

Stationary tail probabilities in exponential server tandems with renewal arrivals

A. Ganesh^a and V. Anantharam^{b,*}

^a*Department of Computer Science, University of Edinburgh,
Edinburgh EH9 3JZ, UK*

E-mail: ajg@dcs.ed.ac.uk

^b*EECS Department, University of California, Berkeley, CA 94720, USA*

E-mail: ananth@eecs.berkeley.edu

Received 6 April 1995; revised 9 August 1995

The problem considered is that of estimating the tail stationary probability for two exponential server queues in series fed by renewal arrivals. We compute the tail of the marginal queue length distribution at the second queue. The marginal at the first queue is known by the classical result for the $GI/M/1$ queue. The approach involves deriving necessary and sufficient conditions on the paths of the arrival and virtual service processes in order to get a large queue size at the second queue. We then use large deviations estimates of the probabilities of these paths, and solve a constrained convex optimization problem to find the most likely path leading to a large queue size. We find that the stationary queue length distribution at the second queue has an exponentially decaying tail, and obtain the exact rate of decay.

Keywords: Tandem queues, large deviations, rare events.

1. Introduction

Networks of queues are widely used to model communications systems. Bit streams representing either voice, video or data, arrive at the nodes of the network, and need to be transmitted to other nodes, either directly or via intermediate nodes. The bit streams are usually broken up into cells or packets for transmission. The arrival process is modeled as a stochastic process; the packet sizes may also be random. Packets need to queue at the nodes to await transmission, because link capacities are finite. Questions of interest in the design of communication networks relate to the probability of packet loss due to inadequate storage capacity at a node of the network, or the probability of large queueing delays within the network, which may be unacceptable in applications involving voice or video transmission. Estimating these probabilities requires knowledge of the tail stationary distribution in the associated queueing network model. Such distributions are known at present only for a few special types of networks.

There has recently been considerable interest in the use of large deviations techniques to estimate tail stationary probabilities. The problem of fast simulation

* Research supported in part by NSF grant NCR 88-57731 and the AT & T Foundation.

of tail probabilities in $M/M/1$ tandems was studied by Parekh and Walrand [22]. The model is of exponential servers in series with Poisson arrivals. The system is started empty, and the problem considered is that of estimating the probability that the total number in system exceeds a large number, N , before the system is again empty after having had customers. Large deviations ideas were used to relate the problem heuristically to a variational problem involving the most likely path to the build-up of large queue sizes. This variational problem was solved for a system of two queues in tandem. The solution was extended to an arbitrary number of tandem queues by Frater et al. [14]. An alternative approach using the time reversal of Markov chains was followed by Anantharam et al. [3] to find the most likely path to large queue size. The heuristic derived in [22] was made rigorous by Tsoucas [23] by relating the path of the queue length process to that of the 'free' process which is obtained as the difference of the arrival and service processes at each queue. The connection was made explicit for two queues in series, and the general idea sketched for an arbitrary number of queues in series, though the detailed analysis becomes cumbersome as the number of queues increases. More recently, the problem of estimating the tail distribution in a single queue with several multiplexed traffic streams at the input has been studied by several authors, see for example [10, 13, 17]. Solutions were obtained for fairly general arrival and service distributions. The estimation of tail probabilities inintree networks using fast simulation was studied by Chang et al. [9].

In this paper, we consider the problem of estimating the tail of the stationary distribution in a tandem of exponential server queues with renewal arrivals. We suspect that some of the techniques developed here will be relevant in extending the result to more general network models. We use ideas from large deviations theory to estimate the most likely path along which the system builds up to a large queue size. More precisely, we use large deviations results to estimate the probability of a path leading to large queue sizes, and solve a constrained optimization problem to find the most likely such path. We find that the stationary queue length distributions decay exponentially, and obtain the exact exponential rate of decay. For an illustration of the key ideas in the simpler case of a single exponential-server queue with renewal arrivals, see [16].

One approach to estimating the tail of the queue length distribution at the second queue would be to characterize the large deviations behaviour of the output process of the first queue and use the known results for a single queue (see [10], [13] or [17]). This approach has been followed successfully by a number of authors, see [5, 7, 8, 11, 21]. Chang [8] estimates the tail stationary probability inintree networks of deterministic service nodes, for quite general arrival processes. O'Connell [21] characterizes the large deviations behaviour of the departures from a queue with fairly general arrival and service processes and multiple customer classes. Independently of our work, Bertsimas et al. [5] have recently extended Chang's result to acyclic networks of queues with general independent service times and quite general arrival processes. The techniques used in [5] are quite different from ours, and while

the setting of [5] is more general, the asymptotics of the tail probability are characterized more sharply in our work. We have verified that the result obtained here agrees with the specialization of the result of Bertsimas et al. to our model.

2. Problem formulation and result

The model we consider consists of two queues in series. Customers arrive into the first queue according to a renewal process, move to the second queue after receiving service, and leave the system after being served at the second queue. There is a single server at each queue, and the service discipline at each queue is first-come-first-served (FCFS). Let $\tau_i, i = \dots, -1, 0, 1, \dots$, denote a sequence of arrival epochs with $\tau_0 = 0$, and define $T_i = \tau_i - \tau_{i-1}$ to be the corresponding sequence of inter-arrival times. Then, the T_i are independent, identically distributed (*i.i.d.*) random variables. Let T denote a random variable with their common distribution, denoted F . The service times of the customers at the two queues are *i.i.d.* exponential random variables, with mean $1/\mu_1$ at the first queue and $1/\mu_2$ at the second queue. The arrival process and the service processes at the two queues are mutually independent.

Let $X(t) = (X^1(t), X^2(t)), t \in \mathbb{R}$ denote the queue length process. Here $X^1(t)$ and $X^2(t)$ denote the queue lengths in the first and second queues respectively at time t . We are interested in the queue length process seen by arrivals. Denote $X_n = X(\tau_n)$ to be the queue lengths seen by the n th arrival, $X_n = (X_n^1, X_n^2)$. Then, X_n is a Markov chain because of the memoryless property of the service time distributions. We construct the process $X(t)$ as described below.

Let $V(t) = (V^1(t), V^2(t)), t \in \mathbb{R}$ denote a pair of Poisson processes with rates μ_1, μ_2 , defined on the same sample space as the arrival process. These are taken to represent the virtual service processes at the first and second queue respectively. It is assumed that the arrival process, the virtual service process at the first queue, and the virtual service process at the second queue are mutually independent. A virtual service at either queue results in an actual service if the corresponding queue is non-empty, and is wasted otherwise. Such a construction of the service process is possible because of the memoryless nature of the exponential distribution. We define $V_n = V(\tau_{n+1}) - V(\tau_n)$ to be the number of virtual services at the two queues between the n^{th} and $(n + 1)^{\text{th}}$ arrival epochs. Note that

$$a_{jk} \triangleq P(V_0^1 = j, V_0^2 = k) = \int_0^\infty \frac{(\mu_1 t)^j}{j!} \frac{(\mu_2 t)^k}{k!} e^{-(\mu_1 + \mu_2)t} dF(t). \tag{1}$$

We assume that $\mu_1 ET > 1$ and $\mu_2 ET > 1$, that is, the service rate at each queue is faster than the arrival rate. This is the stability criterion, and is needed to ensure that the queue sizes do not blow up. Under this assumption, it is known

(see [18]) that the waiting time process for each arriving customer, and hence the queue length process, can be constructed from the above description; the queue lengths and waiting times are finite a.s., and the queue length and waiting time processes are stationary and ergodic. We make the additional assumption that the paths of the queue length process are continuous from the left. That suffices to completely and uniquely specify the queue length processes (up to sets of measure zero).

We wish to compute the stationary distribution of the queue length process seen by arrivals. We shall confine ourselves to computing the marginals of this distribution. That is, with the system started at $-\infty$, we want to compute $P(X_n^1 = j)$ and $P(X_n^2 = k)$, for all j, k , and an arbitrary n , say $n = 0$.

Define F^* to be the Laplace transform of the inter-arrival distribution.

$$F^*(s) = E[e^{-sT}] = \int_0^\infty e^{-st} dF(t).$$

Since T is a positive random variable, there is $-\infty \leq \sigma \leq 0$ such that $F^*(s)$ is finite on (σ, ∞) , and infinite on $(-\infty, \sigma)$ (if $-\infty < \sigma$). $F^*(\sigma)$ could be either finite or infinite. The following properties of F^* are well-known, see for example Lemma 2.2.5, p. 27, in [12]. $F^*(s)$ is a convex, decreasing function of s , and in fact strictly convex and strictly decreasing if T is not identically zero, as we assume. F^* has derivatives of all orders on (σ, ∞) . Finally, $\log F^*(s)$ is also a convex function of s .

Consider the pair of equations

$$x = F^*(\mu_i - \mu_i x), \quad i = 1, 2. \tag{2}$$

We show that each equation has two solutions, ν_i and 1, and that $\nu_i < 1$. Clearly, $x = 1$ is a solution for each i . Furthermore,

$$\left. \frac{d}{dx} F^*(\mu_i - \mu_i x) \right|_{x=1} = -\mu_i \left. \frac{d}{ds} F^*(s) \right|_{s=0} = \mu_i ET > 1, \quad i = 1, 2,$$

where the inequality follows from the stability criterion. That is, the slope of the curve $y = F^*(\mu_i - \mu_i x)$ exceeds that of the line $y = x$, at $x = 1$. At $x = 0$, we have $F^*(\mu_i - \mu_i x) > 0 = x$. But $F^*(\mu_i - \mu_i x)$ is a continuous function of $x \in (0, 1)$. Therefore, (2) has a solution $\nu_i \in (0, 1)$, for $i = 1, 2$. Further, by the strict convexity of $F^*(s)$, (2) has no solutions other than ν_i and 1. We also note the following properties of the solutions.

LEMMA 1

$$\left. \frac{d}{dx} F^*(\mu_i - \mu_i x) \right|_{x=\nu_i} < 1.$$

Proof

Recall that $F^*(s)$ is a strictly convex function of s , and analytic on (σ, ∞) , for some $\sigma \leq 0$. Also, $F^*(\mu_i - \mu_i x) - x$ takes the value zero at $x = \nu_i$ and at $x = 1$. From this, we have

$$\frac{d}{dx} F^*(\mu_i - \mu_i x) = 1 \quad \text{for some } x \in (\nu_i, 1).$$

Hence, by the strict convexity of $F^*(\mu_i - \mu_i x)$,

$$\left. \frac{d}{dx} F^*(\mu_i - \mu_i x) \right|_{x=\nu_i} < 1. \quad \square$$

LEMMA 2

Let ν_1, ν_2 be defined as above. If $\mu_1 > \mu_2$, then $\nu_1 < \nu_2$, and vice versa.

Proof

We saw above that each equation in (2) has ν_i and 1 as its only solutions, with $\nu_i \in (0, 1)$. Also, $F^*(\mu_i - \mu_i x)$ is a strictly increasing function of x , and $F^*(\mu_i - \mu_i \cdot 0) > 0$. Thus, by continuity of F^* ,

$$\nu_i = \min\{x > 0 : F^*(\mu_i - \mu_i x) \leq x\}. \quad (3)$$

Now suppose $\mu_1 > \mu_2$. Then, for any $x \in (0, 1)$, $\mu_1(1 - x) > \mu_2(1 - x)$ and so $F^*(\mu_1 - \mu_1 x) < F^*(\mu_2 - \mu_2 x)$. Therefore, it is clear from (3) that $\nu_1 < \nu_2$. The argument when $\mu_1 < \mu_2$ follows along parallel lines. \square

The stationary distribution of an embedded GI/M/1 queue is known explicitly (see [4]); thus

$$P(X_0^1 \geq N) = \nu_1^N.$$

In this paper, we compute the asymptotics of $P(X_0^2 \geq N)$. Our result is summarized in the theorem below.

THEOREM 3

The stationary distribution of the queue length of the second queue, as seen by arrivals, satisfies

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P(X_0^2 \geq N) = \beta,$$

where β is defined as follows:

1. If $\mu_1 \geq \mu_2$ and $\frac{\mu_1}{\mu_2 \nu_2} \frac{d}{dx} F^*(\mu_2 - \mu_2 x)|_{x=\nu_2} > 1$, then $\beta = \log \nu_2$.
2. If $\mu_1 \leq \mu_2$ and $\frac{\mu_2}{\mu_1 \nu_1} \frac{d}{dx} F^*(\mu_1 - \mu_1 x)|_{x=\nu_1} > 1$, then $\beta = \log \frac{\mu_1 \nu_1}{\mu_2}$.
3. If the conditions in neither (1) nor (2) are satisfied, then $\beta = \log v_2$, where (v_1, v_2) is any solution of the pair of equations

$$F^*(\mu_1 - \mu_1 x_1 + \mu_2 - \mu_2 x_2) = x_1 x_2, \quad (4)$$

$$\frac{d}{dx} F^*(\mu_1 - \mu_1 x + \mu_2 - \mu_2 x_2) \Big|_{x=x_1} = x_2, \quad (5)$$

which satisfies the requirements

$$v_1 \geq 1, \quad v_1 \leq v_1 v_2 < 1, \quad \frac{d}{dx} F^*(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 x) \Big|_{x=v_2} < v_1. \quad (6)$$

The existence of a solution of eqs. (4) and (5) satisfying (6) is proved in Appendix A; see also Lemma 11 in section 4.

The next four sections deal with the proof of the above theorem. The proof consists of deriving upper and lower bounds on the logarithm of the desired probability as N goes to ∞ . The problem of estimating this probability is reduced, in section 3, to a constrained optimization problem which is solved in section 4. The solution is used to derive the upper bound in section 5 and the lower bound in section 6.

It is usually not possible to get closed form solutions for β . However it can be computed numerically from the description in Theorem 3, see [15] for details. If the arrival process is Poisson with rate λ which is less than μ_1 and μ_2 , then it is a classical result, (see [4] for example), that the stationary queue length distribution is given by

$$P(X_0^1 \geq N_1, X_0^2 \geq N_2) = \left(\frac{\lambda}{\mu_1}\right)^{N_1} \left(\frac{\lambda}{\mu_2}\right)^{N_2}, \quad (7)$$

and so

$$P(X_0^2 \geq N) = \left(\frac{\lambda}{\mu_2}\right)^N. \quad (8)$$

We show below that the result obtained using Theorem 3 is consistent with (8).

Suppose the arrival process is Poisson of rate λ . That is, the inter-arrival times are *i.i.d.*, exponentially distributed with mean $1/\lambda$. The Laplace transform of the inter-arrival distribution is

$$F^*(s) = \int_0^\infty \lambda e^{-\lambda t} e^{-st} dt = \frac{\lambda}{\lambda + s}.$$

The stability condition is that $\mu_1/\lambda > 1$ and $\mu_2/\lambda > 1$. We assume that λ satisfies these requirements. Solving the fixed point equations in (2) and taking the solution other than 1, we obtain

$$\nu_1 = \frac{\lambda}{\mu_1}, \quad \nu_2 = \frac{\lambda}{\mu_2}. \tag{9}$$

We shall consider separately the cases when μ_1 is bigger than, less than, and equal to μ_2 . Suppose first that $\mu_1 > \mu_2$. Then

$$\frac{\mu_1}{\mu_2 \nu_2} \frac{d}{dx_2} F^*(\mu_2 - \mu_2 x_2) \Big|_{x_2=\nu_2} = \frac{\mu_1}{\mu_2 \nu_2} \frac{\mu_2 \lambda}{(\lambda + \mu_2 - \mu_2 \nu_2)^2} = \frac{\mu_1}{\mu_2},$$

where the last equality is obtained using (9). Thus, $\mu_1 > \mu_2$ implies that condition 1 of Theorem 3 holds. So, by Theorem 3,

$$P(X_0^2 \geq N) \approx \left(\frac{\lambda}{\mu_2}\right)^N,$$

where the approximate equality means asymptotically exponentially equal. Suppose next that $\mu_1 < \mu_2$. Then,

$$\frac{\mu_2}{\mu_1 \nu_1} \frac{d}{dx_1} F^*(\mu_1 - \mu_1 x_1) \Big|_{x_1=\nu_1} = \frac{\mu_2}{\mu_1 \nu_1} \frac{\mu_1 \lambda}{(\lambda + \mu_1 - \mu_1 \nu_1)^2} = \frac{\mu_2}{\mu_1},$$

where the last equality is due to (9). Thus, $\mu_2 > \mu_1$ implies that condition 2 of Theorem 3 holds. So, by Theorem 3,

$$P(X_0^2 \geq N) \approx \left(\frac{\mu_1 \lambda}{\mu_2 \mu_1}\right)^N = \left(\frac{\lambda}{\mu_2}\right)^N.$$

Finally, suppose $\mu_1 = \mu_2$, so that $\nu_1 = \nu_2$. Then,

$$\frac{\mu_1}{\mu_2 \nu_2} \frac{d}{dx_2} F^*(\mu_2 - \mu_2 x_2) \Big|_{x_2=\nu_2} = \frac{1}{\nu_2} \frac{\mu_2 \lambda}{(\lambda + \mu_2 - \mu_2 \nu_2)^2} = 1,$$

where the last equality follows from (9). Clearly, the same equality holds if we interchange 1 and 2 in the subscripts. Therefore, neither 1 nor 2 holds in Theorem 3. Hence, by condition 3 of Theorem 3, we need to solve

$$\begin{aligned}\lambda(\lambda + \mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2)^{-1} &= v_1 v_2, \\ \lambda \mu_1 (\lambda + \mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2)^{-2} &= v_2,\end{aligned}$$

subject to

$$v_1 \geq 1, \quad v_1 \leq v_1 v_2 < 1, \quad \lambda \mu_2 (\lambda + \mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2)^{-2} < v_1.$$

It is easy to verify that $(1, v_2)$ is a solution, keeping in mind that $\mu_1 = \mu_2$, and using (9) and the stability condition. So, by Theorem 3,

$$P(X_0^2 \geq N) \approx \left(\frac{\lambda}{\mu_2}\right)^N.$$

Thus, we see that in all cases, the approximate result obtained using Theorem 3 is consistent with the exact result obtained by traditional methods.

3. Reduction to an optimization problem

Consider a tandem of two queues with exponential servers and renewal arrivals as described in the last section. We are interested in estimating $P(X_0^2 \geq N)$, where we use the notation $X_n = X(\tau_n)$. The above probability is the same as the stationary probability at the arrival instants. We describe below certain necessary conditions for the event $\{X_0^2 \geq N\}$ and use an estimate of their probability to obtain an upper bound on $P(X_0^2 \geq N)$.

It was shown in [18] that, under the stability assumption, each queue empties infinitely often, with probability one. Thus, on the path leading to $X^2(\tau_0) \geq N$, there is a last time $t \leq \tau_0$ at which the second queue is empty, and $t > -\infty$ almost surely. More precisely, define

$$t = \inf\{s : X^2(u) > 0 \forall u \in (s, \tau_0]\}.$$

It follows that every virtual service at the second queue in $[t, \tau_0)$ results in an actual service. Also, by left continuity of the paths of $X(t)$, we have $X^2(t) = 0$. Define

$$n = \sup\{k : t \leq \tau_{-k}\}.$$

We now have the following upper bounds for the queue length of the second queue

at time τ_0 :

$$\begin{aligned}
 X^2(\tau_0) &\leq V^1([t, \tau_{-n})) - V^2([t, \tau_{-n})) + \sum_{i=-n}^{-1} (V_i^1 - V_i^2), \\
 X^2(\tau_0) &\leq X^1(\tau_{-n-1}) + 1 - V^2([t, \tau_{-n})) + \sum_{i=-n}^{-1} (V_i^1 - V_i^2), \\
 X^2(\tau_0) &\leq X^1(\tau_{-n-1}) + 1 - V^2([t, \tau_{-n})) + \sum_{i=-n}^{-1} (1 - V_i^2).
 \end{aligned}$$

The first equation uses the fact that the total number of arrivals into the second queue during $[t, \tau_0)$ is bounded above by the number of virtual services at the first queue during this period. The second equation uses the total number in the first queue at time τ_{-n-1} , plus the arrival at τ_{-n-1} , as an upper bound on the number of arrivals into the second queue during $[t, \tau_{-n})$; the corresponding upper bound for the interval $[\tau_{-n}, \tau_0)$ is provided by the number of virtual services at the first queue during this period. In the third equation, the number of arrivals into the second queue in $[t, \tau_0)$ is bounded above by the number at the first queue at time τ_{-n-1} plus the number of arrivals into the first queue in $[\tau_{-n-1}, \tau_0)$. All three equations use the fact that $X^2(t)$ is zero, and that the number of actual services at the second queue during $[t, \tau_0)$ is the same as $V^2([t, \tau_0))$, the number of virtual services.

Hence, for a given number N , a necessary condition for $X_0^2 \geq N$ is that there exist an $n \geq 0$, a time $t \in [\tau_{-n-1}, \tau_{-n})$, and $K, L, M \geq 0$ such that

$$V^1([t, \tau_{-n})) = K, \quad V^2([t, \tau_{-n})) = L, \tag{10}$$

$$X^1(\tau_{-n-1}) + 1 = M, \tag{11}$$

$$\begin{aligned}
 K - L + \sum_{i=-n}^{-1} (V_i^1 - V_i^2) &\geq N, \\
 M - L + \sum_{i=-n}^{-1} (V_i^1 - V_i^2) &\geq N, \\
 M - L + \sum_{i=-n}^{-1} (1 - V_i^2) &\geq N.
 \end{aligned} \tag{12}$$

Observe that, for fixed n, K, L, M , the events in (10), (11), (12) are independent. The reason is that the event in (10) depends only on the virtual service processes in the interval $[\tau_{-n-1}, \tau_{-n})$, the event in (11) depends only on the virtual service processes

before time τ_{-n-1} , and the event in (12) depends only on the virtual service processes after time τ_{-n} ; their independence then follows from the Poisson nature of the virtual service processes.

An upper bound on $P(X_0^2 \geq N)$ is given by the probability of the union over all $(n, K, L, M) \geq 0$ of the intersection of the events in (10), (11) and (12). By the independence of said events, the probability of the intersection is the product of the individual probabilities. The probability of the union may be bounded above by the sum of the corresponding probabilities. We thus get

$$\begin{aligned}
 &P(X_0^2 \geq N) \\
 &\leq \sum_{n,K,L,M=0}^{\infty} P\left(\exists t \in [\tau_{-n-1}, \tau_{-n}) : V^1([t, \tau_{-n})) = K, V^2([t, \tau_{-n})) = L\right) \\
 &\quad \cdot P(X^1(\tau_{-n-1}) + 1 = M) \\
 &\quad \cdot P\left(\begin{array}{l} K - L + \sum_{i=-n}^{-1} (V_i^1 - V_i^2) \geq N \\ M - L + \sum_{i=-n}^{-1} (V_i^1 - V_i^2) \geq N \\ M - L + \sum_{i=-n}^{-1} (1 - V_i^2) \geq N \end{array}\right). \tag{13}
 \end{aligned}$$

In order to estimate the above probabilities, we need the following result.

DEFINITION 4

The Kullback-Leibler distance, or relative entropy, $D(\mathbf{q};\mathbf{p})$, between two probability distributions \mathbf{q} and \mathbf{p} on a countable set I is defined as

$$D(\mathbf{q};\mathbf{p}) = \sum_{i \in I} q_i \log \frac{q_i}{p_i},$$

where the logarithm is to base e , and $0 \log 0$ and $0 \log (0/0)$ are defined to be 0. If I has cardinality two, then the distributions are completely specified by giving the probability of one of the points, and we shall make use of the notation

$$D_2(\beta; \alpha) = \beta \log \frac{\beta}{\alpha} + (1 - \beta) \log \frac{1 - \beta}{1 - \alpha}.$$

We also note the following properties of the Kullback-Leibler distance. $D(\mathbf{q};\mathbf{p})$ is well-defined and non-negative, though it could possibly be $+\infty$. It is zero if and only if $\mathbf{q} \equiv \mathbf{p}$. $D(\mathbf{q};\mathbf{p})$ is convex in each of its arguments. For a proof of these properties and the theorem below, see [6].

THEOREM 5 (Sanov)

Let X_1, \dots, X_n be *i.i.d.* random variables with common distribution \mathbf{p} , taking values in a countable set \mathcal{X} . Denote the type (or empirical distribution) of X_1, \dots, X_n

to be the probability distribution on \mathcal{X} defined as

$$q(x) = \frac{1}{n} \sum_{i=1}^n 1\{X_i = x\}.$$

Let $P_n(\mathcal{A})$ denote the probability that X_1, \dots, X_n has its type lying in a convex set \mathcal{A} of probability distributions on \mathcal{X} . Then

$$P_n(\mathcal{A}) \leq \exp\left(-n \inf_{\mathbf{q} \in \mathcal{A}} D(\mathbf{q}; \mathbf{p})\right).$$

We now estimate each of the terms in the product in (13). The first term is bounded above by the probability that there exists a time $t < \tau_{-n}$ such that $V^1([t, \tau_{-n})) = K$ and $V^2([t, \tau_{-n})) = L$. Since V^1, V^2 are Poisson processes, this is the same as the probability of getting K heads and L tails in $K + L$ tosses of a biased coin with $P(\text{head}) = \mu_1 / (\mu_1 + \mu_2)$. Thus, by Sanov’s theorem,

$$\begin{aligned} &P\left(\exists t \in [\tau_{-n-1}, \tau_{-n}) : V^1([t, \tau_{-n})) = K, V^2([t, \tau_{-n})) = L\right) \\ &\leq (K + L)D_2\left(\frac{K}{K + L}; \frac{\mu_1}{\mu_1 + \mu_2}\right). \end{aligned} \tag{14}$$

The second term is given by the known stationary distribution for a single $GI/M/1$ queue, since, for fixed n , τ_{-n-1} is just an arbitrary arrival time and the system is in stationarity. Thus,

$$P(X^1(\tau_{-n-1}) + 1 = M) = (1 - \nu_1)\nu_1^{M-1} 1\{M \geq 1\}. \tag{15}$$

We now estimate the last term. Fix n, K, L, M non-negative integers and let \mathbf{p} denote a probability distribution on \mathbb{Z}_+^2 . Define

$$\begin{aligned} g_1(\mathbf{p}, n, K, L, M) &= N + L - K - n \sum_{j,k=0}^{\infty} (j - k)p_{jk}, \\ g_2(\mathbf{p}, n, K, L, M) &= N + L - M - n \sum_{j,k=0}^{\infty} (j - k)p_{jk}, \\ g_3(\mathbf{p}, n, K, L, M) &= N + L - M - n \sum_{j,k=0}^{\infty} (1 - k)p_{jk}, \end{aligned} \tag{16}$$

if $\sum_{j,k=0}^{\infty} jp_{jk}$ and $\sum_{j,k=0}^{\infty} kp_{jk}$ are finite. Otherwise, define

$$g_1(\cdot) = g_2(\cdot) = g_3(\cdot) = +\infty.$$

Let \mathcal{A} denote the set of probability distributions \mathbf{p} on \mathbb{Z}_+^2 defined as

$$\mathcal{A} = \{\mathbf{p} : g_i(\mathbf{p}, n, K, L, M) \leq 0, i = 1, 2, 3\}. \tag{17}$$

It is clear that $(V_i, -n \leq i \leq -1)$ satisfies the constraints in (13) if and only if its type lies in the set \mathcal{A} defined above (the finiteness condition on $\sum j p_{jk}$ and $\sum k p_{jk}$ are automatically satisfied by the empirical distributions since these involve only a finite number of the V_i). It is also easy to see that \mathcal{A} is a convex set. Hence, by Sanov’s theorem, the probability of the last term in the product in (13) is bounded above by $\exp(-n \inf_{\mathbf{p} \in \mathcal{A}} D(\mathbf{p}; \mathbf{a}))$, where $\mathbf{a} = \{a_{jk}\}$ is given by (1).

Now, using the estimates computed above, we can rewrite (13) as

$$P(X_0^2 \geq N) \leq \sum_{n, K, L, M=0}^{\infty} \exp[-\hat{f}(n, K, L, M)], \tag{18}$$

where

$$\begin{aligned} \hat{f}(n, K, L, M) &= (K + L)D_2\left(\frac{K}{K + L}; \frac{\mu_1}{\mu_1 + \mu_2}\right) - \log\left(\frac{1 - \nu_1}{\nu_1}\right) \\ &\quad - M \log \nu_1 + \chi(M) + n \inf_{\mathbf{p} \in \mathcal{A}} D(\mathbf{p}; \mathbf{a}). \end{aligned} \tag{19}$$

Here $\chi(M)$ denotes the function which takes value $+\infty$ at $M = 0$ and zero at all other M , and the set \mathcal{A} is defined in (17) (note that the definition of \mathcal{A} involves the values of n, K, L , and M , though this dependence has not been made explicit in the notation).

In section 5, we derive an upper bound on the sum in (18) in terms of the maximum value of the summand. We now turn to finding this maximum value, or equivalently, the infimum of $\hat{f}(n, K, L, M)$ over $(n, K, L, M) \in \mathbb{Z}_+^4$. For notational convenience, we shall drop the term $\log(1 - \nu_1/\nu_1)$, remembering that it gives rise to a multiplicative constant in the final estimate. Also, we compute the infimum over $(n, K, L, M) \in \mathbb{R}_+^4$. Clearly, this is no larger than the infimum over \mathbb{Z}_+^4 , and hence provides an upper bound for the maximum value of the summand in (18). Likewise, we drop the term $\chi(M)$, noting that this too does not increase the estimate of the infimum of \hat{f} . We are thus left with the following constrained optimization problem.

$$\inf f(\mathbf{p}, n, K, L, M) = nD(\mathbf{p}; \mathbf{a}) - M \log \nu_1 + (K + L)D_2\left(\frac{K}{K + L}; \frac{\mu_1}{\mu_1 + \mu_2}\right) \tag{20}$$

subject to the constraints

$$g_i(\mathbf{p}, n, K, L, M) \leq 0, \quad i = 1, 2, 3. \tag{21}$$

4. Solution of the optimization problem

A reformulation of the above problem using the change of variables $\mathbf{q} = n\mathbf{p}$ yields a convex optimization problem. This can be solved by introducing a Lagrangian and using standard techniques. The solution procedure is described in detail in [15]. We omit most of the details here for brevity, but merely describe sufficient conditions for a point to achieve the minimum in (20) subject to (21). We then find a point satisfying these conditions.

THEOREM 6

Let $x^0 = (\mathbf{p}^0, n^0, K^0, L^0, M^0)$ have n^0, K^0, L^0 and M^0 non-negative. Then, \mathbf{p}^0 is a probability distribution and x^0 minimizes f in (20) subject to (21) if there exists $y^0 \in \mathbb{R}_+^3$ such that the following conditions are satisfied.

- (C1) $f(x^0) < +\infty$.
- (C2) $g_i(x^0) \leq 0, i = 1, 2, 3$.
- (C3) $\langle y^0, g(x^0) \rangle = 0$, where $\langle \cdot, \cdot \rangle$ denotes the inner product on \mathbb{R}^3 .
- (C4) $p_{jk}^0 = a_{jk} v_1^{j-1} v_2^{k-1} \forall (j, k) \in \mathbb{Z}_+^2$, where $v_1 = e^{y_1^0 + y_2^0}, v_2 = e^{-(y_1^0 + y_2^0 + y_3^0)}$.
- (C5) $F^*(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2) = v_1 v_2$.
- (C6) $\log v_1 + y_2^0 + y_3^0 \leq 0, M^0(\log v_1 + y_2^0 + y_3^0) = 0$.
- (C7) $K^0 = L^0 = 0$ and, for all $K, L > 0$,

$$(K + L)D_2\left(\frac{K}{K + L}; \frac{\mu_1}{\mu_1 + \mu_2}\right) - y_1^0 K + (y_1^0 + y_2^0 + y_3^0)L \geq 0.$$

Proof

Let $x^0 = (\mathbf{p}^0, n^0, K^0, L^0, M^0)$ and $y^0 \in \mathbb{R}_+^3$ be such that the conditions in the theorem are satisfied. Then \mathbf{p}^0 is a probability distribution. Indeed, by (C4) and the fact that $y^0 \in \mathbb{R}_+^3, p_{jk}^0$ is positive for all j, k , while

$$\begin{aligned} \sum_{j,k=0}^{\infty} p_{jk}^0 &= \sum_{j,k=0}^{\infty} a_{jk} v_1^{j-1} v_2^{k-1} \\ &= (v_1 v_2)^{-1} \sum_{j,k=0}^{\infty} \int_0^{\infty} \frac{(\mu_1 v_1 t)^{j-1}}{j!} \frac{(\mu_1 v_1 t)^{k-1}}{k!} e^{-(\mu_1 + \mu_2)t} dF(t) \\ &= (v_1 v_2)^{-1} \int_0^{\infty} \exp[-(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2)t] dF(t) \\ &= (v_1 v_2)^{-1} F^*(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2) = 1, \end{aligned} \tag{22}$$

where the second equality comes from the definition of \mathbf{a} in (1) and the last from (C5).

Let $x = (\mathbf{p}, n, K, L, M)$ have $f(x) < +\infty$ and $g_i(x) \leq 0$, $i = 1, 2, 3$. We shall show that $f(x) \geq f(x^0)$ for any such x . This establishes the theorem, since, if x is such that $f(x) = +\infty$, then $f(x) > f(x^0)$ by (C1). Since $g(x) \leq 0$ and $y^0 \in \mathbb{R}_+^3$, observe from (C3) that $\langle y^0, g(x) - g(x^0) \rangle \leq 0$. Therefore, $f(x) \geq f(x^0)$ if $f(x) + \langle y^0, g(x) \rangle \geq f(x^0) + \langle y^0, g(x^0) \rangle$. But

$$\begin{aligned} & f(x) + \langle y^0, g(x) \rangle - f(x^0) - \langle y^0, g(x^0) \rangle \\ &= n \sum_{j,k=0}^{\infty} p_{jk} \left[\log \frac{p_{jk}}{a_{jk}} - (y_1^0 + y_2^0)(j-1) + (y_1^0 + y_2^0 + y_3^0)(k-1) \right] \\ &\quad - n^0 \sum_{j,k=0}^{\infty} p_{jk}^0 \left[\log \frac{p_{jk}^0}{a_{jk}} - (y_1^0 + y_2^0)(j-1) + (y_1^0 + y_2^0 + y_3^0)(k-1) \right] \\ &\quad + (K+L)D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \\ &\quad - y_1^0 K + (y_1^0 + y_2^0 + y_3^0)L - (M - M^0)(\log \nu_1 + y_2^0 + y_3^0), \end{aligned}$$

where we have used the fact that $K^0 = L^0 = 0$ by (C7). Since K , L and M are non-negative, observe from (C6), (C7) and the above equality that $f(x) + \langle y^0, g(x) \rangle \geq f(x^0) + \langle y^0, g(x^0) \rangle$ if

$$\begin{aligned} & n \sum_{j,k=0}^{\infty} p_{jk} \left[\log \frac{p_{jk}}{a_{jk}} - (y_1^0 + y_2^0)(j-1) + (y_1^0 + y_2^0 + y_3^0)(k-1) \right] \\ &\quad - n^0 \sum_{j,k=0}^{\infty} p_{jk}^0 \left[\log \frac{p_{jk}^0}{a_{jk}} - (y_1^0 + y_2^0)(j-1) + (y_1^0 + y_2^0 + y_3^0)(k-1) \right] \geq 0. \quad (23) \end{aligned}$$

Observe from the definition of \mathbf{p}^0 in (C4) that the left hand side above reduces to

$$n \sum_{j,k=0}^{\infty} p_{jk} \log \frac{p_{jk}}{p_{jk}^0} - n^0 \sum_{j,k=0}^{\infty} p_{jk}^0 \log \frac{p_{jk}^0}{p_{jk}^0} = nD(\mathbf{p}; \mathbf{p}^0).$$

Therefore, by the non-negativity of the Kullback-Leibler distance, the inequality in (23) holds. But this was shown to imply that $f(x) \geq f(x^0)$, which establishes the claim of the theorem. \square

It may be noted that y^0 is the vector of Lagrange multipliers for our constrained optimization problem. Our next task is to find x^0 and y^0 which satisfy the conditions in the theorem above. We first show that (C7) follows from (C1)–(C6) if K^0 and L^0 are both zero.

LEMMA 7

Let $y^0 \in \mathbb{R}_+^3$ be related to v_1, v_2 as in (C4), where (v_1, v_2) is any solution of (C5). Suppose also that y^0 satisfies (C6); that is, $\log v_1 + y_2^0 + y_3^0 \leq 0$. Then, for all $K, L > 0$,

$$(K + L)D_2\left(\frac{K}{K + L}; \frac{\mu_1}{\mu_1 + \mu_2}\right) - y_1^0 K + (y_1^0 + y_2^0 + y_3^0)L \geq 0. \tag{24}$$

Proof

It is clear that equality holds above if $K = L = 0$. Suppose K and L are not both zero. Then $K + L > 0$, so we can divide throughout by $(K + L)$ in (24). Letting $p = K/(K + L)$, and defining

$$h(p) = D_2\left(p; \frac{\mu_1}{\mu_1 + \mu_2}\right) - y_1^0 p + (y_1^0 + y_2^0 + y_3^0)(1 - p),$$

we see that (24) holds if $h(p) \geq 0$ for all $p \in [0, 1]$. We show that this is indeed the case by showing that, under the conditions of the lemma, $\inf_{p \in [0, 1]} h(p) \geq 0$. Setting the derivative of h equal to zero gives the equation

$$\log \frac{p}{1 - p} - \log \frac{\mu_1}{\mu_2} - (2y_1^0 + y_2^0 + y_3^0) = 0.$$

Letting \hat{p} denote the solution, we get

$$\hat{p} = \frac{\mu_1 z}{\mu_1 z + \mu_2}, \quad \text{where} \quad z = e^{2y_1^0 + y_2^0 + y_3^0}.$$

Since h is convex, the minimum of $h(p)$ for $p \in [0, 1]$ is attained at \hat{p} . Also

$$\begin{aligned} h(\hat{p}) &= \hat{p} \log \left[\frac{(\mu_1 + \mu_2)z}{\mu_1 z + \mu_2} \right] + (1 - \hat{p}) \log \left[\frac{\mu_1 + \mu_2}{\mu_1 z + \mu_2} \right] - \hat{p} \cdot \log z + (y_1^0 + y_2^0 + y_3^0) \\ &= \log \left(\frac{\mu_1 + \mu_2}{\mu_1 z + \mu_2} \right) - \log v_2 \\ &= \log \left(\frac{\mu_1 + \mu_2}{\mu_1 z v_2 + \mu_2 v_2} \right), \end{aligned} \tag{25}$$

where the second equality follows from the fact that $\log v_2 = -(y_1^0 + y_2^0 + y_3^0)$ by (C4). We wish to show that the last quantity is non-negative. From (C4), $v_1 v_2 = \exp(-y_3^0) \leq 1$ since $y_3^0 > 0$. Recall that $F^*(s) \leq 1$ if, and only if, $s \geq 0$. Since

(v_1, v_2) are assumed to solve (C5), it follows that

$$\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2 \geq 0. \tag{26}$$

Now, z was defined as $\exp(2y_1^0 + y_2^0 + y_3^0)$. Therefore, by the definition of v_1 and v_2 in (C4), and the fact that $y^0 \in \mathbb{R}_+^3$, we get

$$z v_2 = \exp(y_1^0) \leq \exp(y_1^0 + y_2^0) = v_1. \tag{27}$$

It follows from (26) and (27) that $\mu_1 + \mu_2 \geq \mu_1 z v_2 + \mu_2 v_2$. Hence, by (25), $h(\hat{p}) \geq 0$. Since h achieves its minimum over $p \in [0, 1]$ at \hat{p} , it follows that $h(p) \geq 0$ for all $p \in [0, 1]$. \square

Let \mathbf{p}^0 be given as in (C4), where the a_{jk} were specified in (1). If $\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2$ is less than σ , the left limit of finiteness of $F^*(s)$, then $F^*(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2) = +\infty$, so (C5) cannot hold. If either v_1 or v_2 is less than zero, then \mathbf{p}^0 is not non-negative, and hence cannot be a probability distribution. Thus, if we let D denote

$$D = \{ (x_1, x_2) : x_1 > 0, x_2 > 0, \mu_1 x_1 + \mu_2 x_2 < \mu_1 + \mu_2 - \sigma \}, \tag{28}$$

then we may certainly restrict attention to $(v_1, v_2) \in \bar{D}$, where \bar{D} denotes the closure of D . We may, in fact, say more : for (C5) to hold, it is not possible to have $v_1 v_2 = 0$, and it is only possible to have $\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2 = \sigma$ if $F^*(\sigma)$ is finite.

We shall need to make use of the quantities $\sum_{j,k=0}^{\infty} j p_{jk}^0$, and $\sum_{j,k=0}^{\infty} k p_{jk}^0$, which we now compute. We have

$$\begin{aligned} \sum_{j,k=0}^{\infty} j p_{jk}^0 &= \sum_{j,k=0}^{\infty} j a_{jk} v_1^{j-1} v_2^{k-1} = \frac{1}{v_2} \frac{\partial}{\partial v_1} \sum_{j,k=0}^{\infty} a_{jk} v_1^j v_2^k \\ &= \frac{1}{v_2} \frac{\partial}{\partial v_1} \sum_{j,k=0}^{\infty} \int_0^{\infty} \frac{(\mu_1 t)^j}{j!} \frac{(\mu_2 t)^k}{k!} v_1^j v_2^k e^{-(\mu_1 + \mu_2)t} dF(t) \\ &= \frac{1}{v_2} \frac{\partial}{\partial v_1} F^*(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2). \end{aligned} \tag{29}$$

We likewise get

$$\sum_{j,k=0}^{\infty} k p_{jk}^0 = \frac{1}{v_1} \frac{\partial}{\partial v_2} F^*(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2). \tag{30}$$

If $\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2$ is equal to σ , the left limit of finiteness of $F^*(s)$, and $F^*(\sigma)$

is finite, then the above derivatives have to be interpreted as one-sided derivatives and could possibly be infinite. We also note that

$$\frac{dF^*(s)}{ds} \Big|_{s=0} = \int_0^\infty -t dF(t) = -ET. \tag{31}$$

We assume in the following that $K^0 = L^0 = 0$. We shall find x^0, y^0 satisfying conditions (C1)–(C6) of Theorem 6. Then, by Lemma 7, they also satisfy (C7). Hence, by Theorem 6, x^0 solves the constrained optimization problem in (20), (21). We shall see that the constraint $g^3(x^0) \leq 0$ holds tightly (i.e., with equality). We distinguish four cases depending on which of the other two constraints, $g_i(x^0) \leq 0, i = 1, 2$, is tight, and analyze each case separately.

LEMMA 8

There exists (x^0, y^0) satisfying (C1)–(C6) with $g_1(x^0) < 0$ and $g_2(x^0) < 0$ if μ_1, μ_2 and F satisfy

$$\mu_1 \geq \mu_2 \quad \text{and} \quad \frac{\mu_1}{\mu_2 \nu_2} \frac{d}{dx_2} F^*(\mu_2 - \mu_2 x_2) \Big|_{x_2 = \nu_2} > 1. \tag{32}$$

If the above condition is satisfied, then one such (x^0, y^0) is described as follows:

- 1 $y_1^0 = y_2^0 = 0, y_3^0 = -\log \nu_2.$
- 2 $\mathbf{p}^0 = a_{jk} \nu_2^{k-1}.$
- 3 $M^0 = 0$ and n^0 solves $n^0 \sum_{j,k=0}^\infty (1-k) a_{jk} \nu_2^{k-1} = N.$

The global minimum value is given by $f(x^0) = -\beta N$, with $\beta = \log \nu_2$.

Proof

Suppose (32) holds and x^0, y^0 are specified by conditions 1–3 of the lemma. Then $y^0 \in \mathbb{R}_+^3$ since $\nu_2 < 1$. We shall show that n^0 is positive, $g_1(x^0)$ and $g_2(x^0)$ are negative, and that (x^0, y^0) satisfies (C1)–(C6).

Observe from conditions 1 and 2 of the lemma that (C4) holds with $(v_1, v_2) = (1, \nu_2)$, and that (C5) holds by definition of ν_2 . This in turn implies that \mathbf{p}^0 is a probability distribution, see (22). Since $\mu_1 \geq \mu_2$ by (32), $\nu_1 \leq \nu_2$ by Lemma 2. As $M^0 = 0$ by condition 3 of the lemma, we see that (C6) is satisfied.

Note that v_1, v_2 are positive and $\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2$ is also positive, hence bigger than σ , the left limit of finiteness of $F^*(s)$. Therefore (v_1, v_2) is in D , and F^* and its derivatives are finite at $\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2$. Hence,

$$\sum_{j,k=0}^\infty j a_{jk} \nu_2^{k-1} = \frac{\mu_1}{\mu_2 \nu_2} \frac{d}{dx_2} F^*(\mu_2 - \mu_2 x_2) \Big|_{x_2 = \nu_2} > 1 = \sum_{j,k=0}^\infty a_{jk} \nu_2^{k-1}, \tag{33}$$

where the first equality is due to (29) and the inequality is due to (32). Likewise,

$$\sum_{j,k=0}^{\infty} ka_{jk} \nu_2^{k-1} = \frac{d}{dx_2} F^*(\mu_2 - \mu_2 x_2) \Big|_{x_2=\nu_2} < 1 = \sum_{j,k=0}^{\infty} a_{jk} \nu_2^{k-1}, \tag{34}$$

where the first equality comes from (30) and the inequality from Lemma 1. The quantities above are finite since F^* was seen to have finite derivative at $\mu_2 - \mu_2 \nu_2$.

Observe from (34) that $\sum_{j,k=0}^{\infty} (1-k)a_{jk} \nu_2^{k-1}$ is positive and finite. Since N is assumed to be positive, n^0 given by condition 3 of the lemma is positive and finite. Recalling that K^0 and L^0 were assumed to be zero, we see from (16) and condition 3 of the lemma that $g_3(x^0) = 0$. Since $n^0 > 0$ and $\sum_{j,k=0}^{\infty} (j-1)a_{jk} \nu_2^{k-1} > 0$ by (33), we have from (16) that $0 = g_3(x^0)$ is bigger than $g_1(x^0)$ and $g_2(x^0)$. Together with conditions 1, this implies (C2) and (C3). Finally, since K^0, L^0 and M^0 are zero,

$$f(x^0) = n^0 \sum_{j,k=0}^{\infty} p_{jk}^0 \log \frac{p_{jk}^0}{a_{jk}} = \log \nu_2 \cdot n^0 \sum_{j,k=0}^{\infty} (k-1)a_{jk} \nu_2^{k-1}.$$

Hence, by condition 3 of the lemma,

$$f(x^0) = -\beta N, \quad \text{where} \quad \beta = \log \nu_2. \tag{35}$$

In particular, $f(x^0)$ is finite, so (C1) holds. This completes the proof that (x^0, y^0) specified by conditions 1–3 of the lemma satisfies (C1)–(C6). Consequently, by Theorem 6, x^0 is a global minimizer and $f(x^0)$ the global minimum value for the optimization problem posed in (20), (21). \square

LEMMA 9

There does not exist (x^0, y^0) with $K^0 = L^0 = 0$ satisfying (C1)–(C6) with $g_1(x^0) < 0$ and $g_2(x^0) = 0$.

Proof

Observe from the definition of the g_i in (16), and the fact that $K^0 = L^0 = 0$, that $g_2(x^0) \leq g_1(x^0)$. Clearly then, we cannot have $g_1(x^0) < 0$ and $g_2(x^0) = 0$. \square

LEMMA 10

There exists (x^0, y^0) satisfying (C1)–(C6) with $g_1(x^0) = 0$ and $g_2(x^0) < 0$ if μ_1, μ_2 and F satisfy

$$\mu_1 \leq \mu_2 \quad \text{and} \quad \frac{\mu_2}{\mu_1 \nu_1} \frac{d}{dx_1} F^*(\mu_1 - \mu_1 x_1) \Big|_{x_1=\nu_1} > 1. \tag{36}$$

In that case, one such (x^0, y^0) is described as follows:

- 1 $y_1^0 = \log \frac{\mu_2}{\mu_1}, y_2^0 = 0, y_3^0 = -\log \nu_1.$
- 2 $p_{jk}^0 = a_{jk}v_1^{j-1}v_2^{k-1},$ with $v_1 = \frac{\mu_2}{\mu_1}, v_2 = \frac{\mu_1\nu_1}{\mu_2}.$
- 3 n^0 solves $n^0 \sum_{j,k=0}^{\infty} (j-k)p_{jk}^0 = N.$
- 4 M^0 is given by $M^0 + n^0 \sum_{j,k=0}^{\infty} (1-k)p_{jk}^0 = N.$

The global minimum value is given by $f(x^0) = -\beta N,$ where $\beta = \log(\mu_1\nu_1/\mu_2).$

Proof

Suppose (32) holds and x^0, y^0 are specified by conditions 1–4 of the lemma. Then $y^0 \in \mathbb{R}_+^3$ since $\mu_1 \leq \mu_2$ and $\nu_1 < 1.$ We shall show that n^0 and M^0 are positive, $g_1(x^0)$ is zero, $g_2(x^0)$ is negative, and that (x^0, y^0) satisfies (C1)–(C6).

Observe from conditions 1 and 2 of the lemma, and the definition of $\nu_1,$ that (C4) and (C5) are satisfied. Consequently, p^0 is a probability distribution, see (22). It is also clear from condition 1 that (C6) holds. Now, v_1 and v_2 are positive and $\mu_1 - \mu_1v_1 + \mu_2 - \mu_2v_2$ is bigger than zero, hence also than $\sigma,$ the left limit of finiteness of $F^*(s).$ Therefore, $(v_1, v_2) \in D$ and F^* is finite, with finite derivative, at $\mu_1 - \mu_1v_1 + \mu_2 - \mu_2v_2.$ Hence,

$$\begin{aligned} \sum_{j,k=0}^{\infty} ja_{jk}v_1^{j-1}v_2^{k-1} &= \frac{\mu_2}{\mu_1\nu_1} \frac{d}{dx_1} F^*(\mu_1 - \mu_1x_1) \Big|_{x_1=\nu_1} \\ &> 1 = \sum_{j,k=0}^{\infty} a_{jk}v_1^{j-1}v_2^{k-1}, \end{aligned} \tag{37}$$

where the first equality is due to (29), and the inequality is due to (36). Likewise,

$$\sum_{j,k=0}^{\infty} ka_{jk}v_1^{j-1}v_2^{k-1} = \frac{d}{dx_1} F^*(\mu_1 - \mu_1x_1) \Big|_{x_1=\nu_1} < 1 = \sum_{j,k=0}^{\infty} a_{jk}v_1^{j-1}v_2^{k-1}, \tag{38}$$

where the first equality comes from (30), and the inequality from Lemma 1. The quantities above are finite, since it was noted that F^* has finite derivative at $\mu_1 - \mu_1\nu_1.$

Now, $\sum_{j,k=0}^{\infty} (j-k)a_{jk}v_1^{j-1}v_2^{k-1} > 0$ by (37) and (38), and it is finite. Since N is assumed to be positive, n^0 specified by conditions 2 and 3 of the lemma is positive and finite. Also, by conditions 3 and 4,

$$M^0 = n^0 \sum_{j,k=0}^{\infty} (j-1)p_{jk}^0 > 0, \tag{39}$$

where the inequality is due to (37) and the fact that $n^0 > 0$. Since $K^0 = L^0 = 0$, applying conditions 3 and 4 of the lemma to the definition of g in (16) gives $g_1(x^0) = 0$ and $g_3(x^0) = 0$. Since $M^0 > 0$, we also get $g_2(x^0) < g_1(x^0) = 0$. (C2) and (C3) now follow from the definition of y^0 in condition 1. Finally,

$$\begin{aligned}
 f(x^0) &= n^0 \sum_{j,k=0}^{\infty} p_{jk}^0 \log \frac{p_{jk}^0}{a_{jk}} - M^0 \log \nu_1 \\
 &= n^0 \sum_{j,k=0}^{\infty} p_{jk}^0 \left[(j-1) \log \left(\frac{\mu_2}{\mu_1} \right) + (k-1) \log \left(\frac{\mu_1 \nu_1}{\mu_2} \right) \right] - M^0 \log \nu_1 \\
 &= n^0 \sum_{j,k=0}^{\infty} (j-k) p_{jk}^0 \cdot \log \frac{\mu_2}{\mu_1} - \left(M^0 + n^0 \sum_{j,k=0}^{\infty} (1-k) p_{jk}^0 \right) \cdot \log \nu_1 \\
 &= N \log \left(\frac{\mu_2}{\mu_1} \right) - N \log \nu_1 \\
 &= -\beta N, \quad \text{where } \beta = \frac{\mu_1 \nu_1}{\mu_2}. \tag{40}
 \end{aligned}$$

The second equality above holds by condition 2, and the fourth by conditions 3 and 4 of the lemma. Thus, $f(x^0)$ is finite, so (C1) is satisfied. We have thus shown that (x^0, y^0) specified by conditions 1–4 of the lemma satisfies (C1)–(C6). Therefore, by Theorem 6, x^0 achieves the minimum in the optimization problem posed in (20), (21). □

Let $D \subseteq \mathbb{R}_+^2$ be defined as in (28), where σ is the left limit of finiteness of $F^*(s)$. Define

$$f(x_1, x_2) = F^*(\mu_1 - \mu_1 x_1 + \mu_2 - \mu_2 x_2) - x_1 x_2.$$

In order to describe (x^0, y^0) satisfying (C1)–(C6) with $g_1(x^0)$ and $g_2(x^0)$ both equal to zero, we need the following result, which is proved in Appendix A.

LEMMA 11

Consider the pair of equations

$$f(x_1, x_2) = 0, \tag{41}$$

$$\frac{\partial}{\partial x_1} f(x_1, x_2) = 0, \tag{42}$$

for $(x_1, x_2) \in \mathbb{R}_+^2$, and note that these are identical to (4) and (5) respectively. If neither (32) nor (36) holds, then this pair of equations has a solution $(v_1, v_2) \in D$

satisfying the conditions

$$v_1 \geq 1, \quad \nu_1 \leq v_1 v_2 < 1, \quad \frac{\partial}{\partial x_2} f(x_1, x_2) \Big|_{(x_1, x_2)=(v_1, v_2)} < 0, \tag{43}$$

which are identical to (6).

Proof

See Appendix A.

LEMMA 12

If neither (32) nor (36) holds, then there exists (x^0, y^0) satisfying (C1)–(C6) with $g_1(x^0) = g_2(x^0) = 0$. We describe one such (x^0, y^0) below, in terms of $(v_1, v_2) \in D$ solving (41) and (42) subject to (43). There is such a (v_1, v_2) by Lemma 11.

- 1 $y_1^0 = \log v_1, y_2^0 = 0, y_3^0 = -\log(v_1 v_2).$
- 2 $p_{jk}^0 = a_{jk} v_1^{j-1} v_2^{k-1}.$
- 3 $M^0 = 0, n^0$ solves $n^0 \sum_{j,k=0}^{\infty} (1-k) a_{jk} v_1^{j-1} v_2^{k-1} = N.$

The global minimum value is given by $f(x^0) = -\beta N$, where $\beta = \log v_2$.

Proof

Suppose neither (32) nor (36) holds. Let $(v_1, v_2) \in D$ solve (41) and (42) subject to (43), as in Lemma 11. We shall show that (x^0, y^0) specified in conditions 1-3 of the lemma has $n^0 > 0$ and satisfies (C1)–(C6) with $g_1(x^0) = g_2(x^0) = 0$.

(C4) is immediate from conditions 1 and 2 of the lemma, while (C5) holds by the definition of f , and the fact that (v_1, v_2) solve (41). Consequently, by (22), \mathbf{p}^0 is a probability distribution. (C6) holds because (v_1, v_2) satisfy (43). Since (v_1, v_2) is in D , F^* and its derivatives are finite at $\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2$. Hence,

$$\begin{aligned} \sum_{j,k=0}^{\infty} j a_{jk} v_1^{j-1} v_2^{k-1} &= \frac{1}{v_2} \frac{\partial}{\partial v_1} F^*(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2) \\ &= 1 = \sum_{j,k=0}^{\infty} a_{jk} v_1^{j-1} v_2^{k-1}, \end{aligned} \tag{44}$$

where the first equality is by (29), and the second by the definition of f and the fact that (v_1, v_2) solves (42). Likewise, by (30) and the fact that (v_1, v_2) satisfy (43), we have

$$\begin{aligned} \sum_{j,k=0}^{\infty} k a_{jk} v_1^{j-1} v_2^{k-1} &= \frac{1}{v_1} \frac{\partial}{\partial v_2} F^*(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2) \\ &< 1 = \sum_{j,k=0}^{\infty} a_{jk} v_1^{j-1} v_2^{k-1}. \end{aligned} \tag{45}$$

The quantities above are finite, since F^* was seen to have finite derivative at $\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2$.

Since N is assumed to be positive, and $\sum_{j,k=0}^{\infty} (1-k) a_{jk} v_1^{j-1} v_2^{k-1}$ is positive and finite by (45), n^0 given by condition 3 of the lemma is positive and finite. Condition 3, together with the assumption that $K^0 = L^0 = 0$, also implies that $g_3(x^0) = 0$, where the g_i were defined in (16). Combining this with (44), we get

$$g_1(x^0) = g_2(x^0) = g_3(x^0) = 0. \tag{46}$$

Hence, (C2) and (C3) are satisfied. Finally,

$$f(x^0) = n^0 \sum_{j,k=0}^{\infty} p_{jk}^0 \log \frac{p_{jk}^0}{a_{jk}} = n^0 \sum_{j,k=0}^{\infty} p_{jk}^0 \left[(j-1) \log v_1 + (k-1) \log v_2 \right].$$

Therefore, by (44) and the definition of \mathbf{p}^0 in condition 2,

$$f(x^0) = -\beta N, \quad \text{where} \quad \beta = \log v_2. \tag{47}$$

Since $(v_1, v_2) \in D$, v_2 is positive. Therefore $f(x^0)$ is finite, i.e., (C1) holds. Thus, (x^0, y^0) specified by conditions 1-4 of the lemma has n^0 positive and satisfies (C1)–(C6). Hence, by Theorem 6, $f(x^0)$ is the global minimum value for the optimization problem posed in (20), (21). \square

5. The upper bound

In the last section we obtained a solution to the optimization problem posed in (20) and (21) at the end of section 3. We shall now use this solution to compute an upper bound on $P(X_0^2 \geq N)$. A lower bound will be derived in the next section, thereby completing the proof of Theorem 3.

An upper bound on $P(X_0^2 \geq N)$ is provided in (13) in terms of an infinite sum. Each term in the sum is bounded above by $\exp(-\hat{f}(n, K, L, M))$, see (18), where the function \hat{f} is defined in (17) and (19). An alternative upper bound on the summands in (13) is given by

$$\nu_1^{M-1} \exp \left[(K+L) D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \right] P \left(\sum_{i=-n}^{-1} (1 - V_i^2) \geq N + L - M \right), \tag{48}$$

as follows from (13), (14) and (15).

Our approach to obtaining the upper bound will be as follows. Recall that the solution to the constrained optimization problem posed in section 3 provides a lower bound to $\hat{f}(n, K, L, M)$ for all (n, K, L, M) , after we re-introduce the constant term $-\log [(1 - \nu_1)/\nu_1]$ which was dropped from (19) for convenience in discussing the optimization problem. That is, letting $f(x^0)$ denote the constrained infimum of f , which was defined in (20), we have

$$\hat{f}(n, K, L, M) \geq f(x^0) - \log \frac{1 - \nu_1}{\nu_1} \quad \forall (n, K, L, M) \in \mathbb{R}_+^4.$$

Consequently, $\nu_1^{-1} \exp(-f(x^0))$ is an upper bound on each of the summands in (13). We shall use this upper bound when n, K, L and M are small, namely, when each is less than a constant times N . Otherwise, we shall use upper bounds which are derived below from (48).

Let V_i^2 denote the number of virtual services at the second queue during the i^{th} inter-arrival period. Let $M(\theta) = E[\exp \theta(1 - V_1^2)]$. Note that $M(0) = 1$ and

$$M'(0) = E[1 - V_1^2] = 1 - \mu_2 ET < 0.$$

Here T denotes a typical inter-arrival time, and the derivative should be interpreted as a derivative from the right if $M(\theta) = \infty$ for all $\theta < 0$. The inequality is due to the stability assumption. Thus, we may choose $0 < \theta < -\log \nu_1$, and $\alpha > 0$, so that

$$M(\theta) = \exp(-\alpha). \tag{49}$$

We also note that, for any $n \geq 1$

$$\begin{aligned} &P\left(\sum_{i=-n}^{-1} (1 - V_i^2) \geq N + L - M\right) \\ &= P\left(\exp\left[\theta \sum_{i=-n}^{-1} (1 - V_i^2)\right] \geq \exp[\theta(N + L - M)]\right) \\ &\leq \exp[-\theta(N + L - M)] \cdot [E \exp(\theta(1 - V_1^2))]^n \\ &\leq \exp[-\theta(N + L - M)] \cdot \exp[-\alpha n], \end{aligned} \tag{50}$$

with θ, α as in (49).

LEMMA 13

There exist constants c_0, c_1 and λ strictly positive, such that, for all $L^* \geq 0$ and $0 \leq L \leq L^*$,

$$\sum_{K=\lceil c_1 L^* \rceil}^{\infty} \exp\left[-(K + L)D_2\left(\frac{K}{K + L}; \frac{\mu_1}{\mu_1 + \mu_2}\right)\right] \leq c_0 \exp(-c_1 \lambda L^*),$$

and for all $L \geq 0$,

$$\sum_{K=0}^{\infty} \exp \left[-(K+L) D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \right] \leq c_0 + c_1 L.$$

Proof

We write

$$\begin{aligned} D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) &= \frac{K}{K+L} \log \frac{\mu_1 + \mu_2}{\mu_1} + \frac{L}{K+L} \log \frac{\mu_1 + \mu_2}{\mu_2} \\ &\quad - H_2 \left(\frac{K}{K+L} \right), \end{aligned}$$

where $H_2(p) \triangleq -p \log p - (1-p) \log (1-p)$ denotes the binary entropy of p . Let

$$\lambda = \frac{1}{2} \log \frac{\mu_1 + \mu_2}{\mu_1 \vee \mu_2},$$

where $\mu_1 \vee \mu_2$ denotes $\max\{\mu_1, \mu_2\}$. Then $0 < \lambda < \log 2$. Consequently, $H_2(p)$ is equal to λ for two values of p ; call the larger of these \bar{p} . Note that

$$H_2(p) \leq \lambda \quad \forall \quad \bar{p} \leq p \leq 1.$$

Define $c_1 = \bar{p}/(1-\bar{p})$. Then, for $0 \leq L \leq L^*$ and $K \geq c_1 L^*$, we have $K/(K+L) \geq \bar{p}$. Consequently, $H_2(K/(K+L)) \leq \lambda$, and

$$D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \geq \log \frac{\mu_1 + \mu_2}{\mu_1 \vee \mu_2} - H_2 \left(\frac{K}{K+L} \right) \geq 2\lambda - \lambda = \lambda.$$

It follows that, for $0 \leq L \leq L^*$,

$$\begin{aligned} \sum_{K=\lceil c_1 L^* \rceil}^{\infty} \exp \left[-(K+L) D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \right] &\leq \sum_{K=\lceil c_1 L^* \rceil}^{\infty} e^{-\lambda K} \\ &= (1 - e^{-\lambda})^{-1} \exp(-\lambda c_1 L^*). \end{aligned}$$

Setting $c_0 = (1 - \exp(-\lambda))^{-1}$ gives the first claim of the lemma.

Next, let $L \geq 0$ be given, and observe that

$$\begin{aligned} & \sum_{K=0}^{\infty} \exp \left[-(K+L)D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \right] \\ & \leq \sum_{K=0}^{\lfloor c_1 L \rfloor} \exp \left[-(K+L)D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \right] \\ & \quad + \sum_{K=\lceil c_1 L \rceil}^{\infty} \exp \left[-(K+L)D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \right] \\ & \leq c_1 L + c_0 \exp(-c_1 \lambda L) \leq c_0 + c_1 L, \end{aligned}$$

which gives the second claim of the lemma. □

A consequence of the above lemma and (50) is that

$$\begin{aligned} & \sum_{K=0}^{\infty} \exp \left[-(K+L)D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \right] \nu_1^{M-1} \\ & \cdot P \left(\sum_{i=-n}^{-1} (1 - V_i^2) \geq N + L - M \right) \\ & \leq \nu_1^{-1} (c_0 + c_1 L) \exp(-\alpha n - \theta(N + L - M) + M \log \nu_1). \end{aligned} \tag{51}$$

Define

$$z_1 = \exp(-\alpha), \quad z_2 = \exp(-\theta), \quad z_3 = \exp(\log \nu_1 + \theta). \tag{52}$$

Since $\alpha > 0, \theta > 0$, and $\theta < -\log \nu_1$, we have $z_i < 1, i = 1, 2, 3$. Let a_1, a_2 and a_3 be arbitrary positive constants, and define

$$\mathcal{K} = \{(n, L, M) : n \leq a_1 N, L \leq a_2 N, M \leq a_3 N\}.$$

It now follows from (51) that

$$\begin{aligned} & \sum_{(n,L,M) \notin \mathcal{K}} \sum_{K=0}^{\infty} \exp \left[-(K+L)D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \right] \nu_1^{M-1} \\ & \cdot P \left(\sum_{i=-n}^{-1} (1 - V_i^2) \geq N + L - M \right) \\ & \leq \nu_1^{-1} z_2^N \sum_{(n,L,M) \notin \mathcal{K}} (c_0 + c_1 L) z_1^n z_2^L z_3^M \\ & \leq \nu_1^{-1} z_2^N \frac{(c_0 + c_1)(z_1^{a_1 N} + z_3^{a_3 N}) + (c_0 + c_1 a_2 N)z_2^{a_2 N}}{(1 - z_1)(1 - z_2)^2(1 - z_3)}. \end{aligned} \tag{53}$$

We also note from Lemma 13 that, if $(n, L, M) \in \mathcal{K}$, then

$$\sum_{K=c_1 a_2 N}^{\infty} \exp \left[-(K+L) D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \right] \leq c_0 \exp(-c_1 a_2 \lambda N). \tag{54}$$

Define $a_4 = c_1 a_2$, and $z_4 = \exp(-\lambda)$, and note that $0 < z_4 < 1$. We now get from (54) that

$$\begin{aligned} & \sum_{(n,L,M) \in \mathcal{K}} \sum_{K=a_4 N}^{\infty} \exp \left[-(K+L) D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \right] \nu_1^{M-1} \\ & \quad \cdot P \left(\sum_{i=-n}^{-1} (1 - V_i^2) \geq N + L - M \right) \\ & \leq \sum_{(n,L,M) \in \mathcal{K}} c_0 z_4^{a_4 N} \nu_1^{M-1} \leq \frac{a_1 a_2 N^2 \cdot c_0 z_4^{a_4 N}}{\nu_1 (1 - \nu_1)}. \end{aligned} \tag{55}$$

From the solution to the optimization problem, we also have

$$\begin{aligned} & \sum_{(n,L,M) \in \mathcal{K}} \sum_{K=0}^{a_4 N} \exp \left[-(K+L) D_2 \left(\frac{K}{K+L}; \frac{\mu_1}{\mu_1 + \mu_2} \right) \right] \nu_1^{M-1} \\ & \quad \cdot P \left(\begin{array}{l} K - L + \sum_{i=-n}^{-1} (V_i^1 - V_i^2) \geq N \\ M - L + \sum_{i=-n}^{-1} (V_i^1 - V_i^2) \geq N \\ M - L + \sum_{i=-n}^{-1} (1 - V_i^2) \geq N \end{array} \right) \\ & \leq \nu_1^{-1} a_1 a_2 a_3 a_4 N^4 \exp(\beta N). \end{aligned} \tag{56}$$

Let $\gamma = z_2^{-1} e^\beta$. Suppose a_1, a_2, a_3 and a_4 are chosen sufficiently large, so that

$$\frac{z_1^{a_1}}{1 - z_1} \leq \gamma, \quad \frac{z_2^{a_2}}{(1 - z_2)^2} \leq \gamma, \quad \frac{z_3^{a_3}}{1 - z_3} \leq \gamma, \quad z_4^{a_4} \leq e^\beta,$$

which is possible since $z_i < 1, i = 1, \dots, 4$. We see from (53), (55) and (56) that the sum in (13) is bounded above by

$$(c_0 + c_1)(2 + a_2 N) e^{\beta N} + \frac{a_1 a_2 N^2 e^{\beta N}}{\nu_1 (1 - \nu_1)} + \frac{a_1 a_2 a_3 a_4 N^4 e^{\beta N}}{\nu_1}.$$

Consequently, by (13),

$$P(X_0^2 \geq N) \leq cN^4 \exp(\beta N), \tag{57}$$

where

$$c = (c_0 + c_1)(2 + a_2) + \frac{a_1 a_2}{\nu_1(1 - \nu_1)} + \frac{a_1 a_2 a_3 a_4}{\nu_1}.$$

This completes the derivation of the upper bound.

6. The lower bound

In section 4, we obtained $x^0 = (\mathbf{p}^0, n^0, K^0, L^0, M^0)$ that solves the optimization problem posed in section 3. We found that $K^0 = L^0 = 0$. The quantities \mathbf{p}^0, n^0, M^0 have the following interpretation. The probability distribution \mathbf{p}^0 is the twisted distribution which describes the most likely evolution of the system to large queue size. That is, in order for the queue size to build up, it is necessary that the virtual service processes during each inter-arrival period have the distribution \mathbf{p}^0 rather than their original distribution \mathbf{a} , that this behavior be sustained over n^0 inter-arrival periods, and that we start with M^0 customers in the first queue. We shall make use of the intuition sketched above in proving the lower bound.

We derive sufficient conditions for the event $\{X_0^2 \geq N\}$, and estimate their probability. Define

$$m_1 = \sum_{j,k=0}^{\infty} j p_{jk}^0 \quad m_2 = \sum_{j,k=0}^{\infty} k p_{jk}^0. \tag{58}$$

Let $\epsilon > 0$ be given. Consider the system at time $\tau_{-(1+\epsilon)n^0}$ having at least $(1 + \epsilon)M^0$ customers in the first queue, i.e., $X_{-(1+\epsilon)n^0}^2 \geq (1 + \epsilon)M^0$. Suppose that the virtual service process during $[\tau_{-(1+\epsilon)n^0}, \tau_0)$ evolves in a tube around a straight line with gradient (m_1, m_2) . More precisely, let $\delta > 0$ be a given constant, and suppose that the evolution of the system during the period $[\tau_{-(1+\epsilon)n^0}, \tau_0)$ satisfies the following constraints. For all $k \in \{1, \dots, (1 + \epsilon)n^0\}$,

$$\frac{m_1 k}{(1 + \epsilon)n^0} - \delta < \frac{1}{(1 + \epsilon)n^0} \sum_{i=-(1+\epsilon)n^0}^{-(1+\epsilon)n^0+k-1} V_i^1 < \frac{m_1 k}{(1 + \epsilon)n^0} + \delta, \tag{59}$$

$$\frac{m_2 k}{(1 + \epsilon)n^0} - \delta < \frac{1}{(1 + \epsilon)n^0} \sum_{i=-(1+\epsilon)n^0}^{-(1+\epsilon)n^0+k-1} V_i^2 < \frac{m_2 k}{(1 + \epsilon)n^0} + \delta. \tag{60}$$

If these constraints are satisfied for $\delta = c\epsilon$ with c sufficiently small, then we show that $X_0^2 \geq N$.

Let $S_k = (S_k^1, S_k^2)$ denote the actual number of services at the first and second queue respectively during the interval $[\tau_k, \tau_{k+1})$. Clearly, then

$$X_0^2 \geq \sum_{i=1}^{(1+\epsilon)n^0} (S_{-i}^1 - S_{-i}^2). \tag{61}$$

We shall obtain a lower bound on the sum of the S_{-i}^1 using (59), an upper bound on the sum of the S_{-i}^2 using (60), and thereby get a lower bound on X_0^2 .

The number of actual services at the second queue during $[\tau_{-(1+\epsilon)n^0}, \tau_0)$ is bounded above by the number of virtual services during this same period. Hence, if (60) holds, we have

$$\sum_{i=1}^{(1+\epsilon)n^0} S_{-i}^2 \leq (1 + \epsilon)(m_2 + \delta)n^0. \tag{62}$$

Suppose the first queue is never empty during the period $(\tau_{-(1+\epsilon)n^0}, \tau_0]$. Then, every virtual service results in an actual service. Hence, by (59),

$$\sum_{i=1}^{(1+\epsilon)n^0} S_{-i}^1 \geq (1 + \epsilon)(m_1 - \delta)n^0. \tag{63}$$

Next, suppose the first queue does empty during $(\tau_{-(1+\epsilon)n^0}, \tau_0]$. Let $\tau_{-(1+\epsilon)n^0+\kappa}$ denote the last time that the first queue is empty during this period. Note that $1 \leq \kappa \leq (1 + \epsilon)n^0$. The first queue is never empty during $(\tau_{-(1+\epsilon)n^0+\kappa}, \tau_0]$, and so the number of services at the first queue during this period is the same as the number of virtual services. The latter, by (59), is at least $(1 + \epsilon)(m_1 - 2\delta)n^0 - m_1\kappa$. The number of services at the first queue during $[\tau_{-(1+\epsilon)n^0}, \tau_{-(1+\epsilon)n^0+\kappa})$ is equal to the number of external arrivals during this time plus the number originally in the queue at time $\tau_{-(1+\epsilon)n^0}$, since the first queue is empty at the end of this period. But this is equal to $\kappa + (1 + \epsilon)M^0$. Thus, we get

$$\sum_{i=1}^{(1+\epsilon)n^0} S_{-i}^1 \geq (1 + \epsilon)M^0 + (1 + \epsilon)(m_1 - 2\delta)n^0 + (1 - m_1)\kappa.$$

But $m_1 \geq 1$ by (33), (37) and (44). We also noted above that $\kappa < (1 + \epsilon)n^0$. Combining these observations with the above expression, we get

$$\sum_{i=1}^{(1+\epsilon)n^0} S_{-i}^1 \geq (1 + \epsilon)M^0 + (1 + \epsilon)(1 - 2\delta)n^0. \tag{64}$$

From (61)–(64), we see that either

$$X_0^2 \geq (1 + \epsilon)(m_1 - m_2 - 2\delta)n^0 \tag{65}$$

or

$$X_0^2 \geq (1 + \epsilon)[M^0 + (1 - m_2 - 3\delta)n^0]. \tag{66}$$

Since x^0 solves the constrained optimization problem in (20), (21), $g_i(x^0) \leq 0$ for $i = 1, 2, 3$, where the g_i were defined in (16). Here $x^0 = (\mathbf{q}^0, n^0, K^0, L^0, M^0)$, and we saw that $K^0 = L^0 = 0$. Hence, with m^1, m^2 as in (58), we have

$$g_1(x^0) \leq 0 \Rightarrow (m_1 - m_2)n^0 \geq N, \quad g_3(x^0) \leq 0 \Rightarrow M^0 + (1 - m_2)n^0 \geq N.$$

Therefore, (65) and (66) imply that

$$X_0^2 \geq (1 + \epsilon)(N - 3\delta aN), \tag{67}$$

where we have replaced n^0 by aN , and a is a positive real number defined in Lemmas 8,10 and 12. It is clear from (67) that, if $\delta \leq \epsilon/[3a(1 + \epsilon)]$, then $X_0^2 \geq N$. This is exactly what we set out to prove.

Let $P(\epsilon, \delta, N)$ denote the probability that the virtual service process satisfies the conditions in (59) and (60). We write N rather than n^0 in the notation since the relation between N and n^0 is explicit by Lemmas 8,10 and 12. The probability that the virtual service process satisfies these constraints is given by well-known results in large deviations theory, see [20] or Lemma 5.1.6 of [12]. These results use the additional assumption that F has exponentially decaying tails, which we do not need for a result of the form we seek. We therefore give a proof below using a change of measure argument.

LEMMA 14

Let X_1, \dots, X_n be *i.i.d* \mathbb{Z}_+^2 -valued random variables with distribution \mathbf{a} . Denote $X_i = (X_i^1, X_i^2)$. Let \mathbf{p} be a probability distribution on \mathbb{Z}_+^2 of the form

$$p_{jk} = a_{jk}v_1^{j-1}v_2^{k-1} \quad \forall \quad j, k \in \mathbb{Z}_+.$$

In particular, we require that v_1, v_2 be positive and finite. Suppose the expectation and variance of a random variable with the distribution \mathbf{p} are finite, and denote the expectation by $\underline{m} = (m_1, m_2)$. Define $S_k = X_1 + \dots + X_k$. Let $\|\cdot\|$ denote the Euclidean norm. Then, for all $\delta > 0$,

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P\left(\|S_k - k\underline{m}\| \leq \delta n, 1 \leq k \leq n\right) \geq -D(\mathbf{p}; \mathbf{a}).$$

Proof

Let $n, \delta > 0$ be given. Denote by $\mathcal{A}_{n,\delta}$ the event $\{\|S_k - k\underline{m}\| \leq \delta n, 1 \leq k \leq n\}$. Let P_a, P_p denote the product distribution on $(\mathbb{Z}_+^2)^n$ corresponding to \mathbf{a}, \mathbf{p} respectively. We wish to estimate $P_a(\mathcal{A}_{n,\delta})$, the probability of the event $\mathcal{A}_{n,\delta}$ under the true distribution of X_1, \dots, X_n . But

$$P_a(\mathcal{A}_{n,\delta}) = \sum_{\omega \in \mathcal{A}_{n,\delta}} P_p(\omega) \frac{P_a(\omega)}{P_p(\omega)} \geq P_p(\mathcal{A}_{n,\delta}) \cdot \inf_{\omega \in \mathcal{A}_{n,\delta}} \frac{P_a(\omega)}{P_p(\omega)}. \tag{68}$$

Now, for $\omega = (x_1, \dots, x_n) \in (\mathbb{Z}_+^2)^n$,

$$\log \frac{P_a(\omega)}{P_p(\omega)} = \sum_{i=1}^n [(1 - x_i^1) \log v_1 + (1 - x_i^2) \log v_2].$$

If $\omega \in \mathcal{A}_{n,\delta}$, then $\|\sum_{i=1}^n x_i - n\underline{m}\| \leq \delta n$, and so

$$\log \frac{P_a(\omega)}{P_p(\omega)} \geq (1 - m_1)n \log v_1 + (1 - m_2)n \log v_2 - n\delta(|\log v_1| + |\log v_2|).$$

Hence, by (68),

$$\begin{aligned} \frac{1}{n} \log P_a(\mathcal{A}_{n,\delta}) &\geq \frac{1}{n} \log P_p(\mathcal{A}_{n,\delta}) + (1 - m_1) \log v_1 + (1 - m_2) \log v_2 \\ &\quad - \delta(|\log v_1| + |\log v_2|). \end{aligned} \tag{69}$$

Recall that $E_p[X] = \underline{m}$ by the definition of \underline{m} . That is, if X has the distribution \mathbf{p} , then its expectation is \underline{m} . Hence, by a functional law of large numbers, $P_p(\mathcal{A}_{n,\delta}) \rightarrow 1$ as $n \rightarrow \infty$, see, for example, Corollary 1 of [3]. Consequently,

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P_p(\mathcal{A}_{n,\delta}) = 0. \tag{70}$$

Also note that

$$\begin{aligned} D(\mathbf{p}; \mathbf{a}) &= \sum_{j,k=0}^{\infty} p_{jk} [(j - 1) \log v_1 + (k - 1) \log v_2] \\ &= (m_1 - 1) \log v_1 + (m_2 - 1) \log v_2, \end{aligned} \tag{71}$$

where the second equality uses the fact that $\underline{m} = (m_1, m_2)$ was defined as the expectation of the distribution \mathbf{p} . Since $P_a(\mathcal{A}_{n,\delta}) \geq P_a(\mathcal{A}_{n,\epsilon})$ for all $0 < \epsilon < \delta$ and

all n , we see from (69), (70) and (71) that

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P_a(\mathcal{A}_{n,\delta}) \geq -D(\mathbf{p}; \mathbf{a}) - \epsilon(|\log v_1| + |\log v_2|) \quad \forall \quad 0 < \epsilon < \delta.$$

Note that v_1 and v_2 are positive and finite, so $\log v_1$ and $\log v_2$ are finite. Letting ϵ decrease to zero above completes the proof of the lemma. \square

Let \mathbf{p}^0 and n^0 be as in the solution to the optimization problem. We saw in Lemmas 8, 10 and 12 that $p_{jk}^0 = a_{jk} v_1^{j-1} v_2^{k-1}$, where $(v_1, v_2) \in D$ for the set D defined in (28). By the definition of D , F^* is analytic in a neighborhood of $\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2$. Consequently, \mathbf{p}^0 has finite mean and variance. Let m be defined as in (58). As above, let $P(\epsilon, \delta, N)$ denote the probability that the virtual service process satisfies the conditions in (59) and (60). Then, it follows from the above lemma that

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P(\epsilon, c\epsilon, N) \geq -\frac{(1 + \epsilon)n^0}{N} D(\mathbf{p}^0; \mathbf{a}). \tag{72}$$

Here \mathbf{a} , defined in (1), is the original distribution of the virtual service process.

We have shown above that, if $X_{-(1+\epsilon)n^0}^1 \geq (1 + \epsilon)M^0$, and (59) and (60) hold for δ sufficiently small, then $X_0^2 \geq N$. By the known stationary distribution for the first queue,

$$P\left(X_{-(1+\epsilon)n^0}^1 \geq (1 + \epsilon)M^0\right) = \nu_1^{(1+\epsilon)M^0}. \tag{73}$$

Furthermore, this last event is independent of the event that (59) and (60) hold, since (59) and (60) involve only the virtual service process after time $\tau_{-(1+\epsilon)n^0}$. Hence, by (72) and (73),

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P(X_0^2 \geq N) \geq \frac{1 + \epsilon}{N} \left[M^0 \log \nu_1 - n^0 D(\mathbf{p}^0; \mathbf{a}) \right]. \tag{74}$$

Now, observe from the definition of f in (20), and the fact that $K^0 = L^0 = 0$, that the term in brackets above is precisely $-\inf f$, the optimum value in the constrained optimization problem. Also recall that $\inf f = -\beta N$, for β as defined in Lemmas 8,10 and 12. Finally, since $\epsilon > 0$ is arbitrary, we let $\epsilon \rightarrow 0$ in (74) to obtain

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P(X_0^2 \geq N) \geq \beta. \tag{75}$$

Combining the lower bound in (75) with the upper bound obtained in the previous section in (57), and with the value of β given by Lemmas 8, 10 and 12, the proof of Theorem 3 is complete.

7. Conclusion

We considered the problem of estimating the tail of the stationary queue length distribution in a tandem of exponential server queues with renewal arrivals. We showed that the marginal distribution at each queue decays exponentially, and obtained the exact exponential rate of decay. That is, if X_n^i , $i = 1, 2$, denotes the queue length at the i th queue seen by the n th arrival, then, in stationarity

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P(X_n^2 \geq N) = \beta,$$

where β is defined in Theorem 1. By classical results for the $GI/M/1$ queue, we also have

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P(X_n^1 \geq N) = \log \nu_1,$$

where ν_1 is defined below (2).

A result of this sort impacts several related problems. For instance, in the tandem above, consider how to optimally allocate a total of N buffers so as to minimize some cost function associated with the time to buffer overflow. We suggest the following rule of thumb : allocate $p_1 N$ buffers to the first queue and $p_2 N$ to the second, where $p_1 + p_2 = 1$, such that $p_1 \log \nu_1 = p_2 \beta$. It can be shown that, for any reasonable cost function associated with the time to buffer overflow, the above allocation is approximately optimal in the sense that, if $N_1(N), N_2(N)$ is the exact optimal allocation for that cost function, then

$$\lim_{N \rightarrow \infty} \frac{N_1(N)}{N} = p_1, \quad \lim_{N \rightarrow \infty} \frac{N_2(N)}{N} = p_2.$$

The technique for proving this is parallel to that in [1, 2].

Our solution technique identified the most likely path along which large queue sizes build up in the second queue. The description of this path depends on the values of the system parameters μ_1, μ_2 and F . In terms of the quantities n^0, M^0, v_1 and v_2 given in Lemmas 8, 10 and 12, the most likely way for the second queue to build up a large backlog N when the system is started empty is as follows. If M^0 is non-zero, which is the case only if (36) holds, then there is an initial phase during which the first queue builds up a large backlog M^0 . If (36) does not hold, then this phase is non-existent. In the next phase, the system evolves for n^0 inter-arrival periods as if the service rates were $\mu_1 v_1$ at the first queue and $\mu_2 v_2$ at the second, while the inter-arrival time distribution was

$$\hat{F}(t) = \int_0^t \exp [-(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2)s] dF(s).$$

We note that, if (32) holds, then the service rate at the first queue is unchanged and the first queue remains stable under the modified inter-arrival distribution. Therefore, the second queue evolves as if the modified arrivals feed directly into it. Next, if (36) holds, then the first queue starts this phase with M^0 customers already in it. It remains stable under the modified parameters, and is close to empty after n^0 inter-arrival periods, while the second queue has built up. Finally, if neither (32) nor (36) holds, then the first queue is critical under the modified parameters. In all cases, the second queue is unstable and therefore builds up a large backlog.

The problem of extending our result to an arbitrary number of queues in tandem remains open. It would also be of interest to extend the result to more general service distributions, and to consider arrival and service processes that are autocorrelated.

Acknowledgements

This research was carried out while the authors were at the School of Electrical Engineering, Cornell University. The paper was written while the first author was on a fellowship from BP and the Royal Society of Edinburgh.

Appendix A: Proof of Lemma 11

This appendix deals with the proof of Lemma 11, which was stated in section 4. The function f was defined as

$$f(x_1, x_2) = F^*(\mu_1 - \mu_1x_1 + \mu_2 - \mu_2x_2) - x_1x_2. \tag{76}$$

Here $F^*(s)$ is the Laplace transform of F , which is the probability distribution function of a non-negative random variable, T , which is non-zero with positive probability. Recall that $F^*(s)$ is a strictly decreasing, strictly convex function of s , and that $\log F^*(s)$ is a convex function of s . Hence, both $F^*(\mu_1 - \mu_1x_1 + \mu_2 - \mu_2x_2)$ and its logarithm are convex functions of (x_1, x_2) . Also, $F^*(\mu_1 - \mu_1x_1 + \mu_2 - \mu_2x_2)$ is an analytic function for $(x_1, x_2) \in D$, where the set D was defined in (28). $\sigma \leq 0$ is the left limit of finiteness of $F^*(s)$, and σ could be $-\infty$. Therefore, the restriction of f to D is in $C^\infty(D)$ (for an open set I , we define $C^k(I)$ to be the class of k times continuously differentiable functions on I). Finally, recall that $\mu_i, i = 1, 2$ were assumed to satisfy $\int_0^\infty \mu_i t dF(t) > 1$, and $\nu_i, i = 1, 2$ were defined to be the unique solutions smaller than 1 of $\nu_i = F^*(\mu_i - \mu_i\nu_i)$. We shall make use of the above facts in proving Lemma 11

Note that eq. (41) cannot have solutions with $x_1 = 0$ or $x_2 = 0$ since $F^*(\mu_1 - \mu_1x_1 + \mu_2 - \mu_2x_2) = 0$ is impossible for any finite (x_1, x_2) . Consequently,

we may restrict attention to

$$\mathring{\mathbb{R}}_+^2 = \{(x_1, x_2) \in \mathbb{R}_+^2 : x_1 > 0, x_2 > 0\}.$$

LEMMA 15

Eq. (41) cannot have three collinear solutions for (x_1, x_2) in $\mathring{\mathbb{R}}_+^2$.

Proof

Eq. (41) is equivalent to $F^*(\mu_1 - \mu_1 x_1 + \mu_2 - \mu_2 x_2) = x_1 x_2$. Taking logarithms, we get the equation

$$g(x_1, x_2) = \log F^*(\mu_1 - \mu_1 x_1 + \mu_2 - \mu_2 x_2) - \log x_1 - \log x_2 = 0.$$

Since $\log F^*(\mu_1 - \mu_1 x_1 + \mu_2 - \mu_2 x_2)$ is a convex function of (x_1, x_2) , and $\log x_1 + \log x_2$ is a strictly concave function of (x_1, x_2) , $g(x_1, x_2)$ is strictly convex. Therefore, $g(x_1, x_2) = 0$ cannot have three collinear solutions. \square

LEMMA 16

The set $\mathcal{S} = \{(x_1, x_2) \in \mathring{\mathbb{R}}_+^2 : f(x_1, x_2) = 0\}$ is a bounded subset of \mathbb{R}_+^2 .

Proof

Let $\mu = \min\{\mu_1, \mu_2\}$ and $z = \mu_1 x_1 + \mu_2 x_2$. Recall that F is the distribution of a non-negative random variable T , which is non-zero with positive probability. Thus, there exist positive constants ϵ and δ , such that $P(T \geq \delta) \geq \epsilon$. Therefore, $F^*(s) \geq \epsilon \exp[-s\delta]$ for all $s \leq 0$. In particular, if $z > \mu_1 + \mu_2$, then

$$f(x_1, x_2) \geq \epsilon \exp(z - \mu_1 - \mu_2) - x_1 x_2 \geq (\epsilon \exp(-\mu_1 - \mu_2)) \exp(z) - (\mu^{-1} z)^{-2}.$$

It is clear from above that there exists $X > \mu_1 + \mu_2$ such that, if $z > X$, then $f(x_1, x_2) > 0$. Therefore, we see from the definition of \mathcal{S} in the statement of the lemma that $(x_1, x_2) \in \mathcal{S}$ implies $\mu_1 x_1 + \mu_2 x_2 \leq X$. \square

COROLLARY 17

Let the set \mathcal{S} be defined as in the lemma above. There exists $\delta > 0$ such that

$$(x_1, x_2) \in \mathcal{S} \Rightarrow x_1 > \delta \quad \text{and} \quad x_2 > \delta.$$

Proof

Let $\delta = \mu F^*(\mu_1 + \mu_2)/X$ where X is as in the lemma above, and note that $\delta > 0$. If $x_2 < \delta$, then $x_1 x_2 < \delta \mu^{-1} (\mu_1 x_1 + \mu_2 x_2)$. If $(x_1, x_2) \in \mathcal{S}$, then $\mu_1 x_1 + \mu_2 x_2 \leq X$ by definition of X in the lemma above. Also, x_1, x_2 are positive

by the definition of \mathcal{S} , and $F^*(s)$ is a decreasing function of s , so $F^*(\mu_1 + \mu_2) \leq F^*(\mu_1 - \mu_1 x_1 + \mu_2 - \mu_2 x_2)$. Therefore, if $(x_1, x_2) \in \mathcal{S}$ and $x_2 < \delta$, then

$$F^*(\mu_1 - \mu_1 x_1 + \mu_2 - \mu_2 x_2) \geq F^*(\mu_1 + \mu_2) = \delta \mu^{-1} X \geq \delta \mu^{-1} (\mu_1 x_1 + \mu_2 x_2) > x_1 x_2,$$

where the equality holds by definition of δ . But the above implies $f(x_1, x_2) > 0$, contradicting $(x_1, x_2) \in \mathcal{S}$. A similar argument applies if $x_1 < \delta$. \square

LEMMA 18

Suppose the conditions in neither (32) nor (36) hold. Then,

$$\frac{\mu_1}{\mu_2 \nu_2} \frac{d}{dx_2} F^*(\mu_2 - \mu_2 x_2) \Big|_{x_2=\nu_2} \leq 1.$$

Proof

Suppose first that $\mu_1 \geq \mu_2$. Then, since (32) was assumed not to be satisfied, we see that the conclusion of the lemma holds. Next, suppose $\mu_1 < \mu_2$. Then $\nu_1 > \nu_2$ by Lemma 2, and so $\mu_1 - \mu_1 \nu_1$ is less than $\mu_2 - \mu_2 \nu_2$. Therefore, by the convexity of $\log F^*$,

$$\frac{d \log F^*(s)}{ds} \Big|_{s=\mu_1 - \mu_1 \nu_1} \leq \frac{d \log F^*(s)}{ds} \Big|_{s=\mu_2 - \mu_2 \nu_2}.$$

Since $F^*(\mu_i - \mu_i \nu_i) = \nu_i$, $i = 1, 2$, the above is equivalent to

$$\frac{-1}{\mu_1 \nu_1} \frac{d}{dx_1} F^*(\mu_1 - \mu_1 x_1) \Big|_{x_1=\nu_1} \leq \frac{-1}{\mu_2 \nu_2} \frac{d}{dx_2} F^*(\mu_2 - \mu_2 x_2) \Big|_{x_2=\nu_2}.$$

Note that $dF^*(\mu_i - \mu_i x)/dx > 0$ for $i = 1, 2$ and for all x , because F^* is a decreasing function. Hence, observe from the last inequality above, and the supposition that $\mu_1 < \mu_2$, that

$$\frac{\mu_1}{\mu_2 \nu_2} \frac{d}{dx_2} F^*(\mu_2 - \mu_2 x_2) \Big|_{x_2=\nu_2} \leq \frac{\mu_2}{\mu_1 \nu_1} \frac{d}{dx_1} F^*(\mu_1 - \mu_1 x_1) \Big|_{x_1=\nu_1}.$$

Since $\mu_1 < \mu_2$, and (36) was assumed not to hold, the conclusion of the lemma follows from the above inequality. \square

THEOREM 19

Let $f(x_1, x_2)$ be defined as in (76). So the restriction of f to D is in $C^\infty(D)$.

Suppose $(y_1, y_2) \in D$ satisfies the following conditions.

1. $f(y_1, y_2) = 0, \quad y_2 < 1, \quad y_1 y_2 < 1.$
2. $\frac{\partial}{\partial x_1} f(x_1, x_2) \Big|_{(x_1, x_2) = (y_1, y_2)} \leq 0.$
3. $\frac{\partial}{\partial x_2} f(x_1, x_2) \Big|_{(x_1, x_2) = (y_1, y_2)} < 0.$
4. $\exists \hat{y}_2 > y_2 : f(y_1, \hat{y}_2) = 0.$

Then there is $(v_1, v_2) \in D$ solving (41) and (42) subject to the conditions

$$v_1 \geq y_1, \quad v_2 \leq y_2, \quad \frac{\partial}{\partial x_2} f(x_1, x_2) \Big|_{(x_1, x_2) = (v_1, v_2)} < 0. \tag{77}$$

We shall use the above theorem to prove Lemma 11. The proof of the theorem proceeds through a sequence of lemmas, which we outline below. Note that, if the condition in item 3 above is satisfied, then, by the Implicit Function Theorem (see, e.g., [19]), there is an open neighborhood U of y_1 , an open neighborhood V of y_2 , and a unique function $h : U \mapsto V$ such that $f(x, h(x)) = 0$ for all $x \in U$. Further, $h \in C^\infty(U)$ since $f \in C^\infty(D)$ and $(y_1, y_2) \in D$. Also, $h(y_1) = y_2$ by the uniqueness of h .

The intuition behind the proof of Theorem 19 is as follows. We find a function h as above, defined on a neighborhood of (y_1, y_2) , and satisfying $f(x, h(x)) = 0$ for all x in this neighborhood. That is, $(x, h(x))$ satisfies (41) for all such x . We wish to show that we can define h on a large enough neighborhood that it contains an x such that $(x, h(x))$ also satisfies (42).

LEMMA 20

Let f be defined as in (76), and suppose $(y_1, y_2) \in D$ satisfies conditions 1 through 4 of Theorem 19. Then, it is possible to define an open interval (α_1, α_2) that is the largest open interval satisfying the following conditions:

1. $y_1 \in (\alpha_1, \alpha_2)$, and there is a unique C^∞ function $h : (\alpha_1, \alpha_2) \mapsto \mathbb{R}$ such that $h(y_1) = y_2$, and $f(x, h(x)) = 0$ for all $x \in (\alpha_1, \alpha_2)$.
2. For all $x \in (\alpha_1, \alpha_2)$, $(x, h(x)) \in D$ and

$$\frac{\partial f(x_1, x_2)}{\partial x_2} \Big|_{(x_1, x_2) = (x, h(x))} < 0.$$

Proof

We will show that we can take (α_1, α_2) to be the union of all open intervals (β_1, β_2) containing y_1 which satisfy the above conditions.

Let (α_1, α_2) be so defined. We first note that (α_1, α_2) is non-empty. Indeed, by the implicit function theorem, there is a non-empty open interval (β_1, β_2) satisfying condition 1 of the lemma. By condition 3 of Theorem 19, and the assumption that $(y_1, y_2) \in D$, condition 2 of the lemma is satisfied at (y_1, y_2) . Since D is an open set, $f \in C^\infty(D)$ and $h \in C^\infty(\beta_1, \beta_2)$, condition 2 of the lemma is satisfied on a sufficiently small subset (γ_1, γ_2) of (β_1, β_2) . Thus, conditions 1 and 2 of the lemma are both satisfied on (γ_1, γ_2) , so (α_1, α_2) is non-empty.

Let $x \in (\alpha_1, \alpha_2)$. Then x is in some interval (β_1, β_2) satisfying the conditions of the lemma with some C^∞ function h_1 . We set $h(x) = h_1(x)$ and claim this is a valid definition of h satisfying the conditions of the lemma. To this end, we need to show that, if (γ_1, γ_2) is another interval containing y_1 and x , satisfying the conditions of the lemma with corresponding function h_2 , then $h_1(x) = h_2(x)$.

Suppose not. Observe that $h_1(y_1) = h_2(y_1)$. Since h_1 and h_2 are continuous, there exists $z \in [y_1, x)$ (or $(x, y_1]$ as the case may be) such that $h_1(u) = h_2(u)$ for all $u \in [y_1, z]$, but in every neighborhood of z , there is a u such that $h_1(u) \neq h_2(u)$. Now $f(z, h_1(z)) = 0$, and since

$$\frac{\partial}{\partial x_2} f(x_1, x_2) \Big|_{(x_1, x_2) = (z, h_1(z))} < 0$$

by condition 2 of the lemma, one can apply the implicit function theorem to the function $f(\cdot, \cdot)$ at the point $(z, h_1(z))$. One thus obtains that there is a unique function g on some neighborhood, U , of z such that $f(z, g(z)) = 0$. Both h_1 and h_2 satisfy this requirement, and thus equal g in a neighborhood of z . This contradicts the definition of z . In other words, we must have $h_1(x) = h_2(x)$. This shows that h is well-defined on (α_1, α_2) .

It is clear that, with h as defined above, $f(x, h(x)) = 0$ and condition 2 of the lemma is satisfied, for all $x \in (\alpha_1, \alpha_2)$. It remains to be shown that $h \in C^\infty((\alpha_1, \alpha_2))$. Observe from the way h was defined above that, if x is contained in (β_1, β_2) , then so is some open neighborhood of x . Hence, $h(x) = h_1(x)$ on this neighborhood, and so h inherits the C^∞ property from h_1 on a neighborhood of x . This is true for every $x \in (\alpha_1, \alpha_2)$. Hence, h is a C^∞ function on (α_1, α_2) . \square

LEMMA 21

Let (α_1, α_2) and h be as in Lemma 20 above. Then, $h(\alpha_2)$ defined as $\lim_{x \uparrow \alpha_2} h(x)$ exists and is finite. Hence, h is continuous on $(\alpha_1, \alpha_2]$.

Proof

Suppose $h(x)$ does not approach a limit as x increases to α_2 . Then, there exist constants $a_1 < a_2$, such that h is bigger than a_2 , and h is smaller than a_1 , for infinitely many values of $x \in (\alpha_1, \alpha_2)$. Since h is continuous on this interval, it must take the value $(a_1 + a_2)/2$ infinitely often. That is, $f(x, (a_1 + a_2)/2) = 0$ for infinitely many values of x . Then, these are all collinear solutions of $f(x_1, x_2) = 0$. This contradicts

the conclusion of Lemma 15. Therefore, $h(x)$ must approach a limit as x increases to α_2 .

Observe that, for every $x \in (\alpha_1, \alpha_2)$, $(x, h(x))$ solves $f(x_1, x_2) = 0$. Hence, by Lemma 16, $(x, h(x))$ is in the bounded set $\mathcal{S} \subset \mathbb{R}_+^2$ for every x in (α_1, α_2) . So the limit of $h(x)$ as x increases to α_2 is finite. Hence, h is continuous at α_2 . As already shown, h is also continuous, in fact C^∞ , on (α_1, α_2) . \square

Let α_1, α_2 and h be defined as in Lemmas 20 and 21 above. Define

$$f_i(x) = \frac{\partial f(x_1, x_2)}{\partial x_i} \Big|_{(x_1, x_2) = (x, h(x))}, \quad x \in (\alpha_1, \alpha_2), \quad i = 1, 2. \tag{78}$$

We see from the fact that $f \in C^\infty(D)$ and condition 2 of Lemma 20 that the above derivatives exist and are C^∞ functions on (α_1, α_2) , and also that $f_2(x) < 0$ on this interval.

By the definition of h in Lemma 18, $(x, h(x))$ satisfies (41) for all $x \in (\alpha_1, \alpha_2)$. We would like to show that $(x, h(x))$ also satisfies (42) for some $x \in [y_1, \alpha_2]$. We outline here the steps involved in showing this.

Suppose $(x, h(x))$ does not satisfy (42) for any $x \in [y_1, \alpha_2]$. In particular, (42) doesn't hold at (y_1, y_2) . Then, by condition 2 of Theorem 19 and continuity of f_1 on (α_1, α_2) , we have $f_1(x) < 0$ for all $x \in [y_1, \alpha_2]$. The same is true of $f_2(x)$ by Lemma 20. Now, if $f_2(\alpha_2) < 0$, then by applying the implicit function theorem at $(\alpha_2, h(\alpha_2))$ we can extend the definition of h to an interval around α_2 . This contradicts the maximality in the definition, in Lemma 20, of the interval (α_1, α_2) . Therefore, we conclude that $f_2(\alpha_2)$ is either zero or is undefined. The latter can happen only if $(\alpha_2, h(\alpha_2)) \in \partial D$ and $\mu_1\alpha_2 + \mu_2h(\alpha_2) = \mu_1 + \mu_2 - \sigma$. In either case above, we show a contradiction with f_1 and f_2 being negative on $[y_1, \alpha_2)$. The steps outlined above are made precise in the following.

Define

$$v_1 = \inf \left\{ x \in [y_1, \alpha_2) : f_1(x) = 0 \right\} \wedge \alpha_2, \quad v_2 = h(v_1), \tag{79}$$

where, as usual, the infimum of an empty set is defined to be $+\infty$, and $x \wedge y$ denotes $\min\{x, y\}$. We shall show that $v_1 < \alpha_2$. Hence, $f_1(v_1) = 0$, and (v_1, v_2) solves (41) and (42).

LEMMA 22

Let v_1 and v_2 be defined as above, and D as in (28). Then, for all $x \in [y_1, v_1]$, $F^*(\mu_1 - \mu_1x + \mu_2 - \mu_2h(x)) < 1$, and $(x, h(x)) \in D$. In addition, $f(v_1, v_2) = 0$.

Proof

Let h be defined as in Lemma 20. Then, by the implicit function theorem,

$$h'(x) = -f_1(x)/f_2(x), \quad x \in (\alpha_1, \alpha_2). \tag{80}$$

Now, $f_2 < 0$ on (α_1, α_2) by condition 2 of Lemma 20. Also, $f_1(y_1) \leq 0$ by condition 2 of Theorem 19. Hence, by the definition of v_1 and continuity of f_1 on (α_1, α_2) , $f_1(x) < 0$ for all $x \in [y_1, v_1]$. Thus, by (80), $h'(x) < 0$ for all $x \in [y_1, v_1]$. Consequently, for $z \in [y_1, v_1]$,

$$h(z) = h(y_1) + \int_{y_1}^z h'(x)dx \leq h(y_1), \tag{81}$$

where the equality is because $[y_1, v_1] \subset (\alpha_1, \alpha_2]$, h is differentiable on (α_1, α_2) , and h is continuous at α_2 . Now, $h(y_1)$ is equal to y_2 , which is less than 1 by condition 1 of Theorem 19. Hence,

$$h(z) < 1 \quad \forall \quad z \in [y_1, v_1]. \tag{82}$$

Suppose there exists $x \in [y_1, v_1]$ with $F^*(\mu_1 - \mu_1x + \mu_2 - \mu_2h(x)) \geq 1$. Then $\mu_1 - \mu_1x + \mu_2 - \mu_2h(x) \leq 0$. Note that $\mu_1 - \mu_1y_1 + \mu_2 - \mu_2y_2 > 0$ because $f(y_1, y_2) = 0$ and $y_1y_2 < 1$ by condition 1 of Theorem 19. Hence, by continuity of h on $[y_1, v_1] \subset (\alpha_1, \alpha_2]$, for all $\epsilon > 0$ sufficiently small, there exists $z \in (y_1, x)$ such that $\mu_1 - \mu_1z + \mu_2 - \mu_2h(z) = \epsilon$. Consequently,

$$f_1(z) = -\mu_1 \left. \frac{dF^*(s)}{ds} \right|_{s=\epsilon} - h(z).$$

If $\epsilon > 0$ is sufficiently small, then the first term on the right hand side is bigger than 1 by the stability assumption $\mu_1ET > 1$. The second term is less than 1 by (82). Hence, we have $z \in (y_1, x) \subseteq (y_1, v_1)$ such that $f_1(z) > 0$. This is a contradiction, since we saw above that $f_1(x) < 0$ for all $x \in (y_1, v_1)$. Therefore, we conclude that $F^*(\mu_1 - \mu_1x + \mu_2 - \mu_2h(x)) < 1$ for all $x \in [y_1, v_1]$. This proves the first claim of the lemma.

We note as a consequence that

$$\mu_1 - \mu_1x + \mu_2 - \mu_2h(x) > 0 \geq \sigma \quad \forall \quad x \in [y_1, v_1], \tag{83}$$

where σ denotes the left limit of finiteness of $F^*(s)$. Recall from the corollary to Lemma 16, and the fact that $f(x, h(x)) = 0$ for all $x \in (\alpha_1, \alpha_2)$, that

$$x \geq \delta, \quad h(x) \geq \delta \quad \forall \quad x \in (\alpha_1, \alpha_2).$$

Since h is continuous at α_2 by Lemma 21, the above inequalities hold on $(\alpha_1, \alpha_2]$, and in particular on $[y_1, v_1]$, which is a subset of this interval. It is now clear from (83) and the definition of D in (28) that $(x, h(x))$ is in D for all $x \in [y_1, v_1]$. This completes the proof of the second claim in the lemma.

If $v_1 < \alpha_2$, then $f(v_1, v_2) = 0$ because $f(x, h(x)) = 0$ for all $x \in (\alpha_1, \alpha_2)$. If $v_1 = \alpha_2$, then $f(x, h(x))$ is a continuous function of x at α_2 because h is continuous at α_2 , $f \in C^\infty(D)$, and $(v_1, v_2) \in D$ as proved above. Hence, $f(v_1, v_2) = 0$. \square

Proof of Theorem 19

Let (v_1, v_2) be defined as in (79). We show that (v_1, v_2) solves (41) and (42), and satisfies (43). We begin by showing that $v_1 < \alpha_2$.

Suppose otherwise. Then, by (79), $v_1 = \alpha_2$ and we have from Lemma 22 that $(x, h(x))$ is in D for all $x \in (\alpha_1, \alpha_2]$. Hence, f_1 and f_2 are defined and continuous on this interval. Therefore, by condition 2 of Lemma 20, $f_2(\alpha_2) \leq 0$.

Suppose $f_2(\alpha_2) < 0$. Applying the implicit function theorem at the point $(\alpha_2, h(\alpha_2))$, we see that there is a neighborhood U of α_2 , a neighborhood V of $h(\alpha_2)$, and a unique C^∞ function $g : U \rightarrow V$ such that $f(x, g(x)) = 0$ for all $x \in U$. Also, $(x, g(x)) \in D$ for all $x \in U$ if we choose U small enough, because $(\alpha_2, h(\alpha_2)) \in D$, and D is open. It is also clear that, by choosing U sufficiently small, we can get

$$\frac{\partial}{\partial x_2} f(x_1, x_2) \Big|_{(x_1, x_2) = (x, g(x))} < 0 \quad \forall \quad x \in U.$$

Thus, g satisfies conditions 1 and 2 of Lemma 20 on U . Also, by the uniqueness of g , it agrees with h on $U \cap (\alpha_1, \alpha_2]$. Extend the definition of h to $U \cup (\alpha_1, \alpha_2]$ by defining it to be equal to g on U . Then h satisfies the conditions of Lemma 20 on the open interval $U \cup (\alpha_1, \alpha_2]$ that strictly includes (α_1, α_2) . This contradicts the maximality of (α_1, α_2) defined in Lemma 20. Hence, we cannot have $f_2(\alpha_2) < 0$.

We have thus shown that, if $v_1 = \alpha_2$, then $f_2(\alpha_2) = 0$. Observe from condition 2 of Theorem 19, the definition of v_1 in (79), and the assumption above that $v_1 = \alpha_2$, that $f_1(x) < 0$ for all $x \in [y_1, \alpha_2)$. Hence, $f_1(\alpha_2) \leq 0$.

Suppose first that $f_1(\alpha_2) < 0$. Since $f_2(\alpha_2) = 0$, we have

$$\lim_{x \uparrow \alpha_2} h'(x) = \lim_{x \uparrow \alpha_2} -f_1(x)/f_2(x) = -\infty$$

by the implicit function theorem. Let \hat{y}_2 be as specified in condition 4 of Theorem 19, so that $f(y_1, \hat{y}_2) = 0$. Observe from Lemma 22 and the assumption that $v_1 = \alpha_2$ that $f(\alpha_2, h(\alpha_2)) = 0$. We shall show that there is a point (z_1, z_2) on the straight line joining (y_1, \hat{y}_2) and $(\alpha_2, h(\alpha_2))$ such that $f(z_1, z_2) = 0$. Define

$$g(x) = \hat{y}_2 + \frac{h(\alpha_2) - \hat{y}_2}{\alpha_2 - y_1} (x - y_1).$$

That is, $(x, g(x))$ denotes the straight line joining (y_1, \hat{y}_2) and $(\alpha_2, h(\alpha_2))$. Now, for $\delta > 0$,

$$g(\alpha_2 - \delta) = g(\alpha_2) - \delta \cdot \frac{h(\alpha_2) - \hat{y}_2}{\alpha_2 - y_1},$$

whereas, by the intermediate value theorem,

$$\exists x \in (\alpha_2 - \delta, \alpha_2) : h(\alpha_2 - \delta) = h(\alpha_2) - \delta \cdot h'(x).$$

Since $g(\alpha_2) = h(\alpha_2)$ and $\lim_{x \uparrow \alpha_2} h'(x) = -\infty$, it follows that for $\delta > 0$ sufficiently small,

$$h(\alpha_2 - \delta) > g(\alpha_2 - \delta).$$

But $g(y_1) = \hat{y}_2$, $h(y_1) = y_2$, and so, by condition 4 in Theorem 19, $h(y_1) < g(y_1)$. Since g and h are continuous functions on $[y_1, \alpha_2]$, the last two inequalities imply that there is $z \in (y_1, \alpha_2 - \delta)$ such that $g(z) = h(z)$. Then, (y_1, \hat{y}_2) , $(z, h(z))$ and $(\alpha_2, h(\alpha_2))$ constitute three collinear solutions of $f(x_1, x_2) = 0$. This contradicts Lemma 15, implying that we cannot have $f_1(\alpha_2) < 0$.

Next, suppose $f_1(\alpha_2) = 0$. Let $v_1 = \alpha_2$, $v_2 = h(\alpha_2)$, and define \mathbf{p} by $p_{jk} = a_{jk} v_1^{j-1} v_2^{k-1}$, where the a_{jk} are given in (1). Now $f(v_1, v_2) = 0$ by Lemma 22, so $F^*(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2) = v_1 v_2$ and, by (22), $\sum_{j,k=0}^{\infty} p_{jk} = 1$. Hence, \mathbf{p} is a probability distribution. Also, $f_1(v_1) = 0$ by supposition, and we showed above that $f_2(v_1) = 0$. Since $(v_1, v_2) \in D$, we have from (29) and (30) that

$$\sum_{j,k=0}^{\infty} j p_{jk} = 1, \quad \sum_{j,k=0}^{\infty} k p_{jk} = 1.$$

Therefore,

$$\begin{aligned} D(\mathbf{p}; \mathbf{a}) &= \sum_{j,k=0}^{\infty} p_{jk} \log \frac{p_{jk}}{a_{jk}} \\ &= \sum_{j,k=0}^{\infty} \left[(j-1)p_{jk} \log v_1 + (k-1)p_{jk} \log v_2 \right] \\ &= 0. \end{aligned}$$

Hence, by the properties of the Kullback-Leibler distance, $\mathbf{p} \equiv \mathbf{a}$, that is, v_1 and v_2 are both equal to one. But we have from (81) that $v_2 = h(v_1) \leq y_2$, and hence from condition 1 of Theorem 19 that $v_2 < 1$. Thus, by contradiction, we cannot have $f_1(\alpha_2) = 0$.

To recapitulate, we showed that, if $v_1 = \alpha_2$, then $f_2(v_1) = 0$, and $f_1(v_1) \leq 0$. We showed that either equality or strict inequality in the last expression contradicts $f_2(v_1) = 0$. Thus, we cannot have $v_1 = \alpha_2$. It follows from (79) that $v_1 < \alpha_2$ and $f_1(v_1) = 0$. In other words, (v_1, v_2) solves (42). We also see from (79) that $v_1 \geq y_1$, and from (81) that $v_2 = h(v_1)$ is no larger than $y_2 = h(y_1)$. Since $v_1 \in (\alpha_1, \alpha_2)$, $f_2(v_1) < 0$ by condition 2 of Lemma 20. Thus, (v_1, v_2) is seen to satisfy (77). Finally, (v_1, v_2) is in D and satisfies (41) by Lemma 22. This completes the proof of the theorem. \square

We derive below a couple of additional facts needed to complete the proof of Lemma 11.

LEMMA 23

Let g be defined on the positive reals as $g(x) = \nu_1/x$, and h as in Lemma 20. Then,

$$g(x) = h(x) \quad \Rightarrow \quad x = \nu_1 \quad \text{or} \quad x = \mu_2/\mu_1.$$

Proof

By definition of h , $f(x, h(x)) = 0$. Therefore, if $g(x) = h(x)$, then $f(x, \nu_1/x) = 0$. Hence, by (76), $F^*(\mu_1 - \mu_1 x + \mu_2 - \mu_2 \nu_1/x) = \nu_1$. But $F^*(\mu_1 - \mu_1 \nu_1) = \nu_1$, and F^* is one to one. Therefore,

$$\mu_1 - \mu_1 x + \mu_2 - \mu_2 \frac{\nu_1}{x} = \mu_1 - \mu_1 \nu_1.$$

The above is a quadratic in x , whose only solutions are $x = \nu_1$ and $x = \mu_2/\mu_1$. \square

LEMMA 24

Suppose neither (32) nor (36) holds. Then, the conditions of Theorem 19 are satisfied for (y_1, y_2) equal to either $(1, \nu_2)$ or $(\mu_2/\mu_1, \mu_1 \nu_1/\mu_2)$.

Proof

Suppose neither (32) nor (36) holds. Let $(y_1, y_2) = (1, \nu_2)$. It is clear that (y_1, y_2) is in D , which was defined in (28). With $\hat{y}_2 = 1$, conditions 1 and 4 of Theorem 19 are straightforward to verify. Condition 2 follows from Lemma 18, and condition 3 from Lemma 1.

Next, let $y_1 = \mu_2/\mu_1$ and $y_2 = \mu_1 \nu_1/\mu_2$, and observe that $(y_1, y_2) \in D$. Taking $\hat{y}_2 = \mu_1/\mu_2$, it is easy to see that conditions 1 and 4 of Theorem 19 are satisfied. The left-hand side of condition 2 evaluates to

$$\frac{d}{dx_1} F^*(\mu_1 - \mu_1 x_1) \Big|_{x_1=\nu_1} - \frac{\mu_1 \nu_1}{\mu_2}$$

and is no bigger than zero since (36) was assumed not to hold. The left-hand side in condition 3 is equal to

$$\frac{\mu_2}{\mu_1} \left[\frac{d}{dx_1} F^*(\mu_1 - \mu_1 x_1) \Big|_{x_1=\nu_1} - 1 \right]$$

and is strictly less than zero by Lemma 1. Thus, the conditions of Theorem 19 are satisfied. □

We now make use of the above lemmas and Theorem 19 to complete the proof of Lemma 11.

Proof of Lemma 11

Suppose neither (32) nor (36) holds. We show that (41) and (42) have a solution $(v_1, v_2) \in D$ satisfying (43).

Suppose first that $\mu_1 > \mu_2$. Let $(y_1, y_2) = (1, \nu_2)$, so that, by Lemma 24, the conditions of Theorem 19 are satisfied. Let α_1, α_2 and h be defined as in Lemma 20, corresponding to $(y_1, y_2) = (1, \nu_2)$. Let v_1, v_2 be as in (79). Then, by Theorem 19, (v_1, v_2) is in D and solves (41) and (42) subject to (77). Comparing (77) to (43), it remains to verify that $\nu_1 \leq v_1 v_2 < 1$. Observe from Lemma 22 and (79) that $F^*(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2)$ is less than 1 and $f(v_1, v_2)$ is equal to zero. Hence, $v_1 v_2 < 1$ by (76). We show below that $v_1 v_2 \geq \nu_1$.

Define $g(x) = \nu_1/x$. Since $\mu_2/\mu_1 < 1$ by assumption, and $\nu_1 < 1$, we see from Lemma 23 that $g(x)$ is not equal to $h(x)$ for any $x > 1$. Now $g(1) = \nu_1$, whereas $h(1) = \nu_2$ because $(y_1, y_2) = (1, \nu_2)$. Therefore, by Lemma 2, $g(1) < h(1)$. Now g and h are continuous on $[1, v_1] \subset (\alpha_1, \alpha_2)$, $g(1)$ is less than $h(1)$, and $g(x)$ is not equal to $h(x)$ for any $x > 1$. Therefore, $g(v_1) < h(v_1)$. Since $g(v_1) = \nu_1/v_1$, while $h(v_1) = v_2$ by (79), we conclude that $v_1 v_2 > \nu_1$.

Suppose next that $\mu_1 \leq \mu_2$. Taking $y_1 = \mu_2/\mu_1$ and $y_2 = \mu_1 \nu_1/\mu_2$, we have from Lemma 24 that the conditions of Theorem 19 are satisfied. Let α_1, α_2 and h be defined as in Lemma 20, corresponding to the above choice of (y_1, y_2) . Let v_1, v_2 be as in (79). Then, by Theorem 19, (v_1, v_2) is in D and solves (41) and (42) subject to (77). Comparing (77) to (43), it remains to verify that $\nu_1 \leq v_1 v_2 < 1$. Observe from Lemma 22 and (79) that $F^*(\mu_1 - \mu_1 v_1 + \mu_2 - \mu_2 v_2)$ is less than 1 and $f(v_1, v_2)$ is equal to zero. Hence, $v_1 v_2 < 1$ by (76). We show below that $v_1 v_2 \geq \nu_1$.

By the definition of h , we have $h(\mu_2/\mu_1) = \mu_1 \nu_1/\mu_2$. If $v_1 = \mu_2/\mu_1$, then, by (79), $v_1 v_2 = \nu_1$, and we are done. If not, we argue as follows. Let g be defined on the positive reals as $g(x) = \nu_1/x$. Then $g(\mu_2/\mu_1) = h(\mu_2/\mu_1)$. In addition,

$$g' \left(\frac{\mu_2}{\mu_1} \right) = - \left(\frac{\mu_1}{\mu_2} \right)^2 \nu_1, \tag{84}$$

where $g'(x)$ denotes the derivative of g at x . Also, by the implicit function theorem, and the definition of h in Lemma 21,

$$\begin{aligned}
 h'\left(\frac{\mu_2}{\mu_1}\right) &= -\frac{f_1(\mu_2/\mu_1)}{f_2(\mu_2/\mu_1)} \\
 &= -\frac{\frac{d}{dx_1} F^*(\mu_1 - \mu_1 x_1) \Big|_{x_1=\nu_1} - \frac{\mu_1 \nu_1}{\mu_2}}{\frac{\mu_2}{\mu_1} \frac{d}{dx_1} F^*(\mu_1 - \mu_1 x_1) \Big|_{x_1=\nu_1} - \frac{\mu_2}{\mu_1}} \\
 &= -\left(\frac{\mu_1}{\mu_2}\right)^2 \nu_1 \frac{\frac{\mu_2}{\mu_1 \nu_1} \frac{d}{dx_1} F^*(\mu_1 - \mu_1 x_1) \Big|_{x_1=\nu_1} - 1}{\frac{d}{dx_1} F^*(\mu_1 - \mu_1 x_1) \Big|_{x_1=\nu_1} - 1}. \tag{85}
 \end{aligned}$$

Observe from Lemma 1 and the assumptions that $\mu_1 \leq \mu_2$ and that (36) doesn't hold, that both the numerator and the denominator in the last term above are negative. Furthermore, the numerator is smaller in absolute value than the denominator since $\mu_2 \geq \mu_1$ and $\nu_1 < 1$. We thus see from (84) and (85) that $g'(\mu_2/\mu_1) < h'(\mu_2/\mu_1)$. Since $g(\mu_2/\mu_1)$ was seen to be equal to $h(\mu_2/\mu_1)$, we have

$$g\left(\frac{\mu_2}{\mu_1} + \delta\right) < h\left(\frac{\mu_2}{\mu_1} + \delta\right) \quad \text{for all } \delta > 0 \text{ sufficiently small.}$$

Hence observe from the continuity of g and h on $[\mu_2/\mu_1, \nu_1] \subset (\alpha_1, \alpha_2)$ that, if $g(\nu_1) \geq h(\nu_1)$, then $g(x) = h(x)$ for some $x \in (\mu_2/\mu_1, \nu_1]$. Then $x > \nu_1$ as well, since it was assumed that $\mu_2 \geq \mu_1$. Thus, we have arrived at a contradiction of Lemma 23. We therefore conclude that $g(\nu_1) < h(\nu_1)$. But $h(\nu_1) = \nu_2$ by (79), and $g(\nu_1) = \nu_1/\nu_1$. Hence, $\nu_1 \nu_2 > \nu_1$.

The proof of Lemma 11 is now complete. □

References

- [1] V. Anantharam, The optimal buffer allocation problem, *IEEE Trans. Inform. Theory* 35 (1989) 721–725.
- [2] V. Anantharam and A. Ganesh, Correctness within a constant of an optimal buffer allocation rule of thumb, *IEEE Trans. Info. Theory* 40 (1994) 871–882.
- [3] V. Anantharam, P. Heidelberger and P. Tsoucas, Analysis of rare events in continuous time Markov chains via time reversal and fluid approximations, *IBM Research Report RC 16280* (1990).
- [4] S. Asmussen, *Applied Probability and Queues* (Wiley, 1987).
- [5] D. Bertsimas, I. Paschalidis and J. Tsitsiklis, On the large deviations behaviour of acyclic networks of G/G/1 queues, submitted to *Ann. Appl. Prob.*
- [6] R. Blahut, *Principles and Practice of Information Theory* (Addison-Wesley, 1987).

- [7] C.S. Chang, Stability, queue length and delay of deterministic and stochastic queueing networks, *IEEE Trans. Autom. Control* 39 (1994) 913–931.
- [8] C.S. Chang, Sample path large deviations andintree networks, *IBM Research Report RC 19118* (1993).
- [9] C.S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin, Effective bandwidth and fast simulation of ATM intree networks, *Perform. Eval.* 20 (1994) 45–66.
- [10] G. de Veciana and J. Walrand, Effective bandwidths: call admission, traffic policing and filtering for ATM networks, to appear in *Queueing Systems*.
- [11] G. de Veciana and J. Walrand, Decoupling bandwidths for networks: a decomposition approach to resource management, Memorandum No. UCB/ERL M93/50, University of California, Berkeley (1993).
- [12] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications* (Jones and Bartlett, 1993).
- [13] N. Duffield and N. O’Connell, Large deviations and overflow probabilities for the general single server queue, with applications, to appear in *Proc. Camb. Phil. Soc.*
- [14] M.R. Frater, T.M. Lennon and B.D.O. Anderson, Optimally efficient estimation of the statistics of rare events in queueing networks, *IEEE Trans. Autom. Control* 36 (1991) 1395–1405.
- [15] A. Ganesh, Stationary tail probabilities in exponential server tandems with renewal arrivals, Ph.D. Thesis, Cornell University (1995).
- [16] A. Ganesh and V. Anantharam, Stationary tail probabilities in exponential server tandems with renewal arrivals, in: *Stochastic Networks, IMA Volumes in Mathematics and its Applications*, Vol. 71, eds. F. P. Kelly and R. J. Williams (Springer, 1995).
- [17] P.W. Glynn and W. Whitt, Logarithmic asymptotics for steady-state tail probabilities in a single-server queue, to appear in *J. Appl. Prob.*
- [18] R.M. Loynes, The stability of queues with non-independent inter-arrival and service times, *Proc. Camb. Phil. Soc.* 58 (1962) 497–520.
- [19] J.E. Marsden, *Elementary Classical Analysis*, (W.H. Freeman, 1993).
- [20] A. Mogulskii, Large deviations for trajectories of multi-dimensional random walks, *Theor. Prob. Appl.* 21 (1976) 300–315.
- [21] N. O’Connell, Large deviations in queueing networks, *DIAS Technical Report DIAS-APG-9413*.
- [22] S. Parekh and J. Walrand, A quick simulation method for excessive backlogs in networks of queues, *IEEE Trans. Autom. Control* 34 (1989) 54–66.
- [23] P. Tsoucas, Rare events in series of queues, *J. Appl. Prob.* 29 (1992) 168–175.