MATHEMATICS OF OPERATIONS RESEARCH
Vol. 10, No. 2, May 1985
Printed in U.S.A.

ASYMPTOTIC ANALYSIS OF THE TIME DEPENDENT M/M/1 QUEUE*

WILLIAM A. MASSEY

AT&T Bell Laboratories

Using operator analytic techniques, we develop a nonstationary Markovian queueing theory starting with the **M(t)/M(t)/1** queue. We employ an asymptotic approach quite different from the usual large time analysis. Instead, we *uniformly accelerate* the queue length process. That is, we divide the arrival and service rate by a common parameter ϵ. Then, for a fixed time interval, we consider the asymptotics for the distribution, mean, and variance of the queue length process as ϵ goes to zero. The effects of ϵ can be quite different for the given time interval. This gives us a dynamic asymptotic behavior for the queue length process. We can formulate a time dependent traffic intensity parameter that determines when the process is asymptotically stable and when it is asymptotically unstable.

Introduction. Employing operator analytic methods, we develop a nonstationary Markovian queueing theory. By doing so, we establish a theoretical basis to complement and extend the pioneering work of Newell [6], as well as the results of Keller [3], Rider [7], and Rothkopf and Oren [8]. Stationary queueing theory begins with the development of the **M/M/1** queue, so we will consider its time varying analog and refer to it as **M(t)/M(t)/1**. Specifying some initial time t_0, we let $Q(t_0, t)$ be the Markov process that equals the number of people in the system at time $t > t_0$. The process is characterized by the arrival and service rates $\lambda(t)$ and $\mu(t)$ respectively. We will assume that $\lambda(t)$ and $\mu(t)$ are smooth, positive, real analytic functions of time.

Let $p_n(t_0, t) = \Pr\{ Q(t_0, t) = n\}$, then we will have the $p_n(t_0, t)$'s solving the following set of birth and death equations:

$$\frac{d}{dt} p_n(t_0, t) = \lambda(t) p_{n-1}(t_0, t) + \mu(t) p_{n+1}(t_0, t) - (\lambda(t) + \mu(t)) p_n(t_0, t) \qquad (1)$$

when $n \geqslant 1$, otherwise

$$\frac{d}{dt} p_0(t_0, t) = \mu(t) p_1(t_0, t) - \lambda(t) p_0(t_0, t)$$

and, finally, $p_n(t_0, t_0) = \delta_{n, n_0}$.

The type of asymptotics that we will apply to the **M(t)/M(t)/1** system will be called *uniform acceleration*. That is, we introduce a small positive parameter ϵ and look at an associated queueing process $Q(t_0, t; \epsilon)$ with arrival intensity $\lambda(t)/\epsilon$, and service intensity $\mu(t)/\epsilon$. Its queue length distribution then satisfies the following set of birth and death equations:

$$\epsilon \frac{\partial}{\partial t} p_n(t_0, t; \epsilon) = \lambda(t) p_{n-1}(t_0, t; \epsilon) + \mu(t) p_{n+1}(t_0, t; \epsilon) - (\lambda(t) + \mu(t)) p_n(t_0, t; \epsilon) \quad (2)$$

*Received November 15, 1982; revised December 23, 1983.
AMS 1980 subject classification. Primary: 60K25. Secondary: 90B22.
OR/MS Index 1978 subject classification. Primary: 695 Queues/nonstationary. Secondary: 705 Queues/Markovian.
Key words. Markovian queues, nonstationary queues, asymptotic stability, time-ordered exponentials, stochastic dominance.

0364-765X/85/1002/0305$01.25

when $n > 1$, otherwise

$$\epsilon \frac{\partial}{\partial t} p_0(t_0, t; \epsilon) = \mu(t) p_1(t_0, t; \epsilon) - \lambda(t) p_0(t_0, t; \epsilon)$$

and, again, $p_n(t_0, t_0; \epsilon) = \delta_{n,n_0}$. For each time $t > t_0$, we do an asymptotic analysis on the mean, variance and probability distribution of the queue length process as ϵ approaches zero.

It is a valid question to ask why this approach is preferred to the *large time* asymptotic method used on the **M/M/1** queue. The answer is twofold. First, λ and μ are now functions of time so at some specified time it would be unsatisfactory to approximate the mean queue length, say, by future values of λ and μ. Second, if this approach is used on the **M/M/1** queue, then letting $t \to \infty$ in (1) is equivalent to letting $\epsilon \downarrow 0$ in (2). Consequently, these two approaches coincide for the **M/M/1** queue but it is the latter that generalizes better to the nonstationary **M(t)/M(t)/1** system.

The equivalence of these two methods for the **M/M/1** queue shows that the uniform acceleration method is the nonstationary analog of large time analysis for a stationary process. Whereas only one asymptotic distribution (geometric or defective) is associated with a particular **M/M/1** queue, for **M(t)/M(t)/1**, we associate a different asymptotic distribution for each $t > t_0$. It follows that a traffic intensity parameter can be defined indicating at each time t whether the queue is oversaturated (unstable) or undersaturated (stable) when accelerated. This is established by the following theorem, which we will prove in §8:

THEOREM 1. *Given an* **M(t)/M(t)/1** *process with* $\lambda(t)$ *and* $\mu(t)$, *consider the following quantity*:

$$\rho^*(t_0, t) = \sup_{t_* \in (t_0, t)} \frac{\int_{t_*}^{t} \lambda(s)\, ds}{\int_{t_*}^{t} \mu(s)\, ds}.$$

Then $Q(t_0, t)$ *is undersaturated iff* $\rho^*(t_0, t) < 1$. *In this case, for each nonnegative integer* k, *there exist* k l_1-*sequences* $\{p_n^{(1)}(t)\}, \ldots, \{p_n^{(k)}(t)\}$ *such that each sums to zero and*

$$\Pr\{Q(t_0, t; \epsilon) = n\} = (1 - \rho(t))\rho(t)^n + \epsilon p_n^{(1)}(t) + \cdots + \epsilon^k p_n^{(k)}(t) + O(\epsilon^{k+1})$$

where $\rho(t) = \lambda(t)/\mu(t)$; *otherwise* $\lim_{\epsilon \downarrow 0} \Pr\{Q(t_0, t; \epsilon) = n\}$ *does not represent, a probability distribution. Also, if* $\rho^*(t_0, t) > 1$, *then* $\Pr\{Q(t_0, t; \epsilon) = n\} \simeq 0$ *where* $a(\epsilon) \simeq b(\epsilon)$ *if* $a(\epsilon) - b(\epsilon) = o(\epsilon^k)$ *for all positive integers* k.

For any given **M/M/1** process, the traffic intensity parameter ρ is fixed for all time. By contrast, $\rho^*(t_0, t)$ the parameter for the **M(t)/M(t)/1** queue, is a function of time. Therefore, a given **M(t)/M(t)/1** process may progress in time from undersaturation to oversaturation and back again. These are the two basic modes for the process and, after the initial layer, they are patched together by two types of transition layers. The first describes the onset of rush hour. By onset, we mean that the process sees the start, but not necessarily the continuation, of oversaturation. The second type of layer describes the end of rush hour, i.e., the transition from oversaturation to undersaturation. The recurring regions and layers of time are characterized by $\rho^*(t_0, t)$ as follows:

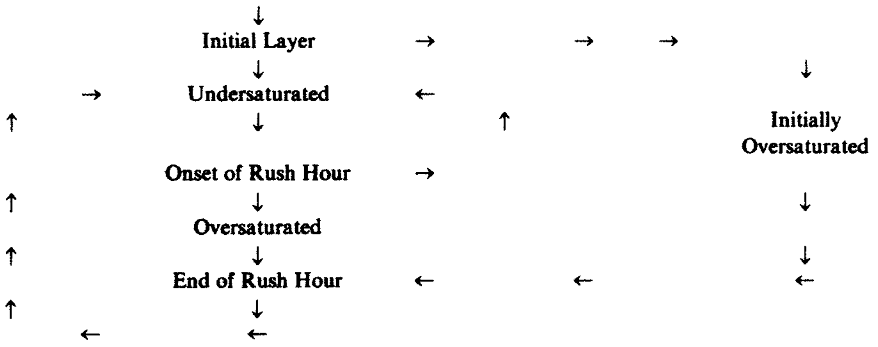
$$\rho^*(t_0, t) < 1 \Rightarrow \text{Undersaturated},$$

$$\rho^*(t_0, t) > 1 \Rightarrow \text{Oversaturated},$$

$$\rho(t) = \rho^*(t_0, t) = 1 \Rightarrow \text{Onset of Rush Hour},$$

$$\rho(t) < \rho^*(t_0, t) = 1 \Rightarrow \text{End of Rush Hour}.$$

The possible evolution of an **M(t)/M(t)/1** process can be described by the following flow chart:



Note that there are two possibilities when the process leaves the initial layer or leaves the onset of rush hour. Otherwise, it loops through a rush hour cycle.

Starting from the top of this diagram, §§1 through 6 will deal with each of these regions or layers in descending order. We will introduce the analytic machinery in §0. Only enough, however, to derive the results that will be discussed in §§1 through 6. §7 will rigorously justify the machinery employed. In the subsequent sections, we derive from basic principles useful properties of the **M(t)/M(t)/1** queue length process. §8 will prove Theorem 1 above and justify our formulation of the traffic intensity parameter. §9 will derive the time reversal formulas for the mean and variance. In §10, we will develop a criterion for stochastic dominance. Finally, in §11, we have an appendix for calculations relevant to §2.

0. Preliminaries. We now present the basic analytic machinery involved in the analysis of the **M(t)/M(t)/1** process. Let

$$\mathbf{p}(t_0,t;\epsilon)=\left[\,p_0(t_0,t;\epsilon)\;p_1(t_0,t;\epsilon)\;\ldots\,\right] \tag{0.1}$$

then $\mathbf{p}(t_0,t;\epsilon)$ belongs to l_1, the Banach space of absolutely summable sequences. We define the vectors $\mathbf{e}_{n_0}$, $\mathbf{f}_{n_0}$ and $\mathbf{1}$ as follows:

$$\mathbf{e}_{n_0}=[0\cdots 0\,1\,0\cdots],\qquad \mathbf{f}_{n_0}=[1\cdots 1\,0\,0\cdots],\qquad \mathbf{1}=[1\cdots 1\,1\,1\cdots], \tag{0.2}$$

where $\mathbf{e}_{n_0}$ has a one only in the n_0 place, $\mathbf{f}_{n_0}$ has ones up to the n_0-1 place and $\mathbf{1}$ is a vector of ones. The birth and death equations (2) can be written compactly as

$$\epsilon\frac{\partial}{\partial t}\mathbf{p}(t_0,t;\epsilon)=\mathbf{p}(t_0,t;\epsilon)\mathbf{A}(t),\qquad \mathbf{p}(t_0,t_0;\epsilon)=\mathbf{e}_{n_0},\qquad \text{where} \tag{0.3}$$

$$\mathbf{A}(t)=\begin{bmatrix} -\lambda(t) & \lambda(t) & 0 & 0 & \cdots \\ \mu(t) & -(\lambda(t)+\mu(t)) & \lambda(t) & 0 & \cdots \\ 0 & \mu(t) & -(\lambda(t)+\mu(t)) & \lambda(t) & \cdots \\ 0 & 0 & \mu(t) & -(\lambda(t)+\mu(t)) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots \end{bmatrix}.$$

It turns out that $\mathbf{p}^*(t_0,t;\epsilon)$, where $p_n^*(t_0,t;\epsilon)=\Pr\{\,Q(t_0,t;\epsilon)>n\}$, will be the more useful quantity to use. It satisfies the equations

$$\epsilon \frac{\partial}{\partial t} \mathbf{p}^*(t_0,t;\epsilon) = \mathbf{p}^*(t_0,t;\epsilon)\mathbf{A}^*(t) + \lambda(t)\mathbf{c}_0, \qquad \mathbf{p}^*(t_0,t_0;\epsilon) = \mathbf{f}_{n_0}, \qquad \text{where} \tag{0.4}$$

$$\mathbf{A}^*(t) = \begin{bmatrix} -(\lambda(t)+\mu(t)) & \lambda(t) & 0 & 0 & \cdots \\ \mu(t) & -(\lambda(t)+\mu(t)) & \lambda(t) & 0 & \cdots \\ 0 & \mu(t) & -(\lambda(t)+\mu(t)) & \lambda(t) & \cdots \\ 0 & 0 & \mu(t) & -(\lambda(t)+\mu(t)) & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \end{bmatrix}.$$

Let us denote the fundamental solution for the homogeneous part of (0.4) by $\mathbf{M}^*(t_0,t;\epsilon)$. It solves the equations

$$\epsilon \frac{\partial}{\partial t} \mathbf{M}^*(t_0,t;\epsilon) = \mathbf{M}^*(t_0,t;\epsilon)\mathbf{A}^*(t), \qquad M^*(t_0,t_0;\epsilon) = \mathbf{I}. \tag{0.5}$$

It is tempting but incorrect to think of $\mathbf{M}^*(t_0,t;\epsilon)$ as $\exp(\epsilon^{-1}\int_{t_0}^{t}\mathbf{A}^*(s)\,ds)$. This view, however, turns out to be useful in characterizing $\mathbf{M}^*(t_0,t;\epsilon)$. Witness the following theorem adapted from T. Kato for the solution of a time inhomogeneous evolution equation (see Yoshida [9, p. 431]).

THEOREM (Kato) 0.1. *Let $T = t - t_0$, then*

$$\mathbf{M}^*(t_0,t;\epsilon) = \lim_{N\to\infty}\left(\prod_{i=0}^{[NT]-1} \exp\left(\frac{1}{\epsilon N}\mathbf{A}^*\left(t_0+\frac{i}{N}\right)\right)\right.$$
$$\left.\times \exp\left(\frac{T-[NT]/N}{\epsilon}\mathbf{A}^*\left(t_0+\frac{[NT]}{N}\right)\right)\right)$$

where the product is ordered from left to right.

The interesting relationship to note here is that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=0}^{[NT]-1}\mathbf{A}^*\left(t_0+\frac{i}{N}\right) + \left(T - \frac{[NT]}{N}\right)\mathbf{A}^*\left(t_0+\frac{[NT]}{N}\right) = \int_{t_0}^{t_0+T}\mathbf{A}^*(s)\,ds.$$

Theorem 0.1 turns out to be crucial in characterizing the asymptotic behavior of $\mathbf{M}^*(t_0,t;\epsilon)$.

A bounded operator on l_1 is said to be *positive* if it maps positive vectors to positive vectors. Given two bounded operators $\mathbf{A}$ and $\mathbf{B}$ on l_1, we say that $\mathbf{A} < \mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is a positive operator. If we let $\mathbf{L}$ denote the left shift operator on row vectors in l_1 and $\mathbf{R}$ the right shift one, then $\mathbf{A}^*(t)$ can be written as follows: $\mathbf{A}^*(t) = \lambda(t)\mathbf{R} + \mu(t)\mathbf{L} - (\lambda(t)+\mu(t))\mathbf{I}$. It is clear that $\mathbf{R}$ and $\mathbf{L}$ are positive operators and

$$\exp(\mathbf{A}^*(t)) = e^{-(\lambda(t)+\mu(t))}\exp(\lambda(t)\mathbf{R} + \mu(t)\mathbf{L}).$$

Hence $\exp(\mathbf{A}^*(t))$ is a positive operator as well or $\exp(\mathbf{A}^*(t)) > 0$. By Theorem 0.1, it follows immediately that $\mathbf{M}^*(t_0,t;\epsilon) > 0$. The fundamental result of §7 will be to derive the following operator upper bound for $\mathbf{M}^*(t_0,t;\epsilon)$:

$$\mathbf{M}^*(t_0,t;\epsilon) < e^{-(\lambda_*+\mu_*)/\epsilon}\left(I_0\left(\frac{2}{\epsilon}\sqrt{\lambda_*\mu_*}\right)\mathbf{I} + \sum_{k=1}^{\infty}\left(\rho_*^{k/2}\mathbf{R}^k + \rho_*^{-k/2}\mathbf{L}^k\right)I_k\left(\frac{2}{\epsilon}\sqrt{\lambda_*\mu_*}\right)\right) \tag{0.6}$$

where $\lambda_* = \int_{t_0}^{t}\lambda(s)\,ds$, $\mu_* = \int_{t_0}^{t}\mu(s)\,ds$, $\rho_* = \lambda_*/\mu_*$, and $I_k(\cdot)$ is the kth modified Bessel function. From this it follows that $\mathbf{M}^*(t_0,t;\epsilon)$ converges to zero in the strong operator toplogy as $\epsilon\downarrow 0$, given $\int_{t_0}^{t}\lambda(s)\,ds < \int_{t_0}^{t}\mu(s)\,ds$. Other useful properties of $\mathbf{M}^*(t_0,t;\epsilon)$ are

$$\epsilon\frac{\partial}{\partial t_0}\mathbf{M}^*(t_0,t;\epsilon) = -\mathbf{A}^*(t_0)\mathbf{M}^*(t_0,t;\epsilon),$$

for all $s \in (t_0,t)$

$$\mathbf{M}^*(t_0,s;\epsilon)\mathbf{M}^*(s,t;\epsilon) = \mathbf{M}^*(t_0,t;\epsilon), \quad \text{and} \tag{0.7}$$

$$|\mathbf{M}^*(t_0,t;\epsilon)|_1 \leq 1. \tag{0.8}$$

We end this section with two representations for $\mathbf{p}^*(t_0,t;\epsilon)$ as a vector in l_1 space and l_∞ space respectively,

$$\begin{aligned} \mathbf{p}^*(t_0,t;\epsilon) &= \frac{1}{\epsilon}\int_{t_0}^{t}\lambda(s)\mathbf{c}_0\mathbf{M}^*(s,t;\epsilon)\,ds + \mathbf{f}_{n_0}\mathbf{M}^*(t_0,t;\epsilon), \\ \mathbf{p}^*(t_0,t;\epsilon) &= \mathbf{1} + (\mathbf{f}_{n_0} - \mathbf{1})\mathbf{M}^*(t_0,t;\epsilon). \end{aligned} \tag{0.9}$$

1. **Initial layer.** Consider the queueing process $Q(t_0,t)$ with $Q(t_0,t_0) = n_0$. Estimating the behavior of Q over a small period of time compatible with the processes $Q(t_0,t;\epsilon)$ suggests that a time interval of length order ϵ is appropriate. Consider $[t_0, t_0 + \epsilon T]$ as the interval of time. The average number of customers arriving in this time period is $\epsilon^{-1}\int_{t_0}^{t_0+\epsilon T}\lambda(s)\,ds$. Expanding in ϵ gives

$$\frac{1}{\epsilon}\int_{t_0}^{t_0+\epsilon T}\lambda(s)\,ds = \lambda(t_0)T + O(\epsilon)$$

as ϵ goes to zero. This is motivation for the following analytical result.

PROPOSITION 1.1 *For all $T > 0$, we have* $\mathbf{M}^*(t_0, t_0 + \epsilon T;\epsilon) = \exp(\mathbf{A}^*(t_0)T) + O(\epsilon)$ *with respect to the operator norm as ϵ goes to zero.*

PROOF. It can be shown that

$$\begin{aligned} &\mathbf{M}^*(t_0, t_0 + \epsilon T;\epsilon) - \exp(\mathbf{A}^*(t_0)T) \\ &\quad = \frac{1}{\epsilon}\int_0^{\epsilon T}\mathbf{M}^*(t_0,t_0+s;\epsilon)\left[\mathbf{A}^*(t_0+s) - \mathbf{A}^*(t_0)\right]\exp(\mathbf{A}^*(t_0)\epsilon T - s)\,ds. \end{aligned}$$

By the mean value theorem, we have

$$\begin{aligned} |\mathbf{A}^*(t_0+s) - \mathbf{A}^*(t_0)|_1 &\leq 2\left[|\lambda(t_0+s) - \lambda(t_0)| + |\mu(t_0+s) - \mu(t_0)|\right] \\ &\leq 2\left[|\lambda'(\xi_1)| + |\mu'(\xi_2)|\right]s \end{aligned}$$

for some ξ_1 and ξ_2 in $(t_0, t_0 + \epsilon T)$. Now apply (0.8) and the rest follows. ∎

This result supports the notion that the process $Q(t_0, t_0 + \epsilon T;\epsilon)$ as ϵ approaches zero converges in distribution to an **M/M/1** queueing process with $\lambda(t_0)$ and $\mu(t_0)$ as the arrival and service rates. Calling this process $Q^{(0)}(T)$, it follows that

$$\begin{aligned} p_n(t_0, t_0 + \epsilon T;\epsilon) &= p_n^{(0)}(T) + O(\epsilon), \\ E(Q(t_0,t_0+\epsilon T;\epsilon)) &= E(Q^{(0)}(T)) + O(\epsilon), \\ \operatorname{Var}(Q(t_0,t_0+\epsilon T;\epsilon)) &= \operatorname{Var}(Q^{(0)}(T)) + O(\epsilon). \end{aligned}$$

So determining the initial layer for the process **M(t)/M(t)/1** reduces to the analysis of the **M/M/1** queue up to order ϵ. Later, it will be clear how this layer matches up with the subsequent undersaturated or oversaturated region.

2. Undersaturated region. In this region, we exploit the fact that $p_n(t_0, t; \epsilon)$ has an asymptotic expansion in ϵ when $\rho^*(t_0, t) < 1$. Formally, we expand $\mathbf{p}(t_0, t; \epsilon)$ as

$$\mathbf{p}(t_0, t; \epsilon) = \mathbf{p}_0(t) + \epsilon \mathbf{p}_1(t) + O(\epsilon^2) \tag{2.1}$$

where $\mathbf{p}_0(t)$ and $\mathbf{p}_1(t)$ belong to l_1. Substituting (2.1) into (0.3), equating coefficients and ignoring initial conditions gives:

$$\mathbf{p}_0(t)\mathbf{A}(t) = 0, \qquad \mathbf{p}_1(t)\mathbf{A}(t) = \frac{d}{dt}\mathbf{p}_0(t).$$

Adding the conditions $\mathbf{p}_0(t) \cdot \mathbf{1}^T = 1$ and $\mathbf{p}_1(t) \cdot \mathbf{1}^T = 0$ determines these two vectors uniquely for $\lambda(t) < \mu(t)$. After quite a few calculations, see the appendix (§11), one gets the following formulas for the distribution, mean, and variance

$$p_n(t_0, t; \epsilon) = (1 - \rho(t))\rho(t)^n + \epsilon \frac{\rho'(t)}{\mu(t)} \left(\frac{\rho(t)}{(1 - \rho(t))^2} - \frac{n(n+1)}{2} \right) \rho(t)^{n-1} + O(\epsilon^2),$$

$$E(Q(t_0, t; \epsilon)) = \frac{\rho(t)}{1 - \rho(t)} - \epsilon \frac{\rho'(t)}{\mu(t)} \frac{1 + \rho(t)}{(1 - \rho(t))^4} + O(\epsilon^2),$$

$$\operatorname{Var}(Q(t_0, t; \epsilon)) = \frac{\rho(t)}{(1 - \rho(t))^2} - \epsilon \frac{\rho'(t)}{\mu(t)} \frac{1 + 4\rho(t) + \rho(t)^2}{(1 - \rho(t))^5} + O(\epsilon^2).$$

It is worthwhile to stress here the difference between being able to formally calculate an asymptotic expansion and having the expansion actually be valid. The expansions above can be derived whenever $\rho(t) < 1$, but they are only *valid* when $\rho^*(t_0, t) < 1$.

3. Initially oversaturated region. The criterion for oversaturation is $\rho^*(t_0, t) > 1$ and the system is initially oversaturated if $\rho^*(t_0, \cdot) > 1$ on (t_0, t). Accelerating such a process, one would not expect to reach any stable equilibrium but rather see the queue length growing.

If a queue is oversaturated, then the server tends rarely to be idle hence the average queue length should merely be the "flow in" minus the "flow out" or $\int_{t_0}^{t} [\lambda(s) - \mu(s)]\, ds$ plus the initial load of customers n_0. Looking at an accelerated process, one would expect $\epsilon^{-1} \int_{t_0}^{t} [\lambda(s) - \mu(s)]\, ds$ to be the dominant term for the average queue length $E(Q(t_0, t; \epsilon))$. The following equations for the mean and variance of $Q(t_0, t; \epsilon)$ provide more evidence for this assertion.

PROPOSITION 3.1. *For any* $\mathbf{M(t)/M(t)/1}$ *process* $Q(t_0, t; \epsilon)$ *we have*

$$\begin{aligned} \epsilon \frac{\partial}{\partial t} E(Q(t_0, t; \epsilon)) &= \lambda(t) - \mu(t) + \mu(t) p_0(t_0, t; \epsilon), \\ \epsilon \frac{\partial}{\partial t} \operatorname{Var}(Q(t_0, t; \epsilon)) &= \lambda(t) + \mu(t) - \mu(t)\left[2E(Q(t_0, t; \epsilon)) + 1\right] p_0(t_0, t; \epsilon). \end{aligned} \tag{3.1}$$

PROOF. These formulas follow simply from (0.4)

$$\begin{aligned} \epsilon \frac{\partial}{\partial t} E(Q(t_0, t; \epsilon)) &= \epsilon \frac{\partial}{\partial t} \mathbf{p}^*(t_0, t; \epsilon) \cdot \mathbf{1}^T \\ &= \mathbf{p}^*(t_0, t; \epsilon)\mathbf{A}^*(t)\mathbf{1}^T + \lambda(t) \\ &= -\mu(t)(1 - \mathbf{p}_0(t_0, t; \epsilon)) + \lambda(t) \end{aligned}$$

and a similar argument is made for $\operatorname{Var}(Q(t_0, t; \epsilon))$. ∎

If the queue is oversaturated with $p_0(t_0,t;\epsilon)\simeq 0$, then $[2E(Q(t_0,t;\epsilon))+1]p_0(t_0,t;\epsilon)\simeq 0$, and so

$$\frac{\partial}{\partial t}E(Q(t_0,t;\epsilon))\simeq\frac{\lambda(t)-\mu(t)}{\epsilon},\qquad \frac{\partial}{\partial t}\operatorname{Var}(Q(t_0,t;\epsilon))\simeq\frac{\lambda(t)+\mu(t)}{\epsilon}. \tag{3.2}$$

Furthermore, by this argument $\epsilon^{-1}\int_{t_0}^{t}\lambda(s)+\mu(s)\,ds$ is the dominant term for the variance.

We can make all of these ideas precise by the following theorem, which we will prove in §9.

THEOREM 3.2. *Given the process $Q(t_0,\cdot;\epsilon)$ on (t_0,t) with $\lambda(\cdot)$ and $\mu(\cdot)$ with $Q(t_0,t;\epsilon)=n_0$, let $\tilde{Q}(t_0,t;\epsilon)$ be an* **M(t)/M(t)/1** *process that starts at time t and evolves backwards in time to t_0 with $\tilde{\lambda}\equiv\mu$, $\tilde{\mu}\equiv\lambda$, and $\tilde{Q}(t,t;\epsilon)=0$ then*

$$E(Q(t_0,t;\epsilon))=\frac{1}{\epsilon}\int_{t_0}^{t}[\lambda(s)-\mu(s)]\,ds+E(\tilde{Q}(t_0,t;\epsilon)\vee n_0), \tag{3.3}$$

$$\begin{aligned}\operatorname{Var}(Q(t_0,t;\epsilon))&=\frac{1}{\epsilon}\int_{t_0}^{t}[\lambda(s)+\mu(s)]\,ds\\&\quad+\frac{2}{\epsilon}\int_{t_0}^{t}(\lambda(s)-\mu(s))(E(\tilde{Q}(s,t_0;\epsilon)\vee n_0)-E(\tilde{Q}(t,t_0;\epsilon)\vee n_0))\,ds\\&\quad+n_0^2+n_0-E(\tilde{Q}(t,t_0;\epsilon)\vee n_0)^2-E(\tilde{Q}(t,t_0;\epsilon)\vee n_0),\end{aligned} \tag{3.4}$$

where $a\vee b$ is the maximum of a and b.

For this initially oversaturated case, we can now derive asymptotics for the mean and variance via these time reversal formulas. They reduce the problem to deriving the asymptotics of the mean of the *reversed* process where in this case the process is undersaturated. Since $\tilde{Q}(t,t_0;\epsilon)$ is undersaturated then

$$\lim_{\epsilon\downarrow 0}\Pr\{\tilde{Q}(t,t_0;\epsilon)=n\}=\left(1-\frac{1}{\rho(t_0)}\right)\left(\frac{1}{\rho(t_0)}\right)^{n}$$

where $1/\rho(t_0)=\mu(t_0)/\lambda(t_0)<1$ so the asymptotic expansion for $E(Q(t_0,t;\epsilon))$ is simply:

$$E(Q(t_0,t;\epsilon))=\frac{1}{\epsilon}\int_{t_0}^{t}[\lambda(s)-\mu(s)]\,ds+n_0+\left(\frac{1}{\rho(t_0)}\right)^{n_0+1}\frac{1}{1-1/\rho(t_0)}+O(\epsilon)$$

where the constant term is merely the average of the maximum of n_0 with a geometrically distributed random variable. This term is interesting in the sense that intuitively we would not expect such a constant term. The analysis speaks otherwise and we get in effect a modification of the initial load. Compare this expansion with the one given in Clarke [1] when $\rho>1$. Note that the constant term as well as the rest of the $O(\epsilon^n)$ terms have no t dependence and so (3.2) is still true.

Doing a little more work with the formulas for the variance (see [5]) gives

$$\operatorname{Var}(Q(t_0,t;\epsilon))=\frac{1}{\epsilon}\int_{t_0}^{t}[\lambda(s)+\mu(s)]\,ds-\frac{\dfrac{4n_0+3}{\rho(t_0)^{n_0+1}}-\dfrac{4n_0+1}{\rho(t_0)^{n_0+2}}+\dfrac{1}{\rho(t_0)^{2n_0+2}}}{(1-1/\rho(t_0))^2}+O(\epsilon).$$

4. **Onset of rush hour.** A system comes to the onset of rush hour at time $t_1>t_0$ if $\rho(\cdot)<1$ on (t_0,t) but $\rho(t_1)=1$, which implies the same type of behavior for $\rho^*(t_0,\cdot)$.

From here $\rho(\cdot)$ may exceed 1, making the queue oversaturated, or come down to below 1 again and return the system to undersaturation. This can be characterized in general by doing a Taylor expansion of $\rho(\cdot)$ about t_1:

$$\rho(t) = 1 + \frac{\rho^{(k)}(t_1)}{k!}(t - t_1)^k + O\left((t - t_1)^{k+1}\right) \tag{4.1}$$

where $\operatorname{sign}(\rho^{(k)}(t_1)) = (-1)^{k+1}$. Although in practice, one would only consider k to be 1 or 2, it will be useful to observe the general effect that the number k has on the asymptotics of the mean and variance. Aside from this general use of k, the development here will run parallel to the arguments put forth by Newell [6].

Recall that when $\lambda = \mu$ for the **M/M/1** system, the leading asymptotic terms for the mean and variance are $2\sqrt{\lambda t/\pi}$ and $(1 - 2/\pi)2\lambda t$ respectively (see Clarke [1]). These same values are attained by the process $|W(2\lambda t)|$ where $W(t)$ is standard Brownian motion. By analogy, the **M(t)/M(t)/1** process in a neighborhood of t_1 should be approximated by an appropriately scaled diffusion process.

This approximation will be localized in a neighborhood about t_1 in such a way as to be compatible with the uniform acceleration method. In a similar manner to the initial layer, consider the interval $[t_1, t_1 + \epsilon^\alpha T]$ where α is an unspecified positive real number and T is the localized time scale. For the process $Q(t_0, t_1 + \epsilon^\alpha T; \epsilon)$, we want to approximate its infinitesimal mean and variance with respect to T, which shall be denoted by $m(T;\epsilon)$ and $\sigma^2(t;\epsilon)$. This is achieved by taking the T-derivative of the contribution to the mean and variance over the interval $[t_1, t_1 + \epsilon^\alpha T]$ and looking at the leading ϵ term. For the mean we get

$$m(t;\epsilon) = \frac{d}{dT}\frac{1}{\epsilon}\int_{t_1}^{t_1+\epsilon^\alpha T}\left[\lambda(s) - \mu(s)\right]ds = \frac{1}{\epsilon^{1-\alpha(k+1)}}\left[\frac{\lambda^{(k)}(t_1) - \mu^{(k)}(t_1)}{k!}T^k + O(\epsilon^\alpha)\right]$$

and similarly the variance gives

$$\sigma^2(t;\epsilon) = \frac{d}{dT}\frac{1}{\epsilon}\int_{t_1}^{t_1+\epsilon^\alpha T}\left[\lambda(s) + \mu(s)\right]ds = \frac{1}{\epsilon^{1-\alpha}}\left[2\mu(t_1) + O(\epsilon^\alpha)\right].$$

To have the proper scaling for a nonreflecting diffusion, the growth of the variance as $\epsilon\downarrow 0$ should be the square of the growth for the mean. Hence $2(1 - \alpha(k+1)) = 1 - \alpha$, which implies that $\alpha = 1/(2k+1)$ and so

$$m(T;\epsilon) = \frac{1}{\epsilon^{k/(2k+1)}}\left[\frac{\lambda^{(k)}(t_1) - \mu^{(k)}(t_1)}{k!}T^k + O(\epsilon^\alpha)\right],$$

$$\sigma^2(T;\epsilon) = \frac{1}{\epsilon^{2k/(2k+1)}}\left[2\mu(t_1) + O(\epsilon^\alpha)\right].$$

Define a process $Q^{(1)}(T)$ which is the reflecting version of a diffusion with infinitesimal mean and variance

$$m_1(T) = \frac{\lambda^{(k)}(t_1) - \mu^{(k)}(t_1)}{k!}T^k, \qquad \sigma_1^2(T) = 2\mu(t_1).$$

We can then say that

$$Q^{(1)}(T) \approx \epsilon^{k/(2k+1)}Q\left(t_0, t_1 + \epsilon^{1/(2k+1)}T; \epsilon\right)$$

in the sense that their infinitesimal means and variances match up for small ϵ. Here, the effects of ϵ and T have been separated and one can approximate the mean and variance of $Q(t_0, t_1; \epsilon)$ by starting with a convenient initial distribution and solving the following Fokker–Planck equation for $p^{(1)}(x,T)$ the density of $Q^{(1)}(T)$:

$$p_T^{(1)}(x,T) = -\frac{\lambda^{(k)}(t_1) - \mu^{(k)}(t_1)}{k!} T^k p_x^{(1)}(x,T) + \mu(t_1) p_{xx}^{(1)}(x,T)$$

with $\int_0^\infty p^{(1)}(x,T)dx = 1$ for all T where x is the space variable and has the scaling dimensions of $x = \epsilon^{k/(2k+1)}n$.

To initialize this density, recall that for $t < t_1$, $p_n(t_0,t;\epsilon) = (1-\rho(t))\rho(t)^n + O(\epsilon)$. Now subject $(1-\rho(t))\rho(t)^n$ to the *diffusion scaling* i.e., $t - t_1 = \epsilon^{1/(2k+1)}T$ and $n = [x/\epsilon^{k/(2k+1)}]$. We then get

$$\lim_{\epsilon\downarrow 0} \frac{1}{\epsilon^{k/(2k+1)}}(1-\rho(t))\rho(t)^n = \left|\frac{\rho^{(k)}(t_1)T^k}{k!}\right| \exp\left(-x\left|\frac{\rho^{(x)}(t_1)T^k}{k!}\right|\right).$$

Therefore, given $T \ll 0$, $p^{(1)}(x,T)$ should look like the exponential distribution given by the above. The Fokker–Planck equation should then be solved for a density on $[0,\infty)$ with the above as the initial distribution.

It is important to point out that this equation need not be solved (numerically that is) each time for different λ and μ, only for different k. If we make the following substitutions by abuse of notation,

$$\text{replace } \left[\frac{|\rho^{(k)}(t_1)|}{\mu(t_1)^k k!}\right]^{-1/(2k+1)} x \text{ by } x \quad\text{and}\quad \left(\frac{\mu(t_1)|\rho^{(k)}(t_1)|^2}{k!^2}\right)^{-1/(2k+1)} T \text{ by } T$$

then the Fokker–Planck equation becomes the dimensionless equation $\psi_T^{(k)}(x,T) = -T^k\psi_x^{(k)}(x,T) + \psi_{xx}^{(k)}(x,T)$ with $\int_0^\infty \psi^{(k)}(x,T)dx = 1$ and for all T and $\psi^{(k)}(x,T) \approx (-T)^k \exp(-x(-T)^k)$ for $T \ll 0$ and then $p^{(1)}(x,T)$ can be expressed in terms of $\psi^{(k)}(x,T)$ as

$$p^{(1)}(x,T) = \left[\frac{|\rho^{(k)}(t_1)|}{\mu(t_1)^k k!}\right]^{1/(2k+1)} \psi^{(k)}\left[\left[\frac{|\rho^{(k)}(t_1)|}{\mu(t_1)^k k!}\right]^{1/(2k+1)} x, \left(\frac{\mu(t_1)|\rho^{(k)}(t_1)|^2}{k!^2}\right)^{1/(2k+1)} T\right].$$

A final comment, if $\lambda \equiv \mu$ on (t_0,t), then $E(Q(t_0,t;\epsilon)) = O(1/\sqrt{\epsilon})$, whereas having $\lambda(t_1) = \mu(t_1)$ suggests that $E(Q(t_0,t;\epsilon)) = O(1/\epsilon^{k/(2k+1)})$.

Note that as $k \to \infty$, the values of $\lambda(\cdot)$ approach the ones for $\mu(\cdot)$ everywhere on (t_0,t) and $k/(2k+1) \uparrow 1/2$.

5. Main oversaturated region.

This is a region where $\rho^*(t_0,t) > 1$ but there is some time in the past s with $t_0 < s < t$ and $\rho^*(t_0,s) < 1$. To obtain an asymptotic formula for the mean, we combine the techniques of the last two sections.

First, we use a generalized form of the time reversal formula for the mean. Take some intermediate time point t_* in the interval (t_0,t), then we have

$$E(Q(t_0,t;\epsilon)) = \frac{1}{\epsilon}\int_{t_*}^{t}[\lambda(s) - \mu(s)]ds + E(\tilde{Q}(t,t_*;\epsilon) \vee Q(t_0,t_*;\epsilon)) \qquad (5.1)$$

where we assume that Q and $\tilde{Q}$ are independent processes and $Q(t_*,t_*;\epsilon) = 0$. We then choose t_* to be the last time that $\rho(t_*) = \rho^*(t_*,t) = 1$. This makes $\tilde{Q}(t,t_*;\epsilon)$ and $Q(t_0,t_*;\epsilon)$ both **M(t)/M(t)/1** processes at the onset of rush hour. Their diffusion approximations both use the same data from λ and μ at time t_*. Assuming that the mean of their maximum differs little from the mean of either one, at least up to order $\epsilon^{k/(2k+1)}$, we then have a way of approximating the mean of the original process.

If we use notation that will be defined in the next section, we can express the qualitative behavior of $E(Q(t_0,t;\epsilon))$ and $\operatorname{Var}(Q(t_0,t;\epsilon))$ as follows:

$$E(Q(t_0,t;\epsilon)) = \frac{1}{\epsilon}\int_{t_*}^{t}[\lambda(s)-\mu(s)]\,ds + O^+\left(\frac{1}{\epsilon^{k/(2k+1)}}\right),$$

$$\operatorname{Var}(Q(t_0,t;\epsilon)) = \frac{1}{\epsilon}\int_{t_*}^{t}[\lambda(s)+\mu(s)]\,ds + O^+\left(\frac{1}{\epsilon^{2k/(2k+1)}}\right).$$

6. End of rush hour. This transitional period is characterized by $\rho^*(t_0,t)=1$ and $\rho(t)<1$. The onset of rush hour has some ambiguity about whether the process would go on from there to undersaturation or oversaturation. For the end of rush hour, we do not have this problem. The backlog effects that help shape the oversaturated asymptotics have no first order influence on the undersaturated asymptotics.

While approximating the process at this time point is difficult, we can at least make some qualitative remarks about the mean and variance. This can be achieved by using what we will call the *order-plus method*. That is, short of demonstrating that some quantity is of order $O(1/\epsilon^\alpha)$, we show that for every $\delta>0$, the quantity is of order $o(1/\epsilon^{\alpha+\delta})$. We shall denote such a quantity as $O^+(1/\epsilon^\alpha)$. To illustrate this usefulness, we apply it to the onset of rush hour period.

PROPOSITION 6.1. *If $\rho(t)=\rho^*(t_0,t)=1$, then*

$$E(Q(t_0,t;\epsilon)) = O^+\left(\frac{1}{\epsilon^{k/(2k+1)}}\right),\qquad \operatorname{Var}(Q(t_0,t;\epsilon)) = O^+\left(\frac{1}{\epsilon^{2k/(2k+1)}}\right).$$

To prove this, we first need the following lemma,

LEMMA 6.2. *Let $\lambda(\cdot)$ be less than $\mu(\cdot)$ on (t_0,t_1) and $\lambda(t_1)=\mu(t_1)$, then $\mathbf{p}^*(t_0,t_1;\epsilon)\simeq\mathbf{p}^*(t_1-\epsilon^\alpha T,t_1;\epsilon)$ as $\epsilon\downarrow 0$, where $\alpha<1/(2k+1)$ and $T>0$.*

PROOF. By (0.9), it is sufficient to show that

$$\frac{1}{\epsilon}\int_{t_0}^{t_1-\epsilon^\alpha T}\lambda(s)\mathbf{c}_0\mathbf{M}^*(s,t_1;\epsilon)\,ds\simeq 0$$

as $\epsilon\downarrow 0$. The problem can be further reduced to showing that $\mathbf{e}_0\mathbf{M}^*(t_1-\epsilon^\alpha T,t_1;\epsilon)\simeq 0$ and by use of the operator inequality (0.6), we can say that

$$|\mathbf{e}_0\mathbf{M}^*(t_1-\epsilon^\alpha T,t_1;\epsilon)|_1 \leqslant \epsilon^{-(\lambda_*(\epsilon)+\mu_*(\epsilon))/\epsilon}\sum_{k=0}^{\infty}\rho_*(\epsilon)^{k/2}\mathbf{I}_k\left(\frac{2}{\epsilon}\sqrt{\lambda_*(\epsilon)\mu_*(\epsilon)}\right)$$

where $\lambda_*(\epsilon)=\int_{t_1-\epsilon^\alpha T}^{t_1}\lambda(s)\,ds$, $\mu_*(\epsilon)=\int_{t_1-\epsilon^\alpha T}^{t_1}\mu(s)\,ds$, and $\rho^*(\epsilon)=\lambda_*(\epsilon)/\mu_*(\epsilon)$. This holds because $\mathbf{e}_0\mathbf{L}^k=0$ and $|\mathbf{e}_0\mathbf{R}^k|_1=1$. Using the properties of modified Bessel functions, we have

$$I_k\left(\frac{2}{\epsilon}\sqrt{\lambda_*(\epsilon)\mu_*(\epsilon)}\right) < e^{(2/\epsilon)\sqrt{\lambda_*(\epsilon)\mu_*(\epsilon)}}.$$

Since $\rho_*(\epsilon)<1$ here, then

$$|\mathbf{e}_0\mathbf{M}^*(t_1-\epsilon^\alpha T,t_1;\epsilon)|_1 \leqslant \frac{\rho_*(\epsilon)}{1-\rho_*(\epsilon)}e^{-(1/\epsilon)\left(\sqrt{\lambda_*(\epsilon)}-\sqrt{\mu_*(\epsilon)}\right)^2}. \tag{6.1}$$

As $\epsilon\downarrow 0$, $\rho_*(\epsilon)$ goes to 1 and $\lambda_*(0)=\mu_*(0)=0$.

Without loss of generality, we will assume $\mu(\cdot)$ to be constant so $\mu_*(\epsilon)=\mu\epsilon^\alpha T$.

Making use of (4.1) for $\rho(\cdot)$, we get

$$\begin{aligned}
\sqrt{\lambda_*(\epsilon)} - \sqrt{\mu_*(\epsilon)} &= \sqrt{\int_{t_1-\epsilon^\alpha T}^{t_1} \lambda(s)\,ds} - \sqrt{\mu\epsilon^\alpha T} \\
&= \sqrt{\mu\epsilon^\alpha T}\left[\sqrt{\frac{1}{\epsilon^\alpha T}\int_{t_1-\epsilon^\alpha T}^{t_1} \rho(s)\,ds} - 1\right] \\
&= \sqrt{\mu\epsilon^\alpha T}\left[\sqrt{1 + \frac{1}{\epsilon^\alpha T}\,\frac{\rho^{(k)}(t_1)}{(k+1)!}\,\epsilon^{\alpha(k+1)}T^{k+1} + O(\epsilon^{\alpha(k+1)})} - 1\right] \\
&= \sqrt{\mu\epsilon^\alpha T}\left[\frac{\rho^{(k)}(t_1)}{2(k+1)!}\,\epsilon^{\alpha k}T^{k+1} + O(\epsilon^{\alpha(k+1)})\right] \\
&= O(\epsilon^{\alpha(k+1/2)}),
\end{aligned}$$

$$\left(\sqrt{\lambda_*(\epsilon)} - \sqrt{\mu_*(\epsilon)}\right)^2 = O(\epsilon^{\alpha(2k+1)}).$$

By (6.1), we get exponential decay whenever $\alpha(2k+1) < 1$ or $\alpha < 1/(2k+1)$. ∎

PROOF OF PROPOSITION 6.1. By Lemma 6.2, $E(Q(t_0, t_1; \epsilon)) \simeq E(Q(t_1 - \epsilon^\alpha T, t_1; \epsilon))$ for $T > 0$ and $\alpha < 1/(2K+1)$. At this point, we employ a *stochastic dominance* argument. On the interval $(t_1 - \epsilon^\alpha T, t_1)$, $\lambda(\cdot)$ is close to $\mu(\cdot)$ since $\lambda(t_1) = \mu(t_1)$, but $\lambda(\cdot)$ is less than $\mu(\cdot)$. Without loss of generality, we let $\mu(\cdot) \equiv \mu$ a constant, and consider $Q^\dagger$, an **M/M/1** queueing process on the time interval $(t_1 - \epsilon^\alpha T, t_1)$ with μ for the arrival and service rates. We then have a system with the same service rate as Q but a larger arrival rate. One would then *expect* $Q^\dagger$ to be larger than Q. This turns out to be the case stochastically which we shall prove in §10, but for now we will say that $E(Q(t_1 - \epsilon^\alpha T, t_1; \epsilon)) \leqslant E(Q^\dagger(t_1 - \epsilon^\alpha T, t_1; \epsilon))$. But $Q^\dagger$ is a stationary process, so we can say by abuse of notation that

$$E\big(Q^\dagger(t_1 - \epsilon^\alpha T, t_1; \epsilon)\big) = E\big(Q^\dagger(\epsilon^{\alpha-1}T)\big)$$

and $\alpha < 1/(2k+1)$ so as $\epsilon \downarrow 0$,

$$E\big(Q^\dagger(\epsilon^{\alpha-1}T)\big) = \epsilon^{(\alpha-1)/2}\sqrt{\frac{2\mu T}{\pi}} - \frac{1}{2} + O\left(\sqrt{\frac{\epsilon^{1-\alpha}}{T}}\right).$$

Therefore

$$\begin{aligned}
\limsup_{\epsilon\downarrow 0} \epsilon^{(1-\alpha)/2} E(Q(t_0, t; \epsilon)) &= \limsup_{\epsilon\downarrow 0} \epsilon^{(1-\alpha)/2} E(Q(t_1 - \epsilon^\alpha T, t_1; \epsilon)) \\
&\leqslant \lim_{\epsilon\downarrow 0} \epsilon^{(1-\alpha)/2} E\big(Q^\dagger(\epsilon^{\alpha-1}T)\big) \\
&\leqslant \sqrt{\frac{2\mu T}{\pi}}\,.
\end{aligned}$$

But $T > 0$ is arbitrary so

$$\lim_{\epsilon\downarrow 0} \epsilon^{(1-\alpha)/2} E(Q(t_0, t; \epsilon)) = 0,$$

and since $\alpha < 1/(2k+1)$, this holds for $(1-\alpha)/2 > k/(2k+1)$.

For $\operatorname{Var}(Q(t_0,t;\epsilon))$, we follow a similar argument, making use of $\mathbf{n}$. ∎

Compare the results of Proposition 6.1 to the informal arguments of §4. We now use the order plus method for the end of rush hour.

PROPOSITION 6.3. *If $\rho(t) < \rho^*(t_0,t) = 1$, then*

$$E(Q(t_0,t;\epsilon)) = O^+\left(\frac{1}{\sqrt{\epsilon}}\right), \qquad \operatorname{Var}(Q(t_0,t;\epsilon)) = O^+\left(\frac{1}{\epsilon}\right).$$

To do the proof, we first require two lemmas.

LEMMA 6.4. *If $\rho(t) < \rho^*(t_0,t) = 1$, then there exists a time t_* such that $\lambda(t_*) = \mu(t_*)$, $\int_{t_*}^{t}\lambda(s)\,ds = \int_{t_*}^{t}\mu(s)\,ds$, and for all τ, where $\tau_* < \tau < t$, $\int_{t_*}^{\tau}\lambda(s)\,ds > \int_{t_*}^{\tau}\mu(s)\,ds$.*

PROOF. Since $\rho^*(t_0,t) = 1$ and $\rho(t) < 1$, there exist a τ and a $\delta > 0$ such that $\int_{\tau}^{t}\lambda(s)\,ds = \int_{\tau}^{t}\mu(s)\,ds$ holds, and it implies that $\tau \leqslant t - \delta$. Let τ_* be the largest such τ. Since $\int_{\tau}^{t}\lambda(s)\,ds < \int_{\tau}^{t}\mu(s)\,ds$ from all τ in (t_*,t), it holds that $\int_{t_*}^{\tau}\lambda(s)\,ds > \int_{t_*}^{\tau}\mu(s)\,ds$.

Finally, $\lambda(t_*) = \mu(t_*)$, otherwise one could find a t' to t_* such that $\int_{t'}^{t}\lambda(s)\,ds > \int_{t'}^{t}\mu(s)\,ds$ which would contradict $\rho^*(t_0,t) = 1$. ∎

LEMMA 6.5. *The time reversal formula for the variance* (3.4) *can be written as*

$$\begin{aligned}\operatorname{Var}(Q(t_0,t;\epsilon)) &= \frac{1}{\epsilon}\int_{t_0}^{t}[\lambda(s) + \mu(s)]\,ds \\ &\quad + \frac{2}{\epsilon^2}\int_{t_0}^{t}\left(\int_{t_0}^{s}[\lambda(\xi) - \mu(\xi)]\,d\xi\right)\mu(s)p_0(t_0,s;\epsilon)\,ds \\ &\quad + n_0^2 + n_0 - E(\tilde{Q}(t,t_0;\epsilon) \vee n_0)^2 - E(\tilde{Q}(t,t_0;\epsilon) \vee n_0).\end{aligned}$$

PROOF. By combining (3.1) with (3.3), we can show that

$$\mu(t)p_0(t_0,t;\epsilon) = \epsilon\frac{\partial}{\partial t}E(\tilde{Q}(t,t_0;\epsilon) \vee n_0)$$

therefore

$$\begin{aligned}&\frac{2}{\epsilon}\int_{t_0}^{t}(\lambda(s) - \mu(s))(E(\tilde{Q}(s,t_0;\epsilon) \vee n_0) - E(\tilde{Q}(t,t_0;\epsilon) \vee n_0))\,ds \\ &\quad = -\frac{2}{\epsilon^2}\int_{t_0}^{t}(\lambda(s) - \mu(s))\int_{s}^{t}\mu(\xi)p_0(t_0,\xi;\epsilon)\,d\xi\,ds \\ &\quad = -\frac{2}{\epsilon^2}\int_{t_0}^{t}\frac{d}{ds}\left(\int_{t_0}^{s}\lambda(\xi) - \mu(\xi)\,d\xi\right)\cdot\int_{s}^{t}\mu(\xi)p_0(t_0,\xi;\epsilon)\,d\xi\,ds \\ &\quad = -\frac{2}{\epsilon^2}\int_{t_0}^{t}\mu(s)p_0(t_0,s;\epsilon)\left(\int_{t_0}^{s}\lambda(\xi) - \mu(\xi)\,d\xi\right)ds\end{aligned}$$

and the rest follows. ∎

PROOF OF PROPOSITION 6.3. By the lemma above, it is sufficient to consider the case of $\lambda(t_0) = \mu(t_0)$, $\int_{t_0}^{t}\lambda(s)\,ds = \int_{t_0}^{t}\mu(s)\,ds$, and $\int_{t_0}^{\tau}\lambda(s)\,ds > \int_{t_0}^{\tau}\mu(s)\,ds$ whenever $t_0 < \tau < t$. Since $\lambda(t_0) = \mu(t_0)$, we expand $\rho(\cdot)$ about t_0 as $\rho(t) = 1 + (\rho^{(k)}(t_0)/k!)(t - t_0)^k + O((t - t_0)^{k+1})$.

As $\epsilon\downarrow 0$, we can say that $\mathbf{f}_{n_0}\mathbf{M}^*(t_0,t;\epsilon)\mathbf{1}^T = O(1)$ at least, so we will ignore initial conditions and say that

$$E(Q(t_0,t;\epsilon)) = \frac{1}{\epsilon}\int_{t_0}^{t}\lambda(s)\mathbf{c}_0M^*(s,t;\epsilon)\mathbf{1}^T\,ds + O(1).$$

We know by remarks in the proof of Lemma 6.2, that if $t_0 < \tau < t$, then $\mathbf{e}_0\mathbf{M}^*(\tau, t; \epsilon)\mathbf{1}^T \simeq 0$. So in fashion similar to the proof of Lemma 6.2, we introduce a local time scale T and try to determine an upper bound for α such that

$$\frac{1}{\epsilon}\int_{t_0}^{t}\lambda(s)\mathbf{e}_0\mathbf{M}^*(s,t;\epsilon)\mathbf{1}^T ds \simeq \frac{1}{\epsilon}\int_{t_0}^{t_0+\epsilon^\alpha T}\lambda(s)\mathbf{e}_0\mathbf{M}^*(s,t;\epsilon)\mathbf{1}^T ds.$$

Just as in Lemma 6.2, we can show that

$$|\mathbf{e}_0\mathbf{M}^*(t_0+\epsilon^\alpha T, t;\epsilon)|_1 \leqslant e^{-(\lambda_*(\epsilon)+\mu_*(\epsilon))/\epsilon}\sum_{k=0}^{\infty}\rho_*(\epsilon)^{k/2} I_k\left(\frac{2}{\epsilon}\sqrt{\lambda_*(\epsilon)\mu_*(\epsilon)}\right)$$

where $\lambda_*(\epsilon) = \int_{t_0+\epsilon^\alpha T}^{t}\lambda(s)\,ds$, $\mu_*(\epsilon) = \int_{t_0+\epsilon^\alpha T}^{t}\mu(s)\,ds$, and $\rho_*(\epsilon) = \lambda_*(\epsilon)/\mu_*(\epsilon)$. And as before, we need only determine the order of $(\sqrt{\lambda_*(\epsilon)} - \sqrt{\mu_*(\epsilon)}\,)^2$. Let $\Gamma = \int_{t_0}^{t}\lambda(s)\,ds = \int_{t_0}^{t}\mu(s)\,ds$, then

$$\begin{aligned}\sqrt{\lambda_*(\epsilon)} - \sqrt{\mu_*(\epsilon)} &= \sqrt{\Gamma - \int_{t_0}^{t_0+\epsilon^\alpha T}\lambda(s)\,ds} - \sqrt{\Gamma - \int_{t_0}^{t_0+\epsilon^\alpha T}\mu(s)\,ds}\\ &= \frac{\int_{t_0}^{t_0+\epsilon^\alpha T}\mu(s) - \lambda(s)\,ds}{\sqrt{\Gamma - \int_{t_0}^{t_0+\epsilon^\alpha T}\lambda(s)\,ds} + \sqrt{\Gamma - \int_{t_0}^{t_0+\epsilon^\alpha T}\mu(s)\,ds}}\\ &= O\left(\int_{t_0}^{t_0+\epsilon^\alpha T}\mu(s) - \lambda(s)\,ds\right)\\ &= O(\epsilon^{\alpha(k+1)}).\end{aligned}$$

Therefore, $(\sqrt{\lambda_*(\epsilon)} - \sqrt{\mu_*(\epsilon)}\,)^2 = O(\epsilon^{2\alpha(k+1)})$ and by (6.1), we get exponential decay whenever $2\alpha(k+1) < 1$ or $\alpha < 1/2(k+1)$.

Now by (0.7), for $t_0 < s < t_0 + \epsilon^\alpha T$ we have

$$\mathbf{M}^*(s,t;\epsilon) = \mathbf{M}^*(s, t_0+\epsilon^\alpha T;\epsilon)\mathbf{M}^*(t_0+\epsilon^\alpha T, t;\epsilon)$$

and so by (0.8), $\mathbf{e}_0\mathbf{M}^*(s,t;\epsilon)\mathbf{1}^T \leqslant \mathbf{e}_0\mathbf{M}^*(s,t_0+\epsilon^\alpha T;\epsilon)\mathbf{1}^T$. Hence,

$$\begin{aligned}\frac{1}{\epsilon}\int_{t_0}^{t_0+\epsilon^\alpha T}\lambda(s)\mathbf{e}_0\mathbf{M}^*(s,t;\epsilon)\mathbf{1}^T ds &\leqslant \frac{1}{\epsilon}\int_{t_0}^{t_0+\epsilon^\alpha T}\lambda(s)\mathbf{e}_0\mathbf{M}^*(s,t_0+\epsilon^{\alpha T};\epsilon)\mathbf{1}^T ds\\ &\simeq \frac{1}{\epsilon}\int_{t_0}^{t_0+\epsilon^\alpha T}[\lambda(s) - \mu(s)]\,ds + O(1).\end{aligned}$$

Therefore,

$$E(Q(t_0,t;\epsilon)) = O\left(\frac{T^{k+1}}{\epsilon^{1-\alpha(k+1)}}\right)$$

But T is arbitrary, so

$$E(Q(t_0,t;\epsilon)) = O\left(\frac{1}{\epsilon^{1-\alpha(k+1)}}\right)$$

which holds whenever $1 - \alpha(k+1) > 1/2$, given our constraint on α.

We now use (5.1), the generalization of the time reversal formula where t_* is chosen to match the specifications of Lemma 6.4. We then get $E(Q(t_0,t;\epsilon)) = E(\tilde{Q}(t,t_*;\epsilon) \vee Q(t_0,t_*;\epsilon))$. In comparing $\tilde{Q}(t,t_*;\epsilon)$ to $Q(t_0,t_*;\epsilon)$, we note the mean of the latter is equal to $O^+(1/\epsilon^{k/(2k+1)})$ by Proposition 6.1. Whereas for $\tilde{Q}$, we use the time reversal formula again to get $E(\tilde{Q}(t,t_*;\epsilon)) = E(Q(t_*,t;\epsilon)) = O^+(1/\sqrt{\epsilon})$. This suggests that as $\epsilon\downarrow 0$, $\tilde{Q}(t,t_*;\epsilon)$ dominates $Q(t_0,t_*;\epsilon)$ and $E(Q(t_0,t;\epsilon))$ takes on $O^+(1/\sqrt{\epsilon})$ behavior.

For $\operatorname{Var}(Q(t_0,t;\epsilon))$, we appeal to a generalized time reversal formula,

$$\begin{aligned}\operatorname{Var}(Q(t_0,t;\epsilon)) = {} & \frac{1}{\epsilon}\int_{t_*}^{t}[\lambda(s)+\mu(s)]\,ds + \frac{2}{\epsilon^2}\int_{t_*}^{t}\left(\int_{t_*}^{s}\lambda(\xi)-\mu(\xi)\,d\xi\right)\mu(s)p_0(t_*,s;\epsilon)\,ds\\ & - E(\tilde{Q}(t,t_*;\epsilon)\vee Q(t_0,t_*;\epsilon))^2 - E(\tilde{Q}(t,t_*;\epsilon)\vee Q(t_0,t_*;\epsilon))\\ & + E(Q(t_0,t_*,\epsilon))^2 + E(Q(t_0,t_*;\epsilon))\end{aligned}$$

where t_* is the same time point that we chose for the mean.

By the previous analysis, it is clear for all except the second term, that each summand has order $O(1/\epsilon)$. We are left with the analysis of

$$\frac{2}{\epsilon^2}\int_{t_*}^{t}\left(\int_{t_*}^{s}\lambda(\xi)-\mu(\xi)\,d\xi\right)\mu(s)p_0(t_*,s;\epsilon)\,ds.$$

Whenever $t_* < s < t$, we have $p_0(t_*,s;\epsilon)\simeq 0$, so we need only consider the behavior about the endpoints.

Case 1. $s\approx t$. This is like the onset of rush hour period so we can restrict s to the interval $(t_*,t_*+\epsilon^\alpha T)$ where $\alpha < 1/(2k+1)$. Now consider an **M(t)/M(t)/1** process $Q^\dagger(t_*,t_*+\epsilon^\alpha T;\epsilon)$ where $\lambda^\dagger(\cdot)\equiv\mu^\dagger(\cdot)\equiv\mu(\cdot)$. It turns out that $Q^\dagger$ is a stationary process and since $\mu(\cdot)<\lambda(\cdot)$ in this region, $Q^\dagger \leqslant Q$ in the sense of stochastic dominance so

$$\Pr\{Q^\dagger(t_*,t_*+\epsilon^\alpha T)>0\} \leqslant \Pr\{Q(t_*,t_*+\epsilon^\alpha T)>0\}$$

or in other words

$$p_0(t_*,t_*+\epsilon^\alpha T) \leqslant p_0^\dagger(t_*,t_*+\epsilon^\alpha T).$$

But $p_0^\dagger(t_*,t_*+\epsilon^\alpha T) = O(\epsilon^{(1-\alpha)/2})$ and so $p_0(t_*,t_*+\epsilon^\alpha T)$ has the same order. Since $\lambda(t_*)=\mu(t_*)$ and $\rho(t_*)$ can be expected as in (4.1), we get

$$\begin{aligned}\frac{2}{\epsilon^2}\int_{t_*}^{t_*+\epsilon^\alpha T}\left(\int_{t_*}^{s}\lambda(\xi)-\mu(\xi)\,d\xi\right)\mu(s)p_0(t_*,s;\epsilon)\,ds &= O\left(\frac{1}{\epsilon^2}\epsilon^{\alpha(k+2)}\epsilon^{(1-\alpha)/2}T^{2k+2}\right)\\ &= O(\epsilon^{\alpha(k+3/2)-3/2}T^{2k+2})\end{aligned}$$

and $3/2-\alpha(k+3/2) > 2k/(2k+1)$ so this term is certainly $O^+(1/\epsilon)$.

Case 2. $t_*\approx t$. Here it is like dealing with the end of rush hour period. We would calculate the upper bound for α just as we did for the mean. The only difference would be that $\lambda(t)\neq\mu(t)$, but this would be as if we set $k=0$ and so we must have $\alpha < 1/2$. Since $\int_{t*}^{t}\lambda(s)\,ds = \int_{t*}^{t}\mu(s)\,ds$, we get

$$\begin{aligned}&\frac{2}{\epsilon^2}\int_{t_*-\epsilon^\alpha T}^{t}\left(\int_{t_*}^{s}\lambda(\xi)-\mu(\xi)\,d\xi\right)\mu(s)p_0(t_*,s,\epsilon)\,ds\\ &\quad = \frac{-2}{\epsilon^2}\int_{t_*-\epsilon^\alpha T}^{t}\left(\int_{s}^{t}\lambda(\xi)-\mu(\xi)\,d\xi\right)\mu(s)p_0(t_*,s;\epsilon)\,ds\\ &\quad = O(\epsilon^{2(\alpha-1)}T^2)\end{aligned}$$

but $2(1-\alpha)>1$ and T is arbitrary so we get $O^+(1/\epsilon)$ for this term and we are done.

∎

COROLLARY TO PROPOSITION 6.3. *For the main oversaturated case we can prove that*

$$\operatorname{Var}(Q(t_0,t;\epsilon)) = \frac{1}{\epsilon}\int_{t_*}^{t}[\lambda(s)+\mu(s)]\,ds + O^+\left(\frac{1}{\epsilon^{2k/(2k+1)}}\right).$$

PROOF. We merely repeat the work done for $\operatorname{Var}(Q(t_0,t;\epsilon))$ above and note that Case 2 does not apply here. ∎

7. The asymptotics of the fundamental solution operator. In this section, the following fundamental theorem will be proved.

THEOREM 7.1. *If $\int_{t_0}^{t}\lambda(s)\,ds < \int_{t_0}^{t}\mu(s)\,ds$, then for all $\mathbf{g}$ in l_1, $\lim_{\epsilon\downarrow 0}\mathbf{g}\mathbf{M}^*(t_0,t;\epsilon) = 0$ and if $\mathbf{g}$ is a finite dimensional vector, then the rate of convergence is exponential. If we merely have $\rho(t_0) < 1$, then $\mathbf{g}\mathbf{M}^*(t_0,t;\epsilon) = O(\epsilon^n)$ provided $\mathbf{g} = \mathbf{g}^+ - \mathbf{g}^-$, where $\mathbf{g}^+$ and $\mathbf{g}^-$ are positive and both belong to the range of $\mathbf{A}^*(t_0)^n$.*

The proof will use the following three lemmas:

LEMMA 7.2. *Let $\mathbf{R}$ and $\mathbf{L}$ be respectively the right and left shift operators, then*

$$\begin{aligned}&\exp(\lambda(\mathbf{R}-\mathbf{I}))\exp(\mu(\mathbf{L}-\mathbf{I}))\\ &\quad= e^{-(\lambda+\mu)}\left[I_0\left(2\sqrt{\lambda\mu}\right)\mathbf{I} + \sum_{k=0}^{\infty}\left(\left(\frac{\mu}{\lambda}\right)^{k/2}\mathbf{L}^k + \left(\frac{\lambda}{\mu}\right)^{k/2}\mathbf{R}^k\right)I_k\left(2\sqrt{\lambda\mu}\right)\right].\end{aligned}$$

PROOF. Let $\mathbf{S}$ be an operator with both $\mathbf{S}$ and $\mathbf{S}^{-1}$ bounded. Then

$$\begin{aligned}\exp(\lambda(\mathbf{S}-\mathbf{I}))\exp\left(\mu(\mathbf{S}^{-1}-\mathbf{I})\right) &= e^{-(\lambda+\mu)}\exp(\lambda\mathbf{S})\exp(\mu\mathbf{S}^{-1})\\ &= e^{-(\lambda+\mu)}\sum_{k=-\infty}^{\infty}\left(\frac{\lambda}{\mu}\right)^{k/2}\mathbf{S}^k I_k\left(2\sqrt{\lambda\mu}\right)\end{aligned}$$

since $e^{yx/2}\cdot e^{yx^{-1}/2} = \sum_{k=-\infty}^{\infty}x^k I_k(y)$.

Now $\mathbf{L}$ is a right inverse operator for $\mathbf{R}$, that is $\mathbf{RL} = \mathbf{I}$. Since $\mathbf{L}$ is only applied to the right in this lemma, we get the same result that we did with $\mathbf{S}$. We merely replace $\mathbf{S}^k$ by $\mathbf{R}^k$ and $\mathbf{S}^{-k}$ by $\mathbf{L}^k$ for k a positive integer. ∎

LEMMA 7.3. *If $\lambda < \mu$, then for all $\mathbf{g}$ in l_1, $\lim_{\epsilon\downarrow 0}\mathbf{g}\exp(\lambda(\mathbf{R}-\mathbf{I}))\exp(\mu(\mathbf{L}-\mathbf{I})) = 0$.*

PROOF. Let $\mathbf{g} = [\,g_0 \ldots g_N 0 \ldots\,]$ be a finite dimensional vector and for simplicity, let $\rho = \lambda/\mu$. Then

$$\begin{aligned}&\mathbf{g}\exp(\lambda(\mathbf{R}-\mathbf{I}))\exp(\mu(\mathbf{L}-\mathbf{I}))\\ &\quad= e^{-(\lambda+\mu)}\left[I_0\left(2\sqrt{\lambda\mu}\right)\mathbf{g} + \sum_{k=0}^{\infty}\left(\left(\frac{\mu}{\lambda}\right)^{k/2}\mathbf{g}\mathbf{L}^k + \left(\frac{\lambda}{\mu}\right)^{k/2}\mathbf{g}\mathbf{R}^k\right)I_k\left(2\sqrt{\lambda\mu}\right)\right].\end{aligned}$$

However, $\mathbf{g}\mathbf{L}^k = 0$ for $k > N$ and $|\mathbf{g}\mathbf{L}^k|_1, |\mathbf{g}\mathbf{R}^k|_1 \leqslant |\mathbf{g}|_1$. Combining this with $\rho < 1$ and the key inequality $I_k(x) \leqslant e^x$ for $x \geqslant 0$ gives

$$\begin{aligned}\left|\mathbf{g}\exp\left(\frac{\lambda}{\epsilon}(\mathbf{R}-\mathbf{I})\right)\exp\left(\frac{\mu}{\epsilon}(\mathbf{L}-\mathbf{I})\right)\right|_1 &\leqslant |\mathbf{g}|_1\sum_{k=-N}^{\infty}\rho^{k/2}\exp\left[-\frac{\left(\sqrt{\lambda}-\sqrt{\mu}\right)^2}{\epsilon}\right]\\ &\leqslant \frac{|\mathbf{g}|_1}{\rho^{N/2}\left(1-\sqrt{\rho}\right)}\exp\left[-\frac{\left(\sqrt{\lambda}-\sqrt{\mu}\right)^2}{\epsilon}\right].\end{aligned}$$

Thus we have shown that the lemma holds for all finite dimensional $\mathbf{g}$. In fact, we have shown for all such $\mathbf{g}$ that the rate of convergence is exponential. Since such vectors are dense in l_1, the rest follows. ∎

LEMMA 7.4. *Let* $\mathbf{A}_i^* = \lambda_i\mathbf{R} + \mu_i\mathbf{L} - (\lambda_i + \mu_i)\mathbf{I}$ *for* $i = 1, \ldots, n$; *then*

$$\exp(\mathbf{A}_1^*) \ldots \exp(\mathbf{A}_n^*) \leqslant \exp((\lambda_1 + \cdots + \lambda_n)(\mathbf{R} - \mathbf{I}))\exp((\mu_1 + \cdots + \mu_n)(\mathbf{L} - \mathbf{I})).$$

PROOF. Use induction on n.

($n = 1$) We have $\mathbf{RL} = \mathbf{I}$, but $\mathbf{LR}$ is a diagonal operator with a zero in the first diagonal entry and ones thereafter so $\mathbf{LR} \leqslant \mathbf{RL}$. The function e^x has a power series representation that converges absolutely for all x. Moreover, the coefficients of this series are positive. So we expand $\exp(\lambda\mathbf{R} + \mu\mathbf{L})$ by the power series representation. Since it is clearly by induction that $(\lambda\mathbf{R} + \mu\mathbf{L})^k \leqslant \sum_{i=0}^k \binom{k}{i}\lambda^i\mu^{k-i}\mathbf{R}^i\mathbf{L}^{k-i}$, then

$$\begin{aligned}\exp(\lambda\mathbf{R} + \mu\mathbf{L}) &= \sum_{k=0}^{\infty} \frac{1}{k!}(\lambda\mathbf{R} + \mu\mathbf{L})^k \\ &\leqslant \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{i=0}^{k} \binom{k}{i}\lambda^i\mu^{k-i}\mathbf{R}^k\mathbf{L}^{k-i} \\ &\leqslant \sum_{i=0}^{\infty}\sum_{k=i}^{\infty} \frac{\lambda^i}{i!}\mathbf{R}^i \frac{\mu^{k-i}}{(k-i)!}\mathbf{L}^{k-i} \\ &\leqslant \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}\mathbf{R}^i \sum_{j=0}^{\infty} \frac{\mu^j}{j!}\mathbf{L}^j \\ &\leqslant \exp(\lambda\mathbf{R})\exp(\mu\mathbf{L}).\end{aligned}$$

Multiplying both sides by $e^{-(\lambda+\mu)}$, we have demonstrated the hypothesis for the case $n = 1$.

($n \Rightarrow n + 1$) By induction hypothesis, we have

$$\begin{aligned}&\exp(\mathbf{A}_1^*) \ldots \exp(\mathbf{A}_n^*)\exp(\mathbf{A}_{n+1}^*) \\ &\quad \leqslant \exp((\lambda_1 + \cdots + \lambda_n)(\mathbf{R} - \mathbf{I}))\exp((\mu_1 + \cdots + \mu_n)(\mathbf{L} - \mathbf{I})) \\ &\quad\quad \times \exp(\lambda_{n+1}(\mathbf{R} - \mathbf{I}))\exp(\mu_{n+1}(\mathbf{L} - \mathbf{I})).\end{aligned} \tag{7.1}$$

By a similar argument to the case $n = 1$, it holds that

$$\begin{aligned}&\exp((\mu_1 + \cdots + \mu_n)(\mathbf{L} - \mathbf{I}))\exp(\lambda_{n+1}(\mathbf{R} - \mathbf{I})) \\ &\quad \leqslant \exp(\lambda_{n+1}(\mathbf{R} - \mathbf{I}))\exp((\mu_1 + \cdots + \mu_n)(\mathbf{L} - \mathbf{I})).\end{aligned}$$

This allows us to rearrange terms in (7.1) to get the desired upper bound. ∎

PROOF OF THEOREM 7.1. From Lemma 7.4 and by Kato's 'representation for $\mathbf{M}^*(t_0, t; \epsilon)$, it follows that

$$\mathbf{M}^*(t_0, t; \epsilon) \leqslant \exp\left(\frac{1}{\epsilon}\int_{t_0}^{t}\lambda(s)\,ds(\mathbf{R} - \mathbf{I})\right)\exp\left(\frac{1}{\epsilon}\int_{t_0}^{t}\mu(s)(\mathbf{L} - \mathbf{I})\right).$$

Lemma 7.2 proves (0.6) and Lemma 7.3 proves the first part.

For the second part, let $\mathbf{g}$ be positive, if $\rho(t_0) < 1$, then there exists an interval (t_0, t_*) such that $\rho(\cdot) < 1$, given the continuity of $\lambda(\cdot)$ and $\mu(\cdot)$. Also $\tilde{\lambda} = \sup_{s \in (t_0, t_*)}\lambda(s) < \inf_{s \in (t_0, t_*)}\mu(s) = \tilde{\mu}$. Let $\tilde{\mathbf{A}}^* = \tilde{\lambda}\mathbf{R} + \mu\mathbf{L} - (\tilde{\lambda} + \tilde{\mu})\mathbf{I}$, then we can use a stochastic dominance argument (to be proved in §10) to say that

$$|\mathbf{gM}^*(t_0,t;\epsilon)|_1 \leqslant |\mathbf{gM}^*(t_0,t_*;\epsilon)|_1 \leqslant \left|\mathbf{g}\exp\left(\frac{t_*-t_0}{\epsilon}\tilde{\mathbf{A}}^*\right)\right|_1 .$$

We note that the range of any given $\mathbf{A}^*$ is the same as long as $\lambda < \mu$. By induction, we can say the same for $(\mathbf{A}^*)^n$. Since $\tilde{\rho} = \tilde{\lambda}/\tilde{\mu} < 1$, the rest follows (see [5, p. 5] for details). ∎

8. A nonstationary traffic intensity parameter.

We now establish the usefulness of the quantity $\rho^*(t_0,t)$ as a traffic intensity parameter where

$$\rho^*(t_0,t) = \sup_{t_*\in(t_0,t)} \frac{\int_{t_*}^{t}\lambda(s)\,ds}{\int_{t_*}^{t}\mu(s)\,ds} .$$

Proof of Theorem 1. Recall from the definition of $\mathbf{p}^*(t,t;\epsilon)$ that

$$\mathbf{p}(t_0,t;\epsilon) = \left[1,\mathbf{p}^*(t_0,t;\epsilon)\right](\mathbf{I}-\mathbf{L}). \tag{8.1}$$

So it will be sufficient to prove this theorem for $\mathbf{p}^*(t_0,t;\epsilon)$.

[Undersaturation $\Rightarrow \rho^*(t_0,t) < 1$] Suppose that $\rho^*(t_0,t) \geqslant 1$. If the inequality is strict, then there exists a t_* in (t_0,t) such that $\int_{t_*}^{t}\lambda(s)\,ds > \int_{t_*}^{t}\mu(s)\,ds$. By use of (0.9), we can say that

$$p_n^*(t_0,t;\epsilon) = \mathbf{p}^*(t_0,t;\epsilon)\cdot\mathbf{e}_n^T = 1-(1-\mathbf{f}_{n_0})\mathbf{M}^*(t_0,t;\epsilon)\mathbf{e}_n^T \simeq 1,$$

since by (0.7), (0.8), and Theorem 7.1, $|\mathbf{e}_n\mathbf{M}^*(t_0,t;\epsilon)^T|_1 \leqslant |\mathbf{e}_n\mathbf{M}^*(t_*,t;\epsilon)^T|_1 \simeq 0$. But by (8.1), $p_n(t_0,t;\epsilon) = p_{n-1}^*(t_0,t;\epsilon) - p_n^*(t_0,t;\epsilon)$ so $p_n^*(t_0,t;\epsilon) \simeq 0$.

Now suppose that $\rho^*(t_0,t) = 1$, then either $\rho(t) = 1$ or there exists a t_* such that $\int_{t_*}^{t}\lambda(s)\,ds = \int_{t_*}^{t}\mu(s)\,ds$. If the former holds, then we can employ a stochastic dominance argument. Let $Q^\dagger(t_0,t;\epsilon)$ be an $\mathbf{M(t)/M(t)/1}$ process where $\mu^\dagger(\cdot) = \mu(\cdot)$, but $\lambda^\dagger(\cdot) < \lambda(\cdot)$ so that $\rho^\dagger(\cdot) < 1$ on (t_0,t). Then

$$E\big(Q(t_0,t;\epsilon)\big) \geqslant E\big(Q^\dagger(t_0,t;\epsilon)\big)$$

and as $\epsilon\downarrow 0$

$$\lim_{\epsilon\downarrow 0} E\big(Q(t_0,t;\epsilon)\big) \geqslant \frac{\rho^\dagger(t)}{1-\rho^\dagger(t)} .$$

But $\rho(t) = 1$, so we can take $\rho^\dagger(t)$ arbitrarily close to 1 which makes $\lim_{\epsilon\downarrow 0}E\big(Q(t_0,t;\epsilon)\big)$ blow up, showing that $Q(t_0,t)$ is oversaturated.

If the latter case holds, then it is easy to show that $\lim_{\epsilon\downarrow 0}|\mathbf{e}_n\mathbf{M}^*(t_*,t;\epsilon)^T|_1 \leqslant \frac{1}{2}$ and so $\lim_{\epsilon\downarrow 0}p_n^*(t_0,t;\epsilon) \geqslant \frac{1}{2}$. However, $p_n^*(t_0,t;\epsilon) = 1-(p_0(t_0,t;\epsilon)+\cdots+p_n(t_0,t;\epsilon))$, and so we have

$$\lim_{\epsilon\downarrow 0} p_0(t_0,t;\epsilon)+\cdots+p_n(t_0,t;\epsilon) \leqslant \tfrac{1}{2}$$

for all n. This means that $\lim_{\epsilon\downarrow 0}p_n(t_0,t;\epsilon)$ does not represent a probability distribution and, of course, $\lim_{\epsilon\downarrow 0}E\big(Q(t_0,t;\epsilon)\big)$ blows up.

[$\rho^*(t_0,t) < 1 \Rightarrow$ Undersaturation] Define $\mathbf{p}_k^*(t)$ for $k = 0,1,\ldots,n$ as follows:

$$\mathbf{p}_0^*(t) = -\lambda(t)\mathbf{e}_0\mathbf{A}^*(t)^{-1}, \qquad \mathbf{p}_{k+1}^*(t) = \left(\frac{d}{dt}\mathbf{p}_k^*(t)\right)\mathbf{A}^*(t)^{-1}. \tag{8.2}$$

As long as $\rho(t) < 1$, these are well defined l_1-vectors. This follows from the fact that the generating function of each $\mathbf{p}_k^*(t)$ will be a rational function. We then make use of a result derived in [5, Proposition 1.2.2].

It is sufficient here to assume that $n_0 = Q(t_0, t_0; \epsilon) = 0$ and $\rho(\cdot) < 1$ on (t_0, t). This holds since $\rho^*(t_0, t) < 1$ implies that $\rho(\cdot) < 1$ on some subinterval (t_*, t) and then

$$\begin{aligned} \mathbf{p}^*(t_0, t; \epsilon) &= \frac{1}{\epsilon} \int_{t_0}^{t} \lambda(s) \mathbf{e}_0 \mathbf{M}^*(s, t; \epsilon)\, ds + \mathbf{f}_{n_0} \mathbf{M}^*(t_0, t; \epsilon) \\ &= \frac{1}{\epsilon} \int_{t_*}^{t} \lambda(s) \mathbf{e}_0 \mathbf{M}^*(s, t; \epsilon)\, ds + \frac{1}{\epsilon} \int_{t_0}^{t_*} \lambda(s) \mathbf{e}_0 \mathbf{M}^*(s, t; \epsilon)\, ds + \mathbf{f}_{n_0} \mathbf{M}^*(t_0, t; \epsilon) \\ &\simeq \frac{1}{\epsilon} \int_{t_*}^{t} \lambda(s) \mathbf{e}_0 \mathbf{M}^*(s, t; \epsilon)\, ds. \end{aligned}$$

Define an l_1-valued process $\mathbf{r}_n^*(t; \epsilon)$ as $\mathbf{r}_n^*(t; \epsilon) = \mathbf{p}^*(t_0, t; \epsilon) - [\mathbf{p}_0^*(t) + \cdots + \epsilon^n \mathbf{p}_n^*(t)]$. By (8.2), we see that $\mathbf{r}_n^*(t; \epsilon)$ solves the equation

$$\begin{aligned} \epsilon \frac{\partial}{\partial t} \mathbf{r}_n^*(t; \epsilon) &= \mathbf{p}^*(t_0, t; \epsilon) \mathbf{A}^*(t) + \lambda(t) \mathbf{e}_0 \\ &\qquad - \left[\epsilon \frac{d}{dt} \mathbf{p}_0^*(t) + \cdots + \epsilon^n \frac{d}{dt} \mathbf{p}_{n-1}^*(t) \right] - \epsilon^{n+1} \frac{d}{dt} \mathbf{p}_n^*(t) \\ &= \mathbf{r}_n^*(t; \epsilon) \mathbf{A}^*(t) - \epsilon^{n+1} \frac{d}{dt} \mathbf{p}_n^*(t) \end{aligned}$$

and by Duhamel's principle

$$\begin{aligned} \mathbf{r}_n^*(t; \epsilon) &= -\epsilon^n \int_{t_0}^{t} \frac{d}{ds} \mathbf{p}_n^*(s) \mathbf{M}^*(s, t; \epsilon)\, ds + (\mathbf{e}_{n_0} - \mathbf{p}_0^*(t_0) - \cdots - \epsilon^n \mathbf{p}_n^*(t_0)) \mathbf{M}^*(t_0, t; \epsilon) \\ &\simeq -\epsilon^n \int_{t_0}^{t} \frac{d}{ds} \mathbf{p}_n^*(s) \mathbf{M}^*(s, t; \epsilon)\, ds \\ &= O(\epsilon^n). \end{aligned}$$

Therefore $\mathbf{r}_n^*(t; \epsilon) = \epsilon^{n+1} \mathbf{p}_{n+1}^*(t) + \mathbf{r}_{n+1}^*(t; \epsilon) = O(\epsilon^{n+1})$ and so with respect to the l_1 norm, $\mathbf{p}^*(t_0, t; \epsilon) = \mathbf{p}_0^*(\mathrm{t}) + \cdots + \epsilon^n \mathbf{p}_n^*(t) + O(\epsilon^{n+1})$ which proves the theorem. ∎

9. Time reversal formulas. For all s belonging to (t_0, t), define operators $\tilde{\mathbf{M}}^*$ and $\tilde{\mathbf{A}}^*$ as follows

$$\tilde{\mathbf{M}}^*(s, t_0; \epsilon) = \mathbf{M}^*(t_0, s; \epsilon)^T, \qquad \tilde{\mathbf{M}}^*(t, s; \epsilon) = \mathbf{M}^*(s, t; \epsilon)^T, \qquad \tilde{\mathbf{A}}^*(s) = \mathbf{A}^*(s)^T.$$

By these definitions, $\tilde{\mathbf{M}}^*$ satisfies

$$-\epsilon \frac{\partial}{\partial s} \tilde{\mathbf{M}}^*(s, t_0; \epsilon) = -\tilde{\mathbf{A}}^*(s) \tilde{\mathbf{M}}^*(s, t_0; \epsilon), \qquad -\epsilon \frac{\partial}{\partial s} \tilde{\mathbf{M}}^*(t, s; \epsilon) = \tilde{\mathbf{M}}^*(t, s; \epsilon) \tilde{\mathbf{A}}^*(s).$$

$\tilde{\mathbf{M}}^*$ can be viewed as an evolution operator just like $\mathbf{M}^*$ except that it evolves backwards in time. Also, $\tilde{\mathbf{A}}^* = \mu \mathbf{R} + \lambda \mathbf{L} - (\lambda + \mu) \mathbf{I}$ so, in addition to a time reversal, the roles of λ and μ are switched.

Now define a process $\tilde{Q}(t, t_0; \epsilon)$ that starts with an initial load of zero at time t, and progresses backwards in time to time t_0. Corresponding to $\tilde{Q}(t, t_0; \epsilon)$, $\tilde{\mathbf{p}}^*(t, t_0; \epsilon)$ can be written down as

$$\tilde{\mathbf{p}}^*(t, t_0; \epsilon) = \frac{1}{\epsilon} \int_{t_0}^{t} \mu(s) \mathbf{e}_0 \tilde{\mathbf{M}}^*(s, t_0; \epsilon)\, ds.$$

Taking transposes, and using $\mathbf{M}^*$ gives

$$\tilde{\mathbf{p}}^*(t,t_0;\epsilon)^T = \frac{1}{\epsilon}\int_{t_0}^{t} M^*(t_0,s;\epsilon)\mu(s)\mathbf{e}_0^T\,ds.$$

Proof of Theorem 3.2. We first derive the time reversal formula for $E(Q(t_0,t;\epsilon))$.

$$\begin{aligned}
E(Q(t_0,t;\epsilon)) &= \frac{1}{\epsilon}\int_{t_0}^{t}\lambda(s)\mathbf{e}_0\mathbf{M}^*(s,t;\epsilon)\mathbf{1}^T ds + \mathbf{f}_{n_0}\mathbf{M}^*(t_0,t;\epsilon)\mathbf{1}^T\\
&= \frac{1}{\epsilon}\int_{t_0}^{t}\lambda(s)\mathbf{e}_0\left[\mathbf{I} + \frac{1}{\epsilon}\int_{s}^{t}\mathbf{M}^*(s,r;\epsilon)\mathbf{A}^*(r)\,dr\right]\mathbf{1}^T ds\\
&\quad + \mathbf{f}_{n_0}\left[\mathbf{I} + \frac{1}{\epsilon}\int_{t_0}^{t}\mathbf{M}^*(t_0,s;\epsilon)\mathbf{A}^*(s)\,ds\right]\mathbf{1}^T\\
&= \frac{1}{\epsilon}\int_{t_0}^{t}\lambda(s)\,ds + \frac{1}{\epsilon^2}\int_{t_0}^{t}\int_{t_0}^{r}\mathbf{1}\mathbf{A}^*(s)\mathbf{M}^*(s,r;\epsilon)\mathbf{e}_0^T\,ds\,dr\\
&\quad + n_0 - \frac{1}{\epsilon}\mathbf{f}_{n_0}\int_{t_0}^{t}\mathbf{M}^*(t_0,s;\epsilon)\mu(s)\mathbf{e}_0^T\,ds\\
&= \frac{1}{\epsilon}\int_{t_0}^{t}\left[\lambda(s) - \mu(s)\right]ds + \frac{1}{\epsilon}\int_{t_0}^{t}\mathbf{1}\mathbf{M}^*(t_0,r;\epsilon)\mu(r)\mathbf{e}_0^T\,dr\\
&\quad + n_0 - \frac{1}{\epsilon}\mathbf{f}_{n_0}\int_{t_0}^{t}\mathbf{M}^*(t_0,s;\epsilon)\mathbf{e}_0^T\,ds\\
&= \frac{1}{\epsilon}\int_{t_0}^{t}\left[\lambda(s) - \mu(s)\right]ds + |\tilde{\mathbf{p}}^*(t,t_0;\epsilon)|_1 + n_0 - \mathbf{f}_{n_0}\cdot\tilde{\mathbf{p}}^*(t,t_0;\epsilon)^T
\end{aligned}$$

since $\mathbf{f}_{n_0}\cdot\tilde{\mathbf{p}}^*(t,t_0;\epsilon)^T = E(\tilde{Q}(t,t_0;\epsilon)\wedge n_0)$. The rest follows.

For $\mathrm{Var}(Q(t_0,t;\epsilon))$, use the differential formulas for $E(Q(t_0,t;\epsilon))$ and $\mathrm{Var}(Q(t_0,t;\epsilon))$, namely (3.1). By what we have derived so far and (3.1), we can say that

$$\epsilon\frac{\partial}{\partial t}E(\tilde{Q}(t,t_0;\epsilon)\vee n_0) = \mu(t)p_0(t_0,t\epsilon). \tag{9.1}$$

Apply $\partial/\partial t$ to the proposed time reversal formulas for the variance. Eliminating terms involving $\tilde{Q}$ via (9.1) and (3.3) will reduce this expression to the right-hand side of $(\epsilon\partial/\partial t)\mathrm{Var}(Q(t_0,t;\epsilon))$. One sees that this formula then satisfies the differential equation for the variance. It only remains to check that at $t = t_0$, the expression is zero since $\mathrm{Var}(Q(t_0,t_0;\epsilon)) = 0$. ∎

10. Stochastic dominance. Recall that a one-dimensional process $Q_1(t)$ is stochastically dominated by a process $Q_2(t)$ if, for all real numbers x and t, $\Pr\{Q_1(t) > x\} < \Pr\{Q_2(t) > x\}$. This type of ordering is very compatible with Markov processes as demonstrated in Kirstein, Franken, and Stoyan [4]. For comparing two **M(t)/M(t)/1** processes, a simple criterion can be established.

Theorem 10.1. *For $\lambda_1(\cdot) < \lambda_2(\cdot)$ and $\mu_1(\cdot) > \mu_2(\cdot)$ on (t_0,t), we have $\mathbf{p}_1^*(t_0,t;\epsilon) < \mathbf{p}_2^*(t_0,t;\epsilon)$.*

Before proving this theorem, we shall need two lemmas

Lemma 10.2. *With respect to $\mathbf{A}(t)$ as defined in §0, define $\mathbf{M}(t_0,t;\epsilon)$ as $\mathbf{M}^*(t_0,t;\epsilon)$ is defined with respect to $\mathbf{A}^*(t)$ in (0.5). If we think of $(\mathbf{I}-\mathbf{L})^{-1}$, in the sense of an unbounded operator, then $(\mathbf{I}-\mathbf{L})\mathbf{M}(t_0,t;\epsilon)(\mathbf{I}-\mathbf{L})^{-1} > 0$.*

PROOF. It is clear that $(\mathbf{I}-\mathbf{L})\mathbf{L}(\mathbf{I}-\mathbf{L})^{-1}=\mathbf{L}$ and $(\mathbf{I}-\mathbf{L})\mathbf{I}(\mathbf{I}-\mathbf{L})^{-1}=\mathbf{I}$. Just as with $\mathbf{A}^*$, $\mathbf{A}$ can be written in terms of the right and left shift operators as $\mathbf{A}=\lambda\mathbf{R}+\mu\mathbf{L}-\lambda\mathbf{I}-\mu\mathbf{L}\mathbf{R}$. It then follows that

$$\begin{aligned}(\mathbf{I}-\mathbf{L})\mathbf{A}(\mathbf{I}-\mathbf{L})^{-1}&=\lambda(\mathbf{I}-\mathbf{L})(\mathbf{R}-\mathbf{I})(\mathbf{I}-\mathbf{L})^{-1}+\mu(\mathbf{I}-\mathbf{L})(\mathbf{L}-\mathbf{L}\mathbf{R})(\mathbf{I}-\mathbf{L})^{-1}\\&=\lambda(\mathbf{I}-\mathbf{L})\mathbf{R}(\mathbf{I}-\mathbf{L})(\mathbf{I}-\mathbf{L})^{-1}-\mu(\mathbf{I}-\mathbf{L})\mathbf{L}\mathbf{R}(\mathbf{I}-\mathbf{L})(\mathbf{I}-\mathbf{L})^{-1}\\&=\lambda(\mathbf{I}-\mathbf{L})\mathbf{R}-\mu(\mathbf{I}-\mathbf{L})\mathbf{L}\mathbf{R}\\&=\lambda\mathbf{R}+\mu\mathbf{L}^2\mathbf{R}-(\lambda+\mu)\mathbf{L}\mathbf{R}\\&=\lambda\mathbf{R}+\mu\mathbf{L}^2\mathbf{R}+(\lambda+\mu)(\mathbf{I}-\mathbf{L}\mathbf{R})-(\lambda+\mu)\mathbf{I}.\end{aligned}$$

Therefore

$$(\mathbf{I}-\mathbf{L})\exp(\mathbf{A})(\mathbf{I}-\mathbf{L})^{-1}=e^{-(\lambda+\mu)}\exp\big(\lambda\mathbf{R}+\mu\mathbf{L}^2\mathbf{R}+(\lambda+\mu)(\mathbf{I}-\mathbf{L}\mathbf{R})\big)\geqslant 0$$

since $\mathbf{I}-\mathbf{L}\mathbf{R}\geqslant 0$. In a similar fashion to $\mathbf{M}^*$, $\mathbf{M}(t_0,t;\epsilon)$ is the limit of products of exponentials like $\exp(\mathbf{A})$ and so the rest holds. ∎

LEMMA 10.3. *Given* $\mathbf{M}_1(t_0,t;\epsilon)$ *and* $\mathbf{M}_2(t_0,t;\epsilon)$, *we can say that*

$$\mathbf{M}_2(t_0,t;\epsilon)-\mathbf{M}_1(t_0,t;\epsilon)=\frac{1}{\epsilon}\int_{t_0}^{t}\mathbf{M}_2(t_0,s;\epsilon)\big[\mathbf{A}_2(s)-\mathbf{A}_1(s)\big]\mathbf{M}_1(s,t;\epsilon)\,ds.$$

PROOF. We merely notice that

$$\epsilon\frac{\partial}{\partial s}\mathbf{M}_2(t_0,s;\epsilon)\mathbf{M}_1(s,t;\epsilon)=\mathbf{M}_2(t_0,s;\epsilon)\big[\mathbf{A}_2(s)-\mathbf{A}_1(s)\big]\mathbf{M}_1(s,t;\epsilon)$$

and then we integrate. ∎

PROOF OF THEOREM 10.1. First observe that $\mathbf{A}(\mathbf{I}-\mathbf{L})^{-1}=\lambda\mathbf{R}-\mu\mathbf{L}\mathbf{R}$. By Lemma 10.2

$$\begin{aligned}&\mathbf{M}_2(t_0,t;\epsilon)(\mathbf{I}-\mathbf{L})^{-1}-\mathbf{M}_1(t_0,t;\epsilon)(\mathbf{I}-\mathbf{L})^{-1}\\&\quad=\frac{1}{\epsilon}\int_{t_0}^{t}\mathbf{M}_2(t_0,s;\epsilon)\big[\mathbf{A}_2(s)(\mathbf{I}-\mathbf{L})^{-1}\mathbf{A}_1(s)(\mathbf{I}-\mathbf{L})^{-1}\big](\mathbf{I}-\mathbf{L})\mathbf{M}_1(s,t;\epsilon)(\mathbf{I}-\mathbf{L})^{-1}\,ds\\&\quad=\frac{1}{\epsilon}\int_{t_0}^{t}M_2(t_0,s;\epsilon)\big[(\lambda_2(s)-\lambda_2(s))\mathbf{R}-(\mu_2(s)-\mu_1(s))\mathbf{L}\mathbf{R}\big]\\&\qquad\times(\mathbf{I}-\mathbf{L})\mathbf{M}_1(s,t;\epsilon)(\mathbf{I}-\mathbf{L})^{-1}\,ds.\end{aligned}$$

All of these expressions are positive so $\mathbf{M}_1(t_0,t;\epsilon)(\mathbf{I}-\mathbf{L})^{-1}\leqslant\mathbf{M}_2(t_0,t;\epsilon)(\mathbf{I}-\mathbf{L})^{-1}$. Applying $\mathbf{e}_{n_0}$ to both sides gives $\mathbf{p}_1(t_0,t;\epsilon)(\mathbf{I}-\mathbf{L})^{-1}\leqslant\mathbf{p}_2(t_0,t;\epsilon)(\mathbf{I}-\mathbf{L})^{-1}$. By the definition of $\mathbf{p}^*$, $\mathbf{p}_1^*(t_0,t;\epsilon)\leqslant\mathbf{p}_2^*(t_0,t;\epsilon)$ and so we are done. ∎

11. Appendix.

We now calculate the zeroth and the first order terms of the distribution, mean, and variance for the undersaturated region. By §8, we see that it is sufficient to determine $\mathbf{p}_0^*(t)$ and $\mathbf{p}_1^*(t)$ which solve

$$\mathbf{p}_0^*(t)\mathbf{A}^*(t)=-\lambda(t)\mathbf{e}_0,\qquad\mathbf{p}_1^*(t)\mathbf{A}^*(t)=\frac{d}{dt}\mathbf{p}_0^*(t).$$

We shall derive them using generating functions.

If g is an l_1-vector and $g(\sigma)$ is its generating function, then

$$\left(\mathbf{g}(\mathbf{A}^*)^{-1}\right)(\sigma) = \frac{\sigma g(\sigma) - g(1)}{(\mu - \lambda\sigma)(1-\sigma)}\,.$$

Let $p_0^*(t,\sigma)$ $(p_1^*(t,\sigma))$ be the generating function of $\mathbf{p}_0^*(t)$ $(\mathbf{p}_1^*(t))$. Since $\mathbf{p}_0^*(t) = -\lambda(t)\,\mathbf{e}_0\mathbf{A}^*(t)^{-1}$, we have

$$\begin{aligned} p_0^*(t,\sigma) &= -\lambda(t)\cdot\frac{\sigma-1}{(\mu(t)-\lambda(t)\sigma)(1-\sigma)} \\ &= \frac{\rho(t)}{1-\rho(t)\sigma} \\ &= \sum_{n=0}^{\infty}\rho(t)^{n+1}\sigma^n. \end{aligned}$$

It follows that

$$\frac{\partial}{\partial t}\,p_0^*(t,\sigma) = \frac{\rho'(t)}{(1-\rho(t)\sigma)^2}$$

and so if $\mathbf{p}_1^*(t) = (d\mathbf{p}_0^*(t)/dt)\mathbf{A}^*(t)^{-1}$, we then have

$$\begin{aligned} p_1^*(t,\sigma) &= \frac{\sigma\cdot\dfrac{\rho'(t)}{(1-\rho(t)\sigma)^2} - \dfrac{\rho'(t)}{(1-\rho(t))^2}}{(\mu(t)-\lambda(t)\sigma)(1-\sigma)} \\ &= \frac{\rho'(t)}{\mu(t)(1-\rho(t))^2}\cdot\frac{\sigma\left(\dfrac{1-\rho(t)}{1-\rho(t)\sigma}\right)^2 - 1}{(1-\rho(t)\sigma)(1-\sigma)} \\ &= \frac{\rho'(t)}{\mu(t)(1-\rho(t))^2}\,\frac{\sigma(1-\rho(t))^2 - 1 + 2\rho(t)\sigma - \rho(t)^2\sigma^2}{(1-\rho(t)\sigma)^3(1-\sigma)} \\ &= \frac{\rho'(t)}{\mu(t)(1-\rho(t))^2}\,\frac{-\rho(t)^2\sigma^2 + \left(1+\rho(t)^2\right)\sigma - 1}{(1-\rho(t)\sigma)^3(1-\sigma)} \\ &= \frac{\rho'(t)}{\mu(t)(1-\rho(t))^2}\,\frac{\rho(t)^2\sigma - 1}{(1-\rho(t)\sigma)^3} \\ &= \frac{\rho'(t)}{\mu(t)(1-\rho(t))^2}\cdot\sum_{n=0}^{\infty}\left[\frac{n(n+1)}{2}\rho(t) - \frac{(n+1)(n+2)}{2}\right]\rho(t)^n\sigma^n. \end{aligned}$$

Recall that $p_n(t_0, t;\epsilon) = p^*_{n-1}(t_0, t;\epsilon) - p^*_n(t_0, t;, \epsilon)$ where $p^*_{-1}(t_0, t;\epsilon) \equiv 1$. We then have

$$\begin{aligned} p_n(t_0, t;\epsilon) &= \rho(t)^n - \rho(t)^{n+1} \\ &\quad + \epsilon \frac{\rho'(t)}{\mu(t)(1-\rho(t))^2} \left[\frac{n(n-1)}{2}\rho(t) - \frac{n(n+1)}{2} - \frac{n(n+1)}{2}\rho(t)^2 \right. \\ &\qquad \left. + \frac{(n+1)(n+2)}{2}\rho(t) \right] \rho(t)^{n-1} + O(\epsilon^2) \\ &= (1-\rho(t))\rho(t)^n \\ &\quad + \epsilon \frac{\rho'(t)}{\mu(t)(1-\rho(t))^2} \left[(n^2+n+1)\rho(t) \right. \\ &\qquad \left. - \frac{n(n+1)}{2}\left(1+\rho(t)^2\right) \right] \rho(t)^{n-1} + O(\epsilon^2) \\ &= (1-\rho(t))\rho(t)^n + \epsilon \frac{\rho'(t)}{\mu(t)} \left[\frac{\rho(t)}{(1-\rho(t))^2} - \frac{n(n+1)}{2} \right] \rho(t)^{n-1} + O(\epsilon^2). \end{aligned}$$

By use of these generating functions we have

$$\begin{aligned} E(Q(t_0, t;\epsilon)) &= p^*_0(t,1) + \epsilon p^*_1(t,1) + O(\epsilon^2) \\ &= \frac{\rho(t)}{1-\rho(t)} - \epsilon \frac{\rho'(t)}{\mu(t)} \frac{1+\rho(t)}{(1-\rho(t))^4} + O(\epsilon^2) \end{aligned}$$

Moreover, we get

$$E\left(Q(t_0, t;\epsilon)^2\right) = E(Q(t_0, t;\epsilon)) + 2 \cdot \frac{\partial p^*_0}{\partial \sigma}(t,1) + \epsilon \cdot 2 \frac{\partial p^*_1}{\partial \sigma}(t,1) + O(\epsilon^2) \qquad \text{where}$$

$$\frac{\partial p^*_0}{\partial \sigma}(t,\sigma) = \frac{\rho(t)^2}{(1-\rho(t)\sigma)^2} \quad \text{and}$$

$$\frac{\partial p^*_1}{\partial \sigma}(t,\sigma) = \frac{\rho'(t)}{\mu(t)(1-\rho(t))^2} \left[\frac{\rho(t)^2}{(1-\rho(t)\sigma)^3} + \frac{3\rho(t)\left(\rho(t)^2\sigma - 1\right)}{(1-\rho(t)\sigma)^4} \right].$$

Consequently,

$$\frac{\partial p^*_0}{\partial \sigma}(t,1) = \frac{\rho(t)^2}{(1-\rho(t))^2}, \qquad \frac{\partial p^*_1}{\partial \sigma}(t,1) = \frac{-\rho'(t)}{\mu(t)} \frac{2\rho(t)^2 + 3\rho(t)}{(1-\rho(t))^3}.$$

Squaring our expansion for the mean gives us

$$E(Q(t_0, t;\epsilon))^2 = \frac{\rho(t)^2}{(1-\rho(t))^2} - \epsilon \frac{\rho'(t)}{\mu(t)} \frac{2\left(\rho(t) + \rho(t)^2\right)}{(1-\rho(t))^5} + O(\epsilon^2).$$

Finally, combining terms gives us

$$\begin{aligned}\operatorname{Var}(Q(t_0,t;\epsilon)) &= \frac{2\rho(t)^2}{(1-\rho(t))^2}+\frac{\rho(t)}{1-\rho(t)}-\frac{\rho(t)^2}{(1-\rho(t))^2}\\ &\quad+\epsilon\frac{\rho'(t)}{\mu(t)}\,\frac{-4\rho(t)^2-6\rho(t)+\rho(t)^2-1+2\big(\rho(t)+\rho(t)^2\big)}{(1-\rho(t))^5}+O(\epsilon^2)\\ &= \frac{\rho(t)}{(1-\rho(t))^2}-\epsilon\frac{\rho'(t)}{\mu(t)}\,\frac{\rho(t)^2+4\rho(t)+1}{(1-\rho(t))^5}+O(\epsilon^2).\end{aligned}$$

Acknowledgements. The author thanks Joseph B. Keller who, as my thesis advisor at Stanford University, originally suggested this problem. Thanks are also due to B. Gopinath of Bell Communications Research for suggesting the use of stochastic dominance.

Bibliography

[1] Clarke, A. B. (1956). Waiting Line Process of Markov Type. *Ann. Math. Statist.*, **27** 452–459.

[2] Gross, D. and Harris, C. M. (1974). *Fundamentals of Queueing Theory*. John Wiley and Sons, New York.

[3] Keller, J. B. (1982). Time-Dependent Queues. *SIAM Rev.* **24**, 4 401–412.

[4] Kirstein, B. M., Franken, P. and Stoyan, D. (1977). Comparability and Monotonicity of Markov Processes. *Theory of Probab. Appl.* **22** 40–52.

[5] Massey, W. A. (1981). Non-Stationary Queues. Ph.D. Dissertation, Mathematics Department of Stanford University.

[6] Newell, G. F. (1968). Queues with Time-Dependent Arrival Rates. (I, II, III). *J. Appl. Probab.* **5** 436–451 (I); 579–590 (II); 591–606 (III).

[7] Rider, K. L. (1976). A Simple Approximation to the Average Queue Size in the Time Dependent M/M/1 Queue. *J. Assoc. Comput. Mach.* **23**, 2 361–367.

[8] Rothkopf, M. H. and Oren, S. S. (1979). A Closure Approximation for the Nonstationary $M/M/s$ Queue. *Management Sci.* **25** 522–534.

[9] Yoshida, K. (1978). *Functional Analysis*. Fifth Edition. Springer Verlag, New York and Berlin.

AT&T BELL LABORATORIES, 600 MOUNTAIN AVENUE, MURRAY HILL, NEW JERSEY 07974