we argue that the dynamic nature of the congestion problem we study rules out these solutions for our class of problems, and justifies considering responsive pricing, as recommended by Vickrey. This point will become clear at the end of the next section, after we have presented the model and situated it within the literature.

The paper is organized as follows. The next section presents the model. Section 3 analyses the steady-state version of the model and introduces the main themes of the paper. Section 4 analyzes the dynamic version of the model and presents the main results. Section 5 discusses an important extension. Section 6 concludes.

2 Model

We consider a congestible resource and we denote the resource's capacity Q . We treat Q as exogenously given and we assume that all costs are fixed. The marginal cost of serving an additional consumer is zero up to capacity Q and infinite once capacity is reached.

The aim of the model is to capture a class of applications where consumers have some discretion over the amount they consume which could be measured in units of time (Internet access, telephone) or number of rides (theme park, ski resorts). Formally, we make two assumptions: (a) consumers have decreasing marginal valuation for the service and (b) consumers can terminate the service at any time. These assumptions are realistic in the applications just mentioned.

[4] Vickrey focused on consumers' decision to strategically postpone the start of consumption while our model focuses on the decision to strategically postpone the decision to end consumption. In both cases, the issue is whether a single instantaneous price, computed under responsive pricing, is enough to give efficient consumption incentives.

[5] See the seminal work of Boiteux (1956, 1960), and for a recent review, Crew et al. (1995).

There are I types of consumers. A consumer of type i who has already consumed n units gets utility $v^i(n) > 0$ for the marginal unit where v^i is continuous, differentiable, and $\frac{d}{dn}v^i(n) < 0$. The assumption $v^i(n) > 0$ implies that it is never efficient that a consumer terminates consumption if there is capacity available. We start by assuming that consumers have identical demands ($I = 1$) and then discuss how the argument extends to heterogeneous demands ($I > 1$). Consumers have discount factor $0 < \rho < 1$. To simplify, we assume that consumers are risk neutral.

The arrival process is a vector $\epsilon_t = (\epsilon_t^i)_{i=1\ldots I}$. $\epsilon_t^i(\omega)\mathrm{d}t$ is an integrable continuous stochastic process on some probability space with increments distributed over $\Xi = [\epsilon_l^1, \epsilon_h^1] \times \cdots \times [\epsilon_l^I, \epsilon_h^I]$ such that $0 < \epsilon_l^i < \epsilon_h^i < \infty$.[6] Sample path $\omega \in \Omega$ captures an entire history of arrival realizations $\epsilon_t^i(\omega)$ for $t \geq 0$. $\int_0^t \epsilon_x^i(\omega)\mathrm{d}x$ consumers of type i arrive between 0 and t in sample path ω. In the steady state analysis (Sect. 3) we impose the additional assumption $\epsilon_t(\omega) = \epsilon(\omega)$. In the dynamic analysis (Sect. 4) we do not make any further assumption on $\epsilon_t(\omega)$. There could be a seasonal component (distribution of ϵ_t depends on t) and also a random component that could be correlated over time.

To simplify the exposition, the core of the analysis presented in Sect. 4 focuses on the simplest possible formulation of the problem where consumers only decide when to stop consumption. This assumption rules out the possibility to temporarily delay consumption. Section 5 discusses more general consumption rules.

The level of capacity utilization is denoted by $q_t(\omega)$. We normalize $q_0(\omega) = 0$ without loss of generality. The instantaneous price when the level of capacity utilization is q is $r(q)$ where $r(.)$ is an exogenously given, non-negative, continuous, function with support $[0, Q]$ that is differentiable and increasing on the set $\{x$ s.t. $r(x) > 0\}$. This captures the spirit of Vickrey's proposition that " it seems entirely satisfactory to base rates on levels of activity." Finally, we assume that $r(0) < v^i(0)$ for some $i \leq I$ to warranty that consumption takes place.

Throughout the paper, we use subscript to denote the time when a variable is measured and superscript to denote the time when a consumer arrives. A consumption rule is a set of indicator functions $d_t^{i,s}(\omega)$ defined for $s \leq t$ where $d_t^{i,s}(\omega) = 1$ if the consumer of type i who arrived at time s is consuming at time t and $d_t^{i,s}(\omega) = 0$ otherwise. Consumption rule $d_t^{i,s}(\omega)$ is feasible if it is non-increasing in t (to rule out interruptions). The level of capacity utilization at time t is

$$q_t(\omega) = \int_0^t \sum_i d_t^{i,x}(\omega)\epsilon_x^i(\omega)\mathrm{d}x. \tag{1}$$

Finally, $J_t(\omega) = \{\varepsilon_x(\omega) \in E, x \in [0, t]\}$ denotes the realization of the arrival process up to time t in sample path ω. $J_t(\omega) \in \Omega_t$ where Ω_t represents the set of possible realizations up to t.

Perfect Bayesian equilibrium Consumers play a continuous game of incomplete information. Although we present the game in its full generality, it is important to

[6] The assumption that the increments of $\epsilon_t(\omega)\mathrm{d}t$ are positive and bounded greatly simplifies the derivations because it guarantees that all equilibrium outcomes are bounded and continuous functions of time.

keep in mind that matters will greatly simplify in the cases we consider. In particular, consumers will not be able to strategically use their private information, and as a consequence the optimal consumption strategies will follow simple rules. Consumers are privately informed about their arrival time and about their types but they may not know $J_t(\omega)$. In contrast with standard games of incomplete information, consumers do not observe directly other consumers' actions $d_t^{i,s}(\omega)$. This assumption is realistic for the applications we have in mind. Consumers observe only the realized price. We define $p_t(\omega)$ the equilibrium price at time t in sample path ω. A consumer who arrives at s and has consumed till $t \geq s$ observes price history $H_t^s(\omega) = \{p_x(\omega), x \in [s, t]\} \in \aleph_t^s$ where $\aleph_t^s$ is the set of non-negative functions defined on $[s, t]$. We denote $\mu_t^{i,s}(J_t; \omega, H_t^s)$ the belief held at t by a type i consumer (who arrived at s in sample path ω and has observed information $H_t^s \in \aleph_t^s$) that the arrival history is $J_t \in \Omega_t$. We leave the initial belief $\mu_s^{i,s}(J_s; \omega)$ unspecified beyond the assumption that $\mu_s^{i,s}(J_s(\omega); \omega) > 0$ and we restrict to beliefs that are computed according to Bayes rule where possible:

$$\mu_t^{i,s}(J_t; \omega, H_t^s) = \Pr\left(J_t \mid \mu_s^{i,s}(J_s; \omega), H_t^s\right). \tag{2}$$

Utility maximization implies that $d_t^{i,s}$ maximizes for any $t \geq s$ and for any ω

$$U_t^{i,s}(\omega, H_t^s) = E\left(\int_t^{\infty} \rho^{x-s} d_x^{i,s}(\omega)\left(v^i(x-s) - p_x(\omega)\right) \mathrm{d}x \mid \mu_t^{i,s}\right)$$

subject to feasibility and to the condition that $d_t^{i,s}(\omega)$ depends only on $H_t^s(\omega)$. The equilibrium price at time t is

$$p_t(\omega) = r(q_t(\omega)). \tag{3}$$

We say that equilibrium capacity utilization is implementable if $q_t(\omega) \leq Q$ and we restrict to equilibria that satisfy this constraint. If $q_t(\omega) > Q$ then demand is greater than capacity at time t in sample path ω. In such events, one would have to supplement the pricing rule $r(.)$ with a rule to determine how capacity is rationed. In contrast, the implementability constraint narrows down the analysis to equilibrium allocations that are solely defined by responsive pricing. We acknowledge that modeling rationing under responsive pricing is interesting in itself, but we leave this for further work since the issue can be investigated independently of the question of whether responsive pricing can approximate the efficient allocation.

A perfect Bayesian equilibrium is a pair $\left(d_t^{i,s}(\omega), \mu_t^{i,s}(J_t; \omega, H_t^s)\right)$ such that the consumption strategy profile $d_t^{i,s}(\omega)$ maximizes consumer utility, prices $p_t(\omega)$ are given by pricing rule (3), and the level of capacity utilization is implementable $q_t(\omega) \leq Q$.

Efficient consumption rule The social planner discounts the utility of a consumer who arrives at time s by ρ^s. This implies that all consumption that takes place at time t is discounted by ρ^t. The social planner maximizes

$$W(d_t^{i,s}(\omega)) = E\left(\int_0^\infty \rho^t \int_0^t \sum_i d_t^{i,s}(\omega)v(t-s)\epsilon_s^i(\omega)\mathrm{d}s\mathrm{d}t\right)$$

subject to the constraint that $d_t^{i,s}(\omega)$ depends only on $J_t(\omega)$ and is non-increasing in t, and subject to the implementability constraint.

Literature The problem we consider falls within the literature studying how a social planner should price a congestible resource in the presence of consumer heterogeneity and demand uncertainty. Two solutions have been proposed to deal with such problems. The first and standard approach uses mechanism design tools. Nature picks a random draw of consumers who are privately informed about their demands. The designer takes the distribution of possible demand realizations as given, and uses this knowledge to establish the menu of contracts that maximizes efficiency subject to information constraints. Priority pricing takes a short-term view and proposes to ration capacity by pricing differently different levels of service priority (see Harris and Raviv 1981; Mendelson and Whang 1990 for an application related to our problem). Peak load pricing takes a long-term view and prices separately capacity requirements and actual utilization (Oren et al. 1985). See also chaps. 10 and 11 respectively in Wilson (1993) for a detailed review of these two pricing schemes as well as others.

Our problem does not fall within the standard mechanism design approach for two reasons. First, our problem is intrinsically dynamic. Consumers must be handled as they arrive and congestion has to be managed in real time. The designer cannot offer a grand menu of contracts ex-ante and request all consumers to select a contract as is assumed under a mechanism design approach.[7] Second, the designer and users in our setup may know very little about the demand environment. In particular, the designer does not need to know the arrival process. In contrast, under a mechanism design approach, the planner uses this information to compute the optimal pricing scheme. [8] We focus on a specific indirect mechanism and we analyze the welfare properties of this mechanism for any arrival process. A limitation that follows is that our efficiency results hold only in a limit sense.

The second approach proposes to use spot markets to solve the capacity congestion problem (Wilson 1993, p. 260). One possibility could be to open contingent markets in all possible state of the world.[9] More realistically, several rules have

[7] One way to generalize a mechanism design approach to our problem that we do not pursue in this work, would be to assume that the designer offers a dynamic and sequential menu of contracts where the set of contracts offered at a given time depends on the contracts that have been selected by previous consumers.

[8] An extension of mechanism design would consider the possibility that the demand environment (arrival process and distribution of types) is not common knowledge.

[9] To clarify this point, consider a slightly different version of the model that can be interpreted in terms of inter-temporal general equilibrium theory. To start, assume that the arrival history is public information and assume that one can define state contingent claims for future consumption where states are conditional on the realization of $J_t(\omega)$. If state contingent markets were open for consumption in all future dates, or if consumers could continuously trade in a sufficiently large set of intermediate markets, then one could investigate whether the first welfare theorem would apply.

been proposed in the applied literature to set and update the spot prices. See for example the concept of ‘smart market’ (MacKie-Mason and Varian 1995), applications to electrical network pricing (Bohn et al. 1984), as well as the review of the literature on pricing in telecommunication networks by Falkner et al. (2000). Consistent with this second approach, we focus on an indirect mechanism that proposes a pre-determined rule to compute the spot prices. While the literature has focused on static or stationary environments, we explicitly model strategic consumer behavior in a dynamic framework. Consistent with the mechanism design approach, we model demand at the individual decision maker level and we use perfect Bayesian equilibrium concept. In contrast with this approach, however, we study a dynamic model and we exclude menus of contracts.

Our focus on responsive pricing is motivated by the observation that the solutions that have been proposed in the literature are not realistic for the applications we have in mind. Opening future markets in the absence of those consumers who have not yet requested access would be meaningless, or would require the intervention of intermediaries which again is not realistic, at least in some of the applications considered. Similarly, assuming that a designer knows the distribution of demand types and can offer a grand menu of contracts, as under a mechanism design approach, is unrealistic.

3 Steady-state example

In the simplest version of the model, the arrival rate does not vary over time. This benchmark case introduces the different steps we will again follow to solve the dynamic version of the model, and reveals some basic properties of responsive pricing that can be illustrated graphically. To further simplify the analysis, we also assume homogeneous consumer demand ($I = 1$). We later generalize the argument to heterogenous demands. In terms of our notations, this means that we ignore the time subscript as well as the type superscript. The number of consumers who request access per unit of time is $\epsilon(\omega)\mathrm{d}t$. We refer to $\epsilon(\omega)$ as the state of the world.

To start, we derive the efficient allocation. Let $d_x(\omega) = 1$ if consumers are still consuming x units of time after arriving. The social planner sets $d_x(\omega)$ to maximize expected steady-state surplus:

$$W(d_x(\omega)) = E\int_0^\infty d_x(\omega)v(x)\epsilon(\omega)\mathrm{d}x,$$

subject to the constraint that $d_x(\omega)$ is non-increasing and that the level of capacity utilization is implementable $\int_0^\infty d_x(\omega)\epsilon(\omega)\mathrm{d}x \leq Q$. Let $n(\omega) = \int_0^\infty d_x(\omega)\mathrm{d}x$ represent the number of units consumed in steady state. The efficient consumption rule specifies that consumers should equally share the resource

$$n(\omega) = \frac{Q}{\epsilon(\omega)}.$$

Under that consumption rule no capacity is wasted and it is not possible to reallocate capacity to increase welfare.

Next, we derive the equilibrium under responsive pricing. Consumers observe the steady state price $p(\omega)$ and decide how long to consume. They maximize $\int_0^\infty \rho^x d_x(\omega)\,(v(x) - p(\omega))\,\mathrm{d}x$ where $d_x(\omega) \in \{0, 1\}$ and is non-increasing in x. Consumers consume n units of time such that $v(n) = p(\omega)$. The level of capacity utilization is given by (1), $q(\omega) = n\epsilon(\omega)$, and the price is determined by the pricing function (3), $p(\omega) = r(q(\omega))$. After replacement, equilibrium consumption in state $\epsilon(\omega)$ must satisfy

$$v(n) = r(n\epsilon(\omega)).$$

There exists a unique solution, $n(\omega)$, to the above equation. If $n(\omega)$ is such that $q(\omega) = n(\omega)\epsilon(\omega) \leq Q$ for all ω then the equilibrium is well defined. This will hold if and only if $r(Q) \geq v\left(\frac{Q}{\varepsilon_h}\right)$. Under this condition, consumers demand at most $\frac{Q}{\varepsilon_h}$ and capacity is sufficient to meet demand even for the highest possible arrival rate since $q(\varepsilon_h) \leq \varepsilon_h \frac{Q}{\varepsilon_h} = Q$. If this condition does not hold, then the demand in state ε_h is higher than capacity, and the equilibrium is not well defined. Note that consumers' initial beliefs about the state do not play a role because once consumers have observed the price they automatically know the true state.

Higher arrival rates imply that consumers consume less ($\mathrm{d}n/\mathrm{d}\epsilon < 0$), the level of capacity utilization is higher ($\mathrm{d}q/\mathrm{d}\epsilon > 0$), and the price is higher ($\mathrm{d}p/\mathrm{d}\epsilon > 0$). Figure 1 illustrates these properties.

To simplify, the figure assumes that the arrival rate is either high or low. The equilibrium level of capacity utilization is located at the point where the aggregate inverse demand ($v(q/\epsilon)$) and the pricing curve intersect. The realized price is higher in the high state when capacity is scarcer, and consumers respond by sharing the capacity available more (lower n).

To understand what is specific to responsive pricing, we contrast the outcome under responsive pricing with the outcome under fixed pricing. Under fixed price ($r(q) = r$) consumers consume n units such that $v(n) = r$. Length of use does not

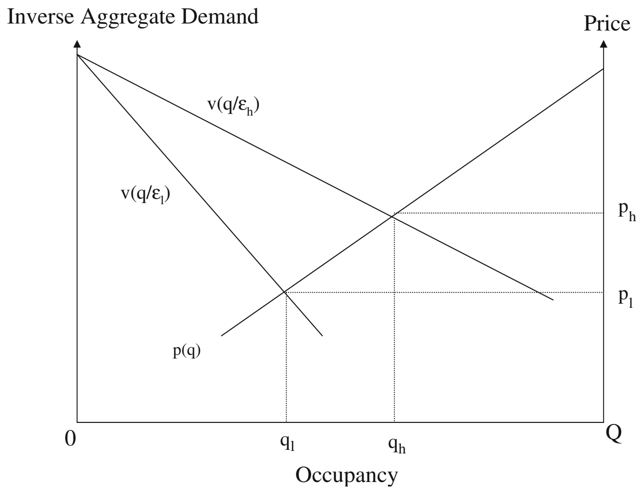

Fig. 1 The 2-states steady state case

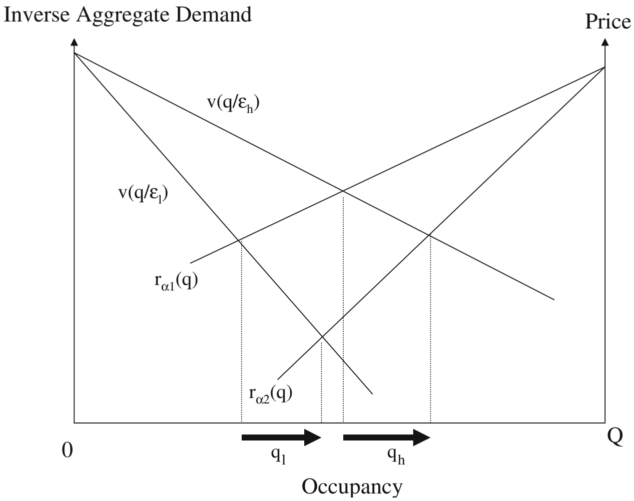

Fig. 2 Increase in responsiveness ($\alpha_2 < \alpha_1$)

depend on the state of the world, $\epsilon(\omega)$, because consumers do not have any incentive to vary consumption as a function of congestion.

To conclude, we investigate the efficiency properties of responsive pricing. To start, note that there does not exist a function $r(.)$ that implements the efficient allocation if there are two states with different arrival rates $\epsilon' > \epsilon''$. The only prices that decentralize the efficient allocation are $p' = v\left(\frac{Q}{\varepsilon'}\right)$ and $p'' = v\left(\frac{Q}{\varepsilon''}\right)$ but the efficient allocation is such that $q'' = q' = Q$. It is not possible to set r such that $r(Q) = v\left(\frac{Q}{\varepsilon'}\right) = v\left(\frac{Q}{\varepsilon''}\right)$.

Next, we show that responsive pricing can implement the efficient outcome in a limit sense. Consider the class of pricing functions $\widetilde{r}_\alpha$ such that $\widetilde{r}_\alpha(q) = 0$ for $q \leq Q - \alpha$ and $\widetilde{r}_\alpha(q) = v(Q/\epsilon_h)(1 - \frac{Q-q}{\alpha})$ otherwise. These functions are equal to zero up to $Q - \alpha$ and then linear with $\widetilde{r}_\alpha(Q) = v(Q/\epsilon_h)$. Since $\widetilde{r}_\alpha(Q) \geq v\left(\frac{Q}{\varepsilon_h}\right)$ the equilibrium is always well defined. The equilibrium level of capacity utilization is given by $\widetilde{r}_\alpha(q(\omega)) = v(q(\omega)/\epsilon(\omega)) > 0$. An upper bound for unused capacity is

$$Q - q(\omega) < \alpha.$$

More responsive schemes (lower α) increase capacity utilization and therefore efficiency (see Fig. 2).

Capacity utilization converges to full utilization as α converges to zero. This limit case corresponds to the consumption rule that maximizes social welfare.[10]

[10] Although there are several ways to define the limit of pricing scheme $\widetilde{r}_\alpha$, the limit does not implement the efficient allocation independently of the concept used. One can define the limit as a correspondence such that $\widetilde{r}(Q) \in \left[0, v\left(\frac{Q}{\varepsilon_h}\right)\right]$. This pricing scheme, however, has little practical interest because it does not identify a unique price when occupancy reaches capacity. Another way to define the limit is $\widetilde{r}(q) = 0$ for $q < Q$ and $\widetilde{r}(Q) = v(Q/\epsilon_h)$. There is no equilibrium for this pricing rule.

The analysis generalizes to the case of heterogeneous consumers. Denote $n^i(p, \omega)$ the number of units consumed by type i when price is p , $v^i(n^i(p, \omega)) = p$. The equilibrium price in state $\epsilon(\omega)$ is uniquely defined by

$$p(\omega) = r\left(\sum_i \epsilon^i(\omega) n^i\left(p(\omega), \omega\right)\right)$$

and the equilibrium level of capacity utilization is given by $q(\omega) = \sum_i \epsilon^i(\omega) n^i(p(\omega), \omega)$. The analysis of efficiency carries through.

The analysis of the steady state version of the model shows that responsive pricing endogenously sets prices in response to demand realizations and implements an outcome that both achieves full capacity utilization and is efficient in the limit. In this version of the model, prices do not vary over time and consumers face a simple decision problem. When the arrival rate changes over time, however, prices continuously change and consumers face a more complex decision problem because they have to anticipate future prices to decide whether to retain access or quit. The rest of this paper generalizes the analysis to non-stationary arrival processes and asks whether the results on efficiency carry through. As we will see, the efficiency analysis carries through for homogeneous consumer demands but not always for heterogeneous demands.

4 Dynamic analysis

We start by focusing on the case where consumers have identical demands ($I = 1$). The analysis mirrors the steady state presentation. We first characterize the first-best consumption rule and then the perfect Bayesian equilibrium. Then, we investigate whether responsive pricing can approach the efficient allocation. We conclude by considering the case of heterogeneous demands ($I > 1$).

4.1 Efficient consumption rule

We reintroduce the time subscript but we ignore type superscript since we assume in this subsection that consumers are homogeneous. Define $\widehat{t}(\omega)$ as the first point in time when capacity is reached if consumers do not terminate consumption $\int_0^{\widehat{t}(\omega)} \epsilon_x(\omega)\mathrm{d}x = Q$ and $\widehat{b}_t(\omega)$ as the solution to

$$\widehat{b}_t(\omega) = \begin{cases} y & \text{such that } \int_y^t \epsilon_x(\omega)\mathrm{d}x = Q \text{ if } t > \widehat{t}(\omega) \\ 0 & \text{if } \quad t \leq \widehat{t}(\omega) \end{cases},$$

$\widehat{b}_t(\omega)$ is increasing in t. It corresponds to the 'oldest' consumer (where consumer a is 'older' than consumer b if a has arrived before b) who can consume at time t if all consumers who have arrived after that consumer are also consuming and capacity utilization is implementable.

Proposition 1 *The efficient consumption rule is*

$$\widehat{d}_t^s(\omega) = \begin{cases} 1 & \text{if } \widehat{b}_t(\omega) \leq s \leq t \\ 0 & \text{if } s < \widehat{b}_t(\omega) \end{cases}.$$

Proof $\widehat{d_t^s}(\omega)$ is feasible and implementable by construction. The proof that $\widehat{d_t^s}(\omega)$ is the only consumption rule that achieves the efficient outcome goes by contradiction. Assume that there exist an alternative consumption rule $\widetilde{d_t^s}(\omega)$ different from $\widehat{d_t^s}(\omega)$ such that $W(\widetilde{d_t^s}(\omega)) \geq W(\widehat{d_t^s}(\omega))$.

Claim There does not exists a sample path ω and a t_0 such that

$$(S) \begin{cases} \int_0^{t_0} \widetilde{d_{t_0}^s}(\omega) v(t_0 - s)\epsilon_s(\omega)\mathrm{d}s > \int_0^{t_0} \widehat{d_{t_0}^s}(\omega) v(t_0 - s)\epsilon_s(\omega)\mathrm{d}s \\ \int_0^{t_0} \widetilde{d_{t_0}^s}(\omega)\epsilon_s(\omega)\mathrm{d}s \leq Q \end{cases}.$$

Since $\int_{\widehat{b_{t_0}}(\omega)}^{t_0} \epsilon_s(\omega)\mathrm{d}s = Q$, the capacity constraint condition (second inequality in S) implies that

$$\int_0^{t_0} \widetilde{d_{t_0}^s}(\omega)\epsilon_s(\omega)\mathrm{d}s \leq \int_{\widehat{b_{t_0}}(\omega)}^{t_0} \epsilon_s(\omega)\mathrm{d}s$$

$$\int_0^{\widehat{b_{t_0}}(\omega)} \widetilde{d_{t_0}^s}(\omega)\epsilon_s(\omega)\mathrm{d}s \leq \int_{\widehat{b_{t_0}}(\omega)}^{t_0} \left(1 - \widetilde{d_{t_0}^s}(\omega)\right)\epsilon_s(\omega)\mathrm{d}s$$

$$v(t_0 - \widehat{b_{t_0}}(\omega)) \int_0^{\widehat{b_{t_0}}(\omega)} \widetilde{d_{t_0}^s}(\omega)\epsilon_s(\omega)\mathrm{d}s \leq v(t_0 - \widehat{b_{t_0}}(\omega)) \int_{\widehat{b_{t_0}}(\omega)}^{t_0} \left(1 - \widetilde{d_{t_0}^s}(\omega)\right)\epsilon_s(\omega)\mathrm{d}s$$

$$\int_0^{\widehat{b_{t_0}}(\omega)} \widetilde{d_{t_0}^s}(\omega) v(t_0 - s)\epsilon_s(\omega)\mathrm{d}s \leq \int_{\widehat{b_{t_0}}(\omega)}^{t_0} \left(1 - \widetilde{d_{t_0}^s}(\omega)\right) v(t_0 - s)\epsilon_s(\omega)\mathrm{d}s$$

$$\int_0^{t_0} \widetilde{d_{t_0}^s}(\omega) v(t_0 - s)\epsilon_s(\omega)\mathrm{d}s \leq \int_{\widehat{b_{t_0}}(\omega)}^{t_0} v(t_0 - s)\epsilon_s(\omega)\mathrm{d}s. \quad (4)$$

A contradiction with S's first inequality.

The above claim rules out the possibility that $W(\widetilde{d_t^s}(\omega)) > W(\widehat{d_t^s}(\omega)$. The only possibility is $W(\widetilde{d_t^s}(\omega)) = W(\widehat{d_t^s}(\omega))$ but this implies that $\int_0^{t_0} \widetilde{d_{t_0}^s}(\omega) v(t_0 - s)\epsilon_s(\omega)\mathrm{d}s = \int_0^{t_0} \widehat{d_{t_0}^s}(\omega) v(t_0 - s)\epsilon_s(\omega)\mathrm{d}s$ for any sample path ω and t_0. Therefore, $\widetilde{d_t^s}(\omega) = \widehat{d_t^s}(\omega)$. A contradiction. □

Efficiency occurs if all consumers who arrive up to $\widehat{t}(\omega)$ consume and for $t > \widehat{t}(\omega)$ only those consumers who arrive between $\widehat{b_t}(\omega)$ and t consume. The intuition for the first best consumption rule in the case of homogeneous demands is simple. Once full capacity utilization is reached, it is efficient to share the capacity so that for every new consumer who arrives, the consumer who has been using the service the longest terminates consumption. Under that allocation, new consumers replace older consumers, who value the service less. Define $\widehat{p}_t(\omega) = v(t - \widehat{b}_t(\omega))$ as the valuation of the consumer who has been using the service for the longest length of time at time t. $\widehat{p}_t(\omega)$ is the marginal social value of capacity for $t \geq \widehat{t}(\omega)$ (the marginal social value of capacity is 0 for $t < \widehat{t}(\omega)$). As we will soon see, $\widehat{p}_t(\omega)$ corresponds to the capacity clearing price, and in the case where consumer demands satisfy the no-crossing condition it also corresponds to the efficient spot price. The issue we investigate next is whether responsive pricing can approximate that price.

4.2 Perfect Bayesian equilibrium

We show that in any equilibrium consumers terminate consumption as soon as their willingness to pay for a unit of consumption falls below the price.

Lemma 1 *In any equilibrium,* $d_t^s(\omega) = 1$ *if and only if* $v(\widetilde{t} - s) \geq p_{\widetilde{t}}(\omega)$ *for* $\widetilde{t} \in [s, t]$.

Proof The 'if' part is obvious. The proof of the 'only if' part goes by contradiction. Assume there exists a pair $s < t$ and a sample path ω such that s receives negative instantaneous net utility at time t, that is, $d_t^s(\omega) = 1$ and $v(t - s) < p_t(\omega)$. Let s_0 denote the consumer that first experiences negative instantaneous net utility ($\exists t$ s.t. $v(t - s_0) < p_t(\omega)$ and $d_t^{s_0}(\omega) = 1$ and $\nexists(\widetilde{s}, \widetilde{t})$ s.t. $\widetilde{t} < t$, $v(\widetilde{t} - \widetilde{s}) < p_{\widetilde{t}}(\omega)$, and $d_{\widetilde{t}}^{\widetilde{s}}(\omega) = 1$).

Claim 1 There exist $\infty \geq t_1 > t_0 > s_0$ and ω such that

$$\begin{cases} v(t_0 - s_0) = p_{t_0}(\omega) & \\ v(t - s_0) \geq p_t(\omega) & \text{for} \quad t \in [s_0, t_0) \\ v(t - s_0) < p_t(\omega) & \text{for} \quad t \in (t_0, t_1) \end{cases} .$$

We only need to show that there exists $t_0 > s_0$ such that the top two conditions hold since the existence of t_1 then follows from the definition of s_0. Assume that there does not exist a $t_0 > s_0$ such that the top two conditions hold. This implies two claims (a) $v(0) \leq p_{s_0}(\omega)$ and (b) $d_s^{s_0}(\omega) = 0$ for $s < s_0$. Claim (b) follows by contradiction. If (b) does not hold, then there exists a consumer who arrived before s_0 and receives negative instantaneous net utility at s_0; a contradiction with the definition of s_0. Claim (b) implies that s_0 is the only consumer consuming at time s_0. The price is $p_{s_0}(\omega) = r(0)$. A contradiction with claim (a) since $r(0) < v(0)$.

Claim 2 $d_{t_0}^s(\omega) = 0$ for $s < s_0$ and $d_{t_0}^s(\omega) = 1$ for $s_0 < s < t_0$.
For $s < s_0$, $v(t_0 - s) < v(t_0 - s_0) = p_{t_0}(\omega)$. Since s_0 is by definition the first consumer who experience negative instantaneous net utility, we must have $d_{t_0}^s(\omega) = 0$. For $s_0 < s < t_0$, $v(t - s) \geq p_t(\omega)$ for $t \in [s, t_0]$ (Claim 1). Consumer s should keep consuming until t_0, that is, $d_{t_0}^s(\omega) = 1$.

Claim 3 For any $t > t_0$, $v(t - s_0) - p_t(\omega) < 0$.
We distinguish two cases. If no consumer has stopped consumption in $[s_0, t]$, that is, $d_t^s(\omega) = 1$ for $s \in [s_0, t]$, then $v(t - s_0) < v(t_0 - s_0) = p_{t_0}(\omega) < p_t(\omega)$ and $v(t - s_0) - p_t(\omega) < 0$. If not, denote $\widetilde{s}$ the last consumer who has stopped

consumption since t, and denote $\widetilde{t}$ the time when $\widetilde{s}$ has stopped consumption. We have $v(t-s_0) < v(t-\widetilde{s}) < v(\widetilde{t}-\widetilde{s}) \leq p_{\widetilde{t}}(\omega) < p_t(\omega)$ where the first inequality holds because $\widetilde{s}$ has arrived after s_0 (Claim 2), the second inequality holds because $t < \widetilde{t}$, and the last inequality holds because no consumer has left between t and $\widetilde{t}$. Again, we have $v(t-s_0) - p_t(\omega) < 0$.
Claim 3 implies that

$$U_{t_0}^{s_0}(\omega, H_{t_0}^{s_0}(\omega)) = E\left(\int_{t_0}^{\infty} \rho^{x-s_0} d_x^{s_0}(\omega)\left(v(x-s_0) - p_x(\omega)\right) \mathrm{d}x \mid \mu_{t_0}^{s_0}\right) < 0$$

Consumer s_0 is better off setting $d_t^{s_0}(\omega) = 0$ for $t > t_0$ in history $H_{t_0}^{s_0}(\omega)$. A contradiction. □

Lemma 1 implies that consumers leave in a first-in first-out fashion in any equilibrium. Formally, $d_t^s(\omega)$ is non-decreasing in s. The reason is simply that consumer s consumes at time t only if $v(\widetilde{t}-s) - p_{\widetilde{t}}(\omega) > 0$ for $\widetilde{t} \in [s, t]$. But this implies that any consumer who arrived after s should also consume since $v(\widetilde{t}-\widetilde{s}) - p_{\widetilde{t}}(\omega) > 0$ for $\widetilde{t} \in [\widetilde{s}, t]$ if $\widetilde{s} > s$. The 'oldest' consumer consuming at time t arrived at $\mathrm{Inf}\left\{s \geq 0,\ \text{s.t. } d_t^s(\omega) = 1\right\}$.[11] Lemma 1 implies that the level of capacity utilization at time t is equal to the mass of consumers who have arrived after the oldest consumer,

$$q_t(\omega) = \int_{\mathrm{Inf}\left\{s,\ \text{s.t. } d_t^s(\omega)=1\right\}}^{t} \epsilon_s(\omega)\mathrm{d}s.$$

The equilibrium does not exist if $q_t(\omega) > Q$. Next we identify the minimum condition that the pricing rule must satisfy to assure that the equilibrium always exists.

Lemma 2 *$q_t(\omega) \leq Q$ for any arrival process if and only if $r(Q) \geq v\left(\frac{Q}{\varepsilon_h}\right)$.*

Proof To start we show that $r(Q) \geq v\left(\frac{Q}{\varepsilon_h}\right)$ is a necessary condition. The proof goes by contradiction. Assume $r(Q) < v\left(\frac{Q}{\varepsilon_h}\right)$ and consider the arrival process $\epsilon_t(\omega) = \epsilon_h$. Consumers consume at least $v^{-1}\left(r(Q)\right) > \frac{Q}{\varepsilon_h}$. The equilibrium level of capacity utilization is at least $\varepsilon_h v^{-1}\left(r(Q)\right) > Q$. A contradiction.

Next, we show that $r(Q) \geq v\left(\frac{Q}{\varepsilon_h}\right)$ is a sufficient condition. The proof again goes by contradiction. Assume there exist ω and t_0 such that $q_{t_0}(\omega) > Q$. Let $s_0 = \mathrm{Inf}\left\{s \text{ s.t. } d_{t_0}^s(\omega) = 1\right\}$. The level of capacity utilization at time t_0 can be expressed as $q_{t_0}(\omega) = \int_{s_0}^{t_0} \epsilon_s(\omega)\mathrm{d}s \leq (t_0 - s_0)\epsilon_h$. This implies that $t_0 - s_0 > \frac{Q}{\varepsilon_h}$. The consumer who arrived at s_0 gets negative instantaneous utility at t_0 since $v(t_0 - s_0) < v\left(\frac{Q}{\varepsilon_h}\right) \leq r(Q)$. A contradiction with Lemma 1. □

[11] We assume without loss of generality that the $\mathrm{Inf}\left\{s,\ \text{s.t. } d_t^s(\omega) = 1\right\}$ exists.

In the rest of this section, we focus on pricing functions that satisfy $r(Q) \geq v\left(\frac{Q}{\varepsilon_h}\right)$. The functions $t(\omega)$ and $b_t(\omega)$ are introduced to characterize the equilibrium consumption strategy profile. Define $t(\omega)$ such that $v(t(\omega)) = r\left(\int_0^{t(\omega)} \epsilon_s(\omega)\mathrm{d}s\right)$ and define the function $b_t(\omega)$ such that

$$b_t(\omega) = \begin{cases} x & \text{such that } v(t-x) = r\left(\int_x^t \epsilon_s(\omega)\mathrm{d}s\right) \text{ if } t > t(\omega) \\ 0 & \text{if } \quad t \leq t(\omega) \end{cases}. \tag{5}$$

By the implicit function theorem, the identity $v(t - b_t(\omega)) = r\left(\int_{b_t(\omega)}^t \epsilon_s(\omega)\mathrm{d}s\right)$ defines a continuously differentiable function for $t{>}t(\omega)$. In addition $b_t(\omega)$ is increasing since

$$\frac{\mathrm{d}}{\mathrm{d}t} b_t(\omega) = \frac{r'\left(\int_{b_t(\omega)}^t \epsilon_s(\omega)\mathrm{d}s\right)\epsilon_t(\omega) - v'(t-b_t(\omega))}{r'\left(\int_{b_t(\omega)}^t \epsilon_s(\omega)\mathrm{d}s\right)\epsilon_{b_t(\omega)}(\omega) - v'(t-b_t(\omega))} > 0.$$

The next proposition characterizes the equilibrium.

Proposition 2 *In any perfect Bayesian equilibrium, the consumption strategy profile is*

$$d_t^s(\omega) = \begin{cases} 1 & \text{if } \; v(t-s) \geq p_t(\omega) \\ 0 & \text{if } \; v(t-s) < p_t(\omega) \end{cases}$$

where $p_t(\omega) = v(t - b_t(\omega))$.

Proof We first show that $d_t^s(\omega)$ is an equilibrium. The level of capacity utilization implied by the consumption strategy profile is $q_t(\omega) = \int_{b_t(\omega)}^t \epsilon_s(\omega)\mathrm{d}s$. Lemma 2 implies that $q_t(\omega)$ is implementable. The equilibrium price satisfies (3) since $p_t(\omega) = v(t - b_t(\omega)) = r(q_t(\omega))$. The consumption strategy profile is optimal since any consumer $s \in [b_t(\omega), t]$ weakly prefers to consume ($v(t-s) \geq v(t - b_t(\omega)) = p_t(\omega)$) and any consumer $s \in [0, b_t(\omega)]$ weakly prefers not to consume ($v(t-s) \leq v(t - b_t(\omega)) = p_t(\omega)$).

Next, we show that $d_t^s(\omega)$ is the unique equilibrium consumption strategy profile. Consider an alternative equilibrium with consumption strategy profile $\widetilde{d}_t^s(\omega)$ and let $\widetilde{p}_t(\omega)$ be the associated price. Define $\widetilde{b}_t(\omega) = \mathrm{Inf}\left\{s, \text{ s.t. } \widetilde{d}_t^s(\omega) = 1\right\}$.

Case a $\widetilde{b}_t(\omega) < b_t(\omega)$. But Lemma 1 implies that $d_t^s(\omega)$ is non-decreasing in s. Therefore, $\widetilde{p}_t(\omega) > p_t(\omega)$, and

$$v(t - \widetilde{b}_t(\omega)) - \widetilde{p}_t(\omega) < v(t - b_t(\omega)) - p_t(\omega) = 0.$$

A contradiction with Lemma 1.

Case b $\widetilde{b}_t(\omega) > b_t(\omega)$ then $\widetilde{p}_t(\omega) < p_t(\omega)$, and

$$v(t - \widetilde{b}_t(\omega)) - \widetilde{p}_t(\omega) > v(t - b_t(\omega)) - p_t(\omega) = 0$$

for $t \in [\widetilde{b}_t(\omega), t]$. The consumer who arrived at $\widetilde{b}_t(\omega) - \eta$, where η is a small positive number, should not have terminated consumption. A contradiction. □

For any $t > t(\omega)$, the price is equal to the marginal valuation of the consumer who arrived at $b_t(\omega)$. This consumer, call it consumer $b_t(\omega)$, is the oldest consumer consuming at time t and is indifferent between continuing and terminating consumption. The equilibrium dynamic consumption strategy profile simplifies to a simple rule specifying that consumers terminate consumption as soon as their instantaneous utility falls below the instantaneous price. Equilibrium strategies are independent of consumers' initial belief $\mu_s^{i,s}(\omega)$. One could generalize the setup by assuming that some consumers receive signals about the arrival process and show that no consumer can benefit from this information although this information could help to predict future prices more accurately.

One may argue that consumers should keep consuming, even if they get negative instantaneous utility, if they expect that prices will decline fast enough so that expected future surpluses eventually outweigh short-term losses. This, however, cannot happen in equilibrium. A consumer may initially believe that she has arrived in a sample path where prices are likely to decrease. But as her net instantaneous utility gets close to zero, that consumer's beliefs have to adjust. In any deviation, a consumer cannot believe that expected future utility could be non negative if net instantaneous utility is negative.

4.3 Pricing responsiveness, capacity utilization, and efficiency

As in Sect. 3, no responsive pricing can implement the efficient allocation. We show, however, that efficiency can be achieved in a limit sense. Let $\{r_\beta(q),\ \beta > 0\}$ be a class of pricing functions indexed by parameter β. Many classes of pricing schemes implement the efficient outcome in the limit. Since our goal is to show only that this is possible, we focus on a very simple subset of such classes. We say that scheme r is α-responsive if $\text{Max}\,\{q \text{ s.t. } r(q) = 0\} \geq Q - \alpha$. For example, scheme $\widetilde{r}_\alpha(q)$ defined earlier is α-responsive.

Consider a class of α-responsive schemes. We ask whether the equilibrium consumption strategy profile under scheme $r_\alpha(q)$ converges to the efficient consumption rule as α converges to 0. We use the notation $q_t(\omega;\alpha)$ to define the equilibrium level of capacity utilization for scheme α and we use the same notations for other equilibrium variables.

Proposition 3 *As α converges to* 0*, $t(\omega;\alpha)$ converges to $\widehat{t}(\omega)$ and $q_t(\omega;\alpha)$ converges to Q for $t > \widehat{t}(\omega)$.*

Proof $t(\omega;\alpha)$ is defined by $v(t(\omega;\alpha)) = r_\alpha\left(\int_0^{t(\omega;\alpha)} \epsilon_s(\omega)\mathrm{d}s\right)$. Since $v(.) > 0$, $Q \geq \int_0^{t(\omega;\alpha)} \epsilon_s(\omega)\mathrm{d}s > Q - \alpha$, and $t(\omega;\alpha)$ converges to $\widehat{t}(\omega)$ as α converges to 0. For $t > \widehat{t}(\omega)$, $v(t - b_t(\omega;\alpha)) = r_t(q_t(\omega;\alpha)) > 0$ and this implies that $q_t(\omega;\alpha) > Q - \alpha$. The claim follows from the observation that $q_t(\omega;\alpha) \leq Q$ in any equilibrium. □

This proposition says that responsive pricing achieves full capacity utilization in the limit. Next, we show that efficiency is achieved in a limit sense.

Proposition 4 *As α converges to* 0*, $b_t(\omega;\alpha)$ converges to $\widehat{b_t}(\omega)$.*

Proof For $t > \widehat{t}(\omega;\alpha)$, $\int_{b_t(\omega;\alpha)}^t \epsilon_s(\omega)\mathrm{d}s > Q - \alpha$. Subtracting $\int_{\widehat{b_t}(\omega)}^t \epsilon_x(\omega)\mathrm{d}x = Q$ on each side gives

$$\alpha > \int_{\widehat{b}_t(\omega)}^{b_t(\omega;\alpha)} \epsilon_s(\omega)\mathrm{d}s$$

$$\alpha > \left(b_t(\omega;\alpha) - \widehat{b}_t(\omega)\right)\epsilon_l$$

In addition, $\int_{\widehat{b}_t(\omega)}^{t} \epsilon_x(\omega)\mathrm{d}x = Q \geq q_t(\omega;\alpha) = \int_{b_t(\omega;\alpha)}^{t} \epsilon_s(\omega)\mathrm{d}s$, which implies $b_t(\omega;\alpha) - \widehat{b}_t(\omega) \geq 0$. Therefore $\alpha > b_t(\omega;\alpha) - \widehat{b}_t(\omega) \geq 0$ and $b_t(\omega;\alpha)$ converges to $\widehat{b}_t(\omega)$. □

In the limit, there is no wasted capacity and responsive pricing approaches the efficient outcome. The price at date t converges to $\widehat{p}_t(\omega) = v(t - \widehat{b}_t(\omega))$ which corresponds to the marginal social value of capacity under the efficient outcome.

The result on efficiency holds for a general class of arrival processes, since we have not made any assumption on ϵ_t besides the support of the increments. Our results apply equally for arrival processes with unexpected demand shocks and for processes with predictable demand shocks. Stated differently, we have shown that responsive pricing could approach the efficient allocation when consumers were only privately informed about their arrival times.

Although responsive pricing approaches the first best allocation, a 'first-in first-out' (FIFO) mechanism that replaces new users with the oldest users implements the first best outcome. All the information the designer needs to implement this alternative mechanism is to be able to keep track of users' arrival times. This equivalence between responsive pricing and FIFO, however, does not hold anymore when consumers are privately informed about their demand types. We show next that for a large class of consumer demands, FIFO does not implement the efficient allocation while responsive pricing does.

4.4 Heterogeneous demands

We turn to the full version of the model. We show that the results presented in the previous section generalize to the case of heterogeneous demands under a 'no-crossing' condition on consumer demands. This condition is important because we show that when it does not hold, inefficiencies can occur.

4.4.1 No-crossing residual demands

We introduce the type superscript to capture heterogeneous demands. We say that the set of demands $\{v^i(.)\}_{i=1\ldots I}$ satisfies the no-crossing condition if for any pair of types $(i, \widetilde{i})$ there do not exist $n, n', \delta \geq 0$ such that

$$v^{\widetilde{i}}(\delta + n) > v^i(n) \quad \text{and} \quad v^{\widetilde{i}}(\delta + n') < v^i(n').$$ [12]

The no-crossing condition has a clear economic interpretation. Define the residual demand of a consumer who has already used the service for some time as the

[12] An example of a class of demands that satisfies the no-crossing condition is the class $v^i(n) = a^i - bn$ where a^i and b are positive numbers.

consumer's willingness to pay for future units. The no-crossing condition says that no two consumers who arrive at different points in time can have residual demands that cross. This condition imposes a fairly strong restriction on the set of demands v^i. In fact, we will see that it is equivalent to say that demands are horizontal shift of one another.

The efficiency analysis generalizes when the v^i satisfy the no-crossing condition. To show that, assume without loss of generality that $v^1(0) \geq v^2(0) \geq \cdots \geq v^I(0)$ and define a^i such that $v^i(a^i) = v^{i+1}(0)$ and $A^i = a^1 + \cdots + a^i$ with $A^0 = 0$. For $\delta \geq 0$, define the function $\tau(\delta)$ as the highest type who values the first unit at least as much as $v^1(\delta)$, $\tau(\delta) = \text{Max}\left\{i \text{ such that } v^i(0) \geq v^1(\delta)\right\}$. Define $q_t(s, \omega)$ as the mass of consumers who have arrived before t and value the service more than $v^1(t-s)$, $q_t(s, \omega) = \sum_{j=1}^{\tau(t-s)} \int_{s+A^{j-1}}^{t} \epsilon_x^j(\omega)\mathrm{d}x$. To characterize the efficient consumption rule, we define the pair of functions $\widehat{t}(\omega)$ and $\widehat{b}_t^1(\omega)$ such that $q_{\widehat{t}(\omega)}(0, \omega) = Q$ and $q_t(\widehat{b}_t^1(\omega), \omega) = Q$ for $t > \widehat{t}(\omega)$.

Proposition 5 *The efficient consumption rule is*

$$\widehat{d}_t^{i,s}(\omega) = \begin{cases} 1 & \text{if} \quad \widehat{b}_t^1(\omega) + A^{i-1} < s \leq t \\ 0 & \text{if} \quad s < \widehat{b}_t^1(\omega) + A^{i-1} \textit{ or if } \widehat{b}_t^1(\omega) + A^{i-1} > t \end{cases}.$$

Proof Before proceeding, we need to establish a preliminary result. The no-crossing condition implies that $v^i(n) = v^1(A^{i-1} + n)$ for $i = 1 \ldots I$. The proof goes by contradiction. Assume that there exist (i, n) such that $i \neq 1$ and $v^i(n) \neq v^1(A^{i-1}+n)$. Assume for example that $v^i(n) > v^1(A^{i-1}+n)$. (The proof is similar if the inequality if reversed.) Then, by continuity $v^i(n) > v^1(A^{i-1}+n-\epsilon)$ for ϵ small. But $v^1(A^{i-1}) = v^i(0)$ implies that $v^1(A^{i-1} - \epsilon) > v^i(0)$. These two inequalities contradict the assumption that v^1 and v^i satisfy the no-crossing condition.

The rest of the proof follows the steps of the proof of Proposition 1. The proof of the claim that there does not exists a sample path ω and a t_0 such that

$$(S)\begin{cases} \int_0^{t_0} \sum\limits_i \widetilde{d}_{t_0}^{i,s}(\omega)v(t_0 - s)\epsilon_s^i(\omega)\mathrm{d}s > \int_0^{t_0} \sum\limits_i \widehat{d}_{t_0}^{i,s}(\omega)v(t_0 - s)\epsilon_s^i(\omega)\mathrm{d}s \\ \int_0^{t_0} \widetilde{d}_{t_0}^s(\omega)\epsilon_s(\omega)\mathrm{d}s \leq Q \end{cases}$$

is established by multiplying the equivalent of (4) by $v^1(t - \widehat{b}_t^1(\omega))$. No consumer values consumption at time t less than $v^1(t - \widehat{b}_t^1(\omega))$ since a consumer of type i who is still consuming in t had to arrive at $\widehat{b}_t^1(\omega) + A^{i-1}$ or after and the lowest valuation among those type i consumers is $v^i(t-(\widehat{b}_t^1(\omega)+A^{i-1})) = v^1(t-\widehat{b}_t^1(\omega))$. □

Under the no-crossing condition, the efficient consumption rule changes slightly. For any $t \geq \widehat{t}(\omega)$, the consumers with the lowest demands are replaced by new consumers, starting with those consumers with highest demands up to the point where no new consumer values consumption more than the marginal consumer. As a result, no consumer terminating consumption ever values consumption more than any consumer retaining consumption.

Similarly, the derivation of the perfect Bayesian equilibrium still holds after straightforward generalizations. Proposition 2 characterizing the equilibrium must take into account the fact that the rule defining the oldest consumer of type 1 consuming at time t, call it $b_t^1(\omega)$, will determine the oldest consumer of type $i \neq 1$ consuming at time t according to,

$$b_t^i(\omega) = \text{Min}(t, b_t^1(\omega) + A^{i-1}).$$

Although consumers of a same type terminate consumption in a FIFO fashion, consumers of different types may not do so. For example, a consumer of type $i \neq 1$ who arrived at t will terminate consumption before a consumer of type $i-1$ who arrived between $t - a_{i-1}$ and t. In the perfect Bayesian equilibrium, the oldest consumer of type one is defined by

$$v^1(t - b_t^1(\omega)) = r\left(q_t(b_t^1(\omega), \omega)\right).$$

The equilibrium price is $p_t(\omega) = r\left(q_t(b_t^1(\omega), \omega)\right)$ and the equilibrium level of capacity utilization is $q_t(\omega) = q_t(b_t^1(\omega), \omega)$. Lemma 2 extends to heterogeneous demands under the condition that $r(Q) \geq \overline{r}$ where $\overline{r}$ is the lowest level of price that rules out excess demand.[13] Propositions 3 and 4 extend, and the equilibrium responsive price at time t converges to $v^1(t - \widehat{b}_t^1(\omega))$ as α converges to 0, so that efficiency can be achieved in a limit sense.

The extension to no-crossing demand is important for the following reason. As mentioned in the previous subsection, a social planner can implement the first best with a FIFO mechanism in the homogeneous demand case if it is possible to keep track of consumers. Alternatively, if the social planner can record the realizations of the arrival rate $\epsilon_t(\omega)$, then the social planner can directly compute the marginal social value of capacity since it is then possible to compute $\widehat{b}_t(\omega)$ and $\widehat{p}_t(\omega) = v(t - \widehat{b}_t(\omega))$. There is no need for responsive pricing. In the heterogeneous demand case, however, consumers have private information about their types (a FIFO mechanism is not efficient anymore), and the history of aggregate arrival rate is not sufficient to compute the marginal social value of capacity. The social planner cannot compute $\widehat{p}_t(\omega)$ without the consumers' private information about their types.

The no-crossing condition is restrictive. This condition is necessary because we have made no restriction on the arrival process ϵ_t. The results would still hold under more general demands if one is willing to impose some restrictions on the arrival process. Stated loosely, the main message of this section is that the results generalize as long as no two consumers who can overlap have residual demands that cross over the length of time over which they overlap. For example, the demand of two consumers who never overlap could cross. Similarly, the demand of two consumers could cross after one terminates consumption. This more general interpretation of the no-crossing condition is important because the analysis does not always hold when this condition is not met, as we show in the next section.

[13] Formally, $\overline{r}$ is uniquely defined by $\sum_i \epsilon_h^i n^i(\overline{r}) = Q$ where $n^i(x)$ is defined as $v^i(n^i(x)) = x$ if $v^i(0) > x$ and 0 otherwise.

4.4.2 An example of inefficiency

The analysis does not follow when the no-crossing condition does not hold. To start, one cannot show anymore that the consumer with the lowest marginal valuation should leave first in the efficient allocation (Proposition 5 does not hold). Similarly, we cannot characterize the equilibrium by focusing on the behavior of the consumer with the lowest marginal valuation. Specifically, the proof of Claim 2 in Lemma 1 does not hold.

We show that when the no-crossing condition does not hold, it is possible that responsive pricing cannot approximate the efficient allocation in the sense defined by Proposition 4. An example is sufficient to establish this claim. For tractability concerns, we present an example with discrete arrival process and step-function demands. It is important to recognize that these features violate some of the continuity assumptions of the model. As we argue later, however, this is not with complete loss of generality.

Time is finite, $t \in [0, 2]$, and we use the terminology period 1 to mean $t \in [0, 1]$, and period 2 for $t \in (1, 2]$. The capacity is 3. A demand is a pair of numbers (see also Table 1). A consumer with demand (a, b) who arrives at t, is willing to pay a from t to $t + 1$ and b from $t + 1$ to $t + 2$ and 0 after $t + 2$. There are four types of consumers $v^1 = (20, 20)$, $v^2 = (25, 0)$, $v^3 = (30, 30)$ and $v^3 = (10, 0)$. To simplify, we assume that consumers do not discount the future.

The arrival process is the following. Consumers arrive only at $t = 0$ or $t = 1$. At $t = 0$, there are two possible states of the world, state π and state $1 - \pi$, which occur with respective probabilities π and $1 - \pi$ with $\pi \in [0, 1]$ and $\pi \neq 1/2$. In state π the arrival realization at date 0 is $\epsilon_0^{\pi} = (2, 4, 0, 0)$ while in state $1 - \pi$ the arrival realization is $\epsilon_0^{1-\pi} = (2, 3, 1, 0)$. At date one, the arrival realizations are $\epsilon_1^{\pi} = (0, 0, 0, 4)$ and $\epsilon_1^{1-\pi} = (0, 3, 0, 0)$. Arrival realization ϵ_0^{π}, for example, means that two consumers of type v^1 and four consumers of type v^2 arrive at date 0 in state π. We denote by v_t^i the consumer of type i who arrive at date t.

The efficient consumption rule maximizes total surplus subject to feasibility and implementability constraints (see also Table 2). In state π, all consumers v_0^1 should consume in both periods, 1 unit of consumers v_0^2 should consume in period 1, and 1 unit of consumers v_1^4 should consume in period 2. In state $1 - \pi$, 2 unit of consumer v_0^2 should consume in period 1, all consumers v_0^3 should consume in both periods, and two unit of consumer v_1^2 should consume in period 2. The expected consumer surplus in the first-best consumption rule is $160 - 45\pi$.

Table 1 Consumer preferences

Type	State π		State $1 - \pi$	
	$t = 0$	$t = 1$	$t = 0$	$t = 1$
$v^1 = (20, 20)$	2	0	2	0
$v^2 = (25, 0)$	4	0	3	3
$v^3 = (30, 30)$	0	0	1	0
$v^4 = (10, 0)$	0	4	0	0

Table 2 Consumption rules and surplus

	Consumption				Expected surplus
	State π		State $1-\pi$		
	$t \in [0, 1]$	$t \in [1, 2]$	$t \in [0, 1]$	$t \in [1, 2]$	
Efficiency equilibrium	$2 \times v_0^1 + 1 \times v_0^2$	$2 \times v_0^1 + 1 \times v_1^4$	$2 \times v_0^2 + 1 \times v_0^3$	$2 \times v_1^2 + 1 \times v_0^3$	$160 - 45\pi$
$\pi > 1/2$	$2 \times v_0^1 + 1 \times v_0^2$	$2 \times v_0^1 + 1 \times v_1^4$	$2 \times v_0^1 + 1 \times v_0^3$	$1 \times v_0^3 + 2 \times v_1^2$	$150 - 35\pi$
$\pi < 1/2$	$3 \times v_0^2$	$3 \times v_1^4$	$2 \times v_0^2 + 1 \times v_0^3$	$2 \times v_1^2 + 1 \times v_0^3$	$160 - 55\pi$

Consider next responsive pricing. Assume that the information structure is common knowledge but consumers privately know their types. This implies that at date zero the consumers of type 1 and 2 do not know the state of the world. The next lemma establishes that responsive pricing cannot approximate the efficient outcome.

Lemma 3 *There does not exist a sequence of state prices* $(p_0^{\pi}, p_0^{1-\pi}, p_1^{\pi}, p_1^{1-\pi})$ *such that if consumers are announced the realized state prices in each period they make efficient consumption decisions.*

Proof Consumer v_0^2 has to be indifferent between consuming and not consuming in both states:

$$25 - p_0^{\pi} = 25 - p_0^{1-\pi} = 0.$$

Since $p_0^{\pi} = p_0^{1-\pi} = 25$, the date 0 price cannot reveal the state of the world. Consumer v_0^1 uses his prior to compute the expected surplus from starting consumption in period 1. Consumer v_0^1 has to weakly prefer to consume in state π:

$$20 - p_0^{\pi} + \pi \text{Max}(20 - p_1^{\pi}, 0) + (1-\pi)\text{Max}(20 - p_1^{1-\pi}, 0) \geq 0$$

and not to consume in state $1-\pi$:

$$20 - p_0^{1-\pi} + \pi \text{Max}(20 - p_1^{\pi}, 0) + (1-\pi)\text{Max}(20 - p_1^{1-\pi}, 0) \leq 0.$$

Since consumer v_1^4 and v_1^2 have to be indifferent between consuming and not consuming in state π and $1-\pi$, respectively, the date 1 prices are $p_1^{\pi} = 10$ and $p_1^{1-\pi} = 25$. Plugging these values in the above inequalities, we have $10\pi - 5 \geq 0 \geq 10\pi - 5$. A contradiction since $\pi \neq 1/2$. □

Lemma 3 shows that is not possible that consumer v_0^1 consumes in state π and not in state $1-\pi$. Therefore, the efficient allocation cannot be arbitrarily approximated. To further illustrate, consider the equilibrium under scheme $\widetilde{r}_{\alpha}$ defined in Sect. 3 where $\widetilde{r}_{\alpha}(Q) = 35$ and α close to 0. To understand the construction of the equilibrium, note first that prices will change only at $t = 0, 1$, and 2 since these are the only dates when new consumers arrive or terminate consumption. Next, consider consumers' consumption decisions. Consumers v_0^3 will consume in state $1-\pi$ because their demand (weakly) dominates any other consumer. Consumers v_0^2's consumption decision is also simple. They are willing to pay 25 and no more than 25 at date 0. Solving the decision problem of consumers v_0^1 is more complicated. How much is a consumer v_0^1 willing to pay at date 0? This decision depends on her expectations about the second period price. In state π (respectively $1-\pi$), she expects that the price will be 10 (respectively 25) in period 2. She expects a period 2 surplus of $20 - 10$ with probability π and of 0 with probability $1-\pi$. A consumer v_0^1 is willing to pay $20 + \pi 10 + (1-\pi)0 = 20 + \pi 10$ at $t = 0$. Since $\pi > 0$ consumers v_0^1 are willing to pay more than their period 1 valuation. When $20 + \pi 10 > 25$, the equilibrium price is 25 in period 1 and all consumers v_0^1 consume. When $20 + \pi 10 < 25$, the price is $20 + \pi 10$ in period 1 and no consumers v_0^1 consume. An inefficiency occurs because consumer v_0^1's decision to consume does not depend on the state of the world as it should under the first best outcome.

The problem identified in the example is general and can be summarized as follows. The no-crossing condition does not hold for consumer v_0^1 and v_0^2. It is not optimal for consumer v_0^1 to terminate consumption when the price is equal to her instantaneous valuation 20. To achieve efficiency, consumer v_0^1 would need to know whether only consumers v_0^2 or also consumers v_0^3 have arrived at $t = 0$. This information, however, is not revealed by the price. More generally, under crossing demands a consumer with high long-term demand may prefer to retain consumption and bear negative instantaneous utility if she believes that (a) there are some consumers with weak long-term demands who are about to terminate consumption, and (b) few consumers are likely to arrive.

Consumers' decision problems differ dramatically when the no-crossing condition holds and when it does not. Under no-crossing, consumers need to know only the current price to decide whether to continue or terminate consumption. The fact that consumers do not know who is consuming at the time they arrive (incomplete information about arrival times and types) does not prevent efficiency from being achieved. When the no-crossing condition does not hold, however, consumers do not decide when to terminate consumption only on the basis of the current price. They have to predict future prices. They do so using their prior belief and the price histories, $H_t^s(\omega)$. As a consequence, consumers' beliefs matter. The example offers an illustration of this point. The period 1 price and the level of inefficiency depend on the consumer v_0^1's initial belief about the likelihood that state π will occur. In the example, we assumed that v_0^1's initial belief was equal to the true probability (common knowledge assumption) but this does not have to be the case.

To conclude, we point out that although the example does not satisfy all the assumptions of the model, it stresses the importance of the no-crossing condition. To illustrate, assume that the no-crossing condition holds as would be the case for example if $v^2 = (25, 25)$. Lemma 3 does not hold since it is possible to define a sequence of state prices $(p_0^{\pi}, p_0^{1-\pi}, p_1^{\pi}, p_1^{1-\pi})$ that implements the efficient allocation. Similarly, responsive pricing approaches the efficient allocation.

5 Consumption interruption

The analysis has assumed so far that consumers never postpone consumption. This was imposed by the restriction that the consumption rules $d_t^{i,s}(\omega)$ had to be non-increasing in t. This section generalizes the analysis in two ways. First, we assume that consumers can interrupt the service (or delay initial start) but have to pay a cost each unit of time they do so. We identify a lower bound on the cost of delaying consumption that rules out interruptions. This formalizes the claim made earlier that the analysis is valid as long as the cost of delaying consumption is sufficiently high. Second, we briefly discuss the case where the opportunity cost of delaying consumption is low.

To simplify the presentation, we return to the case where there is a single consumer type. Consumers have to pay k per unit of time when they delay consumption. This could be because consumers have to physically wait or because there is a cost of monitoring prices. Let $d_t^s(\omega) = 0$ when the consumer who arrives at s delays consumption at t and let $l^s(\omega)$ denote the time when that consumer terminates consumption definitely. A consumer who arrives at s gets expected utility

$$U_s^s(\omega, H_s^s(\omega)) = E\left(\int_s^{l^s(\omega)} \rho^{x-s}\left(d_x^s(\omega)\left(v(x-s) - p_x(\omega)\right) - (1 - d_x^s(\omega))k\right) \mathrm{d}x \mid \mu_s^s\right)$$

under consumption strategy $d_t^s(\omega)$. Let $\widehat{d_t^s}(\omega)$ represent the efficient consumption rule.

Proposition 6 *The efficient consumption rule, $\widehat{d_t^s}(\omega)$, is non-increasing in t for $t > s$ if $k > v(Q/\epsilon_h) - v(Q/\epsilon_l)$.*

Proof Consider the efficient consumption rule under the constraint that interruptions are ruled out. Consumers consume Q/ϵ_l when the arrival rate is fixed at ϵ_l and never consume more than that amount. They consume Q/ϵ_h when the arrival rate is fixed at ϵ_h and never consume less than that amount. The social opportunity cost of capacity varies between $v(Q/\epsilon_h)$ and $v(Q/\epsilon_l)$. The maximum possible social gain from interrupting consumption is $(v(Q/\epsilon_h) - v(Q/\epsilon_l))\,\mathrm{d}t$. Interrupting consumption is never efficient when $v(Q/\epsilon_h) - v(Q/\epsilon_l) < k$. □

It is never efficient for consumers to wait when $v(Q/\epsilon_h) - v(Q/\epsilon_l) < k$. Consider the equilibrium analysis. The pricing function influences the decision to delay consumption. Does there exist a responsive pricing function that rules out waiting and still allocates capacity efficiently? Consider first the conditions that one needs to impose on the pricing function to rule out waiting. The benefit from waiting corresponds to the expected savings from lower prices. This amount is bounded from above by $r(Q) - r(0)$. Consider a pricing rule that sets $r(Q) = v(Q/\epsilon_h)$ and $r(0) = v(Q/\epsilon_l)$. This pricing rule eliminates both excess demand and interruptions since $r(Q) - r(0) < k$. The condition $r(0) = v(Q/\epsilon_l)$ is not restrictive because prices never go below that level in the equilibrium analysis without interruptions. The analysis follows and responsive pricing still implements the efficient consumption rule in a limit sense. This simple extension demonstrates that the analysis presented earlier holds when $v(Q/\epsilon_h) - v(Q/\epsilon_l) < k$.

When $k < v(Q/\epsilon_h) - v(Q/\epsilon_l)$, on the other hand, consumer waiting may occur both under responsive pricing and in the first best consumption rule. To make this point clear, consider the extreme case where the opportunity cost of waiting is zero. Under responsive pricing, consumers will prefer to delay consumption if they anticipate that prices are likely to decrease in the future. But it is not efficient anymore that a consumer terminates consumption for every new consumer who arrives, since there is no welfare cost associated with consumers waiting. More generally, even when consumers have a low but positive cost of waiting, it is not efficient anymore to rule out waiting, since there is a trade-off between the welfare cost of waiting and the opportunity cost of cutting off some consumers.[14] We leave a full treatment of this problem for future research.

6 Summary and conclusions

This paper investigates the efficiency properties of responsive pricing, a simple and easily implementable scheme initially proposed by Vickrey to eliminate inefficiencies that result from last minute demand shocks. Responsive pricing changes prices in real time in response to demand realizations, increasing prices when the resource gets close to congestion and decreasing prices when unused capacity increases, thus promoting full capacity utilization. Consumers only have to decide whether they want to consume. The seller, in turn, only needs to be able to measure congestion and to update prices in real time.

The set of applications we have in mind are characterized (a) by the impossibility to get users to commit to a contract ahead of time and (b) by highly unpredictable last-minute demand shocks that prevent the use of schemes that would be sensitive to the designer's information about the demand environment. As argued by Vickrey, such environments rule out standard pricing schemes considered in the literature (e.g. peak load pricing, priority pricing) as a solution to the allocation problem and justify considering simple indirect mechanisms such as responsive pricing.

An important contribution of this paper is to establish a condition under which the strategic complexity of the game that takes place under responsive pricing dramatically simplifies. Under the no-crossing condition, consumers stop consuming as soon as their willingness to pay for a marginal unit falls below the instantaneous price. Consumers cannot benefit from predicting future prices. When demands can cross, however, consumers may optimally keep consuming even if they receive negative net instantaneous utility. As a result, the equilibrium allocation may depend on consumers' initial beliefs.

We show that responsive pricing can implement the efficient outcome but only in a limit sense and when consumer demands satisfy a no-crossing condition. When this condition is violated the analysis does not follow, and responsive pricing

[14] Positive but low cost of waiting may explain why country clubs and ski resorts do not use prices to allocate capacity although waiting is often observed in equilibrium. In these situations, consumers may have a low cost of waiting and it would be suboptimal to cut some consumers short to free up capacity when there is a sudden arrival flow of consumers. This conclusion is reminiscent of the analysis of ski lifts presented in Barro and Romer (1987).

sometimes fails to achieve efficiency. The problem with responsive pricing is that consumers can bid only for the current unit of consumption, and the equilibrium price does not always aggregate consumers' private information efficiently. An implication for policymaking is that responsive pricing will work well when consumer demands satisfy the no-crossing condition, such as among homogenous populations of consumers.

Undoubtely, this work had to leave several important issues for future work. Most importantly, one could consider more general consumption rules as suggested in the introduction and also more sophisticated information revelation schemes than responsive pricing. We believe, however, that one should focus on simple schemes, such as the one proposed by Vickrey and considered in this work. If one accepts this view, a relevant question for future research is to generalize the class of pricing mechanisms, possibly incorporating more state variables than just current utilization rates, offering partial advance booking, or introducing the possibility of rationing.

Another limitation of this work is that we have focused on a welfare analysis. Our results are relevant to regulated industries considering introducing responsive pricing. Some of the applications discussed in Sect. 1, however, have to do with non-regulated firms concerned about firm surplus rather than total surplus. An important extension would be to derive the profit maximizing pricing scheme and to contrast it with responsive pricing. Would a private firm find it optimal to vary prices as a function of occupancy realizations? Under what conditions?

References

Barro, R., Romer, P.: Ski-lift pricing, with applications to labor and other markets. Am Econ Rev **77**, 875–890 (1987)

Bohn, R.E., Caramanis, M.C., Schweppe, F.C.: Optimal pricing in electrical networks over space and time. RAND J Econ **15**, 360–376 (1984)

Boiteux, M.: Sur la Gestion des Monopoles Publics Astreints a l'Equilibre Budgetaire. Econometrica **24**, 22–40 (1956)

Boiteux, M.: Peak load pricing. J Business **33**, 157–179 (1960)

Borenstein, S.: Frequently asked questions about implementing real-time electricity pricing in California for summer 2001. Mimeo, University of Berkeley, (2001)

Crew, M.A., Fernando, C.S., Kleindorfer, P.R.: The Theory of peak-load pricing: a survey. J Regulat Econ **8**, 215–48 (1995)

Courty, P., Pagliero, M.: Estimating the Welfare Gains From Real Time Pricing: Evidence from an Internet café. CEPR Discussion Paper 4149 (2003)

Falkner, M., Devetsikiotis, M., Lambradaris, I.: An overview of pricing concepts for broadband IP networks. IEEE Commun Surv **3**, 2–13 (2000)

Harris, M., Raviv, A.: A theory of monopoly pricing schemes with demand uncertainty. Am Econ Rev **71**, 347–365 (1981)

Joskow, P., Tirole, J.: Retail electricity competition. Mimeo. Cambridge: MIT (2004)

MacKie-Mason, J., Varian H.R.: Pricing congestible network resources. IEEE JSAC **13**, 1141–49 (1995)

Mendelson, H., Whang, S.: Optimal incentive-compatible priority pricing for the M/M/1 queue. Operat Res **38**, 870–883 (1990)

Oren, S.S., Smith, S.S., Wilson, R.B.: Capacity pricing. Econometrica **53**, 545–567 (1985)

Vickrey, W.: Responsive pricing of public utility services. Bell J Econ Manage Sci **2**, 337–346 (1971)

Wilson, R.B.: Nonlinear Pricing. Oxford: Oxford University Press (1993)