We consider a market with several large-scale service providers that compete on both prices and service levels as measured by the response time to a service request. The literature on many-server approximations provides significant simplifications towards the optimal capacity sizing of large-scale monopolists but falls short of providing similar simplifications for a competition 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. 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 characterization results to provide insights into the price and service level choices in the market and, in particular, into the impact of the market scale on the interdependence between these two strategic decisions.

1. Introduction

In various service industries, an important attribute of the service experience is the delay experienced by customers who are waiting to be served. As a result, customers may consider both the price- and the waiting-time 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 service-levels. We focus on understanding the impact of the market size on the way in which different firms make their pricing and service level choices.
Towards that end, we analyze a competition model 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 needs to 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 so that the task of determining the tradeoff between efficiency and service quality is a non-trivial one.

This tradeoff is a rather complex task 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 vs. waiting-time-related costs is a complex optimization problem. While it can be often solved numerically, numerical solutions fail to provide any structural insights. An alternative to the exact numerical solutions is the use of approximations. The many-server approximations provide a simplified means to approach this problem. 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 §2). The firm’s operational regime dictates how the firm behaves in the face of an increased market size. Some firms 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 waiting time before being served. Some firms are willing to 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 immediately served. An intermediate
regime is the Quality and Efficiency (QED) regime–also known as the Halfin-Whitt regime after the authors that first formalized it (see §2)–that 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 non-trivial fraction (but not all) of the customers begins to 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 choices make the task of identifying equilibria and obtaining quantitative and qualitative results more arduous, the task is further complicated by the fact that the firms demand is not fixed, nor does it depend solely on the firm’s own pricing and service-level choice. 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 the 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) existence: do Nash equilibria exist in the market?, and (b) characterization: given some sort of existence, is it possible to characterize the set of equilibria in order to obtain qualitative insights into the market outcomes? Starting with existence, we note that the concept of Nash equilibria 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, and the simplest supply systems, such as the $M/M/1$ queue (see e.g. Cachon and Harker (2002)). This non-existence 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. Non-existence of Nash equilibria doesn’t 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 and allow us to deal
with relatively general demand functions as well as with 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 non-existence 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 on a single market, but rather on 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 competition 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 can be 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. 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 large-scale approximation. We then relate the equilibria of this new 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 the $\epsilon$-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. While 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, that 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 first implication of this decoupling is the existence of an approximate equilibria in which all firms first set their prices and then set their service levels based on industry standards. In choosing to follow industry standard the firms are only slightly compromising their profitability. Even when imposing greater emphasis on the firm’s optimality level, its service-level choice will still depend on the actions of its competitors mostly through their prices (and not their service levels). The practical implications is that service-level and price choices can be made in a sequential manner rather than jointly. A firm will first choose its price. Based on the chosen price and the initial price choices of the competitors, the firm will make its service-level choice. While the firm might choose to adjust its price at a later stage in response to the actions of the competition, it will not need to revisit its service-level choice. These can remain fixed without any significant compromise to the firm’s profits.

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
Quality and Efficiency Driven (QED) and the Efficiency Driven (ED) regimes can emerge in equilibrium, thus being 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 choice of service levels and prices depends 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 ED regime, firms operating in the QED 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 QED 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 ED 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 disjointed, and some recent work lies at the intersection of the two.

The literature on competition in service industries dates back to the late 1970s. While it initially focused on a single attribute – price or service level (or a simple aggregation of the two), more recent work treats the 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 waiting time 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 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) while most of the literature on competition in services models the supply side via $M/G/1$ queues so that capacity choices are made continuously by adjusting the service rates, we 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 Nash equilibrium intractable for characterization. (ii) In terms of the demand models, 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. While our form of replication is different, our analysis is inspired by his concept. Previous work in game theory has focuses on four types of sequences of games: (i) sequences of games in which the action space is getting finer and finer, and while each game has discrete action space, the limiting game has continuous action space, (ii) sequences of games in which the number of agents grows, (iii) sequence of discrete time games in which the time between periods shrinks to zero along the sequence, and (iv) a sequence of replicated markets with growing market size. We use the fourth framework.

The application of $\epsilon$-Nash in the operations literature is rare. Lu et al. (2007) use this concept in a setting where Nash equilibrium can be shown to exist. The $\epsilon$-Nash equilibrium helps, however, to characterize the equilibrium in a game with a large number of players approaching a continuum of players. Dasci (2003) uses this concept in the context of $\epsilon$-subgame-perfect equilibrium. We are 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). While 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 Gans et al. (2003) and Akşin et al. (2007)) and, more recently, health-care operations.
Halfin and Whitt (1981) were the first to formally identify the QED many-server regime. They consider a sequence of $M/M/N$ queues, all with the same service rate $\mu$. They show that as the demand, $\Lambda$, 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 satisfies a square-root safety staffing rule, i.e,

$$N = R + \beta \sqrt{R} + o(\sqrt{R}),$$ \hspace{1cm} (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 his capacity will utilize his servers very efficiently and, at the same time, have a non-trivial 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 (QED) regime.

While Halfin and Whitt (1981) identified this regime, Borst et al. (2004) placed many-server approximations within a broad economical framework that considers the problem of minimizing capacity and waiting time costs. They show how the QED regime emerges as the optimal economical choice in some cases but also identify conditions under which other regimes, namely the Quality Driven (QD) and Efficiency Driven (ED) regime, 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 in 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 (2008) 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 regime emerge depending on the convexity (or concavity) of the delay-cost function. Similar dependencies will also emerge within the competition setting that we study in this paper.

Other work has also considered large-scale monopolists and obtained staffing and pricing rules in the face of delay-and-price sensitive customers. Notable examples are Armony and Maglaras (2004), Whitt (2003) and Maglaras and Zeevi (2003, 2005).

All of the above results consider a single facility with demand that at most depends on the congestion and price in this single facility. Our paper is the first to show how the different operational regimes emerge in a competition setting and identify the dependencies between operational regimes and pricing choices.
3. The model

We consider a market with a set \( \mathcal{I} = \{1, \ldots, I\} \) of competing service firms, each operating as an \( M/M/N \) facility and serving arriving customers in a First Come First Served (FCFS) manner. Firm \( i \) positions itself in the market by selecting a price \( p_i \) and a waiting-time target \( T_i \). We restrict our attention to service-level guarantees that are given in terms of the customers’ waiting time rather than their whole sojourn time in the system. Having set the waiting-time target \( T_i \), the service provider guarantees that the following Service Level (SL) constraint will be satisfied:

\[
P\{W_i > T_i\} \leq \phi, \tag{2}
\]

where \( W_i \) is the steady-state waiting time and \( 0 < \phi < 1 \) is the satisfaction probability. This form of SL constraint is consistent with the industry practice that commonly uses \( \phi = 0.2 \) (corresponding to 80% of the service requests being answered within target).\(^1\) In this paper we study a competition model where both the prices and the service levels are set simultaneously. We can show that our results continue to hold if the strategies are chosen sequentially (price first or service 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 waiting time of more than a day is unacceptable. Firms choose \( T_i \in [0, \bar{T}] \) and need to adjust their capacity, \( N_i \), so as to guarantee that the SL 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 waiting time in an \( M/M/N \) queue with arrival rate \( \lambda_i \), service rate \( \mu_i \) and \( N \) servers.\(^2\) We write

\[
N_i = R_i + \hat{e}_i(\lambda_i, T_i), \tag{3}
\]

where \( R_i := \lambda_i/\mu_i \) is the offered load given the demand \( \lambda_i \) faced by firm \( i \), and \( \hat{e}_i(\lambda_i, T_i) \) is the excess capacity required to satisfy the service-level target. Naturally, we define \( \hat{e}_i(\lambda_i, T_i) = 0 \)

\(^{1}\)Our results are easily extended to the case where \( \phi \) is allowed to vary between different firms.

\(^{2}\)\( N_i \) 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.
whenever \( \lambda_i = 0 \) but we note that \( \hat{\epsilon}_i(\cdot, \cdot) \) 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 \emph{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 \emph{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.\(^3\) The price \( p_i \) is chosen from a compact interval \([p_i^{\text{min}}, p_i^{\text{max}}], i \in \mathcal{I}\). As each firm will select a price \( p_i \) which results in a non-negative gross profit margin \( p_i - c_i - \gamma_i \mu_i \), we assume, without loss of generality, that

\[
p_i^{\text{min}} = c_i + \frac{\gamma_i}{\mu_i}, \quad i \in \mathcal{I}.
\]  

The upper bound, \( p_i^{\text{max}} \) is allowed to obtain any value in \([p_i^{\text{min}}, \infty)\). We set \( \mathcal{P}_i := [p_i^{\text{min}}, p_i^{\text{max}}] \) 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 waiting time guarantees, i.e., \( \lambda_i = \lambda_i(p, T) \) where \( p = (p_1, \ldots, p_I) \) and \( T = (T_1, \ldots, T_I) \). The following assumption is assumed to hold throughout the rest of the paper.

\begin{assumption}
(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 \( T_i \).
\end{assumption}

Firm-\( i \)'s long-run-average profit \( \Pi_i \), as a function of the prices and service levels in the market, is then given by

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

which, using (3), is re-written as follows:

\[
\Pi_i(p, T) = \lambda_i(p, T) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{\epsilon}_i(\lambda_i, T_i) .
\]  

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

\[
\Lambda_i(p, T) := \Lambda \cdot \lambda_i(p, T),
\]

\[^3\text{see §7 for a discussion of more general capacity-cost models.}\]
be the demand facing firm $i$ in the $\Lambda^{th}$ market. The profit functions in the $\Lambda^{th}$ market are then given by

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

(7)

For future reference we make the following formal definition

**Definition 3.1**  The $\Lambda^{th}$ market game is the $I$-player game with profit functions $\{\Pi^\Lambda_i(\cdot, \cdot), i \in \mathcal{I}\}$ and strategy space $\mathcal{P} \times \Theta$.

As is the case in heavy-traffic analysis, the key idea of our market procedure is to embed the real market (with fixed market size) into a sequence of markets with growing demand. If one is able to get meaningful results for the sequence of markets, these can be applied to the market with fixed size as long as the size is large enough. 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

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

We denote by $T^*_{i,\Lambda}(p, T)$ and $p^*_{i,\Lambda}(p, T)$, respectively, the waiting-time and price components of firm’s $i$ 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.

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$^4$. 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 equilibria is adopted from Tijjs (1981). Rather then defining it in general terms, we provide the definition as it would apply to our setting. To this end, we let

$$ (\hat{p}_i, \hat{T}_i) \uparrow (p, T)_{-i} = ((p_1, T_1), \ldots, (p_{i-1}, T_{i-1}), (\hat{p}_i, \hat{T}_i), (p_{i-1}, T_{i-1}), \ldots, (p_I, T_I)). $$

$^4$It is possible to construct continuous versions of the service-based capacity—see e.g. §4 of Borst et al. (2004). This, however, this would still leave the market game intractable for exact analysis.
Definition 3.2  \( (\epsilon\text{-Nash equilibrium for the } \Lambda^{th} \text{ market game}) \) Fix \( \Lambda \geq 0 \). Let \( \epsilon = (\epsilon_1, \ldots, \epsilon_I) \) 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 \( i \in I \) and any \( \tilde{x}_i \in [p_i^{\min}, p_i^{\max}] \times [0, \bar{T}] \),

\[
\Pi^\Lambda_i(\tilde{x}_i \uparrow x_{-i}) \leq \Pi^\Lambda_i(x) + \epsilon_i.
\]

Note that Nash equilibrium is a special case in which \( \epsilon = 0 \). The \( \epsilon \)-Nash equilibrium is, then, a generalization of Nash equilibrium. The generalization 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 their characterization is hardly tractable. In these cases, and assuming that \( \epsilon \) 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.

Notational conventions and organization of the paper: for two sequences of positive vectors \( \{a^\Lambda\} \) and \( \{b^\Lambda\} \) 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 \( i = 1, \ldots, d \). We say that \( a^\Lambda = o(b^\Lambda) \) if \( \limsup_{\Lambda} a_i^\Lambda/b_i^\Lambda = 0 \) for \( i = 1, \ldots, d \). Finally, we say that \( a^\Lambda \sim b^\Lambda \) if \( a^\Lambda = O(b^\Lambda) \) 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., \( |x| = (|x_1|, \ldots, |x_d|) \). Similarly, the square-root operator should be interpreted componentwise, i.e., for \( x \in \mathbb{R}^d_+ \), \( \sqrt{x} = (\sqrt{x_1}, \ldots, \sqrt{x_d}) \). The notation “\( \rightarrow \)” stands for convergence as \( \Lambda \to \infty \) unless explicitly stated otherwise. We will often use 0 to represent the 0 vector in \( \mathbb{R}^d \) and the dimension of the vector will be always clear from the context. We will also abbreviate by writing \( \Lambda, R_i, T_i^* \) and \( p_i^* \) instead of \( \Lambda_i(p, T), R_i(p, T), T_i^*(p, T) \) or \( p_i^*(p, T) \) when the values of \( (p, T) \) are clear from the context.

The rest of the paper is organized as follows: §4 is concerned with the characterization of the fluid game and the discussion of its implications. In §5, we turn to the diffusion game which is concerned with a refined understanding of the firms’ choices. In §6 we provide a detailed illustration of our results using a linear demand model. Conclusions and directions for future research are discussed in §7. The appendix contains a generalization of the diffusion game framework of §5.
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 e-companion.

4. The fluid game

This section is dedicated to the first-order behavior of the firms in the market. Informally, the first-order characterization corresponds to building a small “envelope” around the sequence of market outcomes. Using the concept of $\varepsilon$-Nash equilibria, we want to be able to say that for all markets large enough, the market prices and service-levels will always lie within some small neighborhood, even if no Nash equilibrium exists. The second-order characterization will then be concerned with obtaining tight bounds on the size of this “envelope”.

Towards that end, we will introduce the fluid game that is motivated by heavy-traffic fluid approximations. In this game we will replace the service facilities (the $M/M/N$ queues) by their fluid counterpart. To motivate the introduction of this game we first argue that the service-based capacity $\hat{e}_i(\cdot,\cdot)$ grows in a lower rate than the volume-based capacity—even for small waiting-time guarantees. This is formalized in the following lemma.

Lemma 4.1 Fix a sequence $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$ with $(p^\Lambda, T^\Lambda) \in \mathcal{P} \times \Theta$. Then, for all $i \in \mathcal{I}$,

$$\frac{\hat{e}_i(\Lambda_i, T^\Lambda_i)}{\Lambda_i} \to 0, \text{ as } \Lambda \to \infty.$$ 

Lemma 4.1 suggests that the profit functions satisfy the following property:

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

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

$$\Pi_i^{\Lambda,P}(p) := \Lambda_i(p, 0) \cdot \left( p_i - c_i - \gamma_i \mu_i \right) = \Lambda \cdot \lambda_i(p, 0) \cdot \left( p_i - c_i - \gamma_i \mu_i \right), \quad i \in \mathcal{I}, \quad (8)$$

and strategy space $\mathcal{P}$ will provide a first-order approximation for the $\Lambda^{th}$ market game. Note that, as division by a common scalar doesn’t change the outcomes of a game, this game is equivalent to
the game with profit functions

$$\Pi^P_i(p) := \lambda_i(p, 0) \cdot \left( p_i - c_i - \gamma_i \frac{\mu_i}{\mu} \right), \ i \in I,$$

(9)

and strategy space $\mathcal{P}$. These observation motivate the introduction of the following game.

**Definition 4.2 (The fluid game)** The fluid game is the $I$-player game with profit functions

$$\Pi^P_i(\cdot) := \lambda_i(p, 0) \cdot \left( p_i - c_i - \gamma_i \frac{\mu}{\mu_i} \right), \ i \in I,$$

and strategy space $\mathcal{P}$.

Note that the fluid game has the original (unscaled) demand functions $\{\lambda_i(\cdot), \ i \in I\}$. The fluid game is a price-competition game, as the players compete only on prices. In particular, the strategy space of each player in this game is a compact subset of $\mathbb{R}_+$ and there exist numerous sufficient conditions for the existence and uniqueness of equilibria. For existence, it suffices to have that $\Pi^P_i(\cdot)$ is continuous and quasi-concave with respect to $p_i$ (see §2.3 of Cachon and Netessine (2004)). This sufficient condition is guaranteed, for example, for attraction models such as the Multinomial Logit demand model or the Cobb Douglas demand model. We will assume that there is a unique equilibrium (and discuss later some concrete examples in which this assumption indeed holds). We formally state these requirements in the following assumption.

**Assumption 4.1 (existence and uniqueness of equilibrium for the fluid game)** The fluid game has a unique Nash equilibrium $p^\ast := (p^\ast_1, \ldots, p^\ast_I)$.

Assumption 4.1 is assumed to hold for the rest of this paper. Henceforth, we will use the notation $p^\ast$ when referring to the unique equilibrium of the fluid game.

The construction of the fluid game was motivated by the prospect that it would serve as a first-order approximation for the original game. Theorems 4.3 and 4.4 below provide the rigorous justification for this intuition.

**Theorem 4.3 (existence)** Let $\{\epsilon^\Lambda, \Lambda \geq 0\}$ be a sequence of vectors in $\mathbb{R}_+^I$ that satisfies

$$\frac{\epsilon_i^\Lambda}{\Lambda} \to 0 \text{ and } \epsilon_i^\Lambda \to \infty, \ i \in I.$$

(10)
Then, there exists a sequence \( T^\Lambda \in \Theta \) with \( T^\Lambda_i \to 0 \) for all \( i \in \mathcal{I} \) such that, for each \( \Lambda \), the vector

\[
(p^*, T^\Lambda) = ((p^*_{1}, T^\Lambda_1), \ldots, (p^*_{I}, T^\Lambda_I)),
\]

is an \( \epsilon^\Lambda \)-Nash equilibrium for the \( \Lambda^{th} \) market game. Moreover, \( T^\Lambda \) can be chosen so that \( T^\Lambda_1 = T^\Lambda_2 = \ldots = T^\Lambda_I \).

**Theorem 4.4 (first-order characterization)** Let \( (p^\Lambda, T^\Lambda) \) be an \( \epsilon^\Lambda \)-Nash equilibrium for the \( \Lambda^{th} \) market game with \( \{\epsilon^\Lambda, \Lambda \geq 0\} \) that satisfies (10). Then, there exists a sequence \( \delta^\Lambda \to 0 \) such that

\[
T^\Lambda_i \in [0, \delta^\Lambda], \quad i \in \mathcal{I},
\]

and

\[
p^\Lambda_i \in [p^*_i - \delta^\Lambda, p^*_i + \delta^\Lambda], \quad i \in \mathcal{I}.
\]

Combined, Theorems 4.3 and 4.4 show that the market prices and service-levels must reside within an increasingly small envelope around \((p^*, 0)\). In particular, they imply that, if a Nash equilibrium \( (p^\Lambda, T^\Lambda) \) exists for the \( \Lambda^{th} \) market game for each \( \Lambda \), then the sequence of these Nash equilibria must converge to \((p^*, 0)\). We note that Theorems 4.3 and 4.4 do not say anything about the convergence rate of \( \delta^\Lambda \) to 0 nor do they relate the quantity \( \delta^\Lambda \) to the bounds \( \epsilon^\Lambda \) on the profit functions. This will be the objective of the analysis of the diffusion game. Before proceeding to the analysis of the diffusion game we would like to discuss the practical implication of the fluid game.

**Remark 4.5 (interpreting \( \epsilon^\Lambda \) as the level of sub-optimality)** One way to interpret \( \epsilon \) is as the level of sub-optimality for a firm if it chooses to price according to the \( \epsilon \)-Nash-equilibrium price. To illustrate this point consider the special case in which the market has a single firm–a monopolist. The implication of Theorem 4.3 for this special case is that the monopolist cannot increase is profit by more than \( \epsilon^\Lambda \) by deviating, i.e,

\[
\Pi^\Lambda_i(p^\Lambda, T^\Lambda) \leq \Pi^\Lambda_i(p^*, T^\Lambda) + \epsilon^\Lambda,
\]

for any sequence of prices and service levels \( \{(p^\Lambda, T^\Lambda), \Lambda \geq 0\} \) as long as \( T^\Lambda \to 0 \). Here \( \epsilon^\Lambda \) is a sequence such that \( \epsilon^\Lambda/\Lambda \to 0 \). In particular, let \( (\tilde{p}^\Lambda, \tilde{T}^\Lambda) \) be the true optimal decision for this monopolist when the market scale is \( \Lambda \). Then, the above implies that

\[
\frac{\Pi^\Lambda_i(p^*, T^\Lambda) - \Pi^\Lambda_i(\tilde{p}, \tilde{T}^\Lambda)}{\Lambda} \to 0 \quad \text{as} \quad \Lambda \to \infty.
\]
This is an instance of the standard notion of fluid-scale asymptotic optimality. Hence, our results with respect to the fluid scale are the game theoretic version of the fluid-scale asymptotic optimality for monopolists and the sequence \( \{ \epsilon^\Lambda \} \) corresponds to the maximal level of suboptimality. In the same spirit, our results for the diffusion game will be a generalization of the diffusion-level asymptotic optimality results for monopolists; see Remark 5.3.

**Remark 4.6 (service-level differentiation:)** A fundamental implication of Theorem 4.3 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. Theorem 4.3 provides one possible explanation (while admitting that other explanations may exist) for the use of industry standards 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. The result can 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.

**5. Second-order analysis: A diffusion game**

In this section we improve on our understanding from the fluid game by introducing and studying a diffusion game. Theorem 4.3 and 4.4 state that the market outcomes \( (p^\Lambda, T^\Lambda) \) converge, as \( \Lambda \to \infty \), to \( (p^*, 0) \). Via a diffusion game we turn now to explicitly identify the convergence rate and, in turn, obtain better estimates for the market outcomes for each \( \Lambda \).

The outline of this section is as follows: we start in §5.1 by motivating and introducing the diffusion game and informally stating our main results with respect to the quality of the approximation that it provides for the original market game. We use the informal statement of our results to discuss some of the qualitative implications of the diffusion game. §5.2 and 5.3 are dedicate to the construction of the pre-requisites for the formal statement of the approximation results. §5.2 focuses on regime characterization and on identifying the order of magnitude of the firms’ service-level choices. §5.3 then provides the remaining structure and is concluded with Theorem 5.11 that formally states that, in a sense, the diffusion game provides good a approximation for the market game.
5.1 The diffusion game

In the monopolist analysis of Borst et al. (2004) a fundamental idea is to replace the original optimization problem that involves integer-valued decision variables with one that is continuous and tractable. Analogously, in this section we introduce a diffusion game that can be interpreted as the game in which the $M/M/N$ service facilities are replaced by their corresponding diffusion approximations. The challenge will be then to relate the equilibria of the diffusion game to those of the original market game.

The following Lemma is the first building block in the construction of the diffusion game.

**Lemma 5.1 (M/M/N Lemma)** Fix a sequence $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$ with $(p^\Lambda, T^\Lambda) \in \mathcal{P} \times \Theta$. Then,

$$
\hat{\epsilon}_i(\Lambda_i, T^\Lambda_i) = \beta_i(\sqrt{R_i T^\Lambda_i}) \sqrt{R_i} + o \left( \beta_i(\sqrt{R_i T^\Lambda_i}) \sqrt{R_i} \right), \quad i \in \mathcal{I},
$$

where, given $T^\Lambda_i$, $\beta_i$ is the unique solution to

$$
P(x)e^{-\mu_i x \sqrt{R_i T^\Lambda_i}} = \phi.
$$

Here

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

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 $[0, \bar{T}]$.

Lemma 5.1 states that the service-based capacity can be written as the sum of $\beta_i(\sqrt{R_i T^\Lambda_i}) \sqrt{R_i}$ and a smaller order term so that the profit functions can be written as follows:

$$
\Pi^\Lambda_i(p, T) := \Lambda_i(p, T) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \beta_i(\sqrt{R_i(p, T) T_i}) \sqrt{R_i(p, T)} + o \left( \beta_i(\sqrt{R_i(p, T) T_i}) \sqrt{R_i(p, T)} \right),
$$

where $R_i(p, T) := \Lambda_i(p, T)/\mu_i$. Even without the smaller order term, the expressions in (13) are complex functions due to the dependence on $p$ and $T$ in the $\sqrt{R_i(\cdot, \cdot)}$ term. By Theorem 4.4 we know, however, that a sequence $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$ of approximate Nash equilibrium must satisfy that $(p^\Lambda, T^\Lambda) \to (p^*, 0)$ and, by the assumed continuity of the demand functions, that

$$
\frac{\sqrt{R_i(p^\Lambda, T^\Lambda)} - \sqrt{R_i(p^*, 0)}}{\sqrt{R_i(p^\Lambda, T^\Lambda)}} \to 0.
$$
These observations motivate the introduction of the diffusion game as an approximation for $\Lambda$th market game for all $\Lambda$ large enough.

**Definition 5.2 (the diffusion game)** Fix $\Lambda \geq 0$. The $\Lambda$th diffusion game has $I$ players, profit functions

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

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

Note that, in defining the new profit function $\hat{\Pi}_i^\Lambda(\cdot, \cdot)$, we have replaced the service-based capacity, $\hat{e}_i(\cdot, \cdot)$, by a simpler term that is independent of the actual price vector $p^\Lambda$ and of the service-levels of the competitors as in $T_{-i}^\Lambda$. Moreover, this term is convex and continuous in $T_i^\Lambda$. This relative simplicity renders the diffusion game tractable for Nash equilibrium analysis in some cases. For example, it suffices to require that, for each $i \in I$, the demand function $\lambda_i(p, T)$ is jointly concave in the decision $(p_i, T_i)$ of firm $i$; see §5.3 for the formal conditions that we impose. The diffusion game was introduced as a means to approximate the original market game. We now proceed to identify the conditions under which it does, indeed, provide a good approximation. Specifically,
we will show that, under certain conditions, if the market game is initialized at a Nash equilibrium of the diffusion game \((p^\Lambda, T^\Lambda)\), profitable deviations will leave all em firms’ actions within a small envelope around \((p^\Lambda, T^\Lambda)\). The challenge is to formally say what “small” means.

First, in §5.2 we will relate the specific structure of the demand model of each firm to its optimal operational regime. The outcome of this initial analysis will be a mapping from firm \(i\)’s demand structure to a quantifier \(y_i^\Lambda\). This quantifier characterizes the order of magnitude of the optimal service-level choice for firm \(i\). Some firms, for example, will have \(y_i^\Lambda = 1/\sqrt{\Lambda}\). 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 \(y_i^\Lambda\) which is significantly larger than \(1/\sqrt{\Lambda}\). These firms will be shown to operate in the ED regime; see the discussion of operational regimes in the introduction.

In §5.3 we state the main result of this section that relates the equilibria of the diffusion game to the original market outcomes through the operational regime quantifiers \(\{y_i^\Lambda, \ i \in \mathcal{I}\}\). Specifically, Theorem 5.11 states that, if \((p^\Lambda, T^\Lambda)\) is an equilibrium for the \(\Lambda^{th}\) diffusion game, then it is also an \(\epsilon^\Lambda\)-Nash equilibria for the \(\Lambda^{th}\) market game with \(\epsilon^\Lambda = o\left(1/y_1^\Lambda, \ldots, 1/y_I^\Lambda\right)\).

In other words, by using the diffusion game to determine the firms’ decision, the compromise in profits is negligible with respect to the cost of the service-based capacity which is, in turn, proportional to \(1/y_i^\Lambda\) (see Lemma 5.5). This is reminiscent of the notion of asymptotic optimality used in the context of monopolists; see Remark 5.3 below.

Moreover, Theorem 5.11 will provide bounds on the maximal profitable deviation in actions. We will show that if \((p^\Lambda, T^\Lambda)\) is an equilibrium for the \(\Lambda^{th}\) diffusion game, then the best responses of firm \(i\) must satisfy that

\[
T_i^{*\Lambda} = T_i^\Lambda + o(y_i^\Lambda), \quad \text{and} \quad p_i^{*\Lambda} = p_i^\Lambda + o\left(B\sqrt{\zeta_i^\Lambda}\right),
\]

where \(\zeta_i^\Lambda = (y_i^\Lambda)^{\alpha_i}\), \(B\) is a certain matrix (defined in §5.3) and \(\zeta_i^\Lambda = (y_i^\Lambda)^{\alpha_i}\).

Thus, initializing the market at the Nash equilibria of the diffusion game, no firm will deviate in its actions more than a quantity that is a function of its own operational regime as reflected in the quantifiers \(y_i^\Lambda\).

**Remark 5.3 (the size of \(\epsilon^\Lambda\) and diffusion-level asymptotic optimality)** The main result of this section can, then, be interpreted as the analogue of the results of Borst et al. (2004) in a competition setting.
To wit, consider the case in which the set $\mathcal{I}$ consists of a single firm, firm 1, which is here a monopolist. The diffusion game then reduces to the problem of maximizing $\hat{\Pi}_i^\Lambda(p_1, T_1)$ where, $\hat{\Pi}_i^\Lambda(\cdot, \cdot)$ is the profit function in the definition of the diffusion game, Definition 5.2. Let
\[
(p_1^\Lambda, T_1^\Lambda) := \arg\max_{p, T} \hat{\Pi}_i^\Lambda(p, T),
\]
i.e $(p_1^\Lambda, T_1^\Lambda)$ is the maximizer of the diffusion-game profit function when the market scale is $\Lambda$ (assuming it exists). The $\epsilon^\Lambda$-Nash equilibria result reduces, in the monopolist setting, to asymptotic optimality in the sense of Borst et al. (2004). Specifically, Theorem 5.11 will imply for this setting that, for any sequence $\{(\tilde{p}_1^\Lambda, \tilde{T}_1^\Lambda), \Lambda \geq 0\}$ of prices and service levels
\[
\frac{\Pi_i^\Lambda(p_1^\Lambda, T_1^\Lambda) - \Pi_i^\Lambda(\tilde{p}_1^\Lambda, \tilde{T}_1^\Lambda)}{1/y_1^\Lambda} \to 0 \text{ as } \Lambda \to \infty,
\]
where $\Pi_i^\Lambda(\cdot, \cdot)$ is now the profit function in the original $\Lambda$th market game; see Definition 3.1. 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 $o(1/y_1^\Lambda)$. If the monopolist has $y_1^\Lambda = 1/\sqrt{\Lambda}$, then the optimality gap is $o(\sqrt{\Lambda})$ which corresponds to the prevalent optimality gap in the literature that considers asymptotic optimality in the Halfin-Whitt regime.

We emphasize that our main result is stronger than asymptotic optimality. We not only provide bounds on the optimality gap with respect to profits but also with respect to the price and service-level decisions. Indeed, Theorem 5.11 shows that, with $(\tilde{p}_1^\Lambda, \tilde{T}_1^\Lambda)$ being an optimal solution for the monopolist when the market scale is $\Lambda$, then the sequence $\{(\tilde{p}_1^\Lambda, \tilde{T}_1^\Lambda), \Lambda \geq 0\}$ must satisfy that
\[
\tilde{T}_1^\Lambda = T_1^\Lambda + o(y_1^\Lambda) \quad \text{and} \quad \tilde{p}_1^\Lambda = p_1^\Lambda + o(B\sqrt{\xi_1^\Lambda}),
\]
where the vector $(p_1^\Lambda, T_1^\Lambda)$ is the optimal solution to the diffusion game.  

Remark 5.4 (level of sub-optimality and freedom in pricing) Interpreting $\epsilon_i^\Lambda$ as the level of sub-optimality for firm $i$ (see Remarks 4.5 and 5.3), Theorem 5.11 provides an insight into the relations between the operational regime of a firm and its pricing decision under a given sub-optimality level. The theorem states that the sub-optimality level of $o(1/y_1^\Lambda)$ is preserved as long as the price distance from the diffusion-game equilibrium price $p_1^\Lambda$ is of order $o((y_1^\Lambda)^{\alpha_i})$ where $\alpha_i$ is the exponent from Assumption 5.1; see §5.2.

To be concrete, consider a firm that operates optimally in the QED regime and has, in particular, $y_1^\Lambda = 1/\sqrt{\Lambda}$. This firm can replace the diffusion game price equilibrium $p_1^\Lambda$ with the price $\tilde{p}_i^\Lambda =$
\( p^λ_i + 1/\Lambda^{1/5} \) without compromising its level of optimality which will remain \( o(\sqrt{\Lambda}) \). In contrast, a firm that operates optimally in the ED regime (which is equivalent to having \( 1/y^λ_i < \sqrt{\Lambda} \)) and has, for example, \( y^λ_i = 1/\Lambda^{1/3} \) will compromise its profits by taking a similar action and will distance itself from optimality by more than \( o(1/y^λ_i) \). In other words, the QED firm has a larger interval from which it can choose its prices without compromising the level of optimality. ■

We now start building the required framework towards the formal statements on the quality of the diffusion-game as an approximation for the original game. These formal statements then appear in Theorems 5.11 and 5.13 which close §5.3.

**5.2 Regime characterization**

A firm chooses its service level trading off the capacity cost and the value of the service level. For a monopolist, this tradeoff 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 tradeoff more subtle.

We will show, however, that, while the actual choice of service level by a firm depends on the characteristics of all firms in the market, its operational regime—Efficiency Drive (ED), Quality Driven (QD) or Quality and Efficiency Driven (QED)—depends only on its own intrinsic properties. Hence, we starts with the regime characterization. Having identified the regimes, we will turn in §5.3 to the explicit characterization of the service levels and prices via a diffusion game.

We henceforth use \((p, T_{-i}, T_i)\) to denote the vector \((p_i, T_i) \uparrow (p, T)_{-i}\). Assume, for sake of the discussion, that \((p^λ, T^λ)\) is a Nash equilibrium in the \(\Lambda^{th}\) market game. Then, the definition of Nash equilibrium implies that

\[
T^λ_i = \arg\max_{x \in [0, T]} \Pi_i^λ(p^λ, T^λ_{-i}, x) = \arg\max_{x \in [0, T]} \Lambda_i(p^λ, T^λ_{-i}, x) \left( p^λ_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, x).
\]

After centering we write

\[
T^λ_i := \arg\max_{x \in [0, T]} \left[ \Lambda_i(p^λ, T^λ_{-i}, x) - \Lambda_i(p^λ, T^λ_{-i}, 0) \right] \left( p^λ_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, x). \tag{15}
\]

It is clear, then, that the order of magnitude of \( T^λ_i \) is determined by optimally balancing the loss of market share due to customer delays, given by \( \Lambda_i(p, T_{-i}, x) - \Lambda_i(p, T_{-i}, 0) \) and the service-based
capacity cost $\gamma_i\hat{e}_i(L_i, x)^5$. For order-of-magnitude analysis it suffices to replace these two elements by crude approximations. Towards that end, we make Assumption 5.1. This assumption states that the demand function is “nicely” behaved in the vicinity of the best waiting-time guarantee $T = 0$. It requires that, as a function of the service level, the demand volume of a firm decreases proportionally to some power of its waiting-time guarantee $T_i$. The power may be different for different firms.

**Assumption 5.1 (behavior around $T = 0$)** For each $i \in \mathcal{I}$ there exists $\alpha_i > 0$ such that
\[
\limsup_{x \to 0} \sup_{p, T - i} \frac{\lambda_i(p, T - i, 0) - \lambda_i(p, T - i, x)}{x^{\alpha_i}} < \infty, \text{ and } \liminf_{x \to 0} \inf_{p, T - i} \frac{\lambda_i(p, T - i, 0) - \lambda_i(p, T - i, x)}{x^{\alpha_i}} > 0.
\]

We impose this assumption for the rest of the paper and we will make recurrent references to the exponents $\alpha_i$. The reader should always interpret them as the exponents arising from this assumption.

Assumption 5.1 plays an important role in identifying the operational regime of a firm. It provides a crude estimate of the rate at which a firm looses market share due to customer delays. This has a direct implications on the order of magnitude of the optimal service level for a firm and, in turn, on its operational regime as shown in Theorem 5.6.

Assumption 5.1 is satisfied by most known demand models, but may not be satisfied in general. The multinomial logit and the Cobb-Douglas demand models are two widely used demand models that satisfy this assumption; see Examples 5.1 and 5.2. A large family of models for which Assumption 5.1 is easily satisfied are the models in which the function $\lambda_i(p, T)$ has a Taylor series expansion around $T_i = 0$. In this cases, using Taylor series expansion around $T_i = 0$ we can write
\[
\lambda_i(p, T - i, x) = \lambda_i(p, T - i, 0) + \sum_{l=1}^{k} \frac{\partial^l}{\partial^l T_i} \lambda_i(p, T) \bigg|_{T_i=0} x^l + o(x^k).
\]

Assumption 5.1 will be satisfied with an exponent that corresponds to the first non-zero 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 for any vector $p$ and $T - i$
\[
\frac{\partial^k}{\partial^k T_i} \lambda_i(p, T) \bigg|_{T_i=0} \in (-\alpha_i, -b), \text{ and } \frac{\partial^l}{\partial^l T_i} \lambda_i(p, T) \bigg|_{T_i=0} = 0, \quad l < k,
\] (16)

\footnote{Here, the loss of market share parallels the role of the waiting-time cost in the monopolist setting of Borst et al. (2004).}
for some $0 < a < b < \infty$. Of course, Assumption 5.1 holds also in many examples in which such a Taylor expansion doesn’t exist; see Example 5.1.

Returning to equation (15) we see that Assumption 5.1 provides an order-of-magnitude handle on the first component of the profit function—the “waiting time cost”. The following Lemma provides a crude approximation for the service-based capacity.

**Lemma 5.5** Fix a sequence $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$ with $(p^\Lambda, T^\Lambda) \in P \times \Theta$ for all $\Lambda \geq 0$. Then,

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

Incorporating Assumption 5.1 and Lemma 5.5 into (15), we see (and will rigorously prove) that the order of magnitude of $T_i^\Lambda$ is determined by solving the optimization problem

$$y_i^\Lambda := \arg\min_{x \in \left[ \frac{1}{\sqrt{\Lambda}}, \bar{T} \right]} \Lambda x^{\alpha_i} + \frac{1}{x}.$$

The optimal solution to this optimization problem is given by

$$y_i^\Lambda = \max \left\{ \frac{1}{\Lambda^{1/\alpha_i}}, \frac{1}{\sqrt{\Lambda}} \right\}, \quad (17)$$

so that $y_i^\Lambda = 1/\sqrt{\Lambda}$ for all $\alpha_i \leq 1$ and $y_i^\Lambda = \Lambda^{-1/\alpha_i}$ otherwise. The reciprocal $1/y_i^\Lambda$ should be interpreted as the order of magnitude of the service-based capacity. In turn, in complete accordance with the literature on monopolists, firms with $y_i^\Lambda \sim 1/\sqrt{\Lambda}$ will operate, in a sense, in the QED regime. This is formally stated in the following theorem that relates $y_i^\Lambda$ to the firms’ operational regime. It is followed by two examples for concrete demand models.

**Theorem 5.6** (regime characterization) Assume that, for each $\Lambda$, a Nash equilibrium $(p^\Lambda, T^\Lambda)$ exists for the $\Lambda^{th}$ market game. Let $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$ be a sequence of Nash equilibria and, for each $\Lambda$, let $W_i^\Lambda$ be the steady-state waiting time at firm $i$ under the equilibrium $(p^\Lambda, T^\Lambda)$. Then,

- **QED regime:** if $\frac{1}{y_i^\Lambda} \sim \sqrt{\Lambda}$ then

  $$\limsup_{\Lambda \to \infty} P\{W_i^\Lambda > 0\} < 1, \quad \text{and} \quad \liminf_{\Lambda \to \infty} P\{W_i^\Lambda > 0\} > 0.$$
• **ED regime:** If \( \frac{1}{y_i} = o(\sqrt{\Lambda}) \) then

\[
\lim_{\Lambda \to \infty} P\{W_i^\Lambda > 0\} = 1.
\]

Theorem 5.6 assumes the existence of Nash equilibrium. A stronger result, that does not require this assumption, appears in Theorem 5.11. We now illustrate Assumption 5.1 and the corresponding regime-characterization results using two well-known demand models.

**Example 5.1 (the multinomial logit (ML) demand model)** Fix \( i \in \mathcal{I} \) and assume that

\[
\lambda_i(p, T) := \frac{e^{a_i(T_i) - b_i p_i}}{v_0 + \sum_j e^{a_j(T_j) - b_j p_j}},
\]

where \( v_0 > 0 \) is a constant and \( a_i(T_i) = a_i - k_i(T_i)^{\alpha_i} \), for \( a_i, k_i \) and \( \alpha_i \) being positive constants. Then, it can be easily verified that \( \alpha_i \) in the definition of \( a_i(T_i) \) plays the role of the exponent in Assumption 5.1. We note that, whenever \( \alpha_i \) is not an integer, a Taylor expansion as in (16) need not exist although Assumption 5.1 holds for some exponent. It is also important to note that in a market in which the demand experienced by each firm is characterized by the Multinomial Logit model, firms may be operating under either the ED regime or the QED, depending on the sensitivity of the attraction values of each firm to its own service-level.

Henceforth, whenever we mention the ML demand model we will be referring to the one in Example 5.1. The following example illustrates a degenerate demand model for which all parameter values lead to the exponent \( \alpha_i = 1 \) and, consequently, to the QED regime.

**Example 5.2 (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 \( v_i(p_i, T_i) := c_i \left( \frac{T_i}{T_i - T} \right)^{-a_i} p_i^{-b_i} \) for strictly positive constants \( a_i, b_i \) and \( c_i \). Here, it is easy to verify that equation (16) holds with \( k = 1 \) so that we always have, \( \alpha_i = 1 \) in Assumption 5.1. Consequently, in a market in which all the firms face a Cobb-Douglas demand model, only the QED regime will emerge as the optimal choice.

24
Remark 5.7 (the Quality-Driven (QD) regime) The QD regime, in which the probability of delay, \( P\{W^\Lambda > 0\} \), approaches 0 as \( \Lambda \to \infty \) doesn’t emerge in our setting. This is a consequence of the special 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 cannot do better than setting \( T_i = 0 \). In this case, Proposition 1 in Halfin and Whitt (1981) tells us that, in order 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 which requires a service-based capacity which is 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 does emerge in competition. 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. ■

We remind the reader that our ultimate goal is to go beyond regime characterization to explicitly characterize the price and service level choices. This is the subject of the next subsection.

5.3 The quality of the approximation

In this section we show that the sequence of diffusion games provides an approximation for the sequence of market games. The approximation is in the sense that, for all \( \Lambda \) large enough, a Nash equilibrium of the diffusion game is an \( \epsilon^\Lambda \)-Nash equilibria for the \( \Lambda^{th} \) market game and \( \epsilon^\Lambda \) is a sequence whose decay rate to 0 is faster than that of \( 1/y^\Lambda \).

First, to allow us to discuss equilibria of the diffusion game we impose the following two conditions.

\[(C1) \text{ for each } i \in \mathcal{I}, \text{ and each } (p_{-i}, T_{-i}), \text{ the demand function } \Lambda_i(p, T) \text{ is jointly concave in } (p_i, T_i). \]

For some of our results we will also require the following differentiability condition:

\[(C2) \text{ for each } i \in \mathcal{I}, \text{ the demand function } \Lambda_i(p, T) \text{ is twice continuously differentiable in } p_i. \]

\(^{6}\)By direct differentiation and using the monotonicity properties of the demand function it can be verified that the concavity of \( \lambda_i(p, T) \) in \( (p_i, T_i) \) implies that of the function \( \lambda_i(p, T)(p_i - c_i - \gamma_i/\mu_i) \).
Condition (C1) guarantees that Nash equilibria exist for the diffusion game. Our most general results require a significantly less stringent restriction on the demand model (see Remark 5.15). We choose to impose it for now to simplify the presentation and discussion of the results.

The definition of $\epsilon$-Nash equilibrium is such that $\epsilon^\Lambda$ provides only a bound on the magnitude of deviations in profits. As we want to obtain estimates on the price and service-level choices, we will have to bound the distance between the real choices of the firms in the market and the choices suggested by the Nash equilibria of the diffusion game. In this section we introduce some notions and results that allow us to translate the bounds from payoff (or profit) space, that we obtain from the $\epsilon$-Nash equilibrium, to bounds in the action space $P \times \Theta$. We cannot perform such translations for the most general models and we thus need to impose certain conditions. Interestingly, however, it suffices to impose conditions on the fluid game which is, as we recall, much simpler than both the original market game and the diffusion game as its strategy space consists only of the price space $P$. For example, we will show that it suffices that the fluid game satisfies a commonly used “Diagonal Dominance” condition for us to be able to perform our bounds-translation.

We will now gradually introduce the notions and conditions that are required for the main result of this section. The conditions that we introduce hold for various demand models and we will illustrate them using the ML demand model in Examples 5.3 and 5.4. To this end, let $\psi_i(p_{-i})$ be the best response of player $i$ (in the fluid game) to prices $p_{-i}$ of the competitors and set

$$\psi(p) := (\psi_1(p_{-1}), \ldots, \psi_I(p_{-I})).$$

The vector $p^*$ is, then, 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 $p^*$. Unfortunately, without imposing additional conditions only the following can be proved.

**Lemma 5.8** Suppose that

(C3) there exists $\delta > 0$ such that for all $p \in P$ and $i \in I$,

$$\frac{\partial^2}{\partial^2 p_i} \Pi_i^P(p) < -\delta.$$ \footnote{This condition can be imposed directly on the demand function by requiring that $\frac{\partial}{\partial p_i} \lambda_i(p, 0)(p_i - c_i - \gamma_i/\mu_i) + 2\frac{\partial}{\partial p_i} \lambda_i(p, 0) < -\delta.$}

There exists constants $C, \bar{\epsilon} > 0$ such that, if $p \in P$ satisfies

$$\Pi_i^P(\psi_i(p_{-i}), p_{-i}) - \Pi_i^P(p) \leq \epsilon_i, \ i \in I \quad (19)$$

26
for some \( \epsilon \in [0, \bar{\epsilon}] \), it must also satisfy that
\[
|p - \psi(p)| \leq C\sqrt{\epsilon}.
\] (20)

Equation (19) states that no firm can increase its profit by using its best response to \( p \). Without additional conditions we only proved that \( p \) would be close to \( \psi(p) \). We would like to show, however, that (19) implies that \( p \) is close to \( p^* \). The next step is, then, to identify conditions under which (20) implies that \( p \) is close to \( p^* \).

We note that this is a question about the solution to a set of (possibly non-linear) equations. By setting \( F_i(p) := p_i - \psi_i(p) \), 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 a statement appear, for example, in the literature on convergence of algorithms for the solution of non-linear equations; see e.g. Gould et al. (2002) and the reference therein.

In some cases, one can characterize such conditions explicitly. Fortunately, in the context of the fluid game, one such condition is the well known “Diagonal Dominance Condition”. The “Diagonal Dominance Condition” stipulates that

(C4) there exists \( C < 1 \) such that
\[
\sum_{k \in I} \left| \frac{\partial}{\partial p_k} \psi_i(p_{-i}) \right| \leq C, \ p \in \mathcal{P}, \ i \in I.
\]

This condition is sufficient for uniqueness of equilibrium \( p^* \) for the fluid game (see e.g. Theorem 5 in Cachon and Netessine (2004)). The following Lemma complements Lemma 5.8 to show that the “Diagonal Dominance Condition” is sufficient (at least for the fluid game) to translate bounds from payoff space to strategy space.

**Lemma 5.9 (linear-continuity of fluid game)** Suppose that (C4) holds. 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, if (C3) holds, there exist constant $C_1, \bar{\epsilon} > 0$ such that,

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

(21)

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

$$|p - p^*| \leq C_1 \sqrt{\epsilon}.$$

We note that Lemma 5.9 is limited to the fluid game while we are interested in performing such translations for the market and the diffusion games. Fortunately, as the proofs show, it suffices to have such translation properties for the fluid game to obtain similar properties for the other, more complicated, games. In the appendix we provide a framework for the continuity of the fluid game which covers also non-linear cases and, in particular, one that doesn’t assume (C4). To keep the presentation of our results as clear as possible, we choose to first present all the result for the case in which (C4) does hold and postpone the general case to the appendix. While the best-response function $\psi$ need not be available in closed form, but rather as an implicit function, it often suffices to apply the implicit function theorem in order to verify (C4). This approach is illustrated in the following example for the multinomial logit demand model. It is followed by a Theorem that uses the linear continuity of the fluid model to relate the market game to the fluid game.

**Example 5.3 (the ML model)** Fix $i \in \mathcal{I}$ with the demand model given in (18). The fluid game is, then, a game with demand functions

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

and payoff functions

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

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

$$(1 - \lambda_i^P(p_{-i}, \psi_i(p_{-i}))) \left( \psi_i(p_{-i}) - c_i - \frac{\gamma_i}{\mu_i} \right) = \frac{1}{b_i}.$$ (22)

In particular, as there exists $\epsilon > 0$ such that $\lambda^P_i(p) < 1 - \epsilon$ for all $p \in \mathcal{P}$, we have that there exists $\delta$ such that $\psi_i(p_{-i}) > c_i + \frac{\Delta_i}{\mu_i} + \delta$ for all $p_{-i}$. Using the implicit function theorem and differentiating
(22) with respect to $p_j$ for $j \neq i$, we get

\[
- \frac{\partial}{\partial p_j} \lambda_i^P(p_{-i}, \psi_i(p_{-i})) \left( \psi_i(p_{-i}) - c_i - \frac{\gamma_i}{\mu_i} \right) 
= \frac{\partial}{\partial p_j} \psi_i(p_{-i}) \left( \frac{\partial}{\partial p_i} \lambda_i^P(p_{-i}, \psi_i(p_{-i})) \left( \psi_i(p_{-i}) - c_i - \frac{\gamma_i}{\mu_i} \right) - (1 - \lambda_i^P(p_{-i}, \psi_i(p_{-i}))) \right),
\]

where $\frac{\partial}{\partial p_j} \lambda_i^P(p_{-i}, \psi_i(p_{-i}))$ and $\frac{\partial}{\partial p_i} \lambda_i^P(p_{-i}, \psi_i(p_{-i}))$ are the partial derivatives with respect to $p_j$ and $p_i$ respectively at the point $p = (p_{-i}, \psi_i(p_{-i}))$. Plugging (22) into (23) as well as the derivatives of $\lambda_i^P(p)$ with respect to $p_i$ and $p_j$ we have that

\[
\sum_{j \in I} \left| \frac{\partial}{\partial p_j} \psi_i(p_{-i}) \right| = \sum_{j \in I} \left| \frac{b_j \lambda_i^P(p_{-i}, \psi_i(p_{-i})) \lambda_j^P(p_{-i}, \psi_i(p_{-i}))}{b_i (1 - \lambda_i^P(p_{-i}, \psi_i(p_{-i})))} \right|.
\]

The right hand side is strictly less than 1 provided that

\[
\frac{\sum_{j \in I} b_j \lambda_j^P(p)}{b_i} < 1, \quad i \in I, \quad p \in P.
\]

This condition is the “Dominant Diagonal Condition” for the ML model. By differentiating $\overline{\Pi}_i^P(p)$ twice one can also verify that (C2) and (C3) hold for the ML model. \hfill \blacksquare

We are now ready to present the main results of this section. We start by improving the bounds on the distance from the fluid game that were obtained in §4. We then proceed to provide the bounds on the distance from the diffusion game.

**Theorem 5.10 (distance from the fluid game)** There exists a sequence $\epsilon^\Lambda = O(1/y_i^\Lambda, \ldots, 1/y_I^\Lambda)$ such that $(p^*, 0)$ is an $\epsilon^\Lambda$ Nash equilibrium for all $\Lambda$. Moreover,

\[
T_i^{*,\Lambda}(p^*, 0) \sim y_i^\Lambda, \quad i \in I : \alpha_i > 1,
\]

\[
T_i^{*,\Lambda}(p^*, 0) = O(y_i^\Lambda), \quad i \in I : \alpha_i = 1,
\]

and

\[
T_i^{*,\Lambda}(p^*, 0) = o(y_i^\Lambda), \quad i \in I : \alpha_i < 1.
\]

If, in addition, (C2)-(C4) hold then

\[
|p_i^{*,\Lambda}(p^*, 0) - p^*| = O(B \sqrt{\delta_i^\Lambda}), \quad i \in I,
\]

with $\delta_i^\Lambda = \frac{1}{\Lambda y_i^\Lambda} + (y_i^\Lambda)^{\alpha_i}$.
Theorem 5.10 characterizes the best responses to the fluid-game equilibrium \((p^*, 0)\). The theorem has three parts. Its first part strengthens our results from §4. While before we only stated that \((p^*, 0)\) is an \(\epsilon^\Lambda\) Nash equilibria for \(\epsilon^\Lambda\) that satisfies \(\epsilon^\Lambda/\Lambda \to 0\), here we provide a tighter characterization of the growth rate of \(\epsilon^\Lambda\) as a function of \(\{y_i^\Lambda, i \in \mathcal{I}\}\). The second part of the theorem—equations (25)-(27)—strengthens our results in Theorem 5.6. Specifically, it shows that, even if Nash equilibria do not exist, firm \(i\) uses a service-based capacity which is of the order of magnitude of \(1/y_i^\Lambda\). In particular, firms with \(\alpha_i \leq 1\) operate in the QED regime and firms with \(\alpha_i > 1\) operate in the ED regime. The last part of the Theorem, equation (28), uses (C4) to obtain bounds for the deviation in prices.

Theorem 5.10 implies that it is enough to consider service level choices \(T_i^\Lambda\) that are in a neighborhood of size \(O(y_i^\Lambda)\) of 0. Next we use the diffusion game to improve this result by providing a characterization that is precise up to \(o(y_i^\Lambda)\). A prerequisite is, however, that the diffusion game itself exhibits some sort of uniqueness at the \(y_i^\Lambda\) scale. To illustrate this point, assume that there is a firm \(i\) so that \(\{(p^\Lambda, T_{-i}^\Lambda, \tilde{T}_i^\Lambda), \Lambda \geq 0\}\) and \(\{(p^\Lambda, T_{-i}^\Lambda, \tilde{T}_i^\Lambda), \Lambda \geq 0\}\) are both Nash equilibria for the diffusion game but

\[
\frac{\tilde{T}_i^\Lambda}{y_i^\Lambda} \to \tilde{\eta} \text{ and } \frac{\tilde{T}_i^\Lambda}{y_i^\Lambda} \to \tilde{\eta},
\]

for \(\tilde{\eta} \neq \tilde{\eta}\). Then, in the diffusion game, firm \(i\) can deviate by \(\tilde{T}_i^\Lambda - \tilde{T}_i^\Lambda = (\tilde{\eta} - \tilde{\eta})y_i^\Lambda + o(y_i^\Lambda)\) without compromising its profits. As the market game is, in a sense, close to the diffusion game one cannot expect to obtain estimates that are precise up to \(o(y_i^\Lambda)\).

The following condition strengthens Assumption 5.1 and is designed to rule out such undesired outcomes.

(C5) there exists a continuous function \(f_i(\cdot, \cdot) : \mathcal{P} \times [0, \bar{T}]^{I-1} \to \mathbb{R}\) such that

\[
\lim_{x \to 0} \frac{\lambda_i(p, T_{-i}, x) - \lambda_i(p, T_{-i}, 0)}{x^{\alpha_i}} \to f_i(p, T_{-i})
\]

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

This strengthened assumption holds for various demand models. This is the case, for example, if the demand of each firm \(i \in \mathcal{I}\) has a Taylor expansion around \(T_i = 0\) and satisfies (16). Example 5.4 below shows that (C5) holds for the ML demand model. With (C5) we can obtain the following result on the relations between the original market game and the diffusion game.
Theorem 5.11 (distance from the diffusion game) Suppose that (C1)-(C5) hold and let \((p^\Lambda, T^\Lambda)\), \(\Lambda \geq 0\) be a sequence such that \((p^\Lambda, T^\Lambda)\) is a Nash equilibrium for the \(\Lambda\)th diffusion game. Then, there exists a sequence \(\epsilon^\Lambda = o\left(1/y_i^\Lambda, \ldots, 1/y_I^\Lambda\right)\), such that \((p^\Lambda, T^\Lambda)\) is an \(\epsilon^\Lambda\)-Nash equilibrium for the \(\Lambda\)th market game. Moreover,

\[
T_i^{*\Lambda} = T_i^\Lambda + o(y_i^\Lambda), \quad \text{and} \quad p_i^{*\Lambda} = p_i^\Lambda + o\left(\theta \sqrt{\zeta^\Lambda}\right),
\]

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

Example 5.4 By Example 5.3, (C1)-(C4) all hold provided that

\[
\sum_{j \in I} b_j \lambda_j^P(p) \leq 1, \quad i \in I, \quad p \in P.
\]

It remains to show that (C5) holds. We claim that (C5) holds with

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

The simple, but detailed, argument is given in the e-companion. The proof can be useful as a guideline towards the verification of (C5) for other demand models. Hence, together with Example 5.3 we hence have that the ML demand model satisfies the conditions of Theorem 5.10 and 5.11.

Theorem 5.11 says that the service level and price choices, \((p^\Lambda, T^\Lambda)\) that we get from the diffusion game provide precise approximations for the real market outcomes, thus mimicking the role of the diffusion approximation in the context of monopolists. Under (C5) we can show an even stronger result in terms of the service-level characterization.

Lemma 5.12 Suppose that (C1)-(C5) hold and let \((p^\Lambda, T^\Lambda)\), \(\Lambda \geq 0\) be a sequence such that \((p^\Lambda, T^\Lambda)\) is a Nash equilibrium for the \(\Lambda\)th diffusion game. Then,

\[
\frac{T_i^\Lambda}{y_i^\Lambda} \to \eta_i^*,
\]

where \(\eta_i^* = 0\) if \(\alpha_i < 1\) and it equals

\[
\eta_i^* = \arg \max_{\eta_i \geq 0} \eta_i^{\alpha_i} f_i(p_i^*, 0) \left( p_i^* - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \beta_i(\eta_i) \left( \frac{\lambda_i(p_i^*, 0)}{\mu_i} \right)^{\frac{1}{\alpha_i+1}},
\]

if \(\alpha_i \geq 1\). Here \(\beta_i(\eta_i)\) is the solution \(x\) to \(P(x) e^{-\mu_i x} \sqrt{\frac{\lambda_i(p_i^*, 0)}{\mu_i}} = \phi\) whenever \(\alpha_i = 1\) and is the solution to \(e^{-\mu_i x \left( \frac{\lambda_i(p_i^*, 0)}{\mu_i} \right)^{\frac{1}{\alpha_i+1}}} = \phi\) when \(\alpha_i > 1\).
Lemma 5.12 implies that the service levels under the stated conditions depend only on the fluid game price equilibrium \( p^* \) rather than on the refinement to the price decisions. It also allows us to provide a more explicit characterization of \( \epsilon^\Lambda \)-Nash equilibria in terms of the equilibria of a much simpler fluid game.

Towards that end let \( \eta^\Lambda = (\eta^\Lambda_1, \ldots, \eta^\Lambda_I) \) and let the fluid game on \( T \) be the \( I \) player game with profit functions

\[
\Pi^{T,P}_i(p) := \lambda_i(p, T) \left( p_i - c_i - \frac{y_i}{\mu_i} \right), \quad i \in \mathcal{I},
\]

and strategy space \( \mathcal{P} \). When \( T = 0 \), this is merely the fluid game from Definition 4.2. Theorem 5.13 below provides then a complete characterization of the \( \epsilon^\Lambda \) Nash equilibrium in terms of the Nash equilibrium of the (simpler) fluid game on \( \eta^\Lambda \).

**Theorem 5.13** Suppose that conditions (C1)-(C5) hold. Assume that, for all \( \Lambda \) large enough the fluid game on \( \eta^\Lambda \) has a unique Nash equilibrium \( p^\Lambda(\eta^\Lambda) \). Then, there exists a sequence \( \epsilon^\Lambda = o \left( \frac{1}{y^\Lambda_1}, \ldots, \frac{1}{y^\Lambda_I} \right) \), such that \( (p^\Lambda(\eta^\Lambda), \eta^\Lambda) \) is an \( \epsilon^\Lambda \)-Nash equilibrium for the \( \Lambda \)th market game. Moreover,

\[
T^*_i,\Lambda = \eta^\Lambda_i + o(y^\Lambda_i), \quad \text{and} \quad p^*_i,\Lambda = p^\Lambda_i(\eta^\Lambda_i) + o \left( B \sqrt{\zeta^\Lambda} \right),
\]

where \( \zeta^\Lambda_i = (y^\Lambda_i)^{\alpha_i} \).

We note that the existence of equilibria for the fluid game on \( \eta^\Lambda \) is guaranteed by virtue of condition (C1). Uniqueness, however, has to be established for the specific demand model in consideration.

**Remark 5.14** (hierarchical decoupling) Lemma 5.12 and Theorem 5.13 justify referring to demand models that satisfy (C5) as demand models that admit a hierarchical decoupling. Indeed, Lemma 5.12 shows that service-level choices depend on the actions of its 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. A firm will first choose its price. Based on the chosen price and the initial price choices of the competitors, the firm will make its service-level choice. While the firm might choose to adjust its price at a later stage in response to the actions of the competition, it will not need to revisit its service-level choice. These can remain fixed without any significant compromise to the firm’s profits.
Remark 5.15 (when (C1) doesn’t hold) Theorem 5.13 assumes that condition (C1) hold and, in particular, that $\lambda_i(p, T)$ is jointly concave in $(p_i, T_i)$. Condition (C1) rules out the multinomial logit demand model. Theorem A.7 in the Appendix to this paper provides the generalizations of Theorem 5.13, respectively, to a case in which (C1) is replaced with a much weaker condition that is imposed only on the fluid game. That condition is shown there to be satisfied, in particular, by the multinomial logit demand model.

6. Example: A linear demand model

In this section we provide a numerical example to illustrate the strength of the approximation in §5. The example is based on the linear demand model that was analyzed in Allon and Federgruen (2007). Specifically, we consider the demand model specified by:

$$\lambda_i(p, T) = \left[ a_i(T_i) - b_i p_i + \sum_{j \neq i} a_{ij}(T_j) + \sum_{j \neq i} \beta_{ij} p_j \right]^+. \quad (30)$$

we assume that

$$a_i(T_i) = a_i - k_i T_i \quad \text{and} \quad a_{ij}(T_j) = k_{ij} T_j \quad (31)$$

for strictly positive constants $a_i, k_i, k_{ij}, i, j \in I$. We further assume that a uniform price increase by all $I$ firms cannot result in an increase in any firm’s demand volume and that a price increase by a given firm cannot result in an increase of the industry’s aggregate demand volume, i.e.,

$$(D) \quad b_i > \sum_{j \neq i} \beta_{ij}, i = 1, \ldots, I; \quad (D') \quad b_i > \sum_{j \neq i} \beta_{ji}, i = 1, \ldots, I. \quad (32)$$

The requirements in (32) guarantee that the ”Dominant Diagonal” condition—Condition (C4) holds for the fluid game of the linear demand model. Equation (32) guarantees that the matrix $I \times I$ matrix $A$, defined by $A_{ii} = 2b_i$ and $A_{ij} = -\beta_{ij}$ for $i \neq j$, is invertible and we let $A^{-1}$ be its inverse. Finally, we make the assumption that $a_i(0) + \sum_{j \neq i} \beta_{ij} p_j \min > b_i p_i \min$ so that

$$\lambda_i(p, T) > 0, \forall (p, T) \in \mathcal{P} \times \Theta. \quad (33)$$

The Nash equilibria of the fluid game for the linear demand model can be characterized in closed form. Specifically, the fluid-game best response function for firm $i$ is given by

$$\psi_i(p_{-i}) = \frac{a_i(0) - \sum_{j \neq i} a_{ij}(0) + \sum_{j \neq i} \beta_{ij} p_j + b_i \left( c_i + \frac{\gamma_i}{\mu_i} \right)}{2b_i}. \quad (33)$$
The invertibility of $A$ then guarantees that the fluid game has a unique equilibrium $p^*$ that solves the system $p^* = \psi(p^*)$.

Using (31), it is easily verified that Assumption 5.1 holds for this linear demand model with $\alpha_i = 1$ for all $i \in I$ and, in particular, that $y_i^\Lambda = 1/\sqrt{\Lambda}$ for all $i \in I$ so that, by Theorems 5.6 and 5.10 all firms will operate in the QED regime. Also, it is easily verified that Conditions (C1)-(C3) and well as Condition (C5) hold for the linear demand model and, consequently, that it satisfies the conditions of Theorem 5.11. What renders this linear demand model suitable for our numerical experiments is that the Nash equilibria of the corresponding diffusion game have a simple, closed-form, characterization. Indeed, the equations for the equilibrium of the diffusion game are given by:

$$
(Ap)_i = a_i - k_i T_i + \sum_{j \neq i} k_{ij} T_j + b_i \left( c_i + \frac{\gamma_i}{\mu_i} \right),
$$

(34)

and

$$
- \left( p - c - \frac{\gamma_i}{\mu_i} \right) k_i - \gamma_i \beta_i(\sqrt{R_i(p^*, 0) T_i}) \sqrt{R_i(p^*, 0)} = 0.
$$

(35)

Consequently, we can derive the following result:

**Theorem 6.1 (the diffusion game of the linear demand model)** For each $\Lambda$, a Nash equilibrium exists for the $\Lambda^{th}$ diffusion game. Let $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$ be a sequence of such equilibria. Then,

$$
p_i^\Lambda = p_i^* + \frac{\delta_i}{\sqrt{R_i(p^*, 0)}}, \quad i \in I,
$$

(36)

and

$$
T_i^\Lambda = \frac{\eta_i}{\sqrt{R_i(p^*, 0)}}, \quad i \in I.
$$

(37)

Here, $\delta := (\delta_1, \ldots, \delta_I)$ and $\eta := (\eta_1, \ldots, \eta_I)$ are the unique solution to the system of equations

$$
(A\delta)_i = k_i \eta_i - \sum_{j \neq i} k_{ij} \eta_j, \quad i \in I,
$$

(38)

$$
\gamma_i \beta_i(\eta_i) = k_i \frac{\mu_i}{\lambda_i(p^*, 0)} \left( p_i^* - c_i - \frac{\gamma_i}{\mu_i} \right), \quad i \in I,
$$

(39)

where (given $\eta_i$) $\beta_i(\eta_i)$ is the unique solution of

$$
P(\beta_i(\eta_i)) e^{-\mu_i \beta_i(\eta_i) \eta_i} = \phi.
$$
The numerical example: We consider an industry with $I = 3$ firms, $\bar{T} = 1$, and cost parameters $c_1 = c_2 = 20, c_3 = 5$, and $\gamma_1 = \gamma_2 = 20, \gamma_3 = 35$. This setting can be interpreted as having firm 3 to be an established local service provider and firms 1 and 2 competitors that have entered the local market more recently from a foreign or remote location, where the capacity costs, $\gamma$, are lower but the per-customer access cost, $c$, is higher. We assume that all firms experience identical price sensitivities. Specifically, we assume that $b_i = 10$ for all $i = 1, 2, 3$ and $\beta_{ij} = 4.75, \forall i \neq j$. Finally, we let $a_1(\theta) = 205 + 0.1\theta_1, a_2(\theta) = 205 + 0.1\theta_2, a_3(\theta) = 295 + 0.1\theta_3, \alpha_{ij} = 0.01\theta_j, \forall i \neq j$.

We then solve the first order conditions (38) and (39) to obtain the vectors $\eta$ and $\delta$ and use (36) and (37) to construct the diffusion-game Nash equilibrium $(p^\Lambda, T^\Lambda)$. We then analyze the market outcomes for different values of the market scale $\Lambda$. Specifically, fixing $\Lambda$, we initialize the $\Lambda^{th}$ market game at the point $(p^\Lambda, T^\Lambda)$ and check the maximal profitable deviation of each firm, $T^*_{i \Lambda}(p^\Lambda, T^\Lambda)$ and $p^*_{i \Lambda}(p^\Lambda, T^\Lambda)$.

![Waiting-Time Guarantees of Firm 1](image1)
![Prices of Firm 1](image2)

Figure 6 displays the maximal profitable deviations for firm 1 (the quality of the approximations is similar for the other two firms). The left-hand graph corresponds to profitable deviation in the service-level dimension. Specifically, the black series depicts the sequence the service-level choice in the diffusion game equilibrium, $T_{1 \Lambda}$, as a function of $\Lambda$. For each value of $\Lambda$ we calculate the best response in the $\Lambda^{th}$ market game, $T^*_{1 \Lambda}(p^\Lambda, T^\Lambda)$. The red and blue series are then, respectively, the upper and lower bounds that this best-responses induce, i.e, the $\Lambda^{th}$ point in the red series corresponds to the value $T_{1 \Lambda} + |T^*_{1 \Lambda} - T_{1 \Lambda}|$ and the $\Lambda^{th}$ point in the blue series to the value $T_{1 \Lambda} - |T^*_{1 \Lambda} - x|$. The fact that the red and blue lines are very close to $T_{1 \Lambda}$ is the numerical illustration of our result in Theorem 5.11 that $T^*_{1 \Lambda} = T_{1 \Lambda} + o(T_{1 \Lambda})$. The right hand graph then repeats the same steps for the sequence of prices $p^\Lambda_i$ obtained from the diffusion game and the corresponding best response sequence $\{p^*_{1 \Lambda}(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$. 
We see, then, that profitable deviations from the Nash equilibrium of the diffusion game \((p^A, T^A)\) are very small for both the service-level and the price choices.

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 \(\varepsilon\)-Nash equilibrium, market replication and heavy-traffic to study market equilibria. The \(\varepsilon\)-Nash framework allows us to go beyond the scope of Nash equilibrium and use relatively general demand and capacity models. The notion of market replication allows us to discuss trends in terms of stability and market outcomes in sequences of markets such as 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 the equilibria behavior and obtain insights for markets in which Nash equilibrium need not necessarily exist.

The framework developed in this paper can be applied to other competitive settings in which congestion and queueing play an important role. The framework is especially relevant in settings that satisfy two conditions: (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 setting 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
\[ \lambda_i(f_1, f_2) = m \frac{a_i e^{b_i f_i}}{v_0 + a_1 e^{b_1 f_1} + a_2 e^{b_2 f_2}}. \]
Here \( m \) plays the role of the market size. Also, \( f_i \) is the full price “charged” by firm \( i \). That is \( f_i = p_i + g_i \) where \( g_i \) 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 to the one considered in Cachon and Harker (2002).

We use the parameters \( c_1 = c_2 = 3.75, a_i = -b_i = 1 \) for \( i = 1, 2 \), as in the example in Figure 3 of Cachon and Harker (2002). As in Cachon and Harker (2002), equilibrium doesn’t 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 will allow us, however, to get a first-order approximation of the equilibria full prices. 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 \( \Pi_i^m(f) = (f_i - c_i)\lambda_i^m(f_1, f_2) \). This fluid game does have the equilibrium \( f^* = (5.725, 5.725) \). Moreover, this equilibrium is unique as it can be easily verified the diagonal dominance condition in equation (24) holds for the above parameters.

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 7. As the equilibria are all symmetric, with \( f_1^m = f_2^m \), each such equilibrium is described by a single point on the blue line. The red line in Figure 7 corresponds to the fluid-game equilibrium \( f^* \). Note that the \( y \) axis covers only a interval of size 0.3 so that the convergence of \( f_i^m \) towards \( f^* \) is very quick. For \( m = 100 \), the distance is less than 0.25 which is, in percentage, less than 4%. The gap for the last point in the graph is less than 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 services non-exponential, the diffusion game and the original game would no longer be identical. Rather, \( G/G/1 \) approxima-
tions 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 (2008)) 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 in 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 recently used in the context of learning and pricing in revenue management (see e.g. Besbes and Zeevi (2008)) and it seems that these can be combined within our framework to characterize the equilibria in these very realistic, but highly intractable, settings.

In this paper we considered only the case of linear capacity costs. We made this choice so as not to distract the attention from the main ideas in the proposed framework. Given our framework, the ability to deal with the general-cost case 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 competition setting.
Acknowledgments

We are grateful to Robert Shumsky for a helpful discussion. We also thank the reviewers and the AE for their valuable comments which lead to significant improvements over the original manuscript.

References


Appendix

The analysis in §5 is restricted to the case in which the satisfies the differentiability condition (C2) and the “Diagonal Dominance” condition (C4) and, in turn, is linearly continuous in the sense of Lemma 5.9. In this appendix we introduce a more general notion of continuity for the fluid game together with the corresponding extensions of the results in §5. The case of the linear continuity that was used in §5 is then obtained as a special case of the more general setting. In addition, Lemma A.6 and Theorem A.7 provide the generalization of Lemma 5.12 and Theorem 5.13 to the case in which condition (C1) fails hold. We start by making the following definition.

**Definition A.1** The fluid game on $T$ is a game with profit functions

$\bar{\Pi}^T_i(p) := \lambda_i(p, T) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right),$

and strategy space $\mathcal{P}$.

The fluid game is then the special case with $T = 0$. We let $\psi^T_i(p_{-i})$ be the best response of player $i$ (in the fluid game on $T$) to prices $p_{-i}$ of the competitors and set

$\psi^T(p) := (\psi^T_1(p_{-1}), \ldots, \psi^T_I(p_{-I})).$

If an equilibrium $p^*(T)$ exists for the fluid game on $T$, then it is given by the unique solution to $p - \psi(p) = 0$. We now introduce a generalization of the linear continuity of the fluid game. The generalization is in two directions (i) we require the continuity to hold in a neighborhood of $T = 0$.
rather than only on $T = 0$ itself, and (ii) we have a general function $g_i(\cdot)$ rather than the linear function defined through the matrix $B$ in §5.

Definition A.2 (uniform g-continuity around $T = 0$) We say that the family of fluid games is uniformly g-continuous around $T = 0$ if there exists $\delta > 0$ such that (a) the fluid game on $T$ has a unique equilibrium for $p^*(T)$ for all $T \in [0, \delta]^I$, and (b) there exists a function $g$ such that for any $T \in [0, \delta]^I$ and $\epsilon \in \mathbb{R}_+$,

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

implies that

$$|p - p^*(T)| \leq g(\epsilon).$$

Note that g-continuity of the fluid game is then the special case with $T = 0$. Within this general terminology, Lemma 5.9 can be re-stated as saying that, under Condition (C4), the fluid game is g-continuous with the function $g(x) := Bx/(1 - C)$.

Lemma A.3 Under conditions (C2)-(C4) the fluid game is uniformly linearly continuous around $T = 0$ with $g(x) = \frac{1}{1-C}|B^{-1}x|$ and $C$ is the constant from (C4).

We remark that, in the proof of this Lemma, the “Diagonal Dominance” condition is used to show the linear continuity of the fluid game. The extension to uniform linear continuity uses the twice continuous differentiability in Condition (C2). In contrast to the case of linear continuity, for which we identified the sufficient condition (C4), for the general setting we provide only the framework and it remains for the user to identify the corresponding functions $g_i(\cdot)$ in the cases that don’t necessarily satisfy the sufficient condition of Lemma A.3.

If one can identify the functions $g_i(\cdot)$ that satisfy Definition A.2, than the following would be the corresponding generalization of Theorem 5.10. We note that the uniform g-continuity is not required for this result, only the g-continuity of the fluid game (with $T = 0$).

Theorem A.4 (bounds on the distance from the fluid game) There exists a sequence $\epsilon^\Lambda = O(1/y_{i_1}^\Lambda, \ldots, 1/y_{i_I}^\Lambda)$ such that $(p^*, 0)$ is an $\epsilon^\Lambda$ Nash equilibrium for all $\Lambda$. Moreover,

$$T_i^{*, \Lambda}(p^*, 0) \sim y_{i_i}^\Lambda, \ i \in \mathcal{I} : \alpha_i > 1,$$

(40)

41
\[ T_i^{p^\Lambda}(p^*, 0) = O(y_i^\Lambda), \ i \in \mathcal{I} : \alpha_i = 1, \]  
(41)

and

\[ T_i^{p^\Lambda}(p^*, 0) = o(y_i^\Lambda), \ i \in \mathcal{I} : \alpha_i < 1. \]  
(42)

If, in addition, the fluid game is g-continuous, then

\[ |p_i^{p^\Lambda}(p^*, 0) - p_i^*| = O(g_i(\sqrt{\delta^\Lambda})), \ i \in \mathcal{I}. \]  
(43)

with \( \delta^\Lambda = C \left( \frac{1}{\Lambda y_i} + (y_i^\Lambda)^{\alpha_i} \right) \) for some constant \( C > 0. \)

The following is the general case of Theorem 5.11.

**Theorem A.5 (bounds on the distance from the diffusion game)** Suppose that (C1)-(C3) and (C5) hold and let \( \{(p^\Lambda, T^\Lambda), \ Lambda \geq 0\} \) be such that \( (p^\Lambda, T^\Lambda) \) is a Nash equilibrium for the \( \Lambda \)th diffusion game. If in addition, the fluid game is uniformly g-continuous around \( T = 0 \), then

\[ T_i^{p^\Lambda}(p^\Lambda, T^\Lambda) = T_i^\Lambda + o(y_i^\Lambda) \text{ and } p_i^{p^\Lambda}(p^\Lambda, T^\Lambda) = p_i^\Lambda + o(g_i(\sqrt{\delta^\Lambda})), \]

where \( \delta^\Lambda = (y_i^\Lambda)^{\alpha_i} \).

We now remove the requirement that (C1) holds. In particular, we do not require that a Nash equilibrium exists for the diffusion model. First, the following is the generalization of Lemma 5.12. Theorem A.7 is then the generalization of Theorem 5.13.

**Lemma A.6** Suppose that conditions (C3) and (C5) hold and let \( (p^\Lambda, T^\Lambda) \) be any sequence with \( (p^\Lambda, T^\Lambda) \to (p^*, 0) \). Then,

\[ \frac{T_i^{p^\Lambda}(p^\Lambda, T^\Lambda)}{y_i^\Lambda} \to \eta_i^*, \]

where \( \eta_i^* = 0 \) if \( \alpha_i < 1 \) and it equals

\[ \eta_i^* = \arg\max_{\eta \geq 0} \eta^{\alpha_i} f_i(p^*, 0) \left( p_i^* - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \beta_i(\eta_i) \left( \frac{\lambda_i(p^*, 0)}{\mu_i} \right)^{\frac{1}{\alpha_i+1}}, \]

if \( \alpha_i \geq 1. \) Here \( \beta_i(\eta_i) \) is the solution \( x \) to \( P(x)e^{-\mu_i x \sqrt{\lambda_i(p^*, 0)/\mu_i}} = \phi \) whenever \( \alpha_i = 1 \) and is the solution to \( e^{-\mu_i x \left( \frac{\lambda_i(p^*, 0)}{\mu_i} \right)^{\frac{1}{\alpha_i+1}}} = \phi \) when \( \alpha_i > 1. \)

For the following, we let \( \eta^\Lambda = (\eta_1^\Lambda y_1^\Lambda, \ldots, \eta_I^\Lambda y_I^\Lambda). \)
Theorem A.7 Suppose that conditions (C2), (C3) and (C5) hold. In addition, assume that (i) for each \( \Lambda \), a unique equilibrium \( p^\Lambda(\eta^\Lambda) \) exists for the fluid game on \( \eta^\Lambda \), and (ii) the fluid game is uniformly \( g \)-continuous around \( T = 0 \). Then, there exists a sequence \( \epsilon^\Lambda = o\left(1/y_1^\Lambda, \ldots, 1/y_I^\Lambda\right) \) such that \((p^\Lambda(\eta^\Lambda), \eta^\Lambda)\) is an \( \epsilon^\Lambda \)-Nash equilibrium for the \( \Lambda \)th diffusion game and for the \( \Lambda \)th market game. Moreover,

\[
T_i^{*,\Lambda} = \eta_i^\Lambda + o(y_i^\Lambda), \quad \text{and} \quad p_i^{*,\Lambda} = p_i^\Lambda(\eta^\Lambda) + o\left(g_i(\sqrt{\zeta^\Lambda})\right),
\]

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

Example A.1 (the multinomial logit case) For the multinomial logit demand model in (18), the pricing game has a Nash equilibrium (see Theorem 2 of Bernstein and Federgruen (2004)). Moreover, for each \( \delta > 0 \) there exists \( \Lambda \) large enough so that \( \eta^\Lambda \in [0,\delta]^I \). Choosing \( \delta > 0 \) small enough, and using the continuity of the multinomial logit demand functions we can conclude that the diagonal dominance condition holds for the pricing game on \( \eta^\Lambda \) provided that it holds for the fluid game. This latter requirement is satisfied by imposing condition (C4). Hence, for all \( \Lambda \) large enough, the equilibrium \( p^\Lambda(\eta^\Lambda) \) of the fluid game on \( \eta^\Lambda \) is unique. \( \blacksquare \)
In this e-companion we provide proofs for all the theorems and lemmas. The proofs of the different results appear in the their order of appearance in the paper. Accordingly, the proofs of the results in the Appendix appear in the second portion of this e-companion. Two auxiliary lemmas—EC.1 and EC.2—appear at the end of the e-companion.

**Proof of Lemma 4.1:** This lemma follows directly from Lemma 5.5 which is proved later in this e-companion.

**Proof of Theorem 4.3:** The proof draws on Definition 3.2 of $\epsilon$-Nash equilibria, Assumption 4.1 on the uniqueness of the equilibrium $p^*$ for the fluid game, and the properties of the demand functions as listed in Assumption 3.1.

We fix a sequence $T^\Lambda$ that satisfies the following three properties:

\[
\begin{align*}
\max_{i \in I} \sup_{p \in P} |\Lambda_i(p, T^\Lambda) - \Lambda_i(p, 0)| &\leq \epsilon^\Lambda / 16, \quad (EC1) \\
\sup_{p \in P} \hat{e}_i(\Lambda_i(p, T^\Lambda), T_i^\Lambda) &\leq \epsilon^\Lambda / 16 \gamma_i, \quad \text{and} (EC2) \\
T^\Lambda &\to 0, \text{ as } \Lambda \to \infty. \quad (EC3)
\end{align*}
\]

Such a sequence exists by the absolute continuity of the demand functions on the compact domain $P \times \Theta$ and by Lemma 5.5. For (EC2) we are using the assumption that $\epsilon^\Lambda \to \infty$.

To show that $(p^*, T^\Lambda)$ is an $\epsilon^\Lambda$-Nash equilibria for the $\Lambda^{th}$ market game, fix a firm $i$ and a sequence $\{(p_i^\Lambda, T_i^\Lambda), \Lambda \geq 0\}$ with $(p_i^\Lambda, T_i^\Lambda) \in P_i \times \Theta$ of prices and service levels for firm $i$ such that $(p_i^\Lambda, T_i^\Lambda) \neq (p^*_i, T_i^\Lambda)$. Define

\[
(\bar{p}^\Lambda, \bar{T}^\Lambda) := (p_i^\Lambda, T_i^\Lambda) \uparrow (p^*_i, T_i^\Lambda)_{-i}.
\]

As $\hat{e}_i(\cdot, \cdot) \geq 0$, we have that

\[
\Pi_i^\Lambda(\bar{p}^\Lambda, \bar{T}^\Lambda) \leq \Lambda_i(\bar{p}^\Lambda, \bar{T}^\Lambda) \left( \bar{p}_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right)
\]
By the choice of $T^\Lambda$, we have that
\[
\left[ \Lambda_i(p^\Lambda, \tilde{T}^\Lambda) \left( \tilde{p}^\Lambda_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \Lambda_i(p^\Lambda, 0) \left( \tilde{p}^\Lambda_i - c_i - \frac{\gamma_i}{\mu_i} \right) \right] \leq \epsilon^\Lambda / 8. \tag{EC4}
\]
Indeed, one writes $\lambda_i(p^\Lambda, \tilde{T}^\Lambda) - \lambda_i(p^\Lambda, 0) = \lambda_i(p^\Lambda, T^\Lambda) - \lambda_i(p^\Lambda, 0) - \Lambda_i(p^\Lambda, T^\Lambda) + \lambda_i(p^\Lambda, \tilde{T}^\Lambda)$. By (EC1) we then have that $|\Lambda_i(p^\Lambda, T^\Lambda) - \Lambda_i(p^\Lambda, 0)| \leq \epsilon^\Lambda / 8$. There are now two cases: if $T^\Lambda_i \leq T^\Lambda_i$ then we can apply (EC1) once again with $T^\Lambda_i$ replaced with $\tilde{T}^\Lambda_i$. If, on the other hand, $T^\Lambda_i > T^\Lambda_i$, then the monotonicity of the demand functions is invoked to have that $\Lambda_i(p^\Lambda, T^\Lambda) - \Lambda_i(p^\Lambda, T^\Lambda) \leq 0$.

Note that (EC4) is independent of the actual values of the sequence $\{(p^\Lambda_i, T^\Lambda_i), \Lambda \geq 0\}$ and depends only on the values of $\{(p^\Lambda, T^\Lambda), \Lambda \geq 0\}$. By (EC1) we have that
\[
\left| \Lambda_i(p^\star, T^\Lambda) \left( p^\star_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \Lambda_i(p^\star, 0) \left( p^\star_i - c_i - \frac{\gamma_i}{\mu_i} \right) \right| \leq \epsilon^\Lambda / 8. \tag{EC5}
\]
By the definition of $p^\star$ as the Nash equilibrium of the fluid game, we have that
\[
\Lambda_i(\tilde{p}^\Lambda, 0) \left( \tilde{p}^\Lambda_i - c_i - \frac{\gamma_i}{\mu_i} \right) \leq \Lambda_i(p^\star, 0) \left( p^\star_i - c_i - \frac{\gamma_i}{\mu_i} \right). \tag{EC6}
\]
Combining (EC4), (EC5) and (EC6) we readily have that,
\[
\Lambda_i(\tilde{p}^\Lambda, \tilde{T}^\Lambda) \left( \tilde{p}^\Lambda_i - c_i - \frac{\gamma_i}{\mu_i} \right) \leq \Lambda_i(p^\star, 0) \left( p^\star_i - c_i - \frac{\gamma_i}{\mu_i} \right) + \frac{\epsilon^\Lambda}{4}.
\]
Using this together with equation (EC5), we then have that
\[
\Pi_i^\Lambda(\tilde{p}^\Lambda, \tilde{T}^\Lambda) \leq \Lambda_i(p^\star, T^\Lambda) \left( p^\star_i - c_i - \frac{\gamma_i}{\mu_i} \right) + \frac{\epsilon^\Lambda}{2}.
\]
By (EC2) $\dot{e}_i(\Lambda_i, T^\Lambda_i) \leq \epsilon^\Lambda / 2\gamma_i$ and we conclude that
\[
\Pi_i^\Lambda(\tilde{p}^\Lambda, \tilde{T}^\Lambda) \leq \Pi_i^\Lambda(p^\star, T^\Lambda) + \epsilon^\Lambda,
\]
so that, for each $\Lambda$, $(p^\star, T^\Lambda)$ is an $\epsilon^\Lambda$-Nash equilibrium. Finally, note that we can also find a sequence $T^\Lambda$ of vectors with equal components, i.e, such that $T^\Lambda_1 = T^\Lambda_2 = \ldots = T^\Lambda_I$, that satisfies (EC1)-(EC3) and repeat the whole argument with this sequence to get the second part of the Theorem.

Proof of Theorem 4.4: We divide the proof into two parts. We first prove the characterization for the equilibrium service-levels in equation (11). We then proceed to prove the equilibrium-price characterization in equation (12).
Proof of (11): Let $(\bar{p}^A, \bar{T}^A)$ be the sequence of $\epsilon^A$-Nash equilibrium. To reach a contradiction, assume that there is no such sequence $\delta^A$ for $T^A$. In particular, there exists a firm $i$ such that $\lim \sup_{A \to \infty} \bar{T}^A_i \geq \delta$, for some $\delta > 0$. We may use the compactness of $\Theta$ to choose a subsequence $\Lambda_j$ such that $\lim_{j \to \infty} \bar{T}^A_i = \bar{\delta} \geq \delta$.

Define $T^A$ by setting $T^A_i = \zeta/\sqrt{\Lambda}$ for this firm $i$ and some $\zeta > 0$ and by setting $T^A_k = T^A_i$ for all $k \neq i$. Since $\bar{T}^A_i \to 0$ as $\Lambda \to \infty$, we can re-choose $j$ large enough so that $\bar{T}^A_i \leq \bar{T}^A_i - \eta$, for some $0 < \eta \leq \delta$. Since $\lambda_i(p, T)$ is strictly decreasing in $T_i$ (see Assumption 3.1), there exists $\epsilon > 0$, such that

$$\lambda_i(\bar{p}^A, \bar{T}^A) \cdot \left(\bar{p}^A_i - c_i - \frac{\gamma_i}{\mu_i}\right) - \lambda_i(p^A, T^A) \cdot \left(\bar{p}^A_i - c_i - \frac{\gamma_i}{\mu_i}\right) \geq 4\epsilon.$$ 

Using the definition of the profit functions we have that

$$\Pi^A_i(\bar{p}^A, \bar{T}^A) - \Pi^A_i(p^A, T^A) \leq \Lambda^j \lambda_i(p^A, \bar{T}^A) \left(p_i - c_i - \frac{\gamma_i}{\mu_i}\right) - \left(\Lambda^j \lambda_i(p^A, \bar{T}^A) \cdot \left(\bar{p}^A_i - c_i - \frac{\gamma_i}{\mu_i}\right) - \gamma_i \hat{e}_i(\Lambda^j, \zeta/\sqrt{\Lambda}) - 4\epsilon \Lambda + \gamma_i \hat{e}_i(\Lambda^j, \zeta/\sqrt{\Lambda}). \right)$$

By Lemma 5.1 we have that $\hat{e}_i(\Lambda, \zeta/\sqrt{\Lambda}) \leq K \sqrt{\Lambda}$ for all $\Lambda$ large enough and some $K > 0$. Hence, we can re-choose $j$, so that

$$\Pi^A_i(\bar{p}^A, \bar{T}^A) - \Pi^A_i(p^A, T^A) \leq -2\epsilon \Lambda^j.$$ 

Firm $i$ can, hence, improve its profit, $\Pi^A_i$, by more than $\epsilon \Lambda$. Since $\epsilon A / \Lambda \to 0$, there exists $j_0$ such that $\epsilon \Lambda^j \leq \epsilon \Lambda$ for all $j \geq j_0$. Consequently, for all $j$ large enough, $(\bar{p}^A, \bar{T}^A)$ can not be an $\epsilon^A$-Nash equilibrium.

Proof of (12): We fix the sequence $(p^A, T^A)$ of $\epsilon^A$-Nash equilibria. To reach a contradiction assume that $\lim \sup_{A \to \infty} \|p^A - p^*\| > 0$. We then say that $p^A$ is asymptotically distinguishable from $p^*$. Note that if $\max_{i \in \mathcal{I}} \lim \sup_{A \to \infty} T^A_i > 0$, the result of the theorem trivially follows from (11). Hence, we assume $T^A_i \to 0$ for all $i \in \mathcal{I}$. We will show that under the assumption that $p^A$ is distinguishable from $p^*$, every limit point $p$ of $p^A$ must be an equilibrium point for the fluid game. Such a limit point exists by the compactness of $\mathcal{P}$. Since $p^A$ is distinguishable from $p^*$, this will imply the existence of multiple equilibria for the fluid game, contradicting Assumption 4.1. It remains, hence, only to show that every limit point $p$ is indeed an equilibrium point for the fluid game. Towards that end, fix a limit point $p$ of $\{p^A, \Lambda \geq 0\}$ and the corresponding
convergent subsequence \( \{p^k, k \geq 0\} \). We claim that \( p \) is an \( \epsilon \)-Nash equilibrium for the fluid game for any \( \epsilon > 0 \). In turn, it is a Nash equilibrium for this game. Define \( \bar{p} := (\bar{p}_i, p_{-i}) \), for some price \( \bar{p}_i \in [p^\text{min}_i, p^\text{max}_i] \) with \( \bar{p}_i \neq p_i \). Then, since \( (p^\Lambda, T^\Lambda) \) is the assumed sequence of \( \epsilon^\Lambda \)-Nash equilibria, we have that for all \( k \) large enough,

\[
\Pi_i^\Lambda(\bar{p}^\Lambda, T^\Lambda) \leq \Pi_i^\Lambda(p^\Lambda, T^\Lambda) + \epsilon/4,
\]

for some \( \epsilon > 0 \). Observe that by Lemma 5.1, \( \hat{e}_i(\Lambda, T^\Lambda)/\Lambda \to 0 \) as \( \Lambda \to \infty \). This, together with the continuity of the demand functions, implies that

\[
\lim_{\Lambda \to \infty} \sum_{i \in I} |\Pi_i^\Lambda(p^\Lambda, T^\Lambda) - \Pi_i^\Lambda(p)| = 0.
\]

In particular,

\[
\lim_{k \to \infty} \sum_{i \in I} |\Pi_i^\Lambda(p^\Lambda, T^\Lambda) - \Pi_i^P(p)| = 0.
\]

Hence,

\[
\Pi_i^\Lambda(\bar{p}^\Lambda, T^\Lambda) \to \Pi_i^P(\bar{p}) \text{ and } \Pi_i^\Lambda(\bar{p}^\Lambda, T^\Lambda) \to \Pi_i^P(p), \text{ as } k \to \infty
\]

where \( \bar{p} = (\bar{p}_i, p_{-i}) \), and we have that

\[
\Pi_i^P(\bar{p}) \leq \Pi_i^P(p) + \epsilon.
\]

Hence, \( p \) is an \( \epsilon \)-Nash equilibrium for the fluid game. As \( \epsilon \) was arbitrary, we have that \( p \) is a Nash equilibrium for the fluid game. Since \( p \neq p^* \) we have a contradiction to the uniqueness of equilibria for the fluid game.

Proof of Lemma 5.1: The first part of the lemma follows directly from Proposition 9.3 in Borst et al. (2004) and from item (ii) in Example 9.4 there. We turn to prove the convexity of the function \( \beta_i(\eta) \).

Let \( f(x, y) := P(x)e^{-\mu_i x y} \). By Lemma B.1 in Borst et al. (2004), the function \( P(\cdot) \) is strictly decreasing convex. Using this property one can easily show that the function \( f(x, y) \) is convex and strictly decreasing in \( x \) and convex and strictly decreasing in \( y \in [0, \bar{T}] \). Also, \( \frac{\partial}{\partial x} \frac{\partial}{\partial y} f < 0 \).

Note that \( \beta_i(\eta) \) is the function that satisfies \( f(\beta_i(\eta), \eta) = \phi \). By differentiating once on both sides of this equality we get:

\[
\frac{\partial}{\partial x} f(\beta_i(\eta), \eta) \cdot \beta_i'(\eta) + \frac{\partial}{\partial y} f(\beta_i(\eta), \eta) = 0.
\]
Using the fact that \( f \) is strictly decreasing in \( y \) and \( x \) we then have that \( \beta'_i(\eta) < 0 \). Differentiating for the second time we have
\[
\frac{\partial^2}{\partial x^2} f(\beta_i(\eta), \eta) \cdot (\beta'_i(\eta))^2 + 2 \frac{\partial}{\partial x} \frac{\partial}{\partial y} f(\beta_i(\eta), \eta) \cdot \beta'_i(\eta) + \frac{\partial^2}{\partial y^2} f(\beta_i(\eta), \eta) = -\beta''_i(\eta) \frac{\partial}{\partial x} f.
\]

The first element on the left-hand side is positive by the convexity of \( f \) in \( x \). The second element is positive by the property \( \frac{\partial}{\partial x} \frac{\partial}{\partial y} f < 0 \). The last element on the left-hand side is positive by the convexity of \( f \) in its \( y \) argument. Dividing both sides by \( \frac{\partial}{\partial x} f \) and using the fact that \( f \) is decreasing in \( x \) we then have that \( \beta''_i(\eta) > 0 \) and the proof is complete. \( \blacksquare \)

**Proof of Lemma 5.5:** By known \( M/M/N \) formulas (see e.g. Chapter 5-9 of Wolff (1989)), we know that \( \hat{e}_i(\Lambda, T_i) \) satisfies
\[
P\{W_i^A > T_i^A\} = P\{W_i^A > 0\} e^{-\mu_i T_i^A \hat{e}_i(\Lambda, T_i^A)} \leq \phi.
\]
As, \( P\{W_i^A > 0\} \leq 1 \) we have that \( \hat{e}_i(\Lambda, T_i^A) \leq -\ln(\phi)/\mu_i T_i^A \). Hence, \( T_i^A \hat{e}_i(\Lambda_i, T_i^A) \leq -\ln(\phi) + o(\ln(\phi)) \), and the result of the Lemma follows. \( \blacksquare \)

**Proof of Theorem 5.6:** The proof of this theorem follows from Theorem A.4. Specifically, let \( \{(p^A, T^A), \Lambda \geq 0\} \) be the assumed sequence of Nash equilibria. Then, as \( (p^A, T^A) \) must be the best response to of firm \( i \) to \( (p^A, T^A) \) we have, by Theorem A.4, that \( T_i^A \sim y_i^A \) if \( 1/y_i^A = o(\sqrt{\Lambda}) \) and \( T_i^A = O(y_i^A) \) if \( 1/y_i^A \sim \sqrt{\Lambda} \). By Lemma 5.5 we then have that \( \hat{e}_i(\Lambda_i, T_i^A) \sim 1/T_i^A \) if \( 1/y_i^A = o(\sqrt{\Lambda}) \) and \( \hat{e}_i(\Lambda_i, T_i^A) \sim \sqrt{\Lambda} \) if \( 1/y_i^A \sim \sqrt{\Lambda} \). By taking subsequences, the result of the Theorem now follows from Proposition 1 in Halfin and Whitt (1981). \( \blacksquare \)

**Proof of Lemma 5.8:** As \( p_{-i} \) is fixed, it suffices to prove the result for a one dimensional function. Specifically, fix a twice continuously differentiable function \( f(x) : \mathbb{X} \rightarrow \mathbb{R}_+ \) that is defined on a compact and convex set \( \mathbb{X} \subseteq \mathbb{R}_+ \) and such that \( \frac{\partial^2}{\partial x^2} f(x) < 0 \) for all \( x \in \mathbb{X} \). Let \( x^* = \arg\max_{x \in \mathbb{X}} f(x) \). We claim that \( |x^* - y| \leq C \sqrt{\epsilon} \) whenever \( \max_{x \in \mathbb{X}} f(x) - f(y) \leq \epsilon \) for some constant \( \epsilon > 0 \). The first part of the Lemma will then follow by setting \( f(x) := \bar{\Pi}_i(x, p_{-i}) \).
To prove our claim for \( f(x) \) consider first the case in which \( x^* \) lies in the interior of \( \mathbb{X} \). In this case \( x^* \) solves the first order condition \( f'(x) = 0 \). Assume that \( |x^* - y| > \sqrt{\epsilon} \). Assume that \( x^* < y \) (the other case is treated similarly). As the second derivative is strictly negative we have
that \( f'(x^* + \sqrt{\epsilon}/2) \leq -\delta \sqrt{\epsilon}/2 \) with \( \delta \) as in the condition of the Lemma. In particular, as \( f'(x) \) is a decreasing function we have that
\[
f(x^*) \geq f(x^* + \sqrt{\epsilon}/2) \geq f(y) + (y - x^* - \sqrt{\epsilon}/2)\delta \sqrt{\epsilon}/2 = f(y) + (y - x^*)\delta \sqrt{\epsilon}/2 + o(\sqrt{\epsilon}).
\]

Hence, \( f(x^*) - f(y) \leq \epsilon \) implies for all \( \epsilon \) small enough that \( y - x^* \leq C \sqrt{\epsilon} \) for some constant \( C > 0 \).

To complete the argument, assume that \( x^* \) is on the boundary of \( X \). Assume that \( x^* \) is the smallest element in \( X \) (the proof is similar for the case in which \( x^* \) is the largest element). Then, we have that \( f'(x) < 0 \) for all \( x > x^* \) and in particular \( f'(x + \sqrt{\epsilon}/2) \leq -\delta \sqrt{\epsilon}/2 \). From here we can apply the same arguments as above.

Proof of Lemma 5.9: This lemma is a direct consequence of Lemma A.3.

Proof of Theorem 5.10: This theorem is a special case of Theorem A.4.

Proof of Theorem 5.11: This theorem is a special case of Theorem A.5.

Proof of Example 5.4: We prove that the ML demand model satisfies condition (C5). To this end, after some basic manipulations we get
\[
\frac{\lambda_i(p, T_{-i}, x) - \lambda_i(p, T_{-i}, 0)}{x^{\alpha_i}} = \frac{(v_i(p_i, x) - v_i(p_i, 0))(1 - \lambda_i(p, T_{-i}, 0))}{1 + \sum_{j \neq i} v_j(p_j, T_j) + v_i(p_i, 0)},
\]
where, for all \( j \in I \), \( v_j(p_j, T_j) := e^{a_j(T_j)p_j} \) and \( a_j(T_j) = a_j - k_j(T_j)^{\alpha_j} \). Using Taylor expansion around \( z = 0 \) for the function \( f(z) := e^{a_i-b_i p_i-k_i z} \) we then have that
\[
\frac{\lambda_i(p, T_{-i}, x) - \lambda_i(p, T_{-i}, 0)}{x^{\alpha_i}} = \frac{k_i v_i(p_i, 0) x^{\alpha_i}(1 - \lambda_i(p, T_{-i}, 0))}{1 + \sum_{j \neq i} v_j(p_j, T_j) + v_i(p_i, 0)} + o(1)
\]
and, consequently,
\[
\frac{\lambda_i(p, T_{-i}, x) - \lambda_i(p, T_{-i}, 0)}{x^{\alpha_i}} \rightarrow \frac{k_i v_i(p_i, 0) (1 - \lambda_i(p, T_{-i}, 0))}{1 + \sum_{j \neq i} v_j(p_j, T_j) + v_i(p_i, 0)}.
\]

Proof of Lemma 5.12: The proof of this lemma follows from that of Lemma A.6–see the comment at the end of the proof of Lemma A.6.
Proof of Theorem 5.13: This theorem is obtained as a special case of Theorem A.7 by noting that the uniform linear continuity of the fluid game around $T = 0$ follows from Lemma A.3. ■

Proof of Lemma A.3: Fix $\epsilon \in \mathbb{R}_+^I$. Using the fact that $p^* = \psi(p^*)$, we write $p_i - \psi_i(p_{-i}) = p_i - p_i^* - (\psi_i(p_{-i}) - \psi(p_i^*))$. Using (C4) we have that

$$|\psi_i(p_{-i}) - \psi_i(p_i^*)| < C\|p_{-i} - p_i^*\|,$$

for $C < 1$. Consequently, $|p - \psi(p)| \leq \epsilon$, implies that

$$\epsilon_i \geq p_i - p_i^* - (\psi_i(p_{-i}) - \psi(p_i^*)) \geq (1 - C)\|p_{-i} - p_i^*\|.$$

Hence, for each $i \in I$, we have that

$$\|p_{-i} - p_i^*\| \leq \frac{\epsilon_i}{1 - C}$$

and, consequently, that

$$|B(p - p^*)| \leq \frac{1}{1 - C} \cdot \epsilon,$$

where $B$ is the matrix with elements $B_{ii} = 0$ and $B_{ij} = 1$ for all $j \neq i$. In particular, the fluid game is $g$-continuous with $g(x) = \frac{1}{1 - C}|B^{-1}x|$. The uniform $g$-continuity now follows from (C2). Indeed, the continuity of the derivatives guarantees that, for all $T$ small enough, there exists $C < 1$ such that

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

Proof of Theorem A.4: The outline of the proof is as follows: we first prove that, given any sequence $\{(p_{i,\Lambda}, T_{i,\Lambda}), \Lambda \geq 0\}$, $T_{i,\Lambda}^*(p_{\Lambda}, T_{\Lambda})$ must satisfy that $T_{i,\Lambda}^*(p_{\Lambda}, T_{\Lambda}) \sim y_i^\Lambda$ if $1/y_i^\Lambda = o(\sqrt{\Lambda})$ and $T_{i,\Lambda}^*(p_{\Lambda}, T_{\Lambda}) = O(y_i^\Lambda)$ otherwise. This will, in particular, establish (40). Having established (40) we will turn to show that $(p^*, 0)$ is an $\epsilon^\Lambda$-Nash equilibrium as claimed and establish the bounds in (43).

We start, then, with the treatment of $T_{i,\Lambda}^*(p_{\Lambda}, T_{\Lambda})$. We divide the proof into two cases: (i) firm $i$ has $\alpha_i \geq 1$ and (ii) $\alpha_i < 1$, where $\alpha_i$ is the exponent in Assumption 5.1. First, consider a firm $i$ with $\alpha_i \geq 1$. It is can be verified directly that, for these values of $\alpha_i$, $y_i^\Lambda \sim z_i^\Lambda$ and that

\[7\]
\( \Lambda f_i(z_i^A) \sim \frac{1}{z_i}. \) Furthermore, by the definition of \( z_i^A \), we have that \( f_i'(z_i^A) = \frac{1}{\Lambda(z_i^A)} \) and we can use the convexity of the function \( f(x) = x^{\alpha_i} \), to write

\[
\Lambda f_i(T_i^A) \geq \Lambda f_i(y_i^A) + \frac{\Lambda}{(y_i^A)^2} (T_i^A - y_i^A),
\]

for any \( T_i^A \geq y_i^A \). In particular, if \( T_i^A/y_i^A \to \infty \) then

\[
\frac{\Lambda f_i(T_i^A) - \Lambda f_i(y_i^A)}{1/y_i^A} \to \infty. \tag{EC7}
\]

Let \( (p_{-i}^A, p_i^*, T_{-i}^A, x) := (p_{-i}^A, x) \uparrow (p^A, T^A)_{-i} \) and note that, by definition,

\[
T_i^* = \arg\max_x \Lambda_i(p_{-i}^A, p_i^*, T_{-i}^A, x) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, x),
\]

and, in particular, that

\[
T_i^* = \arg\max_x \left( \Lambda_i(p_{-i}^A, p_i^*, T_{-i}^A, x) - \Lambda_i(p^A, T_{-i}^A, 0) \right) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, x).
\]

Using Assumption 5.1 we have that

\[
(\Lambda_i(p^A, T_{-i}^A, T_i^A) - \Lambda_i(p^A, T_{-i}^A, 0)) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, T_i^A) \sim -C_1 \Lambda(T_i^A)^{\alpha_i} - C_2 \frac{1}{T_i^A},
\]

for some constants \( C_1 \) and \( C_2 \). Assume that \( T_i^A/y_i^A \to \infty \). Then, it follows from (EC7) that

\[
\frac{(\Lambda_i(p^A, T_{-i}^A, T_i^A) - \Lambda_i(p^A, T_{-i}^A, 0)) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, T_i^A)}{1/y_i^A} \to -\infty. \tag{EC9}
\]

Using (EC8) we have, however, that

\[
\limsup_{\Lambda \to \infty} \frac{(\Lambda_i(p^A, T_{-i}^A, y_i^A) - \Lambda_i(p^A, T_{-i}^A, 0)) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, y_i^A)}{1/y_i^A} > -\infty. \tag{EC10}
\]

which leads to a contradiction to the definition of \( T_i^* \). Hence, for \( i \in I \) with \( \alpha_i \geq 1 \) we must have \( T_i^*(p_{-i}^A, T_{-i}^A) = O(y_i^A) \). To complete the proof, we need to show that \( \liminf_{\Lambda \to \infty} T_i^A/y_i^A > 0 \) for all with \( 1/y_i^A = O(\sqrt{\Lambda}) \). Assume, to reach a contradiction, that \( 1/y_i^A = O(\sqrt{\Lambda}) \) but \( T_i^A/y_i^A \to 0 \). Then, as \( \hat{e}_i(\Lambda_i, T_i^A) \sim \max\{1/T_i^A, \sqrt{\Lambda}\} \) we have that

\[
\frac{\hat{e}_i(\Lambda_i, T_i^A)}{1/y_i^A} \to \infty.
\]

In particular, as \( \Lambda_i(p, T) \) is decreasing in \( T_i \) we will again have (EC9) which is a contradiction to the definition of \( T_i^* \). Indeed, we can always use \( y_i^A \) to obtain a better result as in (EC10). The proof for \( \alpha_i \geq 1 \) is hence complete.
We now turn to consider the case \( \alpha_i < 1 \). First, note that with \( T_i^{*;\Lambda} = 0 \) we have that
\[
\limsup_{\Lambda \to \infty} \frac{(\Lambda_i(p^\Lambda, T_i^\Lambda, 0) - \Lambda_i(p^\Lambda, T_{i,i}^\Lambda, 0)) (p_i - c_i - \gamma_i/\mu_i) - \gamma_i \hat{e}_i(\Lambda_i, 0)}{1/y_i^\Lambda} = \frac{\gamma_i \hat{e}_i(\Lambda_i, 0)}{1/y_i^\Lambda} \geq -\infty.
\]
Fix a sequence \( \{T_i^\Lambda, \Lambda \geq 0\} \) with \( \liminf_{\Lambda \to \infty} T_i^\Lambda/y_i^\Lambda > 0 \). Then, we can apply Assumption 5.1 to get (EC9) once again. Consequently, we must have that \( T_i^{*;\Lambda} = o(y_i^\Lambda) \) and the proof of the first part of the Theorem is complete.

We turn now to show that \( (p^*, 0) \) is an \( \epsilon^\Lambda \) Nash equilibria as well as that the price bounds in (43) hold. To this end, note that
\[
|\Pi_i^\Lambda(p^*_{i,i}, x, 0, T_i^{*;\Lambda}) - \Pi_i^\Lambda(p^*_{i,i}, x, 0, 0)| \leq C \left( \hat{e}_i(\Lambda_i(p^*, 0), 0) + |\Lambda_i(p^*_{i,i}, x, 0, T_i^{*;\Lambda}) - \Lambda_i(p^*_{i,i}, x, 0, 0)| \right),
\]
for some constant \( C > 0 \). Using the first part of the theorem in conjunction with Assumption 5.1 and Lemma 5.5 we then have that
\[
\Pi_i^\Lambda(p^*_{i,i}, x, 0, T_i^{*;\Lambda}) - \Lambda_i(p^*_{i,i}, x, 0, 0) \left( x - c_i - \frac{\gamma_i}{\mu_i} \right) \leq C \frac{1}{y_i^\Lambda}.
\]
Similarly, we have that
\[
\Pi_i^\Lambda(p^*, 0, T_i^{*;\Lambda}) - \Lambda_i(p^*, 0) \left( p_i^* - c_i - \frac{\gamma_i}{\mu_i} \right) \leq C \frac{1}{y_i^\Lambda}.
\]
By the definition of \( p^* \) we have that
\[
\Lambda_i(p^*_{i,i}, x, 0, 0) \left( x - c_i - \frac{\gamma_i}{\mu_i} \right) \leq \Lambda_i(p^*, 0) \left( p_i^* - c_i - \frac{\gamma_i}{\mu_i} \right).
\]
Consequently, we have that, for any \( x \) and \( y \),
\[
\Pi_i^\Lambda(p_i^*, x, 0, y) \leq \Pi_i^\Lambda(p^*, 0) + \frac{C}{y_i^\Lambda},
\]
for some (re-defined) constant \( C > 0 \) so that \( (p^*, 0) \) is the claimed \( \epsilon^\Lambda \) Nash equilibrium for the \( \Lambda^{th} \) market game. We proceed to provide the bounds in (43). By (EC12) and (EC13) we have that
\[
\left| \Pi_i^\Lambda(p^*_{i,i}, x, 0, T_i^{*;\Lambda}) - \Lambda_i(p^*_{i,i}, x, 0, 0) \left( x - c_i - \frac{\gamma_i}{\mu_i} \right) \right| \leq C \frac{1}{y_i^\Lambda},
\]
and, using (EC11) and Assumption 5.1 we then have that
\[
|\Pi_i^P(p^*) - \Pi_i^P(p^*_{i,i}, p_i^{*;\Lambda}(p^*, 0))| \leq C \left( \frac{1}{\Lambda y_i^\Lambda} + (y_i^\Lambda)^{\alpha_i} \right),
\]
for a re-defined constant \( C > 0 \). Equation (43) now follows from the g-continuity of the fluid game and Lemma 5.8. ■
Proof of Theorem A.5: Let \( \{(p^\Lambda, T^\Lambda), \Lambda \geq 0\} \) be a sequence such that \((p^\Lambda, T^\Lambda)\) is a Nash equilibrium for the \( \Lambda^{th} \) diffusion game. The first part of the theorem follows directly from Lemma A.6 and the comment at the end of the proof of that lemma by which we have that, for \( \alpha_i \geq 1 \), both
\[
\frac{T_i^\Lambda}{y_i^\Lambda} \to \eta(p^\star) \quad \text{and} \quad \frac{T_i^{\star,\Lambda}}{y_i^\Lambda} \to \eta(p^\star).
\]
(EC14)

In particular, \( T_i^{\star,\Lambda} = T_i^\Lambda + o(y_i^\Lambda) \). In addition, if \( \alpha_i < 1 \) then both \( T_i^\Lambda \) and \( T_i^{\star,\Lambda} \) are \( o(y_i^\Lambda) \).

We can now use this to prove the second part of the theorem. Note that, by definition of Nash equilibrium for the diffusion game we have that
\[
\Pi_i^A(p_{i-1}^\Lambda, T_{i-1}^\Lambda, p_i^{\star,\Lambda}, T_i^{\star,\Lambda}) - \Pi_i^A(p^\Lambda, T^\Lambda) = \Lambda_i(p_{i-1}^\Lambda, T_{i-1}^\Lambda, p_i^{\star,\Lambda}, T_i^{\star,\Lambda})(p_i^{\star,\Lambda} - c_i - \gamma_i/i\mu_i) - \Lambda_i(p^\Lambda, T^\Lambda)(p_i^\Lambda - c_i - \gamma_i/i\mu_i)
\]
\[
+ \gamma_i\sqrt{R_i(p^\star, 0)}(\beta_i(T_i^{\star,\Lambda}\sqrt{R_i(p^\star, 0)}) - \beta_i(T_i^\Lambda\sqrt{R_i(p^\star, 0)})) \quad \text{(EC15)}
\]

Using lemma 5.1 we have that
\[
\Pi_i^A(p_{i-1}^\Lambda, T_{i-1}^\Lambda, p_i^{\star,\Lambda}, T_i^{\star,\Lambda}) - \Pi_i^A(p^\Lambda, T^\Lambda) = \Lambda_i(p_{i-1}^\Lambda, T_{i-1}^\Lambda, p_i^{\star,\Lambda}, T_i^{\star,\Lambda})(p_i^{\star,\Lambda} - c_i - \gamma_i/i\mu_i)
\]
\[
- \Lambda_i(p^\Lambda, T^\Lambda)(p_i^\Lambda - c_i - \gamma_i/i\mu_i)
\]
\[
+ \gamma_i\sqrt{R_i(p^\star, 0)}(\beta_i(T_i^{\star,\Lambda}\sqrt{R_i(p^\star, 0)}) - \beta_i(T_i^\Lambda\sqrt{R_i(p^\star, 0)})) + o\left(\gamma_i\sqrt{R_i(p^\star, 0)}(\beta_i(T_i^{\star,\Lambda}\sqrt{R_i(p^\star, 0)}) - \beta_i(T_i^\Lambda\sqrt{R_i(p^\star, 0)}))\right) \quad \text{(EC16)}
\]

By (EC14) we have that both \( T_i^\Lambda = \eta(p^\star)y_i^\Lambda + o(y_i^\Lambda) \) and \( T_i^{\star,\Lambda} = \eta(p^\star)y_i^\Lambda + o(y_i^\Lambda) \). Using Lemma EC.2 we then have that
\[
\gamma_i\sqrt{R_i(p^\star, 0)}(\beta_i(T_i^{\star,\Lambda}\sqrt{R_i(p^\star, 0)}) - \beta_i(T_i^\Lambda\sqrt{R_i(p^\star, 0)})) = o(1/y_i^\Lambda).
\]

Also, we can fix \( \delta > 0 \) so that for all \( \Lambda \) large enough \( (\eta_i - \delta)y_i^\Lambda \leq T_i^{\star,\Lambda} \leq (\eta_i + \delta)y_i^\Lambda \) and then use equation (EC17) in the proof of Lemma 5.12 to have that both
\[
\frac{\Lambda_i(p_{i-1}^\Lambda, T_{i-1}^\Lambda, p_i^{\star,\Lambda}, (\eta_i + \delta y_i^\Lambda)) - \Lambda_i(p_{i-1}^\Lambda, T_{i-1}^\Lambda, p_i^{\star,\Lambda}, 0)}{1/y_i^\Lambda} \to (\eta_i + \delta)^{\alpha_i} f_i(p^*, 0),
\]
and
\[
\frac{\Lambda_i(p_{i-1}^\Lambda, T_{i-1}^\Lambda, p_i^{\star,\Lambda}, (\eta_i - \delta y_i^\Lambda)) - \Lambda_i(p_{i-1}^\Lambda, T_{i-1}^\Lambda, p_i^{\star,\Lambda}, 0)}{1/y_i^\Lambda} \to (\eta_i - \delta)^{\alpha_i} f_i(p^*, 0).
\]

Since \( \delta \) is arbitrary we have that
\[
\frac{\Lambda_i(p_{i-1}^\Lambda, T_{i-1}^\Lambda, p_i^{\star,\Lambda}, T_i^\Lambda) - \Lambda_i(p_{i-1}^\Lambda, T_{i-1}^\Lambda, p_i^{\star,\Lambda}, 0)}{1/y_i^\Lambda} \to (\eta_i)^{\alpha_i} f_i(p^*, 0).
\]
A similar argument is repeated for $T_{i}^{*A}$ to conclude that

$$\frac{\Lambda_i(p^A, T_{-i}^A, T_{i}^{*,A}) - \Lambda_i(p^A, T_{-i}^A, T_{i}^A)}{1/y_i^A} \rightarrow 0.$$  

Plugging these back into (EC15) and (EC16) we have that

$$\Lambda_i(p^A, T_{-i}^A, T_{i}^{*,A}, T_i^A) \left( p_i^{*,A} - c_i - \frac{\gamma_i}{\mu_i} \right) \leq \Lambda_i(p^A, T^A) \left( p_i^A - c_i - \frac{\gamma_i}{\mu_i} \right) + o(1/y_i^A).$$

Dividing by $\Lambda$ we then have that

$$\left| \Pi_i^{T^A, P} (p_i^{*,A}) - \Pi_i^{T, P} (p_i^A) \right| \leq C \frac{1}{\Lambda y_i^A}.$$  

Note that, as $p_i^A$ is the equilibrium price for firm $i$ in the diffusion game, it must be the best response of firm $i$ to $(p_i^A, T^A)$. In particular, it must be the equal to best response, $\psi^{T^A}(p_i^A)$, in the fluid game on $T^A$ (see Definition A.1). Using Lemma 5.8 and the assumed $g$-continuity of the fluid game on $T^A$ we then have that $|p_i^{*,A} - p_i^A| \leq g_i(\sqrt{\delta^A})$ with $\delta^A = \frac{C}{\Lambda y_i^A}$. Recalling that $\Lambda y_i^A = (y_i^A)^{\alpha_i}$ (see e.g. the proof of Theorem A.4) we have the result of the theorem.  

**Proof of Lemma A.6:** Fix $i \in I$. We divide the proof into two cases: (i) $\alpha_i \geq 1$ and (ii) $\alpha_i < 1$.

**Case (i) $\alpha_i \geq 1$:** Fix a sequence $\{(p^A, T^A), \Lambda \geq 0\}$ such that $(p^A, T_{-i}^A) \in \mathcal{P} \times [0, T]^{I-1}$ and $(p^A, T_{-i}^A) \to (p^*, 0)$. We claim that (C5) implies, that for every $\eta > 0$,

$$\frac{\Lambda_i(p^A, T_{-i}^A, \eta y_i^A) - \Lambda_i(p^A, T_{-i}^A, 0)}{1/y_i^A} \rightarrow \eta^{\alpha_i} f_i(p^*, 0),$$  

**(EC17)**

where $f_i(\cdot, \cdot)$ is the function from (C5). Moreover, the monotonicity (as a function of $\eta$) of both the pre-limit and limit functions, guarantees that the convergence is not only pointwise (for a given $\eta$) but rather on compact sets. We will prove (EC17) at the end of the proof and we turn to prove the main assertion of the lemma for the case $\alpha_i \geq 1$. To this end, let $\{(p^A, T^A), \Lambda \geq 0\}$ be any sequence with $(p^A, T^A) \to (p^*, 0)$ as $\Lambda \to \infty$. We note that the argument in the beginning of the proof of Theorem A.4 can be repeated to show that the best service-level response is of order $O(y_i^A)$ for firm $i$. In particular, $T_{i}^{*,A} = O(y_i^A)$ so that we can find $C > 0$ such that $T_{i}^{*,A} \in [0, Cy_i^A]$ and we can write

$$T_{i}^{*,A} = \argmax_{x \in [0, Cy_i^A]} \left( \Lambda_i(p^A, T_{-i}^A, x) - \Lambda_i(p^A, T_{-i}^A, 0) \right) \left( p_i^A - c_i - \frac{\gamma_i}{\mu_i} \right)$$

$$+ \beta_i(\eta y_i^A \sqrt{R_i(p^*, 0)}) \sqrt{R_i(p^*, 0)} + o(\beta_i(\eta y_i^A \sqrt{R_i(p^*, 0)}) \sqrt{R_i(p^*, 0)}).$$  

**(EC18)**
By Lemma EC.2, the last term disappears when dividing by $1/y_i^\Lambda$ so that using that lemma and (EC17), we have that

$$\frac{T_i^{*,\Lambda}}{y_i^\Lambda} \to \eta(p^*),$$

where $\eta(p^*)$ is as defined in the Lemma. It remains to prove (EC17). To this end, let

$$T_\delta := \{(p, T_{-i}) \in \mathcal{P} \times [0, \delta]^{I-1} : \|p - p^*\| \leq \delta \text{ and } \sum_{j \neq i} T_j \leq \delta\}.$$ 

As $(p^\Lambda, T^\Lambda) \to (p^*, 0)$, we have that

$$\limsup_{\Lambda \to \infty} \frac{\Lambda_i(p^\Lambda, T_{-i}^\Lambda, \eta y_i^\Lambda) - \Lambda_i(p^\Lambda, T_{-i}^\Lambda, 0)}{1/y_i^\Lambda} \leq \sup_{(p^\Lambda, T_{-i}^\Lambda) \in T_\delta} \lim_{\Lambda \to \infty} \frac{\Lambda_i(p, T_{-i}, \eta y_i^\Lambda) - \Lambda_i(p, T_{-i}, 0)}{1/y_i^\Lambda}.$$ 

We now deal with the right hand side. Fixing $(p, T_{-i})$ and recalling that, for $\alpha_i \geq 1$, $y_i^\Lambda = \Lambda^{1/\alpha_i}$ we have that

$$\frac{\Lambda_i(p, T_{-i}, \eta y_i^\Lambda) - \Lambda_i(p, T_{-i}, 0)}{1/y_i^\Lambda} = \eta^{\alpha_i} \frac{\lambda_i(p, T_{-i}, \eta \Lambda^{1/\alpha_i}) - \lambda_i(p, T_{-i}, 0)}{\eta^{\alpha_i} \Lambda^{1/\alpha_i}} \to \eta^{\alpha_i} f_i(p, T_{-i}),$$

where the convergence follows from (C5) and the fact that $1/\Lambda \to 0$. In particular,

$$\limsup_{\Lambda \to \infty} \frac{\Lambda_i(p^\Lambda, T_{-i}^\Lambda, \eta y_i^\Lambda) - \Lambda_i(p^\Lambda, T_{-i}^\Lambda, 0)}{1/y_i^\Lambda} \leq \eta^{\alpha_i} \sup_{(p, T_{-i}) \in T_\delta} f_i(p, T_{-i}).$$

Since $\delta$ is arbitrary and $f_i(\cdot, \cdot)$ is assumed to be continuous we have that

$$\limsup_{\Lambda \to \infty} \frac{\Lambda_i(p^\Lambda, T_{-i}^\Lambda, \eta y_i^\Lambda) - \Lambda_i(p^\Lambda, T_{-i}^\Lambda, 0)}{1/y_i^\Lambda} \leq \eta^{\alpha_i} f_i(p^*, 0).$$

A similar argument can be repeated with $\liminf$ instead of $\limsup$ to conclude that

$$\lim_{\Lambda \to \infty} \frac{\Lambda_i(p^\Lambda, T_{-i}^\Lambda, \eta y_i^\Lambda) - \Lambda_i(p^\Lambda, T_{-i}^\Lambda, 0)}{1/y_i^\Lambda} = \eta^{\alpha_i} f_i(p^*, 0).$$

**Case (ii) $\alpha_i < 1$:** We write, as before,

$$\liminf_{\Lambda \to \infty} \frac{\Lambda_i(p^\Lambda, T_{-i}^\Lambda, \eta y_i^\Lambda) - \Lambda_i(p^\Lambda, T_{-i}^\Lambda, 0)}{1/y_i^\Lambda} \geq \inf_{(p^\Lambda, T_{-i}^\Lambda) \in T_\delta} \lim_{\Lambda \to \infty} \frac{\Lambda_i(p, T_{-i}, \eta y_i^\Lambda) - \Lambda_i(p, T_{-i}, 0)}{1/y_i^\Lambda}.$$ 

We recall that $y_i^\Lambda = 1/\sqrt{\Lambda}$ for all $i \in I$ with $\alpha_i \leq 1$. Fixing $(p, T_{-i}) \in T_\delta$ and $\eta > 0$ we then write

$$\frac{\Lambda_i(p, T_{-i}, \eta y_i^\Lambda) - \Lambda_i(p, T_{-i}, 0)}{1/y_i^\Lambda} = \sqrt{\Lambda} \left( \frac{\eta}{\sqrt{\Lambda}} \right)^{\alpha_i} \frac{\lambda_i(p, T_{-i}, \eta/\sqrt{\Lambda}) - \lambda_i(p, T_{-i}, 0)}{\left( \frac{\eta}{\sqrt{\Lambda}} \right)^{\alpha_i}} \to \infty.$$
Where the divergence follows from the fact that $\alpha_i < 1$ and from Assumption 5.1. Consequently,

$$\liminf_{\Lambda \to \infty} \frac{\Lambda_i(p^\Lambda, T_{-i}^\Lambda, \eta y_i^\Lambda) - \Lambda_i(p^\Lambda, T_{-i}^\Lambda, 0)}{1/y_i^\Lambda} = \infty$$

for all $\eta > 0$. Recalling that $T_i^{*,\Lambda}$ must satisfy (EC18) we must have that $T_i^{\Lambda} = o(1/\sqrt{\Lambda})$—otherwise, we can improve the profit by setting the waiting-time target to $T_i^{\Lambda} = 0$.

We conclude this proof by noting that, if $(p^\Lambda, T^\Lambda)$ is a Nash equilibrium for the $\Lambda$th diffusion game for each $\Lambda$, then the whole argument can be repeated with $T_i^{\Lambda}(p^\Lambda, T_{-i}^\Lambda)$ replacing $T_i^{*,\Lambda}$. Here we use the observation that this sequence of equilibria must satisfy $(p^\Lambda, T^\Lambda) \to 0$ as $\Lambda \to \infty$. In that case

$$T_i^{\Lambda} = \arg\max_{x \in [0, Cy]} (\Lambda_i(p^\Lambda, T_{-i}^\Lambda, x) - \Lambda_i(p^\Lambda, T_{-i}^\Lambda, 0)) \left( p_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) + \beta_i(\eta y_i^\Lambda \sqrt{R_i(p^* 0)}) \sqrt{R_i(p^* 0)},$$

and the proof can be completed identically as the proof for $T_i^{*,\Lambda}$. ■

**Proof of Theorem A.7:** From Lemma A.6 we have, for any sequence $(p^\Lambda, T^\Lambda) \to (p^*, 0)$, that

$$\frac{T_i^{*,\Lambda}(p^\Lambda, T^\Lambda)}{y_i^\Lambda} \to \eta_i^*, \ i \in I.$$  \hspace{1cm} (EC19)

Let $\eta^\Lambda = (\eta_1^*, y_1^\Lambda, \ldots, \eta_N^*, y_N^\Lambda)$ where $\{\eta_i^*, \ i \in I\}$ are the constant from Lemma A.6. Consider now an arbitrary sequence of prices $\{(p^\Lambda, \Lambda \geq 0) \} \to p^*$ and consider the sequence $\{(p^\Lambda, \eta^\Lambda) \Lambda \geq 0\}$. Let $T^\Lambda = T_i^{*,\Lambda}(p^\Lambda, \eta^\Lambda) \uparrow \eta^\Lambda$. Then,

$$\hat{\Pi}_i(p^\Lambda, T^\Lambda) - \hat{\Pi}_i(p^\Lambda, \eta^\Lambda) = \Lambda_i(p^\Lambda, T^\Lambda) \left( p_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) - \Lambda_i(p^\Lambda, \eta^\Lambda) \left( p_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) + \gamma_i \sqrt{R_i(p^*, 0)}(\beta_i(T_i^{*,\Lambda} \sqrt{R_i(p^* 0)}) - \beta_i(\eta_i^* \sqrt{R_i(p^* 0)}))$$

By (EC19) we have that $T_i^{*,\Lambda} = \eta_i^* + o(y_i^\Lambda)$ so that (by Lemma EC.1)

$$\gamma_i \sqrt{R_i(p^*, 0)}(\beta_i(T_i^{*,\Lambda} \sqrt{R_i(p^* 0)}) - \beta_i(T_i^{\Lambda} \sqrt{R_i(p^* 0)})) = o(1/y_i^\Lambda).$$  \hspace{1cm} (EC21)

From equation (EC17) in the proof of Lemma 5.12 we have that

$$\frac{\Lambda_i(p^\Lambda, \eta_{-i}^\Lambda, \eta_i^* y_i^\Lambda) - \Lambda_i(p^\Lambda, \eta_{-i}^\Lambda, 0)}{1/y_i^\Lambda} \to (\eta_i^*)^{\alpha_i} f_i(p^*, 0).$$
In particular,\[\hat{\Pi}^\Lambda_i(p^\Lambda, \eta^\Lambda_{-i}, T_i^{s, \Lambda}) \leq \hat{\Pi}^\Lambda_i(p^\Lambda, \eta^\Lambda) + o(y_i^\Lambda).\]

Consequently, for each sequence \(\{p^\Lambda, \Lambda \geq 0\}\) with \(p^\Lambda \rightarrow p^*\), we have that\[
\hat{\Pi}^\Lambda_i(p^\Lambda, \eta^\Lambda_{-i}, T_i^{s, \Lambda}) \leq \hat{\Pi}^\Lambda_i(p^\Lambda, \eta^\Lambda) + o(y_i^\Lambda),
\] (EC23)

where the first inequality follows from the definition of \(T_i^{s, \Lambda}\) as the best service-level response and the second inequality follows from (EC22).

Let \(p^\Lambda(\eta^\Lambda)\) be an equilibrium of the fluid game on \(\eta^\Lambda\). Then, we claim that
\[
p^\Lambda(\eta^\Lambda) \rightarrow p^*.
\] (EC24)

We will show (EC24) momentarily. We first use it to complete the proof of the theorem. Since \(p^\Lambda(\eta^\Lambda) \rightarrow p^*\), equation (EC23) holds with \(p^\Lambda\) there replaced by \(p^\Lambda(\eta^\Lambda)\). Fix now \(\hat{p}^\Lambda_i \neq p_i^\Lambda(\eta^\Lambda)\) and let \(\tilde{p}^\Lambda = \hat{p}^\Lambda_i \uparrow p^\Lambda\). Then, using (EC23), we have that\[
\hat{\Pi}^\Lambda_i(\tilde{p}^\Lambda, \tilde{T}^\Lambda) \leq \hat{\Pi}^\Lambda_i(\hat{p}^\Lambda, \eta^\Lambda) + o(y_i^\Lambda).
\]

Now, by the definition of \(p^\Lambda\) as the solution to the pricing game on \(\eta^\Lambda\) we have that\[
\hat{\Pi}^\Lambda_i(\tilde{p}^\Lambda, \tilde{T}^\Lambda) \leq \hat{\Pi}^\Lambda_i(\tilde{p}^\Lambda, \eta^\Lambda) + o(y_i^\Lambda) \leq \hat{\Pi}^\Lambda_i(p^\Lambda, \eta^\Lambda) + o(y_i^\Lambda),
\]

so that \((p^\Lambda, \eta^\Lambda)\) is the claimed \(\epsilon^\Lambda\)-Nash equilibrium for the \(\Lambda^{th}\) diffusion game. The bound on the deviation in service levels follows from the (already established) equality \(T_i^{s, \Lambda} = y_i^\Lambda + o(y_i^\Lambda)\).

To establish the bounds on prices, let \(p_i^{s, \Lambda}(p^\Lambda, \eta^\Lambda)\) be the best response of firm \(i\). Then, since we already showed that \((p^\Lambda, \eta^\Lambda)\) is an \(\epsilon^\Lambda\) Nash equilibrium we have that\[
o(y_i^\Lambda) \geq \hat{\Pi}^\Lambda_i(\tilde{p}^\Lambda, \tilde{T}^\Lambda) - \hat{\Pi}^\Lambda_i(p^\Lambda, \eta^\Lambda) = \Lambda_i(\tilde{p}^\Lambda, \tilde{T}^\Lambda) \left(\tilde{p}^\Lambda_i - c_i - \frac{\gamma_i}{\mu_i}\right) - \Lambda_i(p^\Lambda, \eta^\Lambda) \left(p_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i}\right) + \gamma_i \sqrt{R_i(p^\Lambda, 0)}(\beta_i(T_i^{s, \Lambda} \sqrt{R_i(p^\Lambda, 0)}) - \beta_i(y_i^\Lambda \sqrt{R_i(p^\Lambda, 0)})).
\]
Here $T^\Lambda_i = T^\Lambda_i \uparrow \eta^\Lambda$ and $p^\Lambda = p^\ast\Lambda \uparrow p^\Lambda$. Using (EC22) with $\tilde{p}^\Lambda$ and using (EC21) we then have that
\[
\Lambda_i(\tilde{p}^\Lambda, \eta^\Lambda) \left( \tilde{p}^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) - \Lambda_i(p^\Lambda, \eta^\Lambda) \left( p^\Lambda_i - c_i - \frac{\gamma_i}{\mu_i} \right) = o(y^\Lambda_i).
\]
Dividing by $\Lambda$ we then have that
\[
\left| \Pi^\Lambda_{i,P}(p^\ast\Lambda) - \Pi^T\Lambda_{i,P}(p^\Lambda) \right| \leq C \frac{1}{\Lambda y^\Lambda_i},
\]
where $\Pi^\Lambda_{i,P}(\cdot)$ is the profit function of the fluid game on $\eta^\Lambda$ (see Definition A.1). But, by definition, $p^\Lambda$ was the unique equilibrium of the fluid game on $\eta^\Lambda$ so that we may apply Lemma 5.8 and the $g$-continuity of the fluid game on $\eta^\Lambda$ to conclude that $p^*\Lambda_i = p^\Lambda_i(\eta^\Lambda) + o \left( g_i(\sqrt{\zeta^\Lambda}) \right)$ as claimed. Here we used also the fact that $\Lambda y^\Lambda_i = (y^\Lambda_i)^{\alpha_i}$ (see e.g. the proof of Theorem A.4). To establish that $(p^\Lambda(\eta^\Lambda), \eta^\Lambda)$ is also an $\epsilon^\Lambda$-Nash for the $\Lambda$th market game, one repeats the same arguments with the addition of using Lemma 5.1 to relate $\Pi^\Lambda_{j,(\cdot, \cdot)}$ to $\Pi^\Lambda_{i,(\cdot, \cdot)}$.

To complete the proof it remains to establish (EC24). The proof of this claim is very similar to the proof of the second part of Theorem 4.4. We provide the detailed argument for completeness. Towards this end, let $p^\Lambda(\eta^\Lambda)$ be an equilibrium for the fluid game on $\eta^\Lambda$ and consider a convergent subsequence $\{p^\Lambda_j(\eta^M), j \geq 1\}$ of $\{p^\Lambda(\eta^\Lambda), \Lambda \geq 0\}$. Such a sequence exists by the compactness of $\mathcal{P}$. Let $\tilde{p}$ be its limit. By definition, a Nash equilibrium of the fluid game on $\eta^\Lambda$ satisfies, for each $\hat{p}_i \in \mathcal{P}_i$, that
\[
\Pi^\Lambda_{i,P}(\hat{p}^\Lambda) \leq \Pi^\Lambda_{i,P}(p^\Lambda),
\]
where $\hat{p}^\Lambda = \hat{p}_i \uparrow p^\Lambda$. In particular, we can choose $\hat{p}_i \neq \tilde{p}_i$. Since the demand functions are continuous in their arguments and since $\eta^\Lambda \to 0$ we can take limits on both sides of (EC25) to conclude that
\[
\Pi^0_{i,P}(\hat{p}_i \uparrow \tilde{p}) \leq \Pi^0_{i,P}(\tilde{p}).
\]
Since we can repeat this for each $i \in \mathcal{I}$, we have that $\tilde{p}$ is an equilibrium for the fluid game on $T = 0$ which is the fluid game from Definition 4.2. But, by assumption, the fluid game has a unique equilibrium $p^*$ so that we must have $\tilde{p} = p^*$. Since the same argument holds for any convergent subsequence of $\{p^\Lambda(\eta^\Lambda), \Lambda \geq 0\}$ we have that $p^\Lambda(\eta^\Lambda) \to p^*$ as required.

**Lemma EC.1** Let $\{T^\Lambda_i, \Lambda \geq 0\}$ be a sequence such that $T^\Lambda_i \to 0$. Then,
\[
\beta_i(T^\Lambda_i \sqrt{R_i(p^*, 0)}) \sqrt{R_i(p^*, 0)} \sim \min \left\{ \sqrt{\Lambda_i}, \frac{1}{T^\Lambda_i} \right\}
\]
Proof: By definition,
\[ P(\beta_i(T_i^\Lambda \sqrt{R_i(p^*, 0)}) e^{-\mu_i \beta_i(T_i^\Lambda \sqrt{R_i(p^*, 0)}) T_i^\Lambda \sqrt{R_i(p^*, 0)} = \phi.} \tag{EC26} \]

Now, \( \beta_i(x) \) is smaller than \( \beta_i(0) \) and \( \beta_i(0) \) is a strictly positive constant. By Proposition 1 in Halfin and Whitt (1981) we have that \( P(\beta_i(0)) \in (0, 1) \) We now have two cases: first, if \( T_i^\Lambda \sqrt{R_i(p^*, 0)} \to 0 \), then

\[ \beta_i(T_i^\Lambda \sqrt{R_i(p^*, 0)}) \to \beta_i(0), \]

and, in particular,

\[ \beta_i(T_i^\Lambda \sqrt{R_i(p^*, 0)}) \sqrt{R_i(p^*, 0)} \sim \sqrt{\Lambda}. \]

If, on the other hand, \( \liminf_{\lambda \to \infty} T_i^\Lambda \sqrt{R_i(p^*, 0)} \to 0 \), then we can use the fact that \( P(\beta_i(0)) \) is strictly positive--and that, consequently, \( P(\beta_i(x)) \) is strictly positive for any \( x \geq 0 \)–together with (EC26) to conclude that

\[ \mu_i \beta_i(T_i^\Lambda \sqrt{R_i(p^*, 0)}) T_i^\Lambda \sqrt{R_i(p^*, 0)} \in [C_1, C_2], \]

for two strictly positive constants \( C_1 \) and \( C_2 \). From here it follows that

\[ \beta_i(T_i^\Lambda \sqrt{R_i(p^*, 0)}) T_i^\Lambda \sqrt{R_i(p^*, 0)} \sim 1/T_i^\Lambda. \]

\[ \blacksquare \]

Lemma EC.2 Fix \( \alpha_i \geq 1 \) and let \( y_i^\Lambda = \Lambda^{-\frac{1}{1+\alpha_i}} \). Then,

\[ \frac{\beta_i(\sqrt{R_i(p^*, 0)} \eta y_i^\Lambda) \sqrt{R_i(p^*, 0)}}{1/y_i^\Lambda} \to \tilde{\beta}_i(\eta) \left( \frac{\lambda_i(p^*, 0)}{\mu_i} \right) \]

uniformly on compacts sets where, given \( \eta \), \( \tilde{\beta}_i(\eta) \) is the unique solution to

\[ P(\tilde{\beta}_i)e^{-\mu_i \tilde{\beta}_i \eta \sqrt{\frac{\lambda_i(p^*, 0)}{\mu_i}}} = \phi, \]

if \( \alpha_i = 1 \) and it is the unique solution of \( e^{-\mu_i \tilde{\beta}_i \eta} = \phi \) otherwise.

Proof: First, we prove that the convergence holds pointwise, i.e, for any fixed \( \eta \). To this end, assume first that \( \alpha_i = 1 \). In this case we have that \( \eta y_i^\Lambda \sqrt{R_i(p^*, 0)} = \eta \sqrt{\lambda_i(p^*, 0)/\mu_i} \), and the result of the lemma for this case follows from the definition of \( \beta_i(\cdot) \).
Assume now that $\alpha_i > 1$. Then, by Lemma EC.1, we have that $\beta_i(\eta y_i^A \sqrt{R_i(p^*, 0)}) \sqrt{R_i(p^*, 0)} \sim 1/y_i^A$. In particular,

$$\beta_i(\eta y_i^A \sqrt{R_i(p^*, 0)}) \sim \frac{1}{\sqrt{\Lambda y_i^A}} \to 0.$$ 

In particular,

$$P(\beta_i(\eta y_i^A \sqrt{R_i(p^*, 0)})) \to 1,$$

by Proposition 1 in Halfin and Whitt (1981). Using the definition of $\beta_i(\cdot)$ we have that

$$-\mu_i \beta_i \eta y_i^A \sqrt{R_i(p^*, 0)} = \ln(\phi) + o(1),$$

so that

$$\beta_i \sqrt{R_i(p^*, 0)} = -\frac{\ln(\phi) + o(1)}{\mu_i \eta y_i^A}$$

and the result now follows by dividing by $1/y_i^A$. 

$\blacksquare$