The Value of Flexible Customers via Join the Shortest of d Queues

Shuang Tao
School of Operations Research and Information Engineering
Cornell University
293 Rhodes Hall, Ithaca, NY 14853
st754@cornell.edu

Woo-Hyung Cho
School of Operations Research and Information Engineering
Cornell University
288 Rhodes Hall, Ithaca, NY 14853
wc563@cornell.edu

Jamol Pender *
School of Operations Research and Information Engineering
Cornell University
228 Rhodes Hall, Ithaca, NY 14853
jjp274@cornell.edu

April 21, 2020

Abstract

We propose and analyze a multi-server queueing model that captures a performance trade-off between customers valuing flexibility (join the shortest of d queues) or wanting dedicated service (join a specific queue). We are motivated by healthcare platforms like ZocDoc where patients may choose to see a dedicated physician or choose among the readily available physicians. In our stylized model, a fraction p of the customers are flexible and are willing to join the shortest of d queues and the remaining fraction 1-p will only join the queue of their choice. We prove both fluid and diffusion limits for the queueing model for the transient and steady state dynamics. In the fluid model, the steady state distribution satisfies a nonlinear recursion. Thus, we derive its closed form solution and show the sequence of fluid scaled queue length processes converges to a unique invariant state. Moreover, we prove that the diffusion scaled queue length

*Corresponding Author

process converges to an infinite dimensional OU process and its steady state limit coincides with the steady state of the OU process. Our analysis indicates that even a small number of flexible customers can have a large benefit on the system.

1 Introduction

Imagine a patient who has an undiagnosed health concern and would like to see a physician. Oftentimes a patient's Primary Care Physician (PCP) is the first medical practitioner that she will contact to address her concern. Suppose that, unfortunately, the PCP is fully booked for the next couple of weeks and the patient will have to wait for a prolonged period of time to see the PCP. This long of a wait is not an uncommon scenario in the United States' healthcare system. According to a recent survey, the average wait time for a patient to see a doctor for non-emergency issues can be as long as 66 days in a large city [2]. Thus, the patient, in need of seeing a medical professional, might choose to forgo a visit with their PCP to see the next available physician and resolve their medical concern sooner. To this end, a patient would choose to use a patient platform such as ZocDoc where they can choose among a large set of available doctors who specialize in their health issue.

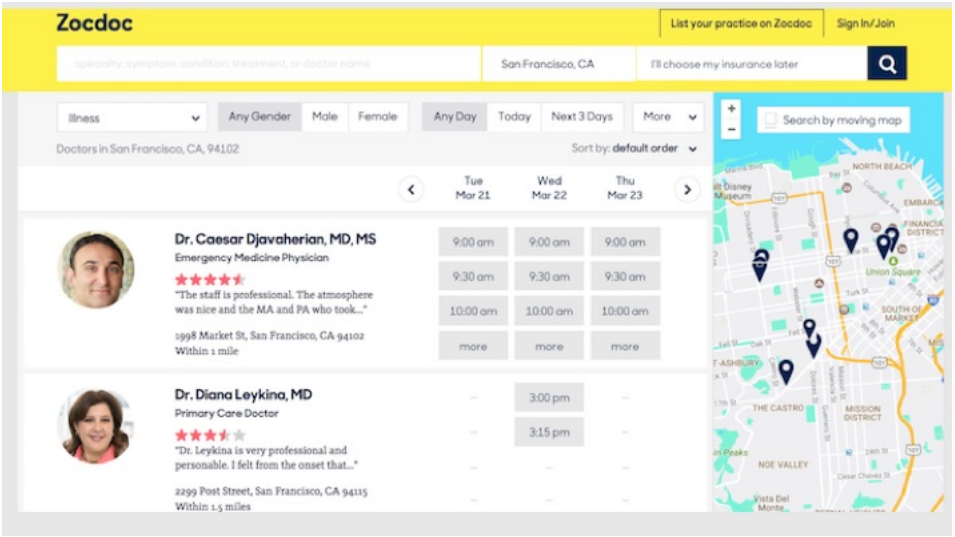

Figure 1: Example of ZocDoc platform where doctors are listed by earliest available appointment time and quality.

ZocDoc is a two-sided medical platform that allows patients to search and view the available appointment times of doctors online and make appointments instantly. ZocDoc's sync technology [1] allows patients to search based on the doctor's location, medical specialty, insurance coverage, and patient ratings. On the ZocDoc patient platform, there are typically 10 doctors listed per page. In Figure 1, we provide an example of the ZocDoc platform where doctors are listed by earliest available appointment time and perceived quality. Appointment booking is not just online, but also can be made via smartphone devices as well. Doctors can also choose to be listed on ZocDoc and allow the platform to access and integrate with their appointment calendars so that their updated calendars can be viewed by patients in

real-time. From a patient perspective, using a service like ZocDoc can help patients book appointments sooner. Earlier appointments typically result in earlier detection of illnesses, which can affect the final cost of healthcare expenses. Thus, we ask the question, what is the value of being able to see another physician on a patient platform like ZocDoc if one is flexible?

In this paper, we abstract the above scenario and model it as a multi-server queueing system under heavy traffic and partial load balancing. Similar types of queueing systems have been studied in the literature, see for example Aghajani et al. [3], Bramson et al. [5, 6], Dai et al. [10], Foley et al. [17], Foss and Stolyar [18], Graham [20, 21, 22, 23], He and Down [24], Lin and Raghavendra [28], Lu et al. [29], Mitzenmacher [30, 31], Mukherjee et al. [32], Tao and Pender [37], Tsitsiklis and Xu [38], Vvedenskaya et al. [40], Whitt [41].

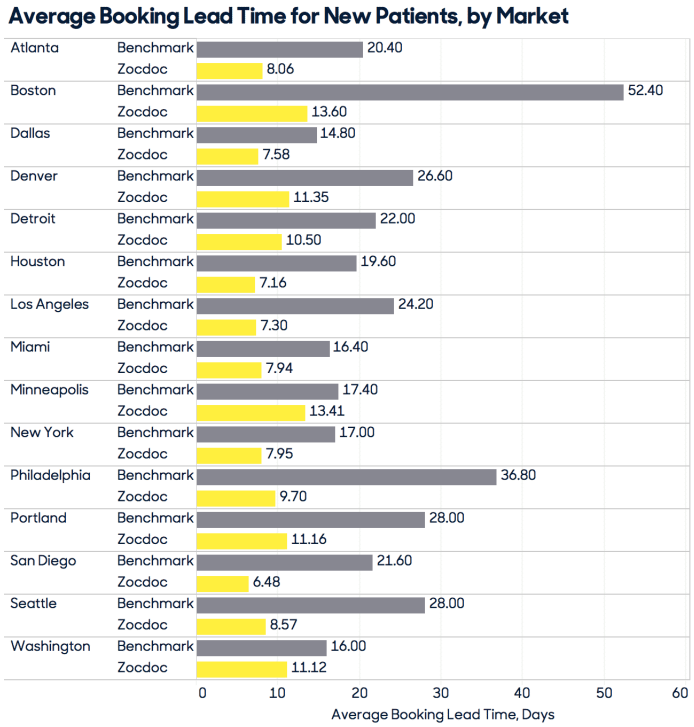

Figure 2: Success of ZocDoc in reducing wait lead times for patients.

However, we analyze the performance of the multi-server queue with the addition of a key feature: patient types, where patients of different types react differently to waiting to see their PCP. For example, some patients are quite particular about only being seen by their PCP for various reasons. These reasons might include familiarity, ease of communication, and accessibility of location. On the other hand, there also exist *flexible* patients who are willing to see another physician other than their PCP if they can have access to a medical professional within a shorter time frame. In fact, these flexible patients might be willing to call several physician's offices, observe waiting times for each physician, and finally *join the queue* by scheduling an appointment with the physician that offers the shortest wait time among those contacted or listed on the ZocDoc platform. In the context of ZocDoc, this flexibility decreases the booking lead time significantly for those who are willing to be

flexible [1]. In Figure 2, we observe that the introduction of ZocDoc as a patient platform has reduced the wait lead time significantly by offering substitute doctors that are willing to see the patient in a earlier time frame. By doing so, flexible patients will acquire queue length information that dedicated patients do not have access to. The question we aim to address in this paper is, how does the overall system perform if only a fraction $p \in [0, 1]$ of the patients are flexible and would be willing to use a platform like ZocDoc?

The model that we consider is highly stylized. We consider a system of N physicians and assume that patients who arrive to the system are one of two types: flexible or dedicated. We fix a flexibility parameter $p \in [0, 1]$, which denotes the probability with which each arriving patient is flexible. We further assume that each patient type has a different policy for joining a physician's queue. The dedicated patients join their designated PCP's queue regardless of queue length i.e. they join one of the N queues uniformly at random. In other words, these patients either have **no information** and do not use a platform like ZocDoc to search for earlier appointments. Hence, they are in some sense loyal to their PCP regardless of the wait they might experience. Flexible patients, on the other hand, are willing to see any physician that reduces their waiting time and are considered impatient. In our model, flexible patients choose d physicians, independently and uniformly at random, and observe the queue lengths of each. Flexible patients subsequently respond to this newly obtained information by joining the shortest queue among the d physician queues sampled. In some of the current literature, the parameter d scales with the number of servers N, see for example Dieker and Suk [13]. However, we assume that d is a fixed constant since the ZocDoc patient platform displays 10 physicians at once on one page and therefore, the value $d = 10$ is a reasonable value for the purposes of our work.

Our goal is to study the performance of the system for varying degrees of "flexibility" and "power of choices", as expressed by parameters p and d, respectively. In doing so, we use a fluid approximation where the queue length dynamics are approximated with a deterministic fluid model as $N \to \infty$ and the fluid model behaves according to an infinite dimensional system of non-linear ordinary differential equations. We are especially interested in studying and deriving an upper bound for the average queue length in the system, which, as we will see, also has some interesting interpretations.

In addition to the healthcare motivation presented by the ZocDoc platform, one can also imagine a supermarket where customers join lines independently without any knowledge of the number of customers at each cashier. Our model is equivalent to having a proportion of informed customers who have the ability to look at d queues and join the shortest among those queues. Thus, our goal is to understand the value that a few "informed" customers can have on the system. We will show in the sequel that even when the proportion of flexible patients is small, these flexible patients can have a large impact on the overall system performance.

1.1 Related Work

There has been a lot of activity in the recent years of researchers analyzing a number of variants of the join the shortest queue model, see for example recent work by Banerjee et al. [4], Braverman [7], Eschenfeldt and Gamarnik [15], Mukherjee et al. [33]. Despite, the large amount of activity in this area, there are relatively few papers that explore the impact of

flexibility or information on the underlying system. In this work, we are inspired by the work by Tsitsiklis and Xu [38] where they explicitly study the trade-off between centralized and distributed processing. In their work, they consider an N-station system where their system designer is given a total amount N of divisible computing resources. Moreover, the system designer in their work can allocate resources to local and central servers. More specifically, for some fraction $p \in (0, 1)$, local servers process tasks at a maximum rate of $1 - p$ tasks per second, while the centralized server, at rate of pN tasks per second. Our work is different from theirs in two main ways. First, we consider a different model where we are joining the shortest of d queues. Second, we do not assume a centralized server processes tasks. In our setting, **flexibility** can be viewed as information each arrival has about the system. Some customers have some partial information about the system and the others do not have any information about the system and join uniformly at random. We also differ from Tsitsiklis and Xu [38] since we also analyze the diffusion scaled system. By studying the diffusion scaled process, we can gain important insights on how the flexibility impacts the fluctuations or variance of the queueing system. This is also helpful in building confidence intervals around the fluid limit.

1.2 Main Contributions of Our Work

The contributions that we make in this work are:

- We develop a new stochastic queueing model that incorporates the structure of dedicated and flexible customers. We explore the trade-off between these types of customers through the parameters p and d, which represent flexibility and the amount of partial information about the system.

- We prove fluid and diffusion limit theorems for the queueing process, thus showing that the fluid limit is an infinite dimensional system of non-linear odes and that the diffusion limit is an infinite dimensional Ornstein-Uhlenbeck process.

- We prove an interchange of limits for the fluid and diffusion scaled processes, thereby showing that the steady state fluid and diffusion limits are good approximations for the original fluid and diffusion scaled processes. In fact, we derive a closed form expression for the steady state distribution using a non-linear recursion. This recursion also allows us to derive new upper and lower bounds on the first and second moments of the queue length in steady state, which converge to each other as $p \to 0$ or $p \to 1$.

- From a mathematical perspective, we derive a new method for proving the global stability of the steady state fluid limit by using a comparison approach. Our approach exploits the fact that if the integral of the difference of two solutions are bounded, then the two solutions converge to the same point. We also derive new infinite horizon bounds for the diffusion scaled process, which are important ingredients for establishing tightness for steady state diffusion limits. The infinite horizon bounds are in general difficult to prove because they must be proved in the appropriate functional space when the sub-generator of an associated birth-death process is not self-adjoint. Moreover, proving these infinite horizon bounds is difficult in our model because the self-adjoint property of the sub-generator depends on the flexibility parameter p.

1.3 Organization of the Paper

The remainder of the paper is as follows. In Section 2, we describe the stochastic model of our paper. In Section 3, we present a fluid model for the tail distribution of the queue length. We prove both a transient and a steady state fluid limit for our stochastic model. The transient fluid limit is proved using martingale techniques and the steady state fluid limit is proved using a new comparison approach. We also prove an interchange of limits result, which shows in a rigorous sense that the steady state limit can be used as an approximation for our stochastic model. In Section 4, we present a diffusion model for the tail distribution of the queue length and prove a transient diffusion limit, a steady state diffusion limit and an interchange of limits for the stochastic model. In Section 5, we prove that the steady state fluid limit can be written in closed form using a nonlinear recursion. We also prove tight upper and lower bounds on the first and second moments of the queue length. We also demonstrate through numerical examples that small values of p can have a large impact on the behavior the system. Finally, in Section 6, we conclude the paper and we move most of the proofs to the Appendix in Section 7.

1.4 Notation

Below in Table 1, we provide a list of the notations that we will use throughout the rest of the paper.

Table 1: Notation

N	# of physicians
λ	Arrival rate of patients
p	Fraction of flexible patients
d	# of physicians flexible patients sample
$Q_i^N(t)$	Number of patients at physician i at time t
$S_i^N(t)$	Fraction of queues with at least i patients at time t
$s_i(t)$	The fluid limit of process $S_i^N(t)$
s^I	The steady state of fluid limit $s(t)$
$D^N(t)$	The fluctuation of $S^N(t)$ around its fluid limit $s(t)$
$D(t)$	The diffusion limit of process $D^N(t)$
ℓ_1	The space of sequences whose series is absolutely convergent
ℓ_2	The space of square-summable sequences
$\mathcal{S}$	$\{s \in [0,1]^{\mathbb{Z}_+} : 1 \geq s_0 \geq s_1 \geq \cdots \geq 0, \sum_{i=0}^{\infty} s_i < \infty\}$

1.5 Preliminaries of Weak Convergence

In this paper, we assume that all random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Moreover, for all positive integers k, we let $\mathcal{D}([0,\infty), \mathcal{S})$ be the space of right continuous functions with left limits (RCLL) in $\mathcal{S}$ that have a time domain in $[0,\infty)$. As is usual, we endow the space $\mathcal{D}([0,\infty), \mathcal{S})$ with the usual Skorokhod J_1 topology, and let M be defined as the Borel σ-algebra associated with the J_1 topology. We also assume

that all stochastic processes are measurable functions from our common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into $(\mathcal{D}([0,\infty), \mathcal{S}), M)$. Thus, if $\{\zeta\}_{n=1}^{\infty}$ is a sequence of stochastic processes, then the notation $\zeta^n \to \zeta$ implies that the probability measures that are induced by the ζ^n's on the space $(\mathcal{D}([0,\infty), \mathcal{S}), M)$ converge weakly to the probability measure on the space $(\mathcal{D}([0,\infty), \mathcal{S}), M)$ induced by ζ. For any $x \in (\mathcal{D}([0,\infty), \mathcal{S}), M)$ and any $T > 0$, we define

$$||x||_{\ell_2} \equiv \sum_{i=0}^{\infty} x_i^2 \tag{1.1}$$

and note that ζ^n converges almost surely to a continuous limit process ζ in the J_1 topology if and only if

$$||\zeta^n - \zeta||_{\ell_2} \to 0 \quad a.s. \tag{1.2}$$

2 The Stochastic Queueing Model

In this section, we present a stochastic queueing model that has N physicians. Each physician operates a single server queue of scheduled patients who are seen in a first in first out manner. We denote the queue length for physician n at time t with $Q_n(t)$ where $n \in \{1, 2, \cdots, N\}$ and $t \geq 0$. Each physician processes the work of their current patients at rate 1 if there are patients in their queue.

For the patients, we assume there are two types of patients: *dedicated* and *flexible*. The two types of patients are split into according to our flexibility parameter p. A patient is flexible with fixed probability $p \in [0, 1]$. We assume that flexible patients are willing to sample d physician queues, independently and uniformly at random, and join the shortest-of-d queues at their time of arrival. This is an abstraction of patients choosing among the available physicians on the ZocDoc platform. Dedicated patients, on the other hand, are only willing to see their designated PCP and are not flexible. Thus, assuming equal popularity among all physicians, this is equivalent to saying that they join any queue at random. Finally, we assume that once a patient joins a queue, the patient is completely locked in and cannot switch to another queue.

Each of the N physicians has a stream of dedicated patients arriving according to independent Poisson processes with a common rate $\lambda(1-p)$, where $\lambda \in [0, 1]$. Thus, the total arrival rate of dedicated patients to the system is $\lambda(1-p)N$. In addition, the overall system also has a stream of flexible patients arriving according to an independent Poisson process with rate λpN.

Once patients are routed to the appropriate physician queue (dedicated patients to their PCP queues and flexible patients to the shortest-of-d physician queues), each physician queue operates as an $M/M/1$ queue. The queue length vector at time t, $(Q_1(t), Q_2(t), \cdots, Q_N(t))$, is a Markov process. In addition, the system is fully symmetric and exchangeable in that the arrival of dedicated patients and patient services are independent and identical, and the arrival of flexible patients depends solely on the length of the physician queues, and not on the specific identity of physicians. Thus, we can use a Markov process $\{S_i^N(t)\}_{i=0}^{\infty}$ to describe the evolution of the system, where we defined $S_i^N(t)$ as

$$S_i^N(t) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{Q_n(t) \geq i\}. \tag{2.3}$$

Here $S_i^N(t)$ represents the fraction of queues with at least i patients. By definition, $S_0^N(t) = 1$ for all values of N and $t \geq 0$. Furthermore, $S_i^N(t)$ is a non-increasing process in the variable i, meaning that

$$1 \geq S_i^N(t) \geq S_{i+1}^N(t) \geq 0$$

for all values of i, N and $t \geq 0$. We define the infinite dimensional vector of this queueing process as $S^N(t) = (S_0^N(t), S_1^N(t), \cdots, S_n^N(t), \cdots S_\infty^N(t))$. Our goal is to study the process $S^N(t)$ in two scenarios. The first is in the transient setting where we let $N \to \infty$ and the second is in the steady state setting where we let both $N \to \infty$ and $t \to \infty$.

3 Fluid Model

Here we summarize the results in this section, which are related to the fluid model of the queueing process $S^N(t)$. For our first result, Theorem 3.1, we prove a functional law of large numbers (LLN) in the transient case for process $S^N(t)$ to its fluid limit $s(t)$. For our second result, Theorem 3.8), we prove an interchange of limits results for the stochastic process model. We use a compactness-uniqueness approach, which show that the limiting point s^I of the fluid limit $s(t)$ is also the limit of the invariant measure $S^N(\infty)$ of $S^N(t)$.

3.1 Transient Analysis of the Fluid Limit

We start with the functional law of large numbers in the transient case for the fluid limit.

Theorem 3.1 (Functional Law of Large Numbers)**.** *Assume that $(S^N(0))_{N\geq d}$ converges in distribution to $s(0)$ in $\mathcal{S}$. Then, $(S^N(t))_{N\geq d}$ converges in probability to the unique solution $s = (s(t))_{t\geq 0}$ i.e. on any compact time interval $t_0 > 0$ and $\epsilon > 0$, we have*

$$\lim_{N\to\infty} \mathbb{P}\left(\sup_{t\leq t_0} \|S^N(t) - s(t)\|_{\ell_2} > \epsilon\right) = 0. \tag{3.4}$$

Moreover, $s(t)$ has initial condition $s(0)$ and is the solution to the following infinite dimensional system of differential equations

$$\frac{ds_i}{dt} = \underbrace{\lambda(1-p)(s_{i-1} - s_i)}_{\text{arrival of dedicated patients}} + \underbrace{\lambda p\left(s_{i-1}^d - s_i^d\right)}_{\text{arrival of flexible patients}} - \underbrace{(s_i - s_{i+1})}_{\text{departure of patients}} \qquad i \geq 1. \tag{3.5}$$

Proof. We prove this result using Doob's inequality for martingales and Gronwall's lemma. We use Proposition 3.2, and Lemma 4.7 in the proof, and they are stated after the proof of Theorem 3.1. To give readers a high-level understanding of the proof idea, we list the essential steps and related theorem numbers below.

1. We decompose queueing process $S^N(t)$ into three parts. The first is the initial condition $S^N(0)$, the second is a martingale $M^N(t)$ term and the final term is an integral of the drift term $\int_0^t F^N(S^N(u))du$. (Equation 3.7)

2. We bound the difference between $S^N(t)$ and its fluid limit $s(t)$ on any finite interval $[0,T]$ by difference in their initial conditions $\|S^N(0)-s(0)\|$, the supremum of martingale $\sup_{u\le T}\|M^N(t)\|$, the difference in drift function and its limit $\int_0^T \|F^N(S^N(u)) - F(S^N(u))\|du$, and finally the difference in limiting drift function evaluated at $S^N(t)$ and $s(t)$ i.e. $\int_0^T \|F(S^N(u)) - F(s(u))\|du$. (Inequality 3.10)

3. We show Lipschitz property of limiting drift function F. (Proposition 3.2)

4. We apply Gronwall's lemma to the difference. (Inequality 3.12)

5. We apply Doob's L_2 martingale inequality to $M^N(t)$ and bounds on quadratic variation. (Inequality 3.19, Lemma 4.7)

6. We prove existence and uniqueness of the fluid limit $s(t)$. (Proposition 3.3)

To start with the proof, we introduce the falling factorial notation $(x)_k = x(x-1)\cdots(x-k+1)$ for $x\in\mathbb{R}$, and define the following mappings for s in c_0 by

$$\begin{aligned}
F_+(s)(i) &= \lambda(1-p)(s_{i-1}-s_i) + \lambda p(s_{i-1}^d - s_i^d),\\
F_-(s)(i) &= (s_i - s_{i+1}),\\
F_+^N(s)(i) &= \lambda(1-p)(s_{i-1}-s_i) + \lambda p\frac{(Ns_{i-1})_d - (Ns_i)_d}{(N)_d}, \quad i\ge 1,\\
F^N(s) &= F_+^N(s) - F_-(s),\\
F(s) &= F_+s) - F_-(s).
\end{aligned} \tag{3.6}$$

Then, the nonlinear differential equation can be written as $\dot{s} = F(s)$, and it is easy to show that $S^N(t)$ is a Markov process, that when in state s, has jump in the i^{th} coordinate of size $+1/N$ with rate $NF_+^N(s)(i)$ and size $-1/N$ with rate $NF_-(s)(i)$, for all $i\ge 1$. Since $S^N(t)$ is a semi-martingale, we have the following decomposition of $S^N(t)$,

$$S^N(t) = \underbrace{S^N(0)}_{\text{initial condition}} + \underbrace{M^N(t)}_{\text{martingale}} + \int_0^t \underbrace{F^N(S^N(u))}_{\text{drift term}}\,du, \tag{3.7}$$

where $S^N(0)$ is the initial condition and $M^N(t)$ is a independent family of martingales. Moreover, $\int_0^t F^N(S^N(u))du$ is the integral of the drift term where the drift term is given by $F^N:\mathcal{S}\to\mathbb{R}^{\mathbb{Z}_+}$ or

$$\begin{aligned}
F^N(s)(k) &= \sum_{x\ne s}(x-s)Q^N(s,x)(k)\\
&= \lambda(1-p)(s_{k-1}-s_k) + \lambda p\frac{(Ns_{k-1})_d - (Ns_k)_d}{(N)_d} - (s_k - s_{k+1}),
\end{aligned}$$

where $Q^N(s,x)(k)$ represents the transition rate from state s to x on the k^{th} coordinate. Now we want to compare $S^N(t)$ with its fluid limit $s(t)$ defined by

$$s(t) = s(0) + \int_0^t F(s(u))du. \tag{3.8}$$

If we let $\|\cdot\|$ denote the ℓ_2 norm in $\mathbb{R}^{\mathbb{Z}_+}$, then

$$\begin{aligned}
\left\|S^N(t) - s(t)\right\| &= \left\|S^N(0) + M^N(t) + \int_0^t F^N(S^N(u))du - s(0) - \int_0^t F(s(u))du\right\| \\
&= \left\|S^N(0) - s(0) + M^N(t) + \int_0^t \left(F^N(S^N(u)) - F(S^N(u))\right)du \right. \\
&\quad \left. + \int_0^t (F(S^N(u)) - F(s(u)))du\right\|. \qquad (3.9)
\end{aligned}$$

Now we define the random function $f^N(t) = \sup_{u\le t}\left\|S^N(u) - s(u)\right\|$, and by the triangle inequality we have

$$\begin{aligned}
f^N(t) &\le \|S^N(0) - s(0)\| + \sup_{u\le t}\|M^N(u)\| + \int_0^t \|F^N(S^N(u)) - F(S^N(u))\|du \\
&\quad + \int_0^t \|F(S^N(u)) - F(s(u))\|du. \qquad (3.10)
\end{aligned}$$

By Proposition 3.2, $F(s)$ is Lipschitz with respect to ℓ_2 norm. Let L be the Lipschitz constant of $F(s)$, then

$$\begin{aligned}
f^N(t) &\le \|S^N(0) - s(0)\| + \sup_{u\le t}\|M^N(u)\| + \int_0^t \|F^N(S^N(u)) - F(S^N(u))\|du \\
&\quad + \int_0^t \|F(S^N(u)) - F(s(u))\|du \\
&\le \|S^N(0) - s(0)\| + \sup_{u\le t}\|M^N(u)\| + \int_0^t \|F^N(S^N(u)) - F(S^N(u))\|du \\
&\quad + L\int_0^t \|S^N(u) - s(u)\|du \\
&\le \|S^N(0) - s(0)\| + \sup_{u\le t}\|M^N(u)\| + \int_0^t \|F^N(S^N(u)) - F(S^N(u))\|du \\
&\quad + L\int_0^t f^N(u)du. \qquad (3.11)
\end{aligned}$$

By Gronwall's lemma,

$$f^N(t) \le \left(\|S^N(0) - s(0)\| + \sup_{u\le t}\|M^N(u)\| + \int_0^t \|F^N(S^N(u)) - F(S^N(u))\|du\right)e^{Lt}. \tag{3.12}$$

Now we proceed to bound $f^N(t)$ term by term. To this end, we define function $\alpha : \mathcal{S} \to \mathbb{R}^{\mathbb{Z}_+}$ as

$$\begin{aligned}\alpha(s)(k) &= \sum_{x \neq s} \|x - s\|^2 Q^N(s,x)(k) \\ &= \frac{1}{N}\left[F_+^N(s)(k) + F_-(s)(k)\right] \\ &= \frac{1}{N}\left[\lambda(1-p)(s_{k-1} - s_k) + \lambda p \frac{(Ns_{k-1})_d - (Ns_k)_d}{(N)_d} + (s_k - s_{k+1})\right]. \quad (3.13)\end{aligned}$$

By Lemma 4.7, we have that $\|\alpha(s)\|_{\ell_2} = \frac{1}{N} O(\|s\|_{\ell_2})$. Thus, there exist a constant $C > 0$ such that $\|\alpha(s)\|_{\ell_2} \leq \frac{C}{N}$ for any s. Now consider the following four sets

$$\begin{aligned}\Omega_0 &= \{\|S^N(0) - s(0)\| \leq \delta\}, \quad (3.14)\\ \Omega_1 &= \left\{\int_0^{t_0} \|F^N(S^N(t)) - F(S^N(t))\| dt \leq \delta\right\}, \quad (3.15)\\ \Omega_2 &= \left\{\int_0^{t_0} \|\alpha(S^N(t))\| dt \leq A(N) t_0\right\}, \quad (3.16)\\ \Omega_3 &= \left\{\sup_{t \leq t_0} \|M^N(t)\| \leq \delta\right\}, \quad (3.17)\end{aligned}$$

where $\delta = \epsilon e^{-Lt_0}/3$. Here the set Ω_0 is for bounding the initial condition, the set Ω_1 is for bounding the drift term F^N and the limit of the drift term F, and the sets Ω_2, Ω_3 are for bounding the martingale $M^N(t)$. Therefore, on the event $\Omega_0 \cap \Omega_1 \cap \Omega_3$,

$$f^N(t_0) \leq 3\delta e^{Lt_0} = \epsilon. \quad (3.18)$$

Consider the stopping time

$$T = t_0 \wedge \inf\left\{t \geq 0 : \int_0^t \alpha(S^N(u)) du > A(N) t_0\right\},$$

by Doob's ℓ_2 martingale inequality,

$$\mathbb{E}\left(\sup_{t \leq T} \|M^N(t)\|^2\right) \leq 4\mathbb{E}\|M^N(T)\|^2 = 4\int_0^T \|\alpha(S^N(u))\| du. \quad (3.19)$$

On Ω_2, we have $T = t_0$, so $\Omega_2 \cap \Omega_3^c \subset \{\sup_{t \leq T} \|M^N(t)\| > \delta\}$. By Chebyshev's inequality,

$$\mathbb{P}(\Omega_2 \cap \Omega_3^c) \leq \mathbb{P}\left(\sup_{t \leq T} \|M^N(t)\| > \delta\right) \leq \frac{\mathbb{E}\left(\sup_{t \leq T} \|M^N(t)\|^2\right)}{\delta^2} \leq 4A(N)t_0/\delta^2. \quad (3.20)$$

Thus, by Equation (3.18), we have the following result,

$$\begin{aligned}\mathbb{P}\left(\sup_{t \leq t_0} \|S^N(t) - s(t)\| > \epsilon\right) &\leq \mathbb{P}(\Omega_0^c \cup \Omega_1^c \cup \Omega_3^c) \\ &\leq \mathbb{P}(\Omega_2 \cap \Omega_3^c) + \mathbb{P}(\Omega_0^c \cup \Omega_1^c \cup \Omega_2^c) \quad (3.21)\\ &\leq 4A(N)t_0/\delta^2 + \mathbb{P}(\Omega_0^c \cup \Omega_1^c \cup \Omega_2^c) \\ &= 36A(N)t_0 e^{2Lt_0}/\epsilon^2 + \mathbb{P}(\Omega_0^c \cup \Omega_1^c \cup \Omega_2^c).\end{aligned}$$

Let $A(N) = \frac{C}{N}$, then $\Omega_2^c = \emptyset$. And since $S^N(0) \xrightarrow{p} s(0)$, $\lim_{N\to\infty} \mathbb{P}(\Omega_2^c) = 0$. Therefore we have

$$\lim_{N\to\infty} \mathbb{P}\left(\sup_{t\le t_0} \|S^N(t) - s(t)\| > \epsilon\right) = \lim_{N\to\infty} \mathbb{P}(\Omega_1^c).$$

By Lemma 4.7, $\lim_{N\to\infty} \mathbb{P}(\Omega_1^c) = 0$. Thus, we proved the final result

$$\lim_{N\to\infty} \mathbb{P}\left(\sup_{t\le t_0} \|S^N(t) - s(t)\| > \epsilon\right) = 0.$$

□

Proposition 3.2 (Lipschitz bound on drift functions)**.** *The mappings F, F_+, F_- are Lipschitz with respect to the ℓ_2 norm.*

Proof. See Appendix for details of the proof. □

Proposition 3.3 (Existence and Uniqueness of the fluid limit)**.** *There exists a unique solution $(s(t))_{t\ge 0} \in \mathcal{S}$ to the differential equation (3.5) and $s(t)$ is continuous in t.*

Proof. This is a direct application from the Lipschitz property of F in Proposition 3.2 and Gronwall's lemma. □

3.2 Steady State Analysis of Fluid Limit

In addition to understanding the transient behavior of the fluid model, it is important to understand the steady state behavior as well. In this section, we outline the steady state analysis of the stochastic queueing model. To begin, we denote the steady state of the queueing model as s^I. Then, s^I satisfies the following equation,

$$\lambda(1-p)(s_{i-1}^I - s_i^I) + \lambda p\left((s_{i-1}^I)^d - (s_i^I)^d\right) - (s_i^I - s_{i+1}^I) = 0, \quad i \ge 1 \tag{3.22}$$

Theorem 3.4. *The steady state of the queueing model s^I has a unique solution given by the following recursion*

$$\begin{aligned} s_0^I &= 1, \\ s_1^I &= \lambda, \\ s_i^I &= \lambda(1-p)s_{i-1}^I + \lambda p(s_{i-1}^I)^d \quad \textit{for all } i \ge 2. \end{aligned}$$

Proof. We prove the result by induction. For $i = 1$,

$$\begin{aligned} s_1^I &= \sum_{i=1}^{\infty}(s_i^I - s_{i+1}^I) \\ &= \sum_{i=1}^{\infty}\left[\lambda(1-p)(s_{i-1}^I - s_i^I) + \lambda p\left((s_{i-1}^I)^d - (s_i^I)^d\right)\right] \\ &= \lambda(1-p)s_0^I + \lambda p(s_0^I)^d \\ &= \lambda. \end{aligned}$$

Now for $i \leq k$, we assume that

$$s_i^I = \lambda(1-p)s_{i-1}^I + \lambda p(s_{i-1}^I)^d.$$

Then, for $i = k+1$,

$$\begin{aligned} s_{k+1}^I &= s_k^I - \lambda(1-p)(s_{k-1}^I - s_k^I) - \lambda p\left((s_{k-1}^I)^d - (s_k^I)^d\right) \\ &= \lambda(1-p)s_k^I + \lambda p(s_k^I)^d. \end{aligned}$$

□

Remark: Note that the existence and uniqueness of the equilibrium point s^I is obtained from the fact that s_i^I is completely determined by s_{i-1}^I and we have the initial condition $s_0^I = 1$ holds.

3.3 Interchanging Limits of Fluid Limit

In this section, we prove an interchange of limits result for the fluid model, i.e. the limiting point s^I of the fluid limit $s(t)$ is also the limit of the invariant measure $S^N(\infty)$ of $S^N(t)$. A visual interpretation of the interchange of limits result corresponds to showing that the following diagram commutes.

$$\begin{array}{ccc} S^N(t) & \xrightarrow{N\to\infty} & s(t) \\ {\scriptstyle t\to\infty}\downarrow & & \downarrow{\scriptstyle t\to\infty} \\ S^N(\infty) & \xrightarrow[N\to\infty]{} & s^I \end{array}$$

We have already proved in Section 3 that $S^N(t) \xrightarrow{p} s(t)$ and the existence and uniqueness of s^I. Now we will show the other two directions of the diagram, which are the existence of invariant measure $S^N(\infty)$ for each $N \geq 1$, and the convergence of the invariant measure $S^N(\infty)$ to s^I. Our method of proof is a modification of the compactness-uniqueness method pioneered by Graham [20]. We can decompose the compactness uniqueness method into three essential steps.

1. Show that the fluid limit (Equation (3.5)) has a globally attractive stable point s^I. (Lemma 3.5, Theorem 3.6)

2. Show that there exists an invariant measure $S^N(\infty)$ for S^N for each $N \geq 1$. (Proposition 3.7, Theorem 3.8 (1))

3. Show that these invariant measures $(S^N(\infty))_{N\geq 1}$ are tight in $\mathcal{S}$. (Theorem 3.8 (2))

In order to prove that the fluid limit has a globally attractive stable point, we will use a comparison result for finiite dimensional ordinary differential equations. This result is outlined below.

Lemma 3.5 (Comparison Result)**.** *Let u and v be two solutions for Equation (3.5) such that $u(0) \leq v(0)$. Then $u(t) \leq v(t)$ for all $t \geq 0$.*

Proof. We first consider the finite dimensional case. For any fixed constant $K \in \mathbb{N}$, we assume WLOG that $u_k(0) < v_k(0), k = 1, \cdots, K$, and that $u_{K+1}(t) < v_{K+1}(t)$ for all $t \geq 0$. We aim to show that $u_k(t) < v_k(t)$ for all $t \geq 0$ and $k = 1, \cdots, K$.

Assume that $u(t) < v(t)$ for $t \in [0, t_0)$ but $u_i(t_0) = v_i(t_0)$ for some $i \in \{1, \cdots, K\}$. Then we know that $u_j(t_0) \leq v_j(t_0)$ for all $j \in \{1, \cdots, K\}$. Now from the fluid limit equation (3.5) we have that

$$\begin{aligned} \dot{u}_i(t_0) &= \lambda(1-p)(u_{i-1}(t_0) - u_i(t_0)) + \lambda p(u^d_{i-1}(t_0) - u^d_i(t_0)) - (u_i(t_0) - u_{i+1}(t_0)) \\ &\leq \lambda(1-p)(v_{i-1}(t_0) - v_i(t_0)) + \lambda p(v^d_{i-1}(t_0) - v^d_i(t_0)) - (v_i(t_0) - v_{i+1}(t_0)) \\ &= \dot{v}_i(t_0), \end{aligned} \tag{3.23}$$

suggesting that $u_i(t) \leq v_i(t)$ for $t \geq t_0$.

Now for any $s(0) \in \mathcal{S}$, there exists a unique solution $s(t) \in \mathcal{S}$ for (3.5). We will show that the solution $s(t)$ can be obtained as the limit of solutions $\{s^K(t)\}_{K=1}^{\infty}$ to (3.5) with $s_{K+1}(t) = 0$.

Denote $s^K(t)$ as the solution to (3.5) with $s^K_{K+1} = 0$. Then we have $s^{K+1}_{K+1}(t) \geq s^K_{K+1}(t) = 0$. By the previous argument, we have that for fixed t and $i \leq K$, $s^{K+1}_i(t) \geq s^K_i(t)$. Then there exists the limit $\lim_{K\to\infty} s^K_i(t) = s_i(t)$ for each i and $s(t) = \{s_i(t)\}_{i=0}^{\infty} \in \bar{\mathcal{S}}$. Notice that $s_i(t)$ satisfies the fluid limit equation (3.5). It follows by uniqueness of the solution that the limit $\lim_{K\to\infty} s^K(t) = s(t)$ is the solution to fluid limit Equation (3.5). Finally, combining the two previous arguments, we conclude the comparison theorem for infinite dimensional case. □

Theorem 3.6 (Global Stability of Fluid Limit)**.** *The fluid limit equation (3.5) has globally attractive stable point s^I. That is, starting from any initial condition $s(0) \in \mathcal{S}$,*

$$\lim_{t\to\infty} s(t) = s^I$$

Proof. The proof is given in the Appendix. □

Now we will construct a coupling which compares the behavior of the system $S^N(t)$ when $d = 1$ vs. $d > 1$. When $d = 1$, the fluid limit equation becomes

$$\dot{s}_i(t) = \lambda(s_{i-1}(t) - s_i(t)) - (s_i(t) - s_{i+1}(t)),$$

which is a system of N i.i.d $M/M/1$ queues. And we know that if and only if $\lambda < 1$, when such system is positive recurrent, with a geometric stationary distribution being

$$s^I_k = \lambda^k, \quad k \in \mathbb{N}.$$

Let's consider coupling three systems with choices between 1 queue, d queues and with probability p of being flexible respectively, and we call them system 0, system 1 and system p. We use $\sigma = \{0, 1, p\}$ to denote quantities related to system σ by superscript σ. We use $c^{N,\sigma}_m(t)$ to denote the number of patients which have at least m patients queueing in front of then at time $t \geq 0$, which can be written as

$$c^{N,\sigma}_m(t) = N \sum_{k\geq m+1} S^{N,\sigma}_k(t), \quad m \in \mathbb{N}.$$

We will first focus on comparing system 0 and system p. We use a single Poisson process of rate $N\lambda$ for arrivals for both systems. At each jump time, we generate a random variable with $Bernoulli(p)$ distribution to decide whether the patient is flexible or not. If he/she is flexible, we choose uniformly $j_1^p < \cdots < j_d^p$ among $1, \cdots, N$ and then j^0 among $j_1^p < \cdots < j_d^p$, and set $j^p = j_d^p$. If the patient is not flexible, we simply choose uniformly j among $\{1, \cdots, N\}$ and set $j^p = j^0 = j$. In system σ, we order the queues by decreasing length (ties are resolved with uniform probability), and let the task join the queue ranking j^σ in this order. Note that $j^0 \leq j^p$.

We use a single Poisson process of rate N for potential departures for both systems. At each jump time, we choose j uniformly in $\{1, \cdots, N\}$. In system σ, we again order the queues by decreasing length, and remove a task from the j^{th} queue in this order if that queue is not empty.

Our goal is to show that performance is ranked as follows (system $1 \leq$ system $p \leq$ system 0) with respect to the number of patients in the system. Our proof of this coupling is a modification of the proof given in Theorem 4 of Turner [39].

Proposition 3.7 (Coupling Result)**.** *For $N \in \mathbb{N}$, if $c_m^{N,1}(0) \leq c_m^{N,p}(0) \leq c_m^{N,0}(0)$ for all $m \in \mathbb{N}$, then*

$$c_m^{N,1}(t) \leq c_m^{N,p}(t) \leq c_m^{N,0}(t), \quad m \in \mathbb{N}, t \geq 0.$$

Proof. See Appendix for details of the proof. □

Theorem 3.8 (Convergence of Stationary Distributions)**.**

1. *The Markov process $S^N(t)$ is positive recurrent for all N, and therefore has a unique stationary distribution $\pi^N \in \mathcal{P}(\bar{S})$ for each N.*

2. *The sequence of stationary distribution π^N of process $S^N(t)$ converges weakly to the Dirac mass at s^I as $N \to \infty$.*

Proof. By Theorem 3.7, the system 1 is empty whenever system 0 is. Therefore system 1 is also positive recurrent when $\lambda < 1$ and have a stationary distribution π^N. Irreducibility implies the uniqueness of the stationary distribution.

Since $\bar{S}$ is compact, so is the set $\mathcal{P}(\bar{S})$ of the probability measures on $\bar{S}$. Therefore the sequence of probability measures $\{\pi^N\}_{N=1}^\infty$ is tight and has limit points. We aim to show that any limit point of $\{\pi^N\}_{N=1}^\infty$ is the Dirac mass at s^I.

Assume that $S^N(0)$ has the same distribution as the stationary distribution π^N, for each N. By Theorem 3.1, let $\pi^\infty(0)$ be the limiting distribution of a subsequence of $(S^N(0))_{N\geq 1}$, and let $\pi^\infty(t)$ be the limiting distribution for the same subsequence of $(S^N)_{N\geq 1}$. For $t \geq 0$ and $N \geq 1$, since the process started with its stationary distribution, we have that $S^N(t)$ also follows distribution π^N. Applying Theorem 3.1, we have that the fluid limit $s(t) = \lim_{N\to\infty} S^N(t)$ has the same distribution as $\pi^\infty(0)$.

Now let $\epsilon > 0$ and V be an open neighborhood of s^I. For $j \in \mathbb{N}$, let P_j be the set of all a in $\mathcal{P}(S)$ such that the solution for the (3.5) starting at a is in V for all times $t \geq j$. Since P_j is measurable, $P_j \subset P_{j+1}$, and by the fact that s^I is a globally attractive point (Theorem 3.6), we have $\mathcal{P}(S) = \cup_j P_j$, therefore there exists k such that $P(\pi^\infty(0) \in P_k) > 1 - \epsilon$. Then

$$P(\pi^\infty(0) \in V) = P(\pi^\infty(k) \in V) \geq P(\pi^\infty(0) \in P_k) > 1 - \epsilon.$$

Since ϵ and V arbitrary, we have $P(\pi^{\infty}(0) = s^I) = 1$. Hence $(S^N(0))_{N\geq 1}$ converges in distribution to the Dirac mass at s^I, and the limiting distribution of $(S^N)_{N\geq 1}$ is the constant s^I.

□

4 Diffusion Model

In this section, we analyze a diffusion scaled version of the queueing process. Since the fluid limit does not capture stochastic fluctuations, the diffusion model can help us gain important insights on the fluctuations of the system, which can be used to build confidence intervals for various performance measures. To do this, we first prove a functional central limit theorem (CLT) in the transient case for the scaled diffusion process $D^N(t) = \sqrt{N}(S^N(t) - s(t))$ to its limit $D(t)$. We identify $D(t)$ as an infinite dimensional Ornstein Uhlenbeck (OU) process. By computing the variance of $D(t)$, we can construct rigorous confidence intervals for characterizing the deviations from the fluid limit in the transient setting. Second, we prove the functional CLT in the equilibrium setting, thereby establishing an interchange of limits result for the diffusion scaled empirical process. We prove the interchange by showing convergence in the appropriate Hilbert spaces and deriving novel infinite horizon bounds for the diffusion scaled process.

4.1 Transient Analysis of the Diffusion Limit

In this section, we derive the diffusion limit of our stochastic queueing model in the transient setting. We define our scaled diffusion process as

$$D^N(t) = \sqrt{N}(S^N(t) - s(t)). \tag{4.24}$$

Theorem 4.1 (Functional Central Limit Theorem)**.** *Consider ℓ_2 with its weak topology and $\mathbb{D}(\mathbb{R}_+, \ell_2)$ with corresponding Skorokhod topology. Let $s(0)$ be in $\mathcal{S} \bigcap \ell_1$, $S^N(0)$ in $\mathcal{S}^N$. If $(D^N(0))_{N\geq d}$ converges in distribution to $D(0)$ and is tight, then $(D^N(t))_{N\geq d}$ is tight and converges in distribution to the unique OU process*

$$D(t) = D(0) + \int_0^t K(s(u))D(u)du + M(t) \tag{4.25}$$

where the infinite dimensional matrix $K(s)$ is given by

$$K(s) = \begin{pmatrix} -\lambda(1-p) - \lambda p d s_1^{d-1} - 1 & 1 & 0 & \cdots \\ \lambda(1-p) + \lambda p d s_1^{d-1} & -\lambda(1-p) - \lambda p d s_2^{d-1} - 1 & 1 & \cdots \\ 0 & \lambda(1-p) + \lambda p d s_2^{d-1} & -\lambda(1-p) - \lambda p d s_3^{d-1} - 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

and the martingale $M(t)$ is defined by the following Doob-Meyer brackets

$$< M_k(t) > \;\; = \;\; \int_0^t [F_+(s(u))(k) + F_-(s(u))(k)]\, du. \tag{4.26}$$

Consider a linearization of Equation (3.5) around a particular solution g, i.e.,

$$d(t) = g(t) - s(t), \tag{4.27}$$

where g is a generic solution to Equation (3.5). Then we have

$$\dot{d}(t) = K(s(t))d(t), \tag{4.28}$$

where K is a matrix in $\mathbb{Z}_+ \times \mathbb{Z}_+$,

$$K(s) = \begin{pmatrix} -\lambda(1-p) - \lambda p d s_1^{d-1} - 1 & 1 & 0 & \cdots \\ \lambda(1-p) + \lambda p d s_1^{d-1} & -\lambda(1-p) - \lambda p d s_2^{d-1} - 1 & 1 & \cdots \\ 0 & \lambda(1-p) + \lambda p d s_2^{d-1} & -\lambda(1-p) - \lambda p d s_3^{d-1} - 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Let $(M_k(t))_{k\in\mathbb{N}}$ be a family of independent, real, continuous centered Gaussian martingales, determined in law by their deterministic Doob-Meyer brackets given by

$$\begin{aligned} < M_k(t) > &= \int_0^t \left[\lambda(1-p)(s_{i-1}(u) - s_i(u)) + \lambda p \left(s_{i-1}(u)^d - s_i(u)^d\right) + (s_i(u) - s_{i+1}(u))\right] du \\ &= \int_0^t \left[F_+(s(u))(k) + F_-(s(u))(k)\right] du. \end{aligned} \tag{4.29}$$

for $t \geq 0$.

To give readers a high-level understanding of the proof idea, we summarize the following list of steps for showing the functional CLT in the transient case,

1. Prove the Lipschitz property for the mappings F, F_+, F_- in ℓ_2. (Theorem 3.2)
2. Prove the Gaussian martingale $M(t)$ is square-integrable in ℓ_2. (Theorem 4.2)
3. Prove the existence and uniqueness of the diffusion limit $D(t)$ by using steps 1 and 2 to show that Equation (4.31) is well-defined and solves the SDE.
4. Show the difference between the drift function $F^N(s)$ and the limiting drift function $F(s)$ is $\frac{1}{N}O(s)$. (Lemma 4.7)
5. Show the finite horizon bound

$$\limsup_{N\to\infty} \mathbb{E}\left(\|D^N(0)\|_{\ell_2}^2\right) < \infty \Rightarrow \limsup_{N\to\infty} \mathbb{E}\left(\sup_{t\leq T} \|D^N(t)\|_{\ell_2}^2\right) < \infty$$

   using Doob's inequality, Gronwall's lemma and steps 1,2, and 4 .
6. Use step 5 to show the tightness of the diffusion process. (Lemma 4.9).
7. Use steps 1-6 to show the functional CLT, i.e. when initial condition converges, the diffusion process D^N converges to the unique OU process solving Equation (4.31). (Theorem 4.1)

Theorem 4.2. *Assume $s(0)$ to be in $\mathcal{S}$. Then, the Gaussian martingale $M(t)$ is square-integrable in ℓ_2.*

Proof. The proof is provided in the Appendix. □

Let $D(t)$ be the diffusion limit for the fluctuations $D^N(t)$, which is a Gaussian perturbation of Equation (4.28), then $D(t)$ satisfies the following SDE for any given $t \geq 0$,

$$D(t) = D(0) + \int_0^t K(s(u))D(u)du + M(t). \tag{4.30}$$

Theorem 4.3 (Existence and Uniqueness of Diffusion Limit)**.** *1. For s in $\mathcal{S}$, the operator $K(s)$ is bounded in ℓ_2 with operator norm uniformly bounded in s.*

2. Let $s(0)$ be in $\mathcal{S} \bigcap \ell_1$. Then there exists a unique strong solution to Equation (4.30) in ℓ_2

$$D(t) = \exp\left\{\int_0^t K(s(u))du\right\} D(0) + \int_0^t \exp\left\{\int_u^t K(s(r))dr\right\} dM(u), \tag{4.31}$$

and

$$\mathbb{E}\left(\|D(0)\|_{\ell_2}^2\right) < \infty \Rightarrow \mathbb{E}\left(\sup_{t\leq T} \|D(t)\|_{\ell_2}^2\right) < \infty.$$

Proof. The proof is provided in the Appendix. □

For the following Lemma 4.4 and Theorem 4.1, the proofs are detailed in subsections 4.1.1 and 4.1.2.

Lemma 4.4 (Finite Horizon Bound)**.** *Let $s(0)$ be in $\mathcal{S} \bigcap \ell_1$ and $S^N(0)$ be in $\mathcal{S}^N$. Then for any $T \geq 0$,*

$$\limsup_{N\to\infty} \mathbb{E}\left(\|D^N(0)\|_{\ell_2}^2\right) < \infty \Rightarrow \limsup_{N\to\infty} \mathbb{E}\left(\sup_{t\leq T} \|D^N(t)\|_{\ell_2}^2\right) < \infty.$$

Theorem 4.5. *Define the two matrices $\mathcal{A}(t) = K(s(t))$, $\mathcal{B}(t) = \left(\frac{d}{dt}\langle M_i(t), M_j(t)\rangle\right)_{ij}$, then the expectation $E(D(t))$ is*

$$\mathbb{E}[D(t)] = e^{\int_0^t \mathcal{A}(s)ds}\mathbb{E}[D(0)], \tag{4.32}$$

and the covariance matrix $\Sigma(t) = \mathrm{Cov}[D(t), D(t)]$ is

$$\Sigma(t) = e^{\int_0^t \mathcal{A}(s)ds}\Sigma(0)e^{\int_0^t \mathcal{A}^\top(s)ds} + \int_0^t e^{\int_s^t \mathcal{A}(u)du}\mathcal{B}(s)e^{\int_s^t \mathcal{A}^\top(u)du}ds. \tag{4.33}$$

Moreover, differentiation with respect to t yields

$$\frac{d\mathbb{E}[D(t)]}{dt} = \mathcal{A}(t)\mathbb{E}[D(t)], \tag{4.34}$$

$$\frac{d\Sigma(t)}{dt} = \Sigma(t)\mathcal{A}(t)^\top + \mathcal{A}(t)\Sigma(t) + \mathcal{B}(t). \tag{4.35}$$

Componentwise, we have

$$\begin{aligned}\frac{d\Sigma_{i,i}(t)}{dt} &= 2\left[\lambda(1-p)+\lambda pds_{i-1}^{d-1}\right]\Sigma_{i,i-1}-2\left[\lambda(1-p)+\lambda pds_i^{d-1}+1\right]\Sigma_{i,i}+2\Sigma_{i,i+1}\\ &\quad+\lambda(1-p)(s_{i-1}-s_i)+\lambda p(s_{i-1}^d-s_i^d)+s_i-s_{i+1}, \qquad (4.36)\end{aligned}$$

$$\begin{aligned}\frac{d\Sigma_{i,j}(t)}{dt} &= \left[2\lambda(1-p)+\lambda pd(s_{i-1}^{d-1}+s_{j-1}^{d-1})\right]\Sigma_{j,i-1}-\left[2\lambda(1-p)+\lambda pd(s_i^{d-1}+s_j^{d-1})+2\right]\Sigma_{j,i}\\ &\quad+2\Sigma_{j,i+1}. \qquad (4.37)\end{aligned}$$

Proof. The proof is provided in the Appendix. □

4.1.1 The Derivation of the Ornstein-Uhlenbeck Process

In this subsection, we introduce a few lemmas which help show the final functional CLT result in the transient case.

Lemma 4.6. *Let $S^N(0)$ be in $\mathcal{S}^N$, s solves Equation (3.5) with $s(0)\in\mathcal{S}$. Then*

$$D^N(t)=D^N(0)+\int_0^t\sqrt{N}(F^N(S^N(u))-F(s(u)))du+M^N(t) \qquad (4.38)$$

defines an independent family of square-integrable martingales M^N independent of $S^N(0)$ with Doob-Meyer brackets given by

$$<M_k^N(t)>=\int_0^t\left(F_+^N(S^N(u)(k))+F_-(S^N(u))(k)\right)du. \qquad (4.39)$$

Proof. This follows from a classical application of Dynkin's formula. □

Lemma 4.7. *Define function $A^N(a)$ for $a\in\mathbb{R}$ and $N\geq d\geq 1$ as*

$$A^N(a)\triangleq\frac{(Na)_d}{(N)_d}-a^d. \qquad (4.40)$$

Then, $A^N(a)=\frac{1}{N}O(a)$ uniformly on $0\leq a\leq 1$ and $A^N(k/N)\leq 0$ for $k=0,1,\cdots,N$.

Proof. See Appendix for details of the proof. □

Lemma 4.8. *For $d\geq 1$ and $a,h\in\mathbb{R}$, define*

$$B(a,h)\triangleq(a+h)^d-a^d-da^{d-1}h=\sum_{i=2}^d\binom{d}{i}a^{d-i}h^i. \qquad (4.41)$$

Then $B(a,h)=0$ for $d=1$ and $B(a,h)=h^2$ for $d=2$. For $d\geq 2$ we have $0\leq B(a,h)\leq h^d+(2^d-d-2)ah^2$ for $a,a+h\in[0,1]$.

Proof. See Appendix for details of the proof. □

4.1.2 Proof of the functional CLT

Now consider the mapping $G^N : \mathcal{S} \to c_0^0$ given by

$$G^N(s)(k) = \lambda p \left(A^N(s_{k-1}) - A^N(s_k) \right), \quad k \geq 1 \tag{4.42}$$

and $H : \mathcal{S} \times c_0^0 \to c_0^0$ given by

$$H(s,x)(k) = \lambda p (B(s_{k-1}, x_{k-1}) - B(s_k, x_k)), \quad k \geq 1 \tag{4.43}$$

so that for $s + x \in \mathcal{S}$, we have

$$F^N = F + G^N, \quad F(s+x) - F(s) = K(s)x + H(s,x). \tag{4.44}$$

Proof of Lemma 4.4 (Finite-horizon bound). By Equations (4.38) and (4.44), we have

$$D^N(t) = D^N(0) + M^N(t) + \sqrt{N} \int_0^t G^N(S^N(u))du + \int_0^t \sqrt{N}(F(S^N(u)) - F(s(u)))du. \tag{4.45}$$

Since Lemma 4.7 indicates that

$$G^N(S^N(u))(k) = \lambda p \left(A^N(S^N(u)(k-1)) - A^N(S^N(u)(k) \right) = \frac{1}{N} O \left(S^N(u)(k-1) + S^N(u)(k) \right),$$

we can conclude that

$$\|G^N(S^N(u))\|_{\ell_2} = \frac{1}{N} O \left(\|S^N(u)\|_{\ell_2} \right). \tag{4.46}$$

By definition of the diffusion process we have

$$\|S^N(u)\|_{\ell_2} \leq \|s(u)\|_{\ell_2} + \frac{1}{\sqrt{N}} \|D(u)^N\|_{\ell_2}. \tag{4.47}$$

Since mappings F_+, F_-, F are Lipschitz with respect to ℓ_2 norm, Gronwall's lemma yields that

$$\|s(u)\|_{\ell_2} \leq L_T \|s(0)\|_{\ell_2} \tag{4.48}$$

for some constant $L_T < \infty$. Then

$$\begin{aligned}
\|D^N(t)\|_{\ell_2} &\leq \|D^N(0)\|_{\ell_2} + \|M^N(t)\|_{\ell_2} + \sqrt{N}\int_0^t \|G^N(S^N(u))\|_{\ell_2} du \\
&\quad + \int_0^t \sqrt{N}(\|F(S^N(u)) - F(s(u))\|_{\ell_2}) du \\
&\leq \|D^N(0)\|_{\ell_2} + \|M^N(t)\|_{\ell_2} + \sqrt{N}\int_0^t \frac{1}{N} O\left(\|S^N(u)\|_{\ell_2}\right) du \\
&\quad + \int_0^t \sqrt{N} L(\|S^N(u) - s(u)\|_{\ell_2}) du \\
&\leq \|D^N(0)\|_{\ell_2} + \|M^N(t)\|_{\ell_2} + \int_0^t \frac{1}{\sqrt{N}} O\left(\|s(u)\|_{\ell_2} + \frac{1}{\sqrt{N}}\|D(u)^N\|_{\ell_2}\right) du \\
&\quad + \int_0^t L(\|D(u)^N\|_{\ell_2}) du \\
&\leq \|D^N(0)\|_{\ell_2} + \|M^N(t)\|_{\ell_2} + \frac{1}{\sqrt{N}} O\left(L_T \|s(0)\|_{\ell_2}\right) \qquad (4.49) \\
&\quad + \int_0^t \left(L + O\left(\frac{1}{N}\right)\right) \|D(u)^N\|_{\ell_2} du. \qquad (4.50)
\end{aligned}$$

By Gronwall's lemma we have

$$\begin{aligned}
&\sup_{0\leq t\leq T} \|D^N(t)\|_{\ell_2} \\
\leq\ & \exp\left\{\left(L + O\left(\frac{1}{N}\right)\right) T\right\} \left(\|D^N(0)\|_{\ell_2} + \sup_{0\leq t\leq T}\|M^N(t)\|_{\ell_2} + \frac{L_T}{\sqrt{N}} O\left(\|s(0)\|_{\ell_2}\right)\right). \qquad (4.51)
\end{aligned}$$

Using Doob's ℓ_2 inequality we know that,

$$\mathbb{E}\left(\sup_{0\leq t\leq T}\|M^N(t)\|_{\ell_2}\right) \leq 2\mathbb{E}\left(\|M^N(T)\|_{\ell_2}\right). \qquad (4.52)$$

By Lemma 4.6 and Lipschitz property of F_+, F_-,

$$\begin{aligned}
\|M_T^N\|_{\ell_2} &= \int_0^T \|F_+^N(S^N(u)) + F_-(S^N(u))\|_{\ell_2} du \\
\text{(Equation (4.44))} &= \int_0^T \|F_+(S^N(u)) + G^N(S^N(u)) + F_-(S^N(u))\|_{\ell_2} du \\
\text{(Equation (4.46))} &\leq \int_0^T \left(2L\|S^N(u)\|_{\ell_2} + \frac{1}{N} O(\|S^N(u)\|_{\ell_2})\right) du \\
\text{(Equation (4.47))} &\leq \int_0^T O\left(\|s(u)\|_{\ell_2} + \frac{1}{\sqrt{N}}\|D(u)^N\|_{\ell_2}\right) du \\
\text{(Equation (4.48))} &= K_T O(\|s(0)\|_{\ell_2}). \qquad (4.53)
\end{aligned}$$

Finally combining all the above equations, we conclude that when

$$\limsup_{N\to\infty} \mathbb{E}\left(\|D^N(0)\|_{\ell_2}^2\right) < \infty,$$

we have

$$\begin{aligned}
& \limsup_{N\to\infty} \mathbb{E}\left(\sup_{0\le t\le T} \|D^N(t)\|_{\ell_2}\right) \\
\le\ & \exp\{O(T)\}\left(\limsup_{N\to\infty} \mathbb{E}\left(\|D^N(0)\|_{\ell_2}\right) + \limsup_{N\to\infty} \mathbb{E}\left(\sup_{0\le t\le T} \|M^N(t)\|_{\ell_2}\right) + \frac{L_T}{\sqrt{N}} O\left(\|s(0)\|_{\ell_2}\right)\right) \\
\le\ & \exp\{O(T)\}\left(\limsup_{N\to\infty} \mathbb{E}\left(\|D^N(0)\|_{\ell_2}\right) + \limsup_{N\to\infty} 2\mathbb{E}\left(\|M^N(T)\|_{\ell_2}\right) + \frac{L_T}{\sqrt{N}} O\left(\|s(0)\|_{\ell_2}\right)\right) \\
\le\ & \exp\{O(T)\}\left(\limsup_{N\to\infty} \mathbb{E}\left(\|D^N(0)\|_{\ell_2}\right) + \left(2K_T + \frac{L_T}{\sqrt{N}}\right) O\left(\|s(0)\|_{\ell_2}\right)\right) \\
<\ & \infty. \qquad (4.54)
\end{aligned}$$

□

Lemma 4.9 (Tightness of the Process)**.** *Consider ℓ_2 with its weak topology and $\mathbb{D}(\mathbb{R}_+;\ell_2)$ with the corresponding Skorokhod topology. Assume $s(0)\in\mathcal{S}\cap\ell_1$ and $S^N(0)\in\mathcal{S}^N$, and D^N as defined in the beginning of the section. If $(D^N(0))_{N\ge d}$ is tight, then $(D^N)_{N\ge d}$ is tight and its limit points are continuous.*

Proof. Since $\mathbb{D}(\mathbb{R}_+;\ell_2)$ is a reflexive Banach space, relatively compact sets are the bounded sets for the norm ℓ_2. Then here a process D^N is tight if and only if for any $\epsilon>0$ there exists $r_\epsilon<\infty$ such that $\mathbb{P}(D^N\in B(r_\epsilon))>1-\epsilon$ for $N\ge 1$. We refer to Ethier and Kurtz [16] the tightness criteria for showing that $(D^N)_{N\ge d}$ is tight. That is, $(D^N)_{N\ge d}$ is tight if

1. For each $T\ge 0$ and $\epsilon>0$ there is a bounded subset $K_{T,\epsilon}\in\ell_2$ such that $\mathbb{P}(D^N\in D([0,T];K_{T,\epsilon}))>1-\epsilon$ for $N\ge d$.

2. For each $k\ge 1$, the k-dimensional process $(D_1^N, D_2^N, \cdots, D_k^N)_{N\ge d}$ are tight.

For Condition 1, it is easy to see that using finite-horizon bound in Lemma 4.4 and Markov inequality, we can derive the tightness of process D^N on $D([0,T];K_{T,\epsilon})$.

For Condition 2, we refer to Graham [23] for the fact that bounds in Lemma 4.4 and that D_k^N has jump size of $\frac{1}{\sqrt{N}}$ classically imply the tightness of the finite-dimensional process. □

Proof of Theorem 4.1 (Functional CLT). Using Lemma 4.9, we know that any subsequence of D^N has a further subsequence that converges to some limit D^∞ with continuous sample path. $D^\infty(0)$ should have the same distribution as $D(0)$. We can rewrite Equation (4.45) as

$$\begin{aligned}
D^N(t) &= D^N(0) + M^N(t) + \int_0^t K(s(u))D^N(u)du \\
&\quad + \sqrt{N}\int_0^t \left(G^N(S^N(u)) + H\left(s(u), D^N(u)/\sqrt{N}\right)\right)du. \qquad (4.55)
\end{aligned}$$

Using Equations (4.46), (4.47), we have that $\sqrt{N}\|G^N(S^N(u))\|_{\ell_2} = \frac{1}{\sqrt{N}} O(\|S^N(u)\|_{\ell_2}) \to 0$ as $N \to \infty$. Using Lemma 4.8, we have

$$\begin{aligned} & \sqrt{N}\|H(s(u), D^N(u)/\sqrt{N})\|_{\ell_2} \\ \leq \; & \sqrt{N}\lambda p \left[\frac{1}{N^{d/2}} \|(D^N(u))^d\|_{\ell_2} + \frac{1}{N}(2^d - d - 2)\|s(u)\|_{\ell_2} \cdot \|D^N(u)\|_{\ell_2}^2 \right] \\ \to \; & 0 \end{aligned} \tag{4.56}$$

as $N \to \infty$. We also have the martingale brackets

$$\begin{aligned} < M_k^N(t) > \; &= \; \int_0^t \left(F_+^N(S^N(u))(k) + F_-(S^N(u))(k) \right) du \\ &\to \; \int_0^t \left(F_+(s(u))(k) + F_-(s(u))(k) \right) du \\ &= \; < M_k(t) > \end{aligned} \tag{4.57}$$

as $N \to \infty$.

By Theorem 4.1 in Ethier and Kurtz [16], together with Lipschitz property of F in Lemma 3.2, finite horizon bounds in Lemma 4.4 and tightness results in Lemma 4.9, we deduce by a martingale characterization that D^∞ has the distribution of the OU process which is the unique solution for (4.30) in ℓ_2 starting at $D^\infty(0)$. Thus, this distribution D^∞ is the unique accumulation point for the relatively compact sequence of distributions of $(D^N)_{N\geq 1}$, therefore itself must then converge to it, proving Theorem 4.1. □

4.2 Steady State Analysis of Diffusion Limit

In this section, we analyze the steady state of the diffusion model. This allows us to gain insights about the long-time behavior of the nonlinear system dynamics appearing at the large N limit. Assume we have $\lambda < 1$, and that $s(0) = s^I$. Define the infinite-dimensional matrix $\mathcal{K} = K(s^I)$. Then we have

$$\mathcal{K} = \begin{pmatrix} -\lambda(1-p) - \lambda p d (s_1^I)^{d-1} - 1 & 1 & 0 & \cdots \\ \lambda(1-p) + \lambda p d (s_1^I)^{d-1} & -\lambda(1-p) - \lambda p d (s_2^I)^{d-1} - 1 & 1 & \cdots \\ 0 & \lambda(1-p) + \lambda p d (s_2^I)^{d-1} & -\lambda(1-p) - \lambda p d (s_3^I)^{d-1} - 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Note that $\mathcal{K} = \mathcal{A}^*$ where $\mathcal{A}$ is the generator of a sub-Markovian birth-death process. We use $\pi = (\pi_k)_{k\geq 1}$ to denote the tail cdf of the the stationary distribution to $\mathcal{A}$. Then, π solves the following balance equations

$$\begin{aligned} \pi_1 \; &= \; 1, \\ \pi_{k+1} \; &= \; \left[\lambda(1-p) + \lambda p d (s_k^I)^{d-1} \right] \pi_k, \quad k \geq 1. \end{aligned} \tag{4.58}$$

Consider the independent and centered Brownian motions $B(t) = (B_k(t))_{k\geq 0}$ such that $B(0) = 0$, and for $k \geq 1$

$$\upsilon_k \triangleq \mathrm{Var}(B_k(1)) = \mathbb{E}(B_k(1)) = 2(s_k^I - s_{k+1}^I)$$

and B has an infinitesimal covariance matrix $\mathrm{diag}(\upsilon)$. The OU process $D(t) = (D_k(t))_{k\in\mathbb{N}}$ solves the affine SDE given for $t \geq 0$ by

$$D(t) = D(0) + \int_0^t \mathcal{K}D(s)ds + B(t) \tag{4.59}$$

which is a Brownian perturbation of the following differential equation

$$\dot{d}(t) = \mathcal{K}d(t). \tag{4.60}$$

Our ultimate goal is to show the interchanging of limits for the diffusion model. However, a main difficulty is that the scalar product for which the operator $\mathcal{K}$ is self-adjoint is too strong for the limit dynamical system and the invariant measures for finite N. Thus, we need to consider appropriate Hilbert spaces in which the operator $\mathcal{K}$ is not self-adjoint and prove the exponential stability of the fluid limit in the newly introduced space. As a result, we introduce the following weighted Hilbert space

$$L_2(w) \triangleq \left\{ x \in \mathbb{R}^{\mathbb{N}} : x(0) = 0, \|x\|^2_{L_2(w)} = \sum_{k\geq 1} x(k)^2 w(k)^{-1} < \infty \right\}.$$

We also consider the following ℓ_1 space with same weights

$$L_1(w) \triangleq \left\{ x \in \mathbb{R}^{\mathbb{N}} : x(0) = 0, \|x\|^2_{L_1(w)} = \sum_{k\geq 1} |x(k)| w(k)^{-1} < \infty \right\}.$$

For easier notation, we denote the sequence $g_\theta = (\theta^k)_{k\geq 1}$. WLOG we assume that $d \geq 2$ and $p \in (0, 1)$ since otherwise the system goes back to JSQ(d) (refer to Graham [23] for their results). Notice that by induction we can show that for $k \geq 2$,

$$\lambda^k(1-p)^{k-1} < s^I_k < \lambda^k \tag{4.61}$$

$$\lambda^{k-1}(1-p)^{k-1} < \pi_k < \lambda^{k-1} \tag{4.62}$$

which means that both s^I and π have exponential decay. In the rest of the paper, we assume that w satisfies the following condition,

$$\exists c, d > 0, \forall k \geq 1, 0 < cw(k+1) \leq w(k) \leq dw(k+1). \tag{4.63}$$

This condition implies $w(1)d(1/d)^k \leq w(k) \leq w(1)c(1/c)^k$, which means w is bounded by geometric sequences.

Theorem 4.10 (Functional Central Limit Theorem in Equilibrium)**.** *Let w satisfies condition (4.63), then*

1. *In $L_2(w)$, the operator $\mathcal{K}$ is bounded, and Equation (4.60) has a unique solution $d_t = e^{\mathcal{K}t}d(0)$. The assumptions and conclusions hold for $w = \pi$ and $w = g_\theta$ for $\theta > 0$.*

2. In addition, let w be such that s^I is in $L_1(w)$. The SDE (4.59) has a unique solution

$$D(t) = e^{\mathcal{K}t}D(0) + \int_0^t e^{\mathcal{K}(t-s)}dB(s)$$

in $L_2(w)$. This is the case for $w = g_\theta$ for $\theta \geq \lambda$ when $d \geq 2$.

Proof. Using the condition in Equation (4.63) and our convexity bounds, we have

$$\begin{aligned}
\|\mathcal{K}x\|_{L_2(w)} &= \sum_{k\geq 1}\left[(\lambda(1-p)+\lambda pd(s^I_{k-1})^{d-1})x_{k-1} - (\lambda(1-p)+\lambda pd(s^I_k)^{d-1}+1)x_k + x_{k+1}\right]^2 w(k)^{-1} \\
&\leq 3\Bigg(\sum_{k\geq 1}(\lambda(1-p)+\lambda pd(s^I_{k-1})^{d-1})^2 x^2_{k-1}w(k)^{-1} + (\lambda(1-p)+\lambda pd(s^I_k)^{d-1}+1)^2x^2_k w(k)^{-1} \\
&\quad + x^2_{k+1}w(k)^{-1}\Bigg) \\
&\leq 3\Bigg(\sum_{k\geq 1}(\lambda(1-p)+\lambda pd)^2 x^2_{k-1}dw(k-1)^{-1} + (\lambda(1-p)+\lambda pd+1)^2x^2_k w(k)^{-1} \\
&\quad + x^2_{k+1}c^{-1}w(k+1)^{-1}\Bigg) \\
&\leq 3(d(\lambda(1-p)+\lambda pd)^2 + (\lambda(1-p)+\lambda pd+1)^2 + c^{-1})\|x\|_{L_2(w)}. \qquad (4.64)
\end{aligned}$$

Then by applying Gronwall's lemma we have the uniqueness result. When B is an Hilbertian Brownian motion, the formula for $D(t)$ yields a well-defined solution. □

4.3 Interchanging limits of Diffusion Limit

Our goal in this section is to prove the following diagram commutes.

$$\begin{array}{ccc}
D^N(t) & \xrightarrow{N\to\infty} & D(t) \\
{\scriptstyle t\to\infty}\downarrow & & \downarrow{\scriptstyle t\to\infty} \\
D^N(\infty) & \xrightarrow[N\to\infty]{} & D(\infty)
\end{array}$$

We have showed in Section 4.1 that $D^N(t) \xrightarrow{d} D(t)$, and the existence and uniqueness of $D(t)$. Now we will show the existence and uniqueness of the equilibrium point $D(\infty)$ of the diffusion limit, and show the weak convergence of invariant measure $D^N(\infty)$ to $D(\infty)$. The proof idea of the interchanging limits of diffusion limits takes the following list of steps:

1. Prove the equilibrium operator $\mathcal{K}$ has bounded spectral gap in the self-adjoint space $L_2(\pi)$, which implies exponential stability of linearized solution d_t in $L_2(\pi)$. (Theorem 4.12)

2. Prove exponential stability of fluid limit $s(t)$ in non self-adjoint space $L_2(g(\theta))$, by constructing a specific birth-death process and obtain exponential stability of its solution $z(t)$ via step 1, then bounding the difference between the fluid limit $s(t)$ and $z(t)$. (Theorem 4.15, Lemma 4.13, Lemma 4.14)

3. Show the infinite horizon bound in space $L_2(g(\theta))$ using the exponential stability result of $s(t)$ in step 2. (Theorem 4.16)

4. Show the weak convergence of stationary distributions $D^N(\infty)$ to the equilibrium point $D(\infty)$ of the diffusion limit $D(t)$. (Theorem 4.17).

Consider $\mathcal{A} = \mathcal{K}^*$, the infinitesimal generator of the sub-Markovian birth death process with birth rates $\lambda_k = \lambda(1-p) + \lambda p d(s_k^I)^{d-1}$ and death rates $\mu_k = 1$ for $k \geq 1$. Let $Q(x) = (Q_n(x))_{n\geq 1}$ denote an eigenvector for $\mathcal{A}$ of eigenvalue $-x$. Then, we have $\lambda_1 Q_2(x) = (\lambda_1 + \mu_1 - x)Q_1(x)$ and $\lambda_n Q_{n+1}(x) = (\lambda_n + \mu_n - x)Q_n(x) - \mu_n Q_{n-1}(x)$ for $n \geq 2$. Such a sequence of polynomials is orthogonal with respect to a probability measure ψ on $\mathbb{R}^+$ such that

$$\text{diag}(\pi^{-1}) = \int_0^\infty Q(x)Q(x)^*\psi(dx).$$

Such a probability measure is called the spectral measure, with its support S called the spectrum. We denote the spectral gap $\gamma = \min S$. The representation formula of Karlin and McGregor [25, 26] yields

$$e^{\mathcal{K}t} = \text{diag}(\pi)\int_0^\infty e^{-xt}Q(x)Q(x)^*\psi(dx). \tag{4.65}$$

Therefore, we have the following lemma which gives the solution of the unique equilibrium point of the OU process.

Lemma 4.11. *The OU process $D(t)$ in Theorem 4.10, its equilibrium point $D(\infty)$, and its covariance matrix $\Sigma(\infty)$ can be written as*

$$D(t) = diag(\pi)\int_S e^{-xt}Q(x)^*\left(D(0) + \int_0^t e^{xs}dB(s)\right)Q(x)\psi(dx) \tag{4.66}$$

$$D(\infty) \;=\; diag(\pi)\int_S \left(Q(x)^*\int_0^\infty e^{-xt}dB(t)\right)Q(x)\psi(dx) \tag{4.67}$$

$$\Sigma(\infty) \;=\; diag(\pi)\int_{S^2}\frac{Q(x)^*diag(v)Q(y)}{x+y}Q(x)Q(y)^*\psi(dx)\psi(dy)diag(\pi). \tag{4.68}$$

Proof. The proof is provided in the Appendix.

□

Theorem 4.12 (Spectral Gap for self-adjoint case)**.** *The operator $\mathcal{K}$ is bounded self-joint in $L_2(\pi)$. The least point γ of the spectrum of $\mathcal{K}$ is such that $0 < \gamma \leq (\sqrt{\lambda(1-p)} - 1)^2$. The solution $d(t) = e^{\mathcal{K}t}d(0)$ for Equation (4.60) in $L_2(\pi)$ satisfies $\|d(t)\|_{L_2(\pi)} \leq e^{-\gamma t}\|d(0)\|_{L_2(\pi)}$.*

Proof. The potential coefficients π solve the detailed balance equations for $\mathcal{A}$ and hence $\mathcal{K} = \mathcal{A}^*$ is self-adjoint in $L_2(\pi)$. It is established in Theorem 5.1 and Theorem 5.3 in Doorn [14] that $\gamma > 0$ if and only if

$$\sigma = \left(\sqrt{\lim_k \lambda_k} - \sqrt{\lim_k \mu_k}\right)^2 = \left(\sqrt{\lambda(1-p)} - 1\right)^2 > 0.$$

For exponential stability, we have $\|d(t)\|^2_{L_2(\pi)} = \left(e^{\mathcal{K}t}d(0), e^{\mathcal{K}t}d(0)\right)_{L_2(\pi)}$ and the fact that $e^{\mathcal{K}t}$ is self-adjoint in $L_2(\pi)$ and the spectral representation yield

$$\begin{aligned}\left(e^{\mathcal{K}t}d(0), e^{\mathcal{K}t}d(0)\right)_{L_2(\pi)} &= \left(d(0), e^{2\mathcal{K}t}d(0)\right)_{L_2(\pi)} = \int_S e^{-2xt}d(0)^*Q(x)Q(x)^*d(0)\psi(dx) \\ &\leq e^{-2\gamma t}\int_S d(0)^*Q(x)Q(x)^*d(0)\psi(dx) = e^{-2\gamma t}(d(0),d(0))_{L_2(\pi)}.\end{aligned} \tag{4.69}$$

□

For the proof of exponential stability for non self-adjoint case, we modify an argument of Graham [23]. We first consider the centered dynamical system $y(t) = s(t) - s^I$, then y solves the centered equation

$$\dot{y}(t) = F(s^I + y) = \mathcal{K}y(t) + H(s^I, y(t)),$$

or

$$\begin{aligned}\dot{y}_k(t) &= [\lambda(1-p) + \lambda pd(s^I_{k-1})^{d-1}]y_{k-1}(t) + \lambda pB\left(s^I_{k-1}, y_{k-1}(t)\right) \\ &\quad -[\lambda(1-p) + \lambda pd(s^I_k)^{d-1} + 1]y_k(t) - \lambda pB\left(s^I_k, y_k(t)\right) + y_{k+1}(t).\end{aligned} \tag{4.70}$$

We also have

$$\dot{y}_k(t) + \dot{y}_{k+1}(t) + \cdots = [\lambda(1-p) + \lambda pd(s^I_{k-1})^{d-1}]y_{k-1}(t) + \lambda pB\left(s^I_{k-1}, y_{k-1}(t)\right) - y_k(t).$$

Lemma 4.13. *Let $\hat{A}$ be the generator of the sub-Markovian birth and death process with birth rate $\hat{\lambda}_k \geq 0$ and death rate 1 for $k \geq 1$. Assume $\sup_k \hat{\lambda}_k < \infty$. Let $z(t)$ solves $\dot{z} = \hat{A}^* z$ in ℓ^0_1. Let $h(t)$ be given in ℓ^0_1 by*

$$h_k(t) = \sum_{i \geq k}(z_i(t) - y_i(t)), \quad k \geq 1$$

Then,

(1) Let $\hat{\lambda}_k \geq [\lambda(1-p) + \lambda pd(s^I_k)^{d-1}] + \lambda p(1 + (2^d - d - 2)s^I_k)$ for $k \geq 1$, $y(0) \geq 0$ and $h(0) \geq 0$. Then $h(t) \geq 0$ for $t \geq 0$.

(2) Let $\hat{\lambda}_k \geq [\lambda(1-p) + \lambda pd(s^I_k)^{d-1}]$ for $k \geq 1$, $y(0) \leq 0$ and $h(0) \leq 0$. Then $h(t) \leq 0$ for $t \geq 0$.

Proof. The proof is provided in the Appendix. □

Lemma 4.14. *For any $0 \leq \theta < 1$ there exists $K_\theta < \infty$ such that for x in $L_2(g_\theta) \subset \ell^0_1$*

$$\|(x_k + x_{k+1} + \cdots)_{k \geq 1}\|_{L_2(g_\theta)} \leq K_\theta \|x\|_{L_2(g_\theta)}.$$

Proof. The proof is provided in the Appendix.

□

Now we finish the proof of Theorem 4.15 using the previous two lemmas.

Theorem 4.15 (Exponential stability for non self-adjoint case)**.** *Let $\lambda \leq \theta < 1$ and s be the solution to Equation (3.5) starting at $s(0)$ in $\mathcal{S} \cap L_2(g_\theta)$. There exists $\gamma_\theta > 0$ and $C_\theta < \infty$ such that*

$$\|s(t) - s^I\|_{L_2(g_\theta)} \leq e^{-\gamma_\theta t} C_\theta \|s(0) - s^I\|_{L_2(g_\theta)}.$$

Proof. The proof is given in the Appendix. □

Theorem 4.16 (Infinite Horizon Bound)**.** *Assume $\lambda \leq \theta < 1$, then*

$$\limsup_{N\to\infty} \mathbb{E}\left(\|D^N(0)\|^2_{L_2(g_\theta)}\right) < \infty \Rightarrow \limsup_{N\to\infty} \sup_{t\geq 0} \mathbb{E}\left(\|D^N(t)\|^2_{L_2(g_\theta)}\right)$$

Proof. The proof is given in the Appendix. □

We now prove that the interchanging of limits is valid, through the following steps:

1) The sequence $(D^N(\infty), N \geq 1)$ is tight.

2) There is a unique possible limit to any convergent subsequence of $(D^N(\infty), N \geq 1)$.

Theorem 4.17. *The stationary distribution $D^N(\infty)$ of the diffusion process $D^N(t)$ converges weakly to the equilibrium point of the diffusion limit $D(\infty)$, whose explicit form is specified in Equation (4.67).*

Proof. Since for any $K > 0$, using Markov inequality we have

$$\begin{aligned} P(\|D^N(\infty)\|_{L_2(g_\theta)} > K) &= \lim_{t\to\infty} P(\|D^N(t)\|_{L_2(g_\theta)} > K) \\ &= \lim_{t\to\infty} \frac{\mathbb{E}[\|D^N(t)\|^2_{L_2(g_\theta)}]}{K^2} \\ &= O\left(\frac{1}{K^2}\right). \end{aligned} \tag{4.71}$$

This shows that $(D^N(\infty), N \geq 1)$ is tight. Now we only need to prove that there is a unique possible limit to any convergent subsequence of $(D^N(\infty), N \geq 1)$. We still denote by $D^N(\infty)$ such a converging subsequence. Its limit is denoted by ν. By properties of Markov processes, $D^N(t)$ with initial condition $D^N(0) = D^N(\infty)$ is a stationary process, hence $D(t) = \nu$ for any t. Then $D(\infty) = \nu$, which proved that any convergent subsequence of $D^N(\infty)$ converge to $D(\infty)$. □

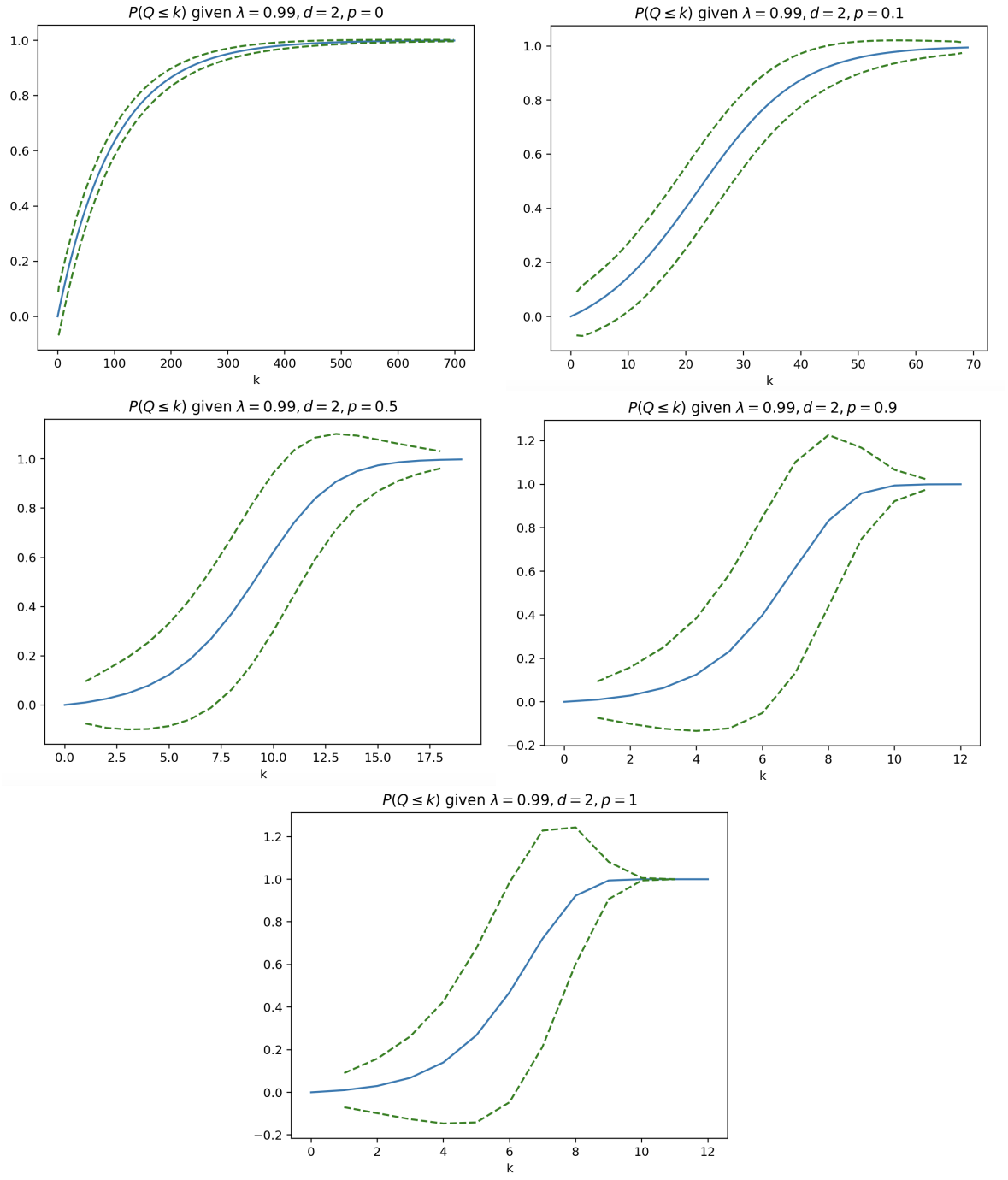

Figure 3: Green dotted line indicates 1 standard deviation computed according to 4.5 with a cut off at 1000 iterations

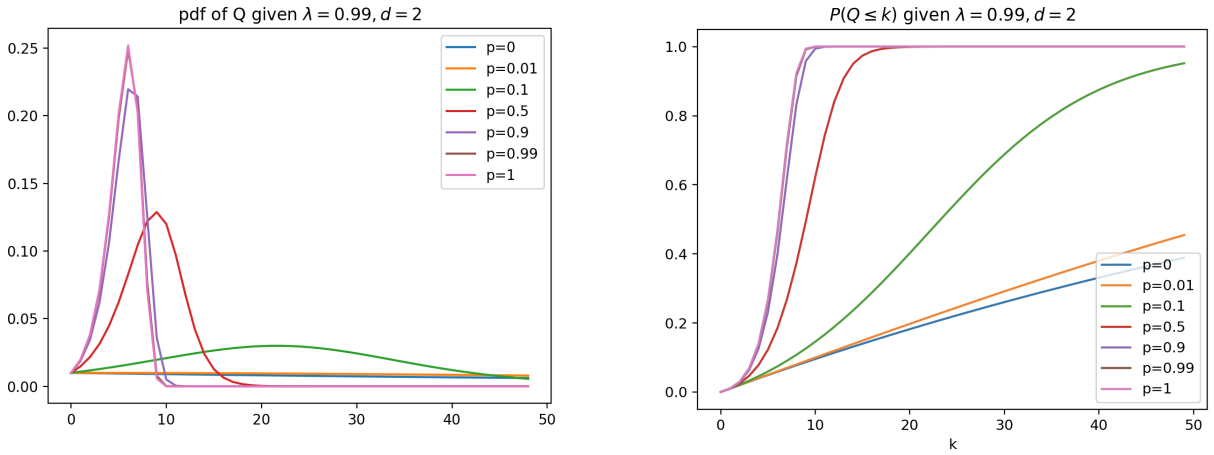

Figure 4: The pdf and cdf of the queue length with $d = 2$, $\lambda = .99$, and $p \in \{0, .01, .1, .5.9, .99, 1\}$.

Now that we have proved both fluid and diffusion limits for the queue length process, we can apply those results to some numerical examples. In Figure 3, we provide five plots where the flexibility parameter p changes throughout each plot. From Figure 3, we observe that p has a large effect on the shape of the distribution. In fact, by increasing p, the distribution develops an inflection point. Moreover, we observe that by having ten percent of flexible customers reduces the max queue length by an order of 10. As one continues to increase p, the max queue length decreases, but not as much as the initial few flexible customers.

In Figure 4, we plot the probability density function and the cdf of the queue length for a variety of values of p. On the left of Figure 4, we observe that as we increase p, the pdf mode moves to the left. Moreover, as p decreases, the pdf becomes more flat. On the right of Figure 4, we see that flat behavior of systems with small p is confirmed since the cdf of the queue length appears to have a linear shape.

5 Insights on Dependence on p and d

In this section, we provide new insights about our model with flexible customers. To this end, we prove two new results and also provide numerical experiments that validate our fluid and diffusion approximations. The first result shows that we can obtain a closed form solution for the tail cdf of the queue length distribution. The second result proves upper and lower bounds on the mean, second moment, and variance of the queue length process. We first start with a closed form solution of the steady state tail cdf.

5.1 Steady State Fluid Limit Solution

The steady state of the fluid limit admits a unique closed-form solution for the tail cdf. In order to show this result, we exploit a similar argument used by Rabinovich et al. [36].

Proposition 5.1 (Closed-form Solution of the Steady State)**.** *The steady state solution of the queueing model s^I satisfies a nonlinear recursion*

$$s_i^I = \lambda(1-p)s_{i-1}^I + \lambda p(s_{i-1}^I)^d \quad \textit{for all } i \geq 2.$$

$$\lambda(1-p)\left(s_{i-1}^I - s_i^I\right) + \lambda p\left(\left(s_{i-1}^I\right)^d - \left(s_i^I\right)^d\right) - \left(s_i^I - s_{i+1}^I\right) = 0$$

and has a unique closed-form solution given by

$$s_i^I = \sum_{k_1=1}^{d}\sum_{k_2=k_1}^{dk_1}\sum_{k_3=k_2}^{dk_2}\cdots\sum_{k_i=k_{i-1}}^{dk_{i-1}}\binom{1}{\frac{k_1-1}{d-1}}\binom{k_1}{\frac{k_2-k_1}{d-1}}\cdots\binom{k_{i-1}}{\frac{k_i-k_{i-1}}{d-1}}(\lambda(1-p))^{1+k_1+k_2+\cdots+k_{i-1}-\frac{k_i-1}{d-1}}(\lambda p)^{\frac{k_i-1}{d-1}}. \tag{5.72}$$

Proof. See Appendix for proof. □

Now that we have a closed form expression for the tail cdf of the queue length process, this result allows us to write the expected queue length $\mathbb{E}[Q]$ explicitly as well.

$$\begin{aligned}
&\mathbb{E}[Q]\\
=&\sum_{i=1}^{\infty}\sum_{k_1=1}^{d}\sum_{k_2=k_1}^{dk_1}\sum_{k_3=k_2}^{dk_2}\cdots\sum_{k_i=k_{i-1}}^{dk_{i-1}}\binom{1}{\frac{k_1-1}{d-1}}\binom{k_1}{\frac{k_2-k_1}{d-1}}\cdots\binom{k_{i-1}}{\frac{k_i-k_{i-1}}{d-1}}(\lambda(1-p))^{1+k_1+k_2+\cdots+k_{i-1}-\frac{k_i-1}{d-1}}(\lambda p)^{\frac{k_i-1}{d-1}}\\
\approx&\sum_{i=1}^{i^*}\sum_{k_1=1}^{d}\sum_{k_2=k_1}^{dk_1}\sum_{k_3=k_2}^{dk_2}\cdots\sum_{k_i=k_{i-1}}^{dk_{i-1}}\binom{1}{\frac{k_1-1}{d-1}}\binom{k_1}{\frac{k_2-k_1}{d-1}}\cdots\binom{k_{i-1}}{\frac{k_i-k_{i-1}}{d-1}}(\lambda(1-p))^{1+k_1+k_2+\cdots+k_{i-1}-\frac{k_i-1}{d-1}}(\lambda p)^{\frac{k_i-1}{d-1}}
\end{aligned}$$

where i^* is the smallest x such that $\mathbb{P}(Q \geq x) < \epsilon$.

In Figure 5, we provide a table of mean queue lengths as a function of the flexibility parameter p and the choice parameter d. We observe that for $d = 2$, the mean queue is decreased by 30% by having 1% of the customers be flexible and a 75% reduction in mean queue length for 10 % of the customers being flexible. Thus, just a small amount of flexibility can go a long way. We also observe that these dramatic improvements are only strengthened when we increase the choice parameter d.

To study the impact of the flexibility and choice parameters on the fluctuations, we provide a table in Figure 5 that describes the variance of the queue length as a function of the flexibility parameter p and the choice parameter d. We observe that for $d = 2$, the variance of the queue length is decreased by 65% by having 1% of the customers be flexible and a 97% reduction in variance queue length for 10 % of the customers being flexible. Thus, the reduction in variance is even better than the mean. Once again just a small amount of flexibility can significantly impact the performance of the system. We also observe for the variance that performance improvements increase when we increase the choice parameter d.

5.2 First and Second Moment Bounds

In this section, we prove upper and lower bounds for the mean, variance, and second moment of the queue length. We show numerically, that these bounds (especially the lower bounds) are quite accurate at approximating the queue length dynamics.

Proposition 5.2 (Moment Estimates)**.** *Let* $\mathbb{E}[Q]$ *denote the expected queue length, then*

$$\frac{\lambda\left(1+\frac{p\lambda^d}{1-\lambda^d(1-p)^d}\right)}{1-\lambda+\lambda p} < \mathbb{E}[Q] < \frac{\lambda\left(1+p\lambda^d\left(\frac{1-p}{1-\lambda^d}+\frac{p}{1-\lambda^{d^2}}\right)\right)}{1-\lambda+\lambda p}. \tag{5.73}$$

	p										
d	0	0.01	0.05	0.1	0.2	0.5	0.8	0.9	0.95	0.99	1
2	99	69	36	24	15	8	6	6	6	5	5
3	99	62	28	18	11	6	4	4	4	4	4
4	99	59	25	15	9	5	4	3	3	3	3
5	99	57	23	14	8	4	3	3	3	3	3
10	99	53	20	12	7	3	3	2	2	2	2
20	99	51	18	10	6	3	2	2	2	2	2
50	99	50	17	10	5	3	2	2	2	2	2
100	99	50	17	9	5	2	2	1	1	1	1

Figure 5: Mean queue length for various values of p and d given $\lambda = 0.99$

	p										
d	0	0.01	0.05	0.1	0.2	0.5	0.8	0.9	0.95	0.99	1
2	9890	3397	544	186	57	11	4	4	3	3	3
3	9890	2849	367	114	32	5	2	2	1	1	1
4	9890	2678	319	96	26	4	1	1	1	1	1
5	9890	2603	299	88	23	3	1	1	1	1	1
10	9890	2506	274	79	20	3	1	1	0	0	0
20	9890	2482	268	76	19	2	1	0	0	0	0
50	9890	2476	266	76	19	2	1	0	0	0	0
100	9890	2474	266	75	19	2	1	0	0	0	0

Figure 6: The variance of the queue length for various values of p and d given $\lambda = 0.99$

Let $\mathbb{E}[Q^2]$ denote the the second moment of queue length, then

$$\mathbb{E}[Q]^2 > \frac{\frac{2\lambda^{d+1}p}{(1-\lambda^d(1-p)^d)^2} + (1+\lambda(1-p))\frac{\lambda\left(1+\frac{p\lambda^d}{1-\lambda^d(1-p)^d}\right)}{1-\lambda+\lambda p}}{1-\lambda+\lambda p}, \tag{5.74}$$

$$\mathbb{E}[Q^2] < \frac{2\lambda^{d+1}\left(\frac{1-p}{(1-\lambda^d)^2} + \frac{p}{(1-\lambda^{d^2})^2}\right) + (1+\lambda(1-p))\frac{\lambda\left(1+p\lambda^d\left(\frac{1-p}{1-\lambda^d}+\frac{p}{1-\lambda^{d^2}}\right)\right)}{1-\lambda+\lambda p}}{1-\lambda+\lambda p}. \tag{5.75}$$

Moreover, let W be the patient waiting time, then

$$\frac{1+\frac{p\lambda^d}{1-\lambda^d(1-p)^d}}{1-\lambda+\lambda p} < \mathbb{E}[W] < \frac{1+p\lambda^d\left(\frac{1-p}{1-\lambda^d}+\frac{p}{1-\lambda^{d^2}}\right)}{1-\lambda+\lambda p}. \tag{5.76}$$

In Figure 8, we plot $\mathbb{E}[Q], \mathbb{E}[Q^2], \mathrm{Var}[Q]$ as well as their upper and lower bounds obtained from Proposition 5.2. We note that our upper and lower bounds are quite accurate at approximating the moment behavior as a function of the flexibility parameter p. In Figure 7, we observe that the compare the wait times of dedicated patients, flexible patients, the average patients, and the system where all the flexible patients are not present. It is clear from Figure 7, that the average wait time decreases by adding flexible patients and the flexible patients benefit themselves from being flexible. Throughout our analysis, one might be tempted to approximate the queue length and waiting time with a model where the flexible patients disappeared i.e. a non-flexible queue where the arrival rate is $\lambda(1-p)$. However, we observe that the wait time is very much under-approximated if one pretends

the flexible patients are not there. Thus, it is still important to capture the flexible patients and they cannot be simply ignored from the performance analysis.

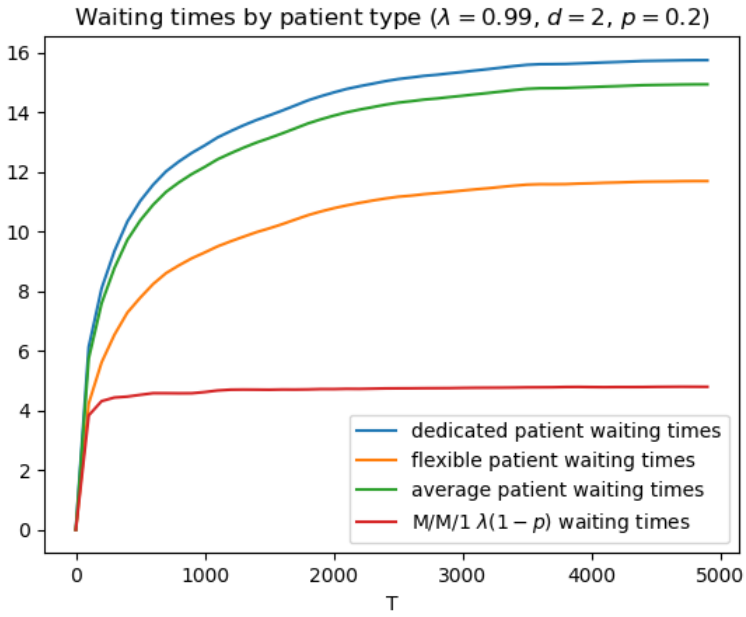

Figure 7: The waiting times by patient type given $\lambda = 0.99, d = 2, p = 0.2$

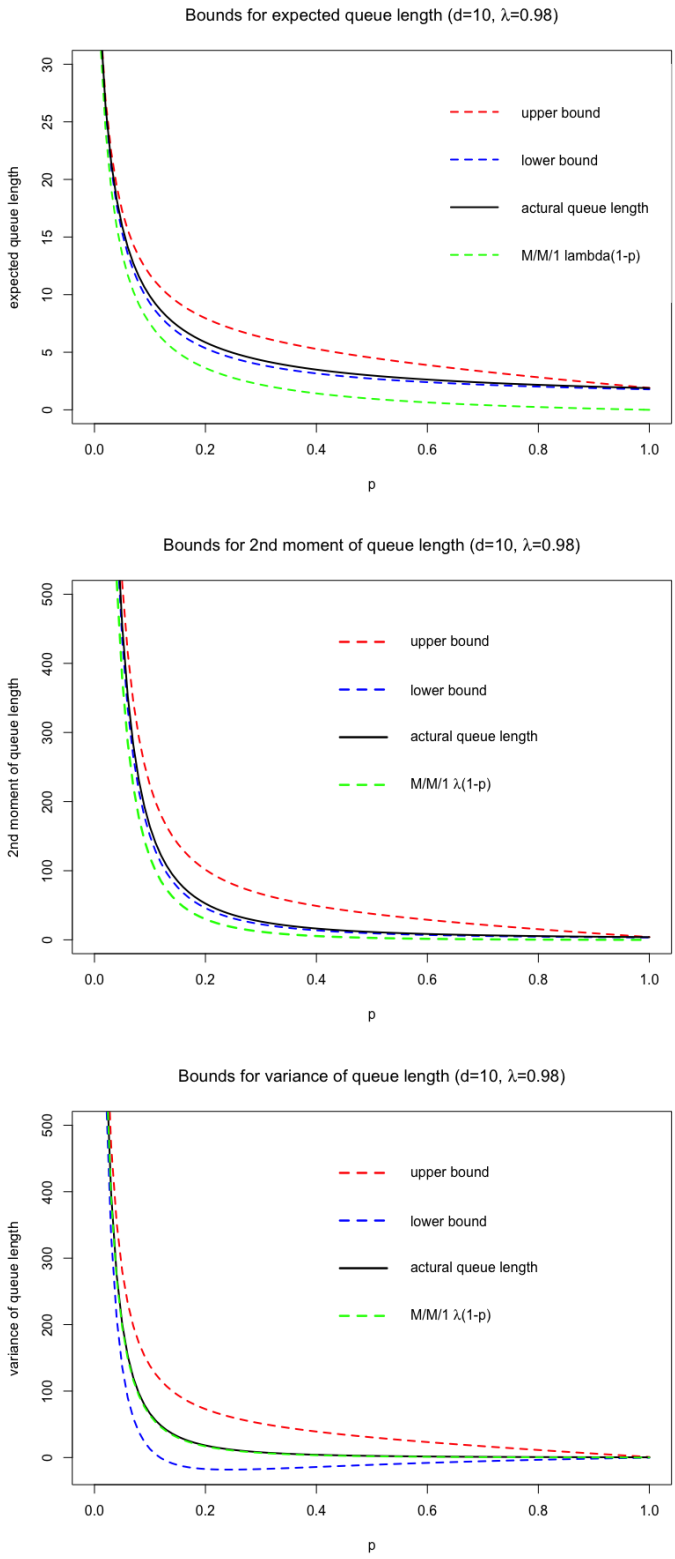

Figure 8: Upper and lower bounds for mean (top), 2nd moment (middle), and variance (bottom).

6 Conclusion

In this paper, we construct a stochastic queueing model that captures the performance trade-off between customers valuing flexibility (or time) vs. customers wanting dedicated services, through setting a fraction p of all customers to be flexible via joining the shortest of d queues. First, we prove the fluid model results in both transient and steady-state behaviors. We show that the scaled queue-length process converges to a unique fluid trajectory on any finite time interval, and that this fluid trajectory converges to a unique steady state s^I, for which a closed-form expression is obtained. We also show that the steady state distribution of the N-physicians system concentrates on s^I as N goes to infinity. Second, we prove the diffusion model results in both transient and steady-state behaviors. We shows that the scaled diffusion process converges to a unique Ornstein-Uhlenbeck process, and that the interchanging of limits $\lim_{t\to\infty}\lim_{N\to\infty} = \lim_{N\to\infty}\lim_{t\to\infty}$ holds for the diffusion limit in equilibrium. Finally, we prove an upper and lower bound for the first and second moment of the expected queue length of the system, and show through numerical examples that having just a small fraction of flexible customers can benefit the system tremendously, both in lowering the mean queue length as well as its variance.
Despite our analysis, there are many future directions for research.

1. The first direction would be to generalize the arrival rate of dedicated patients to each physician to be non-uniform, i.e. taking into account the popularity of different physicians.

2. A second direction would be to generalize our results for non exponential arrival and service distributions, like in the work of Aghajani et al. [3], Bramson et al. [5].

3. It would also be interesting to generalize the work to system of $M/M/c$ queues or system of $M/M/\infty$ queues, and derive new limit theorems in those regimes. One could also incorporate the impact of delayed information to model delays in communicating the queue length to customers. Recent work by Nirenberg et al. [34], Novitzky et al. [35] could help in this regard.

4. Finally, there is recent work that analyzes self-exciting point processes as arrival processes to queues, see for example Castellanosa et al. [8], Chen [9], Daw and Pender [11, 12], Gao and Zhu [19], Koops et al. [27]. It would be interesting to analyze similar JSQ models with Hawkes arrival processes.

We intend to pursue these extensions in future work.

7 Online Appendix

7.1 Proof of Proposition 3.2

Proof. By the identity $u^d - v^d = (u-v)(u^{d-1} + u^{d-2}v + \cdots + v^{d-1}) \leq d(u-v)$, we have the Lipschitz bound

$$\begin{aligned}
\|F_+(u) - F_+(v)\|_{\ell_2}^2 &\leq \sum_{i=1}^{\infty} \big(\|\lambda(1-p)(u_{i-1} - v_{i-1})\|^2 + \|\lambda(1-p)(u_i - v_i)\|^2 \\
&\quad + \|\lambda p(u_{i-1}^d - v_{i-1}^d)\|^2 + \|\lambda p(u_i^d - v_i^d)\|^2 \big) \\
&\leq 4 \sum_{i=0}^{\infty} \left[\lambda^2 (1-p)^2 \|u_i - v_i\|^2 + (\lambda p d)^2 \|u_i - v_i\|^2 \right] \\
&\leq 8\lambda^2 d^2 \|u - v\|_{\ell_2}^2.
\end{aligned}$$

Similarly we can show that

$$\|F_-(u) - F_-(v)\|_{\ell_2}^2 \leq 2\|u - v\|_{\ell_2}^2.$$

Thus, the mappings F, F_+, F_- are Lipschitz with respect to the ℓ_2 norm. □

7.2 Proof of Theorem 3.6

Proof. It is sufficient to show that the conclusion $\lim_{t\to\infty} s(t) = s^I$ holds for any $s(0) \in \mathcal{S}$ for which either $s(0) \leq s^I$ or $s(0) \geq s^I$, since Lemma 3.5 implies that

$$s(t, \min[s(0), s^I]) \leq s(t, s(0)) \leq s(t, \max[s(0), s^I])$$

where we use $s(t,u)$ to denote the solution to Equation (3.5) with initial condition u.

Since the derivative of $s_k(t, s(0))$ is bounded for all k, the convergence of $s(t, s(0)) \to s^I$ will follow from

$$\int_0^{\infty} [s_k(u, s(0)) - s_k^I] du < \infty, \quad \text{where } s(0) \geq s^I \tag{7.77}$$

and from

$$\int_0^{\infty} [s_k^I - s_k(u, s(0))] du < \infty, \quad \text{where } s(0) \leq s^I. \tag{7.78}$$

The proof is similar for both cases so here we only discuss (7.77).

Define $v_k(s(t)) = \sum_{i=k}^{\infty} s_k(t)$, and $v(s) = \{v_i(s)\}_{i=0}^{\infty}$. Then we have for any $k \in \mathbb{N}$ and fix $t \geq 0$,

$$0 \leq v_k(s(t)) \leq v_1(s(t)) = \sum_{i=1}^{\infty} s_i(t) < \infty.$$

We also know that

$$\frac{d(v_1(s(t)) - v_1(s^I))}{dt} = \lambda(1-p)s_0 + \lambda p s_0^d - s_1(t) = \lambda - s_1(t) = s_1^I - s_1(t) \leq 0, \tag{7.79}$$

which implies that $v_1(s(t))$ does not increase with t. Thus, $v_1(s(t))$ is uniformly bounded for all $t \geq 0$. Notice that

$$\begin{aligned}\frac{dv_k(s(t))}{dt} &= \lambda(1-p)s_{k-1}(t) + \lambda p s_{k-1}(t)^d - s_k(t) \\ &= \lambda(1-p)(s_{k-1}(t) - s_{k-1}^I) + \lambda p((s_{k-1}(t))^d - (s_{k-1}^I)^d) - (s_k(t) - s_k^I), \quad (7.80)\end{aligned}$$

which implies that

$$\begin{aligned}& v_k(s(t)) - v_k(s(0)) \\ = & \int_0^t \left[\lambda(1-p)(s_{k-1}(u) - s_{k-1}^I) + \lambda p((s_{k-1}(u))^d - (s_{k-1}^I)^d) - (s_k(u) - s_k^I)\right] du. \quad (7.81)\end{aligned}$$

By the uniform boundedness of $v_k(s(t)) - v_k(s(0))$, we know that

$$\int_0^\infty \left[\lambda(1-p)(s_{k-1}(u) - s_{k-1}^I) + \lambda p((s_{k-1}(u))^d - (s_{k-1}^I)^d) - (s_k(u) - s_k^I)\right] du < \infty.$$

Using an induction argument, we can assume that the integral converges for all $i \leq k-1$, i.e.

$$\int_0^\infty (s_i(t) - s_i^I) dt < \infty, \quad i \leq k-1.$$

Then for $i = k$, again by the uniform boundedness of $v_k(s(t)) - v_k(s(0))$ we have that

$$\int_0^\infty (s_k(t) - s_k^I) dt < \infty,$$

which completes the proof of global stability of the fluid limit $\lim_{t\to\infty} s(t) = s^I$ for any initial condition $s(0)$. □

7.3 Proof of Proposition 3.7

Proof. Let τ be a jump time of the Poisson processes used for arrivals and departures. We first compare system p with system 0. Our goal is to show that assuming

$$c_m^{N,p}(\tau-) \leq c_m^{N,0}(\tau-), \quad m \in \mathbb{N}, \tag{7.82}$$

then we have

$$c_m^{N,p}(\tau-) \leq c_m^{N,0}(\tau), \quad m \in \mathbb{N}.$$

Since we know that

$$c_m^{N,\sigma}(t) = c_{m+1}^{N,\sigma}(t) + N S_{m+1}^{N,\sigma}(t), \quad m \geq 0, t \geq 0.$$

Applying (7.89) to $m = n-1$ and $m = n$ implies

$$c_n^{N,1}(\tau-) = c_n^{N,0}(\tau-) \Rightarrow S_n^{N,1}(\tau-) \leq S_n^{N,0}(\tau-) \text{ and } S_{n+1}^{N,1}(\tau-) = S_{n+1}^{N,0}(\tau-).$$

When τ represent a departure time, let x^σ denote the respective lengths of the queue chosen for potential departure. A patient will depart from the system σ if and only if $x^\sigma > 0$, and there will be one less patient with exactly $x^\sigma - 1$ patients in front of them, therefore

$$c_m^{N,\sigma}(\tau) = c_m^{N,\sigma}(\tau-) - 1, \quad m < x^\sigma, \tag{7.83}$$

and

$$c_m^{N,\sigma}(\tau) = c_m^{N,\sigma}(\tau-), \quad m \geq x^\sigma. \tag{7.84}$$

Assume that there exists $n \geq 0$ such that $c_n^{N,p}(\tau) > c_n^{N,0}(\tau)$, then (7.89), (7.83) and (7.84) imply that it is true if and only if

$$c_n^{N,p}(\tau) = c_n^{N,1}(\tau), \quad x^p \leq n < x^0. \tag{7.85}$$

Now let $j \in \{1, \cdots, N\}$ denotes the rank in decreasing order chosen for departures, then

$$NS_{x^\sigma+1}^{N,\sigma}(\tau-) < j \leq NS_{x^\sigma}^{N,\sigma}(\tau-)$$

which yields in particular that

$$S_{x^1+1}^{N,p}(\tau-) < S_{x^1}^{N,p}(\tau-) \leq S_{x^0}^{N,0}(\tau-).$$

Then combining (7.83), (7.84) and (7.85) yields

$$S_{n+1}^{N,0}(\tau-) \leq S_{n+1}^{N,p}(\tau-) \leq S_{x^1+1}^{N,p}(\tau-) < S_{x^0}^{N,0}(\tau-) \leq S_{n+1}^{N,0}(\tau-)$$

which is a contradiction. Thus $c_m^{N,p}(\tau) \leq c_m^{N,0}(\tau), \quad m \in \mathbb{N}$ holds.

When τ represent an arrival time, let x^σ denote the respective lengths of the queues chosen for either patient. There is a new patient in either system with x^σ patients in front of him, therefore

$$c_m^{N,\sigma}(\tau) = c_m^{N,\sigma}(\tau-) + 1, \quad m \leq x^\sigma \tag{7.86}$$

and

$$c_m^{N,\sigma}(\tau) = c_m^{N,\sigma}(\tau-), \quad m > x^\sigma \tag{7.87}$$

Assume that there exists $n \geq 0$ such that $c_n^{N,p}(\tau) > c_n^{N,0}(\tau)$, then (7.89), (7.86) and (7.87) imply that it is true if and only if

$$c_n^{N,0}(\tau) = c_n^{N,p}(\tau), \quad x^0 \leq n < x^p \tag{7.88}$$

Now let $j^\sigma \in \{1, \cdots, N\}$ denotes the rank in decreasing order of the queue joined by the patient in system σ, then

$$NS_{x^\sigma+1}^{N,\sigma}(\tau-) < j^\sigma \leq NS_{x^\sigma}^{N,\sigma}(\tau-)$$

which yields in particular that

$$S_{x^0+1}^{N,0}(\tau-) < j^0 \leq j^p \leq S_{x^p}^{N,p}(\tau-).$$

Then combining (7.86), (7.87) and (7.88) yields

$$S_n^{N,p}(\tau-) \leq S_n^{N,0}(\tau-) \leq S_{x^0+1}^{N,0}(\tau-) < S_{x^p}^{N,p}(\tau-) \leq S_n^{N,p}(\tau-)$$

which is a contradiction. Thus $c_m^{N,p}(\tau) \le c_m^{N,0}(\tau), \quad m \in \mathbb{N}$ holds.

Similar techniques apply to the case of comparing system 0 and system p, and we have that if

$$c_m^{N,1}(\tau-) \le c_m^{N,p}(\tau-), \quad m \in \mathbb{N}, \tag{7.89}$$

then

$$c_m^{N,1}(\tau-) \le c_m^{N,p}(\tau), \quad m \in \mathbb{N}.$$

□

7.4 Proof of Theorem 4.2

Proof. Because of the Lipschitz property of mappings F_+, F_-, we have

$$\|M(t)\|_{\ell_2} = \int_0^t \|F_+(s(u)) + F_-(s(u))\|_{\ell_2} du \le \int_0^t (2\sqrt{2}\lambda d + \sqrt{2})\|s(u)\|_{\ell_2} du$$

By Gronwall's lemma, we know that $\|s(u)\|_{\ell_2}$ is uniformly bounded on $0 \le u \le t$. Thus, $M(t)$ is square-integrable in ℓ_2.

□

7.5 Proof of Theorem 4.3

Proof. Consider $s \in \mathcal{S}$, we have

$$\begin{aligned}
\|K(s)x\|_{\ell_2}^2 &= \sum_{k\ge1} \left[\lambda(1-p)(x_{k-1}-x_k) + \lambda pd\left(s_{k-1}^{d-1}x_{k-1} - s_k^{d-1}x_k\right) + x_k - x_{k+1}\right]^2 \\
&\le \sum_{k\ge1} \left((\lambda(1-p)+\lambda pd)^2 + (\lambda(1-p)+\lambda pd+1)^2+1)^2 + 1^2\right)(x_{k-1}^2 + x_k^2 + x_{k+1}^2) \\
&\le 6(\lambda(1-p)+\lambda pd+1)^2\|x\|_{\ell_2}^2.
\end{aligned}$$

Then (1) follows. For (2), since the martingale $M(t)$ is square-integrable in ℓ_2 by Theorem 4.2, if $\mathbb{E}\left(\|D(0)\|_{\ell_2}^2\right) < \infty$, then the formula (4.31) for $D(t)$ is well-defined, solves the SDE, and using Gronwall's lemma yields $\mathbb{E}\left(\sup_{t\le T}\|D(t)\|_{\ell_2}^2\right) < \infty$.

□

7.6 Proof of Lemma 4.7

Proof. Since

$$\frac{(Na)_d}{(N)_d} = \prod_{i=0}^{d-1} \frac{Na-i}{N-i} \tag{7.90}$$

$$= \prod_{i=0}^{d-1} \left(a + (a-1)\frac{i}{N-i}\right) \tag{7.91}$$

$$= \sum_{j=1}^{d-1} a^{d-j}(a-1)^j \prod_{1\le i_1<\cdots<i_j\le d-1} \frac{i_1\cdots i_j}{(N-i_1)\cdots(N-i_j)} \tag{7.92}$$

It is obvious that $A^N(a)$ is $\frac{1}{N}O(a)$. For $a = \frac{k}{N}$ where $k = 0, 1, \cdots, N$,

$$\begin{aligned} \prod_{i=0}^{d-1} \frac{Na - i}{N - i} &= \prod_{i=0}^{d-1} \frac{k - i}{N - i} && (7.93) \\ &\leq \prod_{i=0}^{d-1} \frac{k}{N} = a^d && (7.94) \end{aligned}$$

The inequality comes from the fact that each term $\frac{k-i}{N-i}$ is either bounded by a or the product contains a term exactly equal to 0. Thus $A^N(k/N) \leq 0$. □

7.7 Proof of Lemma 4.8

Proof. For $a, a + h \in [0, 1]$,

$$B(a, h) \leq h^d + \sum_{i=2}^{d-1} ah^2 = h^d + (2^d - d - 2)ah^2$$

□

7.8 Proof of Theorem 4.5

Proof. Take expectation on both sides of Equation (4.31), since

$$\mathbb{E}\left[\int_0^t e^{\int_s^t K(s(u))du} dM(s)\right] = 0,$$

we have

$$\mathbb{E}[D(t)] = e^{\int_0^t \mathcal{A}(s)ds}\mathbb{E}[D(0)].$$

Therefore

$$D(t) - \mathbb{E}[D(t)] = e^{\int_0^t \mathcal{A}(s)ds}(D(0) - \mathbb{E}[D(0)]) + \int_0^t e^{\int_s^t \mathcal{A}(u)du} dM(s), \quad (7.95)$$

and

$$\begin{aligned} \Sigma(t) &= E[(D(t) - \mathbb{E}[D(t)])(D(t) - \mathbb{E}[D(t)])^\top] \\ &= e^{\int_0^t \mathcal{A}(s)ds} E[(D(0) - \mathbb{E}[D(0)])(D(0) - \mathbb{E}[D(0)])^\top] \left(e^{\int_0^t \mathcal{A}(s)ds}\right)^\top \\ &+ \left(\int_0^t e^{\int_s^t \mathcal{A}(u)du} dM(s)\right)\left(\int_0^t e^{\int_s^t \mathcal{A}(u)du} dM(s)\right)^\top \\ &= e^{\int_0^t \mathcal{A}(s)ds}\Sigma(0)e^{\int_0^t \mathcal{A}^\top(s)ds} + \int_0^t e^{\int_s^t \mathcal{A}(u)du}\mathcal{B}(s)e^{\int_s^t \mathcal{A}^\top(u)du} ds. \quad (7.96) \end{aligned}$$

□

7.9 Proof of Lemma 4.11

Proof. We have the unique solution $D(t)$ as

$$\begin{aligned} D(t) &= e^{\mathcal{K}t}D(0) + \int_0^t e^{\mathcal{K}(t-s)}dB(s) \\ &= \mathrm{diag}(\pi)\int_0^\infty e^{-xt}Q(x)Q(x)^*\psi(dx)D(0) + \int_0^t \left(\mathrm{diag}(\pi)\int_0^\infty e^{-x(t-s)}Q(x)Q(x)^*\psi(dx)\right)dB(s) \\ &= \mathrm{diag}(\pi)\int_0^\infty e^{-xt}Q(x)^*D(0)Q(x)\psi(dx) + diag(\pi)\int_0^\infty e^{-xt}Q(x)^*\left(\int_0^t e^{xs}dB(s)\right)Q(x)\psi(dx) \\ &= \mathrm{diag}(\pi)\int_S e^{-xt}Q(x)^*\left(D(0) + \int_0^t e^{xs}dB(s)\right)Q(x)\psi(dx). \end{aligned} \tag{7.97}$$

Note that here we define $Q(x) = (Q_1(x), Q_2(x), \cdots, Q_n(x), \cdots)^\top$, which is a infinite dimensional column vector of polynomials. Thus $Q(x)^*D(0)$ and $Q(x)^*dB(s)$ are 1-dimensional numbers and are exchangeable with $Q(x)$ in matrix multiplication.

For the equilibrium point $D(\infty)$ of the OU process, we have

$$\begin{aligned} D(\infty) &= \int_0^\infty e^{\mathcal{K}t}dB(t) \\ &= \int_0^\infty \left(\mathrm{diag}(\pi)\int_0^\infty e^{-xt}Q(x)Q(x)^*\psi(dx)\right)dB(t) \\ &= \mathrm{diag}(\pi)\int_S \left(Q(x)^*\int_0^\infty e^{-xt}dB(t)\right)Q(x)\psi(dx), \end{aligned} \tag{7.98}$$

and its covariance matrix $\Sigma(\infty)$ is as follows,

$$\begin{aligned} \Sigma(\infty) &= \int_0^\infty e^{\mathcal{K}t}\mathbb{E}[B(1), B(1)^*]e^{\mathcal{K}^*t}dt \\ &= \int_0^\infty e^{\mathcal{K}t}\mathrm{diag}(v)e^{\mathcal{K}^*t}dt \\ &= \int_0^\infty \left[\int_S \left(\mathrm{diag}(\pi)e^{-xt}Q(x)Q(x)^*\psi(dx)\right)\mathrm{diag}(v)\int_S \left(\mathrm{diag}(\pi)e^{-yt}Q(y)Q(y)^*\psi(dy)\right)\right]dt \\ &= \mathrm{diag}(\pi)\int_{S^2}\left(\int_0^\infty e^{-(x+y)t}dt\right)Q(x)(Q(x)^*\mathrm{diag}(v)Q(y))Q(y)^*\psi(dx)\psi(dy)\mathrm{diag}(\pi) \\ &= \mathrm{diag}(\pi)\int_{S^2}\frac{Q(x)^*\mathrm{diag}(v)Q(y)}{x+y}Q(x)Q(y)^*\psi(dx)\psi(dy)\mathrm{diag}(\pi). \end{aligned} \tag{7.99}$$

□

7.10 Proof of Theorem 4.13

Proof. We first prove (1). We can assume WLOG that $\hat{\lambda}_k > [\lambda(1-p) + \lambda pd(s_k^I)^{d-1}] + \lambda p(1 + (2^d - d - 2)s_k^I)$ for $k \geq 1$. Since $z(t) = e^{\hat{A}^*t}z(0)$ depends continuously on $z(0)$ in ℓ_1^0, we may

assume $h(0) > 0$. Let $\tau = \inf\{t \geq 0 : \{k \geq 1 : h_k(t) = 0\} = \emptyset\}$ be the first time when $h_k = 0$ for some $k \geq 1$. We know that $\tau > 0$ and the result holds when $\tau = \infty$.

If $\tau < \infty$, we have

$$\begin{aligned}\dot{h}_k(\tau) &= \hat{\lambda}_{k-1} y_{k-1}(\tau) - [\lambda(1-p) + \lambda p d (s^I_{k-1})^{d-1}] y_{k-1}(\tau) - \lambda p B\left(s^I_{k-1}, y_{k-1}(\tau)\right) \\ &\quad + \hat{\lambda}_{k-1}(z_{k-1}(\tau) - y_{k-1}(\tau)) - (z_k(\tau) - y_k(\tau)). \end{aligned} \tag{7.100}$$

Lemma 3.5 and $y(0) \geq 0$ implies that $y(t) \geq 0$ for all $t \geq 0$. Any by Lemma 4.8 we have that

$$B\left(s^I_{k-1}, y_{k-1}(\tau)\right) \leq y^d_{k-1} + (2^d - d - 2) s^I_{k-1} y^2_{k-1} \leq \left(1 + (2^d - d - 2) s^I_{k-1}\right) y_{k-1}.$$

Therefore by the assumption that $\hat{\lambda}_k \geq [\lambda(1-p) + \lambda p d (s^I_k)^{d-1}] + \lambda p (1 + (2^d - d - 2) s^I_k)$ we have that

$$\hat{\lambda}_{k-1} y_{k-1}(\tau) - [\lambda(1-p) + \lambda p d (s^I_{k-1})^{d-1}] y_{k-1}(\tau) - \lambda p B\left(s^I_{k-1}, y_{k-1}(\tau)\right) \geq 0,$$

and equality holds only when $y_{k-1} = 0$. For $k \in \mathcal{Z} = \{k \geq 1 : h_k(\tau) = 0\}$ we have

$$z_{k-1}(\tau) - y_{k-1}(\tau) = h_{k-1}(\tau) - h_k(\tau) = h_{k-1}(\tau) \geq 0,$$

$$z_k(\tau) - y_k(\tau) = h_k(\tau) - h_{k+1}(\tau) = -h_{k+1}(\tau) \leq 0,$$

hence $\dot{h}_k(\tau) \geq 0$ with equality only when $k - 1 \in \mathcal{Z} \cup \{0\}$ and $k + 1 \in \mathcal{Z}$. We also know that $h_k(t) > 0$ for $t < \tau$ and $h_k(\tau) = 0$ which implies $\dot{h}_k(\tau) \leq 0$. Thus $\dot{h}_k(\tau) = 0$ and $z_{k-1}(\tau) = y_{k-1}(\tau) = 0$ and $k - 1 \in \mathcal{Z} \cup \{0\}$ and $k + 1 \in \mathcal{Z}$. By induction we have that $z_k(\tau) = y_k(\tau) = 0$ for all $k \geq 1$, which means $z(t) = y(t) = 0$ for all $t \geq \tau$, thus $h(t) \geq 0$ for $t \geq 0$. The proof for (2) follows similarly. □

7.11 Proof of Theorem 4.14

Proof.

$$\begin{aligned} & \|(x_k + x_{k+1} + \cdots)_{k\geq 1}\|_{L_2(g_\theta)} \\ = & \sum_{k\geq 1} (x_k + x_{k+1} + \cdots)^2 \theta^{-k} \\ \leq & \sum_{k\geq 1} n\left(x_k^2 + x_{k+1}^2 + \cdots + x_{k+n-2}^2 + (x_{k+n-1} + x_{k+n} + \cdots)^2\right)\theta^{-k} \\ \leq & \; n(1 + \theta + \cdots + \theta^{n-2}) \sum_{k\geq 1} x_k^2 \theta^{-k} + n\theta^{n-1} \sum_{k\geq 1} (x_k + x_{k+1} + \cdots)^2 \theta^{-k}. \end{aligned} \tag{7.101}$$

Since this holds for any $n \geq 1$ we can choose n large enough such that $n\theta^{n-1} < 1$, then

$$(1 - n\theta^{n-1}) \|(x_k + x_{k+1} + \cdots)_{k\geq 1}\|_{L_2(g_\theta)} \leq n(1 + \theta + \cdots + \theta^{n-2}) \|x\|_{L_2(g_\theta)}. \tag{7.102}$$

Let $K_\theta = n(1 + \theta + \cdots + \theta^{n-2})/(1 - n\theta^{n-1})$ we proved the lemma. □

7.12 Proof of Theorem 4.15

Proof. Assume $s(0) \in \mathcal{S}$ is in $L_2(g_\theta)$. Denote $s^+(0) = \max\{s(0), s^I\}$ and $s^-(0) = \min\{s(0), s^I\}$, and s^+, s^- as the corresponding solution to 3.5 with such initial condition. Then by the comparison lemma 3.5 we have that $s^+(t) \le s(t) \le s^-(t)$ and $s^+(t) \le s^I \le s^-(t)$ for all $t \ge 0$. Again we use $y(t) = s(t) - s^I$ to denote the solution to the recentered equation, and that $y^+(t) = s^+(t) - s^I, y^-(t) = s^-(t) - s^I$. We also have

$$|y(0)| = \max\{y(0)^+, y(0)^-\}, |y(t)| \le \max\{y(t)^+, -y(t)^-\}, t \ge 0.$$

Now consider a birth death process with generator $\hat{\mathcal{A}}$ where birth rate $\hat{\lambda}_k$ is as follows,

$$\hat{\lambda}_k = \max\{[\lambda(1-p) + \lambda pd(s_k^I)^{d-1}] + \lambda p(1 + (2^d - d - 2)s_k^I), \theta\}, k \ge 1.$$

For $\lambda \le \theta < 1$, we know that for large enough k $\hat{\lambda}_k$ is equal to θ. Using the same method as in the proof of theorem 4.12, we have that the spectral gap $\hat{\gamma}$ for the birth death process with generator $\hat{\mathcal{A}}$ satisfies that $0 < \hat{\gamma} \le \hat{\sigma} = (\sqrt{\theta} - 1)^2$. This means that the solution $z(t)$ to $\dot{z} = \hat{\mathcal{A}}^* z$ has exponential stability, i.e.

$$\|z(t)\|_{L_2(\hat{\pi})} \le e^{-\hat{\gamma}t} \|z(0)\|_{L_2(\hat{\pi})}, t \ge 0$$

where $\hat{\pi}$ is the stationary distribution to $\hat{\mathcal{A}}^*$.

We know that

$$\hat{\pi}_k = \prod_{i=1}^{k-1} \hat{\lambda}_k = \theta^{k-1} \prod_{i=1}^{k-1} \max\{\theta^{-1}[\lambda + \lambda pd(s_k^I)^{d-1} + \lambda p(2^d - d - 2)s_k^I], 1\} \ge \theta^{k-1}$$

and the product converges. Thus $\hat{\pi}_k = O(\theta^k)$ and $\theta^k = O(\hat{\pi}_k)$ and therefore the two norm $L_2(\hat{\pi})$ and $L_2(g_\theta)$ is equivalent, and we have that there exists $c, d > 0$ such that

$$\|z(t)\|_{L_2(g_\theta)} \le d\|z(t)\|_{L_2(\hat{\pi})} \le de^{-\hat{\gamma}t} \|z(0)\|_{L_2(\hat{\pi})} \le cde^{-\hat{\gamma}t} \|z(0)\|_{L_2(g_\theta)}.$$

Let z^+, z^- be the corresponding solutions to $z^+ = \hat{\mathcal{A}}^* z^+$ and $z^- = \hat{\mathcal{A}}^* z^-$ starting $y(0)^+ \ge 0$ and $y(0)^- \le 0$ respectively. Then by lemma 4.13 and lemma 4.14, we have

$$\begin{aligned} \|y^+(t)\|_{L_2(g_\theta)} &\le \|(y_k^+(t) + y_{k+1}^+(t) + \cdots)_{k \ge 1}\|_{L_2(g_\theta)} \\ &\le \|(z_k^+(t) + z_{k+1}^+(t) + \cdots)_{k \ge 1}\|_{L_2(g_\theta)} \\ &\le K_\theta \|z^+(t)\|_{L_2(g_\theta)} \\ &\le cdK_\theta e^{-\hat{\gamma}t} \|y^+(0)\|_{L_2(g_\theta)}, \end{aligned} \tag{7.103}$$

and similarly $\|y^-(t)\|_{L_2(g_\theta)} \le cdK_\theta e^{-\hat{\gamma}t} \|y^-(0)\|_{L_2(g_\theta)}$. Thus let $\gamma_\theta = \hat{\gamma}$ and $C_\theta = c^2 d^2 K_\theta^2$ we have

$$\begin{aligned} \|y(t)\|_{L_2(g_\theta)}^2 \le \|y^+(t)\|_{L_2(g_\theta)}^2 + \|y^-(t)\|_{L_2(g_\theta)}^2 &\le e^{-2\gamma_\theta t} C_\theta \left(\|y^+(0)\|_{L_2(g_\theta)}^2 + \|y^-(0)\|_{L_2(g_\theta)}^2 \right) \\ &= e^{-2\gamma_\theta t} C_\theta \|y(0)\|_{L_2(g_\theta)}^2. \end{aligned} \tag{7.104}$$

□

7.13 Proof of Theorem 4.16

Proof. We consider the case when $s(0) = s^I$. Since s^I is the equilibrium, we have $s(t) = s^I$ for all $t \geq 0$. Let $s(\nu, h)$ be the solution of Equation (3.5) at time $h \geq 0$ with initial value ν. For $t_0 \geq 0$ let $D^N(t_0, h) = \sqrt{N}(S^N(t_0 + h) - s(S^N(t_0), h))$. Then we have $D^N(t_0 + h) = D^N(t_0, h) + \sqrt{N}\left(s(S^N(t_0), h) - s^I\right)$. By Lemma 4.15,

$$\|D^N(t_0 + h)\|_{L_2(g_\theta)} \leq \|D^N(t_0, h)\|_2 + C_\theta e^{-\gamma_\theta h}\|D^N(t_0)\|_{L_2(g_\theta)}. \tag{7.105}$$

The conditional distribution of $D^N(t_0, h)$ given $S^N(t_0) = s$ is the distribution of D^N started with $S^N(t_0) = s(0) = s$. In particular, $D^N(t_0) = D^N(t_0, 0) = 0$.

Following a similar argument as in Equation (4.51), we have that there exists constant $C_T > 0$ such that

$$\begin{aligned}\sup_{0\leq h\leq T}\|D^N(t_0, h)\|_{L_2(g_\theta)} \;\leq\; & C_T\left(\frac{1}{\sqrt{N}}\|s^I\|_{L_2(g_\theta)} + \frac{1}{N}C_\theta\|D^N(t_0)\|_{L_2(g_\theta)}\right. \\ & \left. + \sup_{0\leq h\leq T}\left\|M^N(t_0 + h) - M^N(t_0)\right\|_{L_2(g_\theta)}\right).\end{aligned} \tag{7.106}$$

Combined with (7.105), we have that for some $L_T > 0$ and $0 \leq h \leq T$,

$$\mathbb{E}\left(\|D^N(t_0 + h)\|^2_{L_2(g_\theta)}\right) \leq L_T + 2\left(\frac{C_T}{N} + e^{-\gamma_\theta h}\right)^2 C_\theta^2\mathbb{E}\left(\|D^N(t_0)\|^2_{L_2(g_\theta)}\right). \tag{7.107}$$

Now for fixed T large enough we have $8e^{-2\gamma_\theta T}C_\theta^2 \leq \epsilon < 1$. Then uniformly for $N \geq C_T e^{\gamma_\theta T}$, for $m \in \mathbb{N}$, we have

$$\mathbb{E}\left(\|D^N((m + 1)T)\|^2_{L_2(g_\theta)}\right) \leq L_T + \epsilon\mathbb{E}\left(\|D^N(mT)\|^2_{L_2(g_\theta)}\right). \tag{7.108}$$

By induction,

$$\mathbb{E}\left(\|D^N(mT)\|^2_{L_2(g_\theta)}\right) \leq L_T\sum_{j=1}^{m}\epsilon^{j-1} + \epsilon^m\mathbb{E}\left(\|D^N(0)\|^2_{L_2(g_\theta)}\right) \leq \frac{L_T}{1 - \epsilon} + \mathbb{E}\left(\|D^N(0)\|^2_{L_2(g_\theta)}\right). \tag{7.109}$$

From (7.107), we know that

$$\sup_{0\leq h\leq T}\mathbb{E}\left(\|D^N(mT + h)\|^2_{L_2(g_\theta)}\right) \leq L_T + 8C_\theta^2\mathbb{E}\left(\|D^N(mT)\|^2_{L_2(g_\theta)}\right), \tag{7.110}$$

hence we have the infinite horizon bound

$$\sup_{t\geq 0}\mathbb{E}\left(\|D^N(t)\|^2_{L_2(g_\theta)}\right) \leq L_T + 8C_\theta^2\mathbb{E}\left(\frac{L_T}{1 - \epsilon} + \mathbb{E}\|D^N(0)\|^2_{L_2(g_\theta)}\right). \tag{7.111}$$

Ergodicity and the Fatou Lemma yield that for $D^N(\infty)$

$$\mathbb{E}(\|D^N(\infty)\|^2_{L_2(g_\theta)}) \leq \liminf_{t\geq 0}\mathbb{E}(\|D^N(t)\|^2_{L_2(g_\theta)}) \leq \sup_{t\geq 0}\mathbb{E}(\|D^N(t)\|^2_{L_2(g_\theta)}) < \infty. \tag{7.112}$$

□

7.14 Proof of Theorem 5.1

Proof. By Rabinovich et al. [36], the solution to any first-order recursion equation given by

$$s_{i+1} = P\left(s_i\right)$$

can be written as

$$s_i = \langle e|T^i|s\rangle.$$

Here $|s\rangle = \left\{s_0^j\right\}_{j=0}^{\infty}$ and $\langle e| = \left[\delta_{j1}\right]_{j=0}^{\infty}$ where δ_{jk} is the Kronecker symbol. T is a transfer matrix that transforms the column $\left\{s_i^j\right\}$ to a column $\left\{\left[P\left(s_i\right)\right]^j\right\}$.

In our case,

$$P\left(s_i^I\right) = \lambda(1-p)s_i^I + \lambda p\left(s_i^I\right)^d$$

and $\left\{\left[P\left(s_i^I\right)\right]^j\right\}$ can be expanded as the following:

$$\begin{aligned}\left[\lambda(1-p)s_i^I + \lambda p\left(s_i^I\right)^d\right]^j &= \sum_{l=0}^{j}\binom{j}{l}(\lambda p)^l\left(\left(s_i^I\right)^d\right)^l\left(\lambda(1-p)\right)^{j-l}\left(s_i^I\right)^{j-l}\\ &= \sum_{l=0}^{j}\binom{j}{l}(\lambda p)^l\left(\lambda(1-p)\right)^{j-l}\left(s_i^I\right)^{j+(d-1)l}.\end{aligned}$$

Denoting $k = j + (d-1)l$ so that $l = \frac{k-j}{d-1}$, we have

$$\sum_{k=j}^{dj}\binom{j}{\frac{k-j}{d-1}}(\lambda p)^{\frac{k-j}{d-1}}\left(\lambda(1-p)\right)^{j-\frac{k-j}{d-1}}\left(s_i^I\right)^k. \tag{7.113}$$

Thus the matrix elements T_{jk} are

$$T_{jk} = \binom{j}{\frac{k-j}{d-1}}(\lambda p)^{\frac{k-j}{d-1}}\left(\lambda(1-p)\right)^{j-\frac{k-j}{d-1}}. \tag{7.114}$$

Given $s_0^I = 1$, the solution to our nonlinear recursion $s_i^I = \langle e|T^i|s^I\rangle$ is the sum of all elements in the first row of T^i:

$$\begin{aligned}s_i^I &= \sum_{k_i=0}^{d^i}\left(T^i\right)_{1,k_i}\\ &= \sum_{k_i=0}^{d^i}\sum_{k_{i-1}=0}^{d^i}\cdots\sum_{k_1=0}^{d^i}T_{1,k_1}T_{k_1,k_2}\cdots T_{k_{i-1},k_i}\\ &= \sum_{k_i=0}^{d^i}\sum_{k_{i-1}=0}^{d^i}\cdots\sum_{k_1=0}^{d^i}\binom{1}{\frac{k_1-1}{d-1}}\binom{k_1}{\frac{k_2-k_1}{d-1}}\cdots\binom{k_{i-1}}{\frac{k_i-k_{i-1}}{d-1}}\left(\lambda(1-p)\right)^{1+k_1+\cdots+k_{i-1}-\frac{k_i-1}{d-1}}(\lambda p)^{\frac{k_i-1}{d-1}}\\ &= \sum_{k_1=1}^{d}\sum_{k_2=k_1}^{dk_1}\sum_{k_3=k_2}^{dk_2}\cdots\sum_{k_i=k_{i-1}}^{dk_{i-1}}\binom{1}{\frac{k_1-1}{d-1}}\binom{k_1}{\frac{k_2-k_1}{d-1}}\cdots\binom{k_{i-1}}{\frac{k_i-k_{i-1}}{d-1}}\left(\lambda(1-p)\right)^{1+k_1+k_2+\cdots+k_{i-1}-\frac{k_i-1}{d-1}}(\lambda p)^{\frac{k_i-1}{d-1}}.\end{aligned}$$

□

7.15 Proof of Theorem 5.2

Proof. Since we have the following bounds of the equilibrium s^I for $p \in (0,1)$ and $d \geq 2$,

$$\lambda^k(1-p)^{k-1} < s_k^I < \lambda^k, \quad k \geq 1. \tag{7.115}$$

Applying the above inequality to the recursion again, we get

$$s_k^I < \lambda(1-p)\lambda^{k-1} + \lambda p \lambda^{(k-1)d} = (1-p)\lambda^k + p\lambda^{(k-1)d+1}. \tag{7.116}$$

We can also bound the expected queue length the same way. Denote $x_i = s_i - s_{i+1}$ as the pdf of queue length Q, then

$$\begin{aligned}
\mathbb{E}[Q] &= \sum_{i=1}^{\infty} i x_i = \sum_{i=1}^{\infty} i(s_i^I - s_{i+1}^I) = \sum_{i=1}^{\infty} s_i^I \\
&= \sum_{i=0}^{\infty} \left[\lambda(1-p)s_i^I + \lambda p (s_i^I)^d\right] \\
&= \lambda(1-p)(\mathbb{E}[Q]+1) + \lambda p \left(1 + \sum_{i=1}^{\infty} (s_i^I)^d\right),
\end{aligned}$$

which implies that

$$\mathbb{E}[Q] = \frac{\lambda(1+pZ)}{1-\lambda+\lambda p} \tag{7.117}$$

where $Z = \sum_{i=1}^{\infty}(s_i^I)^d$. Thus, we can obtain an upper bound for Z,

$$\begin{aligned}
Z &= \sum_{i=1}^{\infty} (s_i^I)^d \\
\text{(Inequality (7.116))} &< \sum_{i=1}^{\infty} ((1-p)\lambda^i + p\lambda^{(i-1)d+1})^d \\
\text{(Jensen's Inequality)} &\leq (1-p)\sum_{i=1}^{\infty} \lambda^{id} + p \sum_{i=1}^{\infty} \lambda^{((i-1)d+1)d} \\
&= \lambda^d \left(\frac{1-p}{1-\lambda^d} + \frac{p}{1-\lambda^{d^2}}\right).
\end{aligned} \tag{7.118}$$

Similarly we can obtain an lower bound for Z,

$$\begin{aligned}
Z &= \sum_{i=1}^{\infty} (s_i^I)^d \\
\text{(Inequality (7.115))} &> \sum_{i=1}^{\infty} \left(\lambda^i (1-p)^{i-1}\right)^d \\
&= \frac{\lambda^d}{1-\lambda^d(1-p)^d}.
\end{aligned} \tag{7.119}$$

Combined with equation (7.117), we obtain the upper and lower bound for $\mathbb{E}[Q]$ as follows,

$$\frac{\lambda\left(1+\frac{p\lambda^d}{1-\lambda^d(1-p)^d}\right)}{1-\lambda+\lambda p} < \mathbb{E}[Q] < \frac{\lambda\left(1+p\lambda^d\left(\frac{1-p}{1-\lambda^d}+\frac{p}{1-\lambda^{d^2}}\right)\right)}{1-\lambda+\lambda p}. \tag{7.120}$$

For the second moment, similarly we have that

$$\begin{aligned}
\mathbb{E}[Q^2] &= \sum_{i=1}^{\infty} i^2 x_i = \sum_{i=1}^{\infty} i^2(s_i^I - s_{i+1}^I) \\
&= \sum_{i=1}^{\infty}(i^2-(i-1)^2)s_i^I \\
&= 2\sum_{i=1}^{\infty} i s_i^I - \mathbb{E}[Q] \\
&= 2\sum_{i=0}^{\infty}(i+1)\left[\lambda(1-p)s_i^I + \lambda p(s_i^I)^d\right] - \mathbb{E}[Q] \\
&= 2\lambda(1-p)\left(\sum_{i=0}^{\infty} i s_i^I\right) + 2\mathbb{E}[Q] + 2\lambda p\left(\sum_{i=1}^{\infty} i(s_i^I)^d\right) - \mathbb{E}[Q]
\end{aligned}$$

which implies that

$$\mathbb{E}[Q^2] = 2\sum_{i=1}^{\infty} i s_i^I - \mathbb{E}[Q], \tag{7.121}$$

and

$$\sum_{i=1}^{\infty} i s_i^I = \frac{\lambda p Z_2 + \mathbb{E}[Q]}{1-\lambda+\lambda p} \tag{7.122}$$

where $Z_2 = \sum_{i=1}^{\infty} i(s_i^I)^d$. We can obtain an upper bound for Z_2,

$$\begin{aligned}
Z_2 &= \sum_{i=1}^{\infty} i(s_i^I)^d \\
\text{(Inequality (7.116))} &< \sum_{i=1}^{\infty} i((1-p)\lambda^i + p\lambda^{(i-1)d+1})^d \\
\text{(Jensen's Inequality)} &\leq (1-p)\sum_{i=1}^{\infty} i\lambda^{id} + p\sum_{i=1}^{\infty} i\lambda^{((i-1)d+1)d} \\
&= \lambda^d\left(\frac{1-p}{(1-\lambda^d)^2} + \frac{p}{(1-\lambda^{d^2})^2}\right).
\end{aligned} \tag{7.123}$$

Similarly we can obtain an lower bound for Z_2,

$$\begin{aligned} Z_2 &= \sum_{i=1}^{\infty} i(s_i^I)^d \\ \text{(Inequality (7.115))} &> \sum_{i=1}^{\infty} i\left(\lambda^i(1-p)^{i-1}\right)^d \\ &= \frac{\lambda^d}{(1-\lambda^d(1-p)^d)^2}. \end{aligned} \tag{7.124}$$

Combined with Equations (7.121) and (7.122), we obtain the upper and lower bound for $\mathbb{E}[Q^2]$ as follows,

$$\mathbb{E}[Q^2] = 2 \cdot \frac{\lambda p Z_2 + \mathbb{E}[Q]}{1-\lambda+\lambda p} - \mathbb{E}[Q] = \frac{2\lambda p Z_2 + (1+\lambda(1-p))\mathbb{E}[Q]}{1-\lambda+\lambda p} \tag{7.125}$$

and

$$\mathbb{E}[Q^2] > \frac{\frac{2\lambda^{d+1}p}{(1-\lambda^d(1-p)^d)^2} + (1+\lambda(1-p))\frac{\lambda\left(1+\frac{p\lambda^d}{1-\lambda^d(1-p)^d}\right)}{1-\lambda+\lambda p}}{1-\lambda+\lambda p}, \tag{7.126}$$

$$\mathbb{E}[Q^2] < \frac{2\lambda^{d+1}\left(\frac{1-p}{(1-\lambda^d)^2} + \frac{p}{(1-\lambda^{d^2})^2}\right) + (1+\lambda(1-p))\frac{\lambda\left(1+p\lambda^d\left(\frac{1-p}{1-\lambda^d}+\frac{p}{1-\lambda^{d^2}}\right)\right)}{1-\lambda+\lambda p}}{1-\lambda+\lambda p}. \tag{7.127}$$

If we use subscript U to denote upper bound and subscript L to denote lower bound, then we can obtain an upper bound for $\mathrm{Var}[Q]$,

$$\mathrm{Var}[Q] < \mathrm{Var}_U[Q] = \mathbb{E}_U[Q^2] - \mathbb{E}_L[Q]. \tag{7.128}$$

Similarly we can also obtain a lower bound for $\mathrm{Var}[Q]$,

$$\mathrm{Var}[Q] > \mathrm{Var}_L[Q] = \mathbb{E}_L[Q^2] - \mathbb{E}_U[Q], \tag{7.129}$$

□

References

[1] `https://www.zocdoc.com/about/blog/tech/how-zocdoc-improves-patient-wait-times/`.

[2] `https://money.cnn.com/interactive/economy/average-doctor-wait-times`, 2018.

[3] Reza Aghajani, Xingjie Li, and Kavita Ramanan. The pde method for the analysis of randomized load balancing networks. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):38, 2017.

[4] Sayan Banerjee, Debankur Mukherjee, et al. Join-the-shortest queue diffusion limit in halfin–whitt regime: Tail asymptotics and scaling of extrema. *The Annals of Applied Probability*, 29(2):1262–1309, 2019.

[5] Maury Bramson, Yi Lu, Balaji Prabhakar, et al. Decay of tails at equilibrium for fifo join the shortest queue networks. *The Annals of Applied Probability*, 23(5):1841–1878, 2013.

[6] Maury Bramson et al. Stability of join the shortest queue networks. *The Annals of Applied Probability*, 21(4):1568–1625, 2011.

[7] Anton Braverman. Steady-state analysis of the join the shortest queue model in the halfin-whitt regime. *arXiv preprint arXiv:1801.05121*, 2018.

[8] Antonio Castellanosa, Andrew Dawb, Jamol J Penderb, and Galit B Yom-Tova. The co-production of service: Modeling service times in contact centers using hawkes processes.

[9] Xinyun Chen. Perfect sampling of hawkes processes and queues with hawkes arrivals. *arXiv preprint arXiv:2002.06369*, 2020.

[10] JG Dai, John J Hasenbein, and Bara Kim. Stability of join-the-shortest-queue networks. *Queueing Systems*, 57(4):129–145, 2007.

[11] Andrew Daw and Jamol Pender. Exact simulation of the queue-hawkes process. In *Proceedings of the 2018 Winter Simulation Conference*, pages 4234–4235. IEEE Press, 2018.

[12] Andrew Daw and Jamol Pender. Queues driven by hawkes processes. *Stochastic Systems*, 8(3):192–229, 2018.

[13] AB Dieker and Tonghoon Suk. Randomized longest-queue-first scheduling for large-scale buffered systems. *Advances in Applied Probability*, 47(4):1015–1038, 2015.

[14] Erik A. Van Doorn. Conditions for exponential ergodicity and bounds for the decay parameter of a birth-death process. *Advances in Applied Probability*, 17(3):514–530, 1985. ISSN 00018678. URL `http://www.jstor.org/stable/1427118`.

[15] Patrick Eschenfeldt and David Gamarnik. Join the shortest queue with many servers. the heavy-traffic asymptotics. *Mathematics of Operations Research*, 43(3):867–886, 2018.

[16] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[17] Robert D Foley, David R McDonald, et al. Join the shortest queue: stability and exact asymptotics. *The Annals of Applied Probability*, 11(3):569–607, 2001.

[18] Sergey Foss and Alexander L Stolyar. Large-scale join-idle-queue system with general service times. *Journal of Applied Probability*, 54(4):995–1007, 2017.

[19] Xuefeng Gao and Lingjiong Zhu. Functional central limit theorems for stationary hawkes processes and application to infinite-server queues. *Queueing Systems*, 90(1-2):161–206, 2018.

[20] Carl Graham. Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journal of Applied Probability*, 37(1):198–211, 2000.

[21] Carl Graham. Kinetic limits for large communication networks. In *Modeling in Applied Sciences*, pages 317–370. Springer, 2000.

[22] Carl Graham. Chaoticity results for" join the shortest queue". *CONTEMPORARY MATHEMATICS*, 275:53–68, 2001.

[23] Carl Graham. Functional central limit theorems for a large network in which customers join the shortest of several queues. *Probability Theory and Related Fields*, 131(1):97–120, Jul 2004. ISSN 1432-2064. doi: 10.1007/s00440-004-0372-9. URL `http://dx.doi.org/10.1007/s00440-004-0372-9`.

[24] Yu-Tong He and Douglas G Down. Limited choice and locality considerations for load balancing. *Performance Evaluation*, 65(9):670–687, 2008.

[25] Samuel Karlin and James McGregor. The classification of birth and death processes. *Transactions of the American Mathematical Society*, 86(2):366–400, 1957.

[26] Samuel Karlin and James L McGregor. The differential equations of birth-and-death processes, and the stieltjes moment problem. *Transactions of the American Mathematical Society*, 85(2):489–546, 1957.

[27] David T Koops, Mayank Saxena, Onno J Boxma, and Michel Mandjes. Infinite-server queues with hawkes input. *Journal of Applied Probability*, 55(3):920–943, 2018.

[28] Hwa-Chun Lin and Cauligi S Raghavendra. An approximate analysis of the join the shortest queue (jsq) policy. *IEEE Transactions on Parallel and Distributed Systems*, 7 (3):301–307, 1996.

[29] Yi Lu, Qiaomin Xie, Gabriel Kliot, Alan Geller, James R Larus, and Albert Greenberg. Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation*, 68(11):1056–1071, 2011.

[30] Michael Mitzenmacher. Studying balanced allocations with differential equations. *Combinatorics, Probability and Computing*, 8(5):473–482, 1999.

[31] Michael Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, 12(10):1094–1104, 2001.

[32] Debankur Mukherjee, Sem Borst, Johan Van Leeuwaarden, and Phil Whiting. Universality of power-of-d load balancing schemes. *ACM SIGMETRICS Performance Evaluation Review*, 44(2):36–38, 2016.

[33] Debankur Mukherjee, Sem C Borst, Johan SH Van Leeuwaarden, and Philip A Whiting. Universality of power-of-d load balancing in many-server systems. *Stochastic Systems*, 8(4):265–292, 2018.

[34] Samantha Nirenberg, Andrew Daw, and Jamol Pender. The impact of queue length rounding and delayed app information on disney world queues. In *2018 Winter Simulation Conference (WSC)*, pages 3849–3860. IEEE, 2018.

[35] Sophia Novitzky, Jamol Pender, Richard H Rand, and Elizabeth Wesson. Nonlinear dynamics in queueing theory: Determining the size of oscillations in queues with delay. *SIAM Journal on Applied Dynamical Systems*, 18(1):279–311, 2019.

[36] S Rabinovich, G Berkolaiko, and S Havlin. Solving nonlinear recursions. *Journal of Mathematical Physics*, 37(11):5828–5836, 1996.

[37] Shuang Tao and Jamol Pender. A stochastic analysis of bike sharing systems. *arXiv preprint arXiv:1708.08052*, 2017.

[38] John N. Tsitsiklis and Kuang Xu. On the power of (even a little) resource pooling. *Stochastic Systems*, 2(1):1–66, 2012.

[39] Stephen R.E. Turner. The effect of increasing routing choice on resource pooling. *Probability in the Engineering and Informational Sciences*, 12(1):109–124, 1998.

[40] Nikita Dmitrievna Vvedenskaya, Roland L'vovich Dobrushin, and Fridrikh Izrailevich Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996.

[41] Ward Whitt. Blocking when service is required from several facilities simultaneously. *AT & T Technical journal*, 64(8):1807–1856, 1985.