Doubly Exponential Solution for Randomized Load Balancing Models with General Service Times

Quan-Lin Li
School of Economics and Management Sciences
Yanshan University, Qinhuangdao 066004, P.R. China

September 5, 2010

Abstract

The randomized load balancing model (also called supermarket model) is now being applied to the study of load balancing in data centers and multi-core servers systems. It is very interesting to analyze the general service times in the supermarket model, and specifically understand influence of the heavy-tailed service times on the doubly exponential solution. Since the supermarket model is a complex queueing system, the general service times make its analysis more challenging than the exponential or PH service case. Up to now, it still is an open problem whether or how the heavy-tailed service times can disrupt the doubly exponential structure of the fixed point in the supermarket model.

In this paper, we provide a novel and simple approach to study the supermarket model with general service times. This approach is based on the supplementary variable method used in analyzing stochastic models extensively. We organize an infinite-size system of integral-differential equations by means of the density dependent jump Markov process, and obtain a close-form solution: doubly exponential structure, for the fixed point satisfying the system of nonlinear equations, which is always a key in the study of supermarket models. The fixed point is decomposed into two groups of information under a product form: the arrival information and the service information. Based on this, we indicate two important observations: the fixed point for the supermarket model is different from the tail of stationary queue length distribution for the ordinary M/G/1 queue, and the doubly exponential solution to the fixed point can...
extensively exist even if the service time distribution is heavy-tailed. Furthermore, we analyze the exponential convergence of the current location of the supermarket model to its fixed point, and study the Lipschitz condition in the Kurtz Theorem under general service times. Based on these analysis, one can gain a new understanding how workload probing can help in load balancing jobs with general service times such as heavy-tailed service.

**Keywords:** Randomized load balancing, supermarket model, density dependent jump Markov process, fixed point, doubly exponential solution, heavy-tailed distribution, exponential convergence, Lipschitz condition.

1 Introduction

Randomized load balancing, where a job is assigned to a server from a small subset of randomly chosen servers, is very simple to implement, and can surprisingly deliver better performance (for example reducing collisions, waiting times, backlogs) in a number of applications, such as, data center, hash tables, distributed memory machines, path selection in networks, and task assignment at web servers. One useful model that has been extensively used to study the randomized load balancing schemes is the supermarket model. In the supermarket model, a key result by Vvedenskaya, Dobrushin and Karpelevich [31] indicated that when each Poisson arriving job is assigned to the shortest one of \( d \geq 2 \) randomly chosen queues with exponential service times, the equilibrium queue length can decay doubly exponentially in the limit as the population number \( n \to \infty \), and the stationary fraction of queues with at least \( k \) customers is \( \rho^{\frac{k}{d-1}} \), which is a substantially exponential improvement over the case for \( d = 1 \), where the tail of stationary queue length distribution in the corresponding M/M/1 queue is \( \rho^k \).

The distributed load balancing strategies, in which individual job decisions are based on information on a limited number of other processors, have been studied analytically by Eager, Lazokwska and Zahorjan [5, 6, 7] and through trace-driven simulations by Zhou [33]. Based on this, the supermarket model is developed by queueing theory and Markov processes. Most of recent research applied the density dependent jump Markov processes to deal with a simple supermarket model with Poisson arrival processes and exponential service times, a key result of which illustrates that there exists a unique fixed point which is decreasing doubly exponentially. That approach used in the literature relies
on determining the behavior of the supermarket model as its size grows to infinity, and its behavior is naturally described as a system of differential equations, which leads to a closed form solution: doubly exponential structure, of the fixed point. Readers may refer to, such as, analyzing a basic and simple supermarket model by Azar, Broder, Karlin and Upfal [2], Vvedenskaya, Dobrushin and Karpelevich [31], Mitzenmacher [19, 20].

Certain generalization of the supermarket model have been explored, for example, simple variations by Mitzenmacher and Vöcking [27], Mitzenmacher [21, 22, 25], Vöcking [30], Mitzenmacher, Richa, and Sitaraman [26] and Vvedenskaya and Suhov [32]; and analyzing load information by Mirchandaney, Towsley, and Stankovic [28], Dahlin [3], Mitzenmacher [24, 26]. Furthermore, Martin and Suhov [18], Martin [17], Suhov and Vvedenskaya [29] studied the supermarket mall model by means of the fast Jackson network, where each node in a Jackson network is replaced by $N$ parallel servers, and a job joins the shortest of $d$ randomly chosen queues at the node to which it is directed. Luczak and McDiamid [15, 16] studied the maximum queue length of the original supermarket model with exponential service times when the service speed scales linearly with the number of jobs in the queue. Li, Lui and Wang [11, 12] discussed the supermarket model with PH service times and the supermarket model with Markovian arrival processes, respectively. Readers may refer to an excellent overview by Mitzenmacher, Richa, and Sitaraman [26].

This paper is interested in analyzing the supermarket model with general service times, which is an open problem for determining whether or how the heavy-tailed service times can disrupt the doubly exponential structure of the fixed point. On the other hand, note that the supermarket model is a complex queueing system and has much different characteristics from the ordinary queueing systems, thus the general service times make its analysis more challenging than the exponential or PH service case. Up to now, there has not been an effective method to be able to deal with the supermarket model with general service times.

The main contributions of the paper are threefold. The first one is to provide a novel and simple approach to study the supermarket model with general service times. This approach is based on the supplementary variable method but is described as a new integral-differential structure for expressing and computing the fraction of queues efficiently. Using the new approach, we setup an infinite-size system of integral-differential equations, which makes applications of the density dependent jump Markov processes to be able to deal with the general distributions, such as general service times, involved in the supermarket
The second one is to obtain a close-form solution: doubly exponential structure, for the fixed point satisfying the system of nonlinear equations, which is always a key in the study of supermarket models. Furthermore, this paper analyzes the exponential convergence of the current location of the supermarket model to its fixed point, and studies the Lipschitz condition in the Kurtz Theorem under general service times. Also, this paper provides numerical examples to illustrate the effectiveness of our approach in analyzing the randomized load balancing schemes with the non-exponential service requirements. The third one is to obtain that the fixed point is decomposed into two groups of information under a product form: the arrival information and the service information. Based on this, we indicate three important observations:

(a) The fixed point for the supermarket model is different from the tail of stationary queue length distribution for the ordinary M/G/1 queue, because the fixed point is light-tailed but the stationary queue length is heavy-tailed if the service times are heavy-tailed. Note that such a difference is illustrated in this paper for the first time, while it can not be observed in the literature for the supermarket model with Poisson arrivals and exponential service times, e.g., see Mitzenmacher, Richa, and Sitaraman [26].

(b) The doubly exponential solution to the fixed point can extensively exist even if the service time distribution is heavy-tailed. This is an answer of the above open problem to illustrate the role played by the heavy-tailed service time distribution in the doubly exponential solution to the fixed point.

(c) The doubly exponential solution to the fixed point is not unique for a more general supermarket model. Note that we give three different doubly exponential solutions in the supermarket model with Poisson arrivals and PH service times, thus it is very interesting to provide all the doubly exponential solutions for a more general supermarket model.

Based on this, one can gain the new and important understanding how the workload probing can help in load balancing jobs with general service times such as heavy-tailed service.

The remainder of this paper is organized as follows. In Section 2, we first describe a supermarket model with general service times, which is always useful in the study
of randomized load balancing schemes. Then the supermarket model is expressed as a systems of integral-differential equations in terms of the density dependent jump Markov processes. In Section 3, we first introduce a fixed point of the system of integral-differential equations, and set up a system of nonlinear equations satisfied by the fixed point. Then we provide a close-form solution: doubly exponential structure, to the system of nonlinear equations. In Section 4, we provide a necessary discussion on the key parameter \( \theta \) used in the doubly exponential structure, and indicate that the doubly exponential solution to the fixed point extensively exists even if the service time distribution is heavy-tailed. In Section 5, we give three methods to analyze the supermarket model with Poisson arrivals and PH service times, and provide three different ways to determine the key parameter \( \theta \) and compute the doubly exponential solution to the fixed point. We show that the doubly exponential solution to the fixed point is not unique for a more general supermarket. In Section 6, we study the exponential convergence of the current location of the supermarket model to its fixed point. Not only does the exponential convergence indicates the existence of the fixed point, but it also explains such a convergent process is very fast. In Section 7, we apply the Kurtz Theorem to study the supermarket model with the general service times, and analyze the Lipschitz condition with respect to general service times. Some concluding remarks are given in Section 8.

2 Supermarket Model

In this section, we first describe a supermarket model with general service times. Then we provide a novel and simple approach to setup an infinite-size system of integral-differential equations based on the density dependent jump Markov processes. Note that this approach is based on the supplementary variable method but is described as a new integral-differential structure so that the corresponding boundary conditions are written in a different version.

Let us formally describe the supermarket model, which is abstracted as a multi-server multi-queue stochastic system. Customers arrive at a queueing system of \( n > 1 \) servers as a Poisson process with an average arrival rate \( n \lambda \) for \( \lambda > 0 \). The service time \( \chi_k \) of the \( k \)th customer is general with the distribution function

\[
G(x) = P\{\chi_k \leq x\} = 1 - \exp \left\{ - \int_0^x \mu(y) \, dy \right\},
\]
where all the random variables $\chi_k$ for $k \geq 1$ are i.i.d. with the mean $E[\chi_k] = 1/\mu$. Each arriving customer chooses $d \geq 1$ servers independently and uniformly at random from these $n$ servers, and waits for service at the server which currently contains the fewest number of customers. If there is a tie, servers with the fewest number of customers will be chosen randomly. All customers in any service center will be served in the First-Come-First-Served (FCFS) manner. Figure 1 simply shows such a supermarket model.

In the study of supermarket models, it is necessary for us to study general service time distributions, for example, heavy-tailed distributions. Not only because the general distribution makes analysis of the supermarket models more difficult and challenging than those in the literature for the exponential or PH service case, but it also allows us to model more realistic systems and understand their performance implication under the randomized load balancing strategy. As indicated in [8], the process times of many parallel jobs, in particular, jobs to data centers, tend to be non-exponential. Unless we state otherwise, we assume that all the random variables defined above are independent, and that the system is operating under the condition: $\rho = \lambda/\mu < 1$.

**Lemma 1** The supermarket model with general service times is stable if $\rho = \lambda/\mu < 1$.

**Proof:** Let $Q_O(t)$ and $Q_S(t)$ be the queue lengths of the ordinary M/G/1 queue and of an arbitrary server in the supermarket model at time $t$, respectively. Note that in the supermarket model, each customer chooses $d$ servers independently and uniformly at random, and the queue length of the entering server is currently shorten, it is easy to see that for each $t \geq 0$,

$$0 \leq Q_S(t) \leq Q_O(t).$$

(1)
Since the ordinary M/G/1 queue is stable if $\rho = \lambda/\mu < 1$, it follows from (1) that the supermarket model with general service times is stable if $\rho = \lambda/\mu < 1$. This completes the proof.

For $k \geq 1$, we define $n_k(t, x)dx$ as the number of queues with at least $k$ customers and the residual service time of each server be in the interval $[x, x+dx]$ at time $t \geq 0$. Clearly, $0 \leq n_k(t, x) \leq n$ for $x \geq 0$ and $1 \leq k \leq n$. Let

$$s_{k,n}(t, x) = \frac{n_k(t, x)}{n},$$

which is the density function of the fraction of queues with at least $k$ customers and the residual service time of each server be $x$. We write

$$S_k(t, x) = \lim_{n \to \infty} s_{k,n}(t, x), \text{ for } k \geq 1.$$

We define $n_{0,n}(t)$ as the number of queues with at least 0 customers at time $t \geq 0$. Clearly, $n_{0,n}(t) = n$. Let

$$s_{0,n}(t) = \frac{n_{0,n}(t)}{n}.$$

Then $s_{0,n}(t) = 1$ for all $t \geq 0$ and

$$S_0(t) = \lim_{n \to \infty} s_{0,n}(t) = 1.$$

Let $V(t)$ be the fraction of queues with zero customer at time $t$. Then

$$S_0(t) = V(t) + \int_0^{+\infty} S_1(t, x) \, dx.$$

Thus we have

$$S(t, x) = (S_0(t), S_1(t, x), S_2(t, x), \ldots).$$

The following proposition shows that the sequence $\{S_k(t, x)\}$ is monotone increasing for $k \geq 1$, while its proof is easily by means of the definition of $S_k(t, x)$ for $k \geq 1$.

**Proposition 1** For $1 \leq k < l$

$$S_l(t, x) < S_k(t, x) < S_0(t) = 1$$

and

$$\int_0^{+\infty} S_l(t, x) \, dx < \int_0^{+\infty} S_k(t, x) \, dx < S_0(t) = 1.$$
Now, we setup a system of integral-differential equations by means of the density dependent jump Markov process. To that end, we provide an example with $k \geq 2$ to indicate how to derive the system of integral-differential equations.

Consider the supermarket model with $n$ queues, and determine the expected change in the number of servers with at least $k$ customers and the residual service time of each server be $x$ over a small time period of length $d\,t$. The probability that a customer arriving during this time period is $n\lambda\,d\,t$, and the probability that an arriving customer joins a queue of size $k-1$ is given by

\[
\int_0^{+\infty} s_{k-1, n}^d (t, x) \, dx - \int_0^{+\infty} s_{k, n}^d (t, x) \, dx.
\]

Thus, the probability that during this time period, any arriving customer joins a queue of size $k-1$ is given by

\[
n\lambda \, d\,t \cdot \left[ \int_0^{+\infty} s_{k-1, n}^d (t, x) \, dx - \int_0^{+\infty} s_{k, n}^d (t, x) \, dx \right].
\]

Similarly, the probability that a customer leaves a server of size $k$ is given by

\[
ndt \cdot \left[ \int_0^{+\infty} \mu (x) s_{k,n} (t, x) \, dx - \int_0^{+\infty} \mu (x) s_{k+1,n} (t, x) \, dx \right].
\]

Therefore we can obtain

\[
d \int_0^{+\infty} n_k (t, x) \, dx \bigg|_{dt} = n\lambda \left[ \int_0^{+\infty} s_{k-1,n}^d (t, x) \, dx - \int_0^{+\infty} s_{k,n}^d (t, x) \, dx \right]
+ \left[ \int_0^{+\infty} \mu (x) s_{k,n} (t, x) \, dx - \int_0^{+\infty} \mu (x) s_{k+1,n} (t, x) \, dx \right],
\]

which leads to

\[
\frac{d}{dt} \int_0^{+\infty} s_{k,n} (t, x) \, dx = \lambda \left[ \int_0^{+\infty} s_{k-1,n}^d (t, x) \, dx - \int_0^{+\infty} s_{k,n}^d (t, x) \, dx \right]
+ \left[ \int_0^{+\infty} \mu (x) s_{k,n} (t, x) \, dx - \int_0^{+\infty} \mu (x) s_{k+1,n} (t, x) \, dx \right]. \tag{2}
\]

Taking $n \to \infty$ in the both sides of \[(2)\], we have

\[
\frac{d}{dt} \int_0^{+\infty} S_k (t) \, dx = \lambda \left[ \int_0^{+\infty} S_{k-1} (t, x) \, dx - \int_0^{+\infty} S_k (t, x) \, dx \right]
+ \left[ \int_0^{+\infty} \mu (x) S_k (t, x) \, dx - \int_0^{+\infty} \mu (x) S_{k+1} (t, x) \, dx \right]. \tag{3}
\]

Using a similar analysis to that for deriving Equation \[(3)\], we can easily obtain a system of integral-differential equations for the fraction density vector $S (t, x)$ as follows:

\[
S_0 (t) = 1 \text{ for all } t \geq 0, \tag{4}
\]

\[
\frac{d}{dt} S_0 (t) = -\lambda S_0^d (t) + \int_0^{+\infty} \mu (x) S_1 (t, x) \, dx, \tag{5}
\]

8
\[ \frac{d}{dt} \int_0^{+\infty} S_1(t, x) \, dx = \lambda S_0^d(t) - \lambda \int_0^{+\infty} S_1^d(t, x) \, dx - \int_0^{+\infty} \mu(x) S_1(t, x) \, dx + \int_0^{+\infty} \mu(x) S_2(t, x) \, dx, \quad (6) \]

and for \( k \geq 2, \)

\[ \frac{d}{dt} \int_0^{+\infty} S_k(t, x) \, dx = \lambda \int_0^{+\infty} S_{k-1}^d(t, x) \, dx - \lambda \int_0^{+\infty} S_k^d(t, x) \, dx - \int_0^{+\infty} \mu(x) S_k(t, x) \, dx + \int_0^{+\infty} \mu(x) S_{k+1}(t, x) \, dx. \quad (7) \]

Remark 1 When there are \( n \) servers in the supermarket model, it is necessary to give a finite-size system of integral-differential equations for the fraction density vector \( S^{(n)}(t, x) = (s_{0,n}(t), s_{1,n}(t, x), \ldots, s_{n,n}(t, x)) \) as follows:

\[ s_{0,n}(t) = 1 \text{ for all } t \geq 0, \]

\[ \frac{d}{dt} s_{0,n}(t) = -\lambda s_{0,n}(t) + \int_0^{+\infty} \mu(x) s_{1,n}(t, x) \, dx, \]

\[ \frac{d}{dt} \int_0^{+\infty} s_{1,n}(t, x) \, dx = \lambda s_{0,n}(t) - \lambda \int_0^{+\infty} s_{1,n}^d(t, x) \, dx - \int_0^{+\infty} \mu(x) s_{1,n}(t, x) \, dx + \int_0^{+\infty} \mu(x) s_{2,n}(t, x) \, dx, \]

and for \( n \geq k \geq 2, \)

\[ \frac{d}{dt} \int_0^{+\infty} s_{k,n}(t, x) \, dx = \lambda \int_0^{+\infty} s_{k-1,n}^d(t, x) \, dx - \lambda \int_0^{+\infty} s_{k,n}^d(t, x) \, dx - \int_0^{+\infty} \mu(x) s_{k,n}(t, x) \, dx + \int_0^{+\infty} \mu(x) s_{k+1,n}(t, x) \, dx. \]

3 Doubly Exponential Solution

In this section, we discuss the fixed point of the system of integral-differential equations in Equations (4) to (7), and set up a system of nonlinear equations satisfied by the fixed point. Also, we provide a closed-form solution: doubly exponential structure, to the system of nonlinear equations.

A row vector \( \pi(x) = (\pi_0, \pi_1(x), \pi_2(x), \ldots) \) is called a fixed point of the fraction density vector \( S(t, x) = (S_0(t), S_1(t, x), S_2(t, x), \ldots) \) if there exists a \( t_0 \geq 0 \) such that
\( S_0 (t) = \pi_0 \) and \( S_k (t, x) = \pi_k (x) \) for all \( t \geq t_0 \) and \( k \geq 1 \). It is easy to see that if \( \pi (x) \) is a fixed point of the fraction density vector \( S (t, x) \) for all \( t \geq t_0 \), then
\[
\frac{d}{dt} S_0 (t)_{|t \geq t_0} = 0
\]
and for \( k \geq 1 \)
\[
\frac{d}{dt} S_k (t, x)_{|t \geq t_0} = 0
\]
which leads to
\[
\int_0^{+\infty} \frac{d}{dt} S_k (t, x)_{|t \geq t_0} \, dx = 0. \tag{8}
\]
Since for \( k \geq 1 \)
\[
0 \leq S_k (t, x) \leq 1,
\]
using the Dominated Convergence Theorem we obtain
\[
\frac{d}{dt} \int_0^{+\infty} S_k (t, x)_{|t \geq t_0} \, dx = 0.
\]
Therefore, if \( \pi (x) = (\pi_0, \pi_1 (x), \pi_2 (x), \ldots) \) is a fixed point of the fraction density vector \( S (t, x) = (S_0 (t), S_1 (t, x), S_2 (t, x), \ldots) \) for all \( t \geq t_0 \), then the system of integral-differential equations (4) to (7) can be simplified as
\[
\pi_0 = 1 \tag{9}
\]
\[
- \lambda \pi_0^d + \int_0^{+\infty} \mu (x) \pi_1 (x) \, dx = 0, \tag{10}
\]
\[
\lambda \pi_0^d (t) - \lambda \int_0^{+\infty} \pi_1^d (x) \, dx - \int_0^{+\infty} \mu (x) \pi_1 (x) \, dx + \int_0^{+\infty} \mu (x) \pi_2 (x) \, dx = 0, \tag{11}
\]
and for \( k \geq 2 \),
\[
\lambda \int_0^{+\infty} \pi_k^d (x) \, dx - \lambda \int_0^{+\infty} \pi_{k-1}^d (x) \, dx - \int_0^{+\infty} \mu (x) \pi_k (x) \, dx + \int_0^{+\infty} \mu (x) \pi_{k+1} (x) \, dx = 0. \tag{12}
\]
In what follows we derive a closed-form expression for \( \pi (x) = (\pi_0, \pi_1 (x), \pi_2 (x), \ldots) \).

It follows from Equations (9) and (10) that
\[
\int_0^{+\infty} \mu (x) \pi_1 (x) \, dx = \lambda. \tag{13}
\]
To solve Equation (13), using the fact that \( \int_0^{+\infty} \mu (x) \overline{G} (x) \, dx = 1 \) we have
\[
\pi_1 (x) = \lambda \overline{G} (x) = \rho \cdot \mu \overline{G} (x). \tag{14}
\]
Based on the fact that \( \pi_0 = 1 \) and \( \pi_1 (x) = \rho \cdot \mu G (x) \), it follows from Equations (11) and (13) that
\[
- \lambda \rho^d \cdot \int_0^{+\infty} [\mu G (x)]^d \, dx + \int_0^{+\infty} \mu (x) \pi_2 (x) \, dx = 0.
\]
Let \( \theta = \int_0^{+\infty} [\mu G (x)]^d \, dx \), and we assume that \( 0 < \theta < +\infty \). Then
\[
\int_0^{+\infty} \mu (x) \pi_2 (x) \, dx = \lambda \theta \rho^d. \tag{15}
\]
Using a similar analysis on Equation (15), we have
\[
\pi_2 (x) = \lambda \theta \rho^d \mu G (x) = \theta \rho^{d+1} \cdot \mu G (x). \tag{16}
\]

Based on \( \pi_1 (x) = \rho \cdot \mu G (x) \) and \( \pi_2 (x) = \theta \rho^{d+1} \cdot \mu G (x) \), we can compute
\[
\lambda \int_0^{+\infty} \pi_1^d (x) \, dx = \lambda \theta \rho^d,
\]
\[
\int_0^{+\infty} \mu (x) \pi_2 (x) \, dx = \theta \rho^{d+1} \mu \int_0^{+\infty} \mu (x) G (x) \, dx = \lambda \theta \rho^d
\]
and
\[
\lambda \int_0^{+\infty} \pi_2^d (x) \, dx = \theta \rho^{d+1} \rho^d \int_0^{+\infty} [\mu G (x)]^d \, dx = \lambda \theta^{d+1} \rho^{d+2},
\]
thus it follows from Equation (12) that for \( k = 2 \),
\[
\int_0^{+\infty} \mu (x) \pi_3 (x) \, dx = \lambda \theta^{d+1} \rho^{d+2},
\]
which leads to
\[
\pi_3 (x) = \theta^{d+1} \rho^{d+2} \mu G (x) = \theta^{d+1} \rho^{d+2} \cdot \mu G (x). \tag{17}
\]

Based on the above analysis for the simple expressions \( \pi_k (x) \) for \( k = 1, 2 \) and 3, we can summarize the following theorem.

**Theorem 1** The fixed point \( \pi = (\pi_0, \pi_1 (x), \pi_2 (x), \ldots) \) is given by
\[
\pi_0 = 1,
\]
\[
\pi_1 (x) = \rho \cdot \mu G (x)
\]
and for \( k \geq 2 \),
\[
\pi_k (x) = \theta^{d_k-2+d_k-3+\ldots+1} \rho^{d_k-1+d_k-2+\ldots+1} \cdot \mu G (x), \tag{18}
\]
or
\[
\pi_k (x) = \theta^{d_k-1} \rho^{d_k-1} \cdot \mu G (x). \tag{19}
\]
Proof By induction, one can easily derive the above result.

It is clear from (16) and (17) that Equation (18) or (19) is correct for the cases with \( l = 2, 3 \). Now, we assume that Equation (19) is correct for the cases with \( l = k \). Then

\[
\lambda \int_0^{+\infty} \pi_k^d (x) \, dx = \lambda \theta \frac{d^{k-1}}{d-1} \rho \frac{d^k}{d-1},
\]

and

\[
\mu (x) \pi_k (x) \, dx = \lambda \theta \frac{d^{k-1}}{d-1} \rho \frac{d^k}{d-1},
\]

it follows from Equation (12) that

\[
\lambda \int_0^{+\infty} \pi_k^d (x) \, dx = \lambda \theta \frac{d^{k-1}}{d-1} \rho \frac{d^k}{d-1} + 1,
\]

and

\[
\mu (x) \pi_k^d (x) \, dx = \lambda \theta \frac{d^{k-1}}{d-1} \rho \frac{d^k}{d-1}.
\]

Thus, for \( l = k + 1 \) we have

\[
\pi_{k+1} (x) = \theta \frac{d^{k-1}}{d-1} \rho \frac{d^k}{d-1} \cdot \mu G(x).
\]

This completes the proof.

Let \( \tilde{\theta} = \int_0^{+\infty} \left[ G(x) \right]^d \, dx \). Then \( \theta = \mu^d \tilde{\theta} \). The following corollary provides another expression for the fixed point.

Corollary 2

\[
\pi_0 = 1
\]

and for \( k \geq 1 \)

\[
\pi_k (x) = \lambda \frac{d^{k-1}}{d-1} \left\{ \theta \frac{d^{k-1}}{d-1} \, G(x) \right\}.
\]

It is easy to see from Corollary 2 that the fixed point is decomposed into two groups of information under a product form: the arrival information and the service information. At the same time, the service information indicates that the doubly exponential solution to the fixed point must exist for \( 0 < \mu < + \infty \), even if the service times are heavy-tailed.

The following corollary provides an upper bound for the fixed point.

Corollary 3 For \( k \geq 1 \) and \( x \geq 0 \),

\[
\pi_k (x) < \int_0^{+\infty} \pi_k (x) \, dx < \rho \frac{d^{k-1}}{d-1} \frac{\lambda^k}{\mu}.
\]
Proof: Note that $0 \leq \overline{G}(x) \leq 1$, we have
\[
\bar{\theta} = \int_{0}^{+\infty} [\overline{G}(x)]^d \, dx < \int_{0}^{+\infty} \overline{G}(x) \, dx = \frac{1}{\mu}.
\]
It follows from Corollary 2 that
\[
\pi_k(x) < \int_{0}^{+\infty} \pi_k(x) \, dx < \int_{0}^{+\infty} \frac{d^{k-1}}{d-1} \theta \frac{d^{k-1}}{d-1} \overline{G}(x) \, dx = \rho \frac{d^{k-1}}{d-1} \lambda^d.
\]
This completes the proof.

Now, we compute the expected sojourn time $T_d$ which a tagged arriving customer spends in the supermarket model. For the general service times, a tagged arriving customer is the $k$th customer in the corresponding queue with the following probability
\[
\int_{0}^{+\infty} \pi_k(x) \, dx - \int_{0}^{+\infty} \pi_{k+1}(x) \, dx = \theta \frac{d^{k-1}}{d-1} \rho \frac{d^{k-1}}{d-1} - \theta \frac{d^{k-1}}{d-1} \rho \frac{d^{k-1}}{d-1}.
\]
When $k \geq 1$, the head customer in the queue has been served, and so its service time is residual and is denoted as $X_R$. Under the stationary setting, we have
\[
P\{X_R \leq x\} = \int_{0}^{x} [\mu\overline{G}(y)] \, dy
\]
with
\[
E[X_R] = \int_{0}^{+\infty} \int_{x}^{+\infty} [\mu\overline{G}(y)] \, dy \, dx.
\]
Thus it is easy to see that the expected sojourn time of the tagged arriving customer is given by
\[
E[T_d] = \left[ \pi_0^d - \int_{0}^{+\infty} \pi_1^d(x) \, dx \right] E[X] + \sum_{k=1}^{\infty} \left[ \int_{0}^{+\infty} \pi_k^d(x) \, dx - \int_{0}^{+\infty} \pi_{k+1}^d(x) \, dx \right] \left[ E[X_R] + kE[X] \right]
\]
\[= \left[ 1 - \int_{0}^{+\infty} \pi_1^d(x) \, dx \right] E[X] + \int_{0}^{+\infty} \pi_1^d(x) \, dx E[X_R]
\]
\[+ E[X] \sum_{k=1}^{\infty} k \left[ \int_{0}^{+\infty} \pi_k^d(x) \, dx - \int_{0}^{+\infty} \pi_{k+1}^d(x) \, dx \right]
\]
\[= \{E[X_R] - E[X]\} \int_{0}^{+\infty} \pi_1^d(x) \, dx + E[X] \left\{ 1 + \sum_{k=1}^{\infty} \int_{0}^{+\infty} \pi_k^d(x) \, dx \right\}
\]
\[= \theta \rho^d \{E[X_R] - E[X]\} + E[X] \left[ \sum_{k=1}^{\infty} \theta \frac{d^{k-1}}{d-1} \rho \frac{d^{k-1}}{d-1} \right].
\]
If the service times are exponential, then \( E[X_R] = E[X] \), thus we obtain
\[
E[T_d] = \frac{1}{\mu} \left[ \sum_{k=1}^{\infty} \theta^{d-1} \rho^{d-k} \right],
\]
which is the same as Corollary 3.8 in Mitzenmacher [20].

We consider a computational example for the expected sojourn time in the supermarket model with an Erlang service time distribution \( E(m, \mu) \), where \( m = 2, \mu = 1, d = 2 \). Figure 3 shows how the expected sojourn time depends on the arrival rate.

Figure 2: the expected sojourn time \( E[T_d] \) depends on the arrival rate \( \lambda \)

With the results from Equation (18) or (19), let us now provide some useful discussions on the asymptotic behavior of the fixed point \( \pi(x) = (\pi_0, \pi_1(x), \pi_2(x), \ldots) \). Note that we express \( a_k \sim O(b_k) \) if \( \lim_{k \to \infty} a_k/b_k = c \in (-\infty, 0) \cup (0, +\infty) \).

Remark 2 If the general distribution \( G(x) \) and its mean \( 1/\mu \) are given, then \( \theta = \int_0^{+\infty} [\mu G(x)]^d \, dx \) is a deterministic factor. We have
\[
\frac{\pi_k(x)}{\theta^{d_k-1}} \sim O\left( \rho^{d_k-1} \right) \mu G(x), \text{ as } k \to \infty.
\]
In this case, the heavy traffic should have a bigger influence on the asymptotic behavior of the sequence \( \left\{ \frac{\pi_k(x)}{\theta^{d_k-1}} \right\} \).

Remark 3 If \( \rho \) is given, then
\[
\frac{\pi_k(x)}{\rho^{d_k-1}} \sim O\left( \theta^{d_k-1} \right) \mu G(x), \text{ as } k \to \infty.
\]
In this case, the maximal value $\theta_{\text{max}}$ of the positive number $\int_0^{+\infty} [\mu G(x)]^d \, dx$ should have a bigger influence on the asymptotic behavior of the sequence \( \left\{ \frac{\pi_k(x)}{\rho^{d-1}} \right\} \).

4 A discussion for the key parameter $\theta$

In this section, we provide a necessary discussion for the key parameter $\theta$ in the doubly exponential solution of Theorem 1. Based on this, for the fixed point we give a new and important observation: the doubly exponential solution to the fixed point can extensively exist for $0 < \mu < +\infty$, even if the service time distribution is heavy-tailed.

Note that

$$\theta = \int_0^{+\infty} [\mu G(x)]^d \, dx$$

$$= \frac{\int_0^{+\infty} [G(x)]^d \, dx}{\left[ \int_0^{+\infty} G(x) \, dx \right]^d},$$

it is easy to see that $\theta = 1$ if $d = 1$. Thus, we need to analyze the case for $k \geq 2$ as follows.

Since $0 \leq G(x) \leq 1$, we get that $0 \leq [G(x)]^d \leq G(x) \leq 1$, which leads to

$$\int_0^{+\infty} [G(x)]^d \, dx \leq \int_0^{+\infty} G(x) \, dx = 1/\mu.$$

It is easy to see that $0 < \theta < \mu^{d-1}$, and thus if $0 < \mu < +\infty$, then $0 < \theta < +\infty$.

In what follows we analyze five simple and useful examples. In first two examples, the service time distribution is light-tailed; while in the last three examples, the service time distribution is heavy-tailed. Specifically, the examples with heavy-tailed service times illustrate two important observations: the first one indicates that the fixed point for the supermarket model is different from the tail of stationary queue length distribution for the ordinary M/G/1 queue, and the second one is to show that the doubly exponential solution to the fixed point can exist extensively if the service time mean is non-zero and finite.

**Example one:** Exponential distribution. Let $G(x) = e^{-\mu x}$. Then $\theta = \mu^{d-1}/d$. It is easy to see that when $\mu > \frac{d-1}{d}, \theta > 1$; when $\mu = \frac{d-1}{d}, \theta = 1$; and when $\mu < \frac{d-1}{d}, 0 < \theta < 1$. If $d = 2$, then $\theta$ is a linear function of $\mu$, and if $d = 3$, then $\theta$ is a nonlinear function of $\mu$. Figures 3 and 4 show the functions $\theta = \mu/2$ and $\theta = \mu^2/3$, respectively.
Example two: Erlang distribution $E(m, \mu)$. Let $\mathcal{G}(x) = e^{-\mu x} \sum_{k=0}^{m} \frac{(\mu x)^k}{k!}$. Then $\theta$ is given by

$$
\theta = \left( \frac{\mu}{m} \right)^d \int_0^{+\infty} e^{-\mu dx} \left[ \sum_{k=0}^{m} \frac{(\mu x)^k}{k!} \right]^d dx.
$$

Let $\mu = 1$. Table 1 lists how $\theta$ depends on the parameter pair $(m, d)$. As seen from Table 1, $\theta$ is decreasing for each of the two parameters $m$ and $d$.

| $(m, d)$ | (2, 2) | (2, 5) | (2, 10) | (5, 2) | (10, 2) | (5, 5) | (10, 10) |
|----------|--------|--------|---------|--------|--------|--------|---------|
| $\theta$ | 0.52   | 0.19   | $9.15 \times 10^{-2}$ | $4.13 \times 10^{-2}$ | $9.48 \times 10^{-4}$ | $1.11 \times 10^{-3}$ | $6.51 \times 10^{-10}$ |
Example three: Weibull distribution \( W(\tau, \mu) \). Let \( G(x) = \exp\{- (\mu \tau)^x\} \). It is easy to check that the mean of the Weibull distribution is given by

\[
\frac{1}{\mu} \Gamma \left( 1 + \frac{1}{\tau} \right),
\]

which follows that \( \theta \) is given by

\[
\theta = \frac{\mu^{d-1}}{d^\tau \left[ \Gamma \left( 1 + \frac{1}{\tau} \right) \right]^{d-1}},
\]

where \( \Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx \). Obviously, the Weibull distribution \( W(\tau, \mu) \) is heavy-tailed if \( 0 < \tau < 1 \); and the Weibull distribution \( W(\tau, \mu) \) is light-tailed if \( \tau > 1 \). To indicate the role played by the heavy-tailed parameter \( \tau \) for \( 0 < \tau < 1 \), taking \( \mu = 5 \) and \( d = 2 \) we have

\[
\theta = \frac{5}{2^\tau \Gamma \left( 1 + \frac{1}{\tau} \right)}.
\]

Table 2 indicates how \( \theta \) depends on the heavy-tailed parameter \( \tau \), such as, \( 0 < \theta < 1 \) if \( \tau = 0.2 \); \( \theta > 1 \) if \( \tau = 0.9 \). This example, together with Theorem[I] illustrates an important observation that the fixed point \( \pi = (\pi_0, \pi_1(x), \pi_2(x), \ldots) \) is doubly exponential (clearly, it is light-tailed) even if the service time distribution is heavy-tailed. Based on this, the the fixed point is different from the tail of stationary queue length distribution of the ordinary M/G/1 queue, since for the ordinary M/G/1 queue, the stationary queue length distribution is heavy-tailed if the service time distribution is heavy-tailed, e.g., see Adler, Feldman and Taqqu[I].

| \( \tau \) | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|
| \( \theta \) | \( 1.3 \times 10^{-3} \) | \( 5.3 \times 10^{-2} \) | 0.27 | 0.63 | 1.05 | 1.47 | 1.86 | 2.19 |

Example four: Power law distribution. Let \( \overline{G}(x) = (\mu + x)^{-\alpha} \). If \( 0 < \alpha \leq 1 \), then the power law distribution does not exist the finite mean. In this case, we can not setup the system of integral-differential equations for the fraction density vector \( S(t, x) = (S_0(t), S_1(t, x), S_2(t, x), \ldots) \) which leads to the analysis for the fixed point. Thus we only deal with the case with \( \alpha > 1 \). Note that for each \( \alpha > 1 \)

\[
\int_0^{+\infty} \overline{G}(x) dx = \frac{1}{\mu}
\]
and
\[ \int_{0}^{+\infty} [G(x)]^d \, dx = \frac{1}{\mu}, \]
thus we obtain \( \theta = \mu^{d-1} \). It is easy to see that when \( \mu > 1, \theta > 1 \); when \( \mu = 1, \theta = 1 \); and when \( 0 < \mu < 1, 0 < \theta < 1 \). It follows from Theorem [11] that for \( k \geq 1 \)
\[ \pi_k(x) = \mu^{d-1} - \rho^{d-1} \cdot \mu G(x). \]
This indicates that the fixed point \( \pi = (\pi_0, \pi_1(x), \pi_2(x), \ldots) \) is doubly exponential (of course, it is light-tailed) if the service time distribution is power law.

**Example five:** Almost exponential distribution. Let \( G(x) = \exp \{ -x (\ln x)^{-\alpha} \} \).
Then it is easy to see that the almost exponential distribution is heavy-tailed if \( \alpha > 0 \)
\[ \theta = \frac{\int_{0}^{+\infty} \exp \{ -dx (\ln x)^{-\alpha} \} \, dx}{\left[ \int_{0}^{+\infty} \exp \{ -x (\ln x)^{-\alpha} \} \, dx \right]^d}. \]
Table 3 lists how \( \theta \) depends on the parameter pair \((d, \alpha)\). As seen from Table 3, \( \theta \) is decreasing for each of the two parameters \( d \) and \( \alpha \).

Table 3: \( \theta \) depends on the parameter pair \((d, \alpha)\)

| \((d, \alpha)\) | (2, 2)         | (4, 2)         | (2, 4)         | (4, 4)         |
|----------------|----------------|----------------|----------------|----------------|
| \( \theta \)  | \( 2.24 \times 10^{-2} \) | \( 2.01 \times 10^{-4} \) | \( 3.44 \times 10^{-5} \) | \( 1.18 \times 10^{-13} \) |

5 The key parameter \( \theta \) for PH Service Times

In this section, as an important example we provide three methods to analyze a supermarket model with Poisson arrivals and PH service times. Our purpose is to provide three different ways to determine the key parameter \( \theta \) and compute the doubly exponential solution to the fixed point. Also, we indicate that the doubly exponential solution to the fixed point is not unique for a more general supermarket model.

The supermarket model with Poisson arrivals and PH service times is described as follows. Customers arrive at a queueing system of \( n > 1 \) servers as a Poisson process with arrival rate \( n\lambda \) for \( \lambda > 0 \). The service times of these customers are of phase type with irreducible representation \((\alpha, T)\) of order \( m \). Each arriving customer chooses \( d \geq 1 \) servers independently and uniformly at random from these \( n \) servers, and waits for service at the
server which currently contains the fewest number of customers. If there is a tie, servers with the fewest number of customers will be chosen randomly. All customers in any service center will be served in the FCFS manner. For the PH service time distribution, we use the irreducible representation \((\alpha, T)\) of order \(m\), where the row vector \(\alpha\) is a probability vector whose \(j\)th entry is the probability that a service begins in phase \(j\) for \(1 \leq j \leq m\); and \(T\) is a matrix of order \(m\) whose \((i, j)\)th entry is denoted by \(t_{i,j}\) with \(t_{i,i} < 0\) for \(1 \leq i \leq m\), and \(t_{i,j} \geq 0\) for \(1 \leq i, j \leq m\) and \(i \neq j\). Let \(T^0 = -Te \geq 0\), where \(e\) is a column vector of ones with a suitable dimension in the context. When a PH service time is in phase \(i\), the transition rate from phase \(i\) to phase \(j\) is \(t_{i,j}\), the service completion rate is \(t_i^0\). At the same time, the mean service rate is given by

\[
\mu = -\frac{1}{\alpha T^{-1} e}.
\]

Unless we state otherwise, we assume that all the random variables defined above are independent, and that the system is operating at the stable region: \(\rho = \lambda/\mu < 1\).

We introduce some useful notation. Let \(n_k^{(i)}(t)\) be the number of queues with at least \(k\) customers and the service time in phase \(i\) at time \(t \geq 0\). Clearly, \(0 \leq n_k^{(i)}(t) \leq n\) for \(1 \leq i \leq m\) and \(0 \leq k \leq n\). We define

\[
s_k^{(i)}(t) = \frac{n_k^{(i)}(t)}{n},
\]

which is the fraction of queues with at least \(k\) customers and the service time in phase \(i\).

We write

\[
S_0(t) = (s_0(t))
\]

and for \(k \geq 1\),

\[
S_k(t) = \left( s_k^{(1)}(t), s_k^{(2)}(t), \ldots, s_k^{(m)}(t) \right),
\]

\[
S(t) = (S_0(t), S_1(t), S_2(t), \ldots).
\]

We now introduce Hadamard Product of two matrices \(A = (a_{i,j})\) and \(B = (b_{i,j})\) as follows:

\[
A \odot B = (a_{i,j}b_{i,j}).
\]

Specifically, for \(k \geq 2\) we have

\[
A^{\odot k} = A \odot A \odot \cdots \odot A_{k \text{ matrix } A}.
\]
Let \( a = (a_1, a_2, a_3, \ldots) \). We write
\[
a^{\otimes \frac{1}{2}} = \left( a_1^{\frac{1}{2}}, a_2^{\frac{1}{2}}, a_3^{\frac{1}{2}}, \ldots \right).
\]

Using a similar analysis to that in Equations (4) to (7), we can obtain the following systems of differential vector equations for the fraction density vector \( S(t) = (S_0(t), S_1(t), S_2(t), \ldots) \).

\[
S_0(t) = 1, \quad \text{for} \quad t \geq 0,
\]
\[
\frac{d}{dt} S_0(t) = -\lambda S_0^{\otimes d}(t) + S_1(t) T^0,
\]
\[
\frac{d}{dt} S_1(t) = \lambda_0 S_0^{\otimes d}(t) + \lambda S_1^{\otimes d}(t) + S_1(t) T + S_2(t) T^0 \alpha,
\]
and for \( k \geq 2, \)
\[
\frac{d}{dt} S_k(t) = \lambda S_{k-1}^{\otimes d}(t) - \lambda S_k^{\otimes d}(t) + S_k(t) T + S_{k+1}(t) T^0 \alpha.
\]

If \( \pi = (\pi_0, \pi_1, \pi_2, \ldots) \) is a fixed point of the fraction density vector \( S(t) \), then the system of differential vector equations (21) to (24) can be simplified as

\[
\pi_0 = 1
\]
\[
-\lambda \pi_0^{\otimes d} + \pi_1 T^0 = 0,
\]
\[
\lambda_0 \pi_0^{\otimes d} - \lambda \pi_1^{\otimes d} + \pi_1 T + \pi_2 T^0 \alpha = 0,
\]
and for \( k \geq 2, \)
\[
\lambda \pi_{k-1}^{\otimes d} - \lambda \pi_k^{\otimes d} + \pi_k T + \pi_{k+1} T^0 \alpha = 0.
\]

In what follows we provide three methods to solve the system of nonlinear equations (25) to (28), and give three different doubly exponential solutions to the fixed point.

5.1 The first method

The first method is based on Theorem 1 given in this paper. For the PH service time distribution
\[
\overline{G}(x) = \alpha \exp \{Tx\} e
\]

Let \( \theta = \int_0^{+\infty} \left[ \mu \overline{G}(x) \right]^d dx \), and we assume that \( 0 < \theta < +\infty \). Then the fixed point \( \pi = (\pi_0, \pi_1(x), \pi_2(x), \ldots) \) is given by
\[
\pi_0 = 1,
\]
and for \( k \geq 1 \)
\[
\pi_k(x) = \theta^{\frac{k-1}{d-1}} \rho^{\frac{k}{d-1}} \cdot \mu \overline{G}(x).
\]
5.2 The second method

The second method is proposed in Li, Wang and Liu [11], and the key parameter $\theta$ is based on the stationary probability vector $\omega$ of the irreducible Markov chain $T + T_0\alpha$, that is, $\theta = \omega \odot d_e$.

It follows from Equation (26) that

$$\pi_1 T^0 = \lambda.$$  

Note that

$$\omega T^0 = \mu,$$

$$\frac{\lambda}{\mu} \omega T^0 = \lambda.$$  

Thus, we obtain

$$\pi_1 = \frac{\lambda}{\mu} \omega = \rho \cdot \omega.$$  

Based on the fact that $\pi_0 = 1$ and $\pi_1 = \rho \cdot \omega$, it follows from Equation (27) that

$$\lambda_0 - \lambda \rho^d \cdot \omega \odot d + \rho \cdot \omega T + \pi_2 T^0 \alpha = 0,$$

which leads to

$$\lambda - \lambda \rho^d \cdot \omega \odot d_e + \rho \cdot \omega T e + \pi_2 T^0 = 0.$$  

Note that $\omega T e = -\mu$ and $\rho = \lambda / \mu$, we obtain

$$\pi_2 T^0 = \lambda \rho^d \omega \odot d_e.$$  

Let $\theta = \omega \odot d_e$. Then it is easy to see that $\theta \in (0, 1)$, and

$$\pi_2 T^0 = \theta \rho^d.$$  

Using a similar analysis on Equation (30), we have

$$\pi_2 = \frac{\lambda \theta \rho^d}{\mu} \omega = \theta \rho^{d+1} \cdot \omega.$$  

Based on $\pi_1 = \rho \omega$ and $\pi_2 = \theta \rho^{d+1} \cdot \omega$, it follows from Equation (28) that for $k = 2$,

$$\lambda \rho^d \cdot \omega \odot d - \lambda \theta \rho^{d+2} \cdot \omega \odot d + \theta \rho^{d+1} \cdot \omega T + \pi_3 T^0 \alpha = 0,$$

which leads to

$$\lambda \theta \rho^d - \lambda \theta^{d+1} \rho^{d+2} + \theta \rho^{d+1} \cdot \omega T e + \pi_3 T^0 = 0.$$  

21
thus we obtain
\[ \pi_3 T^0 = \lambda \theta^{d+1} \rho^{d^2+d}. \]

Using a similar analysis on Equation (30), we have
\[ \pi_3 = \frac{\lambda \theta^{d+1} \rho^{d^2+d}}{\mu} \omega = \theta^{d+1} \rho^{d^2+d+1} \cdot \omega. \]

Now, we assume that \( \pi_k = \theta^{d^{k-1}-1} \rho^{d^{k-1}-1} \cdot \omega \) is correct for the cases with \( l = k \). Then it follows from Equation (28) that for \( l = k + 1 \), we have
\[
\begin{align*}
\lambda \theta^{d^k-2+d^k-3+\ldots+d^k-1+d^k-2+\ldots+d^k-1+d^k-1+\ldots+d} \\
+ \theta^{d^k-2+d^k-3+\ldots+d^k-1+d^k-2+\ldots+d^k-1+\ldots+d} \cdot \lambda \theta^{d^k-1+d^k-2+\ldots+d^k-1+d^k-1+\ldots+d} \\
+ \theta^{d^k-2+d^k-3+\ldots+d^k-1+d^k-2+\ldots+d^k-1+\ldots+d} \cdot \omega T + \pi_{k+1} T^0 \alpha = 0,
\end{align*}
\]
which leads to
\[
\begin{align*}
\lambda \theta^{d^k-2+d^k-3+\ldots+d^k-1+d^k-2+\ldots+d^k-1+d^k-1+\ldots+d} \\
+ \theta^{d^k-2+d^k-3+\ldots+d^k-1+d^k-2+\ldots+d^k-1+\ldots+d} \cdot \omega T + \pi_{k+1} T^0 = 0,
\end{align*}
\]
thus we obtain
\[ \pi_{k+1} T^0 = \lambda \theta^{d^{k-1}+d^{k-2}+\ldots+d+1+d^{k-1}+d^{k-2}+\ldots+d+1+d^{k-1}+d^{k-2}+\ldots+d+1} \cdot \omega. \]

By a similar analysis to (30), we have
\[ \pi_{k+1} = \frac{\lambda \theta^{d^{k-1}+d^{k-2}+\ldots+d+1+d^{k-1}+d^{k-2}+\ldots+d+1+d^{k-1}+d^{k-2}+\ldots+d+1}}{\mu} \omega \]
\[ = \theta^{d^{k-1}+d^{k-2}+\ldots+d+1+d^{k-1}+d^{k-2}+\ldots+d+1} \cdot \omega. \]

Therefore, by induction the fixed point \( \pi = (\pi_0, \pi_1, \pi_2, \ldots) \) is given by
\[ \pi_0 = 1, \]
and for \( k \geq 1 \)
\[ \pi_k = \theta^{d^{k-1}-1} \rho^{d^{k-1}-1} \cdot \omega. \] (31)

5.3 The third method

The third method is based on the matrix computation for the system of nonlinear equations (25) to (28), and shows that the key parameter \( \theta \) is based on the initial probability vector \( \alpha \) in the PH service time distribution, that is, \( \theta = 1/\alpha \odot e. \)
It follows from (25) to (28) that

\[
\left(\pi_1^{\otimes d}, \pi_2^{\otimes d}, \pi_3^{\otimes d}, \ldots\right) \begin{pmatrix}
-\lambda & \lambda \\
-\lambda & \lambda \\
-\lambda & \lambda \\
\vdots & \ddots
\end{pmatrix} + \left(\pi_1, \pi_2, \pi_3, \ldots\right) \begin{pmatrix}
T \\
T^0 \alpha \\
T
\end{pmatrix}
\]

\[= - (\lambda \alpha, 0, 0, \ldots),\]

which leads to

\[
(\pi_1, \pi_2, \pi_3, \ldots) = \left(\pi_1^{\otimes d}, \pi_2^{\otimes d}, \pi_3^{\otimes d}, \ldots\right) \begin{pmatrix}
R & V \\
R & V \\
\ddots & \ddots
\end{pmatrix} + \left(\lambda \alpha (-T)^{-1}, 0, 0, \ldots\right),
\]

where

\[V = \lambda (-T)^{-1}\]

and

\[R = \lambda (-I + e\alpha) (-T)^{-1}.\]

Thus we obtain

\[\pi_1 = \lambda \alpha (-T)^{-1} + \pi_1^{\otimes d} \left[\lambda (-I + e\alpha) (-T)^{-1}\right]\]

and for \(k \geq 2\)

\[\pi_k = \pi_k^{\otimes d} \left[\lambda (-T)^{-1}\right] + \pi_k^{\otimes d} \left[\lambda (-I + e\alpha) (-T)^{-1}\right].\]

To omit the term \(\pi_k^{\otimes d} \left[\lambda (-I + e\alpha) (-T)^{-1}\right]\) for \(k \geq 1\), we assume that \(\{\pi_k, k \geq 1\}\) has the following expression

\[\pi_k = r(k) \alpha^{\otimes \frac{1}{2}}.\]

In this case, we have

\[\pi_k^{\otimes d} \left[\lambda (-I + e\alpha) (-T)^{-1}\right] = r^d(k) \alpha \left[\lambda (-I + e\alpha) (-T)^{-1}\right] = 0,
\]

thus it follows from (32) and (33) that

\[\pi_1 = \lambda \alpha (-T)^{-1}\]

(34)
and for $k \geq 2$

$$\pi_k = \pi_{k-1} \left[ \lambda (-T)^{-1} \right].$$

(35)

It follows from (34) that

$$r(1) \alpha^{1/4} = \lambda \alpha (-T)^{-1},$$

which follows that

$$r(1) = \theta \rho,$$

where

$$\theta = \frac{1}{\alpha^{1/4} \rho}.$$

It follows from (35) that

$$r(k) \alpha^{1/4} = r^d(k-1) \alpha \left[ \lambda (-T)^{-1} \right],$$

which follows that

$$r(k) = r^d(k-1) \theta \rho = (\theta \rho)^{\frac{d-1}{d-1}}.$$

Therefore, we can obtain

$$\pi_0 = 1$$

and for $k \geq 1$

$$\pi_k = (\theta \rho)^{\frac{d-1}{d-1}} \cdot \alpha^{1/4}.$$  

(36)

5.4 Non-uniqueness

Based on the above three methods, we can summarize the key parameter and the doubly exponential solution to the fixed point in the following table.

| Method | Key Parameter | Fixed point |
|--------|---------------|-------------|
| Method 1 | $\theta = \frac{\int_0^\alpha \exp(Tx) [\alpha \exp(Tx)]^d dx}{\alpha (-T)^{d-1}}$ | $\pi_k = \theta^{\frac{d}{d-1}} \rho^{\frac{d}{d-1}} \cdot \mu G(x)$ |
| Method 2 | $\theta = \omega^{1/4} \rho$ | $\pi_k = \theta^{\frac{d}{d-1}} \rho^{\frac{d}{d-1}} \cdot \omega$ |
| Method 3 | $\theta = 1/\alpha^{1/4} \rho$ | $\pi_k = (\theta \rho)^{\frac{d}{d-1}} \cdot \alpha^{1/4}$ |

When the PH service time is an $m$-order Erlang distribution with the irreducible representation $(\alpha, T)$, where

$$\alpha = (1, 0, \ldots, 0)$$
and
\[
T = \begin{pmatrix}
-\eta & \eta \\
-\eta & \eta \\
\cdots & \cdots \\
-\eta & \eta \\
-\eta & -\eta
\end{pmatrix}, \quad T^0 = \begin{pmatrix}
0 \\
0 \\
\vdots \\
0 \\
-\eta
\end{pmatrix}.
\]

We have
\[
\alpha^{\frac{d}{s}} = (1, 0, \ldots, 0)
\]
and
\[
\theta = \frac{1}{\alpha^{\frac{d}{s}e}} = 1.
\]
Thus the doubly exponential solution by the third method is given by
\[
\pi_k = \rho^{\frac{d}{s-1}} \cdot (1, 0, \ldots, 0), \quad k \geq 1. \tag{37}
\]

It is clear that
\[
T + T^0 \alpha = \begin{pmatrix}
-\eta & \eta \\
-\eta & \eta \\
\cdots & \cdots \\
-\eta & \eta \\
\eta & -\eta
\end{pmatrix},
\]
which leads to the stationary probability vector of the Markov chain \(T + T^0 \alpha\) as follows:
\[
\omega = \left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right),
\]
\[
\mu = \omega T^0 = \frac{\eta}{m},
\]
\[
\rho = \frac{\lambda}{\mu} = m\frac{\lambda}{\eta}
\]
and
\[
\theta = \omega^{\frac{d}{s}e} = m \left(\frac{1}{m}\right)^d = m^{1-d}.
\]

Thus the doubly exponential solution by the second method is given by
\[
\pi_k = \theta^{\frac{d}{s-1}} \rho^{\frac{d}{s-1}} \left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right)
\]
\[
= \rho^{\frac{d}{s-1}} \cdot \left(m^{-d^{k-1}}, m^{-d^{k-1}}, \ldots, m^{-d^{k-1}}\right), \quad k \geq 1. \tag{38}
\]

It is clear that the three doubly exponential solutions (29), (37) and (38) are different for \(m \geq 2\).
Remark 4 For the supermarket model with Poisson arrivals and PH service times, we have obtained three different doubly exponential solutions to the fixed point. It is interesting but difficult how to be able to find another new doubly exponential solution. We believe that it is an open problem how to give all the doubly exponential solutions to the fixed point for a more general supermarket model including the case with MAP arrivals, PH service times or general service times.

6 Exponential convergence to the fixed point

In this section, we study the exponential convergence of the current location $S(t, x)$ of the supermarket model to its fixed point $\pi(x)$ for $t \geq 0$ and $x \geq 0$. Not only does the exponential convergence indicate the existence of the fixed point, but it also explains such a convergent process is very fast.

For the supermarket model, the initial point $S(0, x)$ can affect the current location $S(t, x)$ for each $t > 0$, since the service process in the supermarket model is under a unified structure. Here, we provide notation for comparison of two vectors. Let $a = (a_1, a_2, a_3, \ldots)$ and $b = (b_1, b_2, b_3, \ldots)$. We write $a \prec b$ if $a_k < b_k$ for all $k \geq 1$; $a \preceq b$ if $a_k \leq b_k$ for all $k \geq 1$.

Now, we can obtain the following useful proposition whose proof is clear from a sample path analysis and is omitted here.

Proposition 2 If $S(0, x) \preceq \tilde{S}(0, x)$, then $S(t, x) \preceq \tilde{S}(t, x)$.

Based on Proposition 2, the following theorem shows that the fixed point $\pi(x)$ is an upper bound of the current location $S(t, x)$ for all $t \geq 0$ and $x \geq 0$.

Theorem 4 For the supermarket model, if there exists some $k$ such that $S_k(0, x) = 0$, then the sequence $\{S_k(t, x)\}$ has a upper bound sequence which decreases doubly exponentially for all $t \geq 0$ and $x \geq 0$, that is, $S(t, x) \preceq \pi(x)$ for all $t \geq 0$ and $x \geq 0$.

Proof Let

$$\tilde{S}_0(0) = \pi_0$$
$$\tilde{S}_k(0, x) = \pi_k(x), \ k \geq 1.$$ 

Then $\tilde{S}_0(t) = \tilde{S}_0(0) = \pi_0$, and for each $k \geq 1$, $\tilde{S}_k(t, x) = \tilde{S}_k(0, x) = \pi_k(x)$ for all $t \geq 0$ and $x \geq 0$, since $\tilde{S}(0)$ is a fixed point in the supermarket model. If $S_k(0, x) = 0$ for
some \(k\), then \(S_k(0, x) < \tilde{S}_k(0, x)\) and \(S_j(0, x) \leq \tilde{S}_j(0, x)\) for all \(j \geq 1\) and \(j \neq k\), thus \(S(0, x) \leq \tilde{S}(0, x)\). It is easy to see from Proposition 2 that \(S_k(t, x) \leq \tilde{S}_k(t, x) = \pi_k(x)\) for all \(k \geq 1, t \geq 0\) and \(x \geq 0\). Thus we obtain that for all \(k \geq 1, t \geq 0\) and \(x \geq 0\)

\[
S_k(t, x) \leq \pi_k(x) = \frac{\rho \cdot 1 - \lambda c_k(t)}{\mu G(x)}.
\]

This completes the proof.

To show the exponential convergence, we use Theorem 4 to define a potential function (or Lyapunov function) \(\Phi(t)\) as follows:

\[
\Phi(t) = \sum_{k=1}^{\infty} w_k \int_{0}^{\infty} [\pi_k(x) - S_k(t, x)] dx,
\]

where \(\{w_k\}\) is a positive scalar sequence with \(w_k > w_{k-1} \geq w_1 = 1\) for \(k \geq 2\). Note that \(\pi_0 = S_0(t) = 1\). It is easy to see from Proposition 2 that \(\Phi(t) \geq 0\) for all \(t \geq 0\).

When \(\int_{0}^{\infty} [\pi_k(x) - S_k(t, x)] dx > 0\) for \(k \geq 1\), we write

\[
\frac{\int_{0}^{\infty} S^d_k(t, x) dx}{\int_{0}^{\infty} [\pi_k(x) - S_k(t, x)] dx} = c_k(t)
\]

and

\[
\frac{\int_{0}^{\infty} \mu(x) S_k(t, x) dx}{\int_{0}^{\infty} [\pi_k(x) - S_k(t, x)] dx} = d_k(t).
\]

The following lemma provide a method to determine the positive scalar sequence \(\{w_k\}\) with \(w_k > w_{k-1} \geq w_1 = 1\) for \(k \geq 2\). This proof is easy by means of some simple computation.

**Lemma 2** If \(\delta\) is a positive constant,

\[
w_1 = 1,
\]

\[
\lambda (w_1 - w_2) c_1(t) = -\delta w_1
\]

and for \(k \geq 2\)

\[
\lambda (w_k - w_{k+1}) c_k(t) + (w_k - w_{k-1}) d_k(t) = -\delta w_k,
\]

then

\[
w_2 = 1 + \frac{\delta}{\lambda c_1(t)},
\]

and for \(k \geq 3\)

\[
w_k = w_{k-1} + \frac{\delta w_{k-1} + (w_{k-1} - w_{k-2}) d_{k-1}(t)}{\lambda c_{k-1}(t)}.
\]
The following theorem measures the distance \( \Phi (t) \) of the current location \( S (t, x) \) for \( t \geq 0 \) and the fixed point \( \pi (x) \) for \( x \geq 0 \), and illustrates that the distance \( \Phi (t) \) to the fixed point from the current location is very close to zero with exponential convergence. Hence, it shows that from any suitable starting point, the supermarket model can be quickly close to the fixed point, that is, there always exists a fixed point in the supermarket model.

**Theorem 5** For \( t \geq 0 \),

\[
\Phi (t) \leq c_0 e^{-\delta t},
\]

where \( c_0 \) and \( \delta \) are two positive constants, and they possibly depend on time \( t \geq 0 \). In this case, the potential function \( \Phi (t) \) is exponentially convergent.

**Proof** Note that

\[
\Phi (t) = \sum_{k=1}^{\infty} w_k \int_{0}^{+\infty} [\pi_k (x) - S_k (t, x)] \, dx,
\]

we have

\[
\frac{d}{dt} \Phi (t) = \frac{d}{dt} \sum_{k=1}^{\infty} w_k \int_{0}^{+\infty} [\pi_k (x) - S_k (t, x)] \, dx
\]

\[
= - \sum_{k=1}^{\infty} w_k \frac{d}{dt} \int_{0}^{+\infty} S_k (t, x) \, dx
\]

by means of the Dominated Convergence Theorem. It follows from (4) to (7) that

\[
\int_{0}^{+\infty} \mu (x) S_1 (t, x) \, dx = \lambda,
\]

(39)
and using (39) we obtain

\[
\frac{d}{dt} \Phi (t) = - \sum_{k=1}^{\infty} w_k \frac{d}{dt} \int_{0}^{+\infty} S_k (t, x) \, dx
\]

\[
= - w_1 \left[ \lambda - \int_{0}^{+\infty} S_1^d (t, x) \, dx \right]
\]

\[
- \int_{0}^{+\infty} \mu (x) S_1 (t, x) \, dx + \int_{0}^{+\infty} \mu (x) S_2 (t, x) \, dx
\]

\[
- \sum_{k=2}^{\infty} w_k \lambda \left[ \int_{0}^{+\infty} S_{k-1}^d (t, x) \, dx - \lambda \int_{0}^{+\infty} S_k^d (t, x) \, dx \right]
\]

\[
= - w_1 \left[ - \lambda \int_{0}^{+\infty} S_1^d (t, x) \, dx + \int_{0}^{+\infty} \mu (x) S_2 (t, x) \, dx \right]
\]

\[
- \sum_{k=2}^{\infty} \lambda \left[ \int_{0}^{+\infty} S_{k-1}^d (t, x) \, dx - \lambda \int_{0}^{+\infty} S_k^d (t, x) \, dx \right]
\]

\[
\leq - \delta w_1
\]

which follows

\[
\frac{d}{dt} \Phi (t) = \lambda (w_1 - w_2) c_1 (t) \int_{0}^{+\infty} \left[ \pi_k (x) - S_k (t, x) \right] \, dx
\]

\[
+ \sum_{k=2}^{\infty} \left[ \lambda (w_k - w_{k+1}) c_k (t) + (w_k - w_{k-1}) d_k (t) \right]
\]

\[
\cdot \int_{0}^{+\infty} \left[ \pi_k (x) - S_k (t, x) \right] \, dx.
\]

Using Lemma 2 we can easily choose a parameter \( \delta > 0 \) and a suitable positive scalar sequence \( \{w_k\} \) with \( w_k > w_{k-1} \geq w_1 = 1 \) for \( k \geq 2 \) such that

\[
\lambda (w_1 - w_2) c_1 (t) \leq -\delta w_1
\]

and for \( k \geq 2 \)

\[
(w_k - w_{k-1}) d_k (t) - \lambda (w_{k+1} - w_k) c_k (t) \leq -\delta w_k,
\]

thus we can obtain

\[
\frac{d}{dt} \Phi (t) \leq -\delta \Phi (t),
\]

which leads to

\[
\Phi (t) \leq c_0 e^{-\delta t}.
\]

This completes the proof.
Remark 5 We have provided an algorithm for computing the positive scalar sequence \( \{w_k\} \) with \( 1 = w_1 \leq w_{k-1} < w_k \) for \( k \geq 2 \) as follows:

Step one:

\[ w_1 = 1. \]

Step two:

\[ w_2 = 1 + \frac{\delta}{\lambda c_1(t)}. \]

Step three: for \( k \geq 2 \)

\[ w_k = w_{k-1} + \frac{\delta w_{k-1} + (w_{k-1} - w_{k-2}) d_{k-1}(t)}{\lambda c_{k-1}(t)}. \]

This illustrates that \( w_k \) is a function of time \( t \). Note that \( \lambda, \delta, c_k(t), d_l(t) > 0 \), it is clear that for \( k \geq 2 \)

\[ 1 = w_1 \leq w_{k-1} < w_k. \]

7 Lipschitz Condition

In this section, we apply the Kurtz Theorem to study the supermarket model with general service times, and analyze the Lipschitz condition with respect to general service times.

The supermarket model can be analyzed by a density dependent jump Markov process, where the density dependent jump Markov process is a Markov process with a single parameter \( n \) which corresponds to the population size. Kurtz’s work provides a basis for the density dependent jump Markov processes in order to relate the infinite-size system of differential equations to the corresponding finite-size system of differential equations. Readers may refer to Kurtz [9] for more details.

In the supermarket model, the states of density dependent jump Markov process can be normalized and interpreted as measuring population densities, so that the transition rates depend only on these densities. Hence, the infinite-size system of differential equations can be regarded as the limiting model of the corresponding finite-size system of differential equations as the population size grows arbitrarily large. When the population size is \( n \), we write

\[ E_n = \{k : k = 0, 1, \ldots, n\}. \]

For \( k \geq 1 \) and \( x \geq 0 \), we write

\[ s_k^{(n)}(x) = \left( \frac{k}{n}, x \right). \]
where $x$ is the residual service time of each server, and
\[ s_{k}^{(n)} = \int_{0}^{+\infty} s_{k}^{(n)}(x) \, dx. \]

Let
\[ S_0 = \lim_{n \to \infty} s_0^{(n)} \]
and for $k \geq 1$
\[ S_k = \lim_{n \to \infty} \int_{0}^{+\infty} s_k^{(n)}(x) \, dx. \]

Let $\{ X_n(t) : t \geq 0 \}$ be a density dependent jump Markov process on the state space $E_n$ whose transition rates are given by
\[ q_{k,k+1}^{(n)} = n \beta_l \left( \frac{k}{n} \right) = n \beta_l \left( s_k^{(n)} \right). \]

In this supermarket model, $\hat{X}_n(t)$ is the unscaled process which records the number of servers with at least $k$ customers for $0 \leq k \leq n$.

Let $a$ and $b$ denote an arrival and a service completion, respectively. Hence taking $l = a$ or $b$ for $a > b > 0$, we write
\[
\begin{align*}
\beta_a \left( s_0^{(n)} \right) &= -\lambda, \\
\beta_b \left( s_0^{(n)} \right) &= \int_{0}^{+\infty} \mu(x) s_1^{(n)}(x) \, dx; \\
\beta_a \left( s_1^{(n)} \right) &= \lambda - \lambda \int_{0}^{+\infty} \left[ s_1^{(n)}(x) \right] d x, \\
\beta_b \left( s_1^{(n)} \right) &= -\int_{0}^{+\infty} \mu(x) s_1^{(n)}(x) \, dx + \int_{0}^{+\infty} \mu(x) s_2^{(n)}(x) \, dx; \\
\end{align*}
\]
and for $n \geq k \geq 2$,
\[
\begin{align*}
\beta_a \left( s_k^{(n)} \right) &= \lambda \int_{0}^{+\infty} \left[ s_{k-1}^{(n)}(x) \right] d x - \lambda \int_{0}^{+\infty} \left[ s_k^{(n)}(x) \right] d x, \\
\beta_b \left( s_k^{(n)} \right) &= -\int_{0}^{+\infty} \mu(x) s_k^{(n)}(x) \, dx + \int_{0}^{+\infty} \mu(x) s_{k+1}^{(n)}(x) \, dx.
\end{align*}
\]

Using Chapter 7 in Kurtz [9] or Subsection 3.4.1 in Mitzenmacher [20], the Markov process $\{ \hat{X}_n(t) : t \geq 0 \}$ with transition rates $q_{k,k+1}^{(n)}$ is given by
\[
\hat{X}_n(t) = \hat{X}_n(0) + \sum_{l=a,b} \left[ \int_{0}^{t} \beta_l \left( \frac{\hat{X}_n(u)}{n} \right) \, du \right], \quad (40)
\]
where $Y_l(x)$ for $l = a$ and $b$ are two independent standard Poisson processes. Clearly, the jump Markov process by Equation (40) at time $t$ is determined by the starting point and the transition rates which are integrated over its history.

Let

$$F(y) = a\beta_a(y) + b\beta_b(y).$$

(41)

Taking $X_n(t) = n^{-1}\hat{X}_n(t)$ which is an appropriate scaled process, we have

$$X_n(t) = X_n(0) + \sum_{l=a,b} \ln^{-1}\hat{Y}_l \left( n \int_0^t \beta_l(X_n(u)) \, du \right) + \int_0^t F(X_n(u)) \, du,$$

(42)

where $\hat{Y}_l(y) = Y_l(y) - y$ is a Poisson process centered at its expectation. Note that in (42), the function $F(y)$ given in (41) is for $y = s^{(n)}_k, 0 \leq k \leq n$.

Taking $X(t) = \lim_{n\to\infty} X_n(t)$ and $x_0 = \lim_{n\to\infty} X_n(0)$, we obtain

$$X(t) = x_0 + \int_0^t F(X(u)) \, du, \quad t \geq 0,$$

(43)

due to the fact that

$$\lim_{n\to\infty} \frac{1}{n} \hat{Y}_l \left( n \int_0^t \beta_l(X_n(u)) \, du \right) = 0$$

by means of the law of large numbers. Note that in (43), the function $F(y)$ given in (41) is for $y = S_k, k \geq 1$. In the supermarket model, the deterministic and continuous process $\{X(t), t \geq 0\}$ is described by the infinite-size system of integral-differential equations (4) to (7), or simply in the below

$$\frac{d}{dt}X(t) = F(X(t))$$

(44)

with the initial condition

$$X(0) = x_0.$$  

(45)

Now, we consider the uniqueness of the limiting deterministic process $\{X(t), t \geq 0\}$ with (44) to (45), or the uniqueness of solution to the infinite-size system of integral-differential equations (4) to (7). To that end, a sufficient condition is Lipschitz, that is, for some constant $M > 0$,

$$|F(y) - F(z)| \leq M|y - z|.$$  

In general, the Lipschitz condition is standard and sufficient for the uniqueness of solution to the finite-size system of differential equations; while for the countable infinite-size case, readers may refer to Theorem 3.2 in Deimling [4] and Subsection 3.4.1 in Mitzenmacher [20] for some generalization.
To check the Lipschitz condition, as $n \to \infty$ we have

$$
\beta_a (S_0) = -\lambda, \\
\beta_b (S_0) = \int_0^{+\infty} \mu (x) S_1 (x) \, dx; \\
\beta_a (S_1) = \lambda - \lambda \int_0^{+\infty} [S_1 (x)]^d \, dx, \\
\beta_b (S_1) = -\int_0^{+\infty} \mu (x) S_1 (x) \, dx + \int_0^{+\infty} \mu (x) S_2 (x) \, dx;
$$

and for $k \geq 2$,

$$
\beta_a (S_k) = \lambda \int_0^{+\infty} [S_{k-1} (x)]^d \, dx - \lambda \int_0^{+\infty} [S_k (x)]^d \, dx, \\
\beta_b (S_k) = -\int_0^{+\infty} \mu (x) S_k (x) \, dx + \int_0^{+\infty} \mu (x) S_{k+1} (x) \, dx.
$$

Let

$$
\zeta_k = \frac{\int_0^{+\infty} [S_k (x)]^d \, dx}{\int_0^{+\infty} S_k (x) \, dx}
$$

and

$$
\eta_k = \frac{\int_0^{+\infty} \mu (x) S_k (x) \, dx}{\int_0^{+\infty} S_k (x) \, dx}.
$$

Then $\zeta_k, \eta_k > 0$ for $k \geq 1$.

The following theorem shows that the supermarket model with general service times satisfies the Lipschitz condition for the infinite-size system of integral-differential equations (1) to (7).

**Theorem 6** The supermarket model with general service times satisfies the Lipschitz condition.

**Proof** Let

$$
\Omega = \{S_k : k \geq 0\}.
$$

For two arbitrary entries $y, z \in \Omega$, we have

$$
|F(y) - F(z)| \leq a|\beta_a(y) - \beta_a(z)| + b|\beta_b(y) - \beta_b(z)|.
$$

Now, we analyze the following four cases for the function $\beta_a(y)$, while the function $\beta_b(y)$ can be analyzed similarly.
Case one: \( y = S_0, z = S_1 \). In this case, we have

\[
|\beta_a (y) - \beta_a (z)| = | - \lambda - \lambda + \lambda \int_0^{+\infty} [S_1 (x)]^d \, dx |
\]

\[
= \lambda | 2 - \zeta_1 \int_0^{+\infty} S_1 (x) \, dx |
\]

\[
= \lambda \left[ 2 - \zeta_1 \int_0^{+\infty} S_1 (x) \, dx \right],
\]

since \( 0 < \zeta_1, \int_0^{+\infty} S_1 (x) \, dx < 1 \). Taking

\[
M_a (0, 1) \geq \frac{2 - \zeta_1 \int_0^{+\infty} S_1 (x) \, dx}{2 - \int_0^{+\infty} S_1 (x) \, dx},
\]

it is clear that

\[
|\beta_a (y) - \beta_a (z)| \leq M_a (0, 1) \lambda \left[ 2 - \int_0^{+\infty} S_1 (x) \, dx \right]
\]

\[
= M_a (0, 1) |y - z|.
\]

Case two: \( y = S_0, z = S_k \) for \( k \geq 2 \). In this case, we have

\[
|\beta_a (y) - \beta_a (z)| = | - \lambda - \lambda + \lambda \int_0^{+\infty} [S_{k-1} (x)]^d \, dx + \lambda \int_0^{+\infty} [S_k (x)]^d \, dx |
\]

\[
= \lambda | - 1 - \zeta_{k-1} \int_0^{+\infty} S_{k-1} (x) \, dx + \zeta_k \int_0^{+\infty} S_k (x) \, dx |
\]

\[
= \lambda \left[ 1 + \zeta_{k-1} \int_0^{+\infty} S_{k-1} (x) \, dx - \zeta_k \int_0^{+\infty} S_k (x) \, dx \right]
\]

due to that \( 0 < \zeta_k, \int_0^{+\infty} S_k (x) \, dx < 1 \). Let

\[
M_a (0, k) \geq \frac{1 + \zeta_{k-1} \int_0^{+\infty} S_{k-1} (x) \, dx - \zeta_k \int_0^{+\infty} S_k (x) \, dx}{1 + \int_0^{+\infty} S_{k-1} (x) \, dx - \int_0^{+\infty} S_k (x) \, dx}.
\]

Then

\[
|\beta_a (y) - \beta_a (z)| \leq M_a (0, k) \lambda \left[ 1 + \int_0^{+\infty} S_{k-1} (x) \, dx - \int_0^{+\infty} S_k (x) \, dx \right]
\]

\[
= M_a (0, k) |y - z|.
\]

Case three: \( y = S_1, z = S_k \) for \( k \geq 2 \). In this case, we have

\[
|\beta_a (y) - \beta_a (z)| = | - \lambda \int_0^{+\infty} [S_1 (x)]^d \, dx - \lambda \int_0^{+\infty} [S_{k-1} (x)]^d \, dx + \lambda \int_0^{+\infty} [S_k (x)]^d \, dx |
\]

\[
= \lambda |1 - \zeta_1 \int_0^{+\infty} S_1 (x) \, dx - \zeta_{k-1} \int_0^{+\infty} S_{k-1} (x) \, dx + \zeta_k \int_0^{+\infty} S_k (x) \, dx |
\]
Let
\[ M_a (1, k) \geq \frac{1 - \zeta_1 \int_0^{+\infty} S_1 (x) \, dx - \zeta_{k-1} \int_0^{+\infty} S_{k-1} (x) \, dx + \zeta_k \int_0^{+\infty} S_k (x) \, dx}{1 - \int_0^{+\infty} S_1 (x) \, dx - \int_0^{+\infty} S_{k-1} (x) \, dx + \int_0^{+\infty} S_k (x) \, dx}. \]

Then
\[ |\beta_a (y) - \beta_a (z)| \leq M_a (1, k) \lambda |1 - \int_0^{+\infty} S_1 (x) \, dx - \int_0^{+\infty} S_{k-1} (x) \, dx + \int_0^{+\infty} S_k (x) \, dx| = M_a (1, k) |y - z|. \]

Case four: \( y = S_l, z = S_k \) for \( k > l \geq 2 \). In this case, we have
\[ |\beta_a (y) - \beta_a (z)| = \lambda \int_0^{+\infty} [S_{l-1} (x)]^d \, dx - \lambda \int_0^{+\infty} [S_l (x)]^d \, dx - \lambda \int_0^{+\infty} [S_{k-1} (x)]^d \, dx + \lambda \int_0^{+\infty} [S_k (x)]^d \, dx \]
\[ = \lambda |\zeta_{l-1} \int_0^{+\infty} S_{l-1} (x) \, dx + \zeta_l \int_0^{+\infty} S_l (x) \, dx - \zeta_{k-1} \int_0^{+\infty} S_{k-1} (x) \, dx - \zeta_k \int_0^{+\infty} S_k (x) \, dx|. \]

Let
\[ M_a (l, k) \geq \frac{|\zeta_{l-1} \int_0^{+\infty} S_{l-1} (x) \, dx + \zeta_l \int_0^{+\infty} S_l (x) \, dx - \zeta_{k-1} \int_0^{+\infty} S_{k-1} (x) \, dx + \zeta_k \int_0^{+\infty} S_k (x) \, dx|}{|\int_0^{+\infty} S_{l-1} (x) \, dx + \int_0^{+\infty} S_l (x) \, dx - \int_0^{+\infty} S_{k-1} (x) \, dx + \int_0^{+\infty} S_k (x) \, dx|.} \]

Then
\[ |\beta_a (y) - \beta_a (z)| \leq M_a (l, k) |y - z|. \]

Based on the above four cases, taking
\[ M_a = \max \{ M_a (l, k) : k > l \geq 0 \} \]
we obtain that for two arbitrary entries \( y, z \in \Omega \),
\[ |\beta_a (y) - \beta_a (z)| \leq M_a |y - z|. \]

Similarly, we can choose a positive number \( M_b \) such that for two arbitrary entries \( y, z \in \Omega \),
\[ |\beta_b (y) - \beta_b (z)| \leq M_b |y - z|. \]

Let \( M = \max \{ aM_a, bM_b \} \). Then for two arbitrary entries \( y, z \in \Omega \),
\[ |F (y) - F (z)| \leq M |y - z|. \]
This completes the proof.

Based on Theorem 6, the following theorem easily follows from Theorem 3.13 in Mitzenmacher [20].

**Theorem 7** In the supermarket model with general service times, \( \{X_n(t)\} \) and \( \{X(t)\} \) are respectively given by (42) and (43), we have

\[
\lim_{n \to \infty} \sup_{u \leq t} |X_n(u) - X(u)| = 0, \quad a.s.
\]

**Proof** It is seen from that in the supermarket model with general service times, the function \( F(y) \) for \( y \in \Omega \) satisfies the Lipschitz condition. At the same time, it is easy to take a subset \( \Omega^* \subset \Omega \) such that

\[
\{X(u) : u \leq t\} \subset \Omega^*
\]

and

\[
a \sup_{y \in \Omega^*} \beta_a(y) + a \sup_{y \in \Omega^*} \beta_a(y) < +\infty.
\]

Thus, this proof can easily be completed by means of Theorem 3.13 in Mitzenmacher [20]. This completes the proof.

Using Theorem 3.11 in Mitzenmacher [20] and Theorem 7, we can obtain the following theorem for the expected sojourn time that a customer spends in an initially empty supermarket model with general service times over the time interval \([0, T]\).

**Theorem 8** In the supermarket model with general service times, the expected sojourn time that a customer spends in an initially empty system over the time interval \([0, T]\) is bounded above by

\[
\theta \rho^d \{E[X_R] - E[X]\} + E[X] \left[ \sum_{k=1}^{\infty} \theta^{\frac{k}{d-1}} \rho^{\frac{k}{d-1}} \right] + o(1),
\]

where \( o(1) \) is understood as \( n \to \infty \).

**8 Concluding remarks**

In this paper, we provide a novel and simple approach to study the randomized load balancing model with general service times, which is described as an infinite-size system of integral-differential equations. This approach is based on the supplementary variable
method, which is always applied in dealing with stochastic models of M/G/1 type, e.g., see Li and Zhao \[13, 14\] and Li \[10\]. We organize an infinite-size system of integral-differential equations by means of the density dependent jump Markov process, and obtain a close-form solution: doubly exponential structure, for the fixed point satisfying the system of nonlinear equations, which is always a key in the study of supermarket models. Since the fixed point is decomposed into two groups of information under a product form, we indicate three important observations:

1. the fixed point for the supermarket model is different from the tail of stationary queue length distribution for the ordinary M/G/1 queue;

2. the doubly exponential solution to the fixed point can exist extensively for \(0 < \mu < +\infty\) even if the service time distribution is heavy-tailed; and

3. the doubly exponential solution to the fixed point is not unique for a more general supermarket model.

Furthermore, we analyze the exponential convergence of the current location of the supermarket model to its fixed point, and study the Lipschitz condition in the Kurtz Theorem under general service times. Finally, we present numerical examples to illustrate the effectiveness of our approach in analyzing the randomized load balancing schemes with the non-exponential service requirements. Based on this analysis, one can gain a new and important understanding how workload probing can help in load balancing jobs with general service times such as heavy-tailed service.

The approach of this paper is useful in analyzing the randomized load balancing schemes in resource allocation in computer networks. We expect that this approach will be applicable to the study other randomized load balancing schemes with general service times, for example, generalizing the arrival process to non-Poisson: the renewal arrival process or the Markovian arrival process.

**Acknowledgements**

The work of Q.L. Li was supported by the National Science Foundation of China under grant No. 10871114 and the National Grand Fundamental Research 973 Program of China under grant No. 2006CB805901.
References

[1] R. Adler, R. Feldman and M.S. Taqqu (1998). *A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions*. Birkhäuser: Boston.

[2] Y. Azar, A.Z. Broder, A.R. Karlin and E. Upfal (1999). Balanced allocations. *SIAM Journal on Computing* **29**, 180–200. A preliminary version of this paper appeared in *Proceedings of the Twenty-Sixth Annual ACM Symposium on the Theory of Computing*, 1994.

[3] M. Dahlin (1999). Interpreting stale load information. In *Proceedings of the Nineteenth Annual IEEE International Conference on Distributed Computing Systems*.

[4] K. Deimling (1977). *Ordinary Differential Equations in Banach Spaces*. Springer-Verlag, Lecture Notes in Math., Vol. 96.

[5] D.L. Eager, E.D. Lazokwska and J. Zahorjan (1986). Adaptive load sharing in homogeneous distributed systems. *IEEE Transactions on Software Engineering* **12**, 662–675.

[6] D.L. Eager, E. D. Lazokwska and J. Zahorjan (1986). A comparison of receiver-initiated and sender-initiated adaptive load sharing. *Performance Evaluation Review* **6**, 53–68.

[7] D.L. Eager, E.D. Lazokwska and J. Zahorjan (1988). The limited performance benefits of migrating active processes for load sharing. *Performance Evaluation Review* **16**, 63–72.

[8] M. harchol-Balter and A.B. Downey. Exploiting process lifetime distributions for dynamic load balancing. *ACM Transactions on Computer Systems* **15**, 253–285, 1997.

[9] T.G. Kurtz (1981). *Approximation of Population Processes*. SIAM.

[10] Q.L. Li (2010). *Constructive Computation in Stochastic Models with Applications: The RG-Factorizations*. Tsinghua Press and Springer.

[11] Q.L. Li, John C.S. Lui and Y. Wang (2010). A matrix-analytic solution for randomized load balancing models with PH service times. Submitted for publication.

[12] Q.L. Li, John C.S. Lui and Y. Wang (2010). Doubly exponential asymptotics for randomized load balancing with Markovian arrival processes. Submitted for publication.
[13] Q.L. Li and Y.Q. Zhao (2004). A MAP/G/1 queue with negative customers. *Queueing Systems* **47**, 5-43.

[14] Q.L. Li, Y. Ying and Y.Q. Zhao (2006). A BMAP/G/1 retrial queue with a server subject to breakdowns and repairs. *Annals of Operations Research* **141**, 233-270.

[15] M. Luczak and C. McDiarmid (2005). On the power of two choices: Balls and bins in continuous time. *The Annals of Applied Probability* **15**, 1733–1764.

[16] M. Luczak and C. McDiarmid (2006). On the maximum queue length in the supermarket model. *The Annals of Probability* **34**, 493–527.

[17] J.B. Martin (2001). Point processes in fast Jackson networks. *Annals of Applied Probability* **11**, 650-663.

[18] J.B. Martin and Y.M Suhov (1999). Fast Jackson networks. *Annals of Applied Probability* **9**, 854–870.

[19] M.D. Mitzenmacher (1996). Load balancing and density dependent jump Markov processes. In *Proceedings of the Thirty-Seventh Annual Symposium on Foundations of Computer Science*, pages 213–222.

[20] M.D. Mitzenmacher (1996). The power of two choices in randomized load balancing. PhD thesis, University of California at Berkeley, Department of Computer Science, Berkeley, CA, 1996.

[21] M. Mitzenmacher (1998). Analyses of load stealing models using differential equations. In *Proceedings of the Tenth ACM Symposium on Parallel Algorithms and Architectures*, pages 212–221.

[22] M. Mitzenmacher (1999). On the analysis of randomized load balancing schemes. *Theory of Computing Systems* **32**, 361–386.

[23] M. Mitzenmacher (1999). Studying balanced allocations with differential equations. *Combinatorics, Probability, and Computing* **8**, 473–482.

[24] M. Mitzenmacher (2000). How useful is old information? *IEEE Transactions on Parallel and Distributed Systems* **11**, 6–20.
[25] M. Mitzenmacher (2001). The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Computing* **12**, 1094-1104.

[26] M. Mitzenmacher, A. Richa, and R. Sitaraman (2001). The power of two random choices: a survey of techniques and results. In *Handbook of Randomized Computing: volume 1*, edited by P. Pardalos, S. Rajasekaran and J. Rolim, pp. 255-312.

[27] M. Mitzenmacher and B. Vöcking (1998). The asymptotics of selecting the shortest of two, improved. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 326–327. A full version is available as Harvard Computer Science TR-08-99.

[28] R. Mirchandaney, D. Towsley, and J.A. Stankovic (1989). Analysis of the effects of delays on load sharing. *IEEE Transactions on Computers* **38**, 1513–1525.

[29] Y.M. Suhov and N.D. Vvedenskaya (2002). Fast Jackson Networks with Dynamic Routing. *Problems of Information Transmission* **38**, 136-153.

[30] B. Vöcking (1999). How asymmetry helps load balancing. In *Proceedings of the Fortieth Annual Symposium on Foundations of Computer Science*, pages 131–140.

[31] N.D. Vvedenskaya, R.L. Dobrushin and F.I. Karpelevich (1996). Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problems of Information Transmissions* **32**, 20–34.

[32] N.D. Vvedenskaya and Y.M. Suhov (1997). Dobrushin’s mean-field approximation for a queue with dynamic routing. *Markov Processes and Related Fields* **3**, 493–526.

[33] S. Zhou (1988). A trace-driven simulation study of dynamic load balancing. *IEEE Transactions on Software Engineering*, 1327–1341.