Multiserver Queueing Systems in Heavy Traffic

by

Patrick Clark Eschenfeldt

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2017

© Massachusetts Institute of Technology 2017. All rights reserved.
Multiserver Queueing Systems in Heavy Traffic

by

Patrick Clark Eschenfeldt

Submitted to the Sloan School of Management
on January 12, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

In the study of queueing systems, a question of significant current interest is that of large scale behavior, where the size of the system increases without bound. This regime has become increasingly relevant with the rise of massive distributed systems like server farms, call centers, and health care management systems. To minimize underutilization of resources, the specific large scale regime of most interest is one in which the work to be done increases as processing capability increases. In this thesis, we characterize the behavior of two such large scale queueing systems.

In the first part of the thesis we consider a Join the Shortest Queue (JSQ) policy in the so-called Halfin-Whitt heavy traffic regime. We establish that a scaled process counting the number of idle servers and queues of length two weakly converges to a two-dimensional reflected Ornstein-Uhlenbeck process, while processes counting longer queues converge to a deterministic system decaying to zero in constant time. This limiting system is similar to that of the traditional Halfin-Whitt model in its basic performance measures, but there are key differences in the queueing behavior of the JSQ model. In particular, only a vanishing fraction of customers will have to wait, but those who do will incur a constant order waiting time.

In the second part of the thesis we consider a widely studied so-called “supermarket model” in which arriving customers join the shortest of \( d \) randomly selected queues. Assuming rate \( n \lambda_n \) Poisson arrivals and rate 1 exponentially distributed service times, our heavy traffic regime is described by \( \lambda_n \uparrow 1 \) as \( n \to \infty \). We give a simple expectation argument establishing that queues have steady state length at least \( i^* = \log_d \frac{1}{1-\lambda_n} \) with probability approaching one as \( n \to \infty \). Our main result for this system concerns the detailed behavior of queues with length smaller than \( i^* \). Assuming \( \lambda_n \) converges to 1 at rate at most \( \sqrt{n} \), we show that the dynamics of such queues does not follow a diffusion process, as is typical for queueing systems in heavy traffic, but is described instead by a deterministic infinite system of linear differential equations, after an appropriate rescaling.

Thesis Supervisor: David Gamarnik
Title: Nanyang Technological University Professor of Operations Research
Acknowledgments

I would like to thank my advisor, David Gamarnik, for his guidance in tackling the questions in this thesis. Thanks also go out to my thesis committee members Devavrat Shah and Sasha Stoylar, along with everyone at the Operations Research Center and all of my teachers over the years.
## Contents

1 Introduction

1.1 Join the shortest queue. ........................................... 14
1.2 The supermarket model. .......................................... 17
1.3 Layout of this thesis. ............................................. 21
1.4 Notations. ......................................................... 21

2 Join the Shortest Queue ........................................... 25

2.1 The model and the main result. ............................... 26
2.2 Integral representation. .......................................... 28
  2.2.1 Lower dimensions. ........................................... 29
  2.2.2 Higher dimensions. .......................................... 38
2.3 Truncation and martingale representation. ................. 41
  2.3.1 Truncation. .................................................. 41
  2.3.2 Random time change. ...................................... 42
  2.3.3 Martingales. ................................................ 43
2.4 Martingale convergence. ....................................... 47
  2.4.1 Fluid limit. .................................................. 51
  2.4.2 Proof of Theorem 2.2 ..................................... 56
2.5 Waiting time. ...................................................... 58

3 The Supermarket Model .......................................... 61

3.1 The model and the main result. ............................... 62
3.2 The model in steady state. ..................................... 68
4 Open Questions and Future Directions
List of Figures

2-1 A simulated sample path of an $M/M/n$-JSQ system, showing the scaled number of idle servers (a) and queues of length at least two ($X_2^n$), three ($X_3^n$), four ($X_4^n$), and five ($X_5^n$). Simulated with $n = 10^5$, $\beta = 2.0$. .................................................. 42

3-1 Simulated steady state expectation for $M/M/n$-Sup\((d)\) with $d = 2$ and $\lambda_n = 1 - \beta n^{-\alpha}$ with $\beta = 2$ and $\alpha = 3/4$. For each line the quantity $\alpha \log_d n$ is rounded to the nearest integer. Each horizontal line indicates $\exp(-\beta d^k/(d - 1))$, which is the conjectured limit of $\mathbb{E}S_{\alpha \log_d n + k}$ as $n$ diverges to infinity. .................................................. 73

3-2 Simulated steady state expectation for $M/M/n$-Sup\((d)\) with $d = 3$ and $\lambda_n = 1 - \beta n^{-\alpha}$ with $\beta = 0.5$ and $\alpha = 3/4$. For each line the quantity $\alpha \log_d n$ is rounded to the nearest integer. Each horizontal line indicates $\exp(-\beta d^k/(d - 1))$, which is the conjectured limit of $\mathbb{E}S_{\alpha \log_d n + k}$ as $n$ diverges to infinity. .................................................. 76

3-3 Simulated steady state expectation for $M/M/n$-Sup\((d)\) with $d = 5$ and $\lambda_n = 1 - \beta n^{-\alpha}$ with $\beta = 1$ and $\alpha = 1/4$. For each line the quantity $\alpha \log_d n$ is rounded to the nearest integer. Each horizontal line indicates $\exp(-\beta d^k/(d - 1))$, which is the conjectured limit of $\mathbb{E}S_{\alpha \log_d n + k}$ as $n$ diverges to infinity. ..................................................
Chapter 1

Introduction

Queueing theory has long been concerned with asymptotic results. Beyond the simplest of models, closed form characterization of the stochastic behavior of a queueing system is not possible, so a standard approach has been to consider regimes that allow for analytical approximation results while also providing valid models for meaningful real world questions. These approximations include the study of large deviations, analysis of heavy traffic situations, and the consideration of heavy tails.

Beyond simply being analytically tractable and related to real world questions, these asymptotic regimes are valuable specifically for the insight analytic results allow that are obfuscated by simulation. When a system is shown to converge to a diffusion process, for example, this not only provides a straightforward means for numerical approximation, it also clearly highlights how system parameters impact the behavior. It further provides a much more direct understanding of the system as a whole, which is invaluable for proper system design. In short, asymptotic results strive to take a complex system and return a simpler picture of its behavior which accurately summarizes the complexity in an appropriate setting. The appropriate asymptotic regime to choose will of course depend on the context of the system under study.

In the world of cloud computing and internet services, it is more vital than ever to consider systems with massive scale while also maintaining the efficiency of resource usage captured by traditional heavy-traffic regimes. With this goal in mind, we consider queueing systems with many parallel servers under a heavy-traffic regime
where the workload scales with the number of servers. Such systems are well understood when a global queue is maintained (i.e. \( M/M/n \) and similar queues [18]), but in many practical situations it may be advantageous to instead maintain parallel queues. Even if a global queue is itself not problematic, it may be necessary to keep queued customers close to the server who will eventually serve them. Consider for example an airport setting with arriving passengers who need to have their passports checked with one of a large number of passport controllers. In this situation having only a global queue can lead to significant walk times between the front of the queue and the server, leaving servers idle while they wait for their next customer. This idle time can be avoided by routing customers to individual queues for each server before earlier customers finish service.

At the same time, a parallel scheme will necessarily allow servers to idle if their own queue is empty, even if customers are waiting in another queue, thus sacrificing some efficiency. To analyze this tradeoff, we will study two types of parallel queueing systems both in a heavy traffic regime. Specifically, we will consider systems with \( n \) parallel queues in which each service time is exponentially distributed with parameter 1 and jobs arrive according to a Poisson process with rate \( \lambda n \). We fix \( \lambda_n < 1 \) for stability and let \( \lambda_n \) increase to 1 as \( n \) diverges to infinity. The two types of systems we study are differentiated primarily by the routing policy for newly arriving customers. In Chapter 2 we consider a system in which each arriving customer is immediately routed to the queue containing the smallest number of customers, namely the Join the Shortest Queue (JSQ) policy. In Chapter 3 we consider the so-called “supermarket model”, in which each arriving customer is immediately routed to the shortest of \( d \geq 2 \) queues chosen uniformly at random with replacement.

These two policies represent different approaches to the tradeoffs inherent in parallel queue systems. In some sense, JSQ is a logical first step for understanding these tradeoffs, because Winston [36] proved that among policies blind to the service time requirement of each job and immediately assigning customers to one of \( n < \infty \) parallel queues, JSQ is optimal in the case of Poisson arrivals and exponential service times. That is, it maximizes, with respect to stochastic order, the number of cus-
tomers served in a given time interval. Weber extended this result to the more general class of service times with non-decreasing hazard rate, with no assumptions on the arrival process. On the other hand, JSQ requires full information about the length of every queue for each routing decision, which may lead to overhead costs similar to those associated with maintaining a global queue. The supermarket model operates with more limited state information as routing decisions are made based on the lengths of only the \( d \) randomly selected queues. Thus this approach accepts some suboptimal assignments of jobs to queues in favor of allowing the routing decisions to be made without polling the state of every queue.

Our results for both models are mainly focused on describing the transient behavior of the system in the limit as the number of queues increases to infinity, under appropriate conditions on \( \lambda_n \) and the starting state of the system. For the JSQ model, we establish that in the heavy traffic regime the queueing process approaches a diffusion limit in which the scaled number of idle servers, and the scaled number of queues with one customer in service and one customer waiting, form a two-dimensional reflected Ornstein-Uhlenbeck process whereas all other queues have one customer in service and none waiting. In particular, assuming the arrival rate \( \lambda_n \) is \( \lambda_n = 1 - O\left(\frac{1}{\sqrt{n}}\right) \), we show that the number of idle servers is \( O\left(\sqrt{n}\right) \), the number of servers with one customer in service and one waiting is \( O\left(\sqrt{n}\right) \), and the number of servers with one customer in service and none waiting is \( n - O\left(\sqrt{n}\right) \).

For the supermarket model we show that in steady state most queues have length of order at least \( \log_d \frac{1}{1-\lambda_n} \), and our main result describes the process level behavior of queues focusing primarily on queues of length at most \( \log_d \frac{1}{1-\lambda_n} \). We show that under a range of appropriate conditions, the scaled process describing such queues converges to a deterministic system of linear differential equations rather than a diffusion process, as is more common for queueing systems in heavy traffic.

We will now describe our results in more detail.
1.1 Join the shortest queue.

In Chapter 2 of this thesis we consider the JSQ system in (Halfin-Whitt) heavy traffic by allowing the arrival rate $\lambda_n$ to depend on $n$, letting the quantity $(1 - \lambda_n)\sqrt{n}$ have a non-degenerate limit, which we denote $\beta > 0$. We denote this system by $M/M/n$-JSQ, distinguishing it from the traditional $M/M/n$ system which maintains a global queue.

Our main result describes the behavior of processes counting the number of idle servers and of queues with one customer waiting to enter service, along with auxiliary processes to count the number of longer queues. We prove that a system that initially has order $\sqrt{n}$ idle servers and order $\sqrt{n}$ queues with customers waiting to be served, appropriately scaled, converges weakly to a diffusion process as $n$ approaches infinity. We establish that the coordinates of this diffusion process representing queues of length at least three are deterministic in the limit and then show that these queues, if present in the initial state, disappear after a constant time and do not form again. The coordinates corresponding to the number of idle servers and number of queues of length exactly two, on the other hand, are shown to correspond to a two-dimensional reflected Ornstein-Uhlenbeck process. The entire limiting system will be defined in terms of a stochastic integral equation which we prove has a unique solution. This existence and uniqueness result is stated in Theorem 2.1 and the weak convergence result is stated in Theorem 2.2 which is our main result.

A novelty in our proof technique includes the introduction of a truncated variant of the $M/M/n$-JSQ system in which no queues of length longer than 2 are created, though such long queues are allowed in the initial condition. This system is more easily analyzed because it limits interactions between coordinates of the process. In this truncated system, we show that the probability the system hits the truncation barrier decreases to zero as $n$ approaches infinity, and thus in the limit the behavior of the truncated and untruncated systems are the same.

One consequence of our result is that over a fixed time interval with high probability both the number of idle servers and the number of queues with exactly one
customer waiting in the $M/M/n$-JSQ system are of the order $O(\sqrt{n})$. This has the following implications for the waiting time experienced by arriving customers. We prove that in the transient system the aggregate waiting time experienced by all customers is of the order $O(\sqrt{n})$, and since the number of customers arriving in that time is order $n$, the waiting time per customer is $O(1/\sqrt{n})$. At the same time, any arriving customer who has to wait incurs a waiting time which is exponentially distributed with parameter 1, which is the service time of the customer in service when they enter a queue. Since any waiting customers must incur a constant order waiting time and the aggregate waiting time is $O(\sqrt{n})$, the fraction of customers who end up waiting is of the order $O(1/\sqrt{n})$. In summary, while a small fraction of customers experience a constant order waiting time, most customers experience no wait at all, and the waiting time per arriving customer is $O(1/\sqrt{n})$, matching the expected waiting time in the $M/M/n$ system with a single queue.

Next we review relevant prior literature. The JSQ model was initially studied in the special case of 2 queues by Haight [16]. Kingman [20] proved stability results along with considering the stationary distribution of the system, and Flatto and McKean [12] also examine the stationary distribution. Further work on the $n = 2$ case includes bounds on the distribution of the number of people in the system by Halfin [17].

Foschini and Salz [13] consider diffusion limits for the heavy traffic case of the $M/M/2$-JSQ system, first proving that the queue-length processes for the two queues are identical in the limit and then deriving the limiting distribution. The limiting behavior of the waiting time is the same as the standard $M/M/2$ system in heavy traffic. Their results extend to the case of $k$ parallel queues, but they do not consider the case where the number of queues grows as the traffic intensity increases. Zhang and Wang [19] and Zhang and Hsu [37] look at a similar problem but drop the assumption of Poisson arrivals and exponential service times, deriving functional central limit theorems for the heavy traffic JSQ system with $s$ servers.

Thus our work is the first study of JSQ systems in the asymptotic regime as $n \to \infty$. Observe that for fixed $\lambda < 1$ as $n$ increases the probability of any customer arriving to find all servers busy will decrease to zero. In this case the JSQ nature
of the system becomes irrelevant as customers will be assigned to an idle server immediately upon arrival. In particular we see that the limiting behavior of the system will essentially be that of the $M/M/\infty$ system. There has been some work on models similar to ours, most notably Tezcan [31], who considers a variant of the JSQ system with multiple pools of servers who each have their own queue. He uses a state-space collapse argument based on a framework of Dai and Tezcan [6] to prove diffusion limits under the Halfin-Whitt heavy traffic regime. In this case that regime has the number of servers and traffic intensity increasing together in the limit, but the number of pools of servers is fixed so the number of queues is also fixed. Therefore our model is similar to Tezcan’s but is not a special case of it. The state-space collapse argument implies that in the limit the system can be fully described by the total number of people in the system (rather than the queue lengths in the individual pools) and the diffusion limit of that process is very similar to the original Halfin and Whitt result [18].

We note also that certain versions of the supermarket model which we discuss below are particularly closely related to JSQ. Specifically, since arriving customers join the shortest queue from among $d$ randomly selected queues, when $d$ is allowed to depend on $n$ the model is very similar to our JSQ model, which essentially sets $d = n$. Brightwell and Luczak [5] give a set of $d$ and $\lambda$ values depending on $n$ for which they prove the steady-state system is usually in a particular state with most queues having the same (known) length. Their conditions require $(1 - \lambda)^{-1} > d$, which excludes the $d = n, (1 - \lambda)\sqrt{n} \to \beta$ case considered in this chapter. Dieker and Suk [8] prove fluid and diffusion limits for queue length processes when $d$ increases to infinity at a rate slower than $n$ and with fixed $\lambda < 1$.

Another concept closely related to JSQ is the so-called “Join-Idle-Queue” (JIQ) policy, introduced by Lu, Xie, Kliot, Geller, Laurus, and Greenberg [22], in which arriving customers either join an idle server or if none are available are routed to a randomly chosen queue. An advantage of this system over JSQ is that routing decisions require information only about which queues are empty, rather than the length of all queues. Since our paper which covers the results on JSQ summarized in
Chapter 2 was posted as a preliminary draft [9], Mukherjee, Borst, van Leeuwaarden, and Whiting [27] extended those results to a class of JIQ policies. Specifically, they consider the policy JIQ\((d)\) in which arriving customers are always routed to an idle server if one exists and if not they are routed to the shortest of \(1 \leq d \leq n\) queues selected at random. For \(d = n\) this coincides with JSQ and for \(d = 1\) it coincides with traditional JIQ. The authors use a stochastic coupling argument to show that under appropriate conditions the diffusion limit for every JIQ\((d)\) policy is the same two dimensional Ornstein-Uhlenbeck process found as the diffusion limit in Chapter 2. Thus the advantages of the JSQ policy can be achieved simply by routing to idle servers when they are available.

1.2 The supermarket model.

In Chapter 3 of this thesis we consider the so-called supermarket model in the heavy traffic regime. The supermarket model is a parallel server queueing system consisting of \(n\) identical servers which process jobs at rate 1 Poisson process. The jobs arrive into the system according to a Poisson process with rate \(n\lambda\) where \(\lambda\) is assumed to be strictly smaller than unity for stability. A positive integer parameter \(d\) is fixed. Each arriving customer chooses \(d\) servers uniformly at random and selects a server with the smallest number of jobs in the corresponding queue, ties broken uniformly at random. The queue within each server is processed according to the First-In-First-Out rule. We denote this system by \(M/M/n\) Sup\((d)\).

The foundational work on this model was done by Dobrushin, Karpelevich and Vvedenskaya [32] and Mitzenmacher [26], who independently showed that when \(\lambda_n = \lambda < 1\) is a fixed constant and \(d \geq 2\), the steady state probability that the customer encounters a queue with length at least \(t\) (and hence experiences the delay at least \(t\)), is of the form \(\lambda^d t^d\). Namely it is doubly exponential in \(t\). This is in sharp contrast with the case \(d = 1\), where each server behaves as an \(M/M/1\) system with load \(\lambda\) and hence the steady state delay has the exponential tail of the form \(\lambda t\). This phenomena has its static counterpart in the form of the so-called Balls-Into-Bins model. In this
model $n$ balls are thrown sequentially into $n$ bins where for each ball $d$ bins are chosen uniformly at random and the bin with the smallest number of balls is chosen. It is well known that for this model the largest bin has $O(\log \log n)$ balls when $d \geq 2$ as opposed to $O(\log n)$ balls when $d = 1$. This known as "Power-of-Two" phenomena.

The development in [32] and [26] is based on the fluid limit approximations for the infinite dimensional process, where each coordinate corresponds to the fraction of servers with at least $i$ jobs. By taking $n$ to infinity, it is shown that the limiting system can be described by a deterministic infinite system of differential equations, which have a unique and simple to describe fixed point satisfying doubly exponential decay rate. Some of the subsequent work that has been performed on the supermarket model and its variations can be found in [2, 3, 4, 8, 14, 21, 23, 24, 25, 33].

In this chapter we consider the supermarket model in the heavy traffic regime described by having the arrival rate parameter $\lambda_n \uparrow 1$. The work of Brightwell and Luczak [5] considers the model which is the closest to the one considered in this chapter. They also assume that $\lambda_n \uparrow 1$, but at the same time they assume that the parameter $d$ diverges to infinity as well. In our setting $d$ remains constant as is the case for the classical supermarket model. More precisely, we assume that as $n$ increases, $d$ is fixed, but $\lambda_n = 1 - \beta/\eta_n$ where $\beta > 0$ is fixed and $\lim_n \eta_n = \infty$. Our goal is conducting the performance analysis of the system both at the process level and in steady state. Unfortunately, the fluid limit approach of [32] and [26] is rendered useless since in this case the corresponding fluid limit trivializes to a system of differential equations describing the critical system corresponding to $\lambda_n = \lambda = 1$. At the same time, however, certain educated guesses can be inferred from the case when $\lambda < 1$ is constant, namely the classical setting. From the $\lambda^{d_i}$ tail behavior which describes the fraction of servers with at least $i$ customers in steady state, it can be inferred that when $\lambda \uparrow 1$, if $i^* = o\left(\log_d \frac{1}{1-\lambda}\right)$ then

$$\lambda^{d_{i^*}} = \lambda^{o\left(\frac{1}{1-\lambda}\right)},$$

which approaches unity as $\lambda \uparrow 1$. Namely, the fraction of servers with at most $i^*$
customers becomes negligible. Of course this does not apply rigorously to our heavy traffic regime as it amounts to first taking the limit in \( n \) and only then taking the limit in \( \lambda \), whereas in the heavy traffic regime this is done simultaneously. Nevertheless, our main results confirm this behavior both at the process level and in the steady state regime. In terms of our notation for \( \lambda \) we show that when \( \omega_n \) is an arbitrary sequence diverging to infinity and

\[
i^*_n = \log_d \frac{1}{1 - \lambda_n} - \omega_n = \log_d \eta_n - \log_d \beta - \omega_n, \tag{1.1}
\]

(note that the term \( \log_d \beta \) is subsumed by \( \omega_n \)) the fraction of queues with at most \( i^*_n \) customers is \( o(1) \) as \( n \) increases. The intuitive explanation for this is as above:

\[
\lambda_n^{d^{i^*_n}} = (1 - \beta/\eta_n)^{d^{-\omega_n}}
= (1 - \beta/\eta_n)^{\frac{\omega_n}{d^{i^*_n}}}
\approx e^{-\frac{\omega_n}{d^{i^*_n}}}
\rightarrow 1,
\]

as \( n \to \infty \).

We now describe our results and our approach at some level of detail. First we give a very simple expectation based argument showing that in steady state the expected fraction of servers with at least \( i \) customers is at least \( 1 - \beta/\eta_n \). Plugging here the value for \( i^*_n \) given in (1.1) the expression becomes asymptotically \( 1 - 1/d^{\omega_n} \to 1 \), as \( n \to \infty \), confirming the claimed behavior in steady state. This immediately implies that the steady state delay experienced by a typical customer is at least \( i^*_n \) with probability approaching 1 as \( n \) diverges. This result is formally stated in Theorem 3.5.

Our main result concerns the detailed process level behavior of queues, with the eye towards queues of length at most \( i^*_n \). As is customary in the heavy traffic theory, the first step is applying an appropriate rescaling step, and thus for each \( i \leq i^*_n \), letting \( S^*_i(t) \) denote the fraction of servers with at most \( i \) jobs at time \( t \), we introduce a rescaled process \( T^*_i(t) \triangleq \eta_n(1 - S^*_i(t)) \). We prove that the sequence
\( T^n = (T^n_i(t), t \geq 0, i \geq 1) \) converges weakly to some deterministic limiting process \( T = (T_i(t), t \geq 0, i \geq 1) \), provided that the system starts in a state where \( T^n_i(0) \) has a non-trivial limit as \( n \to \infty \), and provided \( \eta_n \) grows at most order \( \sqrt{n} \). We show that the process \( T(t) \) is the unique solution to a deterministic infinite system of linear differential equations (given by (3.9) in the body of the chapter).

This is the main technical result of the chapter. This result is perhaps somewhat surprising since the processes arising as heavy traffic limits of queueing systems is usually a diffusion, and not a deterministic process as it is in our case. We further show that the unique fixed point of the process \( T(t) \) is of the form \( \pi_i = \pi_1 (d^i - 1)/(d - 1) \), consistent with our result regarding the lower bound on steady state expectation of \( S^n_i \) discussed above. We also show that this fixed point is an attraction point of the process \( T(t) \) and the convergence occurs exponentially fast.

Our main result regarding the weak convergence of the rescaled process \( T^n \) to \( T \) is obtained by employing several technical steps. The first step is writing the term \( (S^n_{i-1})^d - (S^n_i)^d = (1 - T^n_{i-1}/\eta_n)^d - (1 - T^n_i/\eta_n)^d \) (which corresponds to the likelihood that the arriving job increases the fraction of servers with at least \( i \) jobs), as a sum of a linear function \( dT^n_i/\eta_n - dT^n_{i-1}/\eta_n \) plus plus the correction term \( dg^n_i/\eta_n \), where \( g^n_i \) is the appropriate correction function, and then showing that this correction has a smaller order of magnitude provided \( i \leq i^*_n \). Then, we prove the existence, uniqueness and continuity property of the stochastic integral equation governing the behavior of the rescaled queue length counting process \( (T^n_i, t \geq 0, i \geq 1) \), up to an appropriately chosen stopping time intended to prevent \( T^n_i \) from "growing too much". The stopping time utilized in this theorem is similar to the one employed in Chapter 2. Finally, we apply the martingale method by splitting the underlying stochastic processes into one part which is a martingale and the compensating part which has a non-trivial drift. It is then shown that the martingale part is zero in the limit as \( n \to \infty \), thanks to the nature of the underlying rescaling.
1.3 Layout of this thesis.

In Chapter 2 we state and prove our results about the JSQ system, with the chapter laid out as follows: Section 2.1 defines the model and states our main result. In Section 2.2 we will prove Theorem 2.1, verifying that the integral representation of the limiting system is well defined. This result will also be the key to proving convergence via a continuous mapping theorem (CMT) argument. Section 2.3 will construct a representation of the system as a combination of martingales and reflecting processes. In Section 2.4 we will establish the convergence properties of these martingales, and then apply the CMT to translate the convergence of martingales to the convergence of the scaled queue length processes. This section will conclude our proof of Theorem 2.2. In Section 2.5 we discuss the waiting time in the $M/M/n$-JSQ system.

In Chapter 3 we state and prove our results about the supermarket model, with the chapter laid out as follows: Section 3.1 defines the model and states our main results. In Section 3.2 we prove our results regarding the steady state regime and prove results regarding the properties of the limiting deterministic process $T(t)$. In Section 3.3 we will prove Theorem 3.1 regarding the existence, uniqueness and the continuity property of an infinite dimensional stochastic integral equation system governing the behavior of the rescaled process $T^n(t)$. In Section 3.4 we construct a representation of the system as a combination of martingales and integral terms. Section 3.5 will establish that these martingales converge to zero. This section will also include the conclusion of the proof of Theorem 3.2.

We conclude in Chapter 4 with some discussion of the results and remaining open questions and conjectures.

1.4 Notations.

We use $\Rightarrow$ to denote weak convergence, $1\{A\}$ to denote the indicator function for the event $A$, $(x)^+ = \max(x, 0)$. $\mathbb{R}(\mathbb{R}_+)$ denotes the set of (non-negative) real values. $\mathbb{R}_{\geq 1}$ denotes the set of real values greater than or equal to one. $\mathbb{R}_+ = \mathbb{R}_+ \cup \{\infty\}$.
denotes the extended non-negative real line. \( \mathbb{R}_{\geq 1} = \mathbb{R}_{\geq 1} \cup \{\infty\} \) denotes the extended non-negative real line excluding reals strictly less than 1. We equip \( \mathbb{R}_{+} \) and \( \mathbb{R}_{\geq 1} \) with the order topology, in which neighborhoods of \( \infty \) are those sets which contain a subset of the form \( \{x > a\} \) for some \( a \in \mathbb{R} \). Let \( \mathbb{R}^\infty \) be the space of sequences \( x = (x_0, x_1, x_2, \ldots) \). For \( x \in \mathbb{R}^\infty \) and \( \rho > 1 \) we define the norm

\[
||x||_\rho \overset{\Delta}{=} \sum_{i \geq 0} \rho^{-i} |x_i|,
\]

where \( ||x||_\rho = \infty \) is a possibility, and we define the subspace

\[
\mathbb{R}^{\infty,\rho} \overset{\Delta}{=} \left\{ x \in \mathbb{R}^\infty \text{ s.t. } ||x||_\rho < \infty \right\}.
\]

Let \( D = D([0, \infty), \mathbb{R}) \) be the space of cadlag functions from \([0, \infty)\) to \( \mathbb{R} \). For \( x \in D \) we denote the uniform norm

\[
||x||_t = \sup_{0 \leq s \leq t} |x(s)|.
\]

For \( k \geq 2 \), we treat \( D^k = D([0, \infty), \mathbb{R}^k) \) as the product space \( D \times D \times \cdots \times D \) (see, e.g., [35] §3.3). We will denote the max norm

\[
||(x_1, \ldots, x_k)||_t = \max_{1 \leq i \leq k} ||x_i||_t,
\]

for \( x \in D^k \). Similarly we will use the max norm

\[
|b| = \max_{1 \leq i \leq k} |b_i|,
\]

for \( b \in \mathbb{R}^k \). Let \( D^\infty = D([0, \infty), \mathbb{R}^\infty) \) be the space of cadlag functions from \([0, \infty)\) to \( \mathbb{R}^\infty \). For \( x \in D^\infty \) and \( \rho > 1 \), we define the \( \rho \)-norm by

\[
||x||_{\rho, t} = \sum_{i \geq 0} \rho^{-i} ||x_i||_t,
\]
where $||x||_{\rho,t} = \infty$ is a possibility. We also define the subspace

$$D_{\infty,\rho} = \{x \in D^\infty : ||x||_{\rho,t} < \infty \ \forall t \geq 0\}.$$ 

For $\eta \in \mathbb{R}_+$, let $[-\eta, \eta]^{\infty} = \{x \in \mathbb{R}^\infty : \forall i \geq 0, |x_i| \leq \eta\}$. Let $\tilde{D}^\eta = D([0, \infty), [-\eta, \eta]^{\infty})$ be the space of cadlag functions from $[0, \infty)$ to $[-\eta, \eta]^{\infty}$. Let $\tilde{D}^\infty = D^\infty$ and $\tilde{D}_{\infty,\rho} = D_{\infty,\rho}$. For notational convenience, for $\rho > 1$ and $\eta \in \mathbb{R}_+$ we define $\tilde{D}^{\eta,\rho} \triangleq \tilde{D}^\eta$. 

23
In this chapter we consider perhaps the most natural routing policy for a system with parallel queues: Join the Shortest Queue. Specifically, we consider a system with \( n \) parallel queues in which each arriving customer is routed to the shortest queue. In the Halfin-Whitt heavy traffic regime, here characterized by exponential rate 1 service times and arrival rate \( \lambda_n n \) such that \((1 - \lambda_n)\sqrt{n}\) has a non-degenerate limit, we show that the processes counting the number of idle servers and of queues with exactly two customers converge to a two-dimensional Ornstein-Uhlenbeck process. If longer queues are initially present they disappear in fixed time and do not form again. This convergence implies that the aggregate waiting time experienced by arriving customers is of the order \( O(\sqrt{n}) \) and thus the waiting time per customer is \( O(1/\sqrt{n}) \). This waiting time is comparable to the \( M/M/n \) system in Halfin-Whitt heavy traffic, though the waiting time is distributed differently between customers.

The proof of our main result will follow the martingale method as applied to a truncated variant of the \( M/M/n \)-JSQ system in which no queues of length longer than 2 are allowed to form. After showing that this truncated variant converges to the limiting diffusion process we show that the truncated and untruncated systems are identical with probability approaching one as \( n \) approaches infinity.
2.1 The model and the main result.

We consider a $M/M/n$-JSQ queueing system with $n$ servers where each server maintains a unique queue, with service proceeding according to the First-In-First-Out discipline. Service time is exponentially distributed at each server, with the rate fixed at 1. Arrivals occur in a single stream, as a Poisson process with rate $\lambda_n$, where $0 < \lambda_n < 1$ and

$$\lim_{n \to \infty} \sqrt{n}(1 - \lambda_n) = \beta$$  \hspace{1cm} (2.1)

for fixed $\beta > 0$. Upon arrival, each customer is routed to the server with the shortest queue. In the event of a tie, one of the options is selected uniformly at random.

The state of the system will be represented via the process $Q^n(t) = (Q^n_1(t), Q^n_2(t), \ldots)$, with $Q^n_i(t)$ representing the number of queues with at least $i$ customers (including any customer in service) at time $t \geq 0$. We note that for the system as described we have

$$n \geq Q^n_1(t) \geq Q^n_2(t) \geq \cdots \geq 0, \quad \forall \ t \geq 0,$$  \hspace{1cm} (2.2)

and that we can recover the number of queues with exactly $i$ customers in service via the quantity $Q^n_i(t) - Q^n_{i+1}(t)$, including the number of idle servers $n - Q^n_1(t)$.

To state our weak convergence results, we also introduce a scaled version $X^n(t)$ of this process defined as

$$X^n_1(t) = \frac{Q^n_1(t) - n}{\sqrt{n}} \quad \text{and} \quad X^n_i(t) = \frac{Q^n_i(t)}{\sqrt{n}}, \quad i \geq 2.$$  \hspace{1cm} (2.3)

The $i = 1$ case is treated differently because the number of queues with length 1 behaves differently than the number of queues of all larger lengths. In particular, there will be $O(\sqrt{n})$ idle servers, and thus the number of servers with at least one customer in service will be order $n$.

Our diffusion limit will be the solution to an infinite system of integral equations, so we first introduce this system and prove that it has a unique solution. Furthermore, we prove that for $\rho > 1$ the system defines a continuous map from $\mathbb{R}_+ \times \mathbb{R}^{\infty, \rho} \times D^{\infty, \rho}$ to $D^{\infty, \rho} \times D^2$ with respect to appropriate topologies. We equip $D^{\infty, \rho}$ with the topology
induced by the $\rho$-norm $\|\cdot\|_{\rho,t}$. This continuity, along with the further fact that the function maps continuous functions to continuous functions, allows us to use the CMT to prove weak convergence once we show the weak convergence of the arguments.

**Theorem 2.1.** Given integer $B \in \mathbb{R}_+$, $b \in \mathbb{R}^{\infty,\rho}$, and $y \in D^{\infty,\rho}$, consider the following system:

\[
\begin{align*}
x_1(t) &= b_1 + y_1(t) + \int_0^t (-x_1(s) + x_2(s))ds - u_1(t), \quad (2.4) \\
x_2(t) &= b_2 + y_2(t) + \int_0^t (-x_2(s) + x_3(s))ds + u_1(t) - u_2(t), \quad (2.5) \\
x_i(t) &= b_i + y_i(t) + \int_0^t (-x_i(s) + x_{i+1}(s))ds, \quad i \geq 3, \quad (2.6) \\
x_1(t) &\leq 0, \quad x_2(t) \leq B, \quad t \geq 0, \quad (2.7)
\end{align*}
\]

with $u_1$ and $u_2$ nondecreasing nonnegative functions in $D$ such that

\[
\begin{align*}
\int_0^\infty 1\{x_1(t) < 0\}du_1(t) &= 0, \\
\int_0^\infty 1\{x_2(t) < B\}du_2(t) &= 0.
\end{align*}
\]

Then (2.4)-(2.7) has a unique solution $(x,u) \in D^{\infty,\rho} \times D^2$ so that there is a well defined function $(f,g) : \mathbb{R}_+ \times \mathbb{R}^{\infty,\rho} \times D^{\infty,\rho} \to D^{\infty,\rho} \times D^2$ mapping $(B,b,y)$ into $x = f(B,b,y)$ and $u = g(B,b,y)$. Furthermore, the function $(f,g)$ is continuous on $\mathbb{R}_+ \times \mathbb{R}^{\infty,\rho} \times D^{\infty,\rho}$ with respect to the product topology. Finally, if $y$ is continuous, then so are $x$ and $u$.

We will prove this theorem in Section 2.2. One implication of Theorem 2.1 is that the limiting system we find in our main result below is well defined because, as we will see, it is an application of the function $(f,g)$ with specific arguments $b,y$, and $B = \infty$. Note that $B = \infty$ implies $u_2 = 0$. Our main result is the following:

**Theorem 2.2.** In the sequence of $M/M/n$-JSQ models, suppose there exists a ran-
For some $\rho > 1$ such that
\[ X^n(0) \Rightarrow X(0) \quad \text{in } \mathbb{R}^{\infty,\rho} \text{ as } n \to \infty. \] (2.8)

Then for any $t \geq 0$,
\[ X^n \Rightarrow X \quad \text{in } D^{\infty,\rho} \text{ as } n \to \infty, \]
where $X = (X_1, X_2, \ldots, )$ is the unique solution in $D^{\infty,\rho}$ of the stochastic integral system (2.4)-(2.7) with $B = \infty$, $b = X(0)$, $y_1 = \sqrt{2}W(t) - \beta t$, and $y_i = 0$ for $i \geq 2$, where $W$ is a standard Brownian motion.

Remark 2.1. The integral equations for $i \geq 3$ in the limiting system of Theorem 2.2 are deterministic, and have an explicit solution:
\[ X_i(t) = e^{-t} \left( X_i(0) + \sum_{j=1}^{\infty} \frac{1}{j!} t^j X_{i+j}(0) \right), \quad i \geq 3. \]

Note that $X(0) \in \mathbb{R}^{\infty,\rho}$ implies the sum is finite.

We note that condition (2.8) does place significant but not unreasonable restrictions on the starting state of the finite systems $Q^n$. In particular, $Q^n_1(0) - n = O(\sqrt{n})$ so the number of customers initially in service must be sufficiently near $n$. Similarly, (2.8) requires $Q^n_2(0) = O(\sqrt{n})$ and therefore $Q^n_i(0) = O(\sqrt{n})$ for $i \geq 3$.

Our result shows that the $M/M/n$-JSQ system in the heavy traffic limit becomes essentially a two-dimensional system. If queues of length at least 3 are present initially, they disappear and do not form again. There are $O(\sqrt{n})$ idle servers and $O(\sqrt{n})$ queues of length exactly 2, and the behavior of processes counting these correspond to a two-dimensional Ornstein-Uhlenbeck process.

### 2.2 Integral representation.

We will now prove Theorem 2.1 showing that the representation of the limiting system in Theorem 2.2 is a valid and unique representation. We also show that it defines a continuous map from $\mathbb{R}_+ \times \mathbb{R}^{\infty,\rho} \times D^{\infty,\rho}$ to $D^{\infty,\rho} \times D^2$. The continuity of
the map in the topology of uniform convergence over bounded intervals will allow us to use the continuous mapping theorem (CMT) to demonstrate the convergence $X^n \Rightarrow X$ once we write $X^n$ in the appropriate integral form.

Note that by using $\mathbb{R}_+$ in the domain of this map we allow the upper barrier $B$ for the function $x_2$ to take the value $\infty$, which corresponds to there being no upper barrier on the $\sqrt{n}$ scale.

Our approach to proving Theorem 2.1 will involve two main lemmas, one dealing with the first two dimensions, where reflection plays an important role, and one dealing with dimensions at least three.

2.2.1 Lower dimensions.

To deal with the reflection terms in the first two dimensions of Theorem 2.1 it is convenient to consider those dimensions completely decoupled from the rest of the system. To that end we will prove the following:

Lemma 2.1. Given $B \in \mathbb{R}_+, b \in \mathbb{R}^2$, and $y \in D^2$, consider

\begin{align*}
  x_1(t) &= b_1 + y_1(t) + \int_0^t (-x_1(s) + x_2(s))ds - u_1(t), \\
  x_2(t) &= b_2 + y_2(t) + \int_0^t (-x_2(s))ds + u_1(t) - u_2(t), \\
  x_1(t) &\leq 0, \quad x_2(t) \leq B, \quad t \geq 0,
\end{align*}

with $u_1$ and $u_2$ nondecreasing nonnegative functions in $D$ such that

\begin{align*}
  \int_0^\infty 1\{x_1(t) < 0\}du_1(t) &= 0, \\
  \int_0^\infty 1\{x_2(t) < B\}du_2(t) &= 0.
\end{align*}

Then (2.9)-(2.11) has a unique solution $(x, u) \in D^2 \times D^2$ so that there is a well defined function $(f, g) : \mathbb{R}_+ \times \mathbb{R}^2 \times D^2 \to D^2 \times D^2$ mapping $(B, b, y)$ into $x = f(B, b, y)$ and $u = g(B, b, y)$. Furthermore, the function $(f, g)$ is continuous. Finally, if $y$ is continuous, then so are $x$ and $u$. 

29
We note that there has been extensive study of reflection maps in general and specifically of reflected systems of ordinary differential equations (see, e.g., [1, 29]). In our case we have a possibly discontinuous process and therefore choose to provide a complete argument to prove Lemma 2.1.

The reflection map.

In several places we will make use of the well known one-dimensional reflection map for an upper barrier. Given upper barrier \( \kappa \in \mathbb{R}_+ \), we let \((\phi_\kappa, \psi_\kappa) : D \to D^2\) be the one-sided reflection map with upper barrier at \( \kappa \) (see, e.g., [35] §5.2 and §13.5). In particular given \( y \in D \) with \( y(0) \leq \kappa \) we let \( x = \phi_\kappa(y) \) and \( z = \psi_\kappa(y) \) be the unique solution to the following system:

\[
x = y - z \leq \kappa,
\]

with \( z \) nondecreasing with \( z(0) = 0 \) and

\[
\int_0^\infty \mathbb{1}\{x < \kappa\}dz = 0.
\]

Recall that these functions can be defined explicitly by

\[
\psi_\kappa(x)(t) = \sup_{0 \leq s \leq t} (x(s) - \kappa)^+ \tag{2.12}
\]

and

\[
\phi_\kappa(x)(t) = x(t) - \psi_\kappa(x)(t). \tag{2.13}
\]

We will also make use of a slight variant of the usual Lipschitz condition for these functions to allow for different values of \( \kappa \). In particular, for \( x, x' \in D, \kappa, \kappa' \in \mathbb{R} \), and \( t \geq 0 \) we have

\[
||\psi_\kappa(x) - \psi_\kappa(x')||_t \leq ||x - x'||_t + |\kappa - \kappa'|, \tag{2.14}
\]

\[
||\phi_\kappa(x) - \phi_\kappa(x')||_t \leq 2||x - x'||_t + |\kappa - \kappa'|. \tag{2.15}
\]

30
These follow straightforwardly from (2.12) and (2.13). Note that for $\kappa = \kappa'$ we recover the usual Lipschitz constants of 1 for $\psi_\kappa$ and 2 for $\phi_\kappa$.

We also define a trivial reflection map for $\kappa = \infty$ by letting $(\phi_\infty, \psi_\infty) = (e, 0)$ where $e$ is the identity map. That is, the reflection map leaves the argument unchanged and the regulator is identically zero. We prove the following:

**Lemma 2.2.** The function $(\phi, \psi) : \mathbb{R}_+ \times D \to D^2$ defined by (2.12)-(2.13) for finite $\kappa$ and by $(\phi_\infty, \psi_\infty) = (e, 0)$ for $\kappa = \infty$ is continuous with respect to the product topology when $\mathbb{R}_+$ is equipped with the order topology and $D$ is equipped with the topology of uniform convergence over bounded intervals.

**Proof.** By (2.14)-(2.15) the function is continuous at any finite $\kappa \in \mathbb{R}_+$. For $x \in D$ and $x^\kappa \in D$ such that $x^\kappa \to x$ as $\kappa \to \infty$,

$$
\lim_{\kappa \to \infty} \|\psi_\kappa(x^\kappa)\|_t = \lim_{\kappa \to \infty} \sup_{0 \leq s \leq t} |\psi_\kappa(x^\kappa)|
= \lim_{\kappa \to \infty} \sup_{0 \leq s \leq t} (x^\kappa(s) - \kappa)^+
= \sup_{0 \leq s \leq t} \lim_{\kappa \to \infty} (x^\kappa(s) - \kappa)^+
= 0,
$$

where we have made use of the fact that $\|x\|_t < \infty$. Therefore $\psi_\kappa(x^\kappa) \to \psi_\infty(x)$ and by (2.13) we conclude $\phi_\kappa(x^\kappa) \to \phi_\infty(x)$. Thus the function is continuous at $\kappa = \infty$, completing the proof. \qed

With these facts about the reflection map in hand, we will now prove a result similar to Lemma 2.1 for a related system:

**Lemma 2.3.** Given $B \in \mathbb{R}_+, b \in \mathbb{R}^2$ and $y \in D^2$, consider

$$
w_1(t) = b_1 + y_1(t) + \int_0^t (-\phi_0(w_1)(s) + \phi_B(w_2)(s)) \, ds,
\qquad (2.16)
$$

$$
w_2(t) = b_2 + y_2(t) + \psi_0(w_1)(t) + \int_0^t (-\phi_B(w_2)(s)) \, ds \geq 0.
\qquad (2.17)
$$

31
Then (2.16)-(2.17) has a unique solution $w \in D^2$ so that there is a well defined function $\xi : \mathbb{R}_+ \times \mathbb{R}^2 \times D^2 \to D^2$ mapping $(B, b, y)$ into $w = \xi(B, b, y)$. Furthermore, the function $\xi$ is continuous. Finally, if $y$ is continuous, then so is $w$.

Before proceeding with the proof we introduce a version of Gronwall’s inequality first proved by Greene [15] and proved in the form we use by Das [7]:

**Lemma 2.4** (Gronwall’s inequality). Let $K_1$ and $K_2$ be nonnegative constants, let $h_i$ be real constants, and let $f, g$ be continuous nonnegative functions for all $t \geq 0$ such that

\[
\begin{align*}
  f(t) &\leq K_1 + h_1 \int_0^t f(s)ds + h_2 \int_0^t g(s)ds, \\
  g(t) &\leq K_2 + h_3 \int_0^t f(s)ds + h_4 \int_0^t g(s)ds
\end{align*}
\]

for all $t \geq 0$. Then

\[
\begin{align*}
  f(t) &\leq Me^{ht} \quad \text{and} \quad g(t) \leq Me^{ht}
\end{align*}
\]

for all $t \geq 0$ where $M = K_1 + K_2$ and $h = \max\{h_1 + h_3, h_2 + h_4\}$. In particular, if $K_1, K_2 = 0$, then $f(t), g(t) = 0$ for all $t$.

**Proof of Lemma 2.3.** We will show existence via a contraction mapping argument. First we will show that for $t \geq 0$ there exists a solution $\tilde{w} = (\tilde{w}_1, \tilde{w}_2)$ to the system of integral equations

\[
\begin{align*}
  \tilde{w}_1(t) &= b_1 + y_1(t) + \int_0^t \left(-\phi_0(\tilde{w}_1)(s) + \phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1))(s)\right) ds, \\
  \tilde{w}_2(t) &= b_2 + y_2(t) + \int_0^t \left(-\phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1))(s)\right) ds \geq 0.
\end{align*}
\]

Once we have such a solution, it follows immediately that

\[
w = (w_1, w_2) = (\tilde{w}_1, \tilde{w}_2 + \psi_0(\tilde{w}_1))
\]

is a solution to (2.16)-(2.17).
We first show that the map defined by the right hand side of (2.18)-(2.19) is a contraction for small enough $t$. We define $T : D^2 \to D^2$ by

$$T(\tilde{w})_1(t) = b_1 + y_1(t) + \int_0^t (-\phi_0(\tilde{w}_1)(s) + \phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1))(s)) \, ds,$$

$$T(\tilde{w})_2(t) = b_2 + y_2(t) + \int_0^t (-\phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1))(s)) \, ds.$$  

(2.20)  

(2.21)  

For $\tilde{w}, \tilde{v} \in D^2$ we have

$$||T(\tilde{w})_1 - T(\tilde{v})_1||_t \leq \int_0^t ||-\phi_0(\tilde{w}_1) + \phi_0(\tilde{v}_1)||_s \, ds$$

$$+ \int_0^t ||\phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1)) - \phi_B(\tilde{v}_2 + \psi_0(\tilde{v}_1))||_s \, ds$$

$$\leq 2 \int_0^t ||\tilde{w}_1 - \tilde{v}_1||_s \, ds$$

$$+ 2 \int_0^t ||\tilde{w}_2 + \psi_0(\tilde{w}_1) - \tilde{v}_2 - \psi_0(\tilde{v}_1)||_s \, ds$$

$$\leq 2t ||\tilde{w}_1 - \tilde{v}_1||_t + 2t ||\tilde{w}_2 - \tilde{v}_2||_t + \int_0^t ||\tilde{w}_1 - \tilde{v}_1||_s \, ds$$

$$\leq 2t ||\tilde{w}_1 - \tilde{v}_1||_t + 2t ||\tilde{w}_2 - \tilde{v}_2||_t + t ||\tilde{w}_1 - \tilde{v}_1||_t$$

$$\leq 5t ||\tilde{w} - \tilde{v}||_t$$

and

$$||T(\tilde{w})_2 - T(\tilde{v})_2||_t \leq \int_0^t ||-\phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1)) + \phi_B(\tilde{v}_2 + \psi_0(\tilde{v}_1))||_s \, ds$$

$$\leq 2t ||\tilde{w}_2 + \psi_0(\tilde{w}_1) - \tilde{v}_2 - \psi_0(\tilde{v}_1)||_t$$

$$\leq 2t ||\tilde{w}_2 - \tilde{v}_2||_t + t ||\tilde{w}_1 - \tilde{v}_1||_t$$

$$\leq 3t ||\tilde{w} - \tilde{v}||_t.$$  

We therefore conclude that

$$||T(\tilde{w}) - T(\tilde{v})||_t \leq 5t ||\tilde{w} - \tilde{v}||_t,$$
so for $t_0 < \frac{1}{6}$, $T$ is a contraction on $D([0, t_0], \mathbb{R}^2)$. Therefore by the contraction mapping principle (see, e.g., [30, p.220]), $T$ has a unique fixed point $\tilde{w}$ on $D([0, t_0], \mathbb{R}^2)$ such that $T(\tilde{w}) = \tilde{w}$. This fixed point solves (2.18)-(2.19) for $t \in [0, t_0]$. Now we extend the fixed point argument to $t \in [t_0, 2t_0], [2t_0, 3t_0], \ldots$ and repeat to find a solution $\tilde{w}$ to (2.18)-(2.19) for $t \geq 0$. As noted above, this provides a solution $w$ to (2.16)-(2.17).

To prove uniqueness of this solution, suppose $w$ and $w'$ are two solutions to (2.16)-(2.17). We consider

$$
||w_1 - w'_1||_t \leq \int_0^t ||-\phi_0(w_1) + \phi_0(w'_1) + \phi_B(w_2) - \phi_B(w'_2)||_s ds
\leq 2 \int_0^t (||w_1 - w'_1||_s + ||w_2 - w'_2||_s) ds
$$

(2.22)

and

$$
||w_2 - w'_2||_t \leq ||\psi_0(w_1) - \psi_0(w'_1)||_t + \int_0^t ||\phi_B(w_2) - \phi_B(w'_2)||_s ds
\leq ||w_1 - w'_1||_t + 2 \int_0^t ||w_2 - w'_2||_s ds.
$$

(2.23)

To match the form of Gronwall’s inequality (Lemma 2.4) we rewrite (2.23) as

$$
||w_2 - w'_2||_t - ||w_1 - w'_1||_t \leq 2 \int_0^t ||w_2 - w'_2||_s ds
$$

and note that the right hand side is nonnegative so the inequality remains true as

$$
(||w_2 - w'_2||_t - ||w_1 - w'_1||_t)^+ \leq 2 \int_0^t ||w_2 - w'_2||_s ds.
$$

(2.24)

We now define

$$
w_1(t) = ||w_1 - w'_1||_t,
$$

$$
w_2(t) = (||w_2 - w'_2||_t - ||w_1 - w'_1||_t)^+,
$$

34
and note
\[ ||w_2 - w_2'||_s \leq u_2(s) + u_1(s), \quad s \geq 0. \] \hspace{1cm} (2.25)

Then (2.22), (2.24), and (2.25) imply
\[
\begin{align*}
u_1(t) &\leq 4 \int_0^t u_1(s)ds + 2 \int_0^t u_2(s)ds, \\
u_2(t) &\leq 2 \int_0^t u_1(s)ds + 2 \int_0^t u_2(s)ds.
\end{align*}
\]

Now by Gronwall’s inequality we have
\[ u_1(t) = 0 \quad \text{and} \quad u_2(t) = 0, \]
so by the definition of \( u_1 \) and (2.25) we have
\[ ||w_1 - w_1'||_t = ||w_2 - w_2'||_t = 0 \]
for all \( t \geq 0 \) and therefore the solution \( w \) is unique.

We now establish the continuity of \( \xi \). Suppose
\[ (B^n, b^n, y^n) \to (B, b, y) \quad \text{as } n \to \infty. \]

Fix \( \epsilon > 0 \) and suppose \( w^n \) and \( w \) satisfy (2.16)-(2.17) for \((B^n, b^n, y^n)\) and \((B, b, y)\), respectively. Choose \( N \) such that for all \( n \geq N, \)
\[ |b^n - b| + ||y^n - y||_t + ||\phi_{B^n}(w_2) - \phi_B(w_2)||_t < \delta \]
for some \( \delta > 0 \) which is yet to be determined. Note that such an \( N \) exists by Lemma
and the assumption \( B^n \to B \). We have

\[
\|w^n_1 - w_1\|_t \leq |b^n - b| + \|y^n - y\|_t \\
+ \int_0^t \| -\phi_0(w^n_1) + \phi_0(w_1) + \phi_B(w_2^n) - \phi_B(w_2)\|_s ds \\
\leq \delta + \int_0^t (2\|w^n_1 - w_1\|_s + \|\phi_B(w_2^n) - \phi_B(w_2)\|_s \\
+ \|\phi_B(w_2) - \phi_B(w_1)\|_s) ds \\
\leq \delta + \int_0^t (2\|w^n_1 - w_1\|_s + 2\|w^n_2 - w_2\|_s + \delta) ds \\
\leq \delta(1 + t) + 2 \int_0^t (\|w^n_1 - w_1\|_s + \|w^n_2 - w_2\|_s) ds 
\tag{2.26}
\]

and

\[
\|w^n_2 - w_2\|_t \leq \delta(1 + t) + \|w^n_1 - w_1\|_t + 2 \int_0^t \|w^n_2 - w_2\|_s ds. 
\tag{2.27}
\]

As in the uniqueness argument above, we will apply Gronwall’s inequality, with functions

\[
u_1(t) = \|w^n_1 - w_1\|_t \quad \text{and} \quad u_2(t) = (\|w^n_2 - w_2\|_t - |w^n_1 - w_1|_t^+)\]

Then we have

\[
u_1(t) \leq \delta(1 + t) + 4 \int_0^t u_1(s)ds + 2 \int_0^t u_2(s)ds, \\
u_2(t) \leq \delta(1 + t) + 2 \int_0^t u_1(s)ds + 2 \int_0^t u_2(s)ds,
\]

so Gronwall’s inequality implies

\[
u_1(t) \leq 2\delta(1 + t)e^{6t} \quad \text{and} \quad u_2(t) \leq 2\delta(1 + t)e^{6t}
\]
and we have
\[ ||w_1^n - w_1||_t \leq 2\delta(1 + t)e^{6t} \quad \text{and} \quad ||w_2^n - w_2||_t \leq 4\delta(1 + t)e^{6t}.\]

We choose \( \delta = \frac{1}{44t}e^{-6t} \) to establish the desired continuity.

For the proof of continuity of \( w \) we note
\[
|w_1(t + s) - w_1(t)| \leq |y_1(t + s) - y_1(t)| + \int_t^{t+s} |\phi_B(w_2(z)) - \phi_0(w_1(z))|dz
\]
and
\[
|w_2(t + s) - w_2(t)| \leq |y_2(t + s) - y_2(t)| + |w_1(t + s) - w_1(t)| + \int_t^{t+s} |\phi_B(w_2(z))|dz
\]
The boundedness of \( w_1 \) and \( w_2 \) proved in Lemma 2.9 imply that \( w_1 \) and \( w_2 \) are continuous if \( y_1 \) and \( y_2 \) are continuous.

We are now prepared to prove Lemma 2.1.

**Proof of Lemma 2.1.** Our key insight is to see that a solution is found by setting \( x_1 = \phi_0(w_1) \), \( u_1 = \psi_0(w_1) \), \( x_2 = \phi_B(w_2) \), and \( u_2 = \psi_B(w_2) \) where \( (w_1, w_2) \) is the unique solution defined by Lemma 2.3.

To see that it is unique, note that the conditions on \( u_1 \) and \( u_2 \) imply that they can be written as \( \psi_0(z_1) \) and \( \psi_B(z_2) \) for some functions \( z_1, z_2 \in D_t \). Then \( x_1 \) and \( x_2 \) are \( \phi_0(z_1) \) and \( \phi_B(z_2) \) for the same \( z_1 \) and \( z_2 \). Then (2.9)–(2.11) imply that \( z = (z_1, z_2) \) must be a solution of (2.16)–(2.17). By Lemma 2.3 this solution is unique. In particular, this solution is
\[
x_1 = f_1(b, y) = (\phi_0 \circ \xi_1)(b, y),
\]
\[
u_1 = g_1(b, y) = (\psi_0 \circ \xi_1)(b, y),
\]
\[
x_2 = f_2(b, y) = (\phi_B \circ \xi_2)(b, y),
\]
\[
u_2 = g_2(b, y) = (\psi_B \circ \xi_2)(b, y).
\]
The reflection maps \((\phi_0, \psi_0)\) and \((\phi_B, \psi_B)\) are continuous in the uniform topology and also preserve continuity. Since \(\xi\) also has these properties by Lemma 2.3, we conclude that \((f, g)\) are continuous and preserve continuity.

2.2.2 Higher dimensions.

Because of the lack of reflection terms for \(i \geq 3\), the higher dimensional terms in Theorem 2.1 can be treated fairly simply as a system in \(D^{\infty, \rho}\), indexed starting at \(i = 3\) to avoid confusion with the lower dimensional terms in the full system.

**Lemma 2.5.** Given \(b \in \mathbb{R}^{\infty, \rho}\), and \(y \in D^{\infty, \rho}\), consider the following system of integral equations:

\[
\begin{align*}
x_1(t) &= x_2(t) = 0, \quad (2.28) \\
x_i(t) &= b_i + y_i(t) + \int_0^t -(x_i(s) + x_{i+1}(s)) \, ds, \quad i \geq 3, \quad (2.29) \\
t &\geq 0. \quad (2.30)
\end{align*}
\]

Then (2.28)-(2.30) has a unique solution \(x \in D^{\infty, \rho}\) so that there is a well defined function \(f : \mathbb{R}^{\infty, \rho} \times D^{\infty, \rho} \to D^{\infty, \rho}\) mapping \((b, y)\) into \(x = f(b, y)\). Furthermore, the function \(f\) is continuous. Finally, if \(y\) is continuous, then so is \(x\).

**Proof.** We first prove existence and uniqueness via a contraction mapping argument. Define a map \(\Gamma : D^{\infty, \rho} \to D^{\infty}\) by \(\Gamma(x)_1(t) = \Gamma(x)_2(t) = 0\) and

\[
\Gamma(x)_i(t) = b_i + y_i(t) + \int_0^t (-x_i(s) + x_{i+1}(s)) \, ds, \quad i \geq 3.
\]

Observe

\[
\|\Gamma(x)\|_{\rho, t} \leq \|b\|_{\rho} + \|y\|_{\rho, t} + t \sum_{i \geq 3} \rho^{-i} \|x_i\|_t + t \sum_{i \geq 3} \rho^{-i} \|x_{i+1}\|_t \\
= \|b\|_{\rho} + \|y\|_{\rho, t} + t \sum_{i \geq 3} \rho^{-i} \|x_i\|_t + t \rho \sum_{i \geq 4} \rho^{-i} \|x_i\|_t \\
\leq \|b\|_{\rho} + \|y\|_{\rho, t} + t(1 + \rho) \|x\|_{\rho, t} < \infty,
\]

38
so we have $\Gamma : D^{\infty, \rho} \to D^{\infty, \rho}$. Let $x^1, x^2 \in D^{\infty, \rho}$ and consider

$$
||\Gamma(x^1) - \Gamma(x^2)||_{\rho, t} = \sum_{i \geq 3} \rho^{-i} (||\Gamma(x^1)_i - \Gamma(x^2)_i||_t) \\
\leq \sum_{i \geq 3} \rho^{-i} (t ||x^1_i - x^2_i||_t + t ||x^1_{i+1} - x^2_{i+1}||_t) \\
\leq t(1 + \rho) ||x^1 - x^2||_{\rho, t}
$$

For $t_0 < (1 + \rho)^{-1}$, we have $\Gamma$ is a contraction on

$$D^{\infty, \rho}_{t_0} = \left\{ x \in D \left([0, t_0], \mathbb{R}^\infty\right) : ||x||_{\rho, t} < \infty, \quad \forall t \in [0, t_0] \right\}.
$$

By the contraction mapping principle, $\Gamma$ has a unique fixed point, which solves (2.28)–(2.30) for $t \in [0, t_0]$. We extend the fixed point argument to $t \in [t_0, 2t_0], [2t_0, 3t_0], \ldots$ to construct a unique fixed point for $t \geq 0$.

We next prove continuity. Fix $\epsilon > 0$. Suppose $(b^n, y^n) \Rightarrow (b, y)$, and let $x^n$ be the unique solution for $(b^n, y^n)$ and $x$ be the unique solution for $(b, y)$. Consider

$$
||x^n - x||_{\rho, t} = \sum_{i \geq 3} \rho^{-i} ||x^n_i - x_i||_t \\
\leq \sum_{i \geq 3} \rho^{-i} \left( ||b^n - b||_t + ||y^n - y||_t + \int_0^t \left(||x^n_i - x_i||_s + ||x^n_{i+1} - x_{i+1}||_s\right) ds\right) \\
\leq ||b^n - b||_{\rho} + ||y^n - y||_{\rho, t} + (1 + \rho) \int_0^t ||x^n - x||_s ds.
$$

Choose $N$ such that for $n \geq N$ we have $||b^n - b||_t + ||y^n - y||_{\rho, t} < \delta$ for any $\delta > 0$. We have for $n \geq N$,

$$
||x^n - x||_{\rho, t} < \delta + (1 + \rho) \int_0^t ||x^n - x||_s ds.
$$

By Gronwall’s inequality (Lemma 3.1 with $g = 0$), we have

$$
||x^n - x||_{\rho, t} \leq \delta e^{(1+\rho)t}.
$$
Thus for \( \delta < \varepsilon e^{-(1+\rho)t} \), we have
\[
\|(x^n - x)\|_{\rho,t} < \epsilon
\]
and continuity is established.

To establish the inheritance of continuity, observe for \( i \geq 3 \)
\[
|x_i(t + s) - x_i(t)| \leq |y_i(t + s) - y_i(t)| + \int_t^{t+s} |x_i(u) - x_{i+1}(u)|du.
\]
The continuity then follows. \( \square \)

With this lemma in hand, we can prove Theorem 2.1:

**Proof of Theorem 2.1.** We first prove existence, uniqueness, and preservation of continuity by combining Lemmas 2.3 and 2.5. Lemma 2.5 applied to coordinates \( i \geq 3 \) implies a unique \( x_i \) exists for \( i \geq 3 \) and \( x_i \) is continuous if \( y_i \) is continuous for \( i \geq 3 \).

We now define
\[
\dot{y}_2(t) = y_2(t) + \int_0^t x_3(s)ds,
\]
and apply Lemma 2.3 with \( y_2 = \dot{y}_2 \) to complete the proof of existence, uniqueness, and preservation of continuity.

To verify that the map \((f, g)\) is continuous, suppose \( x^n(t) \) and \( x(t) \) solve (2.4)-(2.7) for \((B^n, b^n, y^n)\) and \((B, b, y)\), respectively, and further suppose \((B^n, b^n, y^n) \to (B, b, y)\) as \( n \to \infty \).

By Lemma 2.5 we have \( x^n_i \to x_i \) for \( i \geq 3 \). This further implies
\[
y_2^n(t) + \int_0^t x_3^n(s)ds \to y_2(t) + \int_0^t x_3(s)ds,
\]
so Lemma 2.1 implies \((x_1^n, x_2^n) \to (x_1, x_2)\), and continuity is established. \( \square \)
2.3 Truncation and martingale representation.

To use Theorem 2.1 in a CMT argument, we want to write the process $X^n$ in the appropriate integral form. Instead of directly considering the full $M/M/n$-JSQ system, however, we will introduce a truncated variant which we will later show has the same behavior as $X^n$ in the limit. It is this truncated system that will be shown to take the integral form of Theorem 2.1.

2.3.1 Truncation.

An important feature of the behavior of the $M/M/n$-JSQ system is that queues with more than one customer waiting are formed only if all queues have at least one customer waiting already. One of our goals is to show that $Q^n_2$, the number of queues with a customer waiting, is of the order $O(\sqrt{n})$ and thus it is unlikely for longer queues to form. With this in mind, as a proof technique, we now introduce a truncated version of the system in which no queues with length greater than 2 are created, though they are allowed to exist in the initial condition. This system will be significantly easier to analyze and will be shown to have stochastic behavior exactly matching the untruncated system with high probability. See Figure 2-1 to see a sample path of an original untruncated system which starts with order $\Theta(\sqrt{n})$ queues of length 4. Note that the number of queues length 3 and length 4 decrease monotonically.

Now we make this more precise. Consider a system in which any arrival that would create a queue of length 3 or longer is rejected. That is, if an arrival occurs when all queues contain at least two customers, that arriving customer does not enter the system. We will denote this system $\hat{Q}^n = (\hat{Q}^n_1, \hat{Q}^n_2, \ldots)$. We also introduce scaled versions

$$\hat{X}^n_1(t) = \frac{\hat{Q}^n_1(t) - n}{\sqrt{n}} \quad \text{and} \quad \hat{X}^n_i(t) = \frac{\hat{Q}^n_i(t)}{\sqrt{n}}, \quad i \geq 2. \quad (2.31)$$

We will now construct the truncated process $\hat{X}^n(t)$ and show that it has the integral form in Theorem 2.1. Our representation will be similar to the first martingale.
representation of [28]; in particular it will rely upon random time changes of rate-1 Poisson processes.

\subsection{Random time change.}

We let $A, D_i$ for $i \geq 1$ be rate-1 Poisson processes and write

\begin{align*}
\hat{Q}_1^n(t) &= Q_1^n(0) + A(\lambda_n t) - D_1 \left( \int_0^t \left( \hat{Q}_1^n(s) - \hat{Q}_2^n(s) \right) ds \right) - \hat{U}_1^n(t), \quad (2.32) \\
\hat{Q}_2^n(t) &= Q_2^n(0) + \hat{U}_1^n(t) - D_2 \left( \int_0^t \left( \hat{Q}_2^n(s) - \hat{Q}_3^n(s) \right) ds \right) - \hat{U}_2^n(t), \quad (2.33) \\
\hat{Q}_i^n(t) &= Q_i^n(0) - D_i \left( \int_0^t \left( \hat{Q}_i^n(s) - \hat{Q}_{i+1}^n(s) \right) ds \right), \quad i \geq 3 \quad (2.34)
\end{align*}

where $\hat{U}_1^n(t)$ is the number of arrivals in $[0, t]$ when every server has at least one customer, and $\hat{U}_2^n(t)$ is the number of arrivals in $[0, t]$ when every server has at least

Figure 2-1: A simulated sample path of an $M/M/n$-JSQ system, showing the scaled number of idle servers (a) and queues of length at least two ($X_2^n$), three ($X_3^n$), four ($X_4^n$), and five ($X_5^n$). Simulated with $n = 10^5, \beta = 2.0$. 

(a) Idle servers  

(b) Longer queues
one customer \textit{and} all $n$ servers have two customers. Formally, we define
\begin{align}
\hat{U}_1^n(t) &= \int_0^t \mathbbm{1}\{\hat{Q}_1^n(s) = n\} dA(\lambda_n n s), \quad (2.35) \\
\hat{U}_2^n(t) &= \int_0^t \mathbbm{1}\{\hat{Q}_1^n(s) = n, \hat{Q}_2^n(s) = n\} dA(\lambda_n n s). \quad (2.36)
\end{align}

We can understand (2.32) term-by-term: first we record the initial state of the system with $Q_1^n(0)$, then arrivals are counted at their full rate $\lambda_n n$. The $D_1$ term represents departures, which occur as a Poisson process with rate equal to the number of customers in service. Since $\hat{Q}_1$ includes queues of length 1 and length 2, however, $\hat{Q}_1$ will only decrease when a customer departs a queue and leaves the server empty. Therefore the instantaneous rate at time $s$ in the $D_1$ term is $\hat{Q}_1^n(s) - \hat{Q}_2^n(s)$, the number of queues of length exactly 1 at time $s$. Through the first three terms of (2.32) we have recorded what the value of $\hat{Q}_1$ would be if it were not constrained to be at most $n$, so the final term will represent this barrier. The process $\hat{U}_1^n$ records any arrival which would increase $\hat{Q}_1$ above $n$, balancing the overcounting we get from $A(\lambda_n n t)$.

We can understand (2.33) in much the same way, with the key difference being in the arrival process. Since arriving customers will always join the shortest available queue, the number of length 2 queues will increase only when all servers are busy. Such arrivals are exactly recorded by $\hat{U}_1^n$, so this will be the process we use to record potential increases to $\hat{Q}_2^n$. The process $\hat{U}_2^n$ provides the upper barrier $n$ on $\hat{Q}_2^n$.

The remaining equation (2.34) is the same except that we record no arrivals, as our truncated approximation does not create queues of length 3 or longer.

As in [28, Lemma 2.1], we can verify that this construction is well defined and generates an element of $D^\infty$ by conditioning on the starting state $Q^n(0)$ and processes $A, D_i$ then constructing recursively. Note that $\|\hat{Q}^n\|_{\rho,t} \leq \sum_{i \geq 1} \rho^{-i} n < \infty$.

2.3.3 Martingales.

Because our approach to (2.32)-(2.34) will be to apply the functional central limit theorem (FCLT) for Poisson processes, we will now rewrite the time changes of Poisson
processes as time changes of scaled Poisson processes. To that end, we define scaled martingales

\[
\hat{M}_0^n(t) = \frac{1}{\sqrt{n}} A(\lambda_n nt) - \lambda_n \sqrt{nt},
\]

(2.37)

\[
\hat{M}_i^n(t) = \frac{1}{\sqrt{n}} D_i \left( \int_0^t \left( \hat{Q}_i^n(s) - \hat{Q}_{i+1}^n(s) \right) ds \right) - \frac{1}{\sqrt{n}} \int_0^t \left( \hat{Q}_i^n(s) - \hat{Q}_{i+1}^n(s) \right) ds, \quad i \geq 1.
\]

(2.38)

Via an argument exactly analogous to that of in §7.1 of [28] leading to Theorem 7.2 we obtain that \( \hat{M}_i^n \) for \( i \geq 0 \) are square-integrable martingales with respect to an appropriate filtration, namely \( F_n = \{ F_{n,t} : t \geq 0 \} \) for

\[
F_{n,t} = \sigma \left( Q^n(0), A(\lambda_n ns), D_i \left( \int_0^s \left( \hat{Q}_i^n(u) - \hat{Q}_{i+1}^n(u) \right) du \right) ; i \geq 1, 0 \leq s \leq t \right), \quad t \geq 0,
\]

augmented by including all null sets. We note for later use that this argument also supplies the predictable quadratic variations

\[
\left\langle \hat{M}_0^n \right\rangle(t) = \lambda nt,
\]

(2.39)

\[
\left\langle \hat{M}_i^n \right\rangle(t) = \frac{1}{n} \int_0^t (\hat{Q}_i^n(s) - \hat{Q}_{i+1}^n(s)) ds.
\]

(2.40)

We now provide an overview of this argument. We begin by considering versions of (2.37) - (2.38) and their associated filtrations which are not indexed by \( n \). Specifically, we consider

\[
\hat{M}_0(t) = A(\lambda nt) - \lambda nt,
\]

\[
\hat{M}_i^n(t) = D_i \left( \int_0^t \left( \hat{Q}_i(s) - \hat{Q}_{i+1}(s) \right) ds \right) - \int_0^t \left( \hat{Q}_i(s) - \hat{Q}_{i+1}(s) \right) ds, \quad i \geq 1,
\]

and the analogous filtration \( F = \{ F_t : t \geq 0 \} \) for

\[
F_t = \sigma \left( Q(0), A(\lambda ns), D_i \left( \int_0^s \left( \hat{Q}_i(u) - \hat{Q}_{i+1}(u) \right) du \right) ; i \geq 1, 0 \leq s \leq t \right), \quad t \geq 0.
\]
We define a multiparameter filtration with a countably infinite number of parameters:

\[
\mathcal{H} = \mathcal{H}(t_i, i \geq 0) = \sigma \left( Q(0), A(s_0), D_i(s_i) : 0 \leq s_i \leq t_i, i \geq 0 \right).
\]

We define nondecreasing nonnegative stochastic processes

\[
I_0(t) = \lambda nt
\]
\[
I_i(t) = \int_0^t \left( \hat{Q}_i(s) - \hat{Q}_{i+1}(s) \right) ds, \quad i \geq 1, t \geq 0
\]

and observe that the vector \((I_i(t), i \geq 0)\) is an \(\mathcal{H}\)-stopping time. Because \(A\) and \(D_i, i \geq 1\) are independent rate-1 Poisson processes, if we define

\[
\bar{M}_0(t) = A(t) - t \quad \text{and} \quad \bar{M}_i(t) = D_i(t) - t, \quad i \geq 1,
\]

and

\[
\bar{M} = (\bar{M}_i, i \geq 0) = \left\{(\bar{M}_i(s_i), i \geq 0) : s_i \geq 0 \right\},
\]

then \(\bar{M}\) is an \(\mathcal{H}\)-multiparameter martingale. By the optional stopping theorem \[\text{[III]}, \]
\((\bar{M}_i \circ I_i, i \geq 0)\) is a martingale with respect to

\[
\mathcal{H}(I(t)) = \sigma \left( Q(0), A(s_0), D_i(s_i) : 0 \leq s_i \leq I_i(t), i \geq 0 \right)
\]
\[
= \sigma \left( Q(0), A(\lambda ns), D_i \left( \int_0^s \left( \hat{Q}_i(u) - \hat{Q}_{i+1}(u) \right) du \right) : i \geq 1, 0 \leq s \leq t \right)
\]
\[
= \mathcal{F}_t.
\]

Crudely bounding \(Q_1(t) \leq Q(0) + A(\lambda nt)\) allows us to show \(\mathbb{E}[I_i(t)] < \infty\) and \(\mathbb{E}[M_i(t)] < \infty\) for \(i \geq 0\) to guarantee that the moment conditions of the optional stopping theorem are satisfied. These results then translate naturally to the \(n\)-indexed and scaled versions defined in \([2.37]-[2.38]\).
We also define
\[ \hat{V}_1^n(t) = \frac{\hat{U}_1^n(t)}{\sqrt{n}} \quad \text{and} \quad \hat{V}_2^n(t) = \frac{\hat{U}_2^n(t)}{\sqrt{n}}. \]

Then we have
\[ \hat{X}_1^n(t) = \frac{\hat{Q}_1^n(t) - n}{\sqrt{n}} \]
\[ = \frac{Q_1^n(0) - n}{\sqrt{n}} + \frac{1}{\sqrt{n}} A(\lambda_nt) \]
\[ - \frac{1}{\sqrt{n}} D_1 \left( \int_0^t (\hat{Q}_1^n(s) - \hat{Q}_2^n(s)) ds \right) - \frac{\hat{U}_1^n(t)}{\sqrt{n}} \]
\[ = X_1^n(0) + \hat{M}_0^n(t) + \lambda_n \sqrt{nt} - \hat{V}_1^n(t) \]
\[ - \hat{M}_1^n(t) - \frac{1}{\sqrt{n}} \int_0^t \left( \hat{Q}_1^n(s) - \hat{Q}_2^n(s) \right) ds \]
\[ = X_1^n(0) + \hat{M}_0^n(t) - \hat{M}_1^n(t) + \lambda_n \sqrt{nt} - \hat{V}_1^n(t) \]
\[ - \sqrt{nt} - \int_0^t \left( \frac{\hat{Q}_1^n(s) - n}{\sqrt{n}} - \frac{\hat{Q}_2^n(s)}{\sqrt{n}} \right) ds \]
\[ = X_1^n(0) + \hat{M}_0^n(t) - \hat{M}_1^n(t) - (1 - \lambda_n) \sqrt{nt} \]
\[ - \int_0^t (\hat{X}_1^n(s) - \hat{X}_2^n(s)) ds - \hat{V}_1^n(t), \]
(2.41)

and
\[ \hat{X}_2^n(t) = X_2^n(0) + \hat{V}_1^n(t) - \hat{M}_2^n(t) - \int_0^t \left( \hat{X}_2^n(s) - \hat{X}_3^n(s) \right) ds - \hat{V}_2^n(t), \]
(2.42)
\[ \hat{X}_i^n(t) = X_i^n(0) - \hat{M}_i^n(t) - \int_0^t \left( \hat{X}_i^n(s) - \hat{X}_{i+1}^n(s) \right) ds, \quad i \geq 3. \]
(2.43)

At this point we can also note that (2.41)-(2.43) put \( \hat{X}^n(t) \) in the integral form of Theorem 2.1. The only difference is the processes \( \hat{V}^n \), which are not described in
exactly the same way. We see, however, that by (2.35) we have

\[
0 = \int_0^\infty \mathbb{1}\{\hat{Q}_1^n(s) < n\} d\hat{U}_1^n(s) \\
= \int_0^\infty \mathbb{1}\{\hat{X}_1^n(s) < 0\} d\hat{U}_1^n(s) \\
= \int_0^\infty \mathbb{1}\{\hat{X}_1^n(s) < 0\} d\hat{V}_1^n(s). \tag{2.44}
\]

Similarly by (2.36) we have

\[
0 = \int_0^\infty \mathbb{1}\{\hat{Q}_1^n(s) < n \text{ or } \hat{Q}_2^n(s) < n\} d\hat{U}_2^n(t) \\
\]

which implies

\[
0 = \int_0^\infty \mathbb{1}\{\hat{Q}_2^n(s) < n\} d\hat{U}_2^n(t) \\
= \int_0^\infty \mathbb{1}\{\hat{X}_2^n(s) < \sqrt{n}\} d\hat{V}_2^n(t). \tag{2.45}
\]

Thus by (2.41)-(2.43) and (2.44)-(2.45) \( \hat{X}^n \) is the unique solution of (2.4)-(2.7) for \( b = X^n(0), y_i = \hat{M}^n_i(t) - \hat{M}^n_i(t) - (1 - \lambda_n) \sqrt{n} t, y_i = -\hat{M}^n_i(t), i \geq 2 \), and \( B = \sqrt{n} \).

Because \( X^n_i(0) \leq n \) for all \( i \geq 1 \), \( ||b||_\rho < \infty \) and thus \( b \in R^{\infty, \rho} \). We will show in Section 2.4 that \( y \in D^{\infty, \rho} \). Then in order to apply the CMT it remains to prove the convergence of the martingales (2.37)-(2.38).

### 2.4 Martingale convergence.

We now prove the convergence of \( \hat{M}^n_i \) to Brownian motions.

**Lemma 2.6.** For the sequences of scaled martingales \( \hat{M}^n_i, i \geq 0 \) defined in Section 2.3.3 we have

\[
\left( \hat{M}^n_0, \hat{M}^n_1, \hat{M}^n_2, \ldots, \hat{M}^n_i, \ldots \right) \Rightarrow (W_1, W_2, 0, \ldots, 0, \ldots) \text{ in } D^{\infty, \rho} \text{ as } n \to \infty, \tag{2.46}
\]
where $W_1$ and $W_2$ are independent standard Brownian motions. Furthermore, for sufficiently large $n$, $\left\| \hat{M}^n \right\|_{\rho,t} < \infty$ almost surely.

To prove this lemma we rely upon the CMT and the FCLT for Poisson processes ([28, Theorem 4.2]), which we restate here convenience.

**Lemma 2.7.** (FCLT for independent Poisson processes) For $k \geq 1$, if $A$ and $D_i$, $1 \leq i \leq k$ are independent rate-1 Poisson processes and

$$M_{C,n}(t) = \frac{C(nt) - nt}{\sqrt{n}}$$

for $C = A, D_i$, then

$$(M_{A,n}, M_{D_1,n}, \ldots, M_{D_k,n}) \Rightarrow (W_1, W_2, \ldots, W_{k+1}) \text{ in } D^{k+1} \text{ as } n \to \infty$$

where $W_i$ are independent standard Brownian motions.

We will use Lemma 2.7 with $k = 2$ for the first three coordinates of $\hat{M}^n$. To apply Lemma 2.7, we define random and deterministic time changes such that the martingales $\hat{M}^n$ can be written as a composition of a time change and the scaled Poisson processes $M_{C,n}$. Specifically, let

$$\Phi_{A,n}(t) = \frac{\lambda_n t}{n},$$

$$\Phi_{D_i,n}(t) = \frac{1}{n} \int_0^t \hat{Q}^n_i(s) ds - \frac{1}{n} \int_0^t \hat{Q}^n_{i+1}(s) ds,$$

so that we have

$$\hat{M}^n_0 = \hat{M}_{A,n} \circ \Phi_{A,n}, \quad \hat{M}^n_i = \hat{M}_{D_i,n} \circ \Phi_{D_i,n}, \quad i \geq 1.$$
where \( e \) is the identity function in \( D \).

Next we note that the terms of (2.48) interleave over \( i \), so for \( i \geq 1 \) we have

\[
\Phi_{D_i,n} \Rightarrow f_i - f_{i+1} \text{ as } n \to \infty,
\]

where \( f_i \) is the limit of \( \tilde{\Phi}_{D_i,n} \) with

\[
\tilde{\Phi}_{D_i,n}(t) = \frac{1}{n} \int_{0}^{t} \hat{Q}_i^n(s) ds.
\]

To find \( f_i \) we first show fluid limits for \( \hat{Q}_i^n \).

**Lemma 2.8.** Let \( \Psi_i^n \) for \( i \geq 1 \) be defined by

\[
\Psi_i^n(t) = \frac{\hat{Q}_i^n(t)}{n}, \quad t \geq 0.
\]

Then for \( i \geq 2 \),

\[
\Psi_i^n \Rightarrow \omega \quad \text{and} \quad \Psi_i^n \Rightarrow 0 \quad \text{as } n \to \infty \quad (2.50)
\]

where \( \omega(t) = 1 \) for \( t \geq 0 \).

The proof of this lemma is found in Section 2.4.1. To use Lemma 2.8 we define a continuous function \( h : D \to D \) by

\[
h(x)(t) = \int_{0}^{t} x(s) ds
\]

for \( t \geq 0 \). Then \( \tilde{\Phi}_{D_1,n} = h \circ \Psi_n \) so by the CMT and Lemma 2.8 we know \( f_1 = h \circ \omega \). Namely,

\[
f_1(t) = \int_{0}^{t} 1 ds = t,
\]

so \( f_1 = e \) is the identity function in \( D \). Therefore we have

\[
\tilde{\Phi}_{D_1,n} \Rightarrow e \quad \text{as } n \to \infty.
\]
For $i \geq 2$ we have $f_i(t) = \int_0^t 0 \, ds = 0$ so $f_i = 0$ on $D$. We conclude that

$$\Phi_{D_{i,n}} \Rightarrow e \quad \text{as } n \to \infty, \quad (2.51)$$

and for $i \geq 2$

$$\Phi_{D_{i,n}} \Rightarrow 0 \quad \text{as } n \to \infty. \quad (2.52)$$

Therefore once we establish Lemma 2.8 we can prove Lemma 2.6:

**Proof of Lemma 2.6.** We apply the CMT with Lemma 2.7 and the limits (2.49), (2.51), and (2.52) to obtain

$$\left(\hat{M}_0^n, \hat{M}_1^n, \hat{M}_2^n\right) = (M_{A,n} \circ \Phi_{A,n}, M_{D_{1,n}} \circ \Phi_{D_{1,n}}, M_{D_{2,n}} \circ \Phi_{D_{2,n}})$$

$$\Rightarrow (W_1 \circ e, W_2 \circ e, W_3 \circ 0)$$

$$= (W_1, W_2, 0).$$

This implies $E \left\| \hat{M}_0^n - W_1 \right\|_t \to 0, E \left\| \hat{M}_1^n - W_2 \right\|_t \to 0,$ and $E \left\| \hat{M}_2^n \right\|_t \to 0$ as $n \to \infty$. We now consider for $i \geq 3$,

$$\left\| \hat{M}_i^n \right\|_t = \sup_{0 \leq u \leq t} \frac{1}{\sqrt{n}} \left| D_i \left( \int_0^t (\hat{Q}_i^n(s) - \hat{Q}_{i+1}^n(s)) \, ds \right) - \int_0^t (\hat{Q}_i^n(s) - \hat{Q}_{i+1}^n(s)) \, ds \right|$$

$$= \sup_{0 \leq u \leq \Phi_{D_{i,n}}(t)} |M_{D_{i,n}}(u)|.$$

Observe that for $i \geq 3$, because no queues of length 3 or longer are ever created,

$$\Phi_{D_{i,n}}(t) \leq \frac{1}{n} \int_0^t \hat{Q}_i^n(s) \, ds$$

$$\leq \frac{1}{n} \int_0^t Q_3^n(0) \, ds \leq \frac{tQ_3^n(0)}{n}.$$

Therefore for $i \geq 3$

$$\left\| \hat{M}_i^n \right\|_t \leq \sup_{0 \leq u \leq tQ_3^n(0)/n} |M_{D_{i,n}}(u)|.$$
Define
\[ \gamma_n = \mathbb{E} \sup_{0 \leq u \leq tQ_3^n(0)/n} |M_{D_{i,n}}(u)|, \]
so we have \( \mathbb{E} \left| \dot{M}_i^n \right|_t \leq \gamma_n \) for all \( i \geq 3 \). Observe that assumption (2.8) implies \( tQ_3^n(0)/n \to 0 \) as \( n \to \infty \), and thus \( \gamma_n \to 0 \). Define \( M = (W_1, W_2, 0, 0, \ldots) \). Then we have
\[
\mathbb{E} \left| \dot{M}^n - M \right|_{\rho,t} = \sum_{i \geq 0} \rho^{-i} \mathbb{E} \left| \dot{M}_i^n - M_i \right|_t \leq \mathbb{E} \left| \dot{M}_0^n - W_1 \right|_t + \rho^{-1} \mathbb{E} \left| \dot{M}_1^n - W_2 \right|_t + \rho^{-2} \mathbb{E} \left| \dot{M}_2^n \right|_t + \sum_{i \geq 3} \rho^{-i} \gamma_n \leq \mathbb{E} \left| \dot{M}_0^n - W_1 \right|_t + \rho^{-1} \mathbb{E} \left| \dot{M}_1^n - W_2 \right|_t + \rho^{-2} \mathbb{E} \left| \dot{M}_2^n \right|_t + \gamma_n \frac{\rho^2}{\rho - 1}.
\]
Each of these terms go to zero as \( n \to \infty \), so we have \( \dot{M}^n \to M \). Further observe that \( \mathbb{E} |W|_t < \infty \) for Brownian motion \( W \), so for sufficiently large \( n \) we have \( \mathbb{E} \left| \dot{M}^n \right|_t < \infty \) for \( i = 0, 1 \), and thus \( \mathbb{E} \left| \dot{M}^n \right|_{\rho,t} < \infty \) for sufficiently large \( n \).

### 2.4.1 Fluid limit.

We prove Lemma 2.8 by showing that \( \dot{X}_i^n \) is stochastically bounded in \( D \) for \( i \geq 1 \). We define a sequence of processes \( Z^n \in D \) to be stochastically bounded if the sequence \( \{||Z^n||_t, n \geq 1\} \) is tight for every \( t > 0 \). For a more complete discussion of stochastic boundedness as we use it here, see §5 of [28].

The stochastic boundedness of \( \dot{X}_1^n \) and \( \dot{X}_2^n \) will follow from the stochastic boundedness of \( \dot{M}_0^n, \dot{M}_1^n, \dot{M}_2^n, \) and \( \dot{X}_3^n \). To see this, we prove

**Lemma 2.9.** Given \( (B_n, X_i^n(0), Y_i^n) \) a random element of \( \mathbb{R}_+ \times \mathbb{R} \times D \) for each \( n \geq 1 \)
and \( i = 1, 2 \), recall that Lemma 2.1 implies that the system

\[
\hat{X}_1^n(t) = X_1^n(0) + Y_1^n(t) + \int_0^t (-\hat{X}_1^n(s) + \hat{X}_2^n(s))ds - \hat{V}_1^n(t),
\]

\[
\hat{X}_2^n(t) = X_2^n(0) + Y_2^n(t) + \int_0^t (-\hat{X}_2^n(s) + \hat{V}_1^n(t) - \hat{V}_2^n(t)) \geq 0,
\]

\[
0 = \int_0^\infty \mathbb{1}\{\hat{X}_1^n(t) < 0\}d\hat{V}_1^n(t),
\]

\[
0 = \int_0^\infty \mathbb{1}\{\hat{X}_2^n(t) < B^n\}d\hat{V}_2^n(t),
\]

has a unique solution \((\hat{X}^n, \hat{V}^n)\). If the sequences \((X^n(0), n \geq 1)\) and \((Y^n_i, n \geq 1)\) are stochastically bounded for \( i = 1, 2 \), then the sequence \((\hat{X}^n, n \geq 1)\) is stochastically bounded in \( D \).

Note that we do not require boundedness for \( B_n \).

**Proof.** We fix \( t > 0 \). We will establish the bound

\[
\left\| \hat{X}^n \right\|_t \leq 8e^{6t} (\|X^n(0)\| + \|Y^n\|_t), \tag{2.53}
\]

from which the result follows.

To show (2.53), we will prove a similar bound for the unreflected process \( W^n \) defined by Lemma 2.3. Then (2.53) will follow from the Lipschitz continuity of the reflection maps \( \phi_0 \) and \( \phi_{B_n} \).

Just as in Lemma 2.1 and Lemma 2.3 we write \( \hat{X}_1^n(t) = \phi_0(W_1^n(t)) \) and \( \hat{X}_2^n(t) = \phi_{B^n}(W_2^n(t)) \) where \( W_1^n(t) \) and \( W_2^n(t) \) satisfy

\[
W_1^n(t) = X_1^n(0) + Y_1^n(t) + \int_0^t (-\phi_0(W_1^n(s)) + \phi_{B^n}(W_2^n(s)))ds, \tag{2.54}
\]

\[
W_2^n(t) = X_2^n(0) + Y_2^n(t) + \int_0^t (-\phi_{B^n}(W_2^n(s)))ds + \psi_0 (W_1^n(t)). \tag{2.55}
\]

We now use Gronwall’s inequality as stated in Lemma 2.4. Using the Lipschitz
property for $\phi_0, \phi_{B_n}$, and $\psi_0$ we have for $t \geq 0$

$$\|W^n_1\|_t \leq |X^n_1(0)| + \|Y^n_1\|_t + 2 \int_0^t (\|W^n_2\|_s + \|W^n_1\|_s) \, ds,$$

$$\|W^n_2\|_t \leq |X^n_2(0)| + \|Y^n_2\|_t + \|\psi_0(W^n_1)\|_t + \int_0^t \|W^n_2\|_s \, ds.$$

Now we note that we have

$$\|\psi_0(W^n_1)\|_t \leq \|W^n_1\|_t.$$

We define

$$u_1(t) = \|W^n_1\|_t \quad \text{and} \quad u_2(t) = (\|W^n_2\|_t - \|W^n_1\|_t)^+.\]

Finally we note

$$\|W^n_2\|_t \leq u_2(t) + u_1(t), \quad (2.56)$$

so we can write the inequalities

$$u_1(t) \leq |X^n_1(0)| + \|Y^n_1\|_t + 4 \int_0^t u_1(s) \, ds + 2 \int_0^t u_2(s) \, ds,$$

$$u_2(t) \leq |X^n_2(0)| + \|Y^n_2\|_t + \int_0^t u_1(s) \, ds + \int_0^t u_2(s) \, ds.$$

Let $|X^n(0)| + \|Y^n\|_t = K$. Then Lemma [2.4] implies

$$u_1(t) \leq 2Ke^{6t} \quad \text{and} \quad u_2(t) \leq 2Ke^{6t}.\]$$

From (2.56) and the definitions of $u_1$ and $u_2$ we obtain

$$\|W^n_1\|_t \leq 2Ke^{6t} \quad \text{and} \quad \|W^n_2\|_t \leq 4Ke^{6t}.$$

Since $\phi_0$ and $\phi_{B_n}$ are Lipschitz continuous with constant 2 this implies

$$\|\dot{X}^n_1\|_t \leq 4Ke^{6t} \quad \text{and} \quad \|\dot{X}^n_2\|_t \leq 8Ke^{6t},$$

53
which proves \( (2.53) \).

Note that this proof also provides the boundedness of \( w \) that we use in the proof of continuity of \( w \) in Lemma \( 2.3 \) and that it does not use any of the continuity properties proved using that boundedness.

In our application of Lemma \( 2.9 \) we will have \( Y_1^n = \hat{M}_0^n - \hat{M}_1^n - (1 - \lambda_n)\sqrt{nt} \) and \( Y_2^n(t) = \hat{M}_2^n(t) + \int_0^t \hat{X}_3^n(s)ds \), so it remains to prove that \( \int_0^t \hat{X}_3^n(s)ds \) and each martingales \( \hat{M}_i^n \) is stochastically bounded.

To show that \( \int_0^t \hat{X}_3^n(s)ds \) is stochastically bounded we need only show that \( \hat{X}_3^n \) is stochastically bounded. This follows from the fact that \( Q_3^n(0) \) is stochastically bounded by assumption \( (2.8) \) and

\[
\hat{X}_3^n(t) = \frac{\hat{Q}_3^n(t)}{\sqrt{n}} \leq \frac{Q_3^n(0)}{\sqrt{n}}
\]

because no queues of length 3 or longer are ever created. An identical argument proves stochastic boundedness of \( \hat{X}_i^n \) for \( i \geq 4 \).

To prove the stochastic boundedness of these martingales we will use the following lemma from \( [28] \):

**Lemma 2.10** \( ([28] \text{ Lemma 5.8}) \). Suppose that, for each \( n \geq 1 \), \( M_n \) is a square integrable martingale with predictable quadratic variation \( \langle M_n \rangle \). If the sequence of random variables \( \langle M_n \rangle(T) \) is stochastically bounded in \( \mathbb{R} \) for each \( T > 0 \), then the sequence of stochastic processes \( M_n \) is stochastically bounded in \( D \).

We now prove that the predictable quadratic variations of \( \hat{M}_i^n \) are stochastically bounded. In the case of \( \hat{M}_0^n \) this is immediate since by \( (2.39) \) the quadratic variation is deterministic.
For \( \hat{M}_1^n \) we refer to (2.40) and apply crude bounds to see
\[
\left\langle \hat{M}_1^n \right\rangle (t) = \frac{1}{n} \int_0^t (\hat{Q}_1^n(s) - \hat{Q}_2^n(s)) ds \\
\leq \frac{1}{n} \int_0^t \hat{Q}_1^n(s) ds \\
\leq \frac{t}{n} \left( Q_1^n(0) + A(\lambda_n nt) \right).
\]

It suffices to show stochastic boundedness of each term in the sum. For \( Q_1^n(0) \) this follows from assumption (2.8).

For \( A(\lambda_n nt) \) we note \( \lambda_n \to 1 \) so by the strong law of large numbers (SLLN) for Poisson processes we have
\[
\frac{A(\lambda_n nt)}{n} \to e(t)
\]
with probability 1, which implies stochastic boundedness, so we conclude that \( \hat{M}_1^n \) is stochastically bounded.

For \( \hat{M}_2^n \) we have
\[
\left\langle \hat{M}_2^n \right\rangle (t) \leq \frac{t}{n} \left( Q_2^n(0) + \hat{U}_1^n(t) \right) \\
\leq \frac{t}{n} \left( Q_2^n(0) + A(\lambda_n nt) \right),
\]
and stochastic boundedness follows.

We now return to the proof of Lemma (2.8)

Proof of Lemma (2.8) We have for \( i \geq 2 \) that
\[
\hat{X}_1^n = \frac{\hat{Q}_1^n}{\sqrt{n}} \quad \text{and} \quad \hat{X}_i^n = \frac{\hat{Q}_i^n}{\sqrt{n}}
\]
are stochastically bounded. Therefore, by, e.g., Lemma 5.9 of [28],
\[
\frac{\hat{X}_i^n}{\sqrt{n}} \Rightarrow 0 \quad \text{in } D \quad \text{as } n \to \infty.
\]
From the definition of $\hat{X}^n$ this is equivalent to

$$\Psi_1^n = \frac{Q_1^n}{n} \Rightarrow \omega \quad \text{and} \quad \Psi_i^n = \frac{Q_i^n}{n} \Rightarrow 0 \quad \text{in} \ D \quad \text{as} \ n \to \infty$$

for $i \geq 2$. \hfill \Box

### 2.4.2 Proof of Theorem 2.2.

Now that we have the convergence of the martingale processes $\hat{M}_i^n$, we can apply the CMT to prove Theorem 2.2.

**Proof of Theorem 2.2.** We first show $\hat{X}^n \Rightarrow X$ before considering $X^n$.

In Theorem 2.1 in the pre-limit regime we set $B_n = \sqrt{n}$, $b_i = X_i^n(0)$ for $i \geq 1$,

$$y_1(t) = \hat{M}_0^n(t) - \hat{M}_1^n(t) - (1 - \lambda_n)\sqrt{nt},$$

and $y_i(t) = -\hat{M}_i^n(t)$ for $i \geq 2$. Equations (2.44) and (2.45) show that $\hat{V}_1^n$ and $\hat{V}_2^n$ are appropriately acting as $u_1$ and $u_2$ in the integral representation, so $x_i(t) = \hat{X}^n_i(t)$ for $i \geq 1$. For application of the CMT we need only determine the limits of $B$, $b$ and $y$.

We have $B_n \to \infty$, so in the limit $u_2 = 0$.

By assumption we have for $i \geq 1$

$$\hat{X}^n_i(0) \Rightarrow X_i(0),$$

so in the limiting system we let $b_i = X_i(0)$. Next we have by (2.44) and (2.46)

$$\hat{M}_0^n(t) - \hat{M}_1^n(t) - (1 - \lambda_n)\sqrt{nt} \Rightarrow W_1(t) - W_2(t) - \beta t$$

$$\overset{d}{=} \sqrt{2}W(t) - \beta t,$$

where $W$ is a standard Brownian motion and $\overset{d}{=}$ indicates equivalence in distribution.

Another application of (2.46) implies

$$-\hat{M}_i^n \Rightarrow 0$$

56
for \( i \geq 2 \) so in the limiting system we have \( y_1(t) = \sqrt{2} W(t) - \beta t \) and \( y_i(t) = 0 \), for \( i \geq 2 \).

The CMT then implies that \( \hat{X}^n \Rightarrow X \) in \( D^{\infty, \rho} \) as \( n \to \infty \).

Now we consider the untruncated system described by \( X^n \). We have

\[
Q_1^n(t) = Q_1^n(0) + A(\lambda_n t) - D_1 \left( \int_0^t (Q_1^n(s) - Q_2^n(s)) \, ds \right) - U_1^n(t),
\]

(2.57)

\[
Q_i^n(t) = Q_i^n(0) + U_{i-1}^n(t) - D_i \left( \int_0^t (Q_i^n(s) - Q_{i+1}^n(s)) \, ds \right) - U_i^n(t), \quad i \geq 2,
\]

(2.58)

where \( U_i^n(t) \) is the number of arrivals in \([0, t]\) when every server has at least \( i \) customers.

Now define

\[
t_n^* = \inf\{ t \geq 0 : Q_2^n(t) = n \}
\]

(2.59)

and note that for \( t \in [0, t_n^*] \) we have \( U_i^n(t) = 0 \) for \( i \geq 2 \). This, implies that for such \( t \), the system (2.57)-(2.58) becomes

\[
Q_1^n(t) = Q_1^n(0) + A(\lambda_n t) - D_1 \left( \int_0^t (Q_1^n(s) - Q_2^n(s)) \, ds \right) - U_1^n(t),
\]

\[
Q_2^n(t) = Q_2^n(0) + U_1^n(t) - D_2 \left( \int_0^t (Q_2^n(s) - Q_3^n(s)) \, ds \right) - U_2^n(t),
\]

\[
Q_i^n(t) = Q_i^n(0) - D_i \left( \int_0^t (Q_i^n(s) - Q_{i+1}^n(s)) \, ds \right), \quad i \geq 3,
\]

which precisely matches (2.32)-(2.34). Thus for \( t \in [0, t_n^*] \), \( X^n(t) \) and \( \hat{X}^n(t) \) are identical.

It only remains to show for all \( t \geq 0 \) that \( P(t_n^* \leq t) \to 0 \) as \( n \to \infty \). Because the systems are identical up to time \( t_n^* \), we can replace \( Q_2^n(t) \) in (2.59) with \( \hat{Q}_2^n(t) \) to see

\[
P(t_n^* \leq t) = P \left( \sup_{0 \leq s \leq t} \hat{Q}_2^n(s) \geq n \right) = P \left( \sup_{0 \leq s \leq t} \hat{X}_2^n(s) \geq \sqrt{n} \right)
\]

\[
\leq P \left( \sup_{0 \leq s \leq t} \hat{X}_2^n(s) \geq C \right)
\]

for constant \( 0 < C \leq \sqrt{n} \). By the weak convergence \( \hat{X}_2^n \Rightarrow X_2 \) and the fact that
\[ \{ \sup_{0 \leq s \leq t} \hat{X}_2^n(s) \geq C \} \text{ is closed, we have} \]

\[
\limsup_n \mathbb{P} \left( \sup_{0 \leq s \leq t} \hat{X}_2^n(s) \geq C \right) \leq \mathbb{P} \left( \sup_{0 \leq s \leq t} X_2(s) \geq C \right)
\]

By continuity of probability we have

\[
\lim_{C \to \infty} \mathbb{P} \left( \sup_{0 \leq s \leq t} X_2(s) \geq C \right) = \mathbb{P} \left( \sup_{0 \leq s \leq t} X_2(s) = \infty \right) = 0.
\]

We therefore have

\[
\limsup_n \mathbb{P}(t_n^* \leq t) \leq \lim_{C \to \infty} \limsup_n \mathbb{P} \left( \sup_{0 \leq s \leq t} \hat{X}_2^n(s) \geq C \right) \leq \lim_{C \to \infty} \mathbb{P} \left( \sup_{0 \leq s \leq t} X_2(s) \geq C \right) = 0
\]

and thus \( \mathbb{P}(t_n^* \leq t) \to 0 \) as \( n \to \infty \). We conclude \( X^n \Rightarrow X \).

\[ \square \]

2.5 Waiting time.

An important performance measure of a queueing system is the expected time that customers will have to wait before entering service. In the \( M/M/n \) system with a single queue in the Halfin-Whitt regime, the expected waiting time is of the order \( O(1/\sqrt{n}) \). We will now show that the \( M/M/n \)-JSQ system has the same order of aggregate waiting time in the transient regime, and thus seems to have a minimal loss of efficiency as measured by waiting time.

Notice that our representation of the system allows us to directly consider the total time any customers in the system will wait. In particular, the instantaneous number of customers waiting to be served at a given time \( t \) is precisely \( \sum_{i \geq 2} Q_i^n(t) \). This quantity can be integrated over time to compute the aggregate waiting time in the system. With this insight, we prove the following:

**Theorem 2.3.** Let \( Z_{i}^{n} \) denote the aggregate waiting time of customers who arrived
over the time period \([0, t]\). If \(\sum_{i \geq 2} X_i(0) < \infty\), then

\[
\lim_{C \to \infty} \limsup_{n} \mathbb{P}(Z^n_t \geq C \sqrt{n}) = 0. \tag{2.60}
\]

Since the total number of arrivals to the system in this time is \(\Theta(n)\), then with high probability, the waiting time per arrival is \(O(1/\sqrt{n})\).

Note that the condition on \(X(0)\) simply insures that the waiting time incurred by customers initially present in the system is not infinite.

**Proof.** As noted above, the aggregate waiting time over the period \([0, t]\) is

\[
Z^n_t = \int_0^t \sum_{i \geq 2} Q^n_i(s) ds.
\]

We consider a scaled version \(Y^n_t = Z^n_t / \sqrt{n}\). Note that \(Y^n_t \leq t \sup_{0 \leq s \leq t} \sum_{i \geq 2} X^n_i(s)\), and thus

\[
\mathbb{P}(Y^n_t \geq C) \leq \mathbb{P}\left(t \sup_{0 \leq s \leq t} \sum_{i \geq 2} X^n_i(s) \geq C\right)
\]

for any constant \(C > 0\). By the weak convergence \(X^n_i \Rightarrow X_i\) and the fact that \(\{t \sup_{0 \leq s \leq t} \sum_{i \geq 2} X^n_i(s) \geq C\}\) is closed, we have

\[
\limsup_n \mathbb{P}\left(t \sup_{0 \leq s \leq t} \sum_{i \geq 2} X^n_i(s) \geq C\right) \leq \mathbb{P}\left(t \sup_{0 \leq s \leq t} \sum_{i \geq 2} X_i(s) \geq C\right)
\]

By continuity of probability and \(\sum_{i \geq 2} X_i(0) < \infty\) we have

\[
\lim_{C \to \infty} \mathbb{P}\left(t \sup_{0 \leq s \leq t} \sum_{i \geq 2} X_i(s) \geq C\right) = 0.
\]

Therefore we conclude

\[
\lim_{C \to \infty} \limsup_n \mathbb{P}(Y^n_t \geq C) \leq \lim_{C \to \infty} \mathbb{P}\left(t \sup_{0 \leq s \leq t} \sum_{i \geq 2} X_i(s) \geq C\right) = 0,
\]

59
which proves (2.60). We note that this implies $Y_t^n = O(1)$ with high probability, and thus $Z_t^n = O(\sqrt{n})$ with high probability.

Because customers arrive according to a Poisson process with rate $\lambda_n = \Theta(n)$, this implies the waiting time per arrival is $O(\sqrt{n}/n) = O(1/\sqrt{n})$ with high probability, completing the proof.

Theorem 2.3 does not directly tell us anything about the distribution of the waiting time. We can see from considering the system that this waiting time is distributed in a qualitatively different way than the standard $M/M/n$ system. Customers immediately enter service if there are any idle servers and otherwise wait a constant order amount of time for the previous customer in their queue to finish service. Because the aggregate waiting time is of the order $O(\sqrt{n})$ and any arriving customers who wait at all incur a constant order waiting time, the total number of customers who have to wait is also $O(\sqrt{n})$. As noted above, the total number of arriving customers is order $n$, so the fraction of customers who have to wait is of the order $O(1/\sqrt{n})$. 

60
Chapter 3

The Supermarket Model

In this chapter we consider the so-called supermarket model, which has been extensively studied as an example of the “power of two choices” phenomenon. Concretely, in the supermarket model arriving customers join the shortest of $d \geq 2$ randomly selected queues. It has been shown [26, 32] that when service is exponentially distributed with rate 1 and arrivals are at rate $\lambda n$ for constant $\lambda < 1$, the expected time a customer spends in the system is improved exponentially by selecting among $d$ choices rather than only one choice. In comparison to the Join the Shortest Queue policy studied in Chapter [2] however, the supermarket model is less efficient because some customers will be routed to queues longer than those not in the set of $d$ randomly selected. Nevertheless, the supermarket model is appealing because it does not depend on global information about the length of every queue for each routing decision, and thus involves significantly less computational overhead.

We consider the supermarket model in which the arrival rate varies with $n$, specifically setting $\lambda_n \uparrow 1$ while holding $d$ constant as $n$ diverges to infinity. We first provide a simple argument that in steady state the expected fraction of queues with at least $i$ customers is at least $1 - (1 - \lambda_n)\frac{d-1}{d-1}$. This implies that if $\omega_n$ is an arbitrary sequence diverging to infinity as $n \rightarrow \infty$ then the steady state fraction of queues with length at least $i^*_n = \log_d \frac{1}{1-\lambda_n} - \omega_n$ goes to 1 as $n \rightarrow \infty$. Our main result primarily concerns queues shorter than this $i^*_n$. We show that so long as $(1 - \lambda_n)\eta_n$ converges to a constant for some $\eta_n = O(\sqrt{n})$, an appropriately scaled process counting the fraction of
queues with length at least \( i \) converges weakly to a deterministic infinite system of integral equations, under reasonable assumptions on the starting state. The proof of these results will rely in part on a stopping time designed to prevent the finite system from “growing too much”.

3.1 The model and the main result.

We consider the supermarket model with \( n \) exponential rate one servers each with their own queue, and Poisson arrivals with rate \( \lambda_n n \) for \( \lambda_n < 1 \) such that \( \lambda_n \uparrow 1 \). Specifically, we assume that there exists some sequence \( \eta_n \) and constant \( \beta > 0 \) such that \( \eta_n \to \infty \) as \( n \to \infty \) and

\[
\lim_{n \to \infty} \eta_n (1 - \lambda_n) = \beta. \tag{3.1}
\]

We assume \( \eta_n \geq 1 \) for all \( n \geq 1 \). Upon arrival customers select \( d \geq 2 \) queues uniformly at random with replacement and join the shortest of these queues, with ties broken uniformly at random.

Let \( 0 \leq S_i^n(t) \leq 1 \) be the fraction of queues with at least \( i \) customers (including the customer in service) at time \( t \). Then the probability that an arriving customer at time \( t \) joins a queue of length exactly \( i - 1 \) is

\[
(S_{i-1}^n(t))^d - (S_i^n(t))^d.
\]

As a result, because the overall arrival rate is \( \lambda_n n \), the instantaneous rate of arrivals to queues of length exactly \( i - 1 \) is

\[
\lambda_n n \left( (S_{i-1}^n(t))^d - (S_i^n(t))^d \right) .
\]

Note that an arrival to a queue of length \( i - 1 \) increases \( S_i^n \) by \( 1/n \), and all other types of arrivals leave \( S_i^n \) unchanged. Similarly, a departure from a queue of length \( i \) decreases \( S_i^n \) by \( 1/n \) and any other departures leave \( S_i^n \) unchanged. The instantaneous
rate of departures from queues of length exactly $i$ at time $t$ is

$$nS_i^n(t) - nS_{i+1}^n(t).$$

For $i \geq 1$ let $A_i$ and $D_i$ be independent rate-1 Poisson processes. Then we can represent the processes $S_i^n$ via random time changes of these Poisson processes. Specifically, we have

$$S_0^n(t) = 1,$$

$$S_i^n(t) = S_i^n(0) + \frac{1}{n} A_i \left( \lambda_n n \int_0^t \left( (S_{i-1}^n(s))^d - (S_i^n(s))^d \right) ds \right)$$

$$- \frac{1}{n} D_i \left( n \int_0^t (S_i^n(s) - S_{i+1}^n(s)) ds \right), \quad i \geq 1. \quad (3.3)$$

We let

$$T_i^n = \eta_n (1 - S_i^n),$$

where $\eta_n$ is defined by (3.1). Observe that

$$0 \leq T_i^n \leq \eta_n.$$ 

For technical reasons we restrict our choice of $\eta_n$ to those for which there exists constant $Q \geq 0$ such that

$$\eta_n \leq Q \sqrt{n}, \quad (3.4)$$

for all $n \geq 1$. That is, we assume $\eta_n = O(\sqrt{n})$. We further define $\eta_\infty = \infty$. We will prove in Theorem 3.2 that under appropriate conditions, $T^n = (T_0^n, T_1^n, \ldots)$ weakly converges to the solution to a certain integral equation, which we first prove in Theorem 3.1 has a unique solution.
For every $\eta \in \mathbb{R}_{\geq 1}$ and $x \in \mathbb{R}$, we let
\[
g^n(x) \triangleq \frac{\eta}{d} \left( 1 - \frac{x}{\eta} \right)^d - \frac{\eta}{d} + x
\]
\[
= \frac{1}{d} \sum_{l=2}^{d} \binom{d}{l} (-1)^l \frac{x^l}{\eta^{l-1}},
\]
and also let $g^\infty = 0$ for $\eta = \infty$.

Given $b \in \mathbb{R}^{\infty,\rho}$, $y \in D^{\infty,\rho}$, $\lambda \in \mathbb{R}$, and $\eta \in \mathbb{R}_{\geq 1}$, consider the following system of integral equations for $t \geq 0$:
\[
T_0(t) = 0,
\]
\[
T_i(t) = b_i + y_i - \lambda d \int_0^t (T_i(s) - T_{i-1}(s) - g^n(T_i(s)) + g^n(T_{i-1}(s))) ds + \int_0^t (T_{i+1}(s) - T_i(s)) ds,
\]
\[
i \geq 1.
\]
An important special case of this system, which will appear as the limiting system in Theorem 3.2 below, is when we set $\eta = \infty$ and $\lambda = 1$, and $y = 0$. In this case $b_i = T_i(0)$ for $i \geq 1$. This system is as follows:
\[
T_0(t) = 0,
\]
\[
T_i(t) = T_i(0) - d \int_0^t (T_i(s) - T_{i-1}(s)) ds + \int_0^t (T_{i+1}(s) - T_i(s)) ds,
\]
\[
i \geq 1.
\]
For any $\eta \in \mathbb{R}_{\geq 1}$, $0 < \alpha < 1/2$, and $\rho > 1$, let $i^* = \frac{\alpha}{2} \log_\lambda \eta$ where $\log_\lambda \infty = \infty$. In fact, for our purposes, $\alpha/2$ can be replaced by any positive number strictly smaller
than $\alpha$. Define a stopping time $t^\ast \in \mathbb{R}_+$ as follows:

\[
t^\ast \triangleq \inf \left\{ s \geq 0 : \exists \ i \text{ s.t. } 1 \leq i \leq i^\ast \text{ and } |T_i(s)| \geq \eta^\alpha \right\}
\]

or

\[
\exists \ i \text{ s.t. } i > i^\ast \text{ and } |T_i(s)| \geq \eta + 1
\]

We also define a subset of $\mathbb{R}^\infty$ related to this stopping time. Let

\[
Z^\eta \triangleq \left\{ x \in \mathbb{R}^\infty \text{ s.t. } \forall 1 \leq i \leq i^\ast, \ |x_i| \leq \eta^\alpha \text{ and } \forall i > i^\ast, \ |x_i| \leq \eta + 1 \right\}
\]

Finally, we define a subset of the product space $\mathbb{R}^{\infty,\rho} \times D^{\infty,\rho} \times \mathbb{R} \times \mathbb{R}_{\geq 1}$ equipped with the product topology which will allow us to limit our attention to certain parameter values. Specifically, let

\[
Z^\alpha_K \triangleq \left\{ (b, y, \lambda, \eta) \in \mathbb{R}^{\infty,\rho} \times D^{\infty,\rho} \times \mathbb{R} \times \mathbb{R}_{\geq 1} \text{ s.t. } b + y(0) \in Z^\eta, \ \|b\|_\rho \leq K, \|y\|_{\rho,t} \leq K \ \forall t \geq 0 \right\}
\]

Observe that for $(b, y, \lambda, \eta)$ which are not in $Z^\alpha_K$ for any $K > 0$, we have $t^\ast = 0$, and that $Z^\alpha_K$ is a closed subset of $\mathbb{R}^{\infty,\rho} \times D^{\infty,\rho} \times \mathbb{R} \times \mathbb{R}_{\geq 1}$. Note that here we equip $\mathbb{R}^{\infty,\rho}$ and $D^{\infty,\rho}$ with the product topology.

Our first result shows that (3.6)-(3.7) has a unique solution on the interval $[0, t^\ast]$ and that it defines a map which satisfies a certain continuity property.

**Theorem 3.1.** For $t \in [0, t^\ast]$, the system (3.6)-(3.7) has a unique solution $T =$
\[(T_i, i \geq 0) \in D^{\infty, \rho}. \text{ For any } t \geq 0, \text{ defining}
\]
\[\hat{T}(t) = \begin{cases} T(t) & t < t^* \\ T(t^*) & t \geq t^* \end{cases}, \quad (3.11)\]
we obtain a function \( f : \mathbb{R}^{\infty, \rho} \times D^{\infty, \rho} \times \mathbb{R} \times [0, \infty] \to D^{\infty, \rho} \) mapping \((b, y, \lambda, \eta)\) to \( \hat{T} = f(b, y, \lambda, \eta) \). Moreover, when the domain is restricted to \( Z_K^\alpha \) for any \( K > 0 \) equipped with the product topology, \( f \) is continuous.

\textit{Remark 3.1.} Note that if \( \eta = \infty \) then \( t^* = \infty \) and thus \( T = \hat{T} \). Further note that for \( \eta < \infty \), the definition of \( t^* \) implies that in fact either \( \hat{T} \in \hat{D}^{\eta+1, \rho} \) or \( t^* = 0 \). In the latter case \( \hat{T} = b + y(0) \) is a constant function.

We prove Theorem \[3.1\] in Section \[3.3\]. We now turn to our main result.

\textbf{Theorem 3.2.} Suppose \( \lambda_n \) satisfies (3.1) for a sequence \( \eta_n \) satisfying (3.4) for some \( Q > 0 \). Suppose there exists \( \rho > 1 \) such that
\[T^n(0) \Rightarrow T(0) \quad \text{in } \mathbb{R}^{\infty, \rho} \text{ as } n \to \infty, \quad (3.12)\]
for some random variable \( T(0) \in \mathbb{R}^{\infty, \rho} \). Furthermore, suppose
\[\limsup_{n \to \infty} \mathbb{E} \left[ ||T^n(0)||_\rho \right] < \infty, \quad (3.13)\]
and there exists \( 0 < \alpha < 1/2 \) such that for all sufficiently large \( n \), almost surely
\[T^n_i(0) \leq \eta^n_\alpha, \quad 0 \leq i \leq i^*. \quad (3.14)\]
Then
\[T^n \Rightarrow T \quad \text{in } D^{\infty, \rho} \text{ as } n \to \infty, \quad \text{where } T \text{ is the unique solution of the system (3.8)- (3.9).} \]

The motivation for the initial condition assumptions (3.13) and (3.14) is as follows: as we will see below, we expect that in steady state the limiting system \( T \) grows like
$T_i = d^i$, so condition (3.13) can be considered as requiring $T^n(0)$ to be consistent with this behavior, with $\rho > d$. Condition (3.14) is similar, as $d^* \approx \eta_n^{\alpha/2}$. We prove Theorem 3.2 in Section 3.5.

Now we turn to results about the solution $T$ which appears in Theorem 3.2.

**Theorem 3.3.** Consider the system of integral equations given by (3.8)-(3.9). This system has a fixed point $\pi = (\pi_i, i \geq 0)$ given by

$$\pi_i = \pi_1 \frac{d^i - 1}{d - 1}.$$  \hspace{1cm} (3.15)

This fixed point is unique up to the constant $\pi_1$.

We also show that this fixed point is attractive:

**Theorem 3.4.** Let $\Phi(t) = \sum_{i \geq 1} d^{-i/2} |T_i(t) - \pi_i|$. If $\Phi(0) < \infty$ then $\Phi$ converges exponentially fast to zero. Specifically, $\Phi(t) \leq \Phi(0)e^{-\left(\sqrt{d-1}\right)^2 t}$ for all $t \geq 0$.

Finally we consider the system in steady state. Via elementary arguments we prove a bound on the expectation of the fraction of short queues in steady state. For $i \geq 0$, let $S_i^n(\infty)$ be the fraction of queues with length at least $i$ in steady state. For the statement below, we assume $\lambda_n$ is given by (3.1) and $\eta_n \to \infty$ is arbitrary. In particular, the assumption (3.4) is no longer needed.

**Theorem 3.5.** For $i \geq 0$ we have

$$\mathbb{E}S_i^n(\infty) \geq 1 - (1 - \lambda_n) \frac{d^i - 1}{d - 1}.$$  \hspace{1cm} (3.16)

As a result, for any sequence $\omega_n$ which diverges to infinity as $n \to \infty$, the fraction of queues with length at least $\log_d \eta_n - \omega_n$ approaches unity with probability approaching one as $n \to \infty$. This further implies that a customer arriving in steady state experiences a delay of at least $\log_d \eta_n - \omega_n$ with probability approaching one as $n \to \infty$.

We prove Theorems 3.3, 3.5 in Section 3.2.
Note that for any sequence \( \omega_n \) which diverges to infinity as \( n \to \infty \), the right hand side of (3.16) diverges to negative infinity for any \( i \geq \log_d \eta_n + \omega_n \). Because (3.16) is only a lower bound this is not useful, but it does suggest that the behavior of \( S_n^i \) is best examined for values of \( i \) near \( \log_d \eta_n \).

### 3.2 The model in steady state.

**Proof of Theorem 3.3.** We set the derivative of \( T_i \) to zero and introduce the notation \( \pi = (\pi_0, \pi_1, \pi_2, \ldots) \) for the desired fixed point. This leads to the recurrence

\[
\pi_{i+1} = (d + 1)\pi_i - d\pi_{i-1}, \quad i \geq 1,
\]

which is solved by

\[
\pi_i = \frac{1}{d - 1} \left( d\pi_0 - \pi_1 + d^i(\pi_1 - \pi_0) \right), \quad i \geq 0.
\]

By (3.8) we have \( T_0(t) = 0 \) for all \( t \geq 0 \) and thus \( \pi_0 = 0 \), so this fixed point reduces to

\[
\pi_i = \pi_1 \frac{d^i - 1}{d - 1}, \quad i \geq 0.
\]

Next we prove Theorem 3.4.

**Proof of Theorem 3.4.** Define \( \epsilon_i(t) = T_i(t) - \pi_i \). We have

\[
\frac{d\epsilon_i}{dt} = d(\epsilon_{i-1} + \pi_{i-1}) - (d + 1)(\epsilon_i + \pi_i) + (\epsilon_{i+1} + \pi_{i+1})
\]

\[
= d\epsilon_{i-1} - (d + 1)\epsilon_i + \epsilon_{i+1}.
\]

Temporarily assume \( \epsilon_i \neq 0 \) for all \( i \) so the derivative \( \frac{d\epsilon_i}{dt} \) is well defined. After providing a basic argument under this assumption we will explain how to remove it.
Now we have

\[
\frac{d\Phi}{dt} = \sum_{i: \epsilon_i > 0} d^{-i/2} (d\epsilon_{i-1} - (d + 1)\epsilon_i + \epsilon_{i+1}) - \sum_{i: \epsilon_i < 0} d^{-i/2} (d\epsilon_{i-1} - (d + 1)\epsilon_i + \epsilon_{i+1}).
\]

Let us now consider the terms involving \(\epsilon_i\). We will first consider \(i \geq 2\). There are several cases, depending on the signs of \(\epsilon_{i-1}, \epsilon_i,\) and \(\epsilon_{i+1}\). First suppose they are all negative, so the term involving \(\epsilon_i\), which we denote \(A_i\), is

\[
A_i = -d^{-(i-1)/2}\epsilon_i + d^{-i/2}(d + 1)\epsilon_i - d^{-(i+1)/2}d\epsilon_i
\]

\[
= d^{-i/2}\epsilon_i \left(-\sqrt{d} + d + 1 - \frac{d}{\sqrt{d}}\right)
\]

\[
= \left(1 - 2\sqrt{d} + d\right) d^{-i/2}\epsilon_i.
\]

We define \(\delta = 1 - 2\sqrt{d} + d = \left(\sqrt{d} - 1\right)^2\) and note that the assumption \(d \geq 2\) implies \(\delta > 0\). Now note that if the sign of \(\epsilon_{i-1}\) or \(\epsilon_{i+1}\) or both is positive and \(\epsilon_i\) remains negative this will simply change the sign of the appropriate coefficient of \(\epsilon_i\) from negative to positive, decreasing \(A_i\). Thus for all cases with \(\epsilon_i\) negative we have

\[
A_i \leq \delta d^{-i/2}\epsilon_i.
\]

If all three signs are positive, we have

\[
A_i = d^{-(i-1)/2}\epsilon_i - d^{-i/2}(d + 1)\epsilon_i + d^{-(i+1)/2}d\epsilon_i
\]

\[
= d^{-i/2}\epsilon_i \left(\sqrt{d} - d - 1 + \frac{d}{\sqrt{d}}\right)
\]

\[
= -\delta d^{-i/2}\epsilon_i.
\]
and if $\epsilon_{i-1}$ or $\epsilon_{i+1}$ is negative we still have

$$A_i \leq -\delta d^{-i/2} \epsilon_i.$$  

Thus for all $i \geq 2$ we have

$$A_i \leq -\delta d^{-i/2} |\epsilon_i|.$$  

To see that this inequality also holds for $i = 1$ note that for that case we simply omit the first term of $A_i$. Using this result for all $i \geq 1$ we conclude

$$\frac{d\Phi}{dt} = \sum_{i \geq 1} A_i \leq -\delta d^{-i/2} |\epsilon_i| = -\delta \Phi.$$  

Thus we have

$$\Phi(t) \leq \Phi(0) e^{-\delta t},$$

and conclude that $\Phi(t)$ converges exponentially.

As in Mitzenmacher [26], because we are interested in the evolution of the system as time increases, we can account for the $\epsilon_i = 0$ case by considering upper right-hand derivatives of $\epsilon_i$, defining

$$\frac{d|\epsilon_i|}{dt} \bigg|_{t=t_0} \triangleq \lim_{t \to t_0^+} \frac{|\epsilon_i(t)|}{t-t_0},$$

and similarly for $\frac{d\Phi}{dt}$. Now if $\epsilon_i(t_0) = 0$ we have $\frac{d|\epsilon_i|}{dt} \bigg|_{t=t_0} \geq 0$, so we can include the $\epsilon_i = 0$ cases in the above proof with the $\epsilon_i > 0$ case now also including the case with $\epsilon_i = 0$ and $\frac{d\Phi}{dt} \geq 0$ and similarly for $\epsilon_i < 0$.  

Proof of Theorem 3.5. From [33] we obtain for $i \geq 1$

$$\mathbb{E}S_i^n(t) = \mathbb{E}S_i^n(0) + \lambda_n \int_0^t \mathbb{E} \left[ (S_{i-1}^n(s))^d - (S_i^n(s))^d \right] ds - \int_0^t (\mathbb{E}S_i^n(s) - \mathbb{E}S_{i+1}^n(s)) ds.$$  

Assuming $(S_i^n(0), i \geq 0)$ has a steady state distribution, the same applies to $(S_i^n(t), i \geq 0)$.
0), implying \( E S^n_i(0) = E S^n_i(t) \). Thus switching to \( S^n_i(\infty) \) for steady state version of \( S^n_i(t) \), we obtain

\[
0 = \lambda_n E \left[ (S^n_{i-1}(\infty))^d - (S^n_i(\infty))^d \right] - (E S^n_i(\infty) - E S^n_{i+1}(\infty)) .
\]

Because \( 0 \leq S^n_i(t) \leq 1 \) for all \( i \geq 0 \) and \( t \geq 0 \) and \( S^n_{i-1}(t) \geq S^n_i \) for all \( i \geq 1 \) and \( t \geq 0 \), we have the bound

\[
(S^n_{i-1}(\infty))^d - (S^n_i(\infty))^d \leq d (S^n_{i-1}(\infty) - S^n_i(\infty)) ,
\]

and thus have

\[
0 \leq \lambda_n d (E S^n_{i-1}(\infty) - E S^n_i(\infty)) - (E S^n_i(\infty) - E S^n_{i+1}(\infty)) .
\]

For \( i \geq 0 \) define

\[
\sigma_i \triangleq E S^n_i(\infty) - E S^n_{i-1}(\infty),
\]

and observe that \( \sigma_i \) is the expected number of queues of length exactly \( i \) in steady state. We now obtain the bound

\[
\sigma_i \leq \lambda_n d \sigma_{i-1}, \quad i \geq 1 ,
\]

which implies

\[
\sigma_i \leq \sigma_0 d^i (\lambda_n)^i .
\]

We have \( S^n_0(t) = 1 \) and use Little's law to observe \( S^n_i(\infty) = \lambda_n \) resulting in

\[
\sigma_i \leq (1 - \lambda_n) d^i \lambda_n^i \leq (1 - \lambda_n) d^i , \quad i \geq 1 .
\]
Now observe

$$\mathbb{E}S_1^n(\infty) = \mathbb{E}S_0^n(\infty) - \sum_{j=0}^{i-1} (\mathbb{E}S_j^n(\infty) - \mathbb{E}S_{j+1}^n(\infty))$$

$$= 1 - \sum_{j=0}^{i-1} \sigma_j$$

$$\geq 1 - (1 - \lambda_n) \sum_{j=0}^{i-1} d^j = 1 - (1 - \lambda_n) \frac{d^i - 1}{d - 1}.$$ 

This establishes (3.16).

Recalling (3.1), for \(i \leq \log_d \eta_n - \omega_n\), we have

$$\mathbb{E}S_i^n(\infty) \geq 1 - (1 - \lambda_n) \frac{\eta_n d^{-\omega_n} - 1}{d - 1} \to 1,$$

as \(n \to \infty\). Since \(S_i^n(\infty) \leq 1\), this implies that as \(n \to \infty\), \(S_i^n(\infty) \to 1\) in probability. Namely, the fraction of queues with length at least \(\log_d \eta_n - \omega_n\) approaches one in probability.

Finally observe that the probability of an arriving customer in steady state joining a queue of length at least \(i\) is \(S_i^n(\infty)^d\). For \(i \leq \log_d \eta_n - \omega_n\) because we have \(S_i^n(\infty)^d \to 1\) in probability, a customer arriving in steady state experiences a delay of at least \(\log_d \eta_n - \omega_n\) with probability approaching one as \(n \to \infty\).

Beyond this elementary bound on the expectation in steady state, Theorem 3.2 suggests that \(T^n(\infty)\) converges to \(\pi\), though formally this is a conjecture because we do not show that the sequence \(T^n(\infty)\) is tight. Establishing this interchange of limits would be a potential direction for future work on this system. For the remainder of this section we will suppose the conjecture is true and consider the implications.

First, treating the fixed point (3.15) as the limit of \(T^n(\infty)\), we have

$$\pi_1 = \lim_{n \to \infty} \eta_n (1 - S_1^n(\infty)).$$
Figure 3-1: Simulated steady state expectation for M/M/n-S up(d) with $d = 2$ and $\lambda_n = 1 - \beta n^{-\alpha}$ with $\beta = 2$ and $\alpha = 3/4$. For each line the quantity $\alpha \log_d n$ is rounded to the nearest integer. Each horizontal line indicates $\exp(-\beta d^k/(d - 1))$, which is the conjectured limit of $\mathbb{E}S_{\alpha \log_d n + k}^n$ as $n$ diverges to infinity.

We use Little’s law to replace $S_1^n(\infty)$ by $\lambda_n$ so we have

$$\pi_1 = \lim_{n \to \infty} \eta_n (1 - \lambda_n) = \beta,$$

and therefore the fixed point becomes

$$\pi_i = \beta \frac{d^{i - 1}}{d - 1}, \quad i \geq 0.$$

Recall that $T_i(t) = \eta_n (1 - S_i(t))$, so this fixed point suggests that in steady state the fraction of servers with at least $i$ jobs can be approximated by

$$S_i^n(\infty) \approx 1 - \frac{\beta d^{i - 1}}{\eta_n d - 1}$$

when $i = \log_d \eta_n - \omega_n$. 
We further conjecture that delay times longer than \( \log_d \eta_n + \omega_n \) are unlikely. This is informed by a heuristic analysis of the heavy traffic supermarket model using the fixed \( \lambda < 1 \) results proved by Mitzenmacher [26] and Vvedenskaya, et. al. [32]. For fixed \( \lambda < 1 \), the system converges to a limiting system which has a unique fixed point at

\[ \pi_i = \lambda^{\frac{d-1}{d}}. \]

If we simply replace \( \lambda \) with \( \lambda_n = 1 - \frac{\beta}{\eta_n} \), then we have

\[ \pi_i = \left(1 - \frac{\beta}{\eta_n}\right)^{\frac{d-1}{d}} \rightarrow \begin{cases} 1 & i \leq \log_d \eta_n - \omega_n \\ e^{-\frac{\beta}{d}k} & i = \log_d \eta_n + k \\ 0 & i \geq \log_d \eta_n + \omega_n, \end{cases} \]

where \( \omega_n \) is any sequence diverging to infinity and \( k \) is any constant. In particular, this heuristic suggests that in steady state all queues will have length \( \log_d \eta_n + O(1) \).

As further evidence for this conjectured behavior, we simulated the system for a variety of values of \( n \) to estimate the expectation in steady state. Figure 3-1 shows that the fraction of queues of length at least \( i = \log_d \eta_n + k \) remains approximately constant as \( n \) increases. Furthermore, these simulated steady state expectations appear to vary around the conjectured limits of \( E S^n_i \) for such \( i \), with the variation primarily introduced by the necessary rounding of \( i \) to an integer value. Figure 3-2 shows that this same behavior holds for another choice of parameters \( \beta \) and \( d \), and Figure 3-3 further demonstrates that the basic behavior is unchanged as \( \alpha \) changes. Note that changing \( \alpha \) does not impact the limiting behavior for given \( \beta \) and \( d \), but rather changes the frequency with which rounding causes a jump in a given \( E S^n_{\log_d \eta_n + k} \) line. All of these simulations were performed by allowing the system to run through a sufficient number of events (arrivals and departures) to reach steady state and continuing to run the system, averaging the values from then on.
Figure 3-2: Simulated steady state expectation for M/M/n-Sup(d) with $d = 3$ and $\lambda_n = 1 - \beta n^{-\alpha}$ with $\beta = 0.5$ and $\alpha = 3/4$. For each line the quantity $\alpha \log_d n$ is rounded to the nearest integer. Each horizontal line indicates $\exp(-\beta d^k/(d - 1))$, which is the conjectured limit of $\mathbb{E}S_{\alpha \log_d n+k}$ as $n$ diverges to infinity.
Figure 3-3: Simulated steady state expectation for M/M/n-Sup(d) with $d = 5$ and
$\lambda_n = 1 - \beta n^{-\alpha}$ with $\beta = 1$ and $\alpha = 1/4$. For each line the quantity $\alpha \log_d n$ is rounded
to the nearest integer. Each horizontal line indicates $\exp(-\beta d^k/(d-1))$, which is the
conjectured limit of $\mathbb{E}S_{\alpha \log_d n + k}$ as $n$ diverges to infinity.
### 3.3 Integral representation.

We prove Theorem 3.1 in this section. We will make use of a version of Gronwall’s inequality, which we state now as a lemma (see, e.g., pg. 498 of [10]).

**Lemma 3.1.** Suppose that $g : [0, \infty) \to [0, \infty)$ is a function such that

$$0 \leq g(t) \leq \epsilon + M \int_0^t g(s) ds, \quad t \geq 0,$$

for some positive finite $\epsilon$ and $M$. Then

$$g(t) \leq \epsilon e^{Mt}, \quad t \geq 0.$$

We begin by establishing two lemmas related to the function $g^n$.

**Lemma 3.2.** Let $0 < \alpha < 1/2$, and $\rho > 1$, and let $\eta^n, \eta \in \mathbb{R}_{\geq 1}$ be such that $\eta^n \to \eta, i_n^* = \frac{\alpha}{2} \log \rho, \eta^n, x^n_i \in D([0, \infty), [-\eta^n, (\eta^n)\alpha])$ for $0 \leq i \leq i_n^*$ and $x^n_i \in D([0, \infty), [\eta^n - 1, \eta^n + 1])$ for $i > i_n^*$. Then for any $i_0 \in \mathbb{N}, t \geq 0$,

$$\left\| g^n(x^n_i) - g^n(x^n_{i_0}) \right\|_t \to 0, \quad i \leq i_0.$$

**Proof.** First suppose $\eta < \infty$. Then for large enough $n$ we have $\eta_n < \infty$. For $i \geq 0$ we have

$$\left\| g^n(x^n_i) - g^n(x^n_{i_0}) \right\|_t = \sup_{0 \leq s \leq t} \frac{1}{d} \left| \sum_{l=2}^d \binom{d}{l} (-1)^l x^n_i(s)^l \left( (\eta^n)^{1-l} - \eta^{1-l} \right) \right|$$

$$\leq \frac{1}{d} \sum_{l=2}^d \binom{d}{l} (\eta^n + 1)^l \left| (\eta^n)^{1-l} - \eta^{1-l} \right| \Rightarrow C_n.$$

Observe that as $\eta^n \to \eta, C_n \to 0$, as desired.
Suppose now $\eta = \infty$ and thus $g^\eta = 0$. For $0 \leq i \leq i^*_n$ we have

$$\left| \left| g^n(x^*_i) - g^n(x_i^n) \right| \right|_t = 1 \leq \sum_{l=2}^{d} \left( \begin{array}{c} d \\ l \end{array} \right) \frac{1}{\eta^n} \frac{1}{(\eta^n)^{1-l}}$$

and

$$\leq \sum_{l=2}^{d} \left( \begin{array}{c} d \\ l \end{array} \right) \frac{1}{\eta^n} \frac{1}{(\eta^n)^{1-l(1-\alpha)}}.$$

As $\eta^n \to \infty$ we have $i^*_n \to \infty$ and thus for large enough $n$, $i_0 < i^*_n$. For $2 \leq l \leq d$ we have

$$(\eta^n)^{1-l(1-\alpha)} \to 0,$$

and thus

$$\left| \left| g^n(x^*_i) \right| \right|_t \to 0.$$

\[\square\]

**Lemma 3.3.** For every $\eta \in \mathbb{R}_{\geq 1}$, $g^\eta$ is a Lipschitz continuous function with constant $4^d$ on $\tilde{D}^{\eta+2}$ equipped with the topology induced by $||\cdot||_{\rho,t}$. That is, for $x^1, x^2 \in \tilde{D}^{\eta+2}$, and for $i \geq 0, t \geq 0$,

$$\left| \left| g^n(x^1_i) - g^n(x^2_i) \right| \right|_t \leq 4^d \left| \left| x^1_i - x^2_i \right| \right|_t$$

and

$$\left| \left| g^n(x^1) - g^n(x^2) \right| \right|_{\rho,t} \leq 4^d \left| \left| x^1 - x^2 \right| \right|_{\rho,t}.$$

**Proof.** Consider the restriction of $g^\eta$ onto $[-\eta - 2, \eta + 2] \to \mathbb{R}$. This function is differentiable and thus is Lipschitz continuous with constant

$$\sup_{y \in [-\eta - 2, \eta + 2]} \left| \frac{g^n(y) - g^n(y')}{y - y'} \right| = \sup_{y \in [-\eta - 2, \eta + 2]} \left| \left( 1 - \frac{y}{\eta} \right)^{d-1} + 1 \right| \leq 4^d.$$

Thus for $y^1, y^2 \in [-\eta - 2, \eta + 2]$ we have $|g^n(y^1) - g^n(y^2)| \leq 4^d |y^1 - y^2|$, which further
implies for $x^1, x^2 \in \tilde{D}^{\eta+2}$ and for $i \geq 0$ we have

$$
\left| g^\eta(x^1_i) - g^\eta(x^2_i) \right|_t = \sup_{0 \leq s \leq t} \left| g^\eta(x^1_i(s)) - g^\eta(x^2_i(s)) \right|
\leq 4^d \sup_{0 \leq s \leq t} \left| x^1_i(s) - x^2_i(s) \right|
\leq 4^d \left| x^1_i - x^2_i \right|_t.
$$

This further implies

$$
\left| g^\eta(x^1) - g^\eta(x^2) \right|_{\rho,t} = \sum_{i \geq 0} \rho^{-i} \left| g^\eta(x^1_i) - g^\eta(x^2_i) \right|_t
\leq \sum_{i \geq 0} \rho^{-i} 4^d \left| x^1_i - x^2_i \right|_t
= 4^d \left| x^1 - x^2 \right|_{\rho,t}.
$$

\[\square\]

**Proof of Theorem 3.1: Existence and uniqueness.** Fix $(b, y, \lambda, \eta) \in \mathbb{R}^{\infty,\rho} \times D^{\infty,\rho} \times \mathbb{R} \times \mathbb{R}_{\geq 1}$.

Suppose first $T(0) = b + y(0) \notin \mathcal{Z}^\eta$. Then $t^* = 0$ and $\dot{T}(t) = T(0)$ for all $t \geq 0$.

We now suppose $T(0) = b + y(0) \in \mathcal{Z}^\eta$, and therefore for all $1 \leq i \leq i^*$ we have $|T_i(0)| \leq \eta^\alpha$ and for all $i > i^*$ we have $|T_i(0)| \leq \eta + 1$. We will show existence and uniqueness via a contraction mapping argument, showing that the map defined by the right hand side of (3.6)-(3.7) is a contraction for small enough $t$. Note that this contraction argument will use the unbounded $\rho$-norm, and uniqueness of the solution with respect to that topology implies uniqueness with respect to the product topology.

We first define the map $\Gamma : D^{\infty,\rho} \rightarrow D^{\infty}$, where for $x \in D^{\infty,\rho}$ and $t \geq 0$,

$$
\Gamma(x)_0(t) = 0,
\Gamma(x)_i(t) = b_i + y_i(t) - \lambda d \int_{0}^{t} (x_i(s) - x_{i-1}(s)) - g^\eta(x_i) + g^\eta(x_{i-1})) ds
+ \int_{0}^{t} (x_{i+1}(s) - x_i(s)) ds, \quad i \geq 1.
$$
Now let
\[ t^* \triangleq \inf \left\{ s \geq 0 : \exists \ i \ s.t. \ 1 \leq i \leq t^* \ \text{and} \ |\Gamma(x)_i(s)| \geq \eta^\alpha \right\} \]

or
\[ \exists \ i \ s.t. \ i > t^* \ \text{and} \ |\Gamma(x)_i(s)| \geq \eta + 1 \right\} \]

and further define \( \hat{\Gamma} : D^{\infty,\rho} \rightarrow D^\infty \) by
\[
\hat{\Gamma}(x)_i(t) = \begin{cases} 
\Gamma(x)_i(t) & t < t^* \\
\Gamma(x)_i(t^*) & t \geq t^*, 
\end{cases}
\]

for all \( i \geq 0 \).

By construction, \( \hat{\Gamma}(x)_i(t) \in [-\eta - 1, \eta + 1] \) for all \( i \geq 0 \), and thus \( \hat{\Gamma} : D^{\infty,\rho} \rightarrow \tilde{D}^{\eta+1} \).

Further note that if \( \eta = \infty \), then \( g^\eta = 0 \) and
\[
\left\| \hat{\Gamma}(x) \right\|_{\rho,t} \leq \left\| b \right\|_\rho + \left\| y \right\|_{\rho,t} + \sum_{i \geq 1} \rho^{-i} \lambda dt \left( \left\| x_i \right\|_t + \left\| x_{i-1} \right\|_t \right)
+ \sum_{i \geq 1} \rho^{-i} t \left( \left\| x_{i+1} \right\|_t + \left\| x_i \right\|_t \right)
\leq \left\| b \right\|_\rho + \left\| y \right\|_{\rho,t} + \lambda dt \sum_{i \geq 1} \rho^{-i} \left\| x_{i-1} \right\|_t 
+ t(\lambda d + 1) \sum_{i \geq 1} \rho^{-i} \left\| x_i \right\|_t 
+ t \sum_{i \geq 1} \rho^{-i} \left\| x_{i+1} \right\|_t.
\]

We bound each of these sums individually. Observe
\[
\sum_{i \geq 1} \rho^{-i} \left\| x_{i-1} \right\|_t = \rho^{-1} \sum_{i \geq 1} \rho^{-(i-1)} \left\| x_{i-1} \right\|_t
= \rho^{-1} \sum_{i \geq 0} \rho^{-i} \left\| x_i \right\|_t = \rho^{-1} \left\| x \right\|_{\rho,t}.
\]
Similarly,
\[
\sum_{i \geq 1} \rho^{-i} \|x_i\|_t \leq \sum_{i \geq 0} \rho^{-i} \|x_i\|_t = \|x\|_{\rho, t},
\]
and
\[
\sum_{i \geq 1} \rho^{-i} \|x_{i+1}\|_t = \rho \sum_{i \geq 1} \rho^{-(i+1)} \|x_{i+1}\|_t
\]
\[
= \rho \sum_{i \geq 2} \rho^{-i} \|x_i\|_t
\]
\[
\leq \rho \sum_{i \geq 0} \rho^{-i} \|x_i\|_t = \rho \|x\|_t.
\]
Therefore we have
\[
\left\| \hat{\Gamma}(x) \right\|_{\rho, t} \leq t \left( \lambda d \rho^{-1} + \lambda d + 1 + \rho \right) \|x\|_{\rho, t} < \infty,
\]
and thus \( \hat{\Gamma} : D^{\infty, \rho} \rightarrow \tilde{D}^{\eta+1, \rho} \).

We now show that for
\[
t_0 < \frac{1}{\lambda d (1 + \rho^{-1})(1 + 4d) + 1 + \rho},
\]
\( \hat{\Gamma} \) is a contraction on \( \tilde{D}^{\eta+1, \rho}_t = D ([0, t], [-\eta + 1, \eta + 1]\infty) \) for all \( t \leq t_0 \). Namely, we claim that there exists \( \gamma < 1 \) such that for all \( t \in [0, t_0] \) and \( x^1, x^2 \in \tilde{D}^{\eta+1, \rho}_t \), we have
\[
\left\| \hat{\Gamma}(x^1) - \hat{\Gamma}(x^2) \right\|_{\rho, t} \leq \gamma \|x^1 - x^2\|_{\rho, t}.
\] (3.17)
Let \( t \leq t_0 \) and \( x^1, x^2 \in \tilde{D}^{\eta+1,\rho}_t \). We have for \( i \geq 1 \)

\[
\left| \left| \hat{\Gamma}(x^1)_i - \hat{\Gamma}(x^2)_i \right| \right|_t \leq \lambda d \int_0^t \left( \left| x^1_i - x^2_i \right|_s + \left| x^1_{i-1} - x^2_{i-1} \right|_s \right.
\]
\[
+ \left| g^\eta(x^1_i) - g^\eta(x^2_i) \right|_s + \left| g^\eta(x^1_{i-1}) - g^\eta(x^2_{i-1}) \right|_s \big) \, ds
\]
\[
+ \int_0^t \left( \left| x^1_{i+1} - x^2_{i+1} \right|_s + \left| x^1_i - x^2_i \right|_s \right) \, ds.
\]

By Lemma \([3.3]\) for \( \eta < \infty \), \( g^\eta \) is Lipschitz when restricted to \( \tilde{D}^{\eta+2,\rho}_t \), with constant \( 4^d \). For \( \eta = \infty \), \( g^\infty = 0 \). Thus we now have

\[
\left| \left| \hat{\Gamma}(x^1) - \hat{\Gamma}(x^2) \right| \right|_t \leq t \lambda d (1 + 4^d) \left| x^1_{i-1} - x^2_{i-1} \right|_t
\]
\[
+ t \left( \lambda d (1 + 4^d) + 1 \right) \left| x^1_i - x^2_i \right|_{\rho, t} + t \left| x^1_{i+1} - x^2_{i+1} \right|_t.
\]

This implies

\[
\left| \left| \hat{\Gamma}(x^1) - \hat{\Gamma}(x^2) \right| \right|_{\rho, t} \leq t \lambda d (1 + 4^d) \sum_{i \geq 1} \rho^{-i} \left| x^1_{i-1} - x^2_{i-1} \right|_t
\]
\[
+ t \left( \lambda d (1 + 4^d) + 1 \right) \sum_{i \geq 1} \rho^{-i} \left| x^1_i - x^2_i \right|_t
\]
\[
+ t \sum_{i \geq 1} \rho^{-i} \left| x^1_{i+1} - x^2_{i+1} \right|_t.
\]

Reindexing and bounding these sums individually gives us

\[
\left| \left| \hat{\Gamma}(x^1) - \hat{\Gamma}(x^2) \right| \right|_{\rho, t} \leq t \left( \lambda d (1 + \rho^{-1})(1 + 4^d) + 1 + \rho \right) \left| x^1 - x^2 \right|_{\rho, t}.
\]  (3.18)

Let

\[
t_0 < \frac{1}{\lambda d (1 + \rho^{-1})(1 + 4^d) + 1 + \rho}.
\]

Then \((3.17)\) holds with

\[
\gamma = t_0 \left( \lambda d (1 + \rho^{-1})(1 + 4^d) + 1 + \rho \right).
\]
By the contraction mapping principle, \( \hat{\Gamma} \) has a unique fixed point \( \hat{T} \) on \( \hat{D}^{q+1,\rho} \) such that \( \hat{\Gamma}(\hat{T}) = \hat{T} \). This fixed point provides a unique solution \( \hat{T} \) to (3.11) for \( t \in [0, t_0] \).

Suppose this fixed solution \( \hat{T} \) is such that \( t^* < t_0 \). Then \( \hat{T} \) is uniquely defined for all \( t \geq 0 \) and the proof is complete. Otherwise, observe for \( t \geq 0 \) and \( i \geq 1 \) we have

\[
T_i(t) = T_i(t_0) + y_i(t) - y_i(t_0) \\
- \lambda d \int_{t_0}^{t} (T_i(s) - T_{i-1}(s) - g^q(T_i(s)) + g^q(T_{i-1}(s))) \, ds \\
+ \int_{t_0}^{t} (T_{i+1}(s) - T_i(s)) \, ds.
\]

Thus if we define a shifted version of \( y \) by \( \tilde{y}(u) = y(u + t_0) \), then \( T_i(t) \) for \( t = u + t_0 \) and \( u \geq 0 \) is the solution to the system

\[
x_0(u) = 0 \\
x_i(u) = T_i(t_0) - y_i(t_0) + \tilde{y}_i(u) \\
- \lambda d \int_{0}^{u} (x_i(s) - x_{i-1}(s) - g^q(x_i(s)) + g^q(x_{i-1}(s))) \, ds \\
+ \int_{0}^{u} (x_{i+1}(s) - x_i(s)) \, ds, \quad i \geq 1.
\]

Observe that this is the system (3.6)-(3.7) with arguments

\[
(T(t_0) - y(t_0), \tilde{y}, \lambda, \eta) \in \mathbb{R}^{\infty,\rho} \times D^{\infty,\rho} \times \mathbb{R} \times \mathbb{R}_{\geq 1},
\]

and furthermore \( x(0) = T(t_0) - y(t_0) + \tilde{y}(0) = T(t_0) \) and \( T(t_0) \in \mathcal{Z}^\eta \) because \( t^* \geq t_0 \).

Thus we can repeat the contraction argument above to find a unique solution \( x \) for \( u \in [0, t_0] \). This unique \( x \) is the unique solution \( \hat{T} \) for \( t \in [t_0, 2t_0] \). If \( t^* < 2t_0 \), then \( \hat{T} \) is uniquely defined for all \( t \geq 0 \) and the proof is complete. Otherwise, the above extension argument can be repeated to find a unique solution \( \hat{T} \) for \([2t_0, 3t_0], [3t_0, 4t_0], \ldots \).

If \( t^* < kt_0 \) for some \( k \geq 3 \) then the argument stops there and we conclude \( \hat{T} \) is uniquely defined for all \( t \geq 0 \). Otherwise it may be extended to any \( t \geq 0 \). 

Before proving continuity, we state and prove a lemma bounding the growth of
Lemma 3.4. For any \( t \geq 0 \), if \( x \) is the solution to (3.11) for arguments \((b, y, \lambda, \eta)\), then
\[
\|x\|_{\rho,t} \leq \left( \|b\|_{\rho} + \|y\|_{\rho,t} \right) e^{(\lambda d(1+4^d)(1+\rho^{-1})+1+\rho)t}
\]

Proof. We have
\[
\|x_i\| t \leq |b_i| + \|y_i\| t + \lambda d \int_0^t (\|x_{i-1}\| s + \|x_i\| s + \|g(x_{i-1})\| s + \|g(x_i)\| s) ds
\]
and thus
\[
\|x\|_{\rho,t} \leq \|b\|_{\rho} + \|y\|_{\rho,t} + \lambda d (1 + 4^d) \sum_{i \geq 1} \rho^{-i} \int_0^t (\|x_{i-1}\| s + \|x_i\| s) ds
\]
\[
+ \sum_{i \geq 1} \rho^{-i} \int_0^t (\|x_i\| s + \|x_{i+1}\| s) ds,
\]
and thus
\[
\|x\|_{\rho,t} \leq \|b\|_{\rho} + \|y\|_{\rho,t} + \lambda d (1 + 4^d) \sum_{i \geq 1} \rho^{-i} \int_0^t (\|x_{i-1}\| s + \|x_i\| s) ds
\]
\[
+ \sum_{i \geq 1} \rho^{-i} \int_0^t (\|x_i\| s + \|x_{i+1}\| s) ds
\]
\[
\leq \|b\|_{\rho} + \|y\|_{\rho,t} + \left( \lambda d (1 + 4^d) (1 + \rho^{-1}) + 1 + \rho \right) \sum_{i \geq 0} \rho^{-i} \int_0^t \|x_i\| s ds
\]
\[
= \|b\|_{\rho} + \|y\|_{\rho,t} + \left( \lambda d (1 + 4^d) (1 + \rho^{-1}) + 1 + \rho \right) \int_0^t \sum_{i \geq 0} \rho^{-i} \|x_i\| s ds
\]
\[
= \|b\|_{\rho} + \|y\|_{\rho,t} + \left( \lambda d (1 + 4^d) (1 + \rho^{-1}) + 1 + \rho \right) \int_0^t \|x\|_{\rho,s} ds.
\]

By Gronwall’s inequality (Lemma 3.1), we have
\[
\|x\|_{\rho,t} \leq \left( \|b\|_{\rho} + \|y\|_{\rho,t} \right) e^{(\lambda d(1+4^d)(1+\rho^{-1})+1+\rho)t}.
\]

Proof of Theorem 3.1: continuity. We now prove continuity for \( f \) restricted to the domain \( Z^a_K \) for any \( K > 0 \). Suppose \( (b^n, y^n, \lambda^n, \eta^n) \to (b, y, \lambda, \eta) \) with respect to the product topology, with \( (b^n, y^n, \lambda^n, \eta^n) \in Z^a_K \) and since the set \( Z^a_K \) is closed, we also have \( (b, y, \lambda, \eta) \in Z^a_K \). Suppose \( x^n \) is the unique solution to (3.11) for \( (b^n, y^n, \lambda^n, \eta^n) \).
and \( x \) is the unique solution for \((b, y, \lambda, \eta)\). Let \( i_n^* = \frac{\log \rho \eta_n}{2} \) and \( i^* = \frac{\log \rho \eta}{2} \). Let \( t_n^* \) and \( t^* \) be the stopping times for \( x^n \) and \( x \), respectively.

Recall that by the definition of \( Z^\alpha_K \), we have \( x^n(0) = b^n + y^n(0) \in \mathbb{Z}^{\eta^n} \). Note that this, along with the definition of \( t_n^* \), implies that \( x^n(t) \in \mathbb{Z}^{\eta^n} \) for any \( t \geq 0 \). Similarly, we have \( x(t) \in \mathbb{Z}^{\eta} \) for any \( t \geq 0 \). We adopt the simplified notation \( g^n \triangleq g^{\eta^n} \) and \( g = g^{\eta} \).

We will show \( x^n \to x \) in \( D^{\infty, \rho} \) equipped with the product topology. Fix \( \epsilon > 0 \) and \( i_0 \in \mathbb{N} \).

Recall the map \( \hat{\Gamma} \) defined in the proof of existence and uniqueness above. We define \( \hat{\Gamma}^n \) and \( \hat{\Gamma} \) analogously for \((b^n, y^n, \lambda^n, \eta^n)\) and \((b, y, \lambda, \eta)\), respectively. Define

\[
t_0 = \frac{1}{2(1 + \lambda) d(1 + \rho^{-1})(1 + 4^d) + 1 + \rho},
\]

and

\[
\gamma = t_0 \left( (1 + \lambda) d(1 + \rho^{-1})(1 + 4^d) + 1 + \rho \right) = 1/2.
\]

By \[3.18\], for \( 0 \leq t \leq t_0 \) and \( x^1, x^2 \in D^{\infty, \rho} \), we have

\[
\left\| \hat{\Gamma}(x^1) - \hat{\Gamma}(x^2) \right\|_{\rho, t} \leq t_0 \left( (1 + \lambda) d(1 + \rho^{-1})(1 + 4^d) + 1 + \rho \right) \left\| x^1 - x^2 \right\|_{\rho, t} = \gamma \left\| x^1 - x^2 \right\|_{\rho, t}.
\]

Furthermore, because \( \lambda^n \to \lambda \), there exists some \( N_\lambda \) such that \( \lambda^n < \lambda + 1 \) for \( n \geq N_\lambda \), and thus for such \( n \),

\[
\left\| \hat{\Gamma}^n(x^1) - \hat{\Gamma}^n(x^2) \right\|_{\rho, t} \leq t_0 \left( \lambda^n d(1 + \rho^{-1})(1 + 4^d) + 1 + \rho \right) \left\| x^1 - x^2 \right\|_{\rho, t} \leq t_0 \left( (1 + \lambda) d(1 + \rho^{-1})(1 + 4^d) + 1 + \rho \right) \left\| x^1 - x^2 \right\|_{\rho, t} = \gamma \left\| x^1 - x^2 \right\|_{\rho, t}.
\]

Thus for \( 0 \leq t \leq t_0 \), \( \hat{\Gamma} \) and \( \hat{\Gamma}^n \) for \( n \geq N_\lambda \) are contractions on the space \( D^{\infty, \rho}_t \) with coefficient \( \gamma = 1/2 \). Recall that \( x \) and \( x^n \) are the fixed points of \( \hat{\Gamma} \) and \( \hat{\Gamma}^n \), respectively, and that therefore each can be found by repeated iteration of an arbitrary point in.
$D_t\infty,\rho$. Specifically, we define

\[
x^0 = 0 \\
x^n = 0 \\
x^r = \hat{\Gamma}(x^{r-1}) \\
x^n.r = \hat{\Gamma}^n(x^{n-1}, r \geq 1).
\]

Then the following inequalities hold:

\[
||x - x^r||_{\rho,t} \leq \frac{\gamma^n}{1 - \gamma} ||x^1 - x^0||_{\rho,t} = 2^{-r+1} ||x^1||_{\rho,t}, \\
||x^n - x^{n,r}||_{\rho,t} \leq \frac{\gamma^n}{1 - \gamma} ||x^{n,1} - x^0||_{\rho,t} = 2^{-r+1} ||x^{n,1}||_{\rho,t}.
\]

(3.19)

Observe

\[
||x^1||_{\rho,t} = ||\hat{\Gamma}(0)||_{\rho,t} = ||b + y||_{\rho,t} \leq ||b||_{\rho} + ||y||_{\rho,t} \leq 2K,
\]

and similarly

\[
||x^{n,1}||_{\rho,t} \leq ||b^n||_{\rho} + ||y^n||_{\rho,t} \leq 2K.
\]

We now argue that there exists $N$ such that for all $n \geq N$ and $i \leq i_0, ||x^n_i - x_i||_t < \epsilon$ for $t \leq t_0$. This will establish continuity of $f$ with respect to the product topology for such $t$. Observe

\[
||x^n_i - x_i||_t \leq ||x^n_i - x^{n,r}_i||_t + ||x^{n,r}_i - x^r_i||_t + ||x^r_i - x_i||_t.
\]

(3.20)

We bound these three terms individually. By (3.19), we have

\[
||x^n_i - x^{n,r}_i||_t \leq \rho^i ||x^n - x^{n,r}||_{\rho,t} \\
\leq 2^{-r+1}\rho^i \left(||b^n||_{\rho} + ||y^n||_{\rho,t}\right) \\
\leq 2^{-r+2}K\rho^{i_0}.
\]

For

\[
r > 2 + \log_2 \frac{3\rho^{i_0}K}{\epsilon},
\]

(3.21)
we have
\[ ||x^n_i - x^{n,r}_i||_t < \epsilon/3 \quad \text{for all } i \leq i_0. \] (3.22)

Via a similar argument, for \( r \) satisfying (3.21), we also have
\[ ||x_i - x^r_i||_t < \epsilon/3 \quad \text{for all } i \leq i_0. \] (3.23)

Finally, we consider
\[
||x^{n,r+1}_i - x^{r+1}_i||_t \leq |b^n_i - b_i| + ||y^n_i - y_i||_t \\
+ \lambda^n d \int_0^t \left( ||x^{n,r}_{i-1} - x^r_{i-1}||_s + ||x^{n,r}_i - x^r_i||_s \right) ds \\
+ ||g^n(x^{n,r}_{i-1}) - g(x^r_{i-1})||_s + ||g^n(x^{n,r}_i) - g(x^r_i)||_s \right) ds \\
+ ||g(x^r_{i-1})||_s + ||g(x^r_i)||_s \right) ds \\
+ \int_0^t \left( ||x^{n,r}_i - x^r_i||_s + ||x^{n,r}_{i+1} - x^r_{i+1}||_s \right) ds.
\]

By Lemma 3.3, \( g \) is Lipschitz continuous when restricted to \( \tilde{D}^{n+2}\rho \), so we have
\[
||g(x^r_{i-1})||_s + ||g(x^r_i)||_s \leq 4^d ||x^r_{i-1}||_s + 4^d ||x^r_i||_s.
\]

Further note the bound
\[
||g^n(x^{n,r}_i) - g(x^r_i)||_s \leq ||g^n(x^{n,r}_i) - g(x^{n,r}_{i})||_s + ||g(x^{n,r}_i) - g(x^r_i)||_s.
\]

Using these pieces, crudely bounding integrals for some terms, and rearranging we
have

\[
\left| x_{i}^{n,r+1} - x_{i}^{r+1} \right|_{t} \leq \left| b_{i}^{n} - b_{i} \right| + \left| y_{i}^{n} - y_{i} \right|_{t}
+ t \left| g^{n}(x_{i-1}^{n,r}) - g(x_{i-1}^{r}) \right|_{t} + t \left| g^{n}(x_{i}^{n,r}) - g(x_{i}^{r}) \right|_{t}
+ |\lambda - \lambda^{n}|t(1 + 4^{d}) \left( \left| x_{i-1}^{r} \right|_{t} + \left| x_{i}^{r} \right|_{s} \right)
+ \lambda^{n} d \int_{0}^{t} \left( \left| x_{i-1}^{r} - x_{i-1}^{n,r} \right|_{s} + \left| x_{i}^{r} - x_{i}^{n,r} \right|_{s} \right) ds
+ \lambda^{n} d \int_{0}^{t} \left( \left| g(x_{i-1}^{n,r}) - g(x_{i-1}^{r}) \right|_{s} + \left| g(x_{i}^{n,r}) - g(x_{i}^{r}) \right|_{s} \right) ds
+ \int_{0}^{t} \left( \left| x_{i}^{n,r} - x_{i}^{r} \right|_{s} + \left| x_{i+1}^{n,r} - x_{i+1}^{r} \right|_{s} \right) ds.
\]

Recall that for any \( t \geq 0 \) we have \( x_{i}^{n,r}(t) \in Z^{\eta} \) and thus \( \left| x_{i}^{n,r}(t) \right| \leq (\eta^{n})^{\alpha} \) for \( i \leq i_{n}^{*} \), and \( \left| x_{i}^{n,r}(t) \right| \leq \eta^{n} + 1 \) for \( i > i_{n}^{*} \), so the conditions of Lemma 3.2 are satisfied and therefore for \( i \leq i_{0} \) we have \( \left| g^{n}(x_{i}^{n,r}) - g(x_{i}^{r}) \right|_{t} \to 0 \), and similarly \( \left| g^{n}(x_{i-1}^{n,r}) - g(x_{i-1}^{r}) \right|_{t} \to 0 \). This, along with \( (b^{n}, y^{n}, \lambda^{n}) \to (b, y, \lambda) \) implies that for any \( \delta > 0 \) we can choose \( N_{\delta} \) such that for \( n \geq N_{\delta} \) we have

\[
\lambda_{n} \leq \lambda + 1,
\]

\[
\eta^{n} \leq \eta + 1,
\]

and for all \( i \leq i_{0} + r + 1 \),

\[
\left| b_{i}^{n} - b_{i} \right| + \left| y_{i}^{n} - y_{i} \right|_{t}
+ t \left| g^{n}(x_{i-1}^{n,r}) - g(x_{i-1}^{r}) \right|_{t} + t \left| g^{n}(x_{i}^{n,r}) - g(x_{i}^{r}) \right|_{t}
+ |\lambda - \lambda^{n}|t(1 + 4^{d}) \left( \left| x_{i-1}^{r} \right|_{t} + \left| x_{i}^{r} \right|_{t} \right)
< \delta.
\]

Observe that \( \eta^{n} \leq \eta + 1 \) implies \( x_{i}^{n,r} \in D^{\eta+2^{d}} \), so

\[
\left| g(x_{i}^{n,r}) - g(x_{i}^{r}) \right|_{s} \leq 4^{d} \left| x_{i}^{n,r} - x_{i}^{r} \right|_{s}.
\]
For $n \geq N\delta$, we have for all $i \leq i_0 + r + 1$

$$
||x_i^{n,r+1} - x_i^{r+1}|| < \delta + \int_0^t (||x_i^{n,r} - x_i^r|| + ||x_i^{n,r} - x_{i+1}^r||) \, ds
$$

$$
+ (1 + 4^d)(1 + \lambda)d \int_0^t (||x_i^{n,r} - x_i^{r-1}|| + ||x_i^{n,r} - x_i^r||) \, ds.
$$

We rewrite this as

$$
||x_i^{n,r+1} - x_i^{r+1}|| < \delta + C \int_0^t \max_{i-1 \leq j \leq i+1} ||x_j^{n,r} - x_j^r|| \, ds,
$$

where $C = (2(1 + 4^d)(1 + \lambda)d + 2)$. For $i \leq i_0$, this can be expanded as

$$
||x_i^{n,r+1} - x_i^{r+1}|| < \delta + Ct \max_{i-2 \leq j \leq i+2} ||x_j^{n,r-1} - x_j^r||
$$

$$
< \delta + Ct \left( \delta + Ct \max_{j=0}^r ||x_j^{n,0} - x_j^0|| \right)
$$

$$
< \delta \sum_{k=0}^r (Ct)^k + (Ct)^{r+1} \max_{(i-r-1) \leq j \leq i+r+1} ||x_j^{n,0} - x_j^0||
$$

Recall that $x_i^{n,0} = x_i^0 = 0$, so for all $i \geq 0$ we have $||x_i^{n,0} - x_i^0|| = 0$, and thus

$$
||x_i^{n,r+1} - x_i^{r+1}|| < \delta \frac{(Ct)^{r+1} - 1}{Ct - 1}.
$$

Reindexing gives, for all $i \leq i_0$,

$$
||x_i^{n,r} - x_i^r|| < \delta \frac{(Ct)^r - 1}{Ct - 1},
$$

and thus for

$$
\delta < \frac{\epsilon}{3} \cdot \frac{Ct - 1}{(Ct)^r - 1},
$$

we have

$$
||x_i^{n,r} - x_i^r|| < \epsilon/3. \tag{3.24}
$$
Thus by plugging (3.22), (3.23), and (3.24) into (3.20), if we choose $r$ and $n$ such that
\[ r > 2 + \log_2 \frac{3\rho^i_0 K}{\epsilon}, \quad \text{and} \quad n \geq \max (N_{\lambda}, N_{\delta}), \]
we have for all $i \leq i_0$,
\[ ||x^n_i - x_i||_t < \epsilon, \]
establishing the continuity of $f$ for $t \leq t_0$.

As in the proof of existence and uniqueness, we define a shifted version of $y$ by
\[ \tilde{y}(u) = y(u + t_0) \]
observing that $x(t)$ for $t = u + t_0$ and $u \geq 0$ is the solution $\hat{z}$ to the system
\[
\begin{align*}
    z_0(u) &= 0 \\
    z_i(u) &= x_i(t_0) - y_i(t_0) + \tilde{y}_i(u) \\
    &\quad - \lambda d \int_0^u (x_i(s) - x_{i-1}(s) - g^\eta(x_i(s)) + g^\eta(x_{i-1}(s))) \, ds \\
    &\quad + \int_0^u (x_{i+1}(s) - x_i(s)) \, ds, \quad i \geq 1
\end{align*}
\]
\[
\hat{z}(u) = \begin{cases} 
    z(u) & u < u^* \\
    z(u^*) & u \geq u^*, 
\end{cases}
\]
where $u^*$ is defined analogously to $t^*$ in (3.10). Observe that this is the system (3.6)-(3.7), (3.11) with arguments
\[
(x(t_0) - y(t_0), \tilde{y}, \lambda, \eta) \in \mathbb{R}^{\infty, \rho} \times D^{\infty, \rho} \times \mathbb{R} \times \mathbb{R}_{\geq 1},
\]
and furthermore $\hat{z}(0) = x(t_0) - y(t_0) + \tilde{y}(0) = x(t_0) \in \mathcal{Z}^\eta$. Also $||\tilde{y}||_{\rho, u} \leq ||y||_{\rho, u + t_0} < K$ and by Lemma 3.4 we have
\[
||x(t_0) - y(t_0)||_{\rho, t} \leq ||y(t_0)||_{\rho} + ||x(t_0)||_{\rho} \\
\leq K + (||b||_{\rho} + ||y||_{t_0, \rho}) e^{(\lambda d(1+4^d)(1+\rho^{-1})+1+\rho)t_0} \\
\leq K + 2K e^{(1+\lambda d(1+4^d)(1+\rho^{-1})+1+\rho)t_0}.
\]
We define
\[ K_1 \triangleq K + 2Ke^{((1+\lambda)d(1+4d)(1+\rho^{-1})+1+\rho)t_0}, \]
so
\[ (x(t_0) - y(t_0), \tilde{y}, \lambda, \eta) \in Z_{K_1}^\alpha. \]
A similar construction allows us to define \( \tilde{y}^n \) and \( \hat{z}^n \), and we have \( \tilde{y}^n \to \tilde{y} \) in \( D_u^{\infty,\rho} \) for any \( u \geq 0 \). For sufficiently large \( n \) we have \( \lambda^n < \lambda + 1 \), and thus
\[
\|x^n(t_0) - y^n(t_0)\|_{\rho,t} \leq K + 2Ke^{(\lambda^n d(1+4d)(1+\rho^{-1})+1+\rho)t_0} \\
\leq K + 2Ke^{((1+\lambda)d(1+4d)(1+\rho^{-1})+1+\rho)t_0} = K_1,
\]
so
\[ (x^n(t_0) - y^n(t_0), \tilde{y}^n, \lambda^n, \eta^n) \in Z_{K_1}^\alpha. \]
Therefore we can repeat the continuity argument above to show \( \hat{z}^n \to \hat{z} \) for \( u \leq t_0 \), which implies \( x^n \to x \) for \( t \in [t_0, 2t_0] \). This extension argument can be repeated to prove \( x^n \to x \) for \([2t_0, 3t_0], [3t_0, 4t_0], \ldots \). Thus for any \( t \geq 0 \) we have \( x^n \to x \) in \( D_t^{\infty,\rho} \) and thus \( f \) is continuous. \( \square \)

### 3.4 Martingale representation.

We now show that the stochastic process underlying the supermarket system stopped at some appropriate time can be written in a form that exactly matches that of \( \hat{T} \) in (3.11). This will allow us to use Theorem 3.1 to prove Theorem 3.2 in the next section.

Before introducing the stopped variant, we will consider the original supermarket system and show that it can be represented by the equations (3.6)-(3.7) for a particular choice of arguments \((b, y, \lambda, \eta)\).

For \( i \geq 1 \), recall the representation (3.3). Given the definition \( T_i^n = \eta_n (1 - S_i^n) \)
we can rewrite this as

\[
T^n_i(t) = T^n_i(0) - \frac{\eta_n}{n} A_i \left( \lambda_n n \int_0^t \left( \left( 1 - \frac{1}{\eta_n} T^n_{i-1}(s) \right)^d - \left( \frac{1}{\eta_n} T^n_i(s) \right)^d \right) ds \right) \\
+ \frac{\eta_n}{n} D_i \left( \frac{n}{\eta_n} \int_0^t \left( 1 - \frac{1}{\eta_n} T^n_i(s) \right) ds \right)
\]

\[
= T^n_i(0) - \frac{\eta_n}{n} A_i \left( \lambda_n \int_0^t \left( \left( 1 - \frac{1}{\eta_n} T^n_{i-1}(s) \right) + \frac{d}{\eta_n} g^n(T^n_{i-1}(s)) \right) ds \right) \\
+ \frac{\eta_n}{n} D_i \left( \frac{n}{\eta_n} \int_0^t \left( T^n_{i-1}(s) - T^n_i(s) \right) ds \right),
\]

(3.25)

where \( g^n \) is defined as in (3.5). We now define scaled martingale processes

\[
M^n_i(t) = \frac{\eta_n}{n} A_i \left( \lambda_n \int_0^t \left( T^n_i(s) - T^n_{i-1}(s) + g^n(T^n_i(s)) - g^n(T^n_{i-1}(s)) \right) ds \right) \\
- \lambda_n \int_0^t \left( T^n_i(s) - T^n_{i-1}(s) + g^n(T^n_i(s)) - g^n(T^n_{i-1}(s)) \right) ds,
\]

(3.26)

\[
N^n_i(t) = \frac{\eta_n}{n} D_i \left( \frac{n}{\eta_n} \int_0^t \left( T^n_{i+1}(s) - T^n_i(s) \right) ds \right) - \int_0^t \left( T^n_{i+1}(s) - T^n_i(s) \right) ds.
\]

(3.27)

Now we can rewrite the system for \( i \geq 1 \) as

\[
T^n_i(t) = T^n_i(0) - M^n_i(t) - \lambda_n \int_0^t \left( T^n_i(s) - T^n_{i-1}(s) + g^n(T^n_i(s)) - g^n(T^n_{i-1}(s)) \right) ds \\
+ N^n_i(t) + \int_0^t \left( T^n_{i+1}(s) - T^n_i(s) \right) ds.
\]

(3.28)

This representation matches (3.4) with \( b = T^n(0), y = -M^n + N^n, \lambda = \lambda_n \) and \( \eta = \eta_n \).

Recall that the assumptions of Theorem 3.2 include constants \( \rho > 1 \) and \( 0 < \alpha < \)
We now define $i^*_n = \frac{\alpha}{2} \log_\rho \eta_n$ and define a stopping time

$$t^*_n = \inf \{ t \geq 0 : \exists i \text{ s.t. } 1 \leq i \leq i^*_n \text{ and } T^n_i(t) \geq \eta^\alpha_n \}.$$ 

Note that compared to (3.10), this stopping time does not contain terms checking $T^n_i(t) \leq - (\eta_n)^\alpha$, or $|T^n_i(t)| \geq \eta_n + 1$. This is because $0 \leq T^n_i(t) \leq \eta_n$ for all $i \geq 0$ and for all $t \geq 0$, so such conditions are never met. Thus $t^*_n$ is equivalent to the stopping time defined by (3.10). We consider the process $\hat{T}^n$ defined by

$$\hat{T}^n_i(t) = \begin{cases} T^n_i(t) & t < t^*_n \\ T^n_i(t^*_n) & t \geq t^*_n. \end{cases}$$

As noted above, the stopped supermarket model $\hat{T}^n$ is the unique solution of the integral equation system (3.11) described in Theorem 3.1, with arguments $b = T^n(0)$, $y = -M^n + N^n$, $\lambda = \lambda_n$, and $\eta = \eta_n$.

### 3.5 Martingale convergence.

Because our sequence of supermarket models indexed by $n$ are all examples of the integral equation system (3.11), and Theorem 3.1 shows that this system defines a continuous map from arguments $(b, y, \lambda, \eta) \in Z^\alpha_K$ for some $K > 0$ to the stopped system $f(b, y, \lambda, \eta) = \hat{T}$, we will find the weak limit of the finite system $\hat{T}^n$ by finding the limits of the arguments $(T^n(0), -M^n + N^n, \lambda_n, \eta_n)$. Though we do not use the continuous mapping theorem because we do not have $(T^n(0), -M^n + N^n, \lambda_n, \eta_n) \in Z^\alpha_K$ almost surely for any non-random $K$, the proof will still rely on the continuity of $f$ and the limits of the arguments.

Three of these limits are given, as (3.12) provides $\hat{T}^n(0) \to T(0)$, and we have $\lambda_n \to 1$ and $\eta_n \to \infty$. We claim $-M^n + N^n \to 0$. We now prove the following:

**Proposition 3.1.** For $M^n$ and $N^n$ as defined in (3.26)-(3.27), if the assumptions of
Theorem 3.2 hold, then for $t \geq 0$,

$$\mathbb{E} ||M^n||_{\rho,t}, \mathbb{E} ||N^n||_{\rho,t} \to 0 \quad \text{as } n \to \infty.$$  

This implies $M^n, N^n \to 0$ in $D^{\infty,\rho}$ equipped with the product topology.

Before proving this proposition, we prove a bound on the unbounded $\rho$-norm of $T^n$ in expectation.

**Lemma 3.5.** For any $\gamma > 1$ we have

$$\mathbb{E} \left[ ||T^n||_{\gamma,t} \right] \leq \mathbb{E} \left[ ||T^n(0)||_{\gamma} \right] e^{\gamma t}. \quad (3.29)$$

**Proof.** By dropping negative terms from (3.25) we have

$$T^n_i(t) \leq T^n_i(0) + \frac{\eta n}{n} D_i \left( \frac{n}{\eta n} \int_0^t T^n_{i+1}(s)ds \right).$$

Then we have

$$\mathbb{E} \left[ ||T^n_i||_t \right] \leq \mathbb{E} T^n_i(0) + \mathbb{E} \left[ \int_0^t T^n_{i+1}(s)ds \right] \leq \mathbb{E} T^n_i(0) + \int_0^t \mathbb{E} \left[ ||T^n_{i+1}||_s \right] ds.$$

This implies

$$\mathbb{E} \left[ ||T^n||_{\gamma,t} \right] = \sum_{i \geq 1} \gamma^{-i} \mathbb{E} \left[ ||T^n_i||_t \right] \leq \mathbb{E} \left[ ||T^n(0)||_{\gamma} \right] + \sum_{i \geq 1} \gamma^{-i} \int_0^t \mathbb{E} \left[ ||T^n_{i+1}||_s \right] ds \leq \mathbb{E} \left[ ||T^n(0)||_{\gamma} \right] + \gamma \int_0^t \mathbb{E} \left[ ||T^n||_{\gamma,s} \right] ds.$$

We can now apply Lemma 3.1 to conclude

$$\mathbb{E} \left[ ||T^n||_{\gamma,t} \right] \leq \mathbb{E} \left[ ||T^n(0)||_{\gamma} \right] e^{\gamma t},$$

94
as desired.

Proof of Proposition 3.4. We first prove the statement for $N^n$. We begin with some observations about $N^n$ and introduce some additional definitions. Let

$$\tau^n_i \triangleq \int_0^t \left( T^n_{i+1}(s) - T^n_i(s) \right) ds, \quad i \geq 0.$$  \hfill (3.30)

Observe for all $i \geq 0$

$$\tau^n_i \leq t \left| | T^n_{i+1} \right|_t.$$  \hfill (3.31)

Now observe

$$\left| | N^n_i \right|_t = \sup_{0 \leq s \leq \tau^n_i} \left| \frac{n}{\eta_n} D_i \left( \frac{n}{\eta_n} \int_0^s \left( T^n_{i+1}(s) - T^n_i(s) \right) ds \right) - \int_0^s \left( T^n_{i+1}(s) - T^n_i(s) \right) ds \right|$$

$$= \sup_{0 \leq s \leq \tau^n_i} \left| \frac{n}{\eta_n} D_i \left( \frac{n}{\eta_n} s \right) - s \right|$$

$$= \sup_{0 \leq u \leq \frac{n \tau^n_i}{\eta_n}} \left| D_i(u) - u \right|.$$  \hfill (3.32)

We claim

$$\lim_{n \to \infty} \max_{1 \leq i \leq \tau^n_i} E \left| | N^n_i \right|_t = 0.$$  \hfill (3.33)

By Lemma 3.5 with $\gamma = \rho$ we have

$$E \left[ \left| | T^n \right|_{\rho,t} \right] \leq E \left[ \left| | T^n(0) \right|_{\rho} \right] e^{\rho t}.$$  

Recall (3.13), fix some $\epsilon > 0$ and let

$$C = \left( \limsup_{n \to \infty} E \left[ \left| | T^n(0) \right|_{\rho} \right] + \epsilon \right) e^{\rho t}.$$  

For all sufficiently large $n$ we have

$$E \left[ \left| | T^n \right|_{\rho,t} \right] \leq C.$$
This further implies that for all \(1 \leq i \leq i^*_n + 1\) we have

\[
\mathbb{E} [||T^n_i||_t] \leq C \rho^i \leq C \rho^*_{n+1}.
\]

Define

\[
\delta_n = \frac{1}{(\log \rho \eta_n)^2},
\]

and define the event

\[
A = \left\{ \exists \ i \leq i^*_n + 1 \text{ s.t. } ||T^n_i||_t \geq \frac{1}{\delta_n} C \rho^*_{n+1} \right\}.
\]

By Markov’s inequality and the union bound

\[
\mathbb{P}(A) \leq \delta_n (i^*_n + 1) = \frac{\alpha}{2 \log \rho \eta_n} + \frac{1}{(\log \rho \eta_n)^2}.
\] (3.34)

Observe

\[
\mathbb{E} [||N^n_i||_t] = \mathbb{E} [||N^n_i||_t 1 \{A\}] + \mathbb{E} [||N^n_i||_t 1 \{A^c\}].
\] (3.35)

We bound these two terms separately. We first consider the second term of (3.35). Recall (3.32) and observe

\[
||N^n_i||_t 1 \{A^c\} \leq \sup_{0 \leq u \leq \frac{\eta_n}{\eta_n}} \frac{\eta_n}{n} |D_i(u) - u|.
\] (3.36)

For \(0 \leq i \leq i^*_n\), \(A^c\) and (3.31) imply

\[
\tau^n_i \leq \frac{t}{\delta_n} C \rho^*_{n+1}
\]

\[
= t \rho C \eta_n^{\alpha/2}/\delta_n.
\]
Applying this for (3.36) we obtain

\[ \| N^n_i \| \mathbb{1}\{ A^c \} \leq \sup \frac{\eta_n}{n} | D_i(u) - u |, \]

where the supremum is over

\[ 0 \leq u \leq t \rho C n \delta_n^{-1} \eta_n^{-1}. \]

We define

\[ \nu_n \triangleq n \delta_n^{-1} \eta_n^{-1} = \frac{n}{\eta_n^{1-\alpha/2}} (\log \eta_n)^2. \]

Recall that \( \eta_n = O(\sqrt{n}) \), so

\[ \eta_n^{1-\alpha/2} = O(n^{1/2-\alpha/4}), \]

so \( \nu_n \to \infty \) as \( n \to \infty \). Thus we have

\[ \| N^n_i \| \mathbb{1}\{ A^c \} \leq \sup_{0 \leq u \leq t \rho C \nu_n} \frac{\eta_n}{n} | D_i(u) - u | \]

\[ = \frac{\eta_n \sqrt{\nu_n}}{n} \sup_{0 \leq u \leq t \rho C \nu_n} \frac{1}{\sqrt{\nu_n}} | D_i(u) - u |. \]

Let

\[ \gamma_n \triangleq \frac{\eta_n \sqrt{\nu_n}}{n} \mathbb{E} \sup_{0 \leq u \leq t \rho C \nu_n} \frac{1}{\sqrt{\nu_n}} | D_i(u) - u |. \]

Thus \( \mathbb{E} \| N^n_i \|_t \mathbb{1}\{ A^c \} \| \leq \gamma_n \) for \( 1 \leq i \leq i^*_n \). By the Functional Central Limit Theorem (FCLT), since \( \nu_n \to \infty \) we have

\[ \sup_{0 \leq u \leq t \rho C \nu_n} \frac{1}{\sqrt{\nu_n}} | D_i(u) - u | \Rightarrow \sup_{0 \leq u \leq t \rho C} | B(u) |, \]

97
where $B$ is a standard Brownian motion. Furthermore, observe that

$$\frac{\eta_n \sqrt{\nu_n}}{n} = n^{-1/2} \eta_n^{1/2+\alpha/4} \log \eta_n.$$ 

By assumption we have $\eta_n = O(\sqrt{n})$, so

$$\eta_n^{1/2+\alpha/4} = O\left(n^{1/4+\alpha/8}\right).$$

Recall that $\alpha < 1/2$ so $1/4 + \alpha/8 < 1/2$ which implies

$$\frac{\eta_n \sqrt{\nu_n}}{n} \to 0,$$

and thus $\gamma_n \to 0$.

We now consider the first term of (3.35). By the Cauchy-Schwarz inequality and (3.34), we have

$$\mathbb{E} \left[ ||N^n_i||_t 1\{A\} \right] \leq \sqrt{\mathbb{E} \left[ ||N^n_i||_t^2 \right]} \sqrt{\mathbb{P}(A)}.$$

By (3.31) and $T_i(t) \leq \eta_n$ for all $i \geq 1$, we have

$$||N^n_i||_t \leq \sup_{0 \leq u \leq nt} \frac{\eta_n}{\sqrt{n}} |D_i(u) - u| \quad \text{for all } i \geq 1.$$

Let

$$w_n \triangleq \frac{\eta_n}{\sqrt{n}} \mathbb{E} \left[ \sup_{0 \leq u \leq nt} \frac{1}{\sqrt{n}} |D_i(u) - u| \right], \quad (3.37)$$

and

$$z_n \triangleq \frac{\eta_n^2}{n} \mathbb{E} \left[ \left( \sup_{0 \leq u \leq nt} \frac{1}{\sqrt{n}} |D_i(u) - u| \right)^2 \right].$$

Then we have, for all $i \geq 1$

$$\mathbb{E} \left[ ||N^n_i||_t \right] \leq w_n \quad \text{and} \quad \mathbb{E} \left[ ||N^n_i||_t^2 \right] \leq z_n.$$
By the FCLT, we have

$$\sup_{0 \leq u \leq nt} \frac{1}{\sqrt{n}} |D_t(u) - u| \Rightarrow \sup_{0 \leq u \leq t} |B(u)|,$$

where $B$ is a standard Brownian motion. We now have

$$\mathbb{E} [||N^n_i||_t \ 1\{A\}] \leq \sqrt{\frac{\alpha}{2 \log \rho \eta_n}} + \frac{1}{(\log \rho \eta_n)^2}.$$

Because $\eta_n = O(\sqrt{n})$, as $n \to \infty$ we have

$$\frac{\eta_n^2}{n \log \rho \eta_n} \to 0,$$

and thus

$$\mathbb{E} [||N^n_i||_t \ 1\{A\}] \leq \sqrt{\frac{z_n \alpha}{2 \log \rho \eta_n}} + \frac{z_n}{(\log \rho \eta_n)^2} \to 0.$$

Returning to (3.35), we obtain

$$\mathbb{E} [||N^n_i||_t] = \mathbb{E} [||N^n_i||_t \ 1\{A\}] + \mathbb{E} [||N^n_i||_t \ 1\{A^c\}]$$

$$\leq \sqrt{\frac{z_n \alpha}{2 \log \rho \eta_n}} + \frac{z_n}{(\log \rho \eta_n)^2} + \gamma_n,$$

and thus

$$\lim_{n \to \infty} \max_{1 \leq i \leq n} \mathbb{E} [||N^n_i||_t] \leq \lim_{n \to \infty} \sqrt{\frac{z_n \alpha}{2 \log \rho \eta_n}} + \frac{z_n}{(\log \rho \eta_n)^2} + \gamma_n = 0,$$

which establishes the claim (3.33).
Now consider

\[
\mathbb{E} \left[ ||N^n||_{\rho,t} \right] = \sum_{i \geq 1} \rho^{-i} \mathbb{E} \left[ ||N^n_i||_t \right]
\]

\[
\leq \max_{1 \leq i \leq i^*_n} \mathbb{E} \left[ ||N^n_i||_t \right] \sum_{i \leq i^*_n} \rho^{-i} + \sum_{i > i^*_n} \rho^{-i} \mathbb{E} \left[ ||N^n_i||_t \right]
\]

\[
\leq \frac{\rho}{\rho - 1} \max_{1 \leq i \leq i^*_n} \mathbb{E} \left[ ||N^n_i||_t \right] + \sum_{i > i^*_n} \rho^{-i} w_n
\]

\[
= \frac{\rho}{\rho - 1} \max_{1 \leq i \leq i^*_n} \mathbb{E} \left[ ||N^n_i||_t \right] + \frac{\rho^{-i^*_n} w_n}{\rho - 1}.
\]

Recall the definition of \( w_n \) given by (3.37) and observe that

\[
\frac{\eta_n}{\sqrt{n} \rho^{-i^* / 2}} = \frac{\eta_n}{\sqrt{n}} \to 0,
\]

as \( n \to \infty \) implies the second term goes to zero. Recalling (3.33), we conclude

\[
\mathbb{E} \left[ ||N^n||_{\rho,t} \right] \to 0.
\]

This implies the convergence \( N^n \to 0 \) in \( D^{\infty,\rho} \) equipped with the product topology.

The argument to show \( \mathbb{E} \left[ ||M^n||_{\rho,t} \right] \to 0 \) is similar: we redefine \( \tau^n_i \) as

\[
\tau^n_i = \lambda_n d \int_0^t \left( T^n_i(s) - T^n_{i-1}(s) + g^n(T^n_i(s)) - g^n(T^n_{i-1}(s)) \right) ds.
\]

By Lemma 3.3, \( g^n \) restricted to \( \tilde{D}^{n+2} \) is Lipschitz continuous with constant \( 4^d \) and that \( \lambda_n \uparrow 1 \), so we have the bound

\[
\tau^n_i \leq td(1 + 4^d) ||T^n_i||_t + td4^d ||T^n_{i-1}||_t.
\]

This bound replaces (3.31) and the rest of the argument proceeds essentially identically to the \( N^n \) case.

We are now prepared to prove our main result:
Proof of Theorem 3.2. We first claim $\hat{T}^n \Rightarrow T$. We established in Section 3.4 that $\hat{T}^n = f(T^n(0), -M^n + N^n, \lambda_n, \eta_n)$ where $f$ is the function defined in Theorem 3.1. In Proposition 3.1 we established $-M^n + N^n \Rightarrow 0$. We prove $f(T^n(0), -M^n + N^n, \lambda_n, \eta_n) \Rightarrow f(T(0), 0, 1, \infty)$ directly. To do that, we choose a closed set $F \subset D^\infty.$ and show that

$$\limsup_n \mathbb{P} \left( f(T^n(0), -M^n + N^n, \lambda_n, \eta_n) \in F \right) \leq \mathbb{P} \left( f(T(0), 0, 1, \infty) \in F \right).$$

(3.38)

Our approach is to choose some large constant $K > 0$ and consider two cases. Let $A_{K,n}$ be the event $\left\{ \max \left( \|T^n(0)\|_{\rho}, \|-M^n + N^n\|_{\rho,t} \right) > K \, \forall t \geq 0 \right\}$ and observe that assumption (3.13) and Proposition 3.1 imply

$$\limsup_n \mathbb{P} (A_{K,n}) \overset{\triangle}{=} \epsilon_K \to 0 \quad \text{as} \quad K \to \infty.$$

Let $Y^n = (T^n(0), -M^n + N^n, \lambda_n, \eta_n)$ and $Y = (T(0), 0, 1, \infty)$. Then

$$\limsup_n \mathbb{P}(f(Y^n) \in F) = \limsup_n \left( \mathbb{P}(f(Y^n) \in F, A_{K,n}) + \mathbb{P}(f(Y^n) \in F, A_{K,n}^c) \right) \leq \epsilon_K + \limsup_n \mathbb{P} \left( f(Y^n) \in F, A_{K,n}^c \right).$$

Recall (3.26)-(3.27) and observe that $M^n(0) = N^n(0) = 0$. Thus $A_{K,n}^c$ and assumption (3.14) imply $Y^n \in Z^K_\alpha$ for sufficiently large $n$. For such $n$, we have

$$\mathbb{P} \left( f(Y^n) \in F, A_{K,n}^c \right) = \mathbb{P} \left( Y^n \in f^{-1}(F), A_{K,n}^c \right) = \mathbb{P} \left( Y^n \in f^{-1}(F) \cap Z^K_\alpha \right).$$

Because both $F$ and $Z^K_\alpha$ are closed and $f$ is continuous on $Z^K_\alpha$, $f^{-1}(F) \cap Z^K_\alpha$ is closed.
Thus the convergence $Y^n \Rightarrow x$ implies

$$\limsup_n \mathbb{P} \left( Y^n \in f^{-1}(F) \cap Z^n_K \right) \leq \mathbb{P} \left( Y \in f^{-1}(F) \right) \leq \mathbb{P} \left( f(Y) \in F \right).$$

Thus

$$\limsup_n \mathbb{P} \left( f(Y^n) \in F, A^n_{K,n} \right) \leq \mathbb{P} \left( f(Y) \in F \right),$$

and

$$\limsup_n \mathbb{P} \left( f(Y^n) \in F \right) \leq \epsilon_K + \mathbb{P} \left( f(Y) \in F \right).$$

Taking the limit $K \to \infty$ on both sides establishes (3.38). Thus

$$\hat{T}^n \Rightarrow f(T(0), 0, 1, \infty).$$

By definition, we have $f(T(0), 0, 1, \infty) = T$. Thus we conclude

$$\hat{T}^n \Rightarrow T.$$

By construction, $\hat{T}^n(t)$ and $T^n(t)$ are identical for $t \in [0, t^n_*]$. Thus it remains to show that for all $t \geq 0$, $\mathbb{P}(t^n_* \leq t) \to 0$ as $n \to \infty$. Let

$$p_n = \mathbb{P}(t^n_* \leq t) = \mathbb{P} \left( \exists \ i \leq i^n_* \ 	ext{s.t.} \ |T^n_i| \geq \eta_n^\alpha \right).$$
We have

$$
\mathbb{E} \left[ \| T^n \|_{\rho, t} \right] \geq \sum_{1 \leq i \leq i^*_n} \rho^{-i} \mathbb{E} \| T^n_i \|_t
$$

$$
\geq \rho^{-i^*_n} \sum_{1 \leq i \leq i^*_n} \mathbb{E} \| T^n_i \|_t
$$

$$
\geq \rho^{-i^*_n} \eta_{n}^\alpha p_n
$$

$$
= \rho^{-\frac{\alpha}{2} \log \eta_n \eta_{n}^\alpha} \eta_{n}^\alpha p_n
$$

$$
= \eta_{n}^{\alpha/2} p_n,
$$

(3.39)

We have $\eta_{n}^{\alpha/2} \to \infty$ as $n \to \infty$ By Lemma 3.5 with $\gamma = \rho$ and assumption 3.13, we have $\lim_{n \to \infty} \mathbb{E} \| T^n \|_{\rho, t} < \infty$, so (3.39) does not diverge to infinity, which implies $p_n \to 0$. This implies $T^n \Rightarrow T$, as desired.
Chapter 4

Open Questions and Future Directions

Theorem 2.2 proves that the behavior of the $M/M/n$-JSQ system in the Halfin-Whitt regime is best understood on the order of $O(\sqrt{n})$. In particular, the numbers of idle servers and waiting customers will both be $O(\sqrt{n})$. If long queues are initially present, they will empty in fixed time, and no additional long queues will be created. All remaining queues will have length one.

For the $M/M/n$-Sup(d) system, Theorem 3.2 similarly describes the dynamics of a small fraction of queues, as the fraction of queues of length at most $\log_d \frac{1}{1-\lambda_n}$ is on the order of $O(1 - \lambda_n)$. In this system, however, the behavior of the remaining queues is not easily characterized, and thus a more complete characterization of queues of length at least $\log_d \frac{1}{1-\lambda_n}$ remains a direction for future work. Perhaps most interesting would be to characterize the behavior of queues of length $\log_d \frac{1}{1-\lambda_n} + O(1)$. We conjecture that in steady state the fraction of queues with this length is non-vanishing and almost all queues are of this type. Our simulation results confirm this conjecture. We also expect that over finite time the process tracking such queues will converge to a diffusion process rather than a deterministic system. The methods used in Chapter 3 do not naturally translate to the “intermediate length queue” regime, partially because as $n \to \infty$ the system becomes infinite dimensional in two directions rather than one. It may also be of some interest to consider the behavior of queues of length
\[
\log_d \frac{1}{1-\lambda_n} + \omega_n \rightarrow \infty, \text{ though we expect these queues to disappear as } n \rightarrow \infty.
\]

Significant questions remain about the steady state behavior of both systems. For JSQ in particular, we do not characterize the distribution of the steady state of the limiting system or show that the steady state of the \( n \)-th system converges to the steady state of the limiting diffusion process (interchange of limits). For \( M/M/n-Sup(d) \), we conjecture that the fixed point in Theorem 3.3 is the limit of the steady state of the finite system. This would suggest that in steady state the fraction of servers with at least \( i \) jobs can be approximated by

\[
S^n_i(\infty) \approx 1 - \frac{\beta}{\eta_n} \frac{d^i - 1}{d - 1}.
\]

Establishing this interchange of limits would be a natural extension of the present result.

Another potential direction for future work is to generalize our results to supermarket models with a wide range of values for \( n, \lambda, \) and \( d \). Though it is not directly comparable to any of the results we establish here, the work of Brightwell and Luczak [5] suggests that it may be possible to establish general frameworks for analyzing the supermarket model as \( n, \lambda, \) and \( d \) are allowed to vary together (see especially [5, pg.7]). Our two cases both have \( \lambda \rightarrow 1 \) with a particular rate as \( n \rightarrow \infty \), while \( M/M/n-Sup(d) \) holds \( d \) constant and \( M/M/n-JSQ \) essentially sets \( d = n \). The drastically different qualitative behavior between these cases invites further investigation of intermediate regimes. A comprehensive characterization connecting more regimes would be invaluable.

Finally it is always of interest to analyze our system for general interarrival and, especially, general service times distribution. For both systems we conjecture that the qualitative behavior established in this paper in the transient domain and the conjectures above regarding the steady-state behavior and the interchange of steady-state limits remain true in this case as well.
Bibliography


