Fluid and Diffusion Approximations of A Two-Station Mixed Queueing Network

Vien Nguyen

OR 274-93 January 1993

Fluid and Diffusion Approximations of A Two-Station Mixed Queueing Network

Viên Nguyen
Sloan School of Management, M.I.T., Cambridge, MA 02139

Abstract

The subject of this paper is a two-station mixed queueing network with two customer types: "Open" customers enter the network at station 1 and depart the system after receiving service. Meanwhile, a fixed number of "closed" customers circulate between stations 1 and 2 indefinitely. Such a mixed queueing network model can represent a single-stage production system that services both make-to-order and make-to-stock customers. We present fluid and diffusion limits for this network under the first-in-first-out service discipline. We find that the heavy traffic limit of the workload process at station 1 is a reflected Brownian motion (RBM) on a finite interval. This result is surprising in light of the behavior of the original mixed network model, in which the workload at station 1 need not be bounded.

KEYWORDS: mixed queueing networks, make-to-order production, make-to-stock production, diffusion approximation, reflected Brownian motion, performance analysis.

Contents:

January, 1993

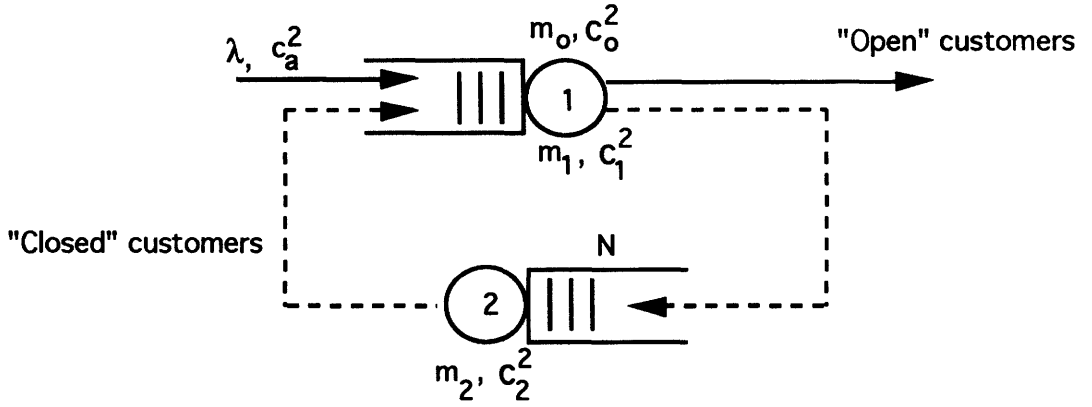

Figure 1: A Two-Station Mixed Queueing Network

1 A Two-Station Mixed Queueing Network

This paper is devoted to the analysis of the network pictured in Figure 1, which consists of two stations serving both "open" and "closed" customers. Open customers enter the network at station 1 and depart the network after being served. Closed customers, on the other hand, circulate between stations 1 and 2 for service. Because there are no external arrivals of closed customers nor are there departures, the number of closed customers in the network remains constant in time. We denote by N the number of closed customers in the system.

Let λ be the arrival rate of open customers at station 1 and let m_0 be the mean service time of these customers. Set m_1 and m_2 to be the mean service times of closed customers at stations 1 and 2, respectively. We will assume throughout this paper that $m_1 < m_2$. From Chen and Mandelbaum [3], one can verify that the relative throughput rate of closed customers is $1/m_2$. Consequently, we can define the "relative" traffic intensities at stations 1 and 2, repsectively, to be

$$\rho_1 = \lambda m_0 + \frac{m_1}{m_2} \tag{1.1}$$

$$\rho_2 = 1. \tag{1.2}$$

For each finite N, the actual throughput rate of closed customers is given by α_N^*/m_2, where α_N^* is a number strictly less than one. Moreover, the actual traffic intensities are given by

$$\rho_1^* = \lambda m_0 + \frac{\alpha_N^*}{m_2} m_1 \tag{1.3}$$

$$\rho_2^* = \alpha_N^*. \tag{1.4}$$

One expects that the traffic intensity at station 2 approaches 1 as the number of closed customers increases; that is, $\alpha_N^* \to 1$ as $N \to \infty$ and accordingly, equations (1.1)-(1.2) can be taken as approximations of (1.3)-(1.4) when N is large.

We are interested in the process $W_1(t)$, defined to be the total amount of work found at station 1 at time t. In addition, let $U_0(t)$ and $U_1(t)$ be that part of the workload corresponding to open and

closed customers, respectively. We will show in this paper that when station 1 is nearly saturated, namely

$$\rho_1 = \lambda m_0 + \frac{m_1}{m_2} \approx 1,$$

the following approximation holds when the closed customer population N is large:

$$\begin{aligned}(W_1^N(\cdot), U_0^N(\cdot), U_1^N(\cdot)) &\equiv \frac{1}{N}(W_1(N^2\cdot), U_0(N^2\cdot), U_1(N^2\cdot)) \\ &\approx (W_1^*(\cdot), U_0^*(\cdot), U_1^*(\cdot)). \end{aligned} \tag{1.5}$$

Here, $W_1^*(\cdot)$ is a reflected Brownian motion on the interval $[0, m_2]$ with drift and variance parameters

$$\theta^* = N\left(\lambda m_0 + \frac{m_1}{m_2} - 1\right) \tag{1.6}$$

$$\sigma^2 = \lambda m_0^2\left(c_a^2 + c_0^2\right) + \frac{m_1^2}{m_2}\left(c_1^2 + c_2^2\right). \tag{1.7}$$

Moreover, the partial workloads corresponding to open and closed customers live in fixed proportions according to

$$U_0^*(t) = \left(\frac{m_2}{m_1} - 1\right) U_1^*(t) = \lambda m_0 W_1^*(t). \tag{1.8}$$

These results seem counterintuitive in light of the behavior of the original mixed queueing network, where the workload process at station 1 need not be bounded above and the partial workloads are not subject to deterministic relationships. On the other hand, neither (1.5) nor (1.8) should be completely surprising if one is familiar with certain properties of queueing networks in heavy traffic. Our goal in this paper is to prove a heavy traffic limit theorem that justifies the approximation stated in (1.5).

For the reader to better understand the contributions of this paper, it is helpful to cast our results within the context of diffusion approximations of open and closed queueing network models. Queueing networks are said to be "multiclass" if the service time distribution as well as the routing of customers at each station can depend on the class designation of the customer. In "single-class" networks, customers at each station are indistinguishable, meaning their service times are identically distributed and all customers at each station follow the same routing mechanism. For single-class networks with Markovian routing, Reiman [19] proved that the diffusion limit of the workload processes is a reflected Brownian motion in the positive orthant. Peterson [18] proved a similar result for multiclass networks in which the routing is deterministic and feedforward. In the same work, Peterson also showed that the class specific workloads at each station are given by fixed proportions of the overall workload at that station. The feedforward structure, which essentially requires that all customers travel from lower numbered stations to higher numbered ones, turns out to be an important restriction. The generalization of Peterson's work to include routing with feedback has proved to be quite difficult and the source of the difficulties contains deep and subtle theoretical issues. In the case of a multiclass single-queue network with feedback,

Reiman [20] was able to prove a theorem to justify the approximation of the workload process by a one-dimensional reflected Brownian motion, and the proof due to Reiman was subsequently simplified by Dai and Kurtz [7]. With insights drawn from these results, Harrison and Nguyen [12, 13] proposed a Brownian system model to approximate a general multiclass queueing network with feedback. The Brownian system model proposed by Harrison and Nguyen is, in essence, a reflected Brownian motion on the nonnegative orthant, and it was generally thought that such an RBM was well defined for any queueing network. Indeed, Dai and Nguyen [8] have shown that if the vector of wokload processes were to converge to any continuous limit, then that limit must be the Brownian system model described in [12, 13]. However, an example by Dai and Wang [9] conclusively verified that there exist queueing networks for which Harrison and Nguyen's Brownian approximation do not exist. Whitt [23] provided another example that further illuminated the irregularities and nonconvergence of the workload process. Due to the work of Taylor and Williams [21], much progress has been made toward identifying sufficient and necessary conditions for the existence and uniqueness of RBM's. However, the convergence of open multiclass queueing networks with feedback remains a wide open question.

Similar progress has been made in the area of diffusion approximations for closed queueing network models. Chen and Mandelbaum [3, 4] have proved fluid and diffusion limit theorems for single class closed queueing networks with Markovian routing. In particular, the diffusion limit of the workload process in such a network is an RBM on a simplex. Extension of Chen and Mandelbaum's work to the multiclass networks involves the same difficulties as in the open counterpart. Dai and Harrison [6] propose a diffusion approximation for a closed manufacturing system, but with the restriction that all job classes which are served at a station share a common service time distribution. From Taylor and Williams [21], one can verify that there exists (in a weak sense) a unique RBM corresponding to the proposed approximation. However, there are no proofs to verify that the workload processes in fact converge to the said RBM.

In light of these results, one may suspect that mixed queueing network models, as a combination of open and closed queueing networks, will exhibit similar properties under the diffusion limits. That is, the diffusion limit of the workload process can be cast in the form of a reflected Brownian motion, and in particular, the workload due to closed customers can be expressed as an RBM on an interval. Moreover, one can conclude from the theory of multiclass queueing networks that the class-specific workload proceses at each station live in fixed proportions, so that the workload processes due to open and closed customers at station 1 are deterministically related. It then follows from these observations that the workload process at station 1 must be an RBM on a finite interval. Because station 1 is in essence a multiclass station, however, the proof of this network is substantially more intricate than the corresponding proof of the single-class open and closed networks. One may therefore view mixed queueing networks as an intermediate stepping stone between the well understood single class networks and the more challenging multiclass queueing models.

The remainder of the paper is organized as follows. Before we turn to problem formulation and proof, we discuss in the next section a make-to-order/make-to-stock production system that is naturally modelled by the mixed queueing network under study. Section 3 defines the processes that we use in our analysis. Our main results are stated in Section 4, and the proof of the limit theorems are then given in Sections 5 and 6. These proofs rely on the properties of a certain pair of mappings, which we discuss in the Appendix. Our approximation scheme is based on a refinement of the Brownian limit, which we discuss in Section 7. Finally, Section 8 contains the results of several numerical experiments.

We end this section with some technical preliminaries. The space $\mathbf{D}^r[0,\infty)$ is the r-dimensional product space of functions $f : [0,\infty) \to \Re^r$ that are right continuous on $[0,\infty)$ and have left limits on $(0,\infty)$. The space $\mathbf{D}^r[0,\infty)$ is endowed with the Skorohod topology [2]. For X^n a sequence of processes in $\mathbf{D}^r[0,\infty)$ and $X \in \mathbf{D}^r[0,\infty)$, we write $X^n \Longrightarrow X$ to mean X^n converges to X in distribution.

All vectors will be envisioned as column vectors. We use the letter e to denote a (column) vector whose components are all ones. The dimension of e should always be clear from context. On occasion, we will also write $e(t)$ to mean the identity map $e(t) = t$. Again, there should be no confusion as to the appropriate interpretation of the letter e.

For $f : [0,\infty) \to \Re$, set

$$||f||_t \equiv \sup_{0 \le s \le t} |f(s)|,$$

and for a vector-valued function $f = (f_1, f_2, \ldots, f_r)' : [0,\infty) \to \Re^r$, we let

$$||f||_t \equiv (||f_1||_t, \ldots, ||f_r||_t)'.$$

A sequence of functions $\{f^n\}$ converges to a function f uniformly on compact sets (u.o.c.) if for each $t \ge 0$, $||f^n - f||_t \to 0$ as $n \to \infty$. We say that f is continuous at x if $x^n \to x$ u.o.c. implies that $f(x^n) \to f(x)$ u.o.c. Finally, for a sequence of functions $\{X^n\}$ on $\mathbf{D}^r[0,\infty)$ and X a process in $\mathbf{D}^r[0,\infty)$, we write $X^n \to X$ u.o.c if almost surely, X^n converges to X uniformly on compact sets.

2 A Make-to-order/Make-to-stock Production system

Production systems are typically categorized as "make-to-order" or "make-to-stock," corresponding to the two scenarios in which new jobs are triggered by customer orders or by replenishment orders for finished goods inventory, respectively. In a make-to-order system, a new job is released into the system each time a customer places an order. A make-to-stock system, on the other hand, maintains a finished goods inventory from which customer demands are filled. Each order fulfillment from inventory triggers a job release in the system; hence, the total number of "jobs" in the system, either in the form of orders waiting to be processed or as finished goods inventory, does not change over time.

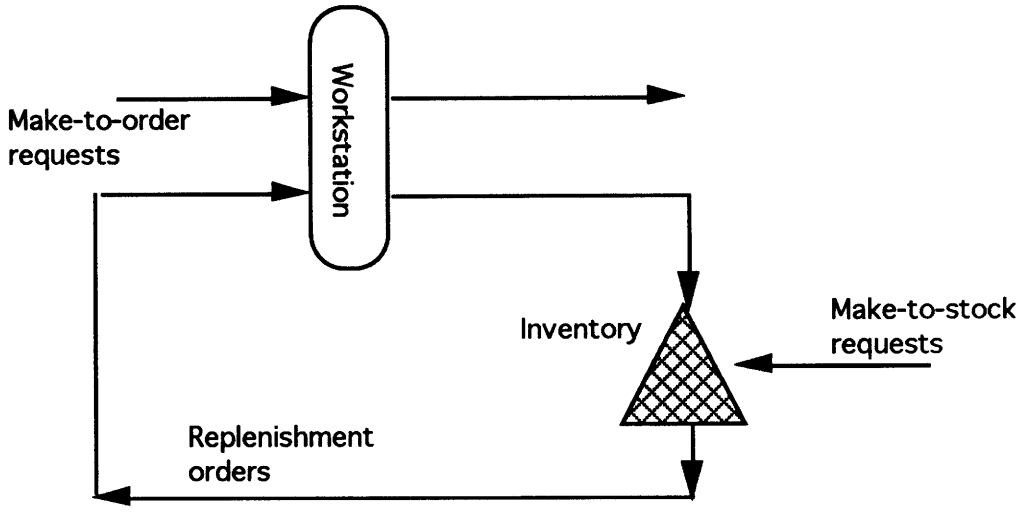

Figure 2: A Make-to-order/Make-to-stock Production System

It is more often the case, however, that production systems employ a combination of both make-to-order and make-to-stock operations. Figure 2 shows such a system with a single processing stage. We make the assumption that orders for make-to-stock products that cannot be filled due to lack of inventory are simply lost (no backlog). One can employ the mixed queueing network in Figure 1 to model this system, where make-to-order products are represented by open customers and closed customers take the place of make-to-stock products. Station 1 naturally represents the production center and we use station 2 to model the finished goods inventory from which make-to-stock orders are serviced. A service at station 2 signals that a make-to-stock request has been filled, which in turn triggers a replenishment order for station 1; that is, a departure from station 2 then proceeds to station 1. In this case, a closed customer at station 1 represents an order for make-to-stock products, whereas a closed customer at station 2 takes the form of a finished good.

We note that station 2 only approximates the demand process for make-to-stock products. In particular, consider a time interval $[t_1, t_2]$ during which the finished goods inventory is empty. During this time period, make-to-stock requests continue arriving with i.i.d. interarrival times. The first order to be filled is the first one to arrive after t_2. Denoting by t^* the arrival time of this order, it is clear that $t^* - t_2$ does not, in general, have the same distribution as other interarrival times. That is, the first service time of a busy period at station 2 should be characterized by an "excess life" distribution. However, this difference is not significant in the sense that both systems can be shown to converge to the same heavy traffic limit (see Iglehart and Whitt [16]).

The subject of this paper is the behavior of such a system as the workstation operates under the first-in-first-out policy. There are, of course, several other policies that should be considered. For example, make-to-order products (or similary, make-to-stock products) may receive higher priority. One can employ an "Order-up-to" policy in which a batch of $N - n$ make-to-stock requests are

sent to the workstation whenever the inventory level n falls below some critical level n^*. Another interesting option is to process the two product types on cyclical basis.

In this example, the parameter λ corresponds to the rate at which make-to-order products are requested. Similarly, m_0 and m_1 are the mean processing times for make-to-order and make-to-stock products, respectively. Finally, the demand rate for make-to-stock products is given by $1/m_2$. It is natural, in the context of this example, to consider only cases in which $m_1 < m_2$.

3 The System Processes

Let $(\Omega, \mathcal{F}, P)$ be a probability space on which are defined four independent sequences of independent and identically distributed (i.i.d.) random variables $\{u(i), i = 1, 2, \ldots\}$, $\{v_k(i), i = 1, 2, \ldots\}$, $k = 0, 1, 2$, where these random variables are positive and have unit mean. We denote by c_a^2 and c_k^2 the squared coefficients of variation (SCV) of $u(i)$ and $v_k(i)$, respectively. We will find it convenient to designate open customers as customers of *class* 0, closed customers at station 1 as *class* 1 customers, and closed customers at station 2 as *class* 2. With this categorization, we set the interarrival time sequence of open customers to be $\{\lambda^{-1}u(i), i = 1, 2, \ldots\}$ and we denote by $m_0 v_0(i)$ the service time of the i^{th} open customer at station 1. Service time sequences for closed customers at stations 1 and 2 are defined as $\{m_1 v_1(i), i = 1, 2, \ldots\}$ and $\{m_2 v_2(i), i = 1, 2, \ldots\}$, respectively. With these definitions, one interprets λ as the arrival rate of open customers and m_0 as the mean service time of open customers at station 1. Similarly, the mean service time of closed customers at stations 1 and 2 are m_1 and m_2, respectively.

Setting $u(0) = 0$, define the arrival process for *open* customers at station 1 as

$$A_0(t) = \max\{i \geq 0 : u(0) + \cdots + u(i) \leq \lambda t\}. \tag{3.1}$$

Next, let $B_j(t)$ be the amount of time that server j has spent working up to time t. At station 1, this time is divided among the two customer types, and we denote by $T_0(t)$ and $T_1(t)$ the amount of time server 1 has devoted to open and closed customers, respectively. (Clearly, $B_1(t) = T_0(t) + T_1(t)$.) Let $S_k(t)$, $k = 0, 1, 2$, be the counting process associated with class k service times,

$$S_k(t) = \max\{i \geq 0 : m_k v_k(0) + \cdots + m_k v_k(i) \leq t\}. \tag{3.2}$$

The arrival processes for class 1 and 2 customers, respectively, are then given by

$$A_1(t) \equiv S_2(B_2(t)), \qquad A_2(t) \equiv S_1(T_1(t)). \tag{3.3}$$

Denote by $V_k(t)$, $k = 0, 1, 2$, the partial sums process for class k service times,

$$V_k(t) \equiv \sum_{i=1}^{\lfloor t \rfloor} m_k v_k(i). \tag{3.4}$$

Setting

$$M_k(t) \equiv V_k(A_k(t)) = m_k v_k(1) + \cdots + m_k v_k(A_k(t)), \tag{3.5}$$

it follows from (3.1)-(3.5) that $M_k(t)$ is the amount of immediate work from class k customers who have arrived to the associated station by time t. If we now define

$$L_1(t) \equiv M_0(t) + M_1(t), \qquad L_2(t) \equiv M_2(t), \tag{3.6}$$

then $L_i(t)$ is the immediate workload input process for all customer classes at station i.

We will assume throughout the paper that $Q_0(0) \equiv Q_1(0) \equiv 0$ and $Q_2(0) \equiv N$; that is, the system starts with all closed customers at station 2 and no open customers in station 1. Letting $W_i(t)$ denote the workload process for station i, defined to be the remaining service time associated with all those customers at station i at time t, either queued or receiving service, we have

$$W_1(0) \equiv 0, \qquad W_2(0) \equiv V_2(N), \tag{3.7}$$

and

$$W_i(t) \equiv W_i(0) + L_i(t) - B_i(t). \tag{3.8}$$

Defining $I_i(t) = t - B_i(t)$ to be the cumulative idleness process at station i and

$$X_i(t) \equiv L_i(t) - t \tag{3.9}$$

to be the workload netflow process, write (3.8) as

$$W_i(t) = W_i(0) + X_i(t) + I_i(t). \tag{3.10}$$

We require the idleness processes satisfy the following properties:

$$I_i \text{ is continuous and nondecreasing with } I_i(0) = 0 \tag{3.11}$$

$$I_i \text{ increases only at times t when } W_i(t) = 0. \tag{3.12}$$

The first statement is a simple consequence of the properties of an idleness process, and the second statement holds for any work-conserving system. That is, it states that the server remains idle only when there is no work to be processed. The vector processes X, W, I, U, Q are then defined in the obvious manner.

It remains to characterize the "allocation" processes $T(t) \equiv (T_0(t), T_1(t))'$. Let $\eta(t)$ denote the arrival time of the customer currently in service at station 1 if $W_1(t) > 0$ and set $\eta(t) = t$ otherwise. With FIFO service discipline, we must have

$$T_k(t) = M_k(\eta(t)) + \epsilon_{1k}(t), \tag{3.13}$$

where $\epsilon_{1k}(t)$ is the amount of service the current customer has received if that task is of class k and $\epsilon_{1k}(t) = 0$ otherwise. The amount of work at station 1 associated with open and closed customers are denoted by U_0 and U_1, respectively, and are given by

$$U_0(t) = M_0(t) - T_0(t) = M_0(t) - M_0(\eta(t)) - \epsilon_{10}(t) \tag{3.14}$$

$$U_1(t) = M_1(t) - T_1(t) = M_1(t) - M_1(\eta(t)) - \epsilon_{11}(t). \tag{3.15}$$

Next, define $Q_k(t)$ to be the queue length process associated with class k customers (including any customer who may be in service). It follows from the previous definitions that

$$Q_0(t) = A_0(t) - A_0(\eta(t)) \tag{3.16}$$

$$Q_1(t) = A_1(t) - A_1(\eta(t)) \tag{3.17}$$

$$Q_2(t) = N + A_2(t) - A_2(\eta(t)). \tag{3.18}$$

As we expect, $Q_1(t)+Q_2(t) = Q_1(0)+Q_2(0) = N$, so the number of closed customers in the network does not fluctuate in time. Furthermore, $Q_2(t)$ is completely determined by $Q_1(t)$. Finally, observe that

$$\eta(t) = t - W_1(\eta(t)) + \epsilon_2(t), \tag{3.19}$$

where $\epsilon_2(t)$ is 0 if $W_1(t) = 0$ and otherwise is equal to the remaining service time of the customer currently occupying station 1.

The limit theorems proved here apply to systems that satisfy conditions of "heavy traffic," and in order to rigorously state these conditions, we require the construction of a "sequence of systems" to be indexed by n. Recall that the interarrival times and service times for the network are defined in terms of the basic sequences of unitized random variables $\{u(i) : i \geq 1\}$, $\{v_k(i) : i \geq 1\}$, $k = 0, 1, 2$. To construct a sequence of networks we further require sequences of positive constants $\{\lambda^{(n)}, n \geq 1\}$, $\{m_k^{(n)}, n \geq 1\}$, $k = 0, 1, 2$. In the n^{th} system of the sequence, the interarrival times and service times are taken to be $u^{(n)}(i) \equiv u(i)/\lambda^{(n)}$ and $v_k^{(n)}(i) \equiv m_k^{(n)} v_k(i)$, respectively. For the n^{th} system, $\lambda^{(n)}$ is the arrival rate of open customers and $m_k^{(n)}$ is the mean service time of the various customer designations. Define the relative traffic intensities $\rho_j^{(n)}$ as in (1.1)-(1.2) using $\lambda^{(n)}$ and $m_k^{(n)}$ in place of λ and m_k. Finally, the closed customer population of each network in the sequence is set to be n, that is, we define $N^{(n)} \equiv n$.

The convention here is to denote a parameter or a process associated with the n^{th} system by the superscript "(n)". For example, $A_0^{(n)}$ refers to the external arrival process of open customers in the n^{th} system. The results in this paper apply to processes that have been "scaled." Let $X^{(n)}$ denote a "generic" process associated with the n^{th} system. The "fluid scaled version" and "diffusion scaled version" of the process $X^{(n)}$, denoted as $\bar{X}^n$ and X^n, respectively, are defined via

$$\bar{X}^n(t) \equiv \frac{1}{n} X^{(n)}(nt) \qquad \text{and} \qquad X^n(t) \equiv \frac{1}{n} X^{(n)}(n^2 t).$$

We also define

$$\bar{\bar{X}}^n(t) \equiv \frac{1}{n^2} X^{(n)}(n^2 t) = \frac{1}{n} X^n(t).$$

It is assumed that the following conditions hold for the input processes of the network. First, the arrival rates and mean service times converge to finite constants, $\lambda^{(n)} \to \lambda$ and $m_k^{(n)} \to m_k$, $k = 0, 1, 2$. This implies that $\rho_1^{(n)} \to \rho_1 = \lambda m_0 + m_1/m_2$. Furthermore, it is assumed that there exists $-\infty < \theta < \infty$ such that

$$n(\rho_1^{(n)} - 1) \longrightarrow \theta \quad \text{as } n \to \infty. \tag{3.20}$$

Condition (3.20) is called the ***heavy traffic condition***. It requires that as the the number of closed customers becomes large, the relative traffic intensity at station 1 must become approximately 1. Moreover, the rate of convergence is "sufficiently fast."

The following two theorems are direct consequences of the functional strong law of large numbers, Donsker's Theorem and the functional central limit theorem for renewal processes:

Theorem 3.1 *As $n \to \infty$, $\bar{W}_2^n(0) \to m_2$ almost surely and*

$$(\bar{A}_0^n, \bar{V}_0^n, \bar{V}_1^n, \bar{V}_2^n, \bar{S}_0^n, \bar{S}_1^n, \bar{S}_2^n) \to (\bar{A}_0^*, \bar{V}_0^*, \bar{V}_1^*, \bar{V}_2^*, \bar{S}_0^*, \bar{S}_1^*, \bar{S}_2^*) \text{ u.o.c.}$$

where $\bar{A}_0^*(t) = \lambda t$, $\bar{V}_0^*(t) = m_0 t$, $\bar{V}_1^* = m_1 t$, $\bar{V}_2^* = m_2 t$, $\bar{S}_0^* = \frac{1}{m_0} t$, $\bar{S}_1^* = \frac{1}{m_1} t$, $\bar{S}_2^* = \frac{1}{m_2} t$.

Theorem 3.2 *$(\hat{A}_0^n, \hat{V}_0^n, \hat{V}_1^n, \hat{V}_2^n) \Longrightarrow (\hat{A}_0^*, \hat{V}_0^*, \hat{V}_1^*, \hat{V}_2^*)$ where $\hat{A}_0^*$ is $(0, \lambda c_a^2)$ Brownian motion and $\hat{V}_k^*$ is $(0, m_k^2 c_k^2)$ Brownian motion, $k = 0, 1, 2$.*

The following result, which establishes that "remainder" terms converge to zero under scaling, will be needed in our proofs. For $j = 1, 2$, let

$$\epsilon_{1j}^n(t) = \frac{1}{n} \epsilon_{1j}^{(n)}(n^2 t), \quad \bar{\epsilon}_{1j}^n(t) = \frac{1}{n} \epsilon_{1j}^{(n)}(nt),$$

and

$$\epsilon_2^n(t) = \frac{1}{n} \epsilon_2^{(n)}(n^2 t), \quad \bar{\epsilon}_2^n(t) = \frac{1}{n} \epsilon_2^{(n)}(nt).$$

Lemma 3.3 *For $j = 1, 2$ and each $t \geq 0$, $\|\epsilon_{1j}^n(\cdot)\|_t \to 0$, $\|\bar{\epsilon}_{1j}^n(\cdot)\|_t \to 0$, $\|\epsilon_2^n(\cdot)\|_t \to 0$, and $\|\bar{\epsilon}_2^n(\cdot)\|_t \to 0$ as $n \to \infty$.*

Proof. It follows from the definitions of $\epsilon_{1j}(t)$ and $\epsilon_2(t)$ that

$$\begin{aligned} 0 \leq \epsilon_{1j}(t) &\leq \max_{1 \leq i \leq A_0(t)} m_0 v_0(i) + \max_{1 \leq i \leq S_2(t)} m_1 v_1(i) \\ 0 \leq \epsilon_2(t) &\leq \max_{1 \leq i \leq S_2(t)} m_1 v_1(i). \end{aligned}$$

An application of Lemma 3.3 from Iglehart and Whitt [16] proves the lemma. ∎

4 The Limit Theorems

Theorem 4.1 (The Fluid Approximation) *If the heavy traffic condition (3.20) holds, then*

$$(\bar{W}^n, \bar{I}^n, \bar{Q}^n, \bar{T}^n) \to (\bar{W}^*, \bar{I}^*, \bar{Q}^*, \bar{T}^*) \quad u.o.c.$$

where

$$\bar{W}_1^*(t) = 0 \tag{4.1}$$

$$\bar{W}_2^*(t) = m_2 \tag{4.2}$$

$$\bar{Q}_0^*(t) = 0, \quad \bar{Q}_1^*(t) = 0, \quad \bar{Q}_2^*(t) = 1 \tag{4.3}$$

$$\bar{I}_1^*(t) = \bar{I}_2^*(t) = 0, \tag{4.4}$$

$$\bar{T}_0^*(t) = \lambda m_0 t, \quad \bar{T}_1^*(t) = \frac{m_1}{m_2} t. \tag{4.5}$$

Theorem 4.2 (The Diffusion Approximation) *If the heavy traffic condition (3.20) holds, then*

$$(W^n, I^n, U^n, Q^n) \Longrightarrow (W^*, I^*, U^*, Q^*)$$

where

$$W_1^*(t) = \xi_1^*(t) + I^*(t) - \frac{m_1}{m_2} I_2^*(t) \tag{4.6}$$

$$\xi_1^* \text{ is a } (\theta, \sigma^2) \text{ Brownian motion} \tag{4.7}$$

$$W_2^*(t) = m_2 - W_1^*(t) \tag{4.8}$$

$$U_0^*(t) = \lambda m_0 W_1^*(t), \quad U_1^*(t) = \frac{m_1}{m_2} W_1^*(t), \quad U_2^*(t) = W_2^*(t) \tag{4.9}$$

$$Q_k^*(t) = \frac{1}{m_k} U_k^*(t), \quad k = 0, 1, 2 \tag{4.10}$$

$$I^* \text{ is continuous and nondecreasing with } I^*(0) = 0 \tag{4.11}$$

$$I_i^* \text{ increases only at times } t \text{ with } W_i^*(t) = 0. \tag{4.12}$$

Equations (4.6)-(4.8),(4.11), and (4.12) characterize W_1^* as a one-dimensional reflected Brownian motion on the interval $[0, m_2]$ with drift θ and variance σ^2, where θ and σ^2 are given by (3.20) and (1.7), respectively. Properties (4.9) and (4.10) express the deterministic relationships between queue lengths, partial workloads, and overall workloads that are characteristic of Brownian limits of queueing networks (for example, see [12, 13]).

In Section 1, we noted that the boundedness of the (limiting) workload process at station 1 can be viewed as a consequence of the "heavy traffic mixing principle." This principle, which is born out in equation (4.9), states that in the heavy traffic limit, the class specific workloads at each station are deterministically proportional to the overall workload at that station. We can develop another rationale for explaining the boundedness of the workload process at station 1, which perhaps may

be more intuitive, by considering the arrival process to this station. Open customers arrive to station 1 at rate λ. While station 2 is not empty, the arrivals of closed customers to station 1 resembles a renewal process with rate $1/m_2$. When all closed customers are at station 1, however, the arrival process for closed customers are temporarily "turned off," and for that period of time, the rate at which work arrives falls below the critical heavy traffic level. That is, during the period of time in which station 2 is empty, station 1 displays "non-heavy traffic behavior."

Let π be the steady state distribution of W_1^*, and set

$$\beta_1 = \lim_{t\to\infty} \frac{1}{t} I_1^*(t)$$
$$\beta_2 = \lim_{t\to\infty} \frac{1}{t} \left(\frac{m_1}{m_2} I_2^*(t)\right).$$

We have the following result from Harrison [10].

Theorem 4.3 (Proposition 5.5.5, [10]) *Set $b \equiv m_2$. If $\theta = 0$, then $\beta_1 = \beta_2 = \sigma^2/2b$ and π is the uniform distribution on $[0, b]$. Otherwise, setting $\kappa \equiv 2\theta/\sigma^2$,*

$$\beta_1 = \frac{\theta}{e^{\kappa b} - 1}, \quad \beta_2 = \frac{\theta}{1 - e^{-\kappa b}}, \tag{4.13}$$

and $\pi(dz) = p(z)dz$ where

$$p(z) = \frac{\kappa e^{\kappa z}}{e^{\kappa b} - 1} \tag{4.14}$$

The following deterministic time change theorem, due to Whitt [22], will be helpful in proving our results.

Theorem 4.4 (Deterministic Time Change Theorem) *Let $\{f_n, n \geq 1\}$ and $\{c_n, n \geq 1\}$ be sequences in $\mathbf{D}$ where c_n is nondecreasing with $c_n(0) = 0$. If (f_n, c_n) converges u.o.c. to a continuous pair (f, c), then $f_n(c_n(t))$ converges u.o.c. to $f(c(t))$.*

5 Proof of the Fluid Approximation

Lemma 5.1 *For each $t \geq 0$, $\|\bar{W}_1^n(\cdot)\|_t \to 0$ as $n \to \infty$.*

Proof. The lemma is proved via a bounding argument in which $W_1^{(n)}$ is bounded above by a sequence of open queues with two customer types 0 and 1. Recall the definitions of the processes $A_0^{(n)}$ and $S_2^{(n)}$ from (3.1) and (3.2). We denote by $A_0^{(n)}$ the arrival process of type 0 customers and we let type 1 customers arrive according to the renewal process $S_2^{(n)}$. The sequence of service

times for class k customers is given by $\{m_k^{(n)} v_k(i), i \geq 0\}$, and as in (3.4), we denote by $V_k^{(n)}$ the associated partial sums process. Defining

$$\chi^{(n)}(t) \equiv V_0^{(n)}(A_0^{(n)}(t)) + V_1^{(n)}(S_2^{(n)}(t)) - t \tag{5.1}$$

$$Z^{(n)}(t) \equiv \chi^{(n)}(t) + Y^{(n)}(t) \tag{5.2}$$

$$Y^{(n)}(t) \equiv \sup_{0 \leq s \leq t} \left\{\chi^{(n)}(s)\right\}^-, \tag{5.3}$$

this queue provides an upper bound for station 1 in the sense that

$$0 \leq W_1^{(n)}(t) \leq Z^{(n)}(t) \quad \text{for all } t \geq 0. \tag{5.4}$$

Applying the fluid scaling to $\chi^{(n)}$, equations (5.1)-(5.3) become

$$\bar{\chi}^n(t) \equiv \bar{V}_0^n(\bar{A}_0^n(t)) + \bar{V}_1^n(\bar{S}_2^n(t)) - t \tag{5.5}$$

$$\bar{Z}^n(t) \equiv \bar{\chi}^n(t) + \bar{Y}^n(t) \tag{5.6}$$

$$\bar{Y}^n(t) \equiv \sup_{0 \leq s \leq t} \{\bar{\chi}^n\}^-. \tag{5.7}$$

It follows from Theorem 3.1 and the Deterministic Time Change Theorem 4.4 that $\|\bar{\chi}^n(\cdot)\|_t \to 0$ almost surely for each $t \geq 0$ as $n \to \infty$. Because the mappings (5.6)-(5.7) are continuous, $\|\bar{Z}^n(\cdot)\|_t \to 0$, and it follows from (5.4) that for each $t \geq 0$, $\|\bar{W}_1^n(\cdot)\|_t \to 0$ almost surely as $n \to \infty$. ∎

Proof of Theorem 4.1 Because $|\bar{B}_i^n(t) - \bar{B}_i^n(s)| < |t-s|$ and $|\bar{T}_i^n(t) - \bar{T}_i^n(s)| < |t-s|$ almost surely, we can conclude from the Arzela-Ascoli theorem that there exists a subsequence n_k, $k = 1, 2, \ldots$ on which $\bar{T}_0^{n_k}$, $\bar{T}_1^{n_k}$, $\bar{B}_1^{n_k}$, and $\bar{B}_2^{n_k}$ converges almost surely to continuous limits and the convergence is uniform on compact sets. Denote the corresponding limits by $\bar{T}_0^*$, $\bar{T}_1^*$, $\bar{B}_1^*$, and $\bar{B}_2^*$, respectively. Observe that $\bar{I}_i^n(t) = t - \bar{B}_i^n(t)$, hence $\bar{I}_i^{n_k} \to \bar{I}_i^*$ u.o.c. where $\bar{I}_i^*(t) = t - \bar{B}_i^*(t)$.

Applying the fluid scaling to equations (3.8) and (3.16)-(3.19), we have

$$\bar{W}_2^n(t) = \bar{W}_2^n(0) + \bar{V}_2^n(\bar{S}_1^n(\bar{B}_1^n(t))) - \bar{B}_2^n(t) \tag{5.8}$$

$$\bar{Q}_0^n(t) = \bar{A}^n(t) - \bar{A}^n(\bar{\eta}^n(t)) \tag{5.9}$$

$$\bar{Q}_1^n(t) = \bar{S}_2^n(\bar{B}_2^n(t)) - \bar{S}_2^n(\bar{B}_2^n(\bar{\eta}^n(t))) \tag{5.10}$$

$$\bar{Q}_2^n(t) = 1 - \bar{Q}_1^n(t) \tag{5.11}$$

$$\bar{\eta}^n(t) = t - \bar{W}_1^n(\bar{\eta}^n(t)) + \bar{\epsilon}_2^n(t). \tag{5.12}$$

From equation (5.12), we have

$$\begin{aligned} \limsup_{n\to\infty} \|\cdot - \bar{\eta}^n(\cdot)\|_t &\leq \limsup_{n\to\infty} \|\bar{W}_1^n(\bar{\eta}^n(\cdot))\|_t + \limsup_{n\to\infty} \|\bar{\epsilon}_2^n(\cdot)\|_t \\ &\leq \limsup_{n\to\infty} \|\bar{W}_1^n(\cdot)\|_t + \limsup_{n\to\infty} \|\bar{\epsilon}_2^n(\cdot)\|_t \\ &\leq 0, \end{aligned} \tag{5.13}$$

where the first inequality follows from the observation that $\bar{\eta}(t) \leq t$ and the last inequality is a result of Lemma 3.3 and Lemma 5.1. Scaling (3.13) in the fluid convention, we obtain

$$\begin{aligned}\bar{T}_0^n(t) &= \bar{V}_0^n(\bar{A}_0^n(\bar{\eta}^n(t))) + \bar{\epsilon}_{10}^n(t) \\ \bar{T}_1^n(t) &= \bar{V}_1^n(\bar{S}_2^n(\bar{B}_2^n(\bar{\eta}^n(t)))) + \bar{\epsilon}_{11}^n(t).\end{aligned}$$

Applying Theorem (3.1), the Deterministic Time Change Theorem 4.4, Lemma 3.3, and (5.13), it follows that $\bar{T}^n \to \bar{T}^*$ u.o.c. where

$$\bar{T}_0^*(t) = \lambda m_0 t \quad \text{and} \quad \bar{T}_1^*(t) = \frac{m_1}{m_2}\bar{B}_2^*(t).$$

Again applying Theorem (3.1) and Theorem 4.4 to (5.8)-(5.11), we get $(\bar{W}_2^{n_k}, \bar{Q}^{n_k}) \to (\bar{W}_2^*, \bar{Q}^*)$ u.o.c. where

$$\bar{W}_2^*(t) = m_2 + \frac{m_2}{m_1}\bar{T}_1^*(t) - \bar{B}_2^*(t) = m_2 \tag{5.14}$$

and

$$\bar{Q}_0^*(t) = 0, \qquad \bar{Q}_1^*(t) = 0, \qquad \bar{Q}_2^*(t) = 1. \tag{5.15}$$

Finally, (5.14) implies that there exists a finite n^* and positive ϵ such that $\bar{W}_2^{n_k}(t) \geq \epsilon$ a.s. for all $t \geq 0$ and $n_k \geq n*$. Because $\bar{I}_2^n$ may increase only at times t for which $\bar{W}_2^n(t) = 0$, it follows that $\bar{I}_2^{n_k}(t) = 0$ for all $t \geq 0$ and $n_k \geq n*$. Hence, $I_2^*(t) = 0$, $B_2^*(t) = t$, from which it follows that $\bar{T}_1^*(t) = \frac{m_1}{m_2}t$, $B_1^*(t) = t$, and $\bar{I}_1^*(t) = \bar{I}_2^*(t) = 0$ for all $t \geq 0$. Because each subsequence of $(\bar{W}^n, \bar{I}^n, \bar{Q}^n, \bar{T}^n)$ contains a convergent subsequence and each of these subsequences in turn converges to the limit described in Theorem 4.1, we may conclude that $(\bar{W}^n, \bar{I}^n, \bar{Q}^n, \bar{T}^n)$ itself converges to the same limit . ∎

6 Proof of the Diffusion Approximation

We begin with the definition of "centered" processes

$$\begin{aligned}\hat{A}_0^{(n)}(t) &\equiv A_0^{(n)}(t) - \lambda^{(n)}t & \qquad \hat{V}_k^{(n)}(t) &\equiv V_k^{(n)}(t) - m_k^{(n)}t \\ \hat{S}_k^{(n)}(t) &\equiv S_k^{(n)}(t) - \frac{1}{m_k^{(n)}}t & \qquad \hat{T}_1^{(n)}(t) &\equiv T_1^{(n)}(t) - \frac{m_1^{(n)}}{m_2^{(n)}}t \\ \hat{\eta}^{(n)}(t) &\equiv \eta^{(n)}(t) - t.\end{aligned}$$

Centering (3.13) and (3.19) in this way, we obtain

$$\begin{aligned}\hat{T}_1(t) &= \hat{V}_1(S_2(B_2(\eta(t)))) + m_1\hat{S}_2(B_2(\eta(t))) - \frac{m_1}{m_2}I_2(\eta(t)) + \frac{m_1}{m_2}\hat{\eta}(t) + \epsilon_{11}(t) \\ &= \hat{V}_1(S_2(B_2(\eta(t)))) + m_1\hat{S}_2(B_2(\eta(t))) - \frac{m_1}{m_2}I_2(\eta(t)) \\ &\quad -\frac{m_1}{m_2}W_1(\eta(t)) + \frac{m_1}{m_2}\epsilon_2(t) + \epsilon_{11}(t)\end{aligned}$$

$$= \hat{V}_1(S_2(B_2(\eta(t)))) + m_1\hat{S}_2(B_2(\eta(t))) - \frac{m_1}{m_2}X_1(\eta(t)) - \frac{m_1}{m_2}I_1(\eta(t))$$
$$-\frac{m_1}{m_2}I_2(\eta(t)) + \frac{m_1}{m_2}\epsilon_2(t) + \epsilon_{11}(t). \tag{6.1}$$

Next, set

$$\xi_1(t) \equiv \hat{V}_0(A_0(t)) + m_0\hat{A}_0(t) + \hat{V}_1(S_2(B_2(t))) + m_1\hat{S}_2(B_2(t)) + \left(\lambda m_0 + \frac{m_1}{m_2} - 1\right)t \tag{6.2}$$

$$\xi_2(t) \equiv W_2(0) + \hat{V}_2(S_1(T_1(t))) + m_2\hat{S}_1(T_1(t)) + \frac{m_2}{m_1}\hat{V}_1(S_2(B_2(\eta(t))))$$
$$+m_2\hat{S}_2(B_2(\eta(t))) - \xi_1(\eta(t)) + \epsilon_2(t) + \frac{m_2}{m_1}\epsilon_{11}(t). \tag{6.3}$$

One can invoke (6.1)-(6.3) together with the observation that $I_1(t) = I_1(\eta(t))$ to express the netput processes (3.9) as

$$X_1(t) = \xi_1(t) - \frac{m_1}{m_2}I_2(t) \tag{6.4}$$

$$X_2(t) = \xi_2(t) - I_1(t) - \left(1 - \frac{m_1}{m_2}\right)I_2(\eta(s)). \tag{6.5}$$

Applying the diffusion scale to (3.10)-(3.12), we have the following expressions for the scaled workload process:

$$W_1^n(t) = \xi_1^n(t) - \frac{m_1^{(n)}}{m_2^{(n)}}I_2^n(t) + I_1^n(t) \tag{6.6}$$

$$W_2^n(t) = \xi_2^n(t) - I_1(t) + \left(1 - \frac{m_1^{(n)}}{m_2^{(n)}}\right)\left(I_2^n(t) - I_2^n(\bar{\bar{\eta}}^n(t))\right) + \frac{m_1^{(n)}}{m_2^{(n)}}I_2^n(t) \tag{6.7}$$

I_i^n is continuous and nondecreasing with $I_i^n(0) = 0$ (6.8)

I_i^n increases only at times t when $W_i^n(t) = 0$. (6.9)

where

$$\xi_1^n(t) = \hat{V}_0^n(\bar{\bar{A}}_0^n(t)) + m_0^{(n)}\hat{A}_0^n(t) + \hat{V}_1^n(\bar{\bar{S}}_2^n(\bar{\bar{B}}_2^n(t))) + m_1^{(n)}\hat{S}_2^n(\bar{\bar{B}}_2^n(t)) +$$
$$\left(\frac{m_1^{(n)}}{m_2^{(n)}} + \lambda^{(n)}m_0^{(n)} - 1\right)nt \tag{6.10}$$

$$\xi_2^n(t) = W_2^n(0) + \hat{V}_2^n(\bar{\bar{S}}_1^n(\bar{\bar{T}}_1^n(t))) + m_2^{(n)}\hat{S}_1^n(\bar{\bar{T}}_1^n(t)) + \frac{m_2^{(n)}}{m_1^{(n)}}\hat{V}_1^n(\bar{\bar{S}}_2^n(\bar{\bar{B}}_2^n(\bar{\bar{\eta}}^n(t))))$$
$$+m_2^{(n)}\hat{S}_2^n(\bar{\bar{B}}_2^n(\bar{\bar{\eta}}^n(t))) - \xi_1^n(\bar{\bar{\eta}}^n(t)) + \epsilon_2^n(t) + \frac{m_2^{(n)}}{m_1^{(n)}}\epsilon_{11}^n(t), \tag{6.11}$$

$$\bar{\bar{\eta}}^n(t) = t + \frac{1}{n}W_1^n(\bar{\bar{\eta}}^n(t)) + \frac{1}{n}\epsilon_2^n(t). \tag{6.12}$$

By the Skorohod representation theorem, we may and will assume henceforth that the convergence in Theorem 3.2 holds u.o.c.

Lemma 6.1 *For each* $t \geq 0, \|\bar{\bar{\eta}}^n(\cdot) - \cdot\|_t \to 0$ *as* $n \to \infty$.

Proof. Let $Z^{(n)}$ be the process defined in Lemma 5.1, and note that $Z^n \to Z^*$ u.o.c. where Z^* is a one-dimensional reflected Brownian motion with drift μ and variance σ^2. Because $W_1^{(n)}(t) \leq Z^{(n)}(t)$,

$$\begin{aligned}
\limsup_{n\to\infty} \|\cdot - \bar{\bar{\eta}}^n(\cdot)\|_t &\leq \limsup_{n\to\infty} \frac{1}{n}\|W_1^n(\bar{\bar{\eta}}^n(\cdot))\|_t + \limsup_{n\to\infty} \frac{1}{n}\|\epsilon_1^n(\cdot)\|_t \\
&\leq \limsup_{n\to\infty} \frac{1}{n}\|W_1^n(\cdot)\|_t \\
&\leq \limsup_{n\to\infty} \frac{1}{n}\|Z_1^n(\cdot)\|_t \\
&= 0.
\end{aligned}$$

∎

Lemma 6.2 $\xi^n \to \xi^*$ *u.o.c. as* $n \to \infty$ *where* ξ_1^* *is* (θ, σ^2) *Brownian motion and* $\xi_2^*(t) = m_2 - \xi_1^*(t)$.

Proof. The convergence of ξ_1^n follows from from (3.20), Theorem 3.2, Theorem 4.4, and Lemma 6.1. Define

$$\zeta_1^n(t) = \hat{V}_2^n(\bar{\bar{S}}_1^n(\bar{\bar{T}}_1^n(t))) + m_2^{(n)}\hat{S}_2^n(\bar{\bar{B}}_2^n(\bar{\bar{\eta}}^n(t))) \tag{6.13}$$

$$\zeta_2^n(t) = \frac{m_2^{(n)}}{m_1^{(n)}}\hat{V}_1^n(\bar{\bar{S}}_2^n(\bar{\bar{B}}_2^n(\bar{\bar{\eta}}^n(t)))) + m_2^{(n)}\hat{S}_1^n(\bar{\bar{T}}_1^n(t)). \tag{6.14}$$

An application of Theorem 1 in Iglehart and Whitt [15], shows that $\hat{S}_i^n \to -m_i{}^{-3/2}V_i^*$ u.o.c. From Theorem 3.2, Theorem 4.1, and the Deterministic Time Change Theorem, we can conclude from (6.13)-(6.14) that $\|\zeta_i^n(\cdot)\|_t \to 0$ as $n \to \infty$ for each $t \geq 0$ and $i = 1, 2$. Writing

$$\xi_2^n(t) = W_2^n(0) + \zeta_1^n(t) + \zeta_2^n(t) - \xi_1^n(\bar{\bar{\eta}}^n(t)) + \epsilon_2^n(t) + \frac{m_2^{(n)}}{m_1^{(n)}}\epsilon_{11}^n(t),$$

the theorem follows as a result of Theorem 3.1, Lemma 3.3, and Lemma 6.1. ∎

Lemma 6.2 implies that $e'\xi^n(t) \geq m_2/2$ for all n sufficiently large. For the purpose of our proof, we may therefore assume that $\xi^n \in \mathbf{D}^2_{m_2/2}$ for all $n \geq 1$. From (6.5), we may also conclude that $\xi_2^n(t) - I_1^n(t) - (1 - m_1^{(n)}/m_2^{(n)})I_2^n(\bar{\bar{\eta}}^n(s))$ has no jumps downward, from which it follows that the assumption of Lemma 9.7 is satisfied because I_1^n is a nondecreasing process. Finally, noting that $\bar{\eta}^n \in \Lambda$, we conclude from Theorem 9.1 that there exists a unique pair of processes (W^n, I^n) satisfying (6.6)-(6.9), and that (W^n, I^n) is given by the mapping $(W^n, I^n) = (\Phi, \Psi)(\xi^n, \bar{\eta}^n)$.

Proof of Theorem 4.2 From Lemmas 6.1-6.2, $(\xi_1^n, \xi_2^n, \overline{\overline{\eta}}^n) \to (\xi_1^*, \xi_2^*, \overline{\overline{\eta}}^*)$ u.o.c. where ξ_1^* is (θ, σ^2) Brownian motion, $\xi_2 j(t) = m_2 - \xi_1^*(t)$, and $\overline{\overline{\eta}}^*(t) = t$. Because Brownian motion is almost surely continuous, it follows from Theorem 9.1 that (Φ, Ψ) is continuous at $(\xi^*, \overline{\overline{\eta}}^*)$, hence $(I^n, W^n) \to (I^*, W^*)$ u.o.c. where $(I^*, W^*) = (\Phi, \Psi)(\xi^*, \overline{\overline{\eta}}^*)$. Specifically, we have from Lemma 9.2,

$$
\begin{aligned}
&W_1^*(t) = \xi_1^*(t) - \frac{m_1}{m_2} I_2^*(t) + I_1^*(t) \\
&W_2^*(t) = m_2 - W_1^*(t) \\
&I_i^* \text{ is nondecreasing and continuous with } I_i^*(0) = 0 \\
&I_i^* \text{ increases only at times } t \text{ with } W_i^*(t) = 0,
\end{aligned}
$$

implying W_1^* is a one-dimensional reflected Brownian motion on the interval $[0, m_2]$ with drift θ and variance σ^2.

Next, observe from (3.19) and (5.13) that $\hat{\eta}^n \to \hat{\eta}^*$ u.o.c. where $\hat{\eta}^*(t) = -W_1^*(t)$. Centering (3.14)-(3.15), we obtain

$$
\begin{aligned}
U_0^n(t) &= \hat{M}_0^n(t) - \hat{M}_0^n(\overline{\overline{\eta}}^n(t)) - \lambda^{(n)} m_0^{(n)} \hat{\eta}^n(t) \\
U_1^n(t) &= \hat{M}_1^n(t) - \hat{M}_1^n(\overline{\overline{\eta}}^n(t)) - \frac{m_1^{(n)}}{m_2^{(n)}} \hat{\eta}^n(t),
\end{aligned}
$$

with

$$
\begin{aligned}
\hat{M}_0^n(t) &= \hat{V}_0^n(\overline{\overline{A}}_0^n(t)) + m_0^{(n)} \hat{A}_0^n(t) \\
\hat{M}_1^n(t) &= \hat{V}_1^n(\overline{\overline{S}}_2^n(\overline{\overline{B}}_2^n(t))) + m_1^{(n)}(\hat{S}_2^n(\overline{\overline{B}}_2^n(t))) - \frac{m_1^{(n)}}{m_2^{(n)}} I_2^n(t).
\end{aligned}
$$

Because Brownian motion is continuous, it follows from Lemma 6.1, Theorem 3.1, Theorem 3.2, and Theorem 4.4 that $\|\hat{M}_i^n(\cdot) - \hat{M}_i^n(\overline{\overline{\eta}}^n(\cdot))\|_t \to 0$ a.s. for each $t \geq 0$ and $i = 1, 2$. Thus, $U_0^n \to U_0^*$ and $U_1^n \to U_1^*$ u.o.c. where

$$
U_0^*(t) = \lambda m_0 W_1^*(t) \quad \text{and} \quad U_1^*(t) = \frac{m_1}{m_2} W_1^*(t).
$$

Similarly, noting that

$$
\begin{aligned}
Q_k^0(t) &= \hat{A}_0^n(t) - \hat{A}_0^n(\overline{\overline{\eta}}^n(t)) - \lambda^{(n)} \hat{\eta}^n(t) \\
Q_k^1(t) &= \hat{A}_1^n(t) - \hat{A}_1^n(\overline{\overline{\eta}}^n(t)) - \frac{1}{m_2^{(n)}} \hat{\eta}^n(t) \\
Q_k^2(t) &= \hat{A}_2^n(t) - \hat{A}_2^n(\overline{\overline{\eta}}^n(t)) - \frac{1}{m_2^{(n)}} \hat{\eta}^n(t)
\end{aligned}
$$

where

$$
\hat{A}_1^n(t) = \hat{S}_2^n(\overline{\overline{B}}_2^n(t)) - \frac{1}{m_2^{(n)}} I_2^n(t) \quad \text{and} \quad \hat{A}_2^n(t) = \hat{S}_1^n(\overline{\overline{T}}_1^n(t)) - \frac{1}{m_1^{(n)}} \hat{T}_1^n(t),
$$

it follows that $Q^n \to Q^*$ u.o.c. with

$$\begin{aligned} Q_0^*(t) &= \lambda W_1^*(t) = \frac{1}{m_0} U_1^*(t) \\ Q_1^*(t) &= \frac{1}{m_2} W_1^*(t) = \frac{1}{m_1} U_1^*(t) \\ Q_2^*(t) &= \frac{1}{m_2} W_2^*(t), \end{aligned}$$

and the proof of Theorem 4.2 is complete. ∎

7 Refining the Brownian Approximation

We now turn to the following important question: Given a mixed network with parameters λ, c_a^2, m_i, and c_i^2, $i = 0, 1, 2$, and a finite number N of closed customers, how do we obtain performance estimates for the network using the theory developed in the previous sections? Theorem 4.2 suggests the following approximation for the workload process at station 1:

$$\frac{1}{N}(W_1(N^2\cdot) \approx W_1^*(\cdot) \tag{7.1}$$

where W_1^* is an RBM on the interval $[0, m_2]$ whose drift θ^* and variance σ^2 are given by (1.6) and (1.7), namely,

$$\begin{aligned} \theta^* &= N\left(\lambda m_0 + \frac{m_1}{m_2} - 1\right) \\ \sigma^2 &= \lambda m_0^2\left(c_a^2 + c_0^2\right) + \frac{m_1^2}{m_2}\left(c_1^2 + c_2^2\right). \end{aligned}$$

By reversing the scaling in equation (7.1), one obtains the approximation

$$W(\cdot) \approx \tilde{W}(\cdot) \tag{7.2}$$

where $\tilde{W}$ is an RBM on the interval $[0, m_2 N]$, whose drift μ is given by

$$\mu = \lambda m_0 + \frac{m_1}{m_2} - 1, \tag{7.3}$$

and whose variance is again σ^2. We then apply Theorem 4.3 to obtain the steady-state distribution of $\tilde{W}$ and other performance measures of interest. In particular, writing $\tilde{W}(\infty)$ to mean the random variable associated with the stationary distribution of the process $\{\tilde{W}(t), t \geq 0\}$ and setting $b \equiv m_2 N$, $\kappa \equiv 2\mu/\sigma^2$, we have

$$\mathbf{E}\tilde{W}(\infty) = \begin{cases} \frac{b}{2} & \mu = 0 \\ \frac{b}{1-e^{-\kappa b}} - \frac{1}{\kappa} & \text{otherwise.} \end{cases} \tag{7.4}$$

Morover, the throughput rate obtained from the Brownian approximation is given by $\tilde{\alpha}/m_2$ where

$$\tilde{\alpha} = \begin{cases} 1 - \frac{m_2}{m_1}\left(\frac{\sigma^2}{2b}\right) & \mu = 0 \\ 1 - \frac{m_2}{m_1}\left(\frac{\mu}{1-e^{-\kappa b}}\right) & \text{otherwise.} \end{cases} \tag{7.5}$$

Equations (7.2), and (7.4)-(7.5) in particular, form one possible approximation for a two-station mixed network. As was noted in previous works on heavy traffic approximations (for example, [6, 12, 13]), however, one typically needs to "refine" the Brownian limit in order to obtain good performance estimates. The approach to developing our refinement is to arrive at an approximation method that, as much as possible, yields estimates that agree with the exact solutions in those special cases for which the exact solutions are known. Henceforth, our results will be benchmarked against the following special case.

From the theory of quasi-reversible queues [17] (see also [1]), the mixed network in Figure 1 has product form solutions if

$$m_0 = m_1 \equiv m, \quad \text{and} \tag{7.6}$$

$$c_a^2 = c_0^2 = c_1^2 = c_2^2 = 1. \tag{7.7}$$

In this case, observe that

$$\sigma^2 = 2m(1+\mu).$$

Denoting by $P(k,l)$ the steady-state probability of k open customers at station 1, l closed customers at station 1, and $N-l$ closed customers at station 2, we have

$$P(k,l) = G\begin{pmatrix} k+l \\ l \end{pmatrix} \eta_0^k \, \eta_1^l \tag{7.8}$$

where

$$\eta_0 = \lambda m, \qquad \eta_1 = \frac{m}{m_2},$$

and G is the normalizing constant

$$G = (1-\eta_0)\sum_{k=0}^{N}\left(\frac{\eta_1}{1-\eta_0}\right)^k.$$

and Let us write $Q_0(\infty)$ to mean the steady-state random variable associated with the process $\{Q_0(t), t \geq 0\}$, and similarly, let us use $Q_1(\infty)$ to mean the steady-state headcount of closed customers at station 1. From (7.8), we obtain the following statistics:

$$\mathbf{E}Q_0(\infty) = G\frac{\eta_0}{(1-\eta_0)^2}\sum_{k=0}^{N}(k+1)\left(\frac{\eta_1}{1-\eta_0}\right)^k \tag{7.9}$$

$$\mathbf{E}Q_1(\infty) = G\frac{1}{1-\eta_0}\sum_{k=0}^{N}k\left(\frac{\eta_1}{1-\eta_0}\right)^k \tag{7.10}$$

and

$$\begin{aligned}\alpha^* &= 1 - \sum_{n=1}^{\infty} P(n,N) \\ &= 1 - G\left(\frac{1}{1-\eta_0}\right)\left(\frac{\eta_1}{1-\eta_0}\right)^N. \end{aligned} \tag{7.11}$$

(Recall that α^*/m_2 is the throughput rate of closed customers.) In particular, if $\rho \equiv \eta_0 + \eta_1 = 1$, equations (7.9)-(7.11) simplify to

$$\mathbf{E}Q_0 = \left(\frac{\eta_0}{1-\eta_0}\right)\left(\frac{N+2}{2}\right) \tag{7.12}$$

$$\mathbf{E}Q_1 = \frac{N}{2} \tag{7.13}$$

$$\alpha^* = \frac{N}{N+1}. \tag{7.14}$$

Finally, when $\rho < 1$, we have the following limit as the number of closed customers in the system increases:

$$\lim_{N\uparrow\infty} \mathbf{E}Q_0 = \frac{\eta_0}{1-\eta_0-\eta_1} \tag{7.15}$$

$$\lim_{N\uparrow\infty} \mathbf{E}Q_1 = \frac{\eta_1}{1-\eta_0-\eta_1}. \tag{7.16}$$

Refinement 1: Replace b by $m_2 N/\tilde{\alpha}$.

We justify this modification via the following argument. The idleness process at station 2 increases whenever that station is empty, or equivalently, at times t when $Q_1(t) = N$. By the "functional" Little's Law which follows from equations (4.9)-(4.10), we have $Q_1(t) = (1/m_2)W(t)$, interpreting $1/m_2$ as the throughput rate of closed customers. This refinement essentially replaces the naive throughput rate by the approximated throughput rate $\tilde{\alpha}/m_2$. In the case $\mu = 0$, it can be verified from (7.5) and (7.14) that $\sigma^2 = 2m$ and the refinement gives

$$\begin{aligned}\tilde{\alpha} &= 1 - \frac{m_2}{m}\left(\frac{\sigma^2}{2b}\right) \\ &= 1 - \frac{\tilde{\alpha}}{N}, \end{aligned} \tag{7.17}$$

from which we obtain

$$\tilde{\alpha} = \frac{N}{N+1}. \tag{7.18}$$

Note that in this case, the Brownian approximation agrees with the exact solution (7.14).

Refinement 2: Set $\tilde{Q}_1(t) = (\tilde{\alpha}/m_2)\tilde{W}(t)/\rho$.

This modification of equations (4.9) and (4.10) consists of two parts. First, the throughput rate used in the so called "functional" Little's Law is taken to be $\tilde{\alpha}/m_2$ rather than $1/m_2$. Second, a weighting factor of $1/\rho$ is then applied to the relationship between queue length and workload processes. As noted in Section 6 of [13], empirical experience suggests that a better approximation may be obtained with such a refinement. Writing $\tilde{Q}_i(\infty)$ to mean the random variable associated with the stationary distribution of the process $\{\tilde{Q}_i(t), t \geq 0\}$, we have in the special case of $\mu = 0$,

$$\begin{aligned} \mathbf{E}\tilde{Q}_1(\infty) &= \frac{\tilde{\alpha}}{m_2\rho}\mathbf{E}\tilde{W}(\infty) \\ &= \frac{\alpha}{m_2\rho}\left(\frac{m_2 N}{2\alpha}\right) \\ &= \frac{N}{2}, \end{aligned}$$

which agrees with (7.13). Moreover, it follows from (7.4) that when $\mu < 0$, the approximation is asymptotically exact as $N \to \infty$:

$$\begin{aligned} \lim_{N\to\infty} \mathbf{E}\tilde{Q}_1(\infty) &= \frac{1}{m_2\rho}\left(\frac{-\sigma^2}{2\mu}\right) \\ &= \frac{\eta_1}{1-\eta_0-\eta_1} = \lim_{N\to\infty} \mathbf{E}Q_1(\infty). \end{aligned}$$

Refinement 3: Set $\tilde{Q}_0(t) = \lambda\tilde{W}(t)/\rho$.

This refinement is essentially identical to the previous one, in that a factor of $1/\rho$ is introduced. It can also be shown that with this modification, the approximation is asymptotically exact as $N \to \infty$ whenever $\mu < 0$,

$$\begin{aligned} \lim_{N\to\infty} \mathbf{E}\tilde{Q}_0(\infty) &= \frac{\lambda}{\rho}\left(\frac{-\sigma^2}{2\mu}\right) \\ &= \frac{\eta_0}{1-\eta_0-\eta_1} = \lim_{N\to\infty} \mathbf{E}Q_0(\infty). \end{aligned}$$

However, note that we do *not* obtain the exact solution for finite N, even when $\mu = 0$, since

$$\begin{aligned} \mathbf{E}\tilde{Q}_0(\infty) &= \frac{\lambda}{\rho}\mathbf{E}\tilde{W}(\infty) \\ &= \frac{\lambda}{\rho}\left(\frac{m_2 N}{2\alpha}\right) \\ &= \frac{\eta_0}{1-\eta_0}\left(\frac{N+1}{2}\right), \end{aligned}$$

where the last equality follows because $\eta_0 + \eta_1 = 1$.

To summarize, our approximation procedure consists of replacing the workload process by $\tilde{W}$, an RBM on the interval $[0, m_2N/\tilde{\alpha}]$ with drift μ and variance σ^2 given by (7.3) and (1.7), respectively. The queue length processes are then approximated via the mappings

$$\tilde{Q}_0(t) = \frac{\lambda}{\rho}\tilde{W}(t) \tag{7.19}$$

$$\tilde{Q}_1(t) = \frac{\tilde{\alpha}}{m_2\rho}\tilde{W}(t). \tag{7.20}$$

Note that the formulation of the approximation includes the quantity $\tilde{\alpha}$, which itself must be approximated. In order for the approximation to be consistent, we must now show that there exists a unique $0 < \tilde{\alpha} < 1$ that satisfies equation (7.5) for *all* ranges of the parameter set λ, c_a^2, m_i, and c_i^2, $i = 0, 1, 2$.

Theorem 7.1 *Let $b \equiv m_2N/\tilde{\alpha}$. Then there exists a unique $0 < \tilde{\alpha} < 1$ that satisifes (7.5).*

Proof. For $\mu = 0$, it is easily verified from (7.5) that the unique solution is given by

$$\tilde{\alpha} = \frac{N}{N + \sigma^2/2m_1}.$$

Consider $\mu < 0$. We need to show that there exists a unique solution $x \in (0, 1)$ to the transcendental equation

$$f(x) \equiv \left(1 - e^{-a/x}\right)(1 - x) - \frac{m_2}{m_1}\mu = 0 \tag{7.21}$$

where

$$a \equiv \frac{2\mu m_2 N}{\sigma^2}. \tag{7.22}$$

First, observe that $f(1) = -(m_2/m_1)\mu > 0$. Next, $f(x) \to -\infty$ as $x \downarrow 0$. Finally, differentiating (7.21), we obtain

$$\frac{df(x)}{dx} = -1 + e^{-a/x}\left(1 - ax^{-2}(1 - x)\right). \tag{7.23}$$

Because $a < 0$, it follows that $df(x)/dx > 0$ for $0 < x < 1$ and consequently there exists a unique solution $x \in (0, 1)$ to (7.21) (see Figure 3). The proof for $\mu > 0$ proceeds similarly. ∎

8 Numerical Examples

The subject of this section is the performance of the refined approximation described in Section 7. Using product-form networks (namely, those whose parameters satisfy conditions (7.6)-(7.7), we compare estimates obtained from the Brownian approximation against exact solutions. Our theory predicts that the approximations are asymptotically exact, that is, as the number of closed

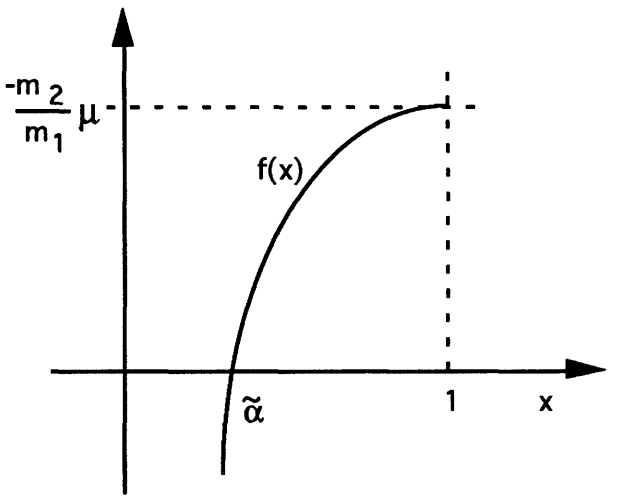

Figure 3: The unique solution $\tilde{\alpha}$

customers N becomes large and as the relative traffic intensity at station 1 ρ_1 approaches unity. We are interested in the accuracy of the estimates for intermediate values of N and ρ, as well as the behavior of the approximations as we take N and ρ toward their respective limits.

An important performance measure in the analysis of mixed (and closed) queueing networks is the throughput rate of closed customers. If m_1 is "much smaller" than m_2, then we can expect the traffic intensity at station 2 to be close to 1 even for small values of N. Consequently, an important test of any approximation scheme is how well it estimates the throughput rate for closed customers when the service times m_1 and m_2 are *approximately equal.* Restricting our attention to product form networks, we set $\lambda = 0.01$, $m_2 = 1$, $c_a^2 = c_0^2 = c_1^2 = c_2^2 = 1$, and consider three systems. In System 1, $m_1 = m_2 = 0.8$; System 2 has $m_1 = m_2 = 0.9$; and in System 3, we let $m_1 = m_2 = 0.95$. The relative traffic intensity at station 1, ρ_1, is then 0.81, 0.91, and .96 for Systems 1, 2, and 3, respectively. For each system, we consider values of N between 1 and 100. (Observe that because $m_2 = 1$, the throughput rate is given by α^*, the fraction of busy time at station 2.)

Figure 4 compares the throughput rate approximations ($\tilde{\alpha}$) against exact solutions (α^*) for System 1 (here, $\rho_1 = .81$). We display throughput rates as a function of N, the number of closed customers in the system. As we expect, the throughput rate approaches 1 as N increases. The throughput rate is clearly less than 1 for small values of N, but it approaches 1 rapidly as N increases. In particular, the throughput rate increases to 0.909, 0.975, and 0.997 for $N = 5$, $N = 10$, and $N = 15$, respectively. The difference between the throughput rate and unity, of course, is even smaller for values of ρ_1 closer to 1. Figure 5 shows the relative percentage error in throughput time approximations for each system (calculated via $(\tilde{\alpha} - \alpha^*)/\alpha^*$). The approximations are reasonably good even for values of $N < 5$, and the accuracy increases dramatically as N increases. The performance of these approximations for intermediate values of N are shown in greater detail in

Figure 6. In all cases, the approximation is within 1% of the exact solution for $N \geq 15$. As we also expect, the quality of the approximation increases as ρ_1 becomes closer to 1.

Figures 7 and 8 display the queue lengths of open and closed customers at station 1, respectively, for System 2. One can see that the approximations are good even for intermediate values of N. The accuracy of these estimates are clarified in Figures 9-12. The queue length of open customers is the performance measure of interest in Figures 9 and 10. As Figure 10 shows, the estimates are with 10% of exact solutions for $N \geq 5$. Finally, Figures 11-12 displays estimates for queue length of closed customers. Figure 11 bears out our expectation that when ρ_1 is closer to 1, larger values of N are required for good estimates. Here, the estimates perform poorly for small values of N, but in all cases, $N \geq 15$ gives estimates that are 10% from exact solutions.

9 Appendix

This section is devoted to characterizing some mappings that are used in the proofs of Sections 5-6. Some of the results here are adapted from the work of Chen and Mandelbaum [5].

Fix $\epsilon > 0$. Let $\mathbf{D}_\epsilon^d[0,t]$ be the set of $x \in \mathbf{D}$ that satisfy the following conditions: (i) $x(0) \geq 0$; (ii) x has no downward jumps; and (iii) $e'x(s) \geq \epsilon$ for all $s \in [0,t]$. Define $\mathbf{C}_\epsilon^d$ to be the continuous functions in $\mathbf{D}_\epsilon^d$, namely,

$$\mathbf{C}_\epsilon^d = \left\{ x \in \mathbf{C}^d : x(0) \geq 0, e'x(s) > \epsilon \text{ for all } t \geq 0 \right\}.$$

Denote by Λ the set of functions $a \in \mathbf{D}$ that have the following properties: (i) a is nondecreasing; (ii) $0 \leq a(t) \leq t$ for all $t \geq 0$; (iii) for each finite t, there is a finite number of subintervals $0 = s_0 < s_1 < \cdots < s_N = t$ and constants $0 \leq a_0 < a_1 < a_2 < \cdots$ such that either $a(t) = t$ or $a(t) = a_i$ for $t \in [s_{i-1}, s_i)$. In particular, observe that $e(t) = t$ is an element of Λ. For $x = (x_1, x_2) \in \mathbf{D}_\epsilon^2$, $a \in \Lambda$, and $0 < c < 1$, let $w = (w_1, w_2)$, $y = (y_1, y_2)$ be a solution of the mapping $(w, y) = (\Phi, \Psi)(x, a)$ defined by the following properties:

$$w_1(t) = \Phi_1(x,a)(t) \equiv x_1(t) + y_1(t) - cy_2(t) \tag{9.1}$$

$$w_2(t) = \Phi_2(x,a)(t) \equiv x_2(t) - y_1(t) + (1-c)(y_2(t) - y_2(a(t))) + cy_2(t) \tag{9.2}$$

$$y_i \text{ are nondecreasing with } y_i(0) = 0 \tag{9.3}$$

$$y_i \text{ increases only at times } t \text{ where } w_i(t) = 0. \tag{9.4}$$

Observe from (9.1)-(9.2) that

$$e'w(t) = e'x(t) + (1-c)(y_2(t) - y_2(a(t))) \geq e'x(t) \tag{9.5}$$

due to the monotonicity of y_2. The main result of this section is the proof of the following:

Theorem 9.1 *For each $x \in \mathbf{D}_\epsilon^2$ and $a \in \Lambda$, there exists a unique pair of processes (w, y) that satisifes (9.1)-(9.4). In other words, the mapping (Φ, Ψ) is well defined on $\mathbf{D}_\epsilon^2 \times \Lambda$. Moreover, if $x \in \mathbf{C}_\epsilon^2$ and $a(t) = t$, then (Φ, Ψ) is continuous at (x, a). Finally, $y = \Psi(x, a)$ is a continuous process if $x_2 - (1-c)(y_2 \circ a)$ has no jumps downward.*

The following is a special case of Theorem 2.5 of Chen and Madelbaum [5].

Lemma 9.2 *Suppose that $x \in \mathbf{D}_\epsilon^2$ and $a(t) = t$. Then (w, y) are uniquely defined by (9.1)-(9.4), and (w, y) is uniquely given by*

$$\begin{aligned} w_1(t) &= x_1(t) + y_1(t) - cy_2(t) \\ w_2(t) &= x_2(t) - y_1(t) + cy_2(t) \\ y_1(t) &= \sup_{0 \le s \le t} \{x_1(s) - cy_2(s)\}^- \\ y_2(t) &= \frac{1}{c} \sup_{0 \le s \le t} \{x_2(s) - y_1(s)\}^- \end{aligned}$$

Moreover, fixing $a(t) = t$, Ψ is continuous on $\mathbf{C}_\epsilon^2$.

Lemma 9.3 *Given $x \in \mathbf{D}$ and $a \in \Lambda$, define*

$$f_1(z)(t) \equiv \sup_{0 \le s \le t} \{x(s) - (1-c)z(a(s))\}^- \tag{9.6}$$

$$f_2(z)(t) \equiv \frac{1}{c} \sup_{0 \le s \le t} \{x(s) + (1-c)(z(s) - z(a(s)))\}^- . \tag{9.7}$$

There exists a unique solution to (9.6) and the solution uniquely satisfies (9.7).

Proof. The key to the proof is the observation that $a(t) \le t$, from which we obtain

$$\begin{aligned} \|f_1(z)(\cdot) - f_1(z')(\cdot)\|_t &\le (1-c)\|z(a(\cdot)) - z'(a(\cdot))\|_t \\ &\le (1-c)\|z(\cdot) - z'(\cdot)\|_t. \end{aligned}$$

Hence, f_1 is a contraction mapping in x and there exists a unique solution to (9.6). Denote by z the fix point solution of (9.6) and observe that

$$\begin{aligned} z(t) &= \sup_{0 \le s \le t} \{x(s) - (1-c)z(a(s))\}^- \\ &= \sup_{0 \le s \le t} \{x(s) + (1-c)(z(s) - z(a(s))) - (1-c)z(s)\}^- \\ &\le \sup_{0 \le s \le t} \{x(s) + (1-c)(z(s) - z(a(s)))\}^- + (1-c) \sup_{0 \le s \le t} \{-z(s)\}^- \\ &= cf_2(z) + (1-c)z(t), \end{aligned}$$

where the first inequality follows because z is nonnegative and the last equality is a result of the monotonicity of z. Therefore, $z(t) \leq f_2(z)(t)$. On the other hand, noting that

$$z(t) + x(s) - (1-c)z(a(s)) \geq 0$$

for all $0 \leq s \leq t$, we have

$$\begin{aligned} f_2(z)(t) &= \frac{1}{c} \sup_{0 \leq s \leq t} \{x(s) - (1-c)z(a(s)) + z(s) - cz(s)\}^- \\ &\geq \frac{1}{c} \sup_{0 \leq s \leq t} \{-cz(s)\}^- \\ &= z(t), \end{aligned}$$

and z is a solution of (9.7). Now let z' be another solution of (9.7). We have

$$\begin{aligned} f_1(z')(t) &= \sup_{0 \leq s \leq t} \{x(s) - (1-c)z'(a(s))\}^- \\ &= \sup_{0 \leq s \leq t} \{x(s) + (1-c)(z'(s) - z'(a(s))) - (1-c)z'(s)\}^- \\ &\leq \sup_{0 \leq s \leq t} \{x(s) + (1-c)(z'(s) - z'(a(s))) + cz'(s)\}^- + \sup_{0 \leq s \leq t} \{-z'(s)\}^- \\ &= z'(t) \end{aligned}$$

and

$$\begin{aligned} f_1(z')(t) &= \sup_{0 \leq s \leq t} \{x(s) + (1-c)(z'(s) - z'(a(s))) - (1-c)z'(s)\}^- \\ &\geq \sup_{0 \leq s \leq t} \{-z'(s)\}^- \\ &= z'(t) \end{aligned}$$

because $x(s) + (1-c)(z'(s) - z'(a(s))) + cz'(s) \geq 0$ for all $s \geq 0$. We have shown that z' is a solution to f_1, but f_1 has a unique solution so we can conclude that $z' = z$. ∎

Remark: Suppose that $a(t) = 0$ for all $t \geq 0$ and $x(0) \geq 0$. It then follows from Lemma 9.3 that

$$z(t) = \sup_{0 \leq s \leq t} \{x(s)\}^- = \frac{1}{c} \sup_{0 \leq s \leq t} \{x(s) + (1-c)z(s)\}^- .$$

Lemma 9.4 *Given $x \in \mathbf{D}_e^2$ and $a \in \Lambda$, suppose that (w, y) is the unique solution of (9.1)-(9.4) on the interval $[0, \tau)$. Then there exists a unique extension of (w, y) to the interval $[0, \tau]$.*

Proof. To extend the definition of (w, y) to the endpoint τ, observe from (9.2)-(9.4) that $(w(\tau) - w(\tau^-), y(\tau) - y(\tau^-))$ satisfy

$$w_1(\tau) = (w_1(\tau^-) + x_1(\tau) - x_1(\tau^-)) + (y_1(\tau) - y_1(\tau^-)) - c\,(y_2(\tau) - y_2(\tau^-)) \tag{9.8}$$

$$w_2(\tau) = (w_2(\tau^-) + x_2(\tau) - x_2(\tau^-)) - (1-c)\,(y_2(\tau^-) - y_2(a(\tau^-)))$$

$$+(1-c)\,(y_2(\tau) - y_2(a(\tau))) - (y_1(\tau) - y_1(\tau^-)) \tag{9.9}$$

$$+c\,(y_2(\tau) - y_2(\tau^-)) \tag{9.10}$$

$$y_i(\tau) - y_i(\tau^-) \geq 0 \tag{9.11}$$

$$w(\tau)'\,(y(\tau) - y(\tau^-)) = 0. \tag{9.12}$$

Because

$$\begin{aligned} & e'(w(\tau^-) + x(\tau) - x(\tau^-)) - (1-c)\,(y_2(\tau^-) - y_2(a(\tau^-))) + (1-c)\,(y_2(\tau) - y_2(a(\tau))) \\ = \;& e'x(\tau) + (y_2(\tau) - y_2(a(\tau))) \\ \geq \;& e'x(\tau) > 0, \end{aligned}$$

where the first inequality is a result of (9.5),it follows from Theorem 4.3 of Chen and Mandelbaum [5] that (9.8)-(9.12) produces a unique solution for $y(\tau) - y(\tau^-)$. ∎

For $x \in \mathbf{D}$ and $a \in \Lambda$, define the mappings

$$u(t) = g_1(x,a)(t) \equiv \sup_{0 \leq s \leq t} \{x(s) + (1-c)(u(s) - u(a(s)))\}^- \tag{9.13}$$

$$v(t) = g_2(x,a)(t) \equiv x(t) + (1-c)(u(t) - u(a(t)) + cu(t). \tag{9.14}$$

For a sequence T_k, $k = 1, 2, \ldots$, let us define the "shifted" processes

$$\begin{aligned} x^k(t) &= v(T_k) + x(t + T_k) - x(T_k) \\ u^k(t) &= u(t + T_k) - u(T_k) \\ (u \circ a)^k(t) &= u(a(t + T_k)) - u(a(T_k)) \\ v^k(t) &= v(t + T_k). \end{aligned}$$

It is straightforward to verify from (9.13)-(9.14) that

$$u^k(t) = g_1^k(x^k, a)(t) \equiv \frac{1}{c} \sup_{0 \leq s \leq t} \left\{x^k(s) + (1-c)(u^k(s) - (u \circ a)^k(s))\right\}^- \tag{9.15}$$

$$v^k(t) = g_2^k(x^k, a)(t) \equiv x^k(t) + (1-c)(u^k(t) - (u \circ a)^k(t)) + cu^k(t). \tag{9.16}$$

Remark: If we define the mappings

$$h_1(x,a)(t) \equiv \frac{1}{c}\sup_{0\le s\le t}\{x(s)\}^- \tag{9.17}$$

$$h_2(x,a)(t) \equiv x(t) + ch_1(x,a)(t), \tag{9.18}$$

it is clear that (h_1, h_2) is a special case of (g_1, g_2) with a equal to e where $e(t) = t$ is the identity map.

Lemma 9.5 *For each $x \in \mathbf{D}_\epsilon^2$ and $a \in \Lambda$, there exists a pair of functions (w, y) that satisifes (9.1)-(9.4).*

Proof. Noting that it suffices to set $\epsilon = 1$, we first prove the lemma for $x \in \mathbf{C}_1^2$. Fix $0 < \delta < 1/2$. We may assume without loss of generality that $w_2(0) \ge 1/2$ (otherwise, it follows from $x \in \mathbf{C}_1^2$ and (9.5) that $w_1(0) \ge 1/2$ and we proceed similarly).

For an increasing sequence of times T_k, $k = 1, 2, \ldots$, and a pair of functions (w, y) satisfying $(w, y) = (\Phi, \Psi)(x, a)$, it will be necessary to refer to the following "shifted" processes:

$$x_i^k(t) \equiv w_i(T_k) + x_i(t + T_k) - x_i(T_k) \tag{9.19}$$

$$y_i^k(t) \equiv y_i(t + T_k) - y_i(T_k) \tag{9.20}$$

$$(y_2 \circ a)^k(t) \equiv y_2(a(t + T_k)) - y_2(a(T_k)) \tag{9.21}$$

$$w_i^k(t) \equiv w(t + T_k). \tag{9.22}$$

It is straightforward to verify that the mappings (9.1)-(9.2) yield

$$w_1^k(t) = x_1^k(t) + y_1^k(t) - cy_2^k(t) \tag{9.23}$$

$$w_2^k(t) = x_2^k(t) - y_1^k(t) + (1-c)(y_2^k(t) - (y_2 \circ a)^k(t)) + cy_2^k(t); \tag{9.24}$$

moreover,

$$e'w^k(t) = e'x^k(t) + (1-c)(y_2^k(t) - (y_2 \circ a)^k(t)) \tag{9.25}$$

and

$$\begin{aligned} e'x^k(t) \;+\; & (1-c)(y_2^k(t) - (y_2 \circ a)^k(t)) = e'w(T_k) + e'x(t + T_k) - e'x(T_k) \\ & (1-c)(y_2(t + T_k) - y_2(T_k)) - (1-c)(y_2(a(t + T_k)) - y_2(a(T_k))) \\ = \;& e'x(t + T_k) + (1-c)(y_2(t + T_k) - y_2(a(t + T_k))), \end{aligned} \tag{9.26}$$

where the equality follows directly from equation (9.5)

Set $T_0 \equiv 0$, $y_2^0(t) = 0$, and observe that by monotonicity, $y_2^0(a(t)) = 0$ for all $t \ge 0$. Define y_1^0 by the mapping (9.17), namely, $y_1^0 = h_1(x_1^0, a)$, and let w^0 be given by equations (9.23)-(9.24) with

$k = 0$. Observe that y_1^0 and hence w^0 are uniquely defined by Theorem 2.2.3 of Harrison [10]. For $k = 0$, set

$$t_{k+1} = \inf\{t \geq 0 : e'x^k(t) + (1-c)(y_2^k(t) - (y_2 \circ a)^k(t)) - w_1^k(t) \leq \delta\}. \tag{9.27}$$

If $t_1 = \infty$ then we are done so let us assume $t_1 < \infty$. First, note from (9.25) that $e'x^0(0) + (1-c)(y_2^0(0) - y_2^0(a(0))) - w_1^0(0) = w_2^0(0) > \delta$. Let $[0, s_1)$ be the first interval associated with the function a. If a takes a constant value over this interval, then $(y_2 \circ a)^k(t) = 0$ for $t < s_1$. On the other hand, if a is the identity map over this interval, then $y_2^k(t) - (y_2 \circ a)^k(t) = 0$ on $[0, s_1)$. In either case, we may conclude that $e'x^k(t) + (1-c)(y_2^k(t) - (y_2 \circ a)^k(t)) - w_1^k(t)$ has no negative jumps on $[0, s_1)$ because x has no downward jumps and y_2^k is nondecreasing. Hence, we may conclude that $t_1 > 0$. Moreover, $w_2^0(s) = e'x^0(s) + (1-c)(y_2^0(s) - y_2^0(a(s))) - w_1^0(s) \geq \delta$ for $x \in [0, t_1]$, and we have shown that $(w, y) = (w^0, y^0)$ is a solution of (9.1)-(9.4) over the time period $t \in [0, t_1]$. Define $T_k = t_1 + \ldots + t_k$ and observe that for $k = 0$,

$$\begin{aligned} w_1(T_{k+1}) &= w_1^k(t_{k+1}) \\ &\geq e'x^k(t_{k+1}) + (1-c)(y_2^k(t_{k+1}) - (y_2 \circ a)^k(t_{k+1})) - \delta \\ &= e'x(T_{k+1}) + (1-c)(y_2(T_{k+1}) - y_2(a(T_{k+1}))) - \delta \\ &\geq 1 - \delta > \delta, \end{aligned} \tag{9.28}$$

where the first equality follows from (9.26) and the last inequality follows from (9.5) and the monotonicity of y_2.

We now shift time 0 to T_1 via the mappings (9.19)-(9.21) setting $k = 1$. Set $y_1^1(t) = 0$ and let y_2^1 be defined by (9.15), namely, $y_2^1 = g_1^k(x_2^k, a)$. The functions w_1^1, y_2^1 are then defined using (9.23)-(9.24) with $k = 1$. Let $[s_{l-1}, s_l)$ denote the interval corresponding to the function a that contains T_1. If $a(t)$ takes a constant value over this interval, then $(y_2 \circ a)^1(t) = 0$ for $t < s_l - T_1$. If, on the other hand, $a(t) = t$, then $y_2^1(t) - (y_2 \circ a)^1(t) = 0$ for $t < s_l - T_1$. Thus y_2^1 is uniquely defined over the interval $[0, s_l - T_1)$ whether $a(t) = t$, in which case we invoke Theorem 2.2.3 of Harrison [10], or $a(t)$ takes on a constant value, in which case the remark following Lemma 9.3 applies. The definition of (w, y) can then be extended to the endpoint $s_l - T_1$ using Lemma 9.4. Using the special structure of a, where a is either a constant or the identity map over intervals of time, one can thus uniquely define (y^1, w^1) for all $t \geq 0$. For $k = 1$ let

$$t_{k+1} = \inf\{t \geq 0 : e'x^k(t) + (1-c)(y_2^k(t) - y_2^k(a(t))) - w_2^k(t) \leq \delta\}. \tag{9.29}$$

By the definition of T_1, we have from (9.28) that $w_1^1(0) = w_1(T_1) > \delta$ and consquently $e'x^1(0) + (1-c)(y_2^1(0) - y_2^1(a(0))) - w_2^1(0) = w_1^1(0) > \delta$. We can use the same argument as before to show that $e'x^k(t) + (1-c)(y_2^k(t) - (y_2 \circ a)^k(t)) - w_w^k(t)$ has no jumps downward, from which we may conclude $t_2 > 0$. In addition, $w_1^1(s) = e'x^1(s) + (1-c)(y_2^1(s) - (y_2 \circ a)^1(s)) - w_2^1(s) \geq \delta$ for $s \in [0, t_2]$. Define for $k = 1$

$$y(t) = \begin{cases} y(t) & 0 \leq t \leq T_k \\ y(T_k) + y^k(t - T_k) & t \geq T_k. \end{cases} \tag{9.30}$$

The pair (w, y) thus constructed is a solution of (9.1)-(9.4) over the time period $t \in [0, T_2]$. Finally, observe that for $k = 1$,

$$\begin{aligned} w_2(T_{k+1}) &= w_2^k(t_{k+1}) \\ &\geq e'x^k(t_{k+1}) + (1-c)(y_2^k(t_{k+1}) - (y_2 \circ a)^k(t_{k+1})) - \delta \\ &= e'x(T_{k+1}) + (1-c)(y_2(T_{k+1}) - y_2(a(T_{k+1}))) - \delta \\ &\geq 1 - \delta > \delta. \end{aligned} \tag{9.31}$$

Iterating in this way, we can construct a pair (w, y) that satisfies (9.1)-(9.4) on the interval $[0, T_k]$. In particular, if k is even, set $y_2^k(t) = 0$ and let $y_1^k = h_1(x_1^k, a)$. If k is odd, we set $y_1^k(t) = 0$ and let $y_2^k = g_1^k(x_2^k, a)$. In either case, w^k is defined according to (9.23)-(9.24). Similarly, we use either (9.27) or (9.29) to define t_{k+1} depending on whether k is even or odd, respectively. The process (w, y) on the interval $[0, T_k]$ is then constructed via the concatenation map given in (9.30) and property (9.22). Our construction for $x \in \mathbf{C}_1^2$ is thus complete if we can show that for each fixed t, there exists finite n^* with $T_{n^*} \geq t$. To do so, let us suppose to the contrary that there is some finite t for which $T_n < t$ for all $n \geq 1$. If k is even, we have the following inequality due to (9.5) and the definition of T_k

$$\begin{aligned} w_1(T_{k+1}) - w_1(T_k) \geq\ & \frac{1}{2} - \delta + e'x(T_{k+1}) - e'x(T_k) + (1-c)(y_2(T_{k+1}) - y_2(a(T_{k+1}))) \\ & -(1-c)(y_2(T_k) - y_2(a(T_k))); \end{aligned}$$

for k odd, we have

$$\begin{aligned} w_2(T_{k+1}) - w_2(T_k) \geq\ & \frac{1}{2} - \delta + e'x(T_{k+1}) - e'x(T_k) + (1-c)(y_2(T_{k+1}) - y_2(a(T_{k+1}))) \\ & -(1-c)(y_2(T_k) - y_2(a(T_k))). \end{aligned}$$

Because $a \in \Lambda$, there are a finite number of intervals partitioning $[0, t]$ such that a is either the identity map or a constant value over each subinterval. From the finiteness of these subintervals, there must be an interval $[s_{l-1}, s_l)$ such that $T_k \in [s_{l-1}, s_l)$ for all $k \geq k^*$. For such k, either

$$(1-c)\left[(y_2(T_{k+1}) - y_2(a(T_{k+1}))) - y_2(T_k) - y_2(a(T_k))\right] = 0$$

or

$$(1-c)\left[(y_2(T_{k+1}) - y_2(a(T_{k+1}))) - (y_2(T_k) - y_2(a(T_k))\right] = (1-c)(y_2(T_{k+1}) - y_2(T_k)) \geq 0,$$

depending on whether a is the indentity map or a constant value on this subinterval, respectively. In either case, we may conclude that for all $k \geq k^*$,

$$\max_j \sup_{T_k \leq s \leq t \leq T_{k+1}} (w_j(t) - w_j(s)) \geq \frac{1}{2} - \delta - 2 \max_j \sup_{T_k \leq s \leq t \leq T_{k+1}} |x_j(t) - x_j(s)|. \tag{9.32}$$

It is straightforward to extend identity 2.8.G. of Chen and Mandelbaum [5] to this setting, which states that

$$\sup_{u\le s\le t\le v}(w(t)-w(s)) \le \sup_{u\le s\le t\le v}|x(t)-x(s)|. \tag{9.33}$$

Substituting (9.33) in (9.32), we have

$$\max_j \sup_{T_k\le s\le t\le T_{k+1}} |x_j(t)-x_j(s)| \ge \frac{1}{3}\left(\frac{1}{2}-\delta\right) > 0. \tag{9.34}$$

However, x is uniformly continuous on $[0,t]$, so there exists $\eta > 0$ such that

$$\max_j |x_j(t)-x_j(s)| < \frac{1}{3}\left(\frac{1}{2}-\delta\right)$$

for all $s,t \in [0,t]$ with $|s-t| < \eta$. The inequality (9.34) together with the assumption that $T_k < t$ for all $k \ge k^*$ imply that $T_k > \eta$ for all $k \ge k^*$. However, this contradicts the finiteness of t.

It remains to extend the construction to $x \in \mathbf{D}_1^2$. This is done by noting that x has only a finite number of jumps over each interval $[0,t]$. Lemma 9.4 is then applied to show that there exists a unique extension at each jump point of x. ∎

Lemma 9.6 *Let x and a satisfy the conditions Theorem 9.1. Then (9.1)-(9.4) have a unique solution.*

Proof. The proof proceeds as in the proof of Proposition 2.4 of Chen and Mandelbaum [5]. Let (w,y) be the process constructed in the proof of Lemma 9.5, and let (w',y') be another process satisfying (9.1)-(9.4). Suppose we can show that

1. y and y' coincide on $[0,\delta]$ for some $\delta > 0$;
2. if $y(\tau) = y'(\tau)$ at some $t \ge 0$, then the also two coincide on $[\tau,\tau+\delta]$ for some positive δ;
3. if $y(t) = y'(t)$ on $t \in [0,\tau)$ then $y(\tau) = y'(\tau)$.

Defining

$$\tau = \sup\{t \ge 0 : y(s) = y'(s) \text{ for all } 0 \le s \le t\},$$

it follows from (1) that $\tau \ge \delta$. Suppose $\tau < \infty$. Then (3) holds hence y and y' coincide beyond τ. This contradicts the definition so we can conclude that $\tau = \infty$. The proof now rests on establishing (1)-(3).

The proof of (1) follows from the construction y^0 in the proof of Lemma 9.5. The proof of (2) then follows from (1) by applying a time shift as in the proof of Lemma 9.5. The proof of (3) is an application of Lemma 9.4 ∎

Lemma 9.7 *If $x_2(\cdot) - (1-c)(y_2 \circ a)(\cdot)$ has no downward jumps where (w, y) satisfies (9.1)-(9.4), then y is continuous.*

Proof. We only need to show that $y(t) = y(t^-)$ for all $t > 0$. Consider the problem posed in (9.8)-(9.12) of Lemma 9.4. As argued in the proof of this Lemma, there exists a unique solution for $y(t) - y(t^-)$. Because x has no downward jumps, $w_1(\tau^-) + x_1(\tau) - x_1(\tau^-) \geq w_1(\tau^-) \geq 0$. Moreover, because we assume that $x_2(\cdot) - (1-c)(y_2 \circ a)(\cdot)$ has no negative jumps,

$$\begin{aligned} & w_2(t^-) + x_2(t) - x_2(t^-) - (1-c)\left(y_2(a(t)) - y_2(a(t^-))\right) + (1-c)\left(y_2(t) - y_2(t^-)\right) \\ \geq\ & w_2(t^-) + (1-c)\left(y_2(t) - y_2(t^-)\right) \\ \geq\ & w_2(t^-) \geq 0. \end{aligned}$$

Hence, $y(t) - y(t^-) = 0$ is the unique solution. ∎

Let us define the modulus of continuity

$$\omega_{\eta,t}(x) = \sup_{0 \leq r,s \leq t, |r-s| < \eta} |x(r) - x(s)|. \tag{9.35}$$

Lemma 9.8 *Let $x, x' \in \mathbf{D}$ and $a, e \in \Lambda$ where $e(t) = t$. Define*

$$(u, v) = (g_1, g_2)(x, a), \quad (u', v') = (g_1, g_2)(x', e),$$

and

$$(u^k, v^k) = (g_1^k, g_2^k)(x^k, a), \quad (u'^k, v'^k) = (g_1^k, g_2^k)(x'^k, e).$$

Fix $t > 0$ and set

$$\eta = \|e(\cdot) - a(\cdot)\|_t.$$

We have the following inequalities:

$$\|u(\cdot) - u(a(\cdot))\|_t \leq \frac{1}{c}\omega_{\eta,t}(x), \tag{9.36}$$

$$\|u^k(\cdot) - (u \circ a)^k(\cdot)\|_t \leq \frac{2}{c}\omega_{\eta,t+T_k}(x), \tag{9.37}$$

$$\|u^k(\cdot) - u'^k(\cdot)\|_t \leq \frac{1}{c}\|x^k(\cdot) - x'^k(\cdot)\|_t + 2\left(\frac{1-c}{c^2}\right)\omega_{\eta,t+T_k}(x), \tag{9.38}$$

$$\|v^k(\cdot) - v'^k(\cdot)\|_t \leq 2\|x^k(\cdot) - x'^k(\cdot)\|_t + 4\left(\frac{1-c}{c}\right)\omega_{\eta,t+T_k}(x). \tag{9.39}$$

Proof. From Lemma 9.3, we have the equivalent representation

$$u(t) = g_1(x, a)(t) = \sup_{0 \leq s \leq t} \{x(s) - (1-c)u(a(s))\}^-. \tag{9.40}$$

With (9.40), we have

$$\begin{aligned} 0 \le u(t) - u(a(t)) &= \sup_{a(t)\le s\le t} \{(x(s) - x(a(t))) - (1-c)(u(a(s)) - u(a(a(t))))\}^- \\ &\le \sup_{a(t)\le s\le t} |x(s) - x(a(t))| + (1-c) \sup_{a(t)\le s\le t} |u(a(s)) - u(a(a(t)))|. \end{aligned}$$

Thus

$$\begin{aligned} \|u(\cdot) - u(a(\cdot))\|_t &\le \omega_{\eta,t}(x) + (1-c)\|u(a(\cdot)) - u(a(a(\cdot)))\|_t \\ &\le \omega_{\eta,t}(x) + (1-c)\|u(\cdot) - u(a(\cdot))\|_t \end{aligned}$$

and equation (9.36) is proved. Equation (9.37) is proved similarly by observing that

$$\begin{aligned} \|u^k(\cdot) - (u\circ a)^k(\cdot)\|_t &\le \|u(\cdot + T_k) - u(a(\cdot + T_k))\|_t + |u(T_k) - u(a(T_k))| \\ &\le 2\|u(\cdot) - u(a(\cdot))\|_{t+T_k} \\ &\le \frac{2}{c}\omega_{\eta,t+T_k}(x) \end{aligned}$$

where the last inequality is an application of (9.36).

Next, we obtain from (9.15) and (9.37)

$$\begin{aligned} \|u^k(\cdot) - u'^k(\cdot)\|_t &\le \frac{1}{c}\|x^k(\cdot) - x'^k(\cdot)\|_t + \frac{1-c}{c}\|u^k(\cdot) - (u\circ a)^k(\cdot)\|_t \\ &\le \frac{1}{c}\|x^k(\cdot) - x'^k(\cdot)\|_t + 2\left(\frac{1-c}{c^2}\right)\omega_{\eta,t+T_k}(x). \end{aligned}$$

Finally, observe that

$$\begin{aligned} \|v^k(\cdot) - v'^k(\cdot)\|_t &\le \|x^k(\cdot) - x'^k(\cdot)\|_t + c\|u^k(\cdot) - u'^k(\cdot)\|_t + +(1-c)\|u^k(\cdot) - (u\circ a)^k(\cdot)\|_t \\ &\le 2\|x^k(\cdot) - x'^k(\cdot)\|_t + 4\left(\frac{1-c}{c}\right)\omega_{\eta,t+T_k}(x) \end{aligned}$$

and the proof of the lemma is finished. ∎

Lemma 9.9 *Suppose that $x \in \mathbf{C}^2_\epsilon$ and $a(t) = t$. Then Ψ is continuous at (x, a).*

Proof. Fix $t > 0$. We will make use of the procedure described in the proof of Lemma 9.5 to construct the processes $(w, y) = (\Phi, \Psi)(x, a)$ in the interval $[0, t]$. Let δ be the postive constant used in the procedure, T_k the sequence of (increasing) times obtained from the construction ($t_k = T_k - T_{k-1}$), and let n^* the (finite) number of iterations required to construct (w, y) up to time t. We may assume that $T_{n^*} = t$. Denote by (w^k, y^k) the shifted processes defined on the k^{th} iteration (starting with iteration 0). We write (v^k, u^k) to mean the processes obtained by applying

to (x', a') the same mappings used in obtaining (w^k, y^k). We then define (v, u) by the concatenation procedure in (9.30):

$$u(t) = \begin{cases} u(t) & 0 \le t \le T_k \\ u(T_k) + u^k(t - T_k) & t \ge T_k \end{cases} \tag{9.41}$$

and $w(t + T_k) = w^k(t)$ for $0 \le t \le t_{k+1}$.

Fix $\epsilon > 0$. We want to show that there exists $\eta > 0$ such that for any $x' \in \mathbf{C}_\epsilon^2$ and $a' \in \Lambda$ with $||x' - x||_t < \eta$ and $||a' - a||_t < \eta$, then

$$||\Psi_i(x', a') - \Psi_i(x, a)||_t < \epsilon.$$

Note that

$$||\Psi_i(x', a') - \Psi_i(x, a)||_t \le ||\Psi_i(x', a') - \Psi_i(x', a)||_t + ||\Psi_i(x', a) - \Psi_i(x, a)||_t. \tag{9.42}$$

From Lemma 9.2, we can conclude that there exists $\eta_1 > 0$ such that

$$||\Psi_i(x', a) - \Psi_i(x, a)||_t \le \frac{\epsilon}{2}$$

for any $||x' - x||_t < \eta_1$. If we can show that there exists $\eta_2 > 0$ such that $||a' - a||_t < \eta_2$ implies

$$||\Psi_i(x', a') - \Psi_i(x', a)||_t \le \frac{\epsilon}{2},$$

then the lemma is proved by setting $\eta = \min(\eta_1, \eta_2)$. Because x' is continuous, there exists $\eta_2 > 0$ such that

$$\omega_{\eta_2, t}(x') < \frac{\epsilon}{2}\left(\frac{c^2}{2(1-c)}\right)\left(\frac{A-1}{A^{n^*}-1}\right) \tag{9.43}$$

and

$$\omega_{\eta_2, t}(x') < \delta\left[8\left(\frac{1-c}{c^2}\right)\left(\frac{A^{n^*}-1}{A-1}\right) + 4\left(\frac{1-c}{c}\right)\right]^{-1}, \tag{9.44}$$

where $A = 1 + 2/c$. We assume henceforth that $||a' - a||_t < \eta_2$.

We first show that (w, y) and (v, u), constructed as described previously, satisfy the following condition:

$$\max_{j=1,2} ||y_j(\cdot) - u_j(\cdot)||_{T_k} < 2\left(\frac{1-c}{c^2}\right)\left(\frac{A^k - 1}{A-1}\right)\omega_{\eta_2, T_k}(x'). \tag{9.45}$$

From (9.45) and (9.43), we can conclude

$$\max_{j=1,2} ||y_j(\cdot) - u_j(\cdot)||_t < 2\left(\frac{1-c}{c^2}\right)\left(\frac{A^{n^*}-1}{A-1}\right)\omega_{\eta_2, t}(x') < \frac{\epsilon}{2},$$

so the lemma is proved if we can establish the processes we constructed satisfy $(v, u) = (\Phi, \Psi)(x', a')$. First, we prove (9.45). Define

$$\begin{aligned} \alpha^k(t) &= w_{j(k)}(T_k) + x_1'(t + T_k) - x_1'(T_k) \\ \beta^k(t) &= v_{j(k)}(T_k) + x_1'(t + T_k) - x_1'(T_k) \end{aligned}$$

where $j(k) = 1$ if k is odd and $j(k) = 2$ if k is even or zero. In either case, observe that $\alpha^k(t) - \beta^k(t) = w_{j(k)}(T_k) - v_{j(k)}(T_k)$ so

$$\begin{aligned} \|\alpha^k(\cdot) - \beta^k(\cdot)\|_t &\leq \|y_1(\cdot) - u_1(\cdot)\|_{T_k} + \|y_1(\cdot) - u_1(\cdot)\|_{T_k} \\ &\leq 2 \max_{j=1,2} \|y_j(\cdot) - u_j(\cdot)\|_{T_k}. \end{aligned} \tag{9.46}$$

We proceed by induction and the observation that

$$\sup_{T_k \leq t \leq T_{k+1}} |y_j(t) - u_j(t)| \leq |y_j(T_k) - u_j(T_k)| + \|y_j^k(\cdot) - u_j^k(\cdot)\|_t.$$

From (9.38) and (9.46), we can conclude

$$\max_{j=1,2} \sup_{T_k \leq t \leq T_{k+1}} |y_j(t) - u_j(t)| \leq \left(1 + \frac{2}{c}\right) \max_{j=1,2} \|y_j(\cdot) - u_j(\cdot)\|_{T_k} + 2\left(\frac{1-c}{c^2}\right) \omega_{\eta_2, T_{k+1}}(x'). \tag{9.47}$$

It is straightforward to verify that

$$\begin{aligned} \max_{j=1,2} \sup_{T_0 \leq t \leq T_1} |y_j(t) - u_j(t)| &= \|y_j(\cdot) - u_j(\cdot)\|_{T_1} \\ &< 2\left(\frac{1-c}{c^2}\right) \omega_{\eta_2, T_1}(x'). \end{aligned}$$

Equation (9.45) thus follows with an inductive argument.

It remains to show that (v, u) as constructed satisfy $(v, u) = (\Phi, \Psi)(x', a')$. To do so, it is enough to show that $v_{j(k)}^k(t) > 0$ for $0 \leq t \leq t_{k+1}$, where $j(k) = 2$ if k is even and $j(k) = 1$ if k is odd. It suffices to do so for k even, for the argument for the case of k being odd proceeds similarly. Because $e'x'(t) = e'w(t)$, we have

$$\begin{aligned} \|e'x'(\cdot + T_k) - v_1^k(\cdot) - w_2^k(\cdot)\|_{t_{k+1}} &= \|w_1^k(\cdot) - v_1^k(\cdot)\|_{t_{k+1}} \\ &= \|g_2^k(\alpha^k, e)(\cdot) - g_2^k(\beta^k, a')(\cdot)\|_{t_{k+1}} \\ &\leq 2\|\alpha^k(\cdot) - \beta^k(\cdot)\|t_{k+1} + 4\left(\frac{1-c}{c}\right) \omega_{\eta_2, T_{k_1}}(x') \\ &\leq 4 \max_{j=1,2} \|y_j(\cdot) - u_j(\cdot)\|_{T_k} + 4\left(\frac{1-c}{c}\right) \omega_{\eta_2, T_{k_1}}(x') \\ &\leq 8\left(\frac{1-c}{c^2}\right)\left(\frac{A^{n^*} - 1}{A - 1}\right) \omega_{\eta_2, t}(x') + 4\left(\frac{1-c}{c}\right) \omega_{\eta_2, t}(x') \\ &< \delta. \end{aligned}$$

Here, the first inequality follows from (9.39); the second inequality is a result of (9.46); and the last two inequalities follow from (9.45) and (9.44), respectively. Hence, by the definition of T_k,

$$e'x'(t + T_k) - v_1(t + T_k) > w_2(t + T_k) - \delta \geq 0$$

for $0 \leq t \leq t_{k+1}$. But

$$\begin{aligned} v_2(t+T_k) &= e'x'(t+T_k) + (1-c)(u_2(t+T_k) - u_2(a(t+T_k))) - v_1(t+T_k) \\ &\geq e'x'(t+T_k) - v_1(t+T_k) \\ &> 0 \end{aligned}$$

and the theorem is proved. ∎

Acknowledgements: I am grateful to Mike Harrison and Larry Wein for their helpful comments during the course of this research. I would like to thank Marty Reiman for sharing the interpretation discussed in Section 4.

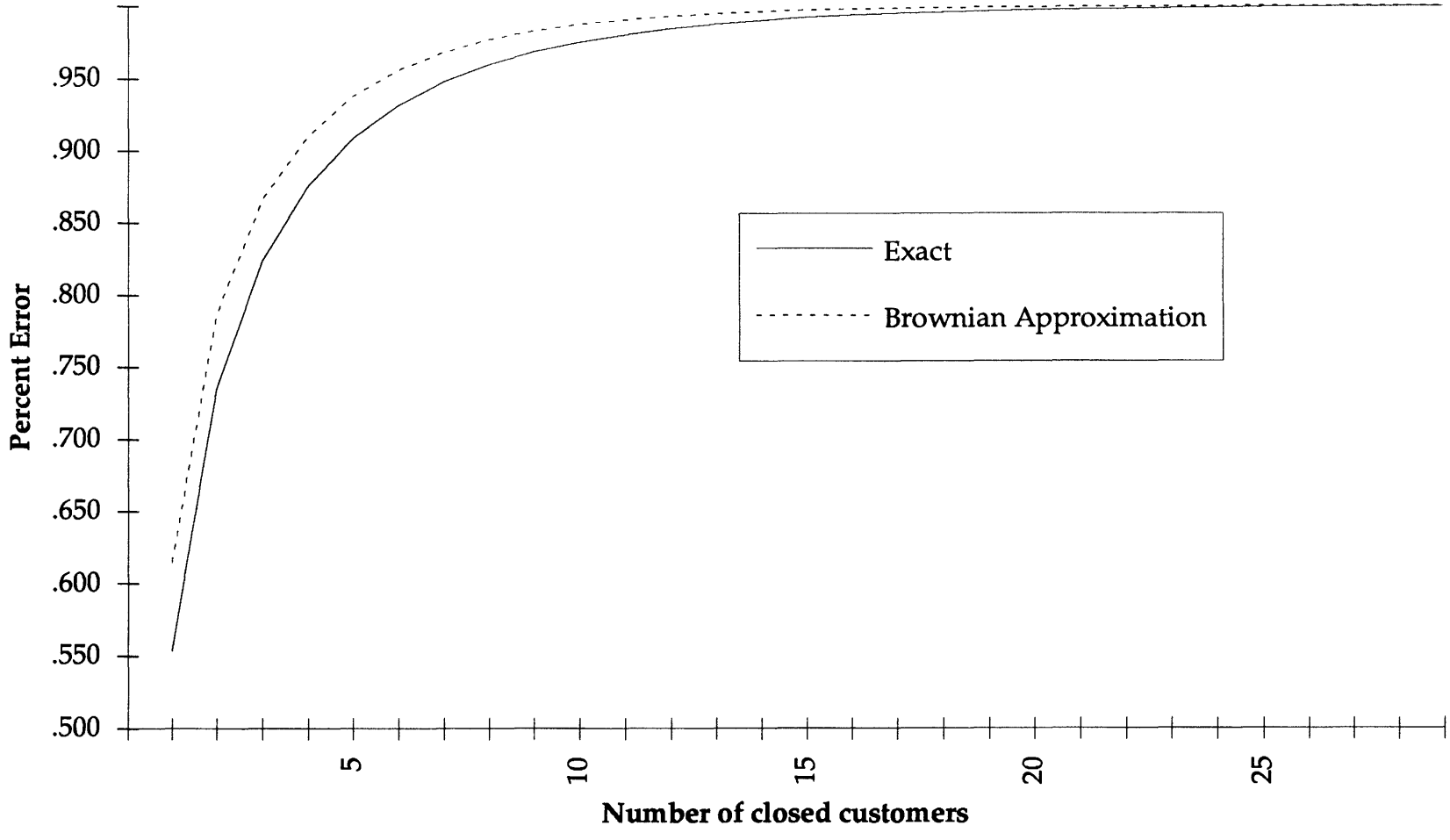

Figure 4: Throughput rate approximations for System 1

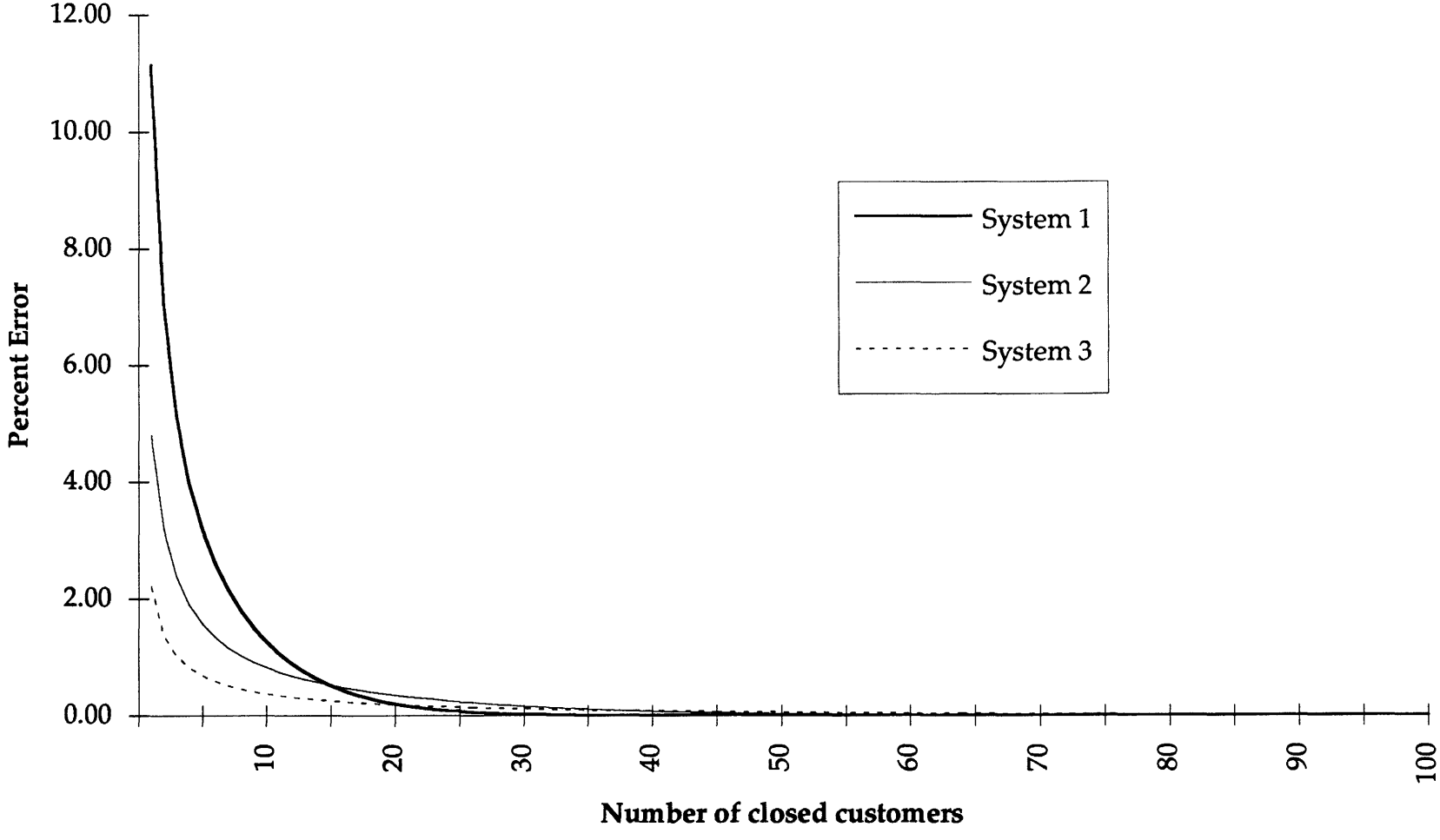

Figure 5: Relative percent error in throughput rate approximations

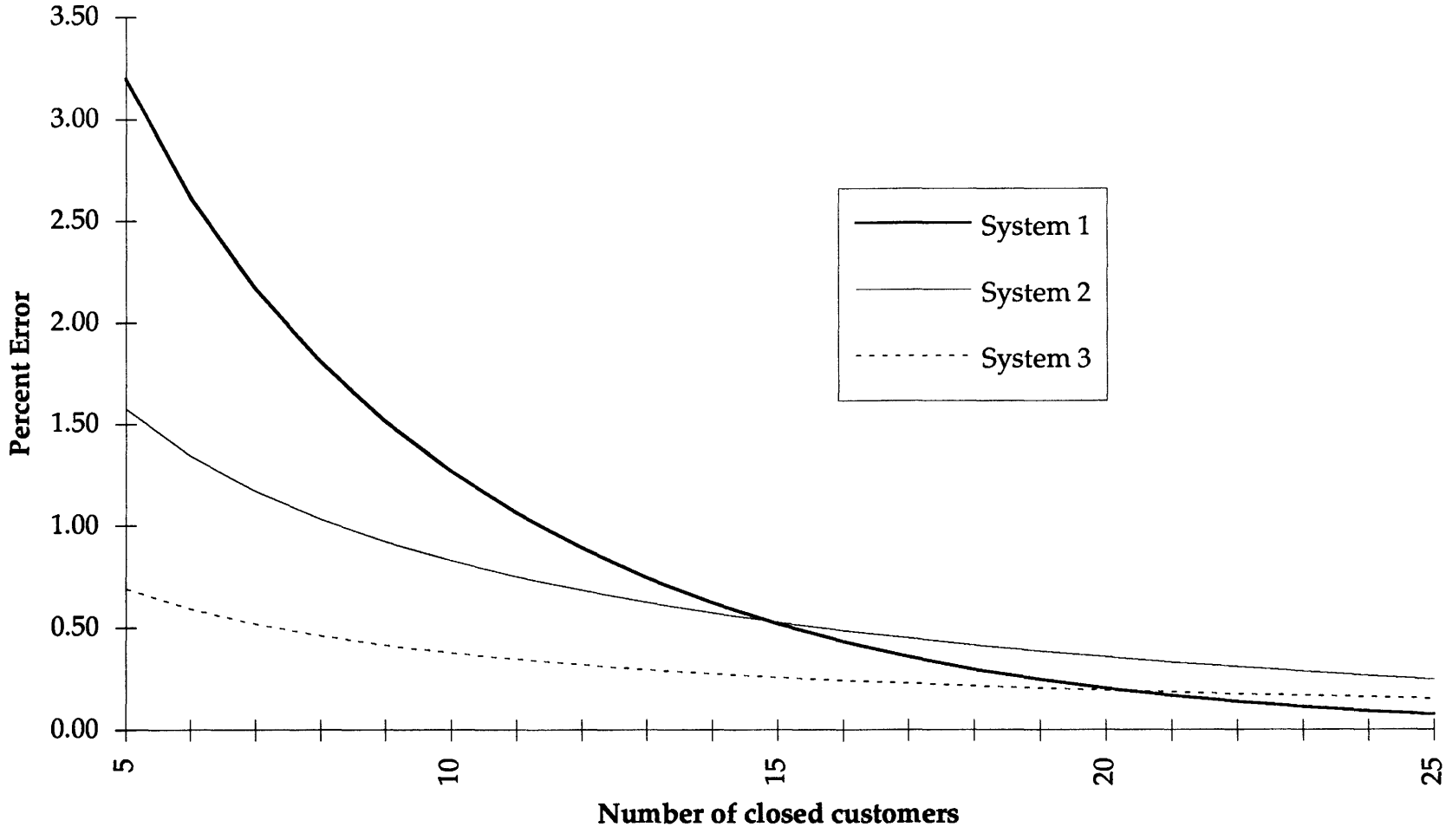

Figure 6: A closer look at relative percent error in throughput rate approximations

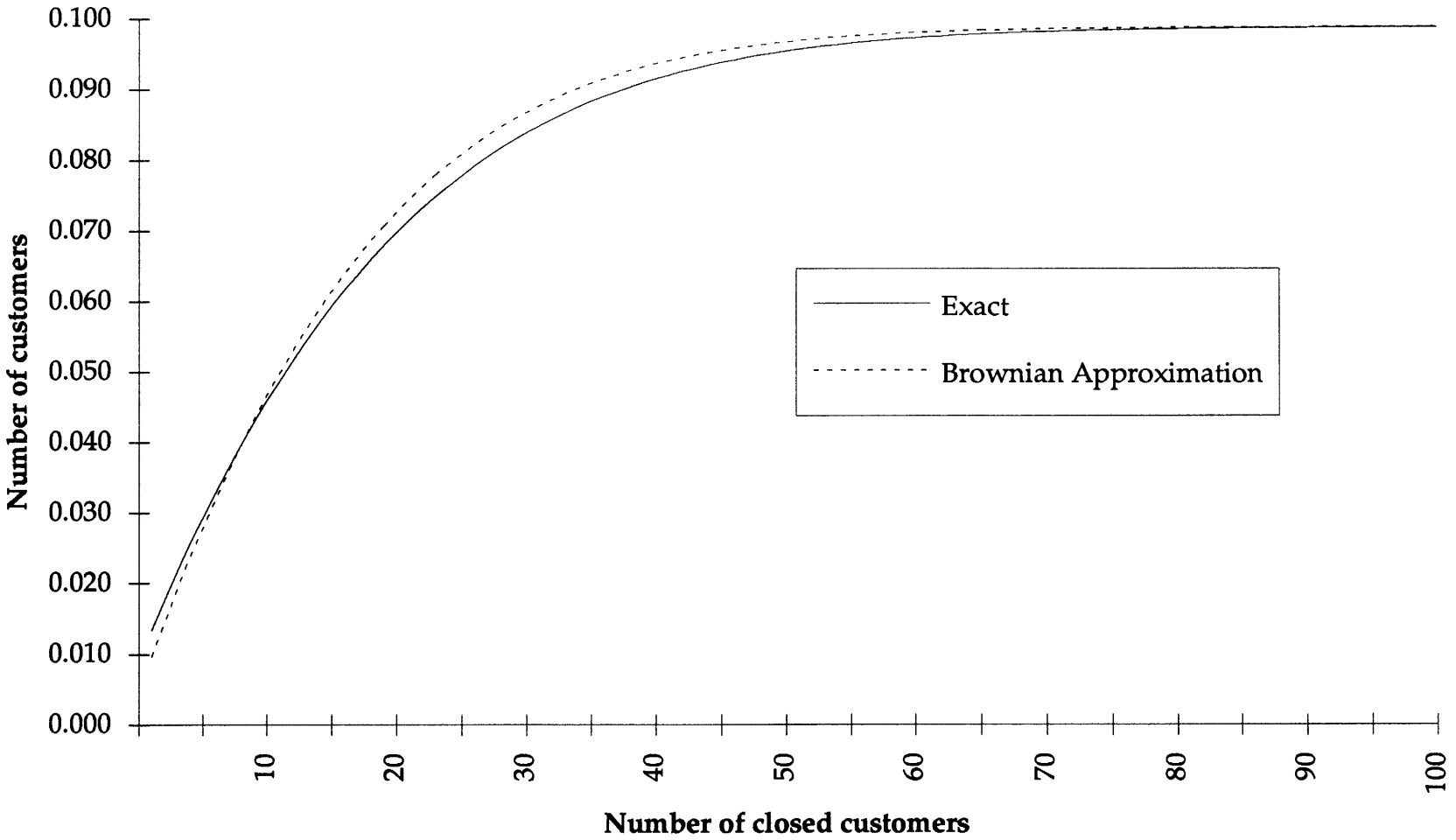

Figure 7: Queue length approximations for open customers in System 2

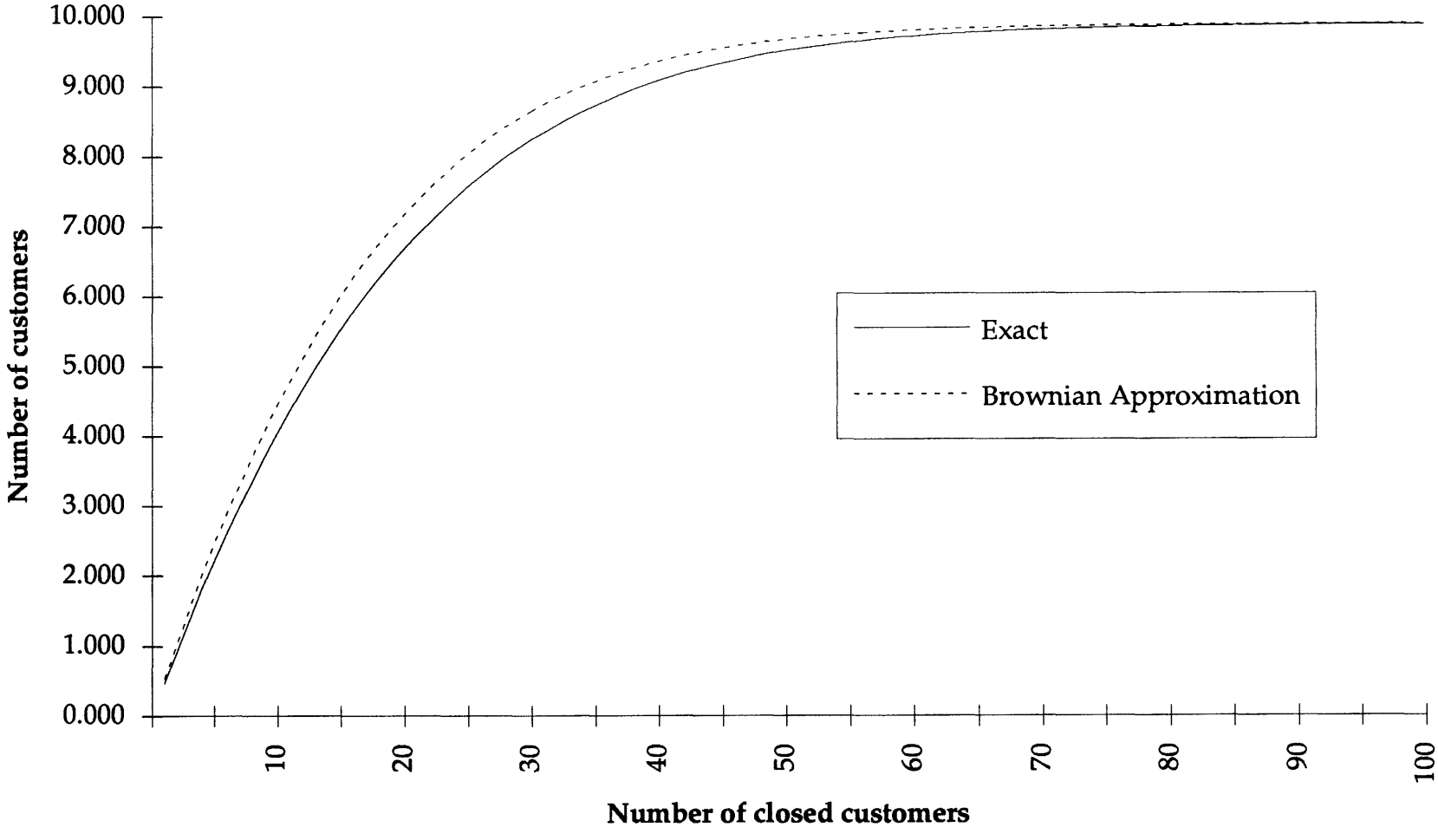

Figure 8: Queue length approximations for closed customers in System 2

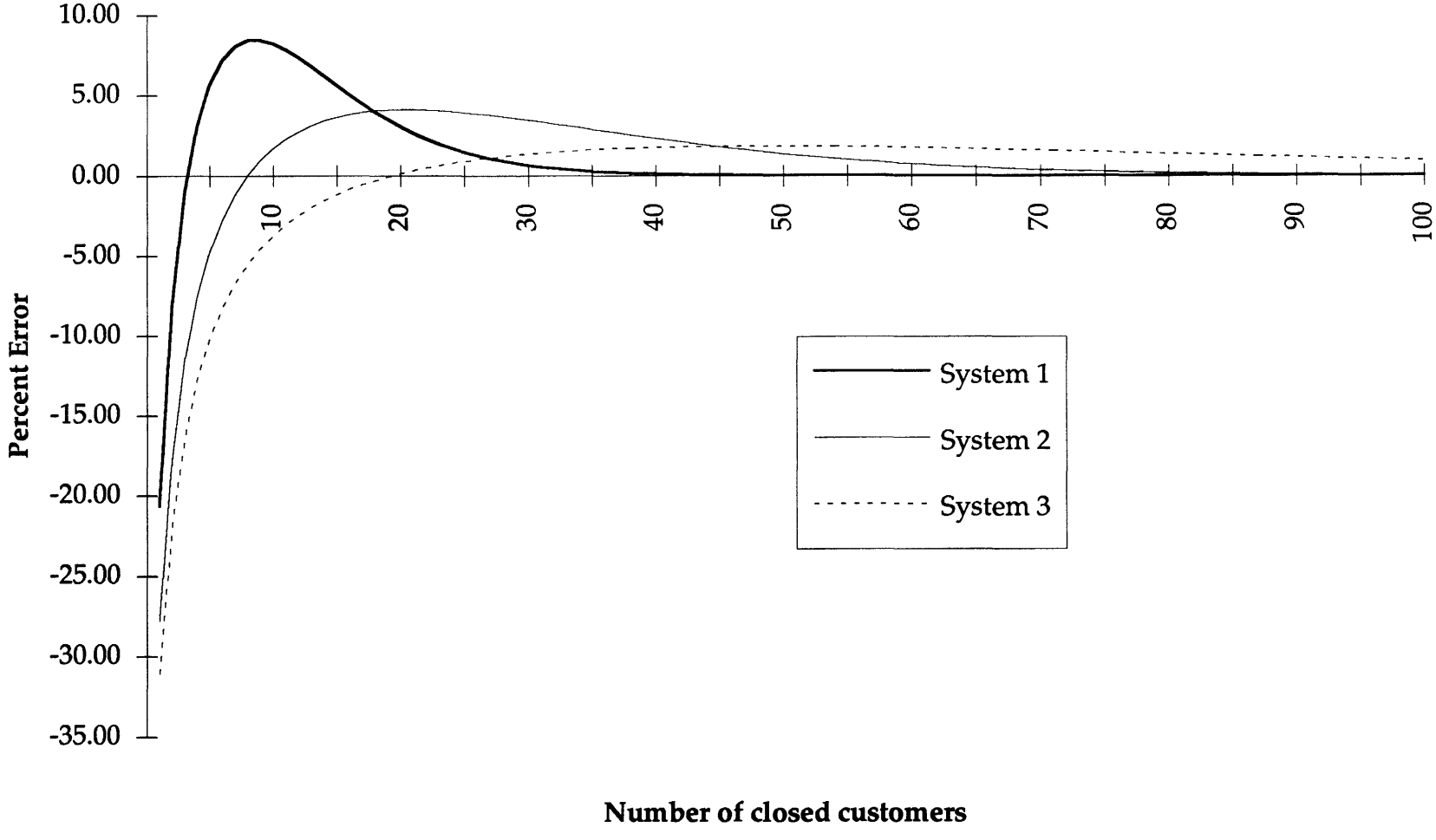

Figure 9: Relative percent error in queue length approximations for open customers

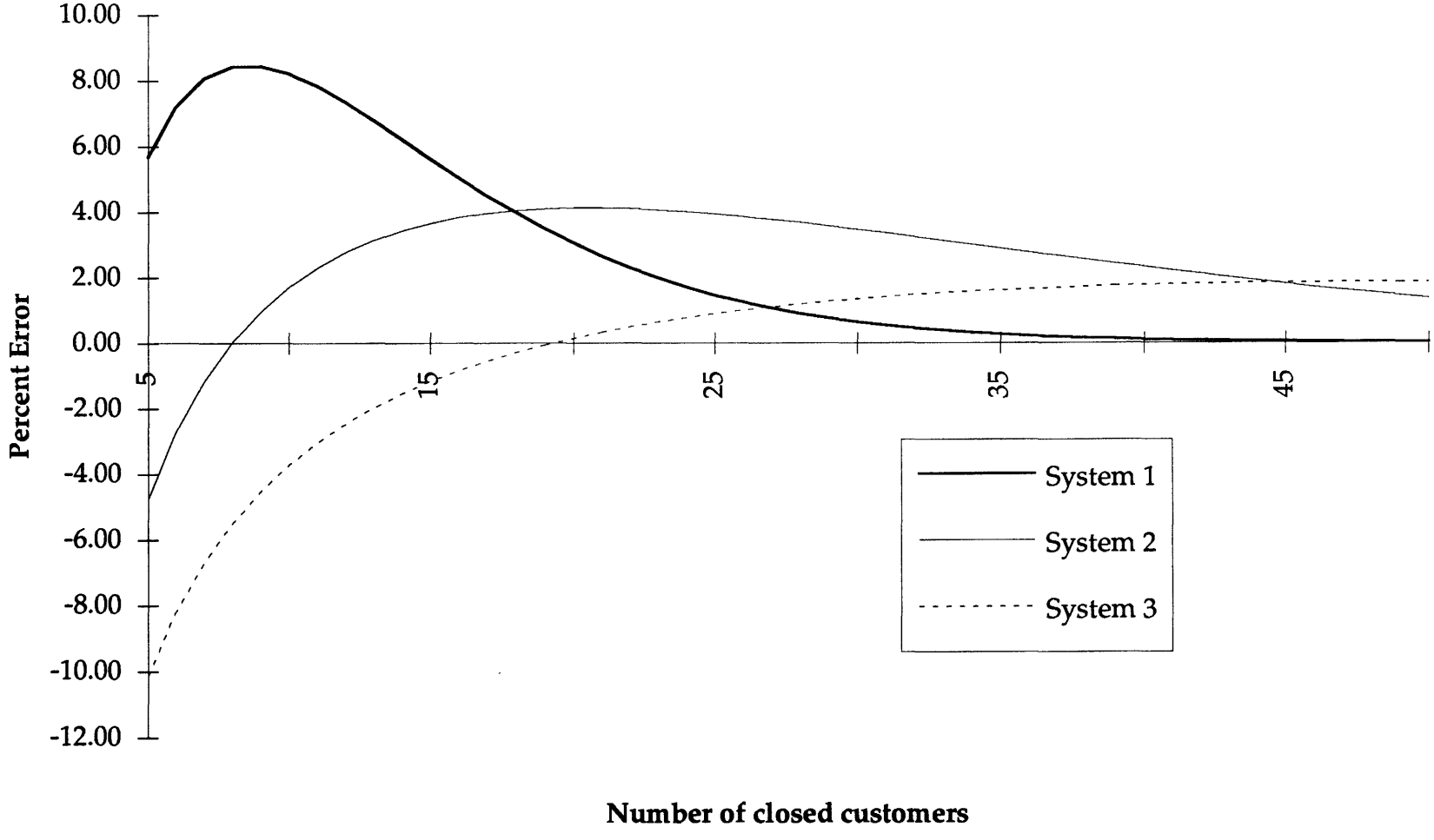

Figure 10: A closer look at relative percent error of queue length approximations for open customers

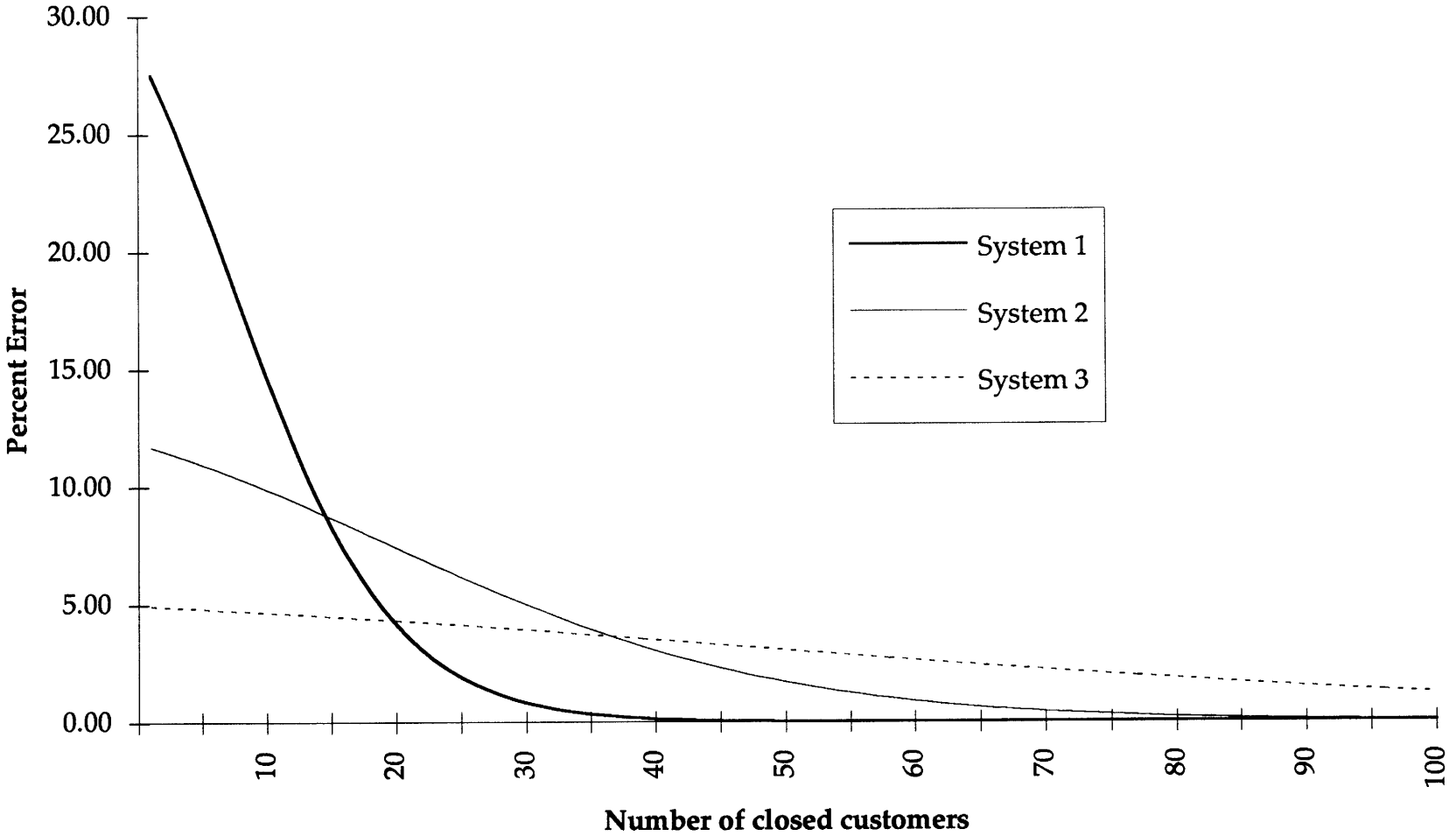

Figure 11: Relative percent error in queue length approximations for closed customers

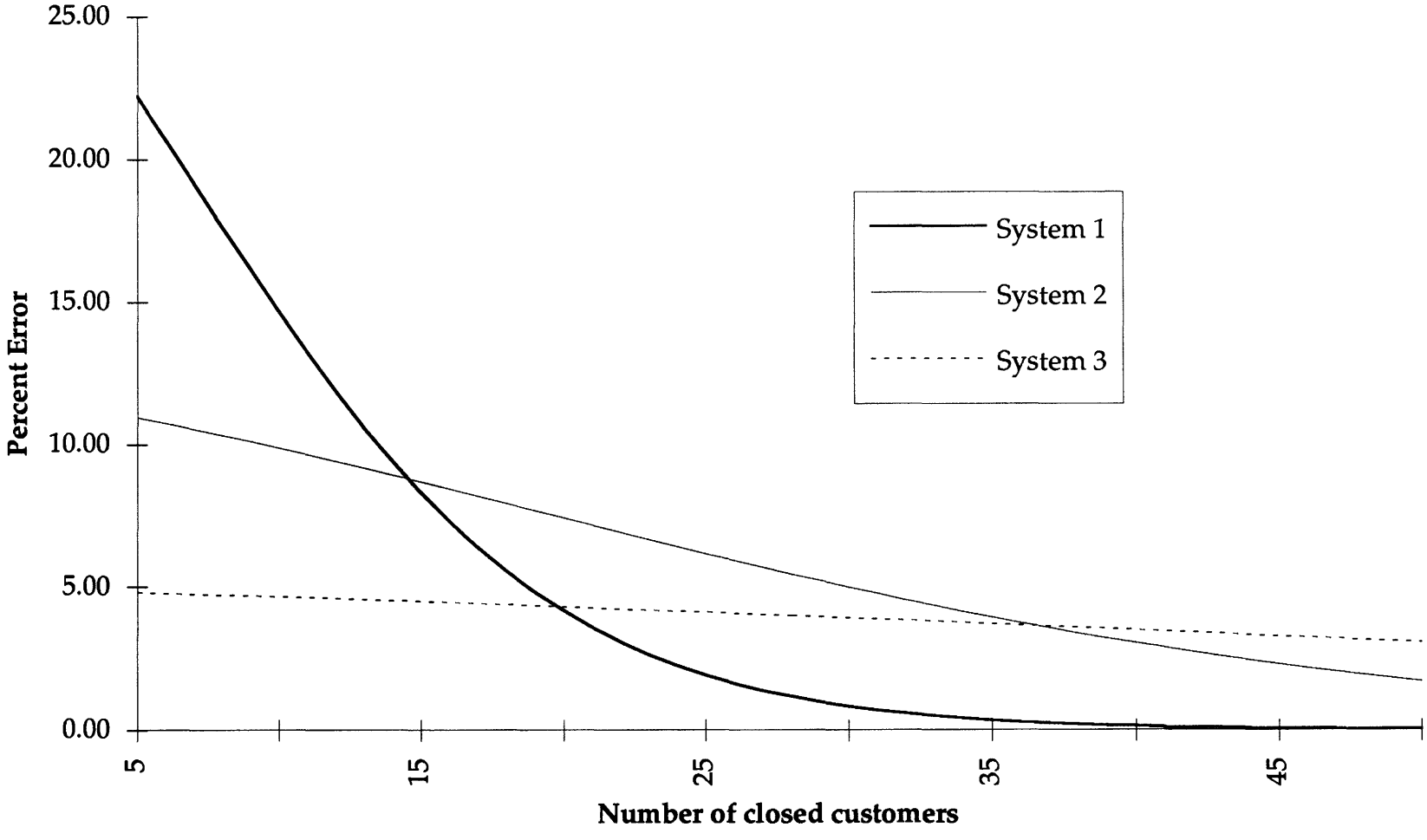

Figure 12: A closer look at relative percent error of queue length approximations for closed customers

References

[1] Baskett, F., Chandy, K. M., Muntz, R. R., and Palacios, F. G. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM* **22**, 248–260 (1975).

[2] Billingsley, P. *Convergence of Probability Measures.* Wiley, New York, 1968.

[3] Chen, H. and Mandelbaum, A. Discrete flow networks: Bottleneck analysis and fluid approximations. *Mathematics of Operations Research* **16**, 408–446 (1991).

[4] Chen, H. and Mandelbaum, A. Stochastic discrete flow networks: Diffusion approximation and bottlenecks. *Annals of Probability* **19**, 1463–1519 (1991).

[5] Chen, H. and Mandelbaum, A. Leontief systems, RBV's and RBM's. *Proceedings of the Imperial College Workshop on Aplied Stochastic Processes,* M. H. A. Davis and R. J. Elliott (eds.), Gordon and Breach Science Publishers (forthcoming).

[6] Dai, J. G. and Harrison J. M. The QNET method for two-moment analysis of closed manufacturing systems. Submitted for publication (1992).

[7] Dai, J. G. and Kurtz T. G. A new proof of the heavy traffic limit theorem for a multi-type one station queue with feedback. Preprint (1992).

[8] Dai, J. G. and Nguyen V. On the convergence of multiclass queueing networks in heavy traffic. Submitted for publication (1992).

[9] Dai, J. G. and Wang Y. Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Systems: Theory and Applications*, to appear (1992).

[10] Harrison, J. M. *Brownian motion and stochastic flow systems.* Wiley, New York, 1985.

[11] Harrison, J. M. Brownian models of queueing networks with heterogeneous customer populations. *Proceedings of the IMA Workshop on Stochastic Differential Systems,* W. Fleming and P. L. Lions (eds.), IMA Volume **10**, Springer-Verlag (1988).

[12] Harrison, J. M. and Nguyen, V. The QNET method for two-moment analysis of open queueing networks. *Queueing Systems: Theory and Applications* **6**, 1–32 (1990).

[13] Harrison, J. M. and Nguyen, V. Brownian models of multiclass queueing networks: Current status and open problems. *Queueing Systems: Theory and Applications* (forthcoming).

[14] Harrison, J. M. and Reiman, M. I. Reflected Brownian motion on an orthant. *Annals of Probability* **9**, 302–308 (1981).

[15] Iglehart, D. L. and Whitt W. The equivalence of functional central limit theorems for counting processes and associated partial sums. Technical Report No. 121, Stanford University, (1969).

[16] Iglehart, D. L. and Whitt W. Multiple channel queues in heavy traffic I. *Advances in Applied Probability* **2**, 150–177 (1970).

[17] Kelly, F. P. *Reversibility and Stochastic Networks*. Wiley, New York, 1979.

[18] Peterson, W. P. A heavy traffic limit theorem for networks of queues with multiple customer types. *Mathematics of Operations Research* **16**, 90–118 (1991).

[19] Reiman, M. I. Open queueing networks in heavy traffic. *Mathematics of Operations Research* **9**, 441–458 (1984).

[20] Reiman, M. I. A multiclass feedback queue in heavy traffic. *Advances in Applied Probability* **20**, 179–207 (1988).

[21] Taylor, L. M. and Williams R. J. Existence and uniqueness of semimartingale reflecting Brownian motion in an orthant. Submitted (1992).

[22] Whitt, W. Some useful functions for functional limit theorems. *Mathematics of Operations Research*, **5** 67–85 (1980).

[23] Whitt, W. Large fluctuations in a deterministic multiclass network of queues. *Management Science*, to appear.