

# Pricing and Dimensioning Competing Large-Scale Service Providers

Gad Allon, Itai Gurvich

Kellogg School of Management, Northwestern University, Evanston, Illinois 60208  
{g-allon@kellogg.northwestern.edu, i-gurvich@kellogg.northwestern.edu}

The literature on many-server approximations provides significant simplifications toward the optimal capacity sizing of large-scale monopolists, but falls short of providing similar simplifications for a competitive setting in which each firm's decision is affected by its competitors' actions. In this paper, we introduce a framework that combines many-server heavy-traffic analysis with the notion of epsilon-Nash equilibrium and apply it to the study of equilibria in a market with multiple large-scale service providers that compete on both prices and response times. In an analogy to fluid and diffusion approximations for queueing systems, we introduce the notions of *fluid game* and *diffusion game*. The proposed framework allows us to provide first-order and second-order characterization results for the equilibria in these markets. We use our results to provide insights into the price and service-level choices in the market and, in particular, into the impact of market scale on the interdependence between these two strategic decisions.

*Key words:* competition; games; approximate equilibrium; asymptotic analysis; heavy traffic; Halfin–Whitt regime; services

*History:* Received: August 2, 2007; accepted August 24, 2009. Published online in *Articles in Advance*.

## 1. Introduction

An important attribute of customer experience in various service industries is the time spent on waiting for service. As a result, customers may consider both the prices and the delay guarantees in choosing which provider to patronize. The purpose of this paper is to study the equilibria that emerge in markets in which large-scale service providers compete on both prices and customer delays. We focus on understanding the impact of market size on the way in which different firms make these pricing and service-level choices.

Toward that end, we analyze a model of competition with multiple large-scale service providers in which the demand faced by each firm depends on the prices and the service levels offered by all firms in the market. Quantitatively, our goal is to characterize the capacity and pricing choices of the firms in the market. Qualitatively, we wish to understand how the strategic positioning of the firm depends on its own characteristics vis-à-vis those of its competitors.

To address these issues, we must first examine the firm's capacity decision. When service levels are measured through delays, a decision to improve the

service level requires an investment in increased capacity. Hence, in positioning itself in the market, a firm must weigh the benefits of high service levels against the associated capacity costs. The benefits of improved service levels are not, however, independent of other competitors' actions, and hence the task of determining the trade-off between efficiency and service quality is a nontrivial one.

This trade-off is a nontrivial one even for a monopolist. Indeed, when capacity is adjusted by determining the number of service representatives (rather than by adjusting the service rates), the problem of optimizing capacity costs versus waiting-time-related costs is a complex optimization problem. Although it can often be solved numerically, numerical solutions fail to provide any structural insights. An alternative to exact numerical solutions is the use of approximations. Many-server approximations provide a simplified means to approach this problem (see, e.g., Borst et al. 2004; additional references are provided in §2). In this type of analysis, one considers a sequence of queueing systems with growing demand (and with capacity that grows accordingly to satisfy

this demand). One then identifies solutions that are asymptotically optimal as the demand grows. The asymptotically optimal solution is nearly optimal for a given system provided that the demand it faces is large enough.

The literature on many-server approximations not only provides a tractable way to characterize nearly optimal capacity and price decisions for monopolists; it also relates a firm's operational regime to the relative significance the firm ascribes to service levels as opposed to capacity costs (see the discussion of Borst et al. 2004 in §2). The firm's operational regime dictates how the firm should optimally respond to an increase in market size. Some firms should use their growth to increase their utilization (and thus their cost efficiency) without improving their service level. These firms are said to operate in the *efficiency-driven* (ED) regime. Their emphasis on efficiency results in a situation in which (when the market is large) almost all customers experience some delay before being served. Some firms will sacrifice efficiency for quality. In response to an increase in market size, these firms will match an increase in utilization with yet a greater improvement in service levels. These operate in the *quality-driven* (QD) regime. In this regime, almost all customers are served immediately. An intermediate regime is the *quality- and efficiency-driven* (QED) regime—also known as the Halfin–Whitt regime after the authors that first formalized it (see §2)—which corresponds to firms for which efficiency and quality are of similar importance. These firms will match the increase in efficiency with a comparable increase in the quality of service. In this regime, a nontrivial fraction (but not all) of the customers receive service immediately, without any delay, but, at the same time, the efficiency is very high.

The regime-characterization results are proved in the literature for service providers that are monopolists in their respective markets. For a monopolist, the many-server approximations provide a tractable way to characterize its optimal capacity choices. The competitive setting is, however, more complex. Not only does the discrete nature of the capacity choice make the task of identifying equilibria and obtaining quantitative and qualitative results more arduous, the task is further complicated by the fact that the demand the firm experiences is not fixed, nor

does it depend solely on the firm's own pricing and service-level choices. Rather, the demand depends on the choices made by all firms in the market. It seems plausible, however, that many-server approximations can be embedded within a game theoretic analysis to characterize equilibria in these markets. We pursue this direction by constructing a formal framework that draws on many-server approximations, as developed for monopolists, and by applying it to the study of equilibria in competitive markets.

Two fundamental questions are central to the study of equilibria in competitive markets: (a) Do Nash equilibria exist in the market (*existence*)? And (b) given some sort of existence, is it possible to characterize the set of equilibria to obtain qualitative insights into market outcomes (*characterization*)? Starting with existence, we note that the concept of Nash equilibrium may be too restrictive for describing service-market behavior. It is known that Nash equilibria need not exist even under the most common demand functions, such as Multinomial Logit (MNL), and the simplest supply systems, such as the  $M/M/1$  queue (see, e.g., Cachon and Harker 2002). This nonexistence is often driven by economies of scale, but is further exacerbated by the lumpy nature of the capacity in settings where the capacity choices are made in a discrete manner, by adjusting the number of service representatives. Nonexistence of Nash equilibria does not rule out the possibility to say something meaningful about the market outcomes. It is desirable in these cases to find a less stringent framework that will allow for some characterization of the market outcomes.

The mathematical framework we propose is designed to address two concerns: (a) in terms of existence, we want to overcome the restrictive nature of the Nash equilibrium in addressing relatively general demand functions as well as supply facilities that are more general than the  $M/M/1$  queue, and (b) in terms of characterization, we want to handle the complex nature of the service system by combining approximations for the queueing dynamics with a game theoretic framework.

Our framework stands on three pillars: (i)  $\epsilon$ -Nash (or approximate) equilibria, (ii) many-server approximations, and (iii) market replication. The introduction of approximate equilibrium is aimed, initially,

to overcome the nonexistence of Nash equilibria. Its eventual benefits, however, go beyond this initial objective when combined with market replication and many-server approximations. We examine the behavior of equilibria not in a single market, but rather in a sequence of markets with increasing aggregate demand—these are referred to as replicated markets. We emphasize that, when characterizing the equilibrium behavior in these markets, we assume that the set of firms is given; in other words, we do not consider the possibility of firms exiting or entering the industry.

Our framework can be thought of as a formalization of the use of fluid and diffusion models of queueing systems in a competitive setting. In the optimization of queueing systems, the original system is often replaced by a deterministic approximation—a *fluid model*—whose analysis sheds light on first-order properties of the underlying queueing system, such as its stochastic stability. In a second step, the original queueing system is replaced by a (more refined) stochastic model, which is often referred to as a *diffusion model* of the queueing system. The latter is often more tractable than the original queueing system and can be used to identify properties that are asymptotically correct for the original queueing system. In particular, the diffusion model is often used to construct nearly optimal solutions for optimization problems that are intractable for the original queueing system.

Analogously, our framework constructs approximate games for the game played among the service providers. We first introduce a *fluid game* that is obtained from the original game by disregarding the stochastic nature of congestion. Building on the analysis of the fluid game, we then introduce a more refined *diffusion game*. This game is obtained from the original one by replacing each of the service providers with its many-server diffusion approximation. We then relate the equilibria of this new approximate game with the outcomes of the original market. As in many-server approximations, the idea is to show that the equilibria of the diffusion game are, in a sense, asymptotically correct for the original game.

The notion of  $\epsilon$ -Nash equilibrium plays a key role in rigorously establishing these approximations. The approximate equilibrium concept provides a formal way to construct *envelopes* for the profits of the firms

in the market. Although a Nash equilibrium might not exist and the market might oscillate, the  $\epsilon$ -Nash identifies a region within which the *profits* of all the firms in the market must reside. The ultimate goal of this paper is, however, to understand market positioning in terms of the *actions* of the different firms (i.e., the prices and service levels that the firms choose). The challenge is, then, to use the envelopes on the profit functions to construct corresponding envelopes in the action space around the approximate price and service-level choices. To our knowledge, there are no general results that, given an  $\epsilon$ -Nash equilibrium, identify the maximum that the firms can deviate in their *actions* without causing a deviation in the *profits* that would compromise the approximate equilibrium. Such results, which characterize the maximal oscillations of the prices and service levels around some point, are thus unique, and are obtained through the framework that we develop by employing the concepts of replicated markets in conjunction with heavy-traffic queueing theory.

Having constructed the analysis framework, we use it to provide an analytical characterization of the approximate equilibria in the market with multiple service providers. The characterization is then used to obtain some insights into the market outcomes. Our insights are concerned with the relationship between the price and service-level choices and, in particular, between the functions in the firm that make these choices—marketing and operations.

We identify a *one-sided decoupling* phenomenon by which the firms can be fairly close to optimality by allowing the price-setting function to “lead,” and the operations function to “follow.” The approximate equilibria—in both the fluid and diffusion level—exhibit a sequential structure: one can first pretend that the customers in the market are entirely insensitive to service levels and solve a simple price competition game. The real price and service level choices, in the approximate equilibrium, are then a function of this “naive” price vector but are, otherwise, independent of each other. This independence allows the marketing function to set the “naive” price vector as an initial estimate for the optimal price and leave for the operations function the task of setting the service levels and adjusting the prices that are eventually offered to the customers.

The analysis of the diffusion game provides a refined understanding of the operational regime of a firm and the implication of this regime on the firm's price choices. We show that both the QED and the ED regimes can emerge in equilibrium, thus, we appear to be the first to show how these different regimes emerge in a competitive market and, in particular, how different demand structures lead to the different regimes. We show that, while the actual choices of service level and price depend on the characteristics of all firms in the market, the operational regime of a firm is determined solely by its own intrinsic properties. Consequently, when different firms have different sensitivities, they may operate in different operational regimes, and thus position themselves differently in the face of increased market size.

We also find that the operational regime of a firm determines the degrees of freedom it has in pricing. We show that, compared with firms that operate in the QED regime, firms operating in the ED regime have greater freedom in choosing the prices they charge. Their freedom is reflected by the fact that they have a larger set from which they can choose their prices with hardly any compromise to their profits. Thus, firms in the ED regime can keep the one-sided decoupling in the sense that the marketing function can pay less attention to the operational side in determining the prices. Firms operating in the QED regime need to pay greater attention to their price choices. For these firms, the decoupling is weaker and a feedback mechanism is required between the manner in which the firm operates (i.e., the operational regime), and its pricing.

## 2. Literature Review

Our work builds on two streams of literature: (a) game theory and its application to competition analysis, and (b) queueing theory and its application to the study of large-scale service systems. These two streams are not disjoint, and some recent work lies at the intersection of the two.

The literature on competition in service industries dates back to the late 1970s; see, e.g., Levhari and Luski (1978). Although it initially focused on a single attribute—price or service level (or a simple aggregation of the two)—more recent work treats prices and waiting-time standards as fully independent

attributes. We follow Allon and Federgruen (2007) in considering a model with *differentiated* services (i.e., a model in which other service attributes matter along with the full price) and in treating delay and price as independent attributes. We refer the reader to Allon and Federgruen (2007) for a systematic discussion of existing results in this context, and to Hassin and Haviv (2003) for a general survey of queueing models with competition.

Allon and Federgruen (2007) and others focus on providing a full analytical characterization of the Nash equilibria that arise in a market in which the market size is fixed. In contrast, we focus on understanding the impact of the market scale on the prices, service levels, and interdependencies between the two. Furthermore, our framework significantly expands the family of models that can be studied. This expansion is in two directions: (i) First, most of the literature on competition in services models the supply side via  $M/G/1$  queues, which implies, in turn, that capacity choices are made continuously by adjusting the service rates. We, in contrast, allow the service provider to adjust its capacity by increasing or decreasing the (integer-valued) number of service representatives, giving rise to an  $M/M/N$  queue, where  $N$  is a decision made by the firm. This is a common method of capacity management of service providers and one that renders the Nash equilibrium intractable for characterization. (ii) Second, in terms of the demand model, our framework allows for significant generality in modeling the customers' sensitivity to service levels and prices.

From the game theoretic perspective, the notion of  $\epsilon$ -Nash equilibria that we use has been used extensively in the economics literature. For the basic definition we rely on Tijs (1981). Dixon (1987) uses the idea of market replication in the context of price competition. Although our form of replication is different—Dixon (1987) increases the number of firms in the market, whereas we increase the aggregate demand volume—our analysis is inspired by his concept. Previous work in game theory has focused on three types of sequences of games: (i) sequences of games in which the action space is getting increasingly finer, and although each game has discrete action space, the limiting game has continuous action space (see, e.g., Whitt 1980); (ii) sequences of games in which

the number of agents grows (Lu et al. 2009); and (iii) a sequence of replicated markets with growing market size (see, e.g., Dixon 1987). We use the third framework.

The application of  $\epsilon$ -Nash in the operations literature is rare. Lu et al. (2009) use this concept in a setting where Nash equilibrium does exist but the  $\epsilon$ -Nash equilibrium concept still helps in characterizing the equilibrium in the game when the number of players is large and approaches a continuum. Dasci (2003) uses this concept in the context of  $\epsilon$ -subgame-perfect equilibrium. We appear to be the first to combine the concepts of  $\epsilon$ -Nash, market replication, and heavy traffic in the context of operational settings. This combination allows us to discuss both stability and trends in markets of competing service providers.

With respect to the relevant queueing literature, our work builds on the literature on many-server approximations of monopolists, starting with the seminal work of Halfin and Whitt (1981). Although many-server approximations existed before, the result of Halfin and Whitt (1981) made such approximations relevant for various applications, such as call-center operations (see the survey paper by Gans et al. 2003) and, more recently, health-care operations (see, e.g., Jennings and de Véricourt 2008).

Halfin and Whitt (1981) consider a sequence of  $M/M/N$  queues and show that, as the demand rate  $\Lambda$  grows, the probability of delay  $P\{W > 0\}$  converges to a number strictly between 0 and 1 if and only if the number of agents grows with  $\Lambda$  according to a *square-root safety-staffing rule*, i.e., if and only if

$$N = R + \beta\sqrt{R} + o(\sqrt{R}), \quad (1)$$

where  $R := \Lambda/\mu$  is the offered load and  $\beta$  is a strictly positive constant. In particular, a service provider that uses the square-root safety-staffing rule to determine its capacity will utilize its servers very efficiently and, at the same time, have a nontrivial fraction of its customers enter service immediately upon their arrival. This combination of high efficiency and high service level provides the justification for the name (quality- and efficiency-driven) regime.

Whereas Halfin and Whitt (1981) identified this regime, Borst et al. (2004) placed many-server approximations within a broad economic framework that

considers the problem of minimizing capacity and waiting-time costs. They showed how the QED regime emerges as the optimal economical choice in some cases, but also identified conditions under which other regimes—namely, the QD and ED regimes—emerge as the optimal choices. A key idea in the framework developed in Borst et al. (2004) is to replace the original optimization problem, which involves the integer-valued number of servers, by a tractable continuous and convex optimization problem. Using similar ideas, we will construct a continuous and tractable game—the *diffusion game*—that will serve as an approximation for the original, relatively intractable one. Recently, Kumar and Randhawa (2009) extended the work of Borst et al. (2004) to a setting in which the customers are price and delay sensitive, and consequently the demand is not fixed. Their work shows how different operational regimes emerge depending on the convexity (or concavity) of the delay cost function. Similar dependencies will also emerge within the competitive setting that we study in this paper.

Additional examples of work that provides staffing and pricing recommendations for large-scale monopolists facing delay- and price-sensitive customers are the papers by Whitt (2003) and Maglaras and Zeevi (2003, 2005).

All the work mentioned above considers a monopolist with a demand rate that may depend only on the congestion experienced and the price charged by this single player. Our paper appears to be the first to show how the different operational regimes emerge in a competitive setting with multiple players and to identify the dependencies between the operational regimes and the price choices of the various players in the market.

## 2.1. Organization of this Paper

The rest of this paper is organized as follows. We present the model in §3. Section 4 is concerned with regime characterization. In §5 we introduce and characterize the fluid game and discuss its implications. In §6 we turn to the diffusion game, which is concerned with a refined understanding of the firms' choices. Conclusions and directions for future research are discussed in §7. The e-companion to this paper contains numerical examples as well as generalizations to some of the results in §§5 and 6.

Our approach in presenting the results is to state them formally within the paper, accompanied by various examples for illustration. Most of the detailed proofs are relegated to the online appendix that can be downloaded from <http://www.kellogg.northwestern.edu/faculty/gurvich/personal/>.

### 3. The Model

We consider a market with a set  $\mathcal{I} = \{1, \dots, I\}$  of competing service firms, each operating as an  $M/M/N$  facility and serving arriving customers in a first-come, first-served manner. Firm  $i$  positions itself in the market by selecting a price  $p_i$  and a delay guarantee  $T_i$ . We restrict our attention to service-level guarantees that are given in terms of the customers' delay rather than their whole sojourn time in the system. Having chosen the delay target  $T_i$ , the service provider guarantees that the following service-level constraint will be satisfied:

$$\mathbb{P}\{W_i > T_i\} \leq \phi, \quad (2)$$

where  $W_i$  is the steady-state delay with the  $i$ th provider, and  $0 < \phi < 1$  is the satisfaction probability. This form of service-level constraint is consistent with the industry practice that commonly uses  $\phi = 0.2$  (corresponding to 80% of the service requests being answered within target; see, e.g., Anton 2001).<sup>1</sup> In this paper, we study a model of competition where both the prices and the service levels are set simultaneously. Our results continue to hold if the strategies are chosen sequentially (price first or service level first).

Service rates are assumed to be fixed and equal to  $\mu_i$  for firm  $i$ , and the capacities are adjusted through the choice of the number of agents (or service representatives), denoted by the integer-valued decision variable  $N_i$ . We assume that there is an upper bound  $\bar{T} > 0$  on the acceptable service levels. For example, in call centers, it is clear that a waiting time of more than a day is unacceptable. Firm  $i$  chooses  $T_i \in [0, \bar{T}]$  and needs to adjust its capacity,  $N_i$ , so as to guarantee that the service-level constraint is satisfied for the chosen target. We let  $\Theta := \times_{i=1}^I [0, \bar{T}]$ . Given the target  $T_i$  and the demand rate  $\lambda_i$ , the required capacity for firm  $i$  is given by

$$N_i = \min\{N \in \mathbb{Z}_+ : P\{W(\lambda_i, \mu_i, N) > T_i\} \leq \phi\},$$

where  $W(\lambda_i, \mu_i, N)$  is the steady-state delay in an  $M/M/N$  queue with arrival rate  $\lambda_i$ , service rate  $\mu_i$ , and  $N$  servers.<sup>2</sup> We write

$$N_i = R_i + \hat{e}_i(\lambda_i, T_i), \quad (3)$$

where  $R_i := \lambda_i/\mu_i$  is the offered load given the demand  $\lambda_i$  faced by firm  $i$ , and  $\hat{e}_i(\lambda_i, T_i) := N_i - R_i$  is the excess capacity required to satisfy the service-level target. Naturally, we define  $\hat{e}_i(\lambda_i, T_i) = 0$  whenever  $\lambda_i = 0$ , but we note that  $\hat{e}_i(\cdot, \cdot)$  must be positive whenever  $\lambda_i > 0$  to guarantee stability. The two terms in (3) represent the two components of the required capacity: The offered load is the *volume-based capacity*, namely, it is the base capacity ensuring that the service process is stable. The second component ensures that the desired service levels are achieved and is referred to as the *service-based capacity*.

Firm  $i$  incurs a cost  $c_i$  per customer served and a cost  $\gamma_i$  per agent, per unit of time. This corresponds to the cost of capacity being linear in the number of agents.<sup>3</sup> The price  $p_i$  is chosen from a compact interval  $[p_i^{\min}, p_i^{\max}]$ . Because firm  $i$  will always select a price  $p_i$ , which results in a nonnegative gross profit margin  $p_i - c_i - \gamma_i/\mu_i$ , we may assume, without loss of generality, that

$$p_i^{\min} = c_i + \frac{\gamma_i}{\mu_i}, \quad i \in \mathcal{I}. \quad (4)$$

The upper bound,  $p_i^{\max}$ , is allowed to obtain any value in  $[p_i^{\min}, \infty)$ . We set  $\mathcal{P}_i := [p_i^{\min}, p_i^{\max}]$  and  $\mathcal{P} := \times_{i=1}^I \mathcal{P}_i$ . In full generality, the demand rates are specified as general functions of all prices and delay guarantees, i.e.,  $\lambda_i \equiv \lambda_i(p, T)$ , where  $p = (p_1, \dots, p_I)$  and  $T = (T_1, \dots, T_I)$ .

**ASSUMPTION 1 (REGULARITY ASSUMPTIONS ON THE DEMAND FUNCTIONS FOR DIFFERENTIATED SERVICES).** For each  $i \in \mathcal{I}$ , the function  $\lambda_i(\cdot, \cdot): \mathcal{P} \times \Theta \mapsto \mathbb{R}_+$  is strictly positive, continuous, and differentiable in all arguments, and strictly decreasing in  $p_i$  and  $T_i$ .

Firm  $i$ 's long-run average profit  $\Pi_i$  is then given by

$$\Pi_i(p, T) = \lambda_i(p, T)(p_i - c_i) - \gamma_i N_i,$$

<sup>2</sup>  $N_i$  can be calculated by iteratively using the Erlang-C formula. Freeware calculators can be found, for example, at <http://iew3.technion.ac.il/serveng/4CallCenters/Downloads.htm> or <http://www.cs.vu.nl/~koole/ccmath/ErlangC>.

<sup>3</sup> See §7 for a discussion of more general capacity-cost models.

<sup>1</sup> Our results are easily extended to the case where  $\phi$  is allowed to vary among different firms.

which, using (3), is rewritten as follows:

$$\Pi_i(p, T) = \lambda_i(p, T) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\lambda_i, T_i). \quad (5)$$

The assumption of large-scale service systems is introduced by considering a family of markets indexed by a market-scale multiplier  $\Lambda \geq 0$  so that the demand grows with the market-scale multiplier in a natural way. Specifically, we let

$$\Lambda_i(p, T) := \Lambda \cdot \lambda_i(p, T) \quad (6)$$

be the demand facing firm  $i$  in the  $\Lambda$ th market. The profit functions in the  $\Lambda$ th market are then given by

$$\Pi_i^\Lambda(p, T) = \Lambda_i(p, T) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, T_i), \quad i \in \mathcal{J}. \quad (7)$$

For future reference, we make the following formal definition.

**DEFINITION 1 (THE MARKET GAME).** The  $\Lambda$ th market game is the  $I$ -player game with profit functions  $\{\Pi_i^\Lambda(\cdot, \cdot), i \in \mathcal{J}\}$  and strategy space  $\mathcal{P} \times \Theta$ .

As is the case in heavy-traffic analysis, the key idea of our market replication procedure is to embed the real market (with fixed market size) into a sequence of markets with growing demand. Looking at the sequence of markets, we are interested in understanding how the stability of the market and the market outcomes change with the increase in market size. Following conventional notation, we let

$$(p, T)_{-i} = ((p_1, T_1), \dots, (p_{i-1}, T_{i-1}), (p_{i+1}, T_{i+1}), \dots, (p_I, T_I)).$$

We denote by  $T_i^{*,\Lambda}(p, T)$  and  $p_i^{*,\Lambda}(p, T)$ , respectively, the delay and price components of firm  $i$ 's best response to  $(p, T)_{-i}$  in the  $\Lambda$ th market game. The existence of a best response for any actions  $(p, T)_{-i}$  follows from the continuity of the demand functions and the compactness of the strategy space. When the best response is not unique, we arbitrarily choose one best response. The way this best response is chosen will be immaterial for our results.

As discussed in the introduction, the market game is intractable for direct Nash equilibria analysis. This is a consequence of the complexity of the expressions for the service-based capacity, the discreteness

of this capacity, and the concavity of the capacity cost function.<sup>4</sup> Instead, we take an indirect approach that exploits the benefits of large-scale asymptotic analysis within an  $\epsilon$ -Nash-equilibrium framework.

### 3.1. $\epsilon$ -Nash Equilibria

The notion of  $\epsilon$ -Nash equilibrium is adopted from Tijs (1981). Rather than defining it in general terms, we provide the definition as it applies to our setting. To this end, given  $(p, T) \in \mathcal{P} \times \Theta$  and  $(\tilde{p}_i, \tilde{T}_i) \in \mathcal{P}_i \times [0, T]$ , we let

$$(\tilde{p}_i, \tilde{T}_i): (p, T)_{-i} = ((p_1, T_1), \dots, (p_{i-1}, T_{i-1}), (\tilde{p}_i, \tilde{T}_i), (p_{i+1}, T_{i+1}), \dots, (p_I, T_I)).$$

**DEFINITION 2 ( $\epsilon$ -NASH EQUILIBRIUM FOR THE  $\Lambda$ TH MARKET GAME).** Fix  $\Lambda \geq 0$ . Let  $\epsilon = (\epsilon_1, \dots, \epsilon_I)$  be a positive vector. We say that  $x \in \mathcal{P} \times \Theta$ , is an  $\epsilon$ -Nash equilibrium of the  $\Lambda$ th market game if, for each  $i \in \mathcal{J}$  and any  $\tilde{x}_i \in \mathcal{P}_i \times [0, \bar{T}]$ ,

$$\Pi_i^\Lambda(\tilde{x}_i; x_{-i}) \leq \Pi_i^\Lambda(x) + \epsilon_i.$$

Nash equilibrium is a special case of  $\epsilon$ -Nash in which  $\epsilon = 0$ . The generalization from Nash to  $\epsilon$ -Nash allows us to construct an “envelope” around the market outcomes, and thus obtain key insights about the market behavior even in cases in which Nash equilibria do not exist. The ability to construct such “envelopes” is useful also in cases in which Nash equilibria do exist but are difficult to characterize. In these cases, if  $\epsilon$  is small enough, the characterization of the  $\epsilon$ -Nash equilibria can shed light on the Nash equilibrium. We will be formally constructing such “envelopes” as well as analyzing the gaps between the  $\epsilon$ -Nash and Nash equilibria whenever the latter exist.

**3.1.1. Notational Conventions.** For two sequences of positive vectors  $\{a^\Lambda, \Lambda \geq 0\}$  and  $\{b^\Lambda, \Lambda \geq 0\}$  with elements in  $\mathbb{R}_+^d$ , we say that  $a^\Lambda = O(b^\Lambda)$  if  $\limsup_\Lambda a_i^\Lambda/b_i^\Lambda < \infty$  for  $i = 1, \dots, d$ . We say that  $a^\Lambda = o(b^\Lambda)$  if  $\limsup_\Lambda a_i^\Lambda/b_i^\Lambda = 0$  for  $i = 1, \dots, d$ . Finally, for  $d = 1$ , we say that  $a^\Lambda \sim b^\Lambda$  if  $a^\Lambda = O(b^\Lambda)$

<sup>4</sup> It is possible to construct continuous versions of the service-based capacity—see, e.g., §4 of Borst et al. (2004). This, however, would still leave the market game intractable for exact analysis.

but  $a^\Lambda \neq o(b^\Lambda)$ . For a vector  $x \in \mathbb{R}^d$ , we let  $\|x\| = \sum_{k=1}^d |x_k|$ . When applied to a vector  $x \in \mathbb{R}^d$ , the absolute value operation should be interpreted componentwise, i.e.,  $|x| = (|x_1|, \dots, |x_d|)$ . Similarly, for  $x \in \mathbb{R}_+^d$ ,  $\sqrt{x} = (\sqrt{x_1}, \dots, \sqrt{x_d})$ . The notation “ $\rightarrow$ ” stands for convergence as  $\Lambda \rightarrow \infty$  unless explicitly stated otherwise. We use 0 to represent the 0 vector in  $\mathbb{R}^d$ , and the dimension of the vector will always be clear from the context. Finally, we use the abbreviated notation  $\Lambda_i$  instead of  $\Lambda_i(p, T)$  whenever the arguments  $p$  and  $T$  are clear from the context.

#### 4. Regime Characterization

In this section, we discuss the optimal operational regimes of the firms in the market (QED, ED, or QD) and relate this regime choice to the underlying demand models. As discussed in the introduction, a firm’s operational regime is the outcome of the firm trading off its capacity cost and the service level it provides. For a monopolist, this trade-off is solely a function of the firm’s own scale economies. In an oligopolistic setting, however, the value of a service level for a given firm depends on its competitors’ decisions, thus making the trade-off more subtle.

The outcome of the regime analysis will be a mapping from firm  $i$ ’s demand structure to a quantifier  $r_i^\Lambda$ . This quantifier characterizes the order of magnitude of the optimal service-level choice for firm  $i$ . Some firms will have  $r_i^\Lambda = 1/\sqrt{\Lambda}$ , and we will show that, for these firms, it is optimal to use a service-based capacity that is of the order of  $\sqrt{\Lambda}$ . Consequently, these firms will operate (in equilibrium) in the QED regime. Other firms will have  $r_i^\Lambda$ , which is significantly larger than  $1/\sqrt{\Lambda}$ . These firms will optimally use a service-based capacity of a magnitude  $o(\sqrt{\Lambda})$  and will operate in the ED regime.

To motivate our results, note that, given a price vector  $p^\Lambda$  and a service-level choice  $T_{-i}^\Lambda$  by its competitors, firm  $i$ ’s best service-level choice is given by

$$\begin{aligned} T_i^\Lambda &\in \arg \max_{x \in [0, \bar{T}]} \Pi_i^\Lambda((p_i^\Lambda, x): (p^\Lambda, T^\Lambda)_{-i}) \\ &= \arg \max_{x \in [0, \bar{T}]} \Lambda_i((p_i^\Lambda, x): (p^\Lambda, T^\Lambda)_{-i}) \\ &\quad \cdot \left( p_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, x). \end{aligned}$$

Equivalently,

$$\begin{aligned} T_i^\Lambda &\in \arg \max_{x \in [0, \bar{T}]} [\Lambda_i((p_i^\Lambda, x): (p^\Lambda, T^\Lambda)_{-i}) \\ &\quad - \Lambda_i((p_i^\Lambda, 0): (p^\Lambda, T^\Lambda)_{-i})] \\ &\quad \cdot \left( p_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, x). \end{aligned} \quad (8)$$

The order of magnitude of  $T_i^\Lambda$  is determined, then, by optimally balancing the loss of market share due to customer delays—which we informally refer to as the “delay cost”—and is given by  $\Lambda_i((p_i^\Lambda, x): (p^\Lambda, T^\Lambda)_{-i}) - \Lambda_i((p_i^\Lambda, 0): (p^\Lambda, T^\Lambda)_{-i})$  against the service-based capacity cost  $\gamma_i \hat{e}_i(\Lambda_i, x)$ .<sup>5</sup> To identify this order of magnitude, Lemma 1 provides us with estimates on the order of magnitude of the service-based capacity, and Assumption 2 allows us to control the delay cost.

LEMMA 1. Fix a sequence  $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$  such that  $(p^\Lambda, T^\Lambda) \in \mathcal{P} \times \Theta$  for all  $\Lambda \geq 0$ . Then,

$$\hat{e}_i(\Lambda_i, T_i^\Lambda) \sim \min \left\{ \frac{1}{T_i^\Lambda}, \sqrt{\Lambda} \right\}.$$

ASSUMPTION 2 (BEHAVIOR AROUND  $T = 0$ ). For each  $i \in \mathcal{I}$  there exists  $\alpha_i > 0$  such that

$$\limsup_{x \rightarrow 0} \sup_{(p, T_{-i}) \in \mathcal{P} \times [0, \bar{T}]^{l-1}} \frac{\lambda_i(p, T_{-i}, 0) - \lambda_i(p, T_{-i}, x)}{x^{\alpha_i}} < \infty,$$

and

$$\liminf_{x \rightarrow 0} \inf_{(p, T_{-i}) \in \mathcal{P} \times [0, \bar{T}]^{l-1}} \frac{\lambda_i(p, T_{-i}, 0) - \lambda_i(p, T_{-i}, x)}{x^{\alpha_i}} > 0.$$

Plugging into (8) Assumption 2 and Lemma 1, we have (informally) that

$$T_i^\Lambda \sim \arg \max_{x \in [0, \bar{T}]} \left[ -\Lambda x^{\alpha_i} - \min \left( \frac{1}{x}, \sqrt{\Lambda} \right) \right].$$

Hence, we should have that

$$T_i^\Lambda \sim \arg \min_{x \in [1/\sqrt{\Lambda}, \bar{T}]} \Lambda x^{\alpha_i} + \frac{1}{x}.$$

A simple calculation then yields that  $T_i^\Lambda \sim r_i^\Lambda$  where

$$r_i^\Lambda = \max \left\{ \frac{1}{\Lambda^{1/(1+\alpha_i)}}, \frac{1}{\sqrt{\Lambda}} \right\}, \quad (9)$$

<sup>5</sup> Here, the loss of market share parallels the role of the delay cost in the monopolist setting of Borst et al. (2004).

so that  $r_i^\Lambda = 1/\sqrt{\Lambda}$  for all  $\alpha_i \leq 1$  and  $r_i^\Lambda = \Lambda^{-1/(1+\alpha_i)}$  otherwise.

Assumption 2 requires that, in the vicinity of  $T = 0$ , the demand volume of a firm decreases proportionally to *some* power of its delay guarantee  $T_i$ . The power may be different for different firms. This assumption is satisfied by most known demand models, but may not be satisfied in general; see Examples 1–3 at the end of this section.

The quantifier  $r_i^\Lambda$ , which depends on Assumption 2 through (9), plays an important role in determining the operational regime of a firm. Our informal discussion above suggests that  $r_i^\Lambda$  provides an order-of-magnitude estimate for the service-level choice of firm  $i$ , and by Lemma 1,  $1/r_i^\Lambda$  then provides an estimate of the service-based capacity for that firm  $i$ . This is formally stated and proved in the following theorem.

**THEOREM 1 (REGIME CHARACTERIZATION).** *Suppose that Assumption 2 holds. Let  $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$  be a sequence such that  $(p^\Lambda, T^\Lambda) \in \mathcal{P} \times \Theta$  for all  $\Lambda \geq 0$ . Then,*

$$T_i^{*,\Lambda}(p^\Lambda, T^\Lambda) \sim r_i^\Lambda, \quad i \in \mathcal{J}: \alpha_i > 1, \quad (10)$$

$$T_i^{*,\Lambda}(p^\Lambda, T^\Lambda) = O(r_i^\Lambda), \quad i \in \mathcal{J}: \alpha_i = 1, \quad (11)$$

and

$$T_i^{*,\Lambda}(p^\Lambda, T^\Lambda) = o(r_i^\Lambda), \quad i \in \mathcal{J}: \alpha_i < 1. \quad (12)$$

Furthermore,

- **QED regime:** if  $1/r_i^\Lambda \sim \sqrt{\Lambda}$ , then  $\hat{e}_i(\Lambda_i, T_i^\Lambda) := N_i^\Lambda - \Lambda_i/\mu_i \sim \sqrt{\Lambda}$ , and

$$\begin{aligned} \limsup_{\Lambda \rightarrow \infty} P\{W_i^\Lambda > 0\} &< 1, \quad \text{and} \\ \liminf_{\Lambda \rightarrow \infty} P\{W_i^\Lambda > 0\} &> 0. \end{aligned} \quad (13)$$

- **ED regime:** If  $1/r_i^\Lambda = o(\sqrt{\Lambda})$ , then  $\hat{e}_i(\Lambda_i, T_i^\Lambda) := N_i^\Lambda - \Lambda_i/\mu_i = o(\sqrt{\Lambda})$ , and

$$\lim_{\Lambda \rightarrow \infty} P\{W_i^\Lambda > 0\} = 1. \quad (14)$$

Here,  $\Lambda_i$  is the demand faced by firm  $i$  when the other firms play  $(p^\Lambda, T^\Lambda)_{-i}$  and firm  $i$  plays the best response  $(T_i^{*,\Lambda}(p^\Lambda, T^\Lambda), p_i^{*,\Lambda}(p^\Lambda, T^\Lambda))$ . The random variable  $W_i^\Lambda$  is the steady-state delay at firm  $i$  under this best response.

Interestingly, Theorem 1 implies that even if the market oscillates between different points in  $\mathcal{P} \times \Theta$ , the operational regime of a firm remains unchanged.

Moreover, it shows that whereas the actual choice of service level by a firm depends on the characteristics of all firms in the market, its operational regime—ED, QD, or QED—depends only on its own intrinsic properties as reflected in the quantifier  $r_i^\Lambda$ .

**REMARK 1 (THE QD REGIME).** The QD regime, in which the probability of delay,  $P\{W^\Lambda > 0\}$ , approaches 0 as  $\Lambda \rightarrow \infty$ , does not emerge in our setting. This is a consequence of the structure of the service-level constraints that we use in our model. Specifically, when a firm's service level is defined via  $P\{W_i > T_i\} \leq \phi$  for  $\phi$  that is strictly positive and exogenously given, it can not do better than setting  $T_i = 0$ . In this case, Proposition 1 in Halfin and Whitt (1981) tells us that, to have  $P\{W_i > 0\} \leq \phi$  for  $\phi \in (0, 1)$ , it suffices to use the square-root safety-staffing rule and, in particular, to use a service-based capacity that is proportional to the square root of the demand. Hence, it cannot be optimal for a firm to operate in the QD regime because this requires the service-based capacity to be orders of magnitude greater than  $\sqrt{\Lambda}$ . The framework that we provide in this paper can, however, be applied to alternative settings in which the QD regime may emerge in equilibrium. We expect, for example, that if service levels are defined via guarantees of the form  $E[W_i] \leq T_i$ , the QD regime will emerge as a possible outcome. Indeed, under some demand models it may be optimal for some firms to guarantee an average delay  $T_i^\Lambda$  such that  $T_i^\Lambda = o(1/\sqrt{\Lambda})$ . These firms will operate in the QD regime; see §9 of Borst et al. (2004).

We conclude this section by pointing out some widely used demand models that satisfy Assumption 2. The multinomial logit and the Cobb–Douglas models are two such examples.

**EXAMPLE 1 (THE MNL DEMAND MODEL).** Let

$$\lambda_i(p, T) := \frac{e^{a_i(T_i) - b_i p_i}}{v_0 + \sum_j e^{a_j(T_j) - b_j p_j}}, \quad i \in \mathcal{J}, \quad (15)$$

where  $v_0 > 0$  is a constant and  $a_i(T_i) = a_i - k_i(T_i)^{\alpha_i}$ , for positive constants  $a_i$ ,  $k_i$ , and  $\alpha_i$ . Then, it is easily verified that  $\alpha_i$  in the definition of  $a_i(T_i)$  plays the role of the exponent in Assumption 2. It is important to note that in a market in which the demand experienced by each firm is characterized by the multinomial logit model, some firms may be operating in the ED regime and others may be operating in the QED regime, depending on the sensitivity of the attraction values of each firm to its own service level.

Henceforth, whenever we mention the MNL demand model we are referring to the one in Example 1.

EXAMPLE 2 (DEMAND MODELS WITH TAYLOR EXPANSION AROUND  $T = 0$ ). A large family of models for which Assumption 2 is satisfied are those in which the function  $\lambda_i(p, T)$  has a Taylor series expansion around  $T_i = 0$ . In these cases, expanding around  $T_i = 0$ , we can write

$$\begin{aligned} \lambda_i(p, T_{-i}, x) \\ = \lambda_i(p, T_{-i}, 0) + \sum_{l=1}^k \frac{\partial^l}{\partial T_i^l} \lambda_i(p, T) \Big|_{T_i=0} \cdot x^l + o(x^k). \end{aligned}$$

Assumption 2 is then satisfied with an exponent that corresponds to the first nonzero derivative with respect to  $T_i$  at the point  $T = 0$ , provided that the derivative is uniformly bounded away from 0. Formally, we will have  $\alpha_i = k$  where  $k$  is such that

$$\begin{aligned} \frac{\partial^k}{\partial T_i^k} \lambda_i(p, T) \Big|_{T_i=0} \in (-a, -b), \quad \text{and} \\ \frac{\partial^l}{\partial T_i^l} \lambda_i(p, T) \Big|_{T_i=0} = 0, \quad l < k, \end{aligned} \quad (16)$$

for some  $0 < a < b < \infty$  and for any vectors  $p \in \mathcal{P}$  and  $T_{-i} \in [0, \bar{T}]^{I-1}$ . Of course, Assumption 2 holds also in many examples in which such a Taylor expansion does not exist; see Example 1 above. In that example, a Taylor expansion need not exist when  $a_i(T_i) = a_i - \kappa_i(T_i)^{\alpha_i}$  for a noninteger exponent  $\alpha_i$ .

The following example illustrates a demand model for which all parameter values lead to the exponent  $\alpha_i = 1$  and, consequently, to the QED regime.

EXAMPLE 3 (THE COBB–DOUGLAS DEMAND MODEL). Fix  $i \in \mathcal{I}$  and assume that

$$\lambda_i(p, T) := \frac{v_i(p_i, T_i)}{v_0 + \sum_j v_j(p_j, T_j)},$$

with  $v_i(p_i, T_i) := c_i(\bar{T}/(\bar{T} - T_i))^{-a_i} p_i^{-b_i}$  for strictly positive constants  $a_i$ ,  $b_i$ , and  $c_i$ . Here, it is easily verified that Equation (16) holds with  $k = 1$  so that we always have  $\alpha_i = 1$  in Assumption 2. Consequently, in a market in which all the firms face a Cobb–Douglas demand model, only the QED regime emerges as the equilibrium choice.

Table 1 Three Different Games

Game	Strategy space	Payoff function
Market	$\mathcal{P} \times \Theta$	$\Pi_i^\Delta(p, T)$ as in (7)
Fluid	$\mathcal{P}$	$\bar{\Pi}_i^p(p) := \lambda_i(p, 0) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right)$
Diffusion	$\mathcal{P} \times \Theta$	$\hat{\Pi}_i^\Delta(p, T) := \lambda_i(p, T) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \Lambda f_i^\Delta(T_i)$

#### 4.1. Roadmap for Rest of this Paper

In §§5 and 6 we introduce the *fluid game* and the *diffusion game*, respectively. The fluid game provides a first-order characterization of the market outcome. It also serves as an essential building block in the introduction and characterization of the (more refined) diffusion game. These two approximate games differ from each other, and from the market game of Definition 1, both in terms of the payoff functions and in terms of the strategy spaces. These differences are outlined in Table 1. The exact derivation of the payoff functions as well the definition of the function  $f_i(\cdot)$  in Table 1 appear in §§5 and 6; see Definitions 3 and 4. In passing, it is important to note that the strategy space of the fluid game is only the price domain  $\mathcal{P}$ , and no service-based capacity cost appears in the payoff function. It is also important that, in the diffusion game, the service-based capacity cost for firm  $i$  is replaced by an expression that depends only on the service-level choice of firm  $i$ . This stands in contrast to the market game in which the service-based capacity is  $\hat{e}_i(\Lambda_i(p, T), T_i)$ , and hence depends on the complete vector  $(p, T)$ .

For each of the above approximate games we will characterize the equilibria: a single Nash equilibrium  $(p^*, 0)$  for the fluid game, and a sequence of Nash equilibria  $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$  for the sequence of diffusion games. We will show that, at different levels of precision, these equilibria approximate the outcomes of the original market game. The precision of the approximation will be a function of the quantifiers  $r_i^\Lambda$  identified earlier in this section. Starting the market in one of the approximate equilibria, we will characterize the maximum profitable deviation that a firm will deviate in its payoffs and actions—service level and price.

## 5. A Fluid Game

In this section we introduce the *fluid game*. This is a simplified game in which only the first-order impact of the firms' actions is modeled. That is achieved by replacing the service facilities (the  $M/M/N$  queues) by their fluid approximations. We will establish that the fluid game does indeed provide a first-order approximation to the original game in that the market prices will always lie within some small neighborhood of the fluid-game equilibrium, and the service levels will be close, in a sense, to  $T = 0$ .

### 5.1. Definition and Characterization

**DEFINITION 3 (THE FLUID GAME).** The fluid game is the  $I$ -player game with profit functions

$$\bar{\Pi}_i^p(\cdot) := \lambda_i(p, 0) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right), \quad i \in \mathcal{F},$$

and strategy space  $\mathcal{P}$ .

Note that the fluid game has the original (unscaled) demand functions  $\{\lambda_i(\cdot), i \in \mathcal{F}\}$ . In the fluid game, the players compete only on prices—this is a pure-price competition game in which the strategy space,  $\mathcal{P}_i$ , of each player is a compact subset of  $\mathbb{R}_+$  so that there exist numerous sufficient conditions for the existence and uniqueness of equilibria. Existence of equilibria is guaranteed under the assumption that  $\bar{\Pi}_i^p(\cdot)$  is continuous and quasi-concave with respect to  $p_i$  (see §2.3 of Cachon and Netessine 2004). This sufficient condition is satisfied, for example, for attraction models such as the multinomial logit demand model or the Cobb–Douglas demand model in Examples 1 and 3, respectively. We will assume that there is a unique equilibrium (and later discuss some concrete examples in which this assumption indeed holds). We formally state this requirement in the following assumption.

**ASSUMPTION 3 (EXISTENCE AND UNIQUENESS OF EQUILIBRIUM FOR THE FLUID GAME).** *The fluid game has a unique Nash equilibrium  $p^* := (p_1^*, \dots, p_I^*)$ .*

For the rest of this paper, whenever Assumption 3 holds, we use the notation  $p^*$  when referring to the unique equilibrium of the fluid game.

### 5.2. The Quality of the Approximation

With the restrictions in Assumptions 1 and 3 we show now that the fluid game serves as a first-order approximation for the original market game. Combined, Theorems 2 and 3 show that approximate equilibria exist, and all such equilibria are concentrated in a small neighborhood of  $(p^*, 0)$  with  $p^*$  as in the Nash equilibrium of the fluid game.

**THEOREM 2 (EXISTENCE OF APPROXIMATE EQUILIBRIA).** *Suppose that Assumptions 1 and 3 hold, and let  $\{\epsilon^\Lambda, \Lambda \geq 0\}$  be a sequence of vectors in  $\mathbb{R}_+^I$  such that, for all  $i \in \mathcal{F}$ ,*

$$\frac{\epsilon_i^\Lambda}{\Lambda} \rightarrow 0 \quad \text{and} \quad \liminf_{\Lambda \geq 0} \frac{\epsilon_i^\Lambda}{\sqrt{\Lambda}} > 0. \quad (17)$$

*Then, there exists a sequence  $\{T^\Lambda, \Lambda \geq 0\}$  such that  $T^\Lambda \in \Theta$ ,  $T_i^\Lambda \rightarrow 0$  for all  $i \in \mathcal{F}$  and, for each  $\Lambda$ , the vector  $(p^*, T^\Lambda) = ((p_1^*, T_1^\Lambda), \dots, (p_I^*, T_I^\Lambda))$  is an  $\epsilon^\Lambda$ -Nash equilibrium for the  $\Lambda$ th market game. Moreover,  $T^\Lambda$  can be chosen so that  $T_1^\Lambda = T_2^\Lambda = \dots = T_I^\Lambda$ .*

**THEOREM 3 (FIRST-ORDER CHARACTERIZATION).** *Suppose that Assumptions 1 and 3 hold, and let  $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$  and  $\{\epsilon^\Lambda, \Lambda \geq 0\}$  be such that  $(p^\Lambda, T^\Lambda)$  is an  $\epsilon^\Lambda$ -Nash equilibrium for the  $\Lambda$ th market game and such that (17) holds. Then, as  $\Lambda \rightarrow \infty$ ,*

$$T_i^\Lambda \rightarrow 0, \quad \text{and} \quad p_i^\Lambda \rightarrow p_i^*, \quad i \in \mathcal{F}. \quad (18)$$

Theorems 2 and 3 are driven by economies of scale. Lemma 1 shows that the service-based capacity  $\hat{e}_i(\cdot, \cdot)$  grows at a lower rate than the volume-based capacity, even for small delay guarantees. Consequently, for any sequence  $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$  with  $(p^\Lambda, T^\Lambda) \in \mathcal{P} \times \Theta$ ,  $\hat{e}_i(\Lambda_i, T_i^\Lambda) = o(\Lambda_i)$  and, in turn, the profit functions satisfy the following property:

$$\Pi_i^\Lambda(p^\Lambda, T^\Lambda) = \Lambda_i(p^\Lambda, T^\Lambda) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) + o(\Lambda_i).$$

Due to the relatively low cost of the service-based capacity, firms will choose to provide relatively high service levels (corresponding to small values of  $T_i$ ). Accordingly, we expect that a game with profit functions

$$\begin{aligned} \Pi_i^{\Lambda, P}(p) &:= \Lambda_i(p, 0) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) \\ &= \Lambda \cdot \lambda_i(p, 0) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right), \quad i \in \mathcal{F}, \end{aligned} \quad (19)$$

and strategy space  $\mathcal{P}$  will provide a first-order approximation for the  $\Lambda$ th market game. Division by the common scalar  $\Lambda$  yields the fluid game in Definition 3.

In a sense, Theorems 2 and 3 show that all sequences of approximate equilibria (and in particular, a sequence of Nash equilibria, if such exist) must converge to  $(p^*, 0)$ . Our ultimate goal in this section is to characterize the convergence rate of  $T_i^\Lambda$  to 0 and of  $p_i^\Lambda$  to  $p_i^*$ , and relate the convergence rate to the bounds  $\epsilon^\Lambda$  on the profit functions. Before moving toward that goal, we discuss some practical implications of the fluid game.

**REMARK 2 (INTERPRETING  $\epsilon^\Lambda$  AS THE LEVEL OF SUB-OPTIMALITY).** One may interpret  $\epsilon$  as the level of sub-optimality for a firm if it chooses to price according to the fluid game price equilibrium  $p^*$ . To illustrate this point, consider the special case in which the market consists of a single firm—a monopolist. The implication of Theorem 2 for this special case is that the monopolist cannot increase its profit by more than  $\epsilon^\Lambda$  by deviating, i.e., that

$$\Pi_i^\Lambda(\tilde{p}^\Lambda, \tilde{T}^\Lambda) \leq \Pi_i^\Lambda(p^*, T^\Lambda) + \epsilon^\Lambda \quad (20)$$

for any sequence of prices and service levels  $\{(\tilde{p}^\Lambda, \tilde{T}^\Lambda), \Lambda \geq 0\}$  as long as  $T^\Lambda \rightarrow 0$ . Here,  $\epsilon^\Lambda$  is a sequence such that  $\epsilon^\Lambda/\Lambda \rightarrow 0$  and  $\liminf_{\Lambda \geq 0} \epsilon^\Lambda/\sqrt{\Lambda} > 0$ . In particular, let  $(\tilde{p}^{*,\Lambda}, \tilde{T}^{*,\Lambda})$  be the true optimal decision for this monopolist when the market scale is  $\Lambda$ . Assuming such a solution exists, Equation (20) then implies that

$$\frac{\Pi_i^\Lambda(p^{*,\Lambda}, T^{*,\Lambda}) - \Pi_i^\Lambda(p^*, T^\Lambda)}{\Lambda} \rightarrow 0 \quad \text{as } \Lambda \rightarrow \infty.$$

Hence,  $(p^*, T^\Lambda)$  satisfies the standard notion of fluid-scale asymptotic optimality. In turn, our results with respect to the fluid game are the game theoretic version of the fluid-scale asymptotic optimality for monopolists. In the same spirit, our equilibrium results for the diffusion game in §6 are a generalization of the diffusion-level asymptotic optimality results for monopolists.

**REMARK 3 (SERVICE-LEVEL DIFFERENTIATION).** A fundamental implication of Theorem 2 is that the market is in an approximate equilibria if all firms set their prices according to  $p^*$  and choose a common

(but very good) service level. Although there might be many plausible explanations for the use of *industry standards*, Theorem 2 provides one such explanation in that it shows that following industry standards is not an irrational choice for firms competing on service levels and prices. In particular, firms need not significantly differentiate themselves in terms of service level. This result can also be interpreted as a one-sided decoupling result between prices and service levels (at least at the first order). The companies may set their prices according to  $p^*$ . Once the prices are fixed, a firm can exploit its large-scale efficiency and, in particular, the relative low cost of the service-based capacity to match the service level of the competitor without moving significantly away from the equilibrium.

Theorems 2 and 3 require only Assumptions 1 and 3. Most standard demand models, however, satisfy also Assumption 2, which, when imposed, allows us to improve on the convergence results in these theorems. To this end, recall that

$$r_i^\Lambda := \max \left\{ \frac{1}{\Lambda^{1/(1+\alpha_i)}}, \frac{1}{\sqrt{\Lambda}} \right\},$$

where  $\{\alpha_i, i \in \mathcal{J}\}$  are as in Assumption 2.

**THEOREM 4 (DISTANCE FROM THE FLUID GAME).** *Suppose that Assumptions 1–3 hold. Then, there exists a sequence  $\epsilon^\Lambda = O(1/r_1^\Lambda, \dots, 1/r_1^\Lambda)$  such that, for each  $\Lambda \geq 0$ ,  $(p^*, 0)$  is an  $\epsilon^\Lambda$ -Nash equilibrium for the  $\Lambda$ th market game. Moreover,*

$$T_i^{*,\Lambda}(p^*, 0) \sim r_i^\Lambda, \quad i \in \mathcal{J}: \alpha_i > 1, \quad (21)$$

$$T_i^{*,\Lambda}(p^*, 0) = O(r_i^\Lambda), \quad i \in \mathcal{J}: \alpha_i = 1, \quad (22)$$

and

$$T_i^{*,\Lambda}(p^*, 0) = o(r_i^\Lambda), \quad i \in \mathcal{J}: \alpha_i < 1. \quad (23)$$

Theorem 4 characterizes the best service-level responses to the fluid-game equilibrium  $(p^*, 0)$ . The first part of the theorem strengthens our results in Theorems 2 and 3 by characterizing the growth rate of  $\epsilon^\Lambda$  as a function of  $r^\Lambda$ . Theorem 4 does not yet provide bounds on the price deviations. For that, we will need to impose additional restrictions on the demand models in consideration. We now gradually introduce the concepts and conditions that are required for that

purpose. To this end, given  $p \in \mathcal{P}$ , let  $\psi_i(p_{-i})$  be a best response of player  $i$  (in the fluid game) to prices  $p_{-i}$  of the competitors. If the best response is not unique, we arbitrarily (but consistently) choose one. We put  $\psi(p) := (\psi_1(p_{-1}), \dots, \psi_I(p_{-I}))$ .

By Assumption 3, the vector  $p^*$  is the unique solution to  $p^* - \psi(p^*) = 0$ . One expects that if  $p$  is a point in which no firm  $i$  can significantly improve its profits by deviating from  $p_i$ , then  $p$  should be close to the unique equilibrium  $p^*$ . Combined, Lemmas 2 and 3 identify conditions under which this intuition is valid.

LEMMA 2. *Suppose that the following two conditions hold:*

(C1) *For each  $i \in \mathcal{I}$  and  $T \in \Theta$ , the demand function  $\lambda_i(p, T)$  is twice continuously differentiable in  $p_i$ .*

(C2) *There exists  $\delta > 0$  such that for all  $p \in \mathcal{P}$  and  $i \in \mathcal{I}$ ,*

$$\frac{\partial^2}{\partial^2 p_i} \bar{\Pi}_i^P(p) \leq -\delta.^6$$

*Then, there exists  $\bar{\epsilon} > 0$  such that if  $p \in \mathcal{P}$  and  $\epsilon \in [0, \bar{\epsilon}]^I$  satisfy*

$$\bar{\Pi}_i^P(\psi_i(p_{-i}), p_{-i}) - \bar{\Pi}_i^P(p) \leq \epsilon_i, \quad i \in \mathcal{I}, \quad (24)$$

*they also satisfy*

$$|p - \psi(p)| \leq M\sqrt{\epsilon} \quad (25)$$

*for some constant  $M > 0$  that is independent of  $p$ .*

Lemma 2 implies that, under the proper conditions on the demand model, if  $p$  is a price vector such that no firm can increase its profit significantly (in the fluid game) by unilaterally deviating from  $p$ , then  $p$  should be close to a corresponding best response vector  $\psi(p)$ . Under some conditions, one can take one step further and show that if a vector  $p$  is close to its best response vector  $\psi(p)$ , then it must be close to the unique equilibrium of the fluid game,  $p^*$ .<sup>7</sup> In Lemma 3 we show

<sup>6</sup> This can be imposed directly on the demand function by requiring that  $(\partial/\partial^2 p_i)\lambda_i(p, 0)(p_i - c_i - \gamma_i/\mu_i) + 2(\partial/\partial p_i)\lambda_i(p, 0) < -\delta$ .

<sup>7</sup> Note that the question of whether  $p$  that is close to  $\psi(p)$  must be close to  $p^*$  is essentially a question about the solution to the set of (possibly nonlinear) equations  $p = \psi(p)$ . Indeed, putting  $F_i(p) := p_i - \psi_i(p)$ , what we want is that, if the set of equations  $F(p) = 0$  has a unique solution  $p^*$ , then any  $p$  that satisfies  $\|F(p)\| \leq \epsilon$  will be close to  $p^*$  in a way that is, to some extent, proportional to  $\epsilon$ . Conditions on the function  $F(\cdot)$  that guarantee the validity of such statements appear, for example, in the literature on convergence of algorithms for the solution of nonlinear equations; see, e.g., Gould et al. (2002) and the references therein.

that one such condition is the “diagonal dominance condition,” which requires the following:

(C3) There exists  $C < 1$  such that

$$\sum_{k \in \mathcal{I}} \left| \frac{\partial}{\partial p_k} \psi_i(p_{-i}) \right| \leq C, \quad p \in \mathcal{P}, i \in \mathcal{I}.$$

Example 4 at the end of this section shows how this condition is verified for the case of the multinomial logit demand model by means of the implicit function theorem. Condition (C3) is well known as a sufficient condition for uniqueness of the equilibrium  $p^*$  for the fluid game (see, e.g., Theorem 5 in Cachon and Netessine 2004), but in our context, it gives us more than that. Whenever Condition (C3) holds, we say that the fluid game is *linearly continuous*. This term is motivated by the following lemma.

LEMMA 3 (LINEAR CONTINUITY OF FLUID GAME). *Suppose that (C1)–(C3) hold. Then, for all  $\epsilon \in \mathbb{R}_+^I$  small enough,*

$$|p - \psi(p)| \leq \epsilon$$

*implies that*

$$|p - p^*| \leq \frac{1}{1-C} B^{-1} \epsilon$$

*for the invertible matrix  $B$ , with  $B_{ii} = 0$  and  $B_{ij} = 1$  for all  $i \neq j$ . Consequently, by Lemma 2, there exist constants  $M, \bar{\epsilon} > 0$  such that*

$$\bar{\Pi}_i^P(\psi_i(p_{-i}), p_{-i}) - \bar{\Pi}_i^P(p) \leq \epsilon_i, \quad i \in \mathcal{I}, \quad (26)$$

*for  $\epsilon \in [0, \bar{\epsilon}]^I$  implies that*

$$|p - p^*| \leq \frac{M}{1-C} B^{-1} \sqrt{\epsilon}. \quad (27)$$

Lemma 3 complements Lemma 2 to show that, under the “diagonal dominance condition,” if  $p \in \mathcal{P}$  is such that no firm (in the fluid game) can increase its profits significantly by unilaterally deviating from it, then  $p$  must be close to  $p^*$ . In other words, under conditions (C1)–(C3), bounds in payoff space (as in Equation (26)) imply bounds in action space (as in Equation (27)).<sup>8</sup> Having introduced Conditions (C1)–(C3), we can now extend the bounds in Theorem 4 to include bounds on deviations in both service level and price. The matrix  $B$  that is used in the statement of the theorem is as in Lemma 3.

<sup>8</sup> In the e-companion, we provide a framework for the continuity of the fluid game that replaces Condition (C3) with a more general condition.

**THEOREM 5 (DISTANCE FROM THE FLUID GAME).** Suppose that Assumptions 1–3 hold. Then, there exists a sequence  $\epsilon^\Lambda = O(1/r_1^\Lambda, \dots, 1/r_i^\Lambda)$  such that, for each  $\Lambda \geq 0$ ,  $(p^*, 0)$  is an  $\epsilon^\Lambda$ -Nash equilibrium for the  $\Lambda$ th market game, and Equations (21)–(23) hold. If, in addition, (C1)–(C3) hold, then

$$|p^{*,\Lambda}(p^*, 0) - p^*| = O(B^{-1}\sqrt{\delta^\Lambda}), \quad i \in \mathcal{J}, \quad (28)$$

with  $\delta_i^\Lambda = 1/(\Lambda r_i^\Lambda) + (r_i^\Lambda)^{\alpha_i}$ .

Theorem 5 uses the linear continuity of the fluid game to relate the market game to the fluid game. It adds to Theorem 4 by establishing the price bounds in (28). Conditions (C1)–(C3) are crucial in the proof of this theorem. To hint into how these condition are used in that proof, note that the bounds on the service levels in (21)–(23) allow us to show that the best response prices  $p^{*,\Lambda}(p^*, 0)$  constitute an  $r^\Lambda$ -Nash equilibrium for the fluid game. Namely, that  $|\bar{\Pi}_i(p^{*,\Lambda}(p^*, 0)) - \bar{\Pi}_i(\psi(p^{*,\Lambda}(p^*, 0)))| \leq r_i^\Lambda$  for all  $i \in \mathcal{J}$ . Having established this property, Conditions (C1)–(C3) and Lemma 3 can be used to obtain the price bounds in (28).

Evidently then, Equation (28) builds heavily on Condition (C3) in obtaining the price bounds from the service-level bounds. Fortunately, various demand models satisfy (C3). We conclude this section with one such example.

**EXAMPLE 4 (THE MNL MODEL).** Consider the demand model in (15). The corresponding fluid game has demand functions

$$\lambda_i^p(p) := \lambda_i(p, 0) = \frac{e^{a_i(0)-b_i p_i}}{v_0 + \sum_{j \in \mathcal{J}} e^{a_j(0)-b_j p_j}}, \quad i \in \mathcal{J},$$

and payoff functions

$$\bar{\Pi}_i^p(p) := \frac{e^{a_i(0)-b_i p_i}}{v_0 + \sum_{j \in \mathcal{J}} e^{a_j(0)-b_j p_j}} \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right).$$

Given  $p_{-i}$ , the best response for firm  $i$  satisfies the equation

$$(1 - \lambda_i^p(p_{-i}, \psi_i(p_{-i}))) \left( \psi_i(p_{-i}) - c_i - \frac{\gamma_i}{\mu_i} \right) = \frac{1}{b_i}. \quad (29)$$

In particular, because there exists  $\epsilon > 0$  such that  $\lambda_i^p(p) < 1 - \epsilon$  for all  $p \in \mathcal{P}$ , we have that there exists  $\delta$  such that  $\psi_i(p_{-i}) > c_i + \gamma_i/\mu_i + \delta$  for all  $p_{-i}$ . Using

the implicit function theorem and differentiating (29) with respect to  $p_j$  for  $j \neq i$ , we get

$$\begin{aligned} & -\frac{\partial}{\partial p_j} \lambda_i^p(p_{-i}, \psi_i(p_{-i})) \left( \psi_i(p_{-i}) - c_i - \frac{\gamma_i}{\mu_i} \right) \\ &= \frac{\partial}{\partial p_j} \psi_i(p_{-i}) \left( \frac{\partial}{\partial p_i} \lambda_i^p(p_{-i}, \psi_i(p_{-i})) \right) \left( \psi_i(p_{-i}) - c_i - \frac{\gamma_i}{\mu_i} \right) \\ & \quad - (1 - \lambda_i^p(p_{-i}, \psi_i(p_{-i}))), \end{aligned} \quad (30)$$

where  $(\partial/\partial p_j)\lambda_i^p(p_{-i}, \psi_i(p_{-i}))$  and  $(\partial/\partial p_i)\lambda_i^p(p_{-i}, \psi_i(p_{-i}))$  are the partial derivatives with respect to  $p_j$  and  $p_i$ , respectively, at the point  $p = (p_{-i}, \psi(p_{-i}))$ . Plugging (29) into (30) as well as the derivatives of  $\lambda_i^p(p)$  with respect to  $p_i$  and  $p_j$ , we have that

$$\sum_{j \in \mathcal{J}} \left| \frac{\partial}{\partial p_j} \psi_i(p_{-i}) \right| = \sum_{j \in \mathcal{J}} \left| \frac{b_j \lambda_i^p(p_{-i}, \psi_i(p_{-i})) \lambda_j^p(p_{-i}, \psi_i(p_{-i}))}{b_i (1 - \lambda_i^p(p_{-i}, \psi_i(p_{-i})))} \right|.$$

The right-hand side is strictly less than 1 provided that

$$\frac{\sum_{j \in \mathcal{J}} b_j \lambda_j^p(p)}{b_i} < 1, \quad i \in \mathcal{J}, p \in \mathcal{P}. \quad (31)$$

This condition is the “dominant diagonal condition” for the MNL model. Hence, Condition (C3) holds for the MNL model provided that (31) holds. By differentiating  $\bar{\Pi}_i^p(p)$  twice, one can also verify that (C1) and (C2) hold for the MNL model.

These example conclude the analysis of the fluid game and we turn to the diffusion game.

## 6. A Diffusion Game

The diffusion game that we introduce in this section is more refined than the fluid game and can be interpreted as a second-order approximation for the market game. The sequence of diffusion games is constructed by replacing the service facilities (the  $M/M/N$  queues) by their corresponding diffusion approximations. We will then show that the diffusion game provides an equilibrium characterization that is asymptotically correct in diffusion scale—thus paralleling the diffusion-scale asymptotic optimality results for monopolists.

### 6.1. Definition and Characterization

In constructing the diffusion game, we first use Lemma 4 below to replace the discontinuous service-based capacity function  $\hat{e}_i(\cdot, \cdot)$  with a continuous function. The lemma relies on Borst et al. (2004) in approximating the delay distribution by an expression that uses the asymptotic version,  $\mathbf{P}(\cdot)$ , for the probability of delay as identified in Halfin and Whitt (1981). Hereafter, we put  $R_i(p, T) := \Lambda_i(p, T)/\mu_i$  for  $(p, T) \in \mathcal{P} \times \Theta$ .

**LEMMA 4 (M/M/N LEMMA).** Fix a sequence  $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$  such that, for each  $\Lambda$ ,  $(p^\Lambda, T^\Lambda) \in \mathcal{P} \times \Theta$ . Then, for all  $i \in \mathcal{I}$ ,

$$\hat{e}_i(\Lambda_i, T_i^\Lambda) = \beta_i\left(\sqrt{R_i(p^\Lambda, T^\Lambda)T_i^\Lambda}\right)\sqrt{R_i(p^\Lambda, T^\Lambda)} + o\left(\beta_i\left(\sqrt{R_i(p^\Lambda, T^\Lambda)T_i^\Lambda}\right)\sqrt{R_i(p^\Lambda, T^\Lambda)}\right)$$

where, given  $y \geq 0$ ,  $\beta_i(y)$  is the unique solution to

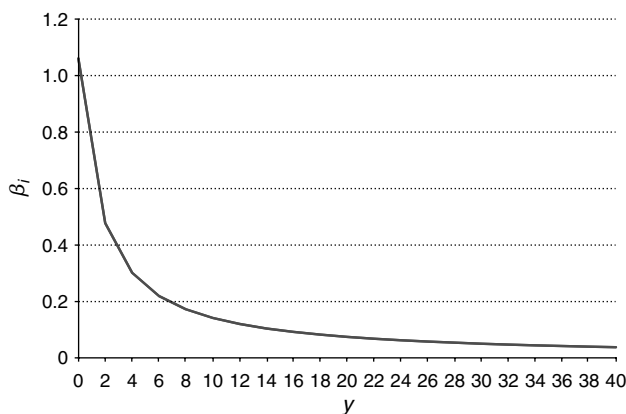
$$\mathbf{P}(x)e^{-\mu_i xy} = \phi.$$

Here,

$$\mathbf{P}(x) = \left[1 + \frac{xZ(x)}{z(x)}\right]^{-1},$$

where  $z(\cdot)$  and  $Z(\cdot)$  are, respectively, the standard normal density function and its cumulative distribution function. Furthermore, the function  $\beta_i(\cdot)$  is a continuously differentiable and convex decreasing function on  $[0, \infty)$ .

**Figure 1** The Function  $\beta_i(\cdot)$  for  $\mu_i = 1$



The function  $\beta_i(\cdot)$  that is defined in Lemma 4 is depicted in Figure 1. Using Lemma 4, we can write

$$\begin{aligned} \Pi_i^\Lambda(p, T) := & \Lambda_i(p, T)\left(p_i - c_i - \frac{\gamma_i}{\mu_i}\right) \\ & - \gamma_i\beta_i\left(\sqrt{R_i(p, T)T_i}\right)\sqrt{R_i(p, T)} \\ & + o\left(\beta_i\left(\sqrt{R_i(p, T)T_i}\right)\sqrt{R_i(p, T)}\right). \end{aligned} \quad (32)$$

The expressions in (32) are still complicated for equilibrium analysis due to the dependence of the  $\sqrt{R_i(\cdot, \cdot)}$  term on the entire vector  $(p, T)$ . Theorem 3 shows, however, that a sequence  $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$  of approximate Nash equilibria must satisfy  $(p^\Lambda, T^\Lambda) \rightarrow (p^*, 0)$  and, by the assumed continuity of the demand functions in Assumption 1, that

$$\frac{\sqrt{R_i(p^\Lambda, T^\Lambda)} - \sqrt{R_i(p^*, 0)}}{\sqrt{R_i(p^\Lambda, T^\Lambda)}} \rightarrow 0 \quad \text{as } \Lambda \rightarrow \infty. \quad (33)$$

These observations motivate the introduction of the diffusion game as an approximation for  $\Lambda$ th market game.

**DEFINITION 4 (THE DIFFUSION GAME).** Fix  $\Lambda \geq 0$ . The  $\Lambda$ th diffusion game has  $I$  players, profit functions

$$\begin{aligned} \hat{\Pi}_i^\Lambda(p, T) := & \Lambda_i(p, T)\left(p_i - c_i - \frac{\gamma_i}{\mu_i}\right) \\ & - \gamma_i\beta_i\left(\sqrt{R_i(p^*, 0)T_i}\right)\sqrt{R_i(p^*, 0)}, \quad i \in \mathcal{I}, \end{aligned}$$

and strategy space  $\mathcal{P} \times \Theta$ .

In defining the new profit functions  $\hat{\Pi}_i^\Lambda(\cdot, \cdot)$ , we have replaced the service-based capacity,  $\hat{e}_i(\cdot, \cdot)$ , by a simpler term that depends on the price equilibrium of the fluid game,  $p^*$ , but is otherwise independent of the actual price vector  $p$  and of the service levels,  $T_{-i}$ , of the competitors. Moreover, by Lemma 4, this term is convex and continuous in  $T_i$ . This relative simplicity renders the diffusion game tractable for Nash equilibrium analysis. For example, it suffices to require that, for each  $i \in \mathcal{I}$ , the demand function  $\lambda_i(p, T)$  is jointly concave in the decision  $(p_i, T_i)$  of firm  $i$ . For future reference, we assign this condition a number:

(C4) For each  $i \in \mathcal{I}$  and each  $(p_{-i}, T_{-i})$ , the demand function  $\Lambda_i(p, T)$  is jointly concave in  $(p_i, T_i)$ .<sup>9</sup>

<sup>9</sup>If the demand functions are twice continuously differentiable, the concavity of  $\lambda_i(p, T)$  in  $(p_i, T_i)$  and the monotonicity in Assumption 1 combined imply the concavity of the function  $\lambda_i(p, T)(p_i - c_i - \gamma_i/\mu_i)$ .

## 6.2. The Quality of the Approximation

We now analyze the quality of the diffusion game as an approximation for the market game. To that end, we need to strengthen Assumption 2 by replacing the lim sup and lim inf there with actual convergence:

(C5) There exists  $0 < \delta \leq \bar{T}$  and a continuous function  $f_i(\cdot, \cdot): \mathcal{P} \times [0, \bar{T}]^{I-1} \rightarrow \mathbb{R}$  such that

$$\lim_{x \rightarrow 0} \frac{\lambda_i(p, T_{-i}, x) - \lambda_i(p, T_{-i}, 0)}{x^{\alpha_i}} \rightarrow f_i(p, T_{-i})$$

for every  $(p, T_{-i}) \in \mathcal{P} \times [0, \delta]^{I-1}$ .

Condition (C5) holds for various demand models; see Example 5 below as well as the linear-demand model in §A1 of the online appendix. With (C5) we have the result in the following theorem. In the statement of the theorem, the matrix  $B$  is as in Lemma 3—it is the  $I \times I$  matrix given by  $B_{ii} = 0$ ,  $i = 1, \dots, I$  and  $B_{ij} = 0$  for all  $i \neq j$ . As before,  $r_i^\Lambda$  is given by

$$r_i^\Lambda := \max \left\{ \frac{1}{\Lambda^{1/(1+\alpha_i)}}, \frac{1}{\sqrt{\Lambda}} \right\}.$$

**THEOREM 6 (DISTANCE FROM THE DIFFUSION GAME).** *Suppose that Assumptions 1 and 3 hold in addition to Conditions (C1)–(C5), and let  $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$  be such that, for each  $\Lambda$ ,  $(p^\Lambda, T^\Lambda)$  is a Nash equilibrium for the  $\Lambda$ th diffusion game. Then, there exists a sequence  $\epsilon^\Lambda = o(1/r_1^\Lambda, \dots, 1/r_I^\Lambda)$  such that  $(p^\Lambda, T^\Lambda)$  is an  $\epsilon^\Lambda$ -Nash equilibrium for the  $\Lambda$ th market game. Moreover,*

$$\begin{aligned} T_i^{*,\Lambda}(p^\Lambda, T^\Lambda) &= T_i^\Lambda + o(r_i^\Lambda), \quad \text{and} \\ p_i^{*,\Lambda}(p^\Lambda, T^\Lambda) &= p_i^\Lambda + o(B^{-1}\sqrt{\zeta^\Lambda}), \end{aligned}$$

where  $\zeta_i^\Lambda = (r_i^\Lambda)^{\alpha_i}$ .

To provide some intuition into the role of Condition (C5) in Theorem 6, assume that a Nash equilibrium  $(p^\Lambda, T^\Lambda)$  does exist for the  $\Lambda$ th market game and recall that, in that case,

$$\begin{aligned} T_i^\Lambda \in \arg \max_{x \in [0, \bar{T}]} & [\Lambda_i((p_i^\Lambda, x): (p^\Lambda, T^\Lambda)_{-i}) \\ & - \Lambda_i((p_i^\Lambda, 0): (p^\Lambda, T^\Lambda)_{-i})] \\ & \cdot \left( p_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, x). \end{aligned}$$

From §4 we know that  $T_i^\Lambda$  will be close to 0; hence, we may heuristically replace the “waiting cost”  $\Lambda_i((p_i^\Lambda, x): (p^\Lambda, T^\Lambda)_{-i}) - \Lambda_i((p_i^\Lambda, 0): (p^\Lambda, T^\Lambda)_{-i})$  by an

approximation of its behavior around 0. Assumption 2 only guarantees that the behavior will be proportional to  $x^{\alpha_i}$ . Although this is sufficient to obtain the relatively crude results of Theorem 4, it is not sufficient for the finer characterization in Theorem 6 above. For this result, we need to identify the functional form of the behavior around  $T = 0$ , and such a functional form is provided by Condition (C5).

Theorem 6 implies that a Nash equilibria,  $(p^\Lambda, T^\Lambda)$ , of the diffusion game provides a refined approximation for the real market outcomes, thus mimicking the role of the diffusion approximation in the context of monopolists. More precisely, by using the diffusion game to determine the firms’ decisions, the compromise in profits is negligible with respect to the cost of the service-based capacity, which is, in turn, proportional to  $1/r_i^\Lambda$ . This is indeed reminiscent of the notion of asymptotic optimality used in the context of monopolists as explained in the following remark.

**REMARK 4 (THE SIZE OF  $\epsilon^\Lambda$  AND DIFFUSION-LEVEL ASYMPTOTIC OPTIMALITY).** Consider the special case in which the set  $\mathcal{F}$  consists of a single firm—a monopolist. Let this be firm 1. An equilibrium of the diffusion game is then a maximizer of  $\hat{\Pi}_1^\Lambda(p_1, T_1)$ , where  $\hat{\Pi}_1^\Lambda(\cdot, \cdot)$  is the profit function in Definition 4. Pick

$$(p_1^\Lambda, T_1^\Lambda) \in \arg \max_{p, T} \hat{\Pi}_1^\Lambda(p, T);$$

i.e.,  $(p_1^\Lambda, T_1^\Lambda)$  is a maximizer of the diffusion-game profit. Theorem 6 implies for this setting that, for any sequence  $\{(\tilde{p}_1^\Lambda, \tilde{T}_1^\Lambda), \Lambda \geq 0\}$  of prices and service levels,

$$\liminf_{\Lambda \rightarrow \infty} \frac{\Pi_1^\Lambda(p_1^\Lambda, T_1^\Lambda) - \Pi_1^\Lambda(\tilde{p}_1^\Lambda, \tilde{T}_1^\Lambda)}{1/r_1^\Lambda} \geq 0,$$

where  $\Pi_1^\Lambda(\cdot, \cdot)$  is the profit function in the  $\Lambda$ th market game; see Definition 1. In other words, the optimality gap for this monopolist, if it chooses to use the outcome of the diffusion game, is of the order of  $o(1/r_1^\Lambda)$ . If the monopolist has  $r_1^\Lambda = 1/\sqrt{\Lambda}$ , then the optimality gap is  $o(\sqrt{\Lambda})$ , which corresponds to the prevalent notion of asymptotic optimality in the Halfin–Whitt (or QED) regime. Thus, the  $\epsilon^\Lambda$ -Nash equilibria result in Theorem 6 reduces, in the monopolist setting, to asymptotic optimality in the sense of Borst et al. (2004).

Theorem 6 does not only provide bounds on the optimality gap with respect to profits, but also with

respect to the price and service-level decisions. Specifically, the theorem shows that, with  $(\tilde{p}_1^\Lambda, \tilde{T}_1^\Lambda)$  being an optimal solution for the monopolist when the market scale is  $\Lambda$ , the sequence  $\{(\tilde{p}_1^\Lambda, \tilde{T}_1^\Lambda), \Lambda \geq 0\}$  must satisfy

$$\tilde{T}_1^\Lambda = T_1^\Lambda + o(r_1^\Lambda) \quad \text{and} \quad \tilde{p}_1^\Lambda = p_1^\Lambda + o(B^{-1}\sqrt{\zeta_1^\Lambda}),$$

where the vector  $(p_1^\Lambda, T_1^\Lambda)$  is the optimal solution to the diffusion game.

It turns out that, under the conditions of Theorem 6, we can be more precise about the service-level characterization. The following lemma shows that the service-level choice under the diffusion-game equilibrium can be characterized in closed form up to an error of size  $o(r_i^\Lambda)$ . Here, the function  $\mathbf{P}(\cdot)$  is as in Lemma 4.

**LEMMA 5.** *Suppose that Assumptions 1 and 3 hold in addition to Condition (C5). Let  $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$  be such that, for each  $\Lambda$ ,  $(p^\Lambda, T^\Lambda) \in \mathcal{P} \times \Theta$   $(p^\Lambda, T^\Lambda) \rightarrow (p^*, 0)$  as  $\Lambda \rightarrow \infty$ . Then,*

$$\frac{T_i^{*,\Lambda}(p^\Lambda, T^\Lambda)}{r_i^\Lambda} \rightarrow \eta_i^* \quad \text{as } \Lambda \rightarrow \infty, i \in \mathcal{J},$$

where  $\eta_i^* = 0$  if  $\alpha_i < 1$  and

$$\eta_i^* = \arg \max_{\eta \geq 0} \eta^{\alpha_i} f_i(p^*, 0) \left( p_i^* - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \tilde{\beta}_i(\eta_i) \left( \frac{\lambda_i(p^*, 0)}{\mu_i} \right)^{1/(\alpha_i+1)}, \quad i \in \mathcal{J},$$

if  $\alpha_i \geq 1$ . Here,  $\tilde{\beta}_i(\eta_i)$  is the unique solution  $x$  to  $\mathbf{P}(x)e^{-\mu_i x (\sqrt{\lambda_i(p^*, 0)/\mu_i})\eta_i} = \phi$  whenever  $\alpha_i = 1$ , and it is the unique solution to  $e^{-\mu_i x (\lambda_i(p^*, 0)/\mu_i)^{1/(\alpha_i+1)}\eta_i} = \phi$  when  $\alpha_i > 1$ .

Lemma 5 allows us to go one step beyond Theorem 6 and replace the requirement of existence of a Nash equilibrium for the diffusion game—Condition (C4)—with a condition on a much simpler “perturbed” fluid game. To this end, let the fluid game on  $T$  be the  $I$ -player game with profit functions

$$\bar{\Pi}_i^{T,p}(p) := \lambda_i(p, T) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right), \quad i \in \mathcal{J},$$

and strategy space  $\mathcal{P}$ . The fluid game on  $T = 0$  is the fluid game from Definition 3. Theorem 7 below provides a characterization of the  $\epsilon^\Lambda$ -Nash equilibrium in terms of the Nash equilibrium of the fluid

game on  $\eta^\Lambda := (\eta_1^* r_1^\Lambda, \dots, \eta_i^* r_i^\Lambda)$  with  $(\eta_1^*, \dots, \eta_i^*)$  as in Lemma 5. Note that, in contrast to Theorem 6, here we do not impose Condition (C4). Instead, we assume uniqueness of equilibrium for the fluid game on  $\eta^\Lambda$ . The matrix  $B$  in the statement of the theorem is as in Lemma 3.

**THEOREM 7.** *Suppose that Assumptions 1 and 3 hold in addition to Conditions (C1)–(C3) and (C5). Assume that, for all  $\Lambda$  large enough, the fluid game on  $\eta^\Lambda := (\eta_1^* r_1^\Lambda, \dots, \eta_i^* r_i^\Lambda)$  has a unique Nash equilibrium  $p^*(\eta^\Lambda)$ . Then, there exists a sequence  $\epsilon^\Lambda = o(1/r_1^\Lambda, \dots, 1/r_i^\Lambda)$  such that  $(p^*(\eta^\Lambda), \eta^\Lambda)$  is an  $\epsilon^\Lambda$ -Nash equilibrium for the  $\Lambda$ th market game. Moreover,*

$$\begin{aligned} T_i^{*,\Lambda}(p^\Lambda(\eta^\Lambda), \eta^\Lambda) &= \eta_i^\Lambda + o(r_i^\Lambda), \quad \text{and} \\ p_i^{*,\Lambda}(p^*(\eta^\Lambda), \eta^\Lambda) &= p_i^*(\eta^\Lambda) + o(B^{-1}\sqrt{\zeta_i^\Lambda}), \end{aligned} \quad (34)$$

where  $\zeta_i^\Lambda = (r_i^\Lambda)^{\alpha_i}$ .

**EXAMPLE 5 (BACK TO THE MULTINOMIAL LOGIT DEMAND).** By Example 4, (C1)–(C3) all hold for the MNL demand model provided that

$$\frac{\sum_{j \in \mathcal{J}} b_j \lambda_j^p(p)}{b_i} < 1, \quad i \in \mathcal{J}, p \in \mathcal{P}. \quad (35)$$

It remains to show that (C5) holds and that a unique equilibrium exists for the fluid game on  $\eta^\Lambda$ . First, we claim that (C5) holds with

$$f_i(p, T_{-i}) := \frac{k_i v_i(p_i, 0)(1 - \lambda_i(p, T_{-i}, 0))}{1 + \sum_{j \neq i} v_j(p_j, T_j) + v_i(p_i, 0)}.$$

The simple but detailed argument is given in the online appendix. The proof can be useful as a guideline toward the verification of (C5) for other demand models. Provided the fluid game on  $T = 0$  satisfies Condition (C3), the fluid game on  $T$  (for  $T$  in a sufficiently small neighborhood of 0) will satisfy Condition (C3) by virtue of the continuity of the best response functions and their derivatives. Condition (C3), in turn, guarantees the uniqueness of equilibria for that game. Because  $\eta^\Lambda \rightarrow 0$  as  $\Lambda \rightarrow \infty$ , we have that the MNL demand model satisfies the conditions of Theorem 7.

**REMARK 5 (HIERARCHICAL DECOUPLING).** Combined, Lemma 5 and Theorem 7 justify referring to demand models that satisfy (C5) as demand models that admit a *hierarchical decoupling*. Indeed, Lemma 5

shows that service-level choices depend on the actions of a firm's competitors mostly through their prices (and not their service levels). Moreover, they depend on these prices only through their fluid-game equilibrium  $p^*$ . Practically, this suggests that service-level and price choices can be made in a sequential manner rather than jointly. The firms will first choose their prices based on the fluid game, i.e., disregarding service-level considerations. Based on these prices the firms will make their service-level choices. Although the firms might choose to adjust their prices at a later stage in response to the actions of the competition, they will not need to revisit their service-level choices. These can remain fixed without any significant compromise to the firms' profits.

**REMARK 6 (GLOBAL STABILITY).** The results stated in this section focus on the existence and, to some extent, uniqueness of approximate equilibria. Accordingly, in the spirit of equilibrium analysis, the focus is on unilateral deviations. In §A3 of the e-companion, we strengthen these results by proving a global stability result. We show that, for any starting point  $(p, T) \in \mathcal{P} \times \Theta$ , the market converges to a neighborhood of the diffusion game equilibrium, and this neighborhood is exactly the one characterized in Theorem 7.

**REMARK 7 (ACTION PREDICTIONS AND FREEDOM IN PRICING).** Note that Theorem 7 is not an if-and-only-if result. It states that a sequence of  $\epsilon^\Lambda$ -Nash equilibria must satisfy the action bounds in (34). It does not say, however, that any sequence that satisfies (34) is a sequence of  $\epsilon^\Lambda$ -Nash equilibria with  $\epsilon^\Lambda = o(r_1^\Lambda, \dots, r_i^\Lambda)$ . The proof of Theorem 7 reveals, however, that any sequence  $(p^\Lambda, T^\Lambda)$  such that

$$|T_i^\Lambda - \eta_i^\Lambda| = o(r_i^\Lambda) \quad \text{and} \quad |p_i^\Lambda - p_i^*(\eta^\Lambda)| = o(\sqrt{(r_i^\Lambda)^{\alpha_i}})$$

is an  $\epsilon^\Lambda$ -Nash equilibrium for  $\epsilon^\Lambda = o(1/r_1^\Lambda, \dots, 1/r_i^\Lambda)$ . There remains a gap then between the "necessary" gap of  $o(B^{-1}\sqrt{(r_i^\Lambda)^{\alpha_i}})$  that is stated in Theorem 7 and the "sufficient" bound of  $o((r_i^\Lambda)^{\alpha_i})$ .

Our asymptotic stability result in §A3 of the e-companion closes this gap by showing that, provided that the fluid game is globally stable, these bounds can be tightened. In that case, we show that

$$p_i^{*,\Lambda}(p^\Lambda(\eta^\Lambda), \eta^\Lambda) = p_i^*(\eta^\Lambda) + o(\zeta_i^\Lambda), \quad (36)$$

with  $\zeta_i^\Lambda = (r_i^\Lambda)^{\alpha_i}$ . Hence, we have that  $(p^\Lambda, \eta^\Lambda)$  constitutes an  $\epsilon^\Lambda$ -Nash equilibrium for

$\epsilon^\Lambda = o(1/r_1^\Lambda, \dots, 1/r_i^\Lambda)$  if and only if  $|p_i^\Lambda - p_i^*(\eta^\Lambda)| = o(\sqrt{(r_i^\Lambda)^{\alpha_i}})$ .

These strengthened price bounds underscore the implication of the operational regime of a firm on the degree of freedom it has in choosing its price. For illustration, consider a market with two firms  $\mathcal{F} = \{1, 2\}$  such that  $\alpha_1 = 1$  and  $\alpha_2 = 2$ . In that case, by (9),  $r_1^\Lambda = 1/\sqrt{\Lambda}$  and  $r_2^\Lambda = 1/\Lambda^{1/3}$ . Theorem 1 states then that firm 1 operates optimally in the QED regime, whereas firm 2 operates optimally in the ED regime. By (36), we then have that firm 1, the QED firm, cannot deviate in price from  $p_i^*(\eta^\Lambda)$  by more than  $o(1/\Lambda^{1/4})$  without compromising its profits. This stands in contrast to firm 2, the ED firm, that can afford deviating from  $p_i^*(\eta^\Lambda)$  by any quantity that is  $o(1/\Lambda^{1/3})$ . Thus, "a QED firm" has to be more precise in setting its price than "an ED firm."

This result is driven by the relative investment in service-based capacity required by QED and ED firms. For example, assume that the two firms considered above decrease their respective prices in a manner that leads to identical increase in demand for both firms. To maintain the same time guarantee, the QED firm will have to increase its capacity by an order of the square root of the increase in demand. The ED firm, in contrast, will be required to make an orders-of-magnitude smaller investment. Consequently, an ED firm can make bigger changes in prices because their impact on its profitability (through the required investment in capacity) will be smaller.

We conclude this section with a brief summary of our results thus far. In §5 we introduced and analyzed the fluid game. Initially, under very mild assumptions on the demand function, we showed that the fluid-game Nash equilibrium provides an  $o(\Lambda)$  approximation for firms' profit and an  $o(1)$  approximation for firms' actions. By introducing additional assumptions on the demand model, we were able to strengthen to bounds and identify how these scale as the market size increases.

In §6.1 we then introduced the more refined diffusion game. The Nash equilibria of the diffusion game provide an asymptotically correct estimate for the market outcome where the asymptotic correctness is analogous to diffusion-scale asymptotic optimality for a monopolist. These results (together with the appropriate assumptions, conditions, and theorems) are summarized in Table 2.

**Table 2** Summary of Results

Game	Equilibrium	Quality of approx. payoff space	Quality of approx. action space	Assumptions and conditions	Theorems
Fluid	$(\rho^*, 0)$	$\epsilon_i^\lambda = o(\lambda)$	$o(1)$	Assumptions 1 and 3	3
Fluid	$(\rho^*, 0)$	$\epsilon_i^\lambda = O(1/r_i^\lambda)$	$O(r^\lambda)$ for service $O(B^{-1}\delta^\lambda)$ for price	Assumptions 1–3 Conditions (C1)–(C3)	4, 5
Diffusion	$(\rho^\lambda, T^\lambda)$	$\epsilon_i^\lambda = o(1/r_i^\lambda)$	$o(r^\lambda)$ for service $o(B^{-1}\sqrt{\xi^\lambda})$ for price	Assumptions 1–3 Conditions (C1)–(C5)	6, 7

## 7. Discussion

In this paper, we study markets with multiple large-scale service providers. To do so, we develop a novel framework that combines the notions of  $\epsilon$ -Nash equilibrium, market replication, and heavy traffic to study market equilibria. The  $\epsilon$ -Nash framework allows us to go beyond the scope of models of competition for which Nash equilibrium exists and use relatively general demand and capacity models. The notion of market replication allows us to discuss trends in the market outcomes along sequences of markets. For example, it allows us to characterize the impact of the market scale on the interdependence between the pricing and service-level decisions. Combined with the notion of heavy traffic, which is well studied for monopolists, this framework allows us to characterize equilibrium behavior and obtain insights for markets in which a Nash equilibrium does not necessarily exist.

The framework developed in this paper can be applied to other competitive settings in which congestion and queueing play important roles. The framework is especially relevant in settings that satisfy two conditions: (a) a Nash equilibrium does not exist or is intractable for characterization, and (b) there are available approximations for the underlying queueing systems that can be used to construct a tractable approximate game. In our setting, the diffusion game—which is based on many-server heavy-traffic approximations—plays the role of this approximate game, but this need not be the case.

Indeed, one can apply the same approach to markets with single-server suppliers in which the service rate, rather than the number of servers, is the capacity decision variable. In these cases, we expect that the so-called *conventional heavy-traffic* approximations—in which the number of servers is kept fixed and the

load approaches one—would play a key role in supplying the approximations that would replace each of the suppliers in the construction of the diffusion game. In these single-server settings our approach can be used to study demand models in which the customers are sensitive to the whole sojourn time rather than solely to the waiting time in queue.

To illustrate this latter claim, we consider a market with two firms such that firm  $i$  faces the following logit demand:

$$\lambda_i(f_1, f_2) = m \frac{a_i e^{bf_i}}{v_0 + a_1 e^{b_1 f_1} + a_2 e^{b_2 f_2}}.$$

Here,  $m$  plays the role of the market size. Also,  $f_i$  is the *full price* “charged” by firm  $i$ , that is,  $f_i = p_i + s_i$ , where  $s_i$  is the average sojourn time of customers served by firm  $i$ . Hence, rather than treating price and service level as independent attributes, this game will consider only their linear combination. Both firms operate through a single server facility so that firm  $i$  adjusts its service rate  $\mu_i$  rather than the number of servers. The cost of capacity is then  $c_i \mu_i$  for  $c_i > 0$ .

$$\Pi_i^m(f_1, f_2) = (f_i - c_i)\lambda_i - 2\sqrt{c_i \lambda_i}, \quad i = 1, 2,$$

where  $c_i$  is the cost of a unit of capacity. This model is very similar (apart from the existence of an outside option) to the one considered in Cachon and Harker (2002).

We use the parameters  $c_1 = c_2 = 3.75$  and  $a_i = -b_i = 1$ , for  $i = 1, 2$ , as in the example of (Cachon and Harker 2002, Figure 3). As in Cachon and Harker (2002), an equilibrium does not exist when  $m = 1$ . For larger values of  $m$ , however, the market seems to have a Nash equilibrium. Still, this Nash equilibrium need not be unique. The fluid game allows us, however, to obtain a first-order approximation of

the full-price equilibrium. Following our approach in this paper, we first define a fluid game by removing the service-based capacity cost  $2\sqrt{c_i\lambda_i}$ . The  $m$ th fluid game is the game with profit functions  $\bar{\Pi}_i^m(f) = (f_i - c_i)\lambda_i^m(f_1, f_2)$ . This fluid game does have the equilibrium  $f^* = (5.725, 5.725)$ . Moreover, it can be easily verified that the diagonal dominance condition in Equation (31) holds for the above parameters, so that this equilibrium is the unique equilibrium of the fluid game.

We numerically compute equilibria for each value of  $m$  for the original game with profit functions  $\Pi_i^m(f)$ . We plot the equilibrium full prices,  $(f_1^m, f_2^m)$ , on the graph in Figure 2. Because the equilibria are all symmetric, with  $f_1^m = f_2^m$ , each such equilibrium is described by a single point on the dashed line. The solid line in Figure 2 corresponds to the fluid-game equilibrium  $f^*$ . Note that the  $y$  axis covers only an interval of size 0.3 so that the convergence of  $f_i^m$  toward  $f^*$  is very quick. For  $m = 100$ , the distance is less than 0.25, which is less than 4%. The gap for the last point in the graph is less than 0.4%.

In this model, in which the service providers are modeled as  $M/M/1$  queues and the competition is only on full price, the diffusion game is identical to the original game, and hence an additional step is not required. Of course, if the arrivals were not Poisson and service times were not nonexponential, the diffusion game and the original game would

no longer be identical. Rather,  $G/G/1$  approximations would be used to construct a diffusion game that provides approximations for the complex original game. Another example of a setting in which our framework is readily applicable is the setting with segmented markets (as considered in Allon and Federgruen 2009), in which each service provider serves multiple customer classes. In this model, we expect that the available approximations for multi-class queueing systems would be used in the construction of the diffusion game.

Yet another model that seems amenable to analysis through our framework (provided that the market under consideration is large) is one in which the competition is incorporated with learning. These are markets in which the demand characteristics as well as the price and service level actions of all the players in the market are not necessarily observable. Large-scale approximations have been used recently in the context of learning and pricing in revenue management (see, e.g., Besbes and Zeevi 2009), and it seems that these can be combined within our framework to characterize the equilibria in these very realistic, but highly intractable, settings.

Finally, in this paper we considered only the case of linear capacity costs. We made this choice so as not to distract attention from the main ideas in the proposed framework. Given our framework, the ability to address more general cost functions, as in Borst et al. (2004), would follow from the ability to do so in a monopolist setting. Hence, the fact that such cost functions are treated in the literature on monopolists suggests that these could also be treated in the competitive setting.

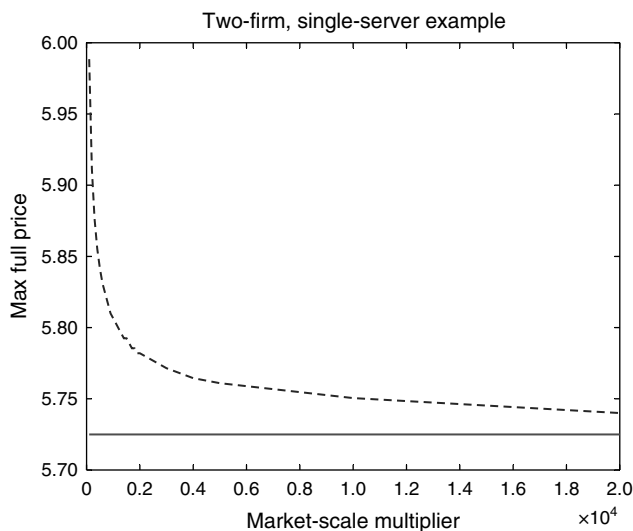
### Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (<http://msom.pubs.informs.org/ecompanion.html>).

### Acknowledgments

The authors are grateful to the reviewers and the associate editor for their numerous comments, which lead to significant improvements over the original manuscript. They also thank Robert Shumsky for helpful discussions in the early stages of this work.

Figure 2 A Single-Server Model



## References

- Allon, G., A. Federgruen. 2007. Competition in service industries. *Oper. Res.* **55**(1) 37–55.
- Allon, G., A. Federgruen. 2009. Competition in service industries with segmented markets. *Management Sci.* **54**(4) 619–634.
- Anton, J. 2001. Call center performance benchmark report. Technical report, Center for Customer-Driven Quality, Purdue University, West Lafayette, IN.
- Besbes, O., A. Zeevi. 2009. Dynamic pricing without knowing the demand function: Risk bounds and near optimal pricing algorithms. *Oper. Res.* Forthcoming.
- Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.
- Cachon, G., S. Netessine. 2004. Game theory in supply chain analysis. D. Simchi-Levi, S. D. Wu, Z. J. Shen, eds. *Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era*. Kluwer Academic Publishers, Boston, 13–59.
- Cachon, P. C., P. T. Harker. 2002. Competition and outsourcing with scale economies. *Management Sci.* **48**(10) 1314–1333.
- Dasci, A. 2003. Dynamic pricing of perishable assets under competition: A two-period model. Working paper, York University, Toronto, Ontario, Canada.
- Dixon, H. D. 1987. Approximate Bertrand equilibria in a replicated industry. *Rev. Econom. Stud.* **54**(1) 47–62.
- Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.
- Gould, N., D. Orban, A. Sartenaer, P. L. Toint. 2002. Component-wise fast convergence in the solution of full-rank systems of nonlinear equations. *Math. Programming* **92**(3) 481–508.
- Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–587.
- Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston.
- Jennings, O., F. de Véricourt. 2008. Dimensioning large-scale membership services. *Oper. Res.* **56**(1) 173–187.
- Kumar, S., R. S. Randhawa. 2009. Exploiting market size in service systems. *Manufacturing Service Oper. Management*. Forthcoming.
- Levhari, D., I. Luski. 1978. Duopoly pricing and waiting lines. *Eur. Econom. Rev.* **11** 17–35.
- Lu, L. X., J. A. Van Mieghem, R. C. Savaskan. 2009. Incentives for quality through endogenous routing. *Manufacturing Service Oper. Management*. **11**(2) 254–273.
- Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Scaling relations and approximate solutions. *Management Sci.* **49**(8) 1018–1038.
- Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* **53**(2) 242–262.
- Tijs, S. H. 1981. Nash Equilibria for noncooperative  $n$ -person games in normal form. *SIAM Rev.* **23**(2) 225–237.
- Whitt, W. 1980. Representation and approximation of noncooperative sequential games. *SIAM J. Control Optim.* **18**(1) 33–48.
- Whitt, W. 2003. How multiserver queues scale with growing congestion-dependent demand. *Oper. Res.* **51**(4) 531–542.