The Error in Steady-State Approximations for the Time-Dependent Waiting Time Distribution

Samuel G. Steckley and Shane G. Henderson

School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853, U.S.A.

Abstract

Arrival processes to queues often exhibit time dependence, but time-dependent queues are notoriously difficult to analyze. Usually one adopts some kind of steady-state approximation for time-dependent performance. We consider a range of queueing models that see wide use in practice, particularly in the modeling of service industry applications, and develop approximations for the error involved in using a steady-state approximation for time-dependent quantities. We focus on the distribution of the waiting time before reaching service, although the approach we use could be extended to many other performance measures. The results reinforce and extend what is known about such errors. For example, customer abandonment leads to dramatic reductions in the error in steady-state approximations as compared to a system without customer abandonment.

Key words: Poisson’s equation, nonhomogeneous Poisson process, call center

1 Introduction

Queueing models are used in a tremendous variety of applications to gain insight and to obtain predictions of performance. A typical application involves selecting the service capacity so as to ensure that customers (jobs) do not wait too long before receiving service. Important examples of such applications include staffing decisions for inbound call centers and for emergency services.
The models typically used in such applications, and the ones we focus on here, model a system where calls arrive to a single queue, are served by one of several servers in FIFO order, and depart the system. We also model customer impatience, where customers leave the system (abandon) without receiving service if they are delayed too long in the queue before reaching a server.

It is often the case that the customer arrival process is well modeled by a Poisson process with a time-varying rate function. Service times have also been observed to vary with time, e.g., Brown et al. [2005], but usually not as dramatically as the arrival rate. It is usually too difficult to compute performance measures for time-varying systems exactly, and so approximations are necessary. The most common approximations involve using steady-state results for queueing models with appropriately chosen parameters. In particular, the pointwise stationary approximation - PSA [Green and Kolesar, 1991] approximates the distribution of queue sizes at time \( t \) by the steady-state distribution of a queueing system with fixed parameters corresponding to the instantaneous arrival rate and service time parameters at time \( t \). (Strictly speaking, the PSA involves integrating a formula for a steady-state performance measure corresponding to instantaneous rates over a time interval.) This method is certainly not exact. Consider, for example, an \( M/M/1 \) queue where the instantaneous arrival rate temporarily exceeds the service capacity. In this case, the approximation suggests an infinite customer population, which clearly does not coincide with what happens in the time-dependent system. Even for models where customer abandonment stabilizes the system, so that an arrival rate exceeding the service capacity is no problem, the approximation is still not exact. Slight modifications to improve the approximation have also been proposed; see Green et al. [2006] for an accessible overview, and for further details see, e.g., Green et al. [2001] and Jennings et al. [1996]. The key observation in this work is that congestion tends to lag behind the peaks in the arrival rate, so one should use a lagged version of the arrival rate in the PSA. A related approximation is developed in Stolletz [2006], where results for loss systems are used to modify the offered load.

When are such approximations reasonable? Whitt [1991] showed that the PSA is exact in the limit as the rates increase. Massey and Whitt [1998] show that the PSA is, in fact, the leading term in an expansion that gives exact time-dependent performance. Green et al. [2001] showed that a slight modification of the PSA involving the lagging operation mentioned above is usually accurate enough for practical purposes, although there are situations when it commits practically significant errors. In this paper we shed further light on the question of when the approximation can be expected to be accurate through a stylized model. Our focus is on the arrival process, although one could also treat a time-varying service rate in similar fashion. We focus on a single performance measure, namely the distribution of customer waiting time in the queue, not counting service. We do so because this is a standard measure used in a number of industries including emergency services and call centers. One could also employ our methodology to explore other performance measures, but we do not do so here.

Performance measures for time-varying systems are usually calculated on a per-period basis. That is, one computes performance measures for each period of the planning horizon. These periods are typically short, e.g., 30 minutes. We focus on a single period denoted \( [0, t] \), and the immediately
preceding period. The customer arrival process is assumed to be nonhomogeneous Poisson, with piecewise constant arrival rate that is constant within each period.

Although our analysis focuses on a piecewise constant arrival rate function, we believe that in most applications the arrival rate function is continuous. Our model should then be viewed as a stylized model that allows us to compute the error approximation, and thereby obtain considerable insight into the quality of the steady-state approximation for time-varying systems. By studying this stylized model with a piecewise constant arrival rate, we obtain results that exhibit the same trends that we have observed in simulations using continuous arrival rate functions. In that sense, we extend our previous understanding of when steady-state approximations are accurate. However, we do not claim that our results apply directly to the continuous-arrival-rate case.

The performance measure we focus on is a standard one in the service industry. We study $\alpha$, the probability that a customer reaches a server within $\tau$ time units, assuming that the customer is willing to wait at least that long. More precisely, suppose there are $c$ servers, and let $X(u)$ be the number of customers in the system at time $u$, $u \geq 0$. Let $(X_u(t) : t \geq u)$ denote the number-in-system process when the customer arrival process is stopped subsequent to time $u$. If we let $V(u)$ be the time until a server becomes free at time $u$ in the system with no further customer arrivals, so that $V(u) = \inf\{s \geq 0 : X_u(u + s) \leq c - 1\}$, then

$$\alpha = \frac{1}{t} \int_0^t P(V(u) \leq \tau) \, du. \tag{1}$$

To the extent possible we allow any $\tau \geq 0$, but in several cases we specialize to the $\tau = 0$ case, which corresponds to the probability that a customer reaches a server immediately.

The remainder of this paper is organized as follows. In Section 2 we describe the model more precisely and develop the error approximation, outlining the key steps in its construction. Section 3 studies the approximation in the context of models with no customer abandonment. These models are particularly appropriate when customers are highly reluctant to abandon, as in emergency call centers for example. Section 4 addresses models with customer abandonment. These models see extensive use in call center modeling. The main conclusions are discussed in Section 5, and Section 6 collects some technical results used to justify the approximation procedure.

2 The Error Approximation

We consider a $c$-server queue where customers arrive according to a nonhomogeneous Poisson process, are served in first-in-first-out order, and have i.i.d. exponentially distributed service times with rate $\mu$ that are independent of the arrival process. The arrival rate is constant over the period $[0, t]$ with rate $\lambda$. In Section 4 we assume, in addition, that customers have limited patience in that each customer is willing to wait an exponentially distributed amount of time with rate $\theta$ after they arrive, independent of all else. If this time elapses before they reach service, then they
abandon the system without receiving service.

Let $X(\cdot)$ and $V(\cdot)$ be the number in system and time-to-reach-service processes respectively, as defined in the introduction. Under the above assumptions, $X(\cdot)$ is a continuous time Markov chain (CTMC) on $[0, t]$. Let $\nu$ indicate the distribution of $X(0)$, and $P_\nu$ and $E_\nu$ represent the probability and expectation associated with the Markov process $X(\cdot)$ with initial distribution $\nu$. Recall that we are interested in computing (1) for some $\tau \geq 0$. We can write

$$\alpha = \frac{1}{t} \int_0^t P_\nu(V(u) \leq \tau) \, du.$$ 

Suppose that $X(\cdot)$ is irreducible and positive recurrent and so possesses a unique stationary probability distribution $\pi$. If $\nu = \pi$, so that $X(\cdot)$ is in steady-state at time 0, then $P_\nu(V(u) \leq \tau)$ is a steady-state quantity that does not depend on $u$. Let $\bar{\alpha}$ denote this quantity, and note that

$$\frac{1}{t} \int_0^t P_\pi(V(u) \leq \tau) \, du = \bar{\alpha}.$$ 

Then $\bar{\alpha}$ is a steady-state approximation for $\alpha$, and the error in the approximation is

$$\alpha - \bar{\alpha} = \frac{1}{t} \int_0^t [P_\nu(V(u) \leq \tau) - \bar{\alpha}] \, du$$

$$\approx \frac{1}{t} \int_0^\infty [P_\nu(V(u) \leq \tau) - \bar{\alpha}] \, du$$

$$\triangleq \Gamma(\nu, t).$$

The expression $\Gamma(\nu, t)$ is our approximation for the error in using the steady state value $\bar{\alpha}$ to approximate the true value $\alpha$. To understand why we replaced $t$ by $\infty$ in (2) note that

$$P_\nu(V(u) \leq \tau) \rightarrow \bar{\alpha}$$

as $u \rightarrow \infty$ (as can be seen by conditioning on $X(u)$). If the convergence is sufficiently fast, then the term

$$\int_0^\infty [P_\nu(V(u) \leq \tau) - \bar{\alpha}] \, du$$

is finite and converges to 0 as $t \rightarrow \infty$, so that the approximation only involves a “small” correction. (This is shown rigorously in Lemma 2 in Section 6). The systems we study in this paper all converge exponentially rapidly to steady state, in the sense that they are $V$-uniformly ergodic (see Meyn and Tweedie [1993, Chapter 15, 16] for the general theory) so that the tail integral (3) converges to 0 at an exponential rate. Furthermore, it is well known that, in general, lightly-loaded systems converge to steady-state very rapidly, so the approximation is very good for such systems. In addition, systems with abandonment behave somewhat like infinite-server systems, which again
converge to steady-state extremely rapidly. So in all of these settings we expect the approximation
to be very good. Having said that, we expect that the approximation is not as good in heavily
loaded systems without abandonment. In that case, the approximation should probably be viewed
as less accurate, but still indicating general trends.

But how do we compute the error approximation \( \Gamma(\nu, t) \)?

For clarity we defer the justification of some of the technical steps to Section 6. Let \( P_x \) and \( E_x \)
denote the probability and expectation when the Markov process \( X(\cdot) \) is initiated at \( X(0) = x \)
with probability 1. Define

\[
\begin{align*}
g(x) &= \int_{0}^{\infty} [P_x(V(u) \leq \tau) - \bar{\alpha}] \, du \\
&= \int_{0}^{\infty} [E_x P(V(u) \leq \tau | X(u)) - \bar{\alpha}] \, du \\
&= \int_{0}^{\infty} [E_x h(X(u)) - \bar{\alpha}] \, du,
\end{align*}
\]

(4)

(5)

where

\[
h(y) = P(V(u) \leq \tau | X(u) = y).
\]

For example, if \( \tau = 0 \), then \( h(y) = I(y \leq c - 1) \), where \( I(\cdot) \) is the indicator function that is 1 if
its argument is true and 0 otherwise. We need to show that \( g \) is well defined, but defer this to
Section 6.

So now,

\[
\begin{align*}
t \Gamma(\nu, t) &= \int_{0}^{\infty} [P_\nu(V(u) \leq \tau) - \bar{\alpha}] \, du \\
&= \int_{0}^{\infty} [E_\nu P(V(u) \leq \tau | X(u)) - \bar{\alpha}] \, du \\
&= \int_{0}^{\infty} [E_\nu h(X(u)) - \bar{\alpha}] \, du \\
&= \int_{0}^{\infty} E_\nu [E[h(X(u)) | X(0)] - \bar{\alpha}] \, du \\
&= E_\nu \int_{0}^{\infty} [E[h(X(u)) | X(0)] - \bar{\alpha}] \, du \\
&= E_\nu g(X(0)).
\end{align*}
\]

(6)

(7)

We need to justify the interchange (6), but defer that to Section 6. Hence, to compute the ap-
proximation \( \Gamma(\nu, t) \) it suffices to compute \( g \), and then compute the expectation (7).

We need to explain how we compute \( g \). We show in Section 6 that \( g \) solves the system of equations

\[
Ag(x) = -(h(x) - \bar{\alpha}) \quad \forall x,
\]

(8)
where $A$ is the rate matrix of the CTMC $X(\cdot)$. This system of equations is known as Poisson’s equation with right-hand side function $h$, and $g$ is therefore the solution to Poisson’s equation. The matrix $A$ is the rate matrix of a birth-death process, so this set of equations can be solved numerically using straightforward recursions. We also need some side conditions to identify the constants that arise in this solution. The key ones we require are that $g$ is $\pi$-integrable, i.e., $E_\pi |g(X(0))| < \infty$ and $E_\pi g(X(0)) = 0$. These results are proved in Section 6.

It remains to specify $\nu$. Recall that we assume a piecewise constant arrival rate function. It seems reasonable to take $\nu$ as the steady-state distribution associated with the parameters of the process in the previous period. Given our focus on errors associated with a time-dependent arrival rate, the most important of these parameters is the arrival rate $\lambda_0$ in that period. Our earlier discussion related to the rate of convergence of systems to steady state applies equally well here, and therefore clarifies when this approximation for $\nu$ should be reasonable.

The agenda for the sections to follow is to consider a variety of models. For each model we first compute $g$ for an arrival rate $\lambda$, compute $\nu$ for the arrival rate $\lambda_0$ prior to time 0, and then compute $E_\nu g(X(0))$. This then gives the error approximation $\Gamma(\nu, t)$ from (7). The result can be obtained analytically in some cases, and numerically in others.

### 3 Models Without Abandonment

#### 3.1 The $M/M/1$ queue

We first consider the $M/M/1$ queue, partly to demonstrate the methodology in a transparent setting, and partly because the conclusions we draw from this model can be extrapolated to more realistic models. We restrict attention to the $\tau = 0$ case here, so that the goal is to immediately answer calls.

The rate matrix $A$ of the CTMC $X(\cdot)$ has non-zero off-diagonal elements $A_{i,i+1} = \lambda$ and $A_{i+1,i} = \mu$, $i \geq 0$. An arriving call is immediately answered if the system is empty, so $h(x) = I(x = 0)$. If $\rho \triangleq \lambda/\mu < 1$ then $X$ has a steady-state distribution $\pi$ where $\pi_i \triangleq \pi(\{i\}) = (1 - \rho)\rho^i$, so that $\bar{\alpha} = 1 - \rho$. Poisson’s equation is then

$$Ag(x) = -(I(x = 0) - (1 - \rho)) \quad \forall x \geq 0.$$  

This set of difference equations has $\pi$-integrable solution $g(x) = \kappa - x/\mu$, and since $\pi g = 0$ the constant $\kappa = \rho/\mu(1 - \rho)$.

This expression for $g$ confirms the intuitive notion that the error is smallest when the initial state is close to typical steady-state values ($g(x) = 0$ for $x = E_\pi X(0)$). For smaller values of $x$, $g(x)$ is positive, showing that if the initial state is small relative to steady-state conditions, then the
true performance is greater than the steady-state approximation. This again is consistent with the intuition that suggests that performance should be better when the system has few customers in it at the start of the period. The reverse applies when \( x \) is larger than typical steady-state values.

But how significant are these biases? To answer that question we compute \( \nu g \triangleq E_{\nu} g(X(0)) \), where \( \nu \) is the stationary distribution for the parameters associated with the previous period. Suppose that the previous arrival rate is \( \lambda_0 \), and let \( \rho_0 = \lambda_0 / \mu \). A direct calculation shows that

\[
\nu g = \frac{\rho - \rho_0}{\mu(1 - \rho)(1 - \rho_0)},
\]

so that the relative error (relative to the steady-state approximation \( \bar{\alpha} \)) is

\[
\frac{\nu g}{\bar{\alpha} t} = \frac{1}{\mu t} \frac{\rho - \rho_0}{1 - \rho_0} \frac{1}{(1 - \rho)^2}.
\] (9)

From (9) we see that if \( \rho \) is close to \( \rho_0 \) then the relative error is negligible. This is expected since there is then little change in the parameters of the system at the start of the period. The sign of the error is the same as the sign of \( \rho - \rho_0 \), which means that true performance is better than the approximation when \( \rho_0 < \rho \), again as expected. When \( \mu t \) is large, i.e., when the expected number of service completions when the server remains busy over the entire period is large, the error is reduced. This latter observation reinforces the results of Whitt [1991]. (Whitt showed only that the error reduced to 0 as the rates grew without bound, so we have been able to quantify that important observation.) The error is magnified by the factor \((1 - \rho)^{-2}\), suggesting that when the system enters heavy traffic the errors can be significant. Interestingly, when the system leaves heavy traffic (so that \( \rho_0 \) is close to 1 but \( \rho \) is not), the error is of the order \((1 - \rho_0)^{-1}\), which is an order of magnitude smaller than when the system enters heavy traffic. We see exactly this behavior in Figure 2 for general \( M/M/c \) queues.

3.2 The \( M/M/c \) queue

The results for the \( M/M/1 \) queue offer insight, but are they representative of multi-server systems? Let us now consider the \( M/M/c \) queue with \( c \) servers, arrival rate \( \lambda \), service rate \( \mu \) and \( \rho = \lambda / c \mu < 1 \). It is important to consider this case, because it is extensively used in the call center industry, and is especially appropriate when customers are reluctant to abandon, as occurs, for example, in emergency services.

Again the number of customers in the system is a CTMC, with a well-known rate matrix (e.g., Wolff [1989, p. 256]). A customer immediately enters service if the number of customers seen on arrival is \( c - 1 \) or less. Accordingly \( h(x) = I(x \leq c - 1) \) and \( \bar{\alpha} = \pi_0 + \cdots + \pi_{c-1} \). It is possible to show that the solution to Poisson’s equation is of the form \( \kappa - x/(c \mu) \) for \( x > c \), with a more complicated form for \( x \leq c \). The expressions are cumbersome and difficult to extract meaning.
from. We instead perform the calculations outlined above numerically. Some care is needed in the calculations due to numerical instabilities.

A representative plot of $g$ is given in Figure 1. Without loss of generality we took $c\mu = 1$. (This is merely a choice of time scale.) We see similar behavior to the $M/M/1$ solution in that $g$ is decreasing and crosses 0 near the steady-state mean, which is approximately 56 for this example. Notice also the clear linear growth beyond $c$. The slight “kink” near $x = 0$ is due to the use of a “guess” to replace unreliable values due to numerical difficulties.

Fig. 1. The solution to Poisson’s equation for the $M/M/c$ queue where $\lambda = 0.94$, $\mu = 0.02$ and $c = 50$

A representative plot of the relative error for various parameter values is given in Figure 2. As with the $M/M/1$ model we see dramatic increases in relative error as $\rho \to 1$ that are tempered near the line $\rho = \rho_0$. The sign of the error coincides with the sign of $\rho - \rho_0$, as intuition suggests. For fixed $\rho$ and increasing $\rho_0$ the (negative) error increases in absolute value, again agreeing with intuition that says that if the queue is very long at the start of the current period then we can expect true performance to be poor even if the steady-state values indicate otherwise. For fixed $\rho_0$ and increasing $\rho$ the relative error grows explosively.

The explosive behavior of the relative error in heavy traffic is similar to that seen in the $M/M/1$ queue, although we cannot quantify the rate of growth with our numerical results. But we can quantify and understand this phenomenon using diffusion approximations. The process giving the number of customers in the system over time in a heavily-loaded $M/G/c$ queue with finite service-time variance can be approximated by a regulated (reflected) Brownian motion with appropriate parameters [Iglehart and Whitt, 1970a,b]. This approximation is established through a rigorous limit theorem that arises as the arrival rate approaches the maximum service capacity $c\mu$. This is known as the “efficiency-driven regime” since it reflects the notion that agents are very heavily utilized. Agents are only very rarely free, and so the behavior of the approximation is dominated by the behavior of the original birth-death process above the level $c$. 

8
Fig. 2. The relative error for the $M/M/c$ queue expressed as a percentage with $\mu = 0.02$, $c = 50$, and values of $\lambda = \rho$ and $\lambda_0 = \rho_0$ in $[0.04, 0.94]$

The approximation is $X(\cdot) \approx Y(\cdot; -\gamma, \sigma^2)$, where $\gamma = c\mu - \lambda$, $\sigma^2 = c\mu(c^2 + \rho)$ and $c^2$ is the squared coefficient of variation of the service times (see, e.g., Whitt [2002, §10.2]). Here $Y(\cdot; -\gamma, \sigma^2)$ is a regulated Brownian motion with drift $-\gamma$ and infinitesimal variance $\sigma^2$. The process $Y$ lives on the state space $[0, \infty)$. We take $h(x) = I(x \leq b)$ for some $b \approx c - 1$. When $\gamma > 0$, $Y$ has a stationary distribution that is exponential with mean $\eta^{-1} \triangleq \sigma^2/2\gamma$. Therefore $\bar{\alpha} = P_\pi(Y(0) \leq b) = 1 - e^{-\eta b}$.

Poisson’s equation is then

$$\frac{\sigma^2}{2} g''(x) - \gamma g'(x) = -[I(x \leq b) - \bar\alpha]$$

$$g'(0) = 0. \tag{10}$$

See Karlin and Taylor [1981, Chapter 15] and Glynn [1990] for background on diffusions and equations of this type.

The solution to (10) is [Glynn and Torres, 1996]

$$g(x) = \frac{b}{\gamma} e^{-\eta b} + \begin{cases} \frac{e^{-\eta b}(e^{\eta x} - \eta x - 1)}{-\gamma^2} & 0 \leq x \leq b, \\ \frac{e^{-\eta b}(e^{\eta b} - \eta b - 1) + \frac{1 - e^{-\eta b}}{-\gamma}(x - b)}{-\gamma} & x > b, \end{cases}$$

where we have used the fact that $\pi g = 0$ to compute an additive constant. Notice that the solution is decreasing, and beyond $b$ is linear, exactly as in the $M/M/c$ model.

Next we compute the relative error by first computing $\nu g$, where $\nu$ corresponds to the stationary distribution associated with the parameters of the previous period. Letting $\eta_0^{-1}$ denote the steady-
state mean occupancy in the previous period, algebra reveals that

\[ \nu_g = \frac{be^{-\eta b}}{\gamma} + \frac{\eta}{\gamma \eta_0} \left( e^{-\eta b} - e^{-(\eta - \eta_0)b} \right). \]

This expression is somewhat complicated. To obtain insight, we use Taylor approximations to approximate the exponential terms. Note that \( \eta = 2c(1 - \rho)/(c_s^2 + \rho) \), with a similar expression for \( \eta_0 \). The approximations \( e^{-\eta b} \approx 1 \) and \( e^{-\eta b} - e^{-\eta_0 b} \approx (\eta - \eta_0)b \) are then valid when \( \rho, \rho_0 \) are close to 1, and give

\[
\nu_g \approx \frac{b}{\gamma} \left( 1 - \frac{\eta}{\eta_0} \right) \approx \frac{b(c_s^2 + 1)}{c_s^2 + \rho} \frac{\rho - \rho_0}{\mu(1 - \rho)(1 - \rho_0)}, \tag{11}
\]

where (11) follows since \( \sigma^2 = c\mu(c_s^2 + \rho) \) and \( \sigma_0^2 = c\mu(c_s^2 + \rho_0) \).

Finally, another Taylor expansion gives

\[
\bar{\alpha} t = (1 - e^{-\eta b})t \\
\approx \eta bt \\
= \frac{2bt(1 - \rho)}{c_s^2 + \rho},
\]

so that

\[
\frac{\nu g}{\bar{\alpha} t} \approx \frac{1 + c_s^2}{2} \frac{1}{c\mu(1 - \rho_0)(1 - \rho)^2}. \tag{12}
\]

There is a close correspondence between the expressions (12) and (9). This suggests that the same observations made for the \( M/M/1 \) model also apply for models that can be approximated by RBM. This includes a large class of multi-server queues that includes the \( M/G/c \) family of queues with finite service time variance. Service time variability is exhibited through the factor \( (1 + c_s^2)/2 \), so that the relative error increases with service time variability.

A criticism of the efficiency-driven regime just explored is that the approximations are most valid when agents are busy almost constantly and customer waiting times are large or excessive. Given a choice, one would usually prefer to operate in the so-called “quality and efficiency driven” regime, which is also known as the Halfin-Whitt regime in honor of Halfin and Whitt [1981]. In this regime, the servers are highly utilized and customers receive a high level of service.

While a study of Figure 2 reveals some insights about this case, further insight can be obtained by studying a diffusion approximation that is relevant in the Halfin-Whitt regime. The study of this model can be further motivated by noting that the RBM approximation given in the previous
section is a special case of the approximation given here when the arrival rate is excessively high relative to the maximum service rate.

The appealing properties of the Halfin-Whitt regime come at a cost: the diffusion approximation applies for only a small family of service-time distributions. We restrict attention to exponential service times here.

The approximation is (see Halfin and Whitt [1981], and Whitt [2002, p. 359] for an update)

\[
X(\cdot) \approx c + \sqrt{c}Y(\cdot; \beta),
\]

where \( \beta = \sqrt{c(1 - \rho)} > 0 \). Here \( Y(\cdot; \beta) \) is a diffusion on \(( -\infty, \infty )\) with drift function

\[
\mu(x) = \begin{cases} 
-\mu \beta & x \geq 0 \\
-\mu(\beta + x) & x < 0 
\end{cases}
\]

and constant infinitesimal variance \( 2\mu \). The approximation is justified by taking a limit as the number of servers \( c \) and the arrival rate \( \lambda \) increase together so that \( \beta \) converges to some positive value. It is therefore most relevant when the number of servers \( c \) is large.

Let \( Y(\infty) \) denote a random variable distributed according to the steady-state distribution of \( Y \). Then

\[
P(Y(\infty) > 0) = \left[ 1 + \frac{\beta \Phi(\beta)}{\phi(\beta)} \right]^{-1},
\]

\[
P(Y(\infty) > x|Y(\infty) > 0) = e^{-\beta x}, \quad x > 0, \text{ and}
\]

\[
P(Y(\infty) \leq x|Y(\infty) \leq 0) = \frac{\Phi(x + \beta)}{\Phi(\beta)} \quad x \leq 0,
\]

where \( \phi \) and \( \Phi \) are, respectively, the density and cumulative distribution function of a (standard) normal random variable with mean 0 and variance 1. Hence, the steady-state distribution \( \pi \) is a mixture of an exponential distribution on \(( 0, \infty )\) and a truncated normal distribution on \(( -\infty, 0]\).

We want to solve Poisson’s equation for the process \( c + \sqrt{c}Y \) and function \( I(x \leq b) \), where \( b \approx c - 1 \). For simplicity we take \( b = c \). We first solve Poisson’s equation for the process \( Y \) and the function \( I(x \leq 0) \) to give \( f \) say, and then set

\[
g(x) = f \left( \frac{x - c}{\sqrt{c}} \right).
\]

We have that \( \bar{\alpha} = P(Y(\infty) \leq 0) \), and Poisson’s equation is

\[
\mu f''(x) + \mu(x)f'(x) = -[I(x \leq 0) - \bar{\alpha}] \quad \forall x.
\]

11
We can solve (14) analytically. Along the way we exploit the fact that $f$ is $\pi$-integrable to establish the value of certain constants. The solution is

$$f(x) = \kappa + \begin{cases} \frac{-\bar{\alpha}x}{\mu\beta} & x > 0 \\ \frac{1-\bar{\alpha}}{\mu} \int_x^0 \frac{\Phi(s+\beta)}{\phi(s+\beta)} \, ds & x \leq 0. \end{cases}$$

The constant $\kappa$ is chosen to ensure that $\pi f = 0$, and is given by

$$\kappa = \frac{\bar{\alpha}(1-\bar{\alpha})}{\mu\beta^2} - \frac{\bar{\alpha}(1-\bar{\alpha})}{\mu\Phi(\beta)} \int_{-\infty}^{\beta} \frac{\Phi^2(s)}{\phi(s)} \, ds.$$ 

Next we compute the relative error by first computing $\nu g$, where $\nu$ corresponds to the stationary distribution associated with the parameters of the previous period. We append a suffix of 0 to parameters for the previous period, and again assume for simplicity that only the arrival rate changes. Some algebra reveals that

$$\nu g = \frac{1}{\mu t} \left( \frac{1-\bar{\alpha}}{\beta} - \frac{1-\bar{\alpha}_0}{\beta_0} \right) + \frac{1-\bar{\alpha}}{\mu t \Phi(\beta)} \int_{-\infty}^{\beta} \Phi(s) \left( \frac{\bar{\alpha}_0 \Phi(\beta)}{\bar{\alpha} \Phi(\beta_0)} \Phi(s+\beta_0-\beta) - \Phi(s) \right) ds.$$ 

Plots of this expression, again evaluated numerically, are virtually identical to Figure 2. To obtain some sense of the magnitude of the relative errors when $\rho_0 \approx \rho$, we use two-term Taylor expansions around $\beta$. After some effort, the result is

$$\frac{(1-\bar{\alpha})(\beta_0-\beta)}{\mu t} \left[ \frac{1}{\beta} + \frac{1-\alpha}{\beta^3} + \frac{1}{\Phi(\beta)} \int_{-\infty}^{\beta} \Phi(s) \, ds + \frac{\alpha}{\beta^2 \Phi(\beta)} \int_{-\infty}^{\beta} \Phi^2(s) \, ds \right].$$

Notice that the coefficient of $\beta_0-\beta$ is positive. As $\beta \to 0$ (i.e., we approach heavy traffic), the dominant term in the coefficient is $(1-\alpha)/\beta^3$ which is of the order $\beta^{-2} = [c(1-\rho)^2]^{-1}$, matching earlier results. Furthermore, as $\beta \to \infty$ (i.e., we approach light traffic conditions) the error decreases to 0 extremely rapidly; it is asymptotically of the order

$$\frac{\phi(\beta)}{\mu t} (\beta_0-\beta).$$

Hence, the Halfin-Whitt regime yields further insight, especially with regard to light-traffic behavior.

The analysis thus far has focused on the case where $\tau = 0$. Let us now turn to the case $\tau > 0$, and return to the CTMC model. We can numerically solve Poisson’s equation (8). The rate matrix $A$ remains the same as before. We have that $V(u)$, the waiting time of a customer in the queue that arrives at time $u$, given that $X(u)$, the number in the system at time $u$, is $x$ where $x \geq c$ has an
Erlang distribution. Then

\[ h(x) = \begin{cases} 
1 & x < c \\
1 - \sum_{k=0}^{x-c} \frac{e^{-c\mu}(c\mu)^k}{k!} & x \geq c, 
\end{cases} \]

and

\[ \bar{\alpha} = 1 - \pi_c \left( \frac{c\mu}{c\mu - \lambda} \right) e^{-\tau(c\mu - \lambda)}. \]

The solution \( g \) in this case has the same shape as in the case \( \tau = 0 \) but the scale is reduced, indicating smaller errors in the steady state approximations. Similarly, plots of relative error \( \nu g/(\bar{\alpha} t) \) at various parameter values are much like those for the case in which \( \tau = 0 \) but with reduced scale.

For an example consider Figure 3, which is a plot of relative error for various values of \( \rho \) and \( \rho_0 \). In this figure, \( \tau = 50/9 \), which is \( 1/9 \)th of the average service time. If we were to rescale time so that the average service time is 3 minutes, \( \tau \) would then be 20 seconds. Comparing Figure 3 to Figure 2 in which \( \tau = 0 \), we see that scale is indeed reduced in Figure 3. Due to numerical difficulties we do not plot the relative error for values of \( \rho < 0.59 \).

To gain some intuition for the reduction in scale of the relative error \( \nu g/(\bar{\alpha} t) \), we reason nonrigorously as follows. Suppose that \( \lambda_0 < \lambda \). An arriving customer who never abandons will wait less than \( \tau > 0 \) seconds if there is an available server upon arrival (Case 1), or if the customers ahead of him in the queue either are served or abandon within \( \tau \) seconds (Case 2). Case 1 contributes the same amount to the error approximation whether \( \tau > 0 \) or \( \tau = 0 \). The difference in the error approximation arises from Case 2. Case 2 can only occur when there are \( c \) or more customers in the system upon arrival, and is more likely under the stationary distribution than under the true time-dependent distribution, since \( \lambda_0 < \lambda \). Therefore, we get a negative contribution to the error approximation in Case 2, thereby reducing the observed scale in our plots.
This intuition is supported by Figure 4, where we plot the (scaled) probability mass functions of $\nu$ and $\pi$ as well as the function $h$ for both $\tau = 0$ and $\tau = 50/9$. In this plot, $\rho_0 = 0.7$ and $\rho = 0.94$. (We have plotted these values as continuous curves even though the function values are only defined at integer values.) The $h$ functions for $\tau = 0$ and $\tau > 0$ differ only for $x \geq c$, as noted above. We expect that the distribution of $X(u)$ will change smoothly from $\nu$ to $\pi$ as $u$ increases from 0, so that the probability mass function of $X(u)$ on values $x \geq c$ will lie under that of $\pi$. Hence the error approximation $\nu g$ will be smaller when $\tau > 0$ than when $\tau = 0$. In addition, for any given $\rho_0$ and $\rho$, $\bar{\alpha} t$ will be larger when $\tau > 0$ than when $\tau = 0$. The relative error $\nu g/(\bar{\alpha} t)$ is thus reduced for $\tau > 0$ as compared to $\tau = 0$.

Fig. 4. Plot of $\nu$ and $\pi$ corresponding to $\lambda_0 = 0.7$ and $\lambda = 0.94$, respectively, and $h$ for $\tau = 0$ and $\tau = 50/9$.

4 Models With Abandonment

The results for the models considered up to now suggest, among other things, that for large traffic intensities the error can be significant. However, those models omit an important aspect of many service systems, namely customer abandonment, that one might suspect may at least reduce the heavy-traffic effect. In this section we consider both the $M/M/c + M$ model, and a diffusion approximation for that model that is valid for large numbers of servers.

The $M/M/c + M$ model is identical to the $M/M/c$ model except that customers have limited patience. Customers are willing to wait an exponentially distributed amount of time (independent of all else) and, if that time passes before they reach service, they depart without receiving service.

Abandonment has the effect of stabilizing the queue lengths, since even if customers arrive faster than they can be served, abandonment rates increase and keep waiting times in the queue small. Again our performance measure is the fraction of arriving customers who immediately receive service, so that unsatisfactory services include customers who abandon, or who wait in the queue.
for a positive amount of time before reaching a server.

Again the process giving the number of customers in the system over time is a CTMC, with rate matrix $A$ given as follows. Let $\lambda, \mu$ and $\theta$ denote the arrival rate, the service rate for a single server, and the abandonment rate for a single customer. The nonzero off-diagonal entries of $A$ are

\[
\begin{align*}
A_{i,i+1} &= \lambda & i & \geq 0 \\
A_{i,i-1} &= i\mu & 1 & \leq i & \leq c \\
A_{i,i-1} &= c\mu + (i-c)\theta & i & > c
\end{align*}
\]

The stationary distribution $\pi$ is easily computed numerically using standard birth-death results; see, e.g., Ross [1996, p. 253]. Then for the case $\tau = 0$ we have $\bar{\alpha} = \pi_0 + \cdots + \pi_{c-1}$, the steady-state probability that there are $c - 1$ or fewer customers in the system. Poisson’s equation is again of the form

\[
Ag(x) = -[I(x \leq c - 1) - \bar{\alpha}].
\]

We solve this equation numerically. Again care is required due to numerical issues. A representative plot of $g$ is given in Figure 5. Notice that it decreases at a sublinear rate; this rate is clarified through the calculations for a diffusion approximation below. It is worthwhile comparing this plot to the solution to Poisson’s equation for the $M/M/c$ queue in Figure 1. Notice the large reduction in scale, suggesting that the errors are much reduced when abandonment is taken into account.

![Figure 5. The Solution to Poisson’s Equation for the $M/M/c + M$ Queue with $\lambda = 0.94$, $\mu = 0.02$, $c = 50$ and $\theta = 0.02$.](image)

This suspicion is confirmed when we look at the plot of the numerically-computed relative error in Figure 6. There is a dramatic reduction in relative error relative to the $M/M/c$ case. Again the error is greatest when $\rho$ is large and $\rho_0$ is small, i.e., when the system becomes very busy after being less so.
As with the $M/M/c$ queue we can obtain further insight by considering an appropriate diffusion approximation. Garnett et al. [2002] show that the process $X$ giving the number of customers in the system over time in the $M/M/c + M$ queue can be approximated by

$$X(\cdot) \approx c + \sqrt{c}Y(\cdot; \beta),$$

where $\beta = \sqrt{c}(1 - \rho)$ is unconstrained in sign and $\rho = \lambda/c\mu$. Here $Y(\cdot; \beta)$ is a diffusion on $(-\infty, \infty)$ with drift function

$$\mu(x) = \begin{cases} 
- (\mu \beta + \theta x) & x \geq 0 \\
- \mu (\beta + x) & x < 0
\end{cases}$$

and constant infinitesimal variance $2\mu$. This approximation is justified by taking a limit as the number of servers $c$ and the arrival rate $\lambda$ increase so that $\sqrt{c}(1 - \rho)$ converges to a value that, unlike the $M/M/c$ case, is not restricted to be positive. (This reflects the fact that customers abandon if they wait too long, so a steady-state exists for any set of parameter values.)

Let $Y(\infty)$ denote a random variable distributed according to the steady-state distribution of $Y$. Let $r = (\theta/\mu)^{1/2}$. Then [Garnett et al., 2002]

$$P(Y(\infty) > 0) = \left[ 1 + \frac{H(\beta r^{-1})}{r^{-1}H(-\beta)} \right]^{-1},$$

$$P(Y(\infty) > x|Y(\infty) > 0) = \frac{\Phi(rx + \beta r^{-1})}{\Phi(\beta r^{-1})}, \quad x > 0,$$

$$P(Y(\infty) \leq x|Y(\infty) \leq 0) = \frac{\Phi(x + \beta)}{\Phi(\beta)} \quad x \leq 0,$$

where $H(\alpha)$ is the right-end-point of the distribution function of $Y(\cdot; \beta)$, $\Phi(\alpha)$ is the distribution function of a normal random variable of mean $0$ and variance $2\mu$, and $\Phi(\alpha)$ is the distribution function of a normal random variable of mean $\beta$ and variance $2\mu$. This approximation is justified by taking a limit as the number of servers $c$ and the arrival rate $\lambda$ increase so that $\sqrt{c}(1 - \rho)$ converges to a value that, unlike the $M/M/c$ case, is not restricted to be positive. (This reflects the fact that customers abandon if they wait too long, so a steady-state exists for any set of parameter values.)

Let $Y(\infty)$ denote a random variable distributed according to the steady-state distribution of $Y$. Let $r = (\theta/\mu)^{1/2}$. Then [Garnett et al., 2002]
where \( \Phi(\cdot) = 1 - \Phi(\cdot) \) is the complementary cdf, and \( H(\cdot) = \phi(\cdot)/\Phi(\cdot) \) is the hazard function of a standard normal random variable. Hence, the steady-state distribution \( \pi \) is a mixture of two truncated normal distributions.

We want to solve Poisson’s equation for the process \( c + \sqrt{c}Y \) and function \( I(x \leq b) \), where \( b \approx c - 1 \). For simplicity we take \( b = c \). We first solve Poisson’s equation for the process \( Y \) and the function \( I(x \leq 0) \) to give \( f \) say, and then set \( g(x) = f \left( \frac{x - c}{\sqrt{c}} \right) \).

We have that \( \bar{\alpha} = P(Y(\infty) \leq 0) \), and Poisson’s equation is

\[
\mu f''(x) + \mu f'(x) = -[I(x \leq 0) - \bar{\alpha}] \quad \forall x. \tag{15}
\]

We can solve (15) analytically. Along the way we exploit the fact that \( f \) is \( \pi \)-integrable to establish the value of certain constants. The solution is

\[
f(x) = \kappa + \begin{cases} 
-\frac{\bar{\alpha}}{\mu r} \int_0^x \Phi(s + \beta r^{-1}) \Phi(\beta) \phi(s) ds & x > 0 \\
\frac{1-\bar{\alpha}}{\mu} \int_0^x \Phi(s + \beta r^{-1}) \phi(s) ds & x \leq 0.
\end{cases}
\]

The constant \( \kappa \) is chosen to ensure that \( \pi f = 0 \), and is given by

\[
\kappa = \frac{\bar{\alpha}(1-\bar{\alpha})}{\mu r^2 \Phi(\beta r^{-1})} \int_{\beta r^{-1}}^{\infty} \Phi(s) ds - \frac{\bar{\alpha}(1-\bar{\alpha})}{\mu \Phi(\beta)} \int_{\beta r^{-1}}^{\infty} \frac{\Phi^2(s)}{\phi(s)} ds.
\]

The growth rate of \(|f|\) is now clear: Since \( x\Phi(x)/\phi(x) \to 1 \) as \( x \to +\infty \), \(|f|\) increases at a logarithmic rate in the right tail. Similarly, the left-hand tail increases at a logarithmic rate.

Next we compute the relative error by first computing \( \nu g \), where \( \nu \) corresponds to the stationary distribution associated with the parameters of the previous period. We append a suffix of 0 to parameters for the previous period, and again assume for simplicity that only the arrival rate changes. Algebra reveals that

\[
\frac{\nu g}{\bar{\alpha}t} = \frac{1 - \bar{\alpha}}{\mu t \Phi(\beta)} \int_{-\infty}^{\beta} \frac{\Phi(s)}{\Phi(s)} \left( \frac{\bar{\alpha}_0 \Phi(\beta)}{\alpha \Phi(\beta_0)} \Phi(s + \beta_0 - \beta) - \Phi(s) \right) ds
\]

\[+ \frac{1 - \bar{\alpha}}{\mu t r^2 \Phi(\beta r^{-1})} \int_{\beta r^{-1}}^{\infty} \frac{\Phi(s)}{\phi(s)} \left( \Phi(s) - \frac{(1 - \bar{\alpha}_0) \Phi(\beta r^{-1})}{(1 - \bar{\alpha}) \Phi(\beta_0 r^{-1})} \Phi(s + \beta_0 - \beta) \right) ds.
\]

Consider the case when \( r = 1 \), i.e., \( \theta = \mu \). This leads to large simplifications, most likely since the queue process is then identical to that of an \( M/M/\infty \) queue, which has a Poisson stationary
distribution with mean $\lambda/\mu$. A linear Taylor expansion of $\Phi(s+\beta_0-\beta)$ shows that
\[
\frac{\nu g}{\bar{\alpha}t} \approx \frac{1-\bar{\alpha}}{\mu t} (\beta_0 - \beta) \left[ \frac{1}{\Phi(\beta)} \int_\beta^\infty \Phi(s) \, ds + \frac{1}{\Phi(\beta)} \int_{-\infty}^\beta \Phi(s) \, ds \right].
\] (16)

In light traffic, i.e., as $\beta \to \infty$, the bracketed term in (16) is of the order $\beta$, while $1 - \bar{\alpha} = \Phi(-\beta)$ is of the order $\phi(\beta)/\beta$, and hence the error is of the order $\phi(\beta)(\beta_0 - \beta)/\mu t$, which converges to 0 very rapidly. In heavy traffic, i.e., as $\beta \to -\infty$, the bracketed term is of the order $-\beta$, while $1 - \bar{\alpha}$ converges to 1, and hence the error is of the order $-\beta(\beta_0 - \beta)/\mu t$. The contrast between models with and without abandonment is particularly noticeable at the point $\beta \approx 0$. For models without abandonment the relative error is of the order $\beta^{-2}/\mu t$, yet with abandonment, substituting $\beta = 0$ in (16) gives
\[
\frac{\sqrt{2} \beta_0}{\sqrt{\pi \mu t}}.
\]

We turn now to the case $\tau > 0$, and return to the CTMC model. We can numerically solve Poisson’s equation (8). The rate matrix $A$ remains the same as above. Riordan [1962] computes $h$ and $\bar{\alpha}$ for this case (Koole [2004] computes $h$ for a more general model) as follows:
\[
h(x) = \begin{cases} 
1 & x < c \\
1 - \frac{c\mu(c\mu+\theta)-c\mu+(x-c)\theta}{(x-c)\theta-\lambda c+c+1} \sum_{k=0}^{x-c} (-1)^k \frac{\theta e^{(c\mu+k\theta)}}{c\mu+k\theta} & x \geq c,
\end{cases}
\]
and
\[
\bar{\alpha} = 1 - \pi c \left( \frac{\theta}{\lambda} \right)^{c\mu/\theta} \frac{c\mu e^{c\mu/\theta}}{\theta} \int_0^{(\lambda/\theta)e^{-\theta} - e^{(c\mu/\theta)-1}e^{-\theta}} g \, dy.
\]
Plots of $g$ and the error approximation for $\tau > 0$ have the same shape but are reduced in magnitude compared to those for the case $\tau = 0$, exactly as we saw for the $M/M/c$ model.

5 Discussion and Conclusions

The error approximations reveal a large qualitative difference in the performance of the approximation for models with and without abandonment, and both of these cases are important to consider in the setting of service systems.

When customers are highly reluctant to abandon, the error can be enormous under heavy loads. Furthermore, when the system is entering heavy traffic, the error is an order of magnitude larger (in terms of powers of $1 - \rho$) than when the system is leaving heavy traffic. The error agrees with our intuition, in that the steady-state approximations overestimate performance when we are leaving heavy traffic, and underestimate performance when entering heavy traffic. When the system is moderately or lightly loaded, the errors are very small, and we derived the asymptotic magnitudes of the errors as the load goes to 0.
For service systems where customers are more willing to abandon, the errors are much smaller. This is in line with intuition that says that systems with abandonment reach a steady state far more quickly than systems without abandonment. The errors are still large when the system is heavily loaded, so care still needs to be exercised in using steady-state approximations. When loads are small, the errors are extremely small, with an asymptotic (as $\rho \rightarrow 0$) value that coincides with a system with no abandonment, as one might expect.

This work immediately suggests a number of avenues for research. For example, we have focused on performance measures related to the fraction of calls answered on time, since this is the industry standard. But other performance measures may make more practical sense and, in time, be adopted. It would be instructive to see whether the conclusions reached here apply more broadly to other performance measures, such as the one considered in Koole [2003], or perhaps some variant of the conditional value-at-risk measure that is receiving a great deal of attention in the risk management community. As another example, we have focused on systems with exponential service and patience times. These assumptions are reasonable given our goal of qualitative insights, but it would still be of great interest to extend some of these ideas to service and patience times that more closely match the distributions we see in practice.

6 Proofs

We collect here proofs of some of the technical results used to justify the approximation and our method for its computation. We only prove these results for the CTMC model. The analogous results for diffusions can be proved using similar, although more involved, techniques. So throughout this section we assume the setup described in Section 2. In cases where there is no abandonment, we assume that $\lambda < c\mu$, so that the system is stable.

**Lemma 1** Suppose that $\nu$ is the stationary distribution for the number of customers in a multiserver queue, where the number of servers, arrival, service and abandonment rates (if any) give rise to a stable system. Then $\nu$ has a tail that decreases at least geometrically fast.

**Proof:** We show that $\nu(n + 1) \leq r\nu(n)$ for all $n \geq n^*$, for some $n^*$ and some $r \in (0, 1)$, proving the geometric decay.

Since $\nu$ is the stationary distribution for a multiserver queue with $c$ servers, and arrival, service and abandonment rates $\lambda, \mu$ and $\theta$ say, it follows that the birth rates $\lambda(\cdot)$ and death rates $\mu(\cdot)$ satisfy

\[
\lambda(i) = \lambda \quad \text{and} \\
\mu(i) = (i \wedge c)\mu + [i - c]^+\theta,
\]
where \( a \wedge b = \min(a, b) \) and \( x^+ = \max\{x, 0\} \). (We take \( \theta = 0 \) if there is no abandonment.) Standard results for birth-death processes then ensure that for \( n \geq c \),

\[
\nu(n+1) = \frac{\lambda}{c \mu + (n-c) \theta} \nu(n).
\]

Now, if there is no abandonment, then take \( r = \lambda/c \mu \) and \( n^* = c \). If there is abandonment, then let \( n^* \) be large enough that \( c \mu + (n^* - c) \theta > 2 \lambda_0 \), and let \( r = 1/2 \).

**Lemma 2** The function \( g \) as defined in (4) is well-defined and bounded by a linear function. Hence, in particular, \( E_\nu |g(X(0))| < \infty \) for \( \nu \) defined as in Lemma 1.

**Proof:** We use coupling theory. Let \( \hat{X}(\cdot) \) be a stationary version of \( X(\cdot) \) constructed on the same probability space as \( X(\cdot) \), and we take \( X(0) = x \) with probability 1. Let \( P_{(x,\pi)} \) and \( E_{(x,\pi)} \) denote probability and expectation on the enlarged probability space. Then \( \tilde{\alpha} = E_{\pi} h(X(0)) = E_{(x,\pi)} h(\hat{X}(u)) \) for any \( u \geq 0 \), and so we can write

\[
|g(x)| = \left| \int_0^\infty E_{(x,\pi)} [h(X(u)) - h(\hat{X}(u))] \, du \right|
\leq \int_0^\infty |E_{(x,\pi)} [h(X(u)) - h(\hat{X}(u))]| \, du.
\]

Now, let \( T = \inf\{u \geq 0 : X(u) = \hat{X}(u)\} \) be the coupling time of \( X(\cdot) \) and \( \hat{X}(\cdot) \), and let the joint construction be such that after the two chains meet, they remain together. Since \( h(\cdot) \in [0,1] \), it follows that

\[
|g(x)| \leq \int_0^\infty |E_{(x,\pi)} [h(X(u)) - h(\hat{X}(u)); T > u]| \, du
\leq \int_0^\infty E_{(x,\pi)} |h(X(u)) - h(\hat{X}(u))| I(T > u) \, du
\leq \int_0^\infty P_{(x,\pi)} (T > u) \, du
= E_{(x,\pi)} T.
\]

So it remains to give a coupling where \( E_{(x,\pi)} T \) grows at most linearly in \( x \). Our coupling is one where \((X(u), \hat{X}(u)) : u \geq 0\) is a continuous time Markov chain on the state space \( \{(x, y) : x, y \in \{0, 1, \ldots\}\} \). We choose the transition rates so that if \( X(0) \leq \hat{X}(0) \) then \( X(u) \leq \hat{X}(u) \) for all \( u \geq 0 \), and vice versa. The marginal chains transition up and down in synchrony, except that the chain in the “higher” state has more down transitions.
More precisely, let \( R((x_1, y_1), (x_2, y_2)) \) denote the transition rate from \((x_1, y_1)\) to \((x_2, y_2)\). The non-zero transition rates are

\[
\begin{align*}
R((x, y), (x + 1, y + 1)) &= \lambda \\
R((x, y), (x - 1, y - 1)) &= (x \wedge c)\mu + [x - c]^+\theta & \text{if } x \leq y \\
R((x, y), (x, y - 1)) &= (y \wedge c - x \wedge c)\mu + ([y - c]^+ - [x - c]^+)\theta & \text{if } x \leq y \\
R((x, y), (x - 1, y - 1)) &= (y \wedge c)\mu + [y - c]^+\theta & \text{if } x > y \\
R((x, y), (x - 1, y)) &= (x \wedge c - y \wedge c)\mu + ([x - c]^+ - [y - c]^+)\theta & \text{if } x > y.
\end{align*}
\]

The coupling time \( T \) is bounded by the first time that the coupled chain hits \((0, 0)\). Let \( T_0 \) and \( \tilde{T}_0 \) be the first times that the marginal chains \( X(\cdot) \) and \( \bar{X}(\cdot) \) each hit \( 0 \). Then, because of the ordering property of our coupling,

\[
E_{(x, \pi)} T \leq E_{(x, \pi)} \max\{T_0, \tilde{T}_0\} \\
\leq E_{(x, \pi)} (T_0 + \tilde{T}_0) \\
= E_x T_0 + E_{\pi} T_0.
\]

Now, \( E_x T_0 \leq a + bx \) for some \( a, b \geq 0 \). (To see why, note that there exists a state \( x^* \) such that \( \lambda < \mu(x^*) \leq \mu(x) \) for all \( x \geq x^* \), where \( \mu(\cdot) \) are the death rates of \( X(\cdot) \). Hence, the time for the chain to reach the state \( x^* \) from a state \( x > x^* \) is stochastically bounded by the corresponding time for a random walk with negative drift. The expectation of this time grows linearly in \( x \). Furthermore, the expected time to reach 0 from \( x^* \) is finite, since \( X(\cdot) \) is positive recurrent.) Furthermore, \( E_{\pi} T_0 < \infty \) since

\[
E_{\pi} T_0 = E_{\pi} E[T_0 | X(0)] \leq E_{\pi}[a + bX(0)] = a + b E_{\pi} X(0) < \infty.
\]

(The final step uses the fact that \( \pi \), being a special case of the distributions \( \nu \) in Lemma 1, has a tail that decays geometrically rapidly, and hence has finite expectation.) \( \square \)

**Lemma 3** The function \( g \) defined in (4) satisfies Poisson’s equation (8). Furthermore,

\[
E_{\pi} g(X(0)) = 0.
\]

**Proof:** For \( x \geq 0 \), \( Ag(x) = \lambda g(x + 1) - (\lambda + \mu(x))g(x) + \mu(x)g(x - 1) \), where \( \mu(x) \) is the death rate in state \( x \). Now, \( \lambda = \lim_{u \to 0} P_{x,x+1}(u)/u, \mu(x) = \lim_{u \to 0} P_{x,x-1}(u)/u, \) and \(-(\lambda + \mu(x)) = \lim_{u \to 0}(P_{x,y}(u) - 1)/u, \) where \( P_{x,y}(u) = P_x(X(u) = y) \). Hence

\[
Ag(x) = \lim_{u \to 0} P_{x,x+1}(u)g(x + 1) + P_{x,x-1}(u)g(x - 1) + (P_{x,x}(u) - 1)g(x).
\]

(17)

Let \( o(u) \) denote a term \( \varphi(u) \) say that is such that \( \varphi(u)/u \to 0 \) as \( u \to 0 \). Then,
\[ E_x g(X(u)) = \sum_y P_{x,y}(u)g(y) \]
\[ = P_{x,x+1}(u)g(x + 1) + P_{x,x-1}(u)g(x - 1) + P_{x,x}(u)g(x) + o(u). \]  
(18)

(Showing that the residual terms are indeed \( o(u) \) is straightforward but tedious. One uses the fact that \( g \) is bounded by a linear function. We omit the details.)

Hence, from (17) and (18), we see that

\[ Ag(x) = \lim_{u \to 0} \frac{E_x g(X(u)) - g(x) + o(u)}{u} \]
\[ = \lim_{u \to 0} \frac{\int_{0}^{\infty} E_x h(X(s)) - \bar{\alpha} \, ds - \int_{0}^{\infty} E_x h(X(s)) - \bar{\alpha} \, ds}{u} \]
\[ = \lim_{u \to 0} - \int_{0}^{\infty} E_x h(X(s)) - \bar{\alpha} \, ds. \]  
(19)

Let \( \Rightarrow \) denote convergence in distribution. We have that \( X(s) \Rightarrow X(0) \) as \( s \to 0 \), and since \( X(\cdot) \) has a discrete state space, this implies that \( h(X(s)) \Rightarrow h(X(0)) \). Since \( h(x) \in [0,1] \) it follows that \( E_x h(X(s)) \to E_x h(X(0)) = h(x) \) as \( s \to 0 \). Hence, the mean value theorem applied to (19) gives that \( Ag(x) = -(h(x) - \bar{\alpha}) \) as required.

To complete the proof, note that

\[ E_\pi g(X(0)) = E_\pi \int_{0}^{\infty} h(X(u)) - \bar{\alpha} \, du \]
\[ = \int_{0}^{\infty} E_\pi h(X(u)) - \bar{\alpha} \, du \]
\[ = 0, \]

where the interchange is justified by the proof of Lemma 2, and the last step follows since \( E_\pi h(X(u)) = E_\pi h(X(0)) = \bar{\alpha} \).

**Acknowledgments**

We would like to thank Ger Koole for a suggestion that led to the generalization to \( \tau \geq 0 \) from \( \tau = 0 \), and the referees and editors for constructive criticism that improved the paper. This research was partially supported by NSF Grant Number DMI-0400287.
References


