Queueing Systems 43, 103–128, 2003

A Diffusion Approximation for a Markovian Queue with Reneging

AMY R. WARD amy@isye.gatech.edu
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA

PETER W. GLYNN glynn@stanford.edu
Department of Management Science & Engineering, Stanford University, Stanford, CA 94305, USA

Received 17 November 2001; Revised 1 August 2002

Abstract. Consider a single-server queue with a Poisson arrival process and exponential processing times in which each customer independently reneges after an exponentially distributed amount of time. We establish that this system can be approximated by either a reflected Ornstein–Uhlenbeck process or a reflected affine diffusion when the arrival rate exceeds or is close to the processing rate and the reneging rate is close to 0. We further compare the quality of the steady-state distribution approximations suggested by each diffusion.

Keywords: Markovian queues, reneging, impatience, deadlines, reflected Ornstein–Uhlenbeck process, reflected affine diffusion, diffusion approximation, steady-state

1. Introduction

It has long been recognized that reneging is an important feature in many real-world queueing contexts. In fact, Palm [19] introduced reneging as a means of modeling the behavior of telephone switchboard customers more than 60 years ago. However, due to the explosive growth of the call center industry, there has been renewed interest in such models in recent years. In the call center setting, customer impatience (amplified by large customer loads) leads naturally to large amounts of reneging. Ignoring the presence of reneging can lead to inappropriate sizing of the system and poor staffing allocation.

Models in which reneging is present are also potentially valuable in problem contexts within which customers arrive with deadlines. When the time-in-system exceeds a given customer's deadline, the customer leaves the queue (and thereby reneges). Deadline-sensitive traffic is of practical interest currently in the wireless context, because certain packets of wireless data (for example, location data) lose their value unless transmitted or received within a given time interval.

In this paper, we study approximations for a class of Markovian queueing models in which reneging is present. Specifically, we approximate the corresponding number-in-system birth–death process by a one-dimensional reflected diffusion with state-dependent coefficients. We show that both a reflected Ornstein–Uhlenbeck (O–U) process and a reflected affine diffusion (a diffusion having linear drift and variance) serve as suitable approximations for such a system. Although the steady-state distribution of a reflected affine diffusion has a simple form, the reflected O–U is in general more tractable.

The contributions of this paper are to (1) prove rigorous weak convergence theorems for Markovian queueing models with reneging (2) introduce the notion of a "universal diffusion approximation" (3) establish weak convergence of steady-state distributions (4) provide a simple formula for the steady-state distribution of a reflected affine diffusion and (5) perform a numerical study evaluating the quality of our proposed approximations. Theorem 1 establishes the limiting regimes in which Markovian reneging models may be approximated by diffusion processes. The most important asymptotic regime is that in which the arrival rate and processing rate are roughly in balance and the reneging rate is small. The resulting reflected O–U process incorporates both the queueing effects associated with high server utilization and the abandonment features associated with a reneging model. In contrast to the work of Garnett et al. [11] (see also related work by Fleming et al. [10]), our theory is developed within a setting in which the number of servers is fixed (whereas they focus on diffusion limits obtained as the number of servers goes to infinity). As a consequence, our approximation can more easily be applied to systems with a small number of servers. We appeal to a semigroup approach in proving our diffusion limits, thereby providing an important illustration of how that body of theory applies to queues. (Another possible proof strategy would be to apply continuous-mapping arguments, based on ideas of Mandelbaum and Pats [17].)

The second contribution of this paper is the introduction of the notion of a "universal diffusion approximation" in the context of heavy-traffic limit theorems for queues. Historically, in developing diffusion approximations for systems involving multiple problem parameters (like arrival rate, processing rate, and reneging rate), different limit theorems (involving different temporal and spatial scalings) are offered for each of the various limiting regimes under consideration. This can create difficulties for practitioners, as it will require the practitioner to decide a priori which diffusion limit is appropriate for a given set of problem parameters. The notion of a universal diffusion approximation is intended to provide a single diffusion approximation, to be used universally across all combinations of the problem parameters. This single approximating process is obtained by "pasting together" the different diffusion limits obtained from the various limiting regimes associated with the model. We propose two universal diffusion approximations for our model: a reflected Ornstein–Uhlenbeck process and a reflected affine diffusion. Theorems 2 and 3 make rigorous the idea that our proposed universal diffusion approximations are indeed consistent with the case-by-case limit theorems known for this class of models. This idea is consistent with work seen in the approximations literature in which a single approximation is proposed in order that it be consistent with several different limiting regimes. For example, Mitra and Morrison [18] develop a "uniform asymptotic approximation" for blocking probabilities in a finite capacity model that is appropriate for overloaded, critical, and underloaded regimes. A second example is [4], in which they derive a uniform asymptotic approximation to the partition function for a single-chain closed product form queueing network.

In addition to our functional limit theorems for the reneging model, we establish the weak convergence of steady-state distributions in the asymptotic regime in which the reflected O–U process appears. Typically, showing this convergence is a non-trivial mathematical issue; see, for example [20]. However, in our setting, such a result is straightforward because our reneging model is a birth and death process with a tractable steady-state distribution. In proposition 1, we obtain a "local limit theorem" for the steady-state distribution under a range of spatial scalings. These scalings identify the range over which the steady-state of the reflected O–U process offers accurate approximations to that of the Markovian reneging model.

Motivated by our reflected affine universal diffusion approximation, we also establish the steady-state distribution of a reflected affine diffusion process. The density of this distribution turns out to have a form similar to that of a gamma density.

The final contribution of this paper is a numerical study providing results on the accuracy of the proposed approximations. The study shows that even for seemingly small probabilities of customer abandonments, ignoring the presence of reneging can lead to huge approximation errors. Fortunately, our weak convergence theory suggests a criterion on the problem data under which the impact of reneging can be ignored (see theorem 1), and our numerical study validates this criterion.

In a forthcoming paper [23], we will show that the reflected O–U process arises as a diffusion limit for queues with renewal arrivals, and general processing and reneging time distributions. Consequently, the reflected O–U process plays the same role in the reneging context as does reflected Brownian motion in the setting of conventional queues. It therefore becomes important to study as much of the structure of the reflected O–U process as is possible. We analyze various (steady-state and transient) properties of reflected O–U in a companion paper [24].

The rest of this paper is organized as follows. In section 2, we specify the model and its steady-state behavior. In section 3, we discuss the concept of a universal diffusion approximation in the context of a single-server queue without reneging. In section 4, we establish the appropriate diffusion approximations in different limiting regimes for our reneging model. In section 5, we show how one diffusion process can consolidate the limiting regimes found in section 4. In section 6, we establish weak convergence of steady-state distributions and find the steady-state distribution of our alternative universal diffusion approximation, a reflected affine diffusion. Finally, in section 7, we numerically study the quality of our proposed diffusion approximations in the context of approximating steady-state performance characteristics.

2. Model description

In this paper, we are concerned with developing approximations for a class of queueing models that can be characterized as birth–death continuous-time Markov chains (CTMCs) $Q = (Q(t)\colon\ t \geqslant 0)$ on $\mathbb{Z}^+ = \{0, 1, 2, \ldots\}$, with birth rates $\lambda_n = \lambda$ $(n \geqslant 0)$ and death rates $\mu_n = \mu + (n-1)\gamma$ $(n \geqslant 1)$, where λ, μ, and γ are positive parameters. The process Q describes the number-in-system process for a system that is fed by a Poisson arrival process having rate λ and in which customer processing times form an independent sequence of i.i.d. exponential random variables having mean μ^{-1}. The server processes available work at unit rate. Customers are served in the order in which they arrive; see remark 5. Each customer independently abandons the system when that customer has spent an exponentially distributed amount of time (having mean γ^{-1}) in the system without receiving service. Our approximations can easily be extended to the case in which s unit rate servers process customers in the order in which they arrive. This class of models describes, in simplified form, a call center environment within which customers abandon (or renege from) the queue after an exponentially distributed amount of time. Thus, this paper makes a contribution to the general literature on Markovian queues with reneging.

Our approximations are also valid for another closely related class of Markovian queueing models in which reneging is present. Specifically, consider a birth–death process Q' on $\mathbb{Z}^+$, with birth rates $\lambda_n = \lambda$ $(n \geqslant 0)$ and death rates $\mu_n = \mu + n\gamma$ $(n \geqslant 1)$. Note that in this model, customers can abandon the system even after service has been initiated. In particular, Q' describes the number-in-system process for a system with Poisson arrivals (having rate λ), exponential processing times (with mean μ^{-1}), and a unit rate server. A customer abandons the system after spending an exponentially distributed amount of time having mean γ^{-1}. The abandonment times in this model can be interpreted as customer deadlines. When a given customer's deadline is exceeded, service of that customer becomes worthless and that customer is immediately dropped from the system. For example, this might be appropriate in a wireless context in which either a packet is transmitted within its deadline time or it becomes too dated to be of value. It should be noted that in this class of models, the exponential assumptions we have made imply that the distribution of Q' is unaffected by the service discipline used (e.g., FIFO, LIFO, processor sharing, etc.).

The CTMC queueing models described above are always irreducible and positive recurrent. Focusing on the first class of models, it follows that $Q(t) \Rightarrow Q(\infty)$ as $t \to \infty$. The birth–death structure implies that the steady-state probabilities $\pi_n = P(Q(\infty) = n)$ can be computed fairly explicitly. In particular,

$$\pi_0 = \left(1 + \sum_{n=1}^{\infty} \frac{\lambda^n}{\prod_{j=0}^{n-1}(\mu + j\gamma)}\right)^{-1}$$

and

$$\pi_n = \frac{\lambda^n}{\prod_{j=0}^{n-1}(\mu + j\gamma)}\pi_0$$

for $n \geqslant 1$. The product $\prod_{j=0}^{n-1}(\mu + j\gamma)$ can be expressed in terms of known special functions. Specifically,

$$\prod_{j=0}^{n-1}(\mu + j\gamma) = \frac{\gamma^n \Gamma(\mu/\gamma + n)}{\Gamma(\mu/\gamma)}$$

and

$$\sum_{n=1}^{\infty} \frac{\lambda^n}{\prod_{j=0}^{n-1}(\mu + j\gamma)} = \left(\frac{\lambda}{\gamma}\right)^{1-(\mu/\gamma)} \exp\left(\frac{\lambda}{\gamma}\right) \Gamma_{\lambda/\gamma}\left(\frac{\mu}{\gamma}\right)$$

where $\Gamma(y) = \int_0^\infty t^{y-1}\mathrm{e}^{-t}\,\mathrm{d}t$ is the gamma function and $\Gamma_x(y) = \int_0^x t^{y-1}\mathrm{e}^{-t}\,\mathrm{d}t$ is the incomplete gamma function; see [2,22].

A number of important steady-state performance measures can easily be expressed in terms of the above steady-state distribution. For example, the mean number-in-system is given by $E[Q(\infty)]$, and Little's Law implies that the mean time-in-system is given by $E[Q(\infty)]/\lambda$. Similarly, the steady-state reneging rate is $\gamma E[Q(\infty) - 1]^+$, whereas the steady-state fraction of customers that renege prior to receiving service is $\gamma E[Q(\infty) - 1]^+/\lambda$.

But more complex performance measures can also be computed in terms of the above steady-state distribution. An important such example is the quantity β that describes the average time-in-system spent by a non-reneging customer in the FIFO system described above. Note that

$$\begin{aligned}\beta &= E[\text{time in system spent by a customer} \mid \text{customer receives service}]\\ &= \frac{E[\xi I(\text{customer does not renege})]}{P(\text{customer does not renege})}\end{aligned}$$

where ξ is the time-in-system spent by a (typical) customer and $I(\cdot)$ is the indicator function. Suppose that such a customer arrives to find n customers in the system. Because Poisson arrivals see time averages, this occurs with probability π_n; see [27]. Such a customer needs to wait for each of these n customers to exit the system before s/he receives service. Suppose we temporarily view the system as consisting of the n queued customers plus the arriving customer. Then, the time required for the system population to decrease from j to $j - 1$ $(1 \leqslant j \leqslant n + 1)$ is exponential with rate parameter $\gamma(j - 1) + \mu$. The probability that the arriving customer does not renege during the interval of time required to drive the population from i to $i - 1$ $(2 \leqslant i \leqslant n + 1)$ is $(\gamma(i - 2) + \mu)/(\gamma(i - 1) + \mu)$. Hence,

$$\begin{aligned}&E\big[\xi I(\text{customer does not renege})\big]\\ &\quad = \pi_0\left(\frac{1}{\mu}\right) + \sum_{n=1}^{\infty} \pi_n \sum_{j=1}^{n+1} \frac{1}{(\gamma(j-1)+\mu)} \prod_{i=2}^{n+1} \frac{\gamma(i-2)+\mu}{\gamma(i-1)+\mu} \qquad (2.1)\\ &\quad = \sum_{n=0}^{\infty} \pi_n \frac{\mu}{\gamma n + \mu} \sum_{j=0}^{n} \frac{1}{(\gamma j + \mu)}, \qquad (2.2)\end{aligned}$$

from which β may then be computed.

In principle, the formulae provided above can be numerically evaluated to compute the requisite performance measures. However, because of the factorial-type products involved in the steady-state probabilities, underflow and overflow issues need to be carefully addressed in such a numerical computation. Consequently, we provide, in section 6, an approximation to the steady-state distribution based on our diffusion approximation ideas. This approximation avoids the numerical issues just mentioned, and provides additional qualitative insight that we believe is more transparent than that exhibited in the exact formulae. Our discussion in section 6 will illustrate, for example, how the steady-state performance measure β just described can be suitably approximated via our diffusion approximation.

Computations for transient performance measures for reneging models generally involve numerical procedures to invert transforms. See [26] for expressions in terms of transforms for some transient performance measures in reneging models and [1] for a description of the Fourier-series method for transform inversion. The introduction of diffusion approximations for reneging models with tractable analytic expressions for transient performance measures eliminates the need for transform inversion. We briefly return to the issue of transient performance measure computation at the end of section 5.

3. Universal diffusion approximations

One of our principal goals in this paper is to establish a heavy-traffic diffusion approximation for the class of Markovian reneging models introduced in section 2. To put our results in context, we briefly review here some of what is known about conventional heavy-traffic theory for queues without reneging.

Consider the M/M/1 number-in-system process $Q_\rho = (Q_\rho(t)\colon\ t \geqslant 0)$, where Q_ρ is a birth–death process on $\mathbb{Z}^+$ with constant birth rates $\lambda_n = \mu\rho$ ($n \geqslant 0$) and constant death rates $\mu_n = \mu > 0$ ($n \geqslant 1$). The parameter ρ can, of course, be identified with the *traffic intensity* of the queue or, equivalently, the server utilization (when $\rho \leqslant 1$). The conventional heavy-traffic limit theorem asserts that when $\rho \uparrow 1$,

$$|1-\rho|Q_\rho\bigl(\cdot/(1-\rho)^2\bigr) \Rightarrow X^R(\cdot) \tag{3.1}$$

in the topology of weak convergence on $D[0,\infty)$; see, for example, [3] for a discussion of this convergence concept. Here, $X^R = (X^R(t)\colon\ t \geqslant 0)$ is a reflected Brownian motion (RBM) with drift $-\mu$ and variance parameter 2μ. (See [13] for a rigorous definition of RBM and derivations of many important properties.) A similar limit theorem holds as $\rho \downarrow 1$, in which case X^R is a RBM with drift μ and variance parameter 2μ; see [15] for such results. These two limit theorems describe the time-dependent behavior of an M/M/1 queue in which the traffic intensity ρ is close to one, when viewed on time scales of $(1-\rho)^{-2}$ and spatial scales of $(1-\rho)$.

But there are other diffusion approximations that can also be developed for the M/M/1 queue. For example, if $\rho > 1$, then

$$\varepsilon\bigl(Q_\rho\bigl(\cdot/\varepsilon^2\bigr) - \mu(\rho-1)e\bigl(\cdot/\varepsilon^2\bigr)\bigr) \Rightarrow \sqrt{\mu(1+\rho)}B(\cdot) \tag{3.2}$$

in $D[0,\infty)$ as $\varepsilon \downarrow 0$, where $e(t) = t$ and $B = (B(t)\colon\ t \geqslant 0)$ is a standard Brownian motion with zero drift and unit variance parameters; see [25]. This approximation establishes that if the queue is over-saturated, then the queue eventually leaves the boundary associated with the idle state, and the reflecting barrier on the Brownian motion can therefore be ignored.

A less well-known diffusion approximation for the M/M/1 queue in *heavy traffic* asserts that if $0 < p < 1$, then

$$|1-\rho|^p Q_\rho\big(\cdot/(1-\rho)^{2p}\big) \Rightarrow X^R(\cdot) \tag{3.3}$$

in $D[0,\infty)$ as $\rho \uparrow 1$, where X^R is a RBM with zero drift and variance parameter 2μ; see [12] for a related result. This heavy-traffic result describes the behavior of the number-in-system process for the M/M/1 queue over time scales of smaller order than $(1-\rho)^{-2}$.

Thus, at least three different types of limit processes arise in the context of the M/M/1 queue. The particular approximation to be used to analyze $Q_\rho(t)$ may depend on the relative magnitudes of t and $|1-\rho|$. Each of the limit processes described above is appropriate in a particular subregion of $(t, |1-\rho|)$ space. In principle, given a specific applications environment, the user of such a diffusion approximation needs to assess which of the three limit theorems (3.1)–(3.3) is most appropriate to the particular combination of t and $|1-\rho|$ arising in the application.

However, it turns out that there is no need to make this assessment. In particular, there is a *universal diffusion approximation* that can be used globally across all combinations of t and $|1-\rho|$ that are consistent with each of the three limit theorems above. Specifically, we may choose to approximate Q_ρ as follows:

$$Q_\rho(\cdot) \overset{D}{\approx} X_\rho^R(\cdot) \tag{3.4}$$

where $X_\rho^R = (X_\rho^R(t)\colon t \geqslant 0)$ is a RBM with drift $\mu(\rho-1)$ and variance parameter $(1+\rho)\mu$. Here, $\overset{D}{\approx}$ means "has approximately the same distribution as," and can be rigorously verified to be accurate in each of the limiting regimes introduced in (3.1)–(3.3). For example, (3.4) suggests that

$$\varepsilon\big(Q_\rho(\cdot/\varepsilon^2) - \mu(\rho-1)e(\cdot/\varepsilon^2)\big) \overset{D}{\approx} \varepsilon\big(X_\rho^R(\cdot/\varepsilon^2) - \mu(\rho-1)e(\cdot/\varepsilon^2)\big)\big). \tag{3.5}$$

Both the left-hand side and right-hand side of (3.5) converge when $\rho > 1$ in $D[0,\infty)$, as $\varepsilon \downarrow 0$, to the limit specified in (3.2), Brownian motion with zero drift and variance parameter $\mu(1+\rho)$. Consequently, (3.4) can be rigorously expected to provide a good approximation (in a relative magnitude sense) when $\rho > 1$ and t is large. A similar argument, taking advantage of (3.1) and (3.3), establishes the rigorous validity of the approximation (3.4) whenever ρ is close to one and t is large; see [12] for additional discussion of these mathematical issues.

Of course, a potential user of the universal diffusion approximation (3.4) is free to use the approximation even in parameter regions of $(t, |1-\rho|)$ space in which rigorous validity of the approximation has not been verified or is questionable. The key point, from our perspective, is that the user has available a single diffusion approximation (provided by (3.4)) that is consistent with all known asymptotic regimes.

Our goal, in this paper, will be to develop an appropriate universal diffusion approximation for Markovian queues with reneging. We obtain these approximations by first developing, in section 4, various diffusion approximations that are valid in different subregions of the model's parameter space. Section 5 then "knits together" these approximations in an effort to provide a universal diffusion approximation for Markovian queues with reneging.

4. Diffusion approximations for queues with reneging

In studying the CTMC model Q introduced in section 2, we note that if the reneging rate γ is zero, then the model reduces to an M/M/1 queue with arrival rate λ and service rate μ. As discussed in section 3, it is well known that when $\rho \overset{\Delta}{=} \lambda/\mu$ is close to one, then Q behaves like an RBM over time scales of order $(1-\rho)^{-2}$ and spatial scales of order $(1-\rho)^{-1}$. More precisely, the limit theorem (3.1) holds.

On the other hand, if $\mu = 0$ in the CTMC model Q', we end up with a model that is identical to the infinite-server M/M/∞ queue with arrival rate λ and processing rate (per server) γ. A well-known diffusion approximation (see [14]) is also available for this model. Specifically, if $Q'_{\lambda,\gamma} = (Q'_{\lambda,\gamma}(t)\colon\ t \geqslant 0)$ is an M/M/∞ queue with arrival rate λ and processing rate (per server) γ with $Q'_{\lambda,\gamma}(0) = \lambda/\gamma$, then

$$\lambda^{-1/2}\big(Q'_{\lambda,\gamma}(\cdot) - \lambda/\gamma\big) \Rightarrow Y(\cdot) \tag{4.1}$$

as $\lambda \to \infty$ in $D[0,\infty)$, where Y is an Ornstein–Uhlenbeck (O–U) diffusion process with infinitesimal drift $-\gamma x$ and infinitesimal variance 2, starting from the origin. Noting that $Q'_{\lambda,\gamma}(\cdot/\gamma) \overset{D}{=} Q'_{\lambda/\gamma,1}(\cdot)$ (where $\overset{D}{=}$ denotes equality in distribution), it follows from (4.1) that

$$\gamma^{1/2}\big(Q'_{\lambda,\gamma}(\cdot/\gamma) - \lambda/\gamma\big) \Rightarrow Y(\cdot) \tag{4.2}$$

as $\gamma \downarrow 0$ in $D[0,\infty)$ where Y is an O–U process with infinitesimal drift $-x$ and infinitesimal variance 2λ. Because of the minor difference in the transition structures of Q and Q', it is easily seen that Q obeys the same limit theorem (4.2) as does Q'. Consequently, in a pure reneging model, Q can be approximated by an O–U process on time scales of order $1/\gamma$ and spatial scales of order $\gamma^{-1/2}$, at least when γ is close to zero.

Our goal is to construct a diffusion approximation for Q that reflects the queueing phenomena that arise both because of limited service capacity and the presence of customer reneging. Thus, we wish to develop a diffusion approximation that describes the behavior of Q in an asymptotic regime that is intermediate to (3.1) and (4.2). Given that (3.1) describes fluctuations of order $(1-\rho)^{-1}$ on time scales of order $(1-\rho)^{-2}$ and (4.2) describes fluctuations of order $\gamma^{-1/2}$ on time scales of order $1/\gamma$, this suggests that an intermediate asymptotic regime (the regime of part 1 of theorem 1) should be one in which $c\gamma^{1/2} \approx (1-\rho)$, where γ (or, equivalently, $(1-\rho)$) is close to zero.

This intuition turns out to be correct. To set the stage for our limit theorem, let $X = (X(t)\colon\ t \geqslant 0)$ be the strong solution to the stochastic differential equation (SDE)

$$\mathrm{d}X(t) = \bigl(\alpha - \gamma X(t)\bigr)\,\mathrm{d}t + \sigma\,\mathrm{d}B(t) + \mathrm{d}L(t) \tag{4.3}$$

subject to $X(0) = x \geqslant 0$, where $L = (L(t)\colon\ t \geqslant 0)$ is the minimal nondecreasing process which makes $X(t) \geqslant 0$ for $t \geqslant 0$. The process L increases only when X is zero, so that

$$\int_{[0,\infty)} I\bigl(X(t) > 0\bigr)\,\mathrm{d}L(t) = 0.$$

The existence of a unique strong solution to (4.3) is guaranteed by a careful extension of the results of Lions and Sznitman [16] ([16] treats only bounded domains). We refer to X as a reflected O–U process with infinitesimal drift $\alpha - \gamma x$ and infinitesimal variance σ^2.

Remark 1. When $\gamma = 0$, the reflected O–U process X reduces to RBM. In the setting of RBM, the process L can be described explicitly in terms of the unreflected Brownian motion with drift; see, for example, [13]. No such explicit representation is possible in the setting of a reflected O–U process, because when $\gamma > 0$, the state-dependent drift implies that the concept of an “unreflected version” of X is meaningless. The lack of an explicit representation for L means that many of the methods widely used in analysis of RBM are inappropriate in the reflected O–U setting.

Let $Q_\gamma = (Q_\gamma(t)\colon\ t \geqslant 0)$ be the birth–death process on $\mathbb{Z}^+$ with birth rates $\lambda_n = \mu\rho$ $(n \geqslant 0)$ and death rates $\mu_n = (\mu + (n-1)\gamma)$ $(n \geqslant 1)$. Theorem 1 provides a description of the behavior of Q_γ in a panorama of limiting regimes. In particular, theorem 1 pertains to the situations in which $\rho \approx 1$ or $\rho > 1$. Whenever $\rho \ll 1$, the server spends positive time in the idle state, and we cannot hope for a diffusion approximation to be rigorously valid.

Part 1 of theorem 1 describes the behavior of Q when $(1-\rho) \approx c\gamma^{1/2}$ with γ small. Note that the approximating process X has both a constant term in the drift (as in RBM) and a linear term in its drift (as in an O–U process). Thus, a reflected O–U process is, in some sense, a blend of the processes discussed in (3.1) and (4.2). Since this is the only asymptotic regime in which both the effects of the server and the effects of customer reneging appear in the limiting diffusion process, part 1 of theorem 1 is the most important asymptotic regime and motivates the first “universal diffusion approximation” we develop in section 5.

We see the parts of the parameter space in which our model behaves as a queue without reneging in parts 2 and 3 of theorem 1. In particular, when $\gamma^{1/2} \ll 1 - \rho$, then the limit process is identical to that obtained in the setting of no reneging, namely RBM; see (3.1). In other words, if $\gamma^{1/2} \ll 1 - \rho$, reneging may be effectively ignored. This is a potentially important qualitative insight that we revisit in our numerical study in section 7. Also, when Q_γ is viewed on time scales of smaller order than those discussed in parts 1 and 2, the limit process is again identical to that obtained in the setting of no reneging; see (3.3).

Finally, parts 4 and 5 of theorem 1 focus on the situation in which $\rho > 1$. In part 4, we obtain an analog to (4.2). In part 5, we establish the approximate behavior of Q_γ when $\rho > 1$ and when viewed under shorter time scales, as in part 3.

Throughout theorem 1, $\Rightarrow$ denotes weak convergence in $D[0, \infty)$.

Theorem 1 (Weak convergence of Q_γ).

1. Suppose that $\rho = \rho(\gamma)$ is such that $\gamma^{-1/2}(1-\rho) \to c$ as $\gamma \downarrow 0$ for some finite constant c. In addition, suppose that $\gamma^{1/2} Q_\gamma(0) \Rightarrow X(0)$ as $\gamma \downarrow 0$. Then,
$$\gamma^{1/2} Q_\gamma(\cdot/\gamma) \Rightarrow X(\cdot)$$
as $\gamma \downarrow 0$, where X is a reflected O–U process with initial position $X(0)$, infinitesimal drift $-c\mu - x$, and infinitesimal variance 2μ.

2. Suppose that $\rho = \rho(\gamma)$ is such that $1 - \rho \downarrow 0$ and $\gamma^{1/2}/(1-\rho) \downarrow 0$ as $\gamma \downarrow 0$. In addition, suppose that $(1-\rho)Q_\gamma(0) \Rightarrow X(0)$ as $\gamma \downarrow 0$. Then,
$$(1-\rho)Q_\gamma\bigl(\cdot/(1-\rho)^2\bigr) \Rightarrow X^R(\cdot)$$
as $\gamma \downarrow 0$, where X^R is a RBM with initial position $X(0)$, drift $-\mu$, and variance 2μ.

3. Suppose that $\rho = \rho(\gamma)$ is such that $\gamma^{-1/2}(1-\rho) \to c$ as $\gamma \downarrow 0$ for some finite constant c. In addition, suppose that $\gamma^{p/2} Q_\gamma(0) \Rightarrow X(0)$ as $\gamma \downarrow 0$. Then, if $0 < p < 1$,
$$\gamma^{p/2} Q_\gamma\bigl(\cdot/\gamma^p\bigr) \Rightarrow X^R(\cdot)$$
as $\gamma \downarrow 0$, where X^R is a RBM with zero drift and variance parameter 2μ, starting from $X(0)$.

4. Suppose $\lambda > \mu$ is fixed and that $\gamma^{1/2}(Q_\gamma(0) - (\lambda-\mu)/\gamma) \Rightarrow Y(0)$ as $\gamma \downarrow 0$. Then,
$$\gamma^{1/2}\biggl(Q_\gamma(\cdot/\gamma) - \frac{\lambda-\mu}{\gamma}\biggr) \Rightarrow Y(\cdot)$$
as $\gamma \downarrow 0$, where Y is an O–U process with initial position $Y(0)$, infinitesimal drift $-x$ and infinitesimal variance 2λ.

5. Suppose that $\lambda > \mu$ is fixed and that $\gamma^{p/2}(Q_\gamma(0) - (\lambda-\mu)/\gamma) \Rightarrow \sqrt{2\lambda}B(0)$ as $\gamma \downarrow 0$, for some $p \in (0, 1)$. Then,
$$\gamma^{p/2}\bigl(Q_\gamma\bigl(\cdot/\gamma^p\bigr) - (\lambda-\mu)/\gamma\bigr) \Rightarrow \sqrt{2\lambda}B(\cdot)$$
as $\gamma \downarrow 0$, where $B = (B(t){:}\ t \geqslant 0)$ is a standard Brownian motion starting from $B(0)$.

The proof of theorem 1 can be found in the appendix. In the following, we make a few remarks of interest.

Remark 2. A similar limit to that described in part 2 of theorem 1 holds when $\gamma^{1/2}/(1-\rho) \uparrow 0$ as $\gamma \downarrow 0$. In this setting, where $\rho > 1$ for the queue, the limit process (under the same normalization) is an RBM X^R with initial position $X(0)$, drift μ, and variance parameter 2μ.

Remark 3. It may at first seem unintuitive that the parameter μ does not appear in the infinitesimal variance parameters given in parts 4 and 5 of theorem 1. Therefore, we offer the following heuristic argument showing that this is the case. The variance for the CTMC Q is:

$$var\bigl(Q(t+h) - Q(t) \mid Q(t) = n\bigr) = \bigl(\lambda + \mu + \gamma\bigl(Q(t) - 1\bigr)\bigr)h + \mathrm{o}(h). \tag{4.4}$$

When $\lambda > \mu$, the "mass-balance" point for this system is: $(\lambda - \mu)/\gamma$. (One can see this by setting the birth rate λ equal to the death rate $\mu + n\gamma$.) Substituting $(\lambda - \mu)/\gamma$ for Q in the equation above, we have:

$$var\bigl(Q(t+h) - Q(t) \mid Q(t) = n\bigr) \approx 2\lambda.$$

Remark 4. Theorem 1 is also valid when we replace $Q = (Q(t)\colon\ t \geqslant 0)$ by the CTMC $Q' = (Q'(t)\colon\ t \geqslant 0)$, with precisely the same spatial and temporal normalizations. In other words, the slight difference in the definitions of the death rates for Q and Q' is irrelevant in the asymptotic regimes we consider above.

Remark 5. Theorem 1 remains valid, when suitably modified, in the setting of multi-server queues with reneging. Specifically, suppose that $Q = (Q(t)\colon\ t \geqslant 0)$ is a birth–death CTMC on $\mathbb{Z}^+$ with birth rates $\lambda_n = \lambda = s\rho\mu$ $(n \geqslant 0)$ and death rates equal to either $\mu_n = \min(s, n)\mu + n\gamma$ or $\mu'_n = \min(s, n)\mu + (n - s)^+\gamma$, for $n \geqslant 1$. These death rates describe a Markovian queue with s servers, in which customers either can or cannot renege while in service. Then, theorem 1 holds as stated, with the parameter $s\mu$ replacing μ in all the limit processes and normalizations that arise.

5. A universal diffusion approximation for queues with reneging

As discussed in section 3, our goal here is to "knit together" the diffusion approximations obtained in theorem 1, in an effort to provide one globally applicable approximating diffusion process. Given the CTMC Q, with associated parameters λ, μ, and γ, we propose the following univeral diffusion approximation to Q:

$$Q(\cdot) \overset{D}{\approx} X(\cdot) \tag{5.1}$$

where $X = (X(t)\colon\ t \geqslant 0)$ is a reflected O–U process with $X(0) = Q(0)$, infinitesimal drift $\lambda - \mu - \gamma(x - 1)$ and infinitesimal variance 2λ.

At a practical level, (5.1) can potentially be used to compute an approximation (involving X) to virtually any performance measure involving Q. Of course, in using such approximations, it is important to be able to identify those subregions of (t, ρ, γ) parameter space within which the approximations can be rigorously validated to be accurate. To obtain such rigorous guarantees requires use of limit theorems.

As in section 4, we consider limit theorems that are expressed in terms of the reneging parameter γ. We view $\rho = \rho(\gamma) = \lambda(\gamma)/\mu$ as a function of γ. For $\gamma > 0$, let X_γ be a reflected O–U process with infinitesimal drift $(\rho(\gamma) - 1)\mu - \gamma(x - 1)$ and infinitesimal variance $2\rho(\gamma)\mu$, and let Q_γ be defined as in section 4.

We shall prove that X_γ and Q_γ have distributions that are, in some sense, "close." Our measure of distance involves the Prohorov metric d defined on the space of probability measures on the function space $D[0, \infty)$. The Prohorov metric is the metric that gives rise to the topology of weak convergence on $D[0, \infty)$; see, for example, [9]. In other words, $\xi_n \Rightarrow \xi$ in $D[0, \infty)$ if and only if $d(P(\xi_n \in \cdot), P(\xi \in \cdot)) \to 0$ as $n \to \infty$. In a (slight) abuse of notation, we shall henceforth write $d(\xi_n, \xi)$ in place of $d(P(\xi_n \in \cdot), P(\xi \in \cdot))$.

Let $\overset{D}{=}$ denote equality in distribution, and assume throughout the following theorem (whose proof can be found in the appendix) that $X_\gamma(0) \overset{D}{=} Q_\gamma(0)$.

Theorem 2 (Universal diffusion approximation).

(i) If $\gamma^{-1/2}(1 - \rho) \to c$ and $\gamma^{p/2}Q_\gamma(0) \Rightarrow Q(0)$ as $\gamma \downarrow 0$, then

$$d\big(\gamma^{p/2}Q_\gamma\big(\cdot/\gamma^p\big), \gamma^{p/2}X_\gamma\big(\cdot/\gamma^p\big)\big) \to 0$$

as $\gamma \downarrow 0$ (for $0 < p \leqslant 1$).

(ii) If $\rho \to 1$, $\gamma^{1/2}/(1 - \rho) \to 0$, and $(1 - \rho)Q_\gamma(0) \Rightarrow Q(0)$ as $\gamma \downarrow 0$, then

$$d\big((1 - \rho)Q_\gamma\big(\cdot/(1 - \rho)^2\big), (1 - \rho)X_\gamma\big(\cdot/(1 - \rho)^2\big)\big) \to 0$$

as $\gamma \downarrow 0$.

(iii) If $\rho > 1$ is fixed and $\gamma^{p/2}(Q_\gamma(0) - (\lambda - \mu)/\gamma) \Rightarrow Q(0)$ as $\gamma \downarrow 0$, then

$$d\left(\gamma^{p/2}\left(Q_\gamma\big(\cdot/\gamma^p\big) - \frac{\lambda - \mu}{\gamma}\right), \gamma^{p/2}\left(X_\gamma\big(\cdot/\gamma^p\big) - \frac{(\lambda - \mu)}{\gamma}\right)\right) \to 0$$

as $\gamma \downarrow 0$ (for $0 < p \leqslant 1$).

Note that the temporal and spatial scalings appearing in theorem 2 are, in each case, identical for X_γ and Q_γ. This guarantees that in each of the asymptotic regimes described by theorem 1, the universal diffusion approximation provided by (5.1) provides approximations with high relative accuracy. This provides a mathematical justification for the assertion that (5.1) legitimately "knits together" the limiting diffusions obtained in theorem 1.

Remark 6. An almost identical universal diffusion approximation exists for Q'. In particular, given the CTMC Q', with associated parameters λ, μ, and γ, we propose the universal diffusion approximation:

$$Q'(\cdot) \overset{D}{\approx} X'(\cdot),$$

where X' is a reflected O–U process with $X'(0) = Q'(0)$, infinitesimal drift $\lambda - \mu - \gamma x$ and infinitesimal variance 2λ. Furthermore, theorem 2 holds with Q' and X' appropriately replacing Q and X.

Remark 7. Suppose that we are given a multi-server queue, possessing s servers, in which reneging is present. As discussed in remark 5, this leads to a CTMC Q with (slightly) modified birth and death parameters. We propose approximating Q by X, where X is a reflected O–U process with infinitesimal drift $\lambda - s\mu - \gamma x$ and infinitesimal variance 2λ.

It turns out that there is an alternative universal diffusion approximation that is consistent with the asymptotic regimes described in theorem 2. To motivate this alternative universal approximation, note that we can easily compute the following analogs to the infinitesimal drift and infinitesimal variance for the CTMC Q. Specifically, for $n \geqslant 1$,

$$E\big(Q(t+h) - Q(t) \mid Q(t) = n\big) = \big(\lambda - \mu - \gamma\big(Q(t) - 1\big)\big)h + \mathrm{o}(h)$$

and

$$var\big(Q(t+h) - Q(t) \mid Q(t) = n\big) = \big(\lambda + \mu + \gamma\big(Q(t) - 1\big)\big)h + \mathrm{o}(h)$$

as $h \downarrow 0$. This suggests that we can construct an alternative universal diffusion approximation to Q by using a diffusion process that matches the infinitesimal mean and variance characteristics of Q, and that exhibits reflection at the origin.

Specifically, let $Z = (Z(t)\colon\ t \geqslant 0)$ be the solution to the SDE

$$\mathrm{d}Z(t) = \big(\lambda - \mu - \gamma\big(Z(t) - 1\big)\big)\,\mathrm{d}t + \sqrt{\lambda + \mu + \gamma\big(Z(t) - 1\big)}\,\mathrm{d}B(t) + \mathrm{d}L(t) \tag{5.2}$$

subject to $Z(0) = z$, where $L = (L(t)\colon\ t \geqslant 0)$ is the minimal nondecreasing process which makes $Z(t) \geqslant 0$ for $t \geqslant 0$. The existence of a unique solution to (5.2) is again guaranteed by a careful extension of the results of Lions and Sznitman [16]. The process Z without reflection is referred to as an affine diffusion in the finance literature; see, for example, [7]. For this reason, we call the process Z a reflected affine diffusion.

Our alternative universal diffusion approximation to Q then takes the form

$$Q(\cdot) \overset{D}{\approx} Z(\cdot) \tag{5.3}$$

where, of course, we require that $Z(0) = X(0)$. This universal approximation can be mathematically justified in the same way as theorem 2 supports (5.1). Specifically, let Z_γ be the universal approximation to Q_γ associated with (5.2), where Q_γ is defined as in section 4, and Z_γ has infinitesimal drift $(\rho(\gamma) - 1)\mu - \gamma(x - 1)$ and infinitesimal variance $\mu(\rho(\gamma) + 1) + \gamma(x + 1)$.

Theorem 3. Theorem 2 holds with Z_γ replacing X_γ throughout the theorem statement.

Given that Z more faithfully reproduces the fine structure of the infinitesimal drift and variance of Q, we suspect that one often obtains better approximations to Q by using Z rather than X. In fact, table 2 in section 7 illustrates that Z estimates steady-state tail probabilities slightly better than X. However, the presence of state dependence in the infinitesimal variance of Z makes it substantially harder to compute transient performance measures for Z than for X. As we show in our companion paper [24], transient performance measure computations involving X are tractable. For this reason, we recommend use of the universal approximation based on X in preference to that based on Z.

6. Reflected O–U and reflected affine diffusion processes: steady-state behavior

Perhaps the single most important performance measure for Q is its steady-state distribution. Given the universal diffusion approximation to Q proposed in (5.1), we expect that

$$Q(\infty) \overset{D}{\approx} X(\infty),$$

where $X(\infty)$ is the steady-state of X.

Let $X = (X(t)\colon t \geqslant 0)$ be a reflected O–U process with infinitesimal drift $\alpha - \gamma x$ and infinitesimal variance σ^2, with $\gamma, \sigma^2 > 0$. Then, X has a unique stationary distribution π with density

$$\begin{aligned} p(x) &= P\big[N\big(\alpha/\gamma, \sigma^2/2\gamma\big) \in dx \mid N\big(\alpha/\gamma, \sigma^2/2\gamma\big) \geqslant 0\big] \\ &= \sqrt{\frac{2\gamma}{\sigma^2}}\frac{\phi(\sqrt{2\gamma/\sigma^2}(x - \alpha/\gamma))}{1 - \Phi(-\sqrt{2\alpha^2/\gamma\sigma^2})} \end{aligned}$$

for $x \geqslant 0$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution of a $N(0, 1)$ random variable, respectively. Computation of the distribution π can be found in [5,6,24].

Weak convergence of $\gamma^{1/2}Q_\gamma$ to X, as established in part 1 of theorem 1, does not imply that the steady-state distribution of $\gamma^{1/2}Q_\gamma$ converges weakly to that of X. In the current setting, establishing weak convergence of the steady-state distribution is quite straightforward, because the process Q_γ is a birth–death process with a steady-state that is very tractable. Our next result uses this tractability to rigorously establish the weak convergence of the steady-state distributions. Actually, our proof technique establishes much more. It yields not only a weak convergence statement, but also a "local limit theorem" in which it is shown that the probability mass function of $Q_\gamma(\infty)$ may be approximated by the density of $X(\infty)$, where $Q_\gamma(\infty)$ and $X(\infty)$ are random variables endowed with the steady-state distributions of Q_γ and X, respectively. In addition, our argument shows that the steady-state distribution of X provides asymptotically accurate approximations to the steady-state of Q_γ over spatial scales as large as $\gamma^{-2/3}$, and that $\gamma^{-2/3}(=\gamma^{-1/2}\gamma^{-1/6})$ is the critical spatial scale at which the steady-state approximation given by X breaks down (see part (ii)).

Proposition 1. Let Q_γ and X be defined as in part 1 of theorem 1, and satisfy the conditions stated there. Assume, in addition, that $1-\rho(\gamma)=c\gamma^{1/2}+\mathrm{o}(\gamma^{2/3})$ as $\gamma\downarrow 0$. Suppose $p(\cdot)$ is the stationary density of X.

(i) If $x_\gamma=\mathrm{o}(\gamma^{-1/6})$, then

$$P\big(Q_\gamma(\infty)=\lfloor\gamma^{-1/2}x_\gamma\rfloor\big)\sim\gamma^{1/2}p(x_\gamma)$$

as $\gamma\downarrow 0$. Also,

$$P\big(\gamma^{1/2}Q_\gamma(\infty)>x_\gamma\big)\sim P\big(X(\gamma)>x_\gamma\big)$$

as $\gamma\downarrow 0$.

(ii) If $x_\gamma\sim x\gamma^{-1/6}$, then

$$P\big(Q_\gamma(\infty)=\lfloor\gamma^{-1/2}x_\gamma\rfloor\big)\sim\gamma^{1/2}p(x_\gamma)\exp\big(x_\gamma^3/6\mu^2\big)$$

as $\gamma\downarrow 0$.

Proof. Recall that $P(Q_\gamma(\infty)=n)\triangleq\pi_\gamma(n)$, where $\pi_\gamma(n)=v_\gamma(n)/\sum_{m=0}^\infty v_\gamma(m)$ and

$$v_\gamma(n)=\rho^n\prod_{j=0}^{n-1}\left(1+\frac{j\gamma}{\mu}\right)^{-1}.$$

Observe that

$$\log\left(1+\frac{j\gamma}{\mu}\right)=\frac{j\gamma}{\mu}-\frac{1}{2}\frac{j^2\gamma^2}{\mu^2}+\mathrm{O}\big(j^3\gamma^3\big)$$

uniformly in $j=\mathrm{o}(1/\gamma)$. Consequently,

$$\sum_{j=0}^{n-1}\log\left(1+\frac{j\gamma}{\mu}\right)^{-1}=\frac{-n(n-1)\gamma}{2\mu}+\frac{\gamma^2}{6\mu^2}(n-1)\left(n-\frac{1}{2}\right)n+\mathrm{O}\big(n^4\gamma^3\big),$$

provided that $n=\mathrm{o}(1/\gamma)$. So,

$$\begin{aligned}v_\gamma(n)&=\exp\big(n\log\big(1-(1-\rho)\big)-\sum_{j=0}^{n-1}\log\left(1+\frac{j\gamma}{\mu}\right)\\&=\exp\bigg(-n(1-\rho)+\mathrm{O}\big(n(1-\rho)^2\big)\\&\qquad-\frac{n(n-1)\gamma}{2\mu}+\frac{\gamma^2}{6\mu^2}n\left(n-\frac{1}{2}\right)(n-1)+\mathrm{O}\big(n^4\gamma^3\big)\bigg)\end{aligned}\qquad(6.1)$$

uniformly in $n = \mathrm{o}(1/\gamma)$. We conclude that

$$v_\gamma(n) = \exp\biggl(-cn\gamma^{1/2} - \frac{(n\gamma^{1/2})^2}{2\mu} + \mathrm{o}(1)\biggr) \tag{6.2}$$

uniformly in $n = \mathrm{O}(\gamma^{-1/2})$. Choose k so large that $P(X(\infty) > k) < \varepsilon$, $\exp(-ck - k^2/2\mu) < \varepsilon$, and $\lambda/(\mu + \gamma^{1/2}\lceil k\gamma^{-1/2}\rceil) < 1/2$. Then, (6.2) implies that for $K(c,\mu) = \sqrt{2\pi\mu}\exp(c^2\mu/2)(1 - \Phi(-c\sqrt{\mu}))$

$$\begin{aligned}\sum_{m=0}^{\lfloor k\gamma^{-1/2}\rfloor} v_\gamma(n) &= \gamma^{-1/2}K(c,\mu)\exp\bigl(\mathrm{o}(1)\bigr)\sum_{n=0}^{\lfloor k\gamma^{-1/2}\rfloor} p\bigl(n\gamma^{1/2}\bigr)\gamma^{1/2}\\ &\sim \gamma^{-1/2}K(c,\mu)\int_{-\infty}^{k} p(y)\,\mathrm{d}y\\ &= \gamma^{-1/2}K(c,\mu)\bigl(1 - \mathrm{O}(\varepsilon)\bigr)\end{aligned}$$

as $\gamma \downarrow 0$. Also,

$$\begin{aligned}\sum_{n=\lceil k\gamma^{-1/2}\rceil}^{\infty} v_\gamma(n) &= v_\gamma\bigl(\lceil k\gamma^{-1/2}\rceil\bigr)\sum_{j=0}^{\infty}\prod_{l=0}^{j}\biggl(\frac{\lambda}{\mu + \gamma(\lceil k\gamma^{-1/2}\rceil + l)}\biggr)\\ &\leqslant 2v_\gamma\bigl(\lceil k\gamma^{-1/2}\rceil\bigr) \leqslant 2\varepsilon\bigl(1 + \mathrm{o}(1)\bigr)\end{aligned}$$

as $\gamma \downarrow 0$. Since $\varepsilon > 0$ can be made arbitrarily small, this yields the conclusion that

$$\sum_{n=0}^{\infty} v_\gamma(n) \sim \gamma^{-1/2}K(c,\mu) \tag{6.3}$$

as $\gamma \downarrow 0$. It follows from (6.1) and (6.3) that

$$\pi_\gamma\bigl(\lfloor\gamma^{-1/2}x_\gamma\rfloor\bigr) \sim \gamma^{1/2}p(x_\gamma)$$

as $\gamma \downarrow 0$, proving the first assertion of part (i). The second assertion of part (i) follows easily from (6.2) and (6.3) if $x_\gamma = \mathrm{O}(1)$. If $x_\gamma \to \infty$, put $r(\gamma) = \lceil x_\gamma\gamma^{-1/2}\rceil$, $s(\gamma) = 2r(\gamma)$, and note that (6.1) and (6.3) together imply that

$$\begin{aligned}\sum_{m=r(\gamma)}^{s(\gamma)-1} \pi_\gamma(n) &= \exp\bigl(\mathrm{o}(1)\bigr)\sum_{m=r(\gamma)}^{s(\gamma)-1} p\bigl(n\gamma^{1/2}\bigr)\gamma^{1/2}\\ &\sim \int_{x_\gamma}^{\infty} p(y)\,\mathrm{d}y\end{aligned}$$

as $\gamma \downarrow 0$, whereas

$$\begin{aligned}\sum_{m=s(\gamma)}^{\infty} \pi_\gamma(n) &\leqslant 2\frac{v_\gamma(s(\gamma))\gamma^{1/2}}{K(c,\mu)}\bigl(1 + \mathrm{o}(1)\bigr)\\ &\sim 2p(2x_\gamma)\end{aligned}$$

as $\gamma \downarrow 0$. Since $p(2x_\gamma)/\int_{x_\gamma}^{\infty} p(y)\,\mathrm{d}y \to 0$ as $\gamma \downarrow 0$, this establishes that

$$P\big(Q_\gamma(\infty) > x_\gamma \gamma^{-1/2}\big) = \sum_{n=r(\gamma)}^{\infty} \pi_\gamma(n) \sim \int_{x_\gamma}^{\infty} p(y)\,\mathrm{d}y$$

as $\gamma \downarrow 0$, completing the proof of the second assertion of part (i).

For part (ii), note that (6.1) shows that if $x_\gamma \sim x\gamma^{-1/6}$, then

$$v_\gamma\big(\lfloor x_\gamma \gamma^{-1/2} \rfloor\big) \sim \exp\bigg(-cx_\gamma - \frac{x_\gamma^2}{2\mu} + \frac{x_\gamma^3}{6\mu^2}\bigg)$$

as $\gamma \downarrow 0$. Combining this asymptotic with (6.3) completes the proof. □

As promised in section 2, we can use the steady-state approximations of this section to approximate various complex performance measures associated with the original reneging queue. As an illustration, consider the quantity $\beta = \beta(\gamma)$ of section 2 that describes the average time-in-system spent by a non-reneging customer in the system associated with Q_γ. Recall that (see (2.2))

$$\beta(\gamma) = E\Bigg[\bigg(\frac{\mu}{\gamma Q_\gamma(\infty) + \mu}\bigg) \sum_{j=0}^{Q_\gamma(\infty)} \frac{1}{\gamma j + \mu}\Bigg]\bigg(1 - \frac{\gamma E[Q(\infty) - 1]^+}{\lambda}\bigg)^{-1}.$$

Straightforward analysis then shows that

$$\beta(\gamma) = \gamma^{-1/2}\frac{EX(\infty)}{\mu} - \frac{var\,X(\infty)}{\mu^2} - \frac{1}{2}\frac{EX^2(\infty)}{\mu^2} + \mathrm{o}(1)$$

as $\gamma \downarrow 0$. The leading term is the time-in-system that one typically sees in a queue without reneging. The correction terms, which are negative in sign, are the reduction in time-in-system for a given customer that is contributed by those customers "ahead" of the given customer that choose to renege before receiving service.

We close this section with a discussion of the steady-state behavior of our alternative approximating diffusion, the reflected affine diffusion Z. We start with a nonrigorous argument that yields an appropriate differential equation for Z's stationary density.

Assuming that a stationary distribution π (with density p) exists, Echeverria [8] shows that p ought to satisfy

$$\int_{[0,\infty)} (Af)(x)p(x)\,\mathrm{d}x = 0 \tag{6.4}$$

for all functions f that are twice continuously differentiable on $[0, \infty)$ with compact support and satisfying the boundary condition $f'(0) = 0$, where

$$A = \frac{1}{2}(\beta + \gamma x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + (\alpha - \gamma x)\frac{\mathrm{d}}{\mathrm{d}x}.$$

Suppose that the stationary density p is four times continuously differentiable (with bounded derivatives) and satisfies $xp(x) \to 0$ and $xp'(x) \to 0$ as $x \to \infty$. Integrating by parts twice then yields the relation

$$\begin{aligned}\int_0^\infty \frac{1}{2}(\beta+\gamma x)f''(x)p(x)\,\mathrm{d}x = \int_0^\infty f(x)\left[\gamma p'(x) + \frac{1}{2}(\beta+\gamma x)p''(x)\right]\mathrm{d}x \\ + \frac{\gamma}{2}p(0)f(0) + \frac{1}{2}\beta p'(0)f(0). \qquad (6.5)\end{aligned}$$

On the other hand, integration by parts once gives us the equality

$$\begin{aligned}\int_0^\infty (\alpha-\gamma x)f'(x)p(x)\,\mathrm{d}x = -\int_0^\infty f(x)\big[(\alpha-\gamma x)p'(x) - \gamma p(x)\big]\,\mathrm{d}x \\ - \alpha p(0)f(0). \qquad (6.6)\end{aligned}$$

Substituting (6.5) and (6.6) into (6.4) yields

$$\begin{aligned}\int_0^\infty \left[\frac{1}{2}(\beta+\gamma x)p''(x) + (\gamma-\alpha+\gamma x)p'(x) + \gamma p(x)\right]f(x)\,\mathrm{d}x \\ + \left[\frac{\beta}{2}p'(0) + \left(\frac{\gamma}{2}-\alpha\right)p(0)\right]f(0) = 0. \qquad (6.7)\end{aligned}$$

If we can find a probability density p satisfying the second-order differential equation

$$\frac{1}{2}(\beta+\gamma x)p''(x) + (\gamma-\alpha+\gamma x)p'(x) + \gamma p(x) = 0 \qquad (6.8)$$

subject to the boundary condition

$$\frac{\beta}{2}p'(0) + \left(\frac{\gamma}{2}-\alpha\right)p(0) = 0, \qquad (6.9)$$

then (6.7) is clearly satisfied. Thus, we conclude that we may compute the stationary density of Z by solving (6.8), subject to (6.9).

The key to solving (6.8) is to note that it can be re-written as

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\big[(\beta+\gamma x)p(x)\big] - \frac{\mathrm{d}}{\mathrm{d}x}\big[(\alpha-\gamma x)p(x)\big] = 0. \qquad (6.10)$$

Integration of (6.10) therefore yields a first order linear differential equation (with non-constant coefficients) that may be solved explicitly. The general solution is

$$p(x) = \exp(-2x)(\beta+\gamma x)^\nu\left[2C_1\int_0^x \exp(2u)(\beta+\gamma u)^{-(\nu+1)}\,\mathrm{d}u + C_2\right]$$

where $\nu = (2\alpha+2\beta-\gamma)/\gamma$ and C_1, C_2 are arbitrary constants of integration. In order that p integrates to unity and satisfies (6.9), we conclude that

$$p(x) = \frac{\exp(-2x)(\beta+\gamma x)^\nu}{\int_0^\infty \exp(-2y)(\beta+\gamma y)^\nu\,\mathrm{d}y}. \qquad (6.11)$$

Note that p is non-negative so that (6.11) yields a legitimate density.

Having obtained a candidate stationary distribution through the above argument, it remains only to rigorously verify that this distribution is indeed the stationary distribution for Z. This argument is similar to that of proposition 1 in [24], and is therefore omitted. We summarize this discussion with the following result.

Proposition 2. Suppose that γ and β are positive. Then, the reflected diffusion Z has a unique stationary distribution π with density given by (6.11).

The diffusion Z has strikingly different tail behavior in its stationary distribution as compared to X. Nevertheless, under the limiting regimes described in sections 4 and 5, the two stationary distributions have identical asymptotic behavior. We discuss this issue further in section 7.

7. Numerical study of the quality of the suggested approximations

We conclude this paper with a brief numerical investigation of the accuracy of the two universal diffusion approximations proposed in section 5 for our Markovian reneging model. In particular, we study the relative error associated with approximating the steady-state of the continuous-time Markov chain Q by both the steady-state of the reflected O–U process X, and the steady-state of the reflected affine diffusion process Z. The processes X and Z are related to Q via (5.1) and (5.3), respectively.

In many real-world service industry applications of queueing theory (such as call centers, fast food restaurants, etc.), reneging is present. Of course, if the reneging rate is small, one might be tempted to ignore the presence of reneging and model the system as a normal queue (without reneging). In such settings, the RBM approximation to the queue becomes relevant. Therefore, we also take this opportunity to present the relative error associated with approximating the continuous-time Markov chain Q via the reflected Brownian motion process X^R having infinitesimal drift $\lambda - \mu$ and infinitesimal variance $\lambda + \mu$; see [13] for its steady-state distribution.

Because the magnitude of the parameter γ itself is difficult to interpret, table 1 also provides the "steady-state reneging probability" (i.e. the long-run fraction of arriving customers that eventually renege). It is striking to observe the degree to which tiny amounts of reneging can have a substantial impact on queueing performance characteristics. For example, at a nominal traffic intensity of 0.98 (with $\lambda = 0.98$ and $\mu = 1$) and with only 0.36% of the customers choosing to renege, the RBM relative error (in which reneging is ignored by the diffusion approximation) is already 32 times larger than that associated with the reflected O–U approximation (in which reneging is incorporated). Given that a modeler may well be tempted to ignore the presence of reneging at such small reneging rates, the numbers in table 1 make clear the perils associated with such shortcuts.

The numbers in table 1 also confirm the heuristic suggested in section 4 as to when RBM provides a reasonable approximation for reneging models. The heuristic is that when $\sqrt{\gamma} \ll 1 - \rho$, we may effectively ignore the presence of reneging. Notice that

Table 1
Relative error calculations for $E[Q(\infty)]$.

			Approximate values and percent relative error		
γ	P[renege]	$E[Q(\infty)]$	$E[X(\infty)]$	$E[Z(\infty)]$	$E[X^R(\infty)]$
			$\lambda = 0.98$ $\mu = 1$		
0.0001	0.0036	36.74	37.02 (00.8%)	37.18 (01.2%)	49.50 (0034.7%)
0.0010	0.0184	19.03	19.07 (00.2%)	19.40 (01.9%)	49.50 (0160.1%)
0.0100	0.0655	07.34	07.25 (01.2%)	07.67 (04.5%)	49.50 (0574.4%)
0.1000	0.1814	02.58	02.44 (05.5%)	02.87 (11.2%)	49.50 (1818.6%)
1.0000	0.3625	00.98	00.79 (19.4%)	01.19 (21.4%)	49.50 (4951.0%)
			$\lambda = 0.9$ $\mu = 1$		
0.0001	0.0009	8.84	9.33 (05.5%)	9.34 (05.7%)	9.50 (007.5%)
0.0010	0.0077	7.84	8.22 (04.8%)	8.31 (06.0%)	9.50 (021.1%)
0.0100	0.0452	4.93	5.07 (02.8%)	5.32 (07.9%)	9.50 (092.6%)
0.1000	0.1578	2.18	2.13 (02.3%)	2.50 (14.6%)	9.50 (335.8%)
1.0000	0.3406	0.90	0.74 (17.5%)	1.12 (24.4%)	9.50 (955.6%)
			$\lambda = 0.85$ $\mu = 1$		
0.0001	0.0006	5.62	6.12 (08.8%)	6.12 (08.9%)	6.17 (009.8%)
0.0010	0.0052	5.30	5.74 (08.2%)	5.79 (09.2%)	6.17 (016.4%)
0.0100	0.0363	3.90	4.13 (05.9%)	4.33 (10.9%)	6.17 (058.0%)
0.1000	0.1439	1.95	1.96 (02.2%)	2.29 (17.3%)	6.17 (216.1%)
1.0000	0.3264	0.85	0.72 (15.8%)	1.07 (26.4%)	6.17 (625.5%)
			$\lambda = 0.7$ $\mu = 1$		
0.0001	0.0002	2.33	2.82 (21.4%)	2.83 (21.4%)	2.83 (021.5%)
0.0010	0.0023	2.30	2.78 (21.0%)	2.79 (21.5%)	2.83 (023.1%)
0.0100	0.0197	2.07	2.45 (18.6%)	2.53 (22.5%)	2.83 (036.8%)
0.1000	0.1062	1.37	1.51 (10.6%)	1.75 (27.6%)	2.83 (106.6%)
1.0000	0.2808	0.70	0.64 (09.1%)	0.94 (33.8%)	2.83 (304.3%)

when either $\rho = 0.85$ or $\rho = 0.9$ and $\gamma = 0.0001$, the RBM approximation has a relative error less than 10%. Of course, as theorems 2 and 3 suggest, the error associated with the reflected O–U and reflected affine approximations is small for these parameter combinations as well. Since in practical situations it is often hard to confirm $\sqrt{\gamma} \ll 1 - \rho$, in systems where reneging is present (even if at small rates), the safe approach is to explicitly model the reneging.

With regard to the quality of our two universal diffusion approximations, table 2 suggests that Z outperforms X in some regions of the tail distribution of the steady-state. Given that X and Z differ only in the asymptotic behavior of their corresponding infinitesimal variances, it is perhaps not surprising that tail probabilities approximated via Z perform better than those obtained from X (because its infinitesimal variance reproduces more faithfully that of Q than does X). Nevertheless, our conclusion is that tables 1 and 2 support the use of the approximation based on X over that based on Z. The approximations based on X beat those of Z in virtually all entries of table 1 and X is competitive with Z in approximating tail probabilities, except when the tail probabilities

Table 2
Relative error calculations for tail probabilities.

				Percent relative error	
x	$P(Q > x)$	$P(X > x)$	$P(Z > x)$	X	Z
		$\lambda = 0.98$ $\mu = 1$ $\gamma = 0.0001$			
1	9.53×10^{-1}	9.76×10^{-1}	9.76×10^{-1}	2.40%	2.42%
50	2.64×10^{-1}	2.68×10^{-1}	2.71×10^{-1}	1.31%	2.65%
100	5.71×10^{-2}	5.69×10^{-2}	5.87×10^{-2}	0.40%	2.89%
500	6.22×10^{-11}	3.72×10^{-11}	6.55×10^{-11}	0.40%	5.33%
1000	3.19×10^{-31}	2.06×10^{-32}	3.56×10^{-31}	93.53%	11.60%
		$\lambda = 0.98$ $\mu = 1$ $\gamma = 0.001$			
1	9.25×10^{-1}	9.61×10^{-1}	9.62×10^{-1}	3.98%	4.00%
50	5.04×10^{-2}	5.06×10^{-2}	5.35×10^{-2}	0.43%	6.13%
100	3.22×10^{-4}	2.65×10^{-4}	3.50×10^{-4}	17.81%	8.79%
500	3.68×10^{-53}	1.82×10^{-61}	1.17×10^{-52}	$\geqslant$100%	$\geqslant$100%
1000	1.49×10^{-178}	3.76×10^{-232}	1.89×10^{-174}	$\geqslant$100%	$\geqslant$100%
		$\lambda = 0.98$ $\mu = 1$ $\gamma = 0.01$			
1	8.33×10^{-1}	9.13×10^{-1}	9.12×10^{-1}	9.62%	9.55%
50	1.37×10^{-6}	2.81×10^{-7}	1.94×10^{-6}	79.53%	41.81%
100	2.52×10^{-19}	2.10×10^{-24}	9.09×10^{-19}	$\geqslant$100%	$\geqslant$100%
500	2.98×10^{-256}	1.17×10^{-558}	1.15×10^{-221}	$\geqslant$100%	$\geqslant$100%
1000	2.66×10^{-722}	5.15×10^{-2223}	3.39×10^{-564}	$\geqslant$100%	$\geqslant$100%
		$\lambda = 0.98$ $\mu = 1$ $\gamma = 0.1$			
1	6.08×10^{-1}	7.90×10^{-1}	7.80×10^{-1}	29.81%	28.18%
50	2.87×10^{-14}	9.72×10^{-56}	1.89×10^{-23}	$\geqslant$100%	$\geqslant$100%
100	6.44×10^{-74}	1.88×10^{-220}	6.99×10^{-58}	$\geqslant$100%	$\geqslant$100%
500	2.21×10^{-660}	7.26×10^{-5525}	6.37×10^{-381}	$\geqslant$100%	$\geqslant$100%
1000	2.81×10^{-1601}	7.50×10^{-22126}	4.99×10^{-804}	$\geqslant$100%	$\geqslant$100%

themselves become very small (in which case both approximations tend to do poorly). The fact that X is analytically more tractable than Z makes the case for X even stronger. Of course, both approximations consistently outperform the RBM approximation (that effectively ignores the presence of reneging effects).

Acknowledgements

We thank Tom Kurtz, Ward Whitt, Alan Bain the two anonymous referees, and the associate editor for helpful comments relating to the issues addressed in this paper.

Appendix A. Proofs of theorems 1–3

We will prove our weak convergence theorems by appealing to semigroup convergence methods. This requires working with certain function spaces. To this end, let $B(S)$

be the Banach space of bounded real-valued measurable functions on S, equipped with supremum norm. Furthermore, for $-\infty \leqslant r_1 < r_2 \leqslant +\infty$, let $\overline{C}(r_1, r_2)$ be the space of continuous functions on (r_1, r_2) having finite limits at r_1 and r_2, $C^2(r_1, r_2)$ be the space of twice continuously differentiable functions on (r_1, r_2), and let $\widehat{C}(r_1, r_2)$ be the space of continuous functions vanishing at the infinite boundaries of (r_1, r_2).

Proof of theorem 1. *Parts 1–3.* For part 1, we start by identifying the generator of the limit process X. Set

$$A = -(c\mu + x)\frac{\mathrm{d}}{\mathrm{d}x} + \mu\frac{\mathrm{d}^2}{\mathrm{d}x^2}$$

and let $f \in \overline{C}(0, \infty) \cap C^2(0, \infty)$ with $Af \in \overline{C}(0, \infty)$ and $\lim_{x\downarrow 0} f'(x) = 0$. Put $f(0) = \lim_{x\downarrow 0} f(x)$. Note that for such an f, $f'(0)$ exists and equals zero. To see this, realize $f(h) - f(0) = f(h) - f(\varepsilon_h) + f(\varepsilon_h) - f(0) = f'(\xi)(h - \varepsilon_h) + \mathrm{O}(h^2)$, where $\xi \in (\varepsilon_h, h)$, provided we choose $\varepsilon_h < h^2$ small enough so that $|f(\varepsilon_h) - f(0)| \leqslant h^2$. Furthermore, we can repeat this argument to establish that $f''(0)$ exists (using the fact that $f'(0) = 0, Af \in \overline{C}(0, \infty)$, and $f(h) - f(\varepsilon_h) = f'(0)h + f''(\xi)(h^2/2)$, with $\xi \in (\varepsilon_h, h)$ and $\varepsilon_h < h^3$ chosen so that $|f(\varepsilon_h) - f(0)| \leqslant h^3$). Consequently, Itô's formula and the properties of L establish that for such an f,

$$f\bigl(X(t)\bigr) - f\bigl(X(0)\bigr) = \int_0^t (Af)\bigl(X(s)\bigr)\,\mathrm{d}s + \sqrt{2\mu}\int_0^t f'\bigl(X(s)\bigr)\,\mathrm{d}B(s).$$

Because f and Af are clearly bounded, it follows that

$$E_x f\bigl(X(h)\bigr) - f(x) = (Af)(x)h + \mathrm{o}(h)$$

as $h \downarrow 0$. We now apply remark 8.1.3 and corollary 8.1.2 of Ethier and Kurtz [9] to conclude that $\{(f, Af)\colon\ f \in \widehat{C}(0, \infty) \cap C^2(0, \infty),\ f'(0) = 0,\ Af \in \widehat{C}(0, \infty)\}$ generates a Feller semigroup on $\widehat{C}(0, \infty)$.

We now appeal to theorems 1.6.1 and 4.2.11 of Ethier and Kurtz [9]. To do this, observe that $\widetilde{Q}_\gamma(\cdot) = \gamma^{1/2} Q_\gamma(\cdot/\gamma) = (r_\gamma \circ X_\gamma)(\cdot)$, where $r_\gamma : \mathbb{Z}^+ \to \Re^+$ is defined by $r_\gamma(k) = \gamma^{1/2}k$ and X_γ is a CTMC on $\mathbb{Z}^+$ with birth rates $\lambda_n = (\rho\mu)/\gamma$ ($n \geqslant 0$) and death rates $\mu_n = (\mu + (n-1)\gamma)/\gamma$ ($n \geqslant 1$). We must show that for each pair (f, Af) generating the semigroup, there exists $(f_\gamma\colon\ f_\gamma \in B(\mathbb{Z}^+),\ \gamma > 0)$ such that

$$\sup_{k\geqslant 0}\bigl|f_\gamma(k) - f\bigl(\gamma^{1/2}k\bigr)\bigr| \to 0 \tag{A.1}$$

and

$$\sup_{k\geqslant 0}\bigl|(A_\gamma f_\gamma)(k) - (Af)\bigl(\gamma^{1/2}k\bigr)\bigr| \to 0 \tag{A.2}$$

where A_γ is the generator of X_γ.

Select $\tau = \tau(\gamma) \to \infty$ so that

$$\bigl|(\rho - 1)\gamma^{-1/2} - c\bigr| \sup_{0 \leqslant x \leqslant \tau} \bigl|f'(x)\bigr| \to 0, \tag{A.3}$$

$$\gamma^{1/2}\tau \sup_{0 \leqslant x \leqslant \tau} \bigl|f''(x)\bigr| \to 0, \tag{A.4}$$

$$\gamma^{1/2} \sup_{\substack{0 \leqslant x, y \leqslant \tau \\ |x-y| \leqslant 1}} \bigl|f''(x) - f''(y)\bigr| \to 0. \tag{A.5}$$

Note that f, Af lie in $\widehat{C}(0, \infty)$, so that $f(x)$ and $(Af)(x) \to 0$ as $x \to \infty$. Consequently, $f(x) - f(x-1) = f'(\xi_x) \to 0$ for some $\xi_x \in (x-1, x)$. Also, $f(\xi_x) - f(\xi_x - 1) - f'(\xi_x) = -f''(\tilde{\xi}_x)/2 \to 0$ as $x \to \infty$ for some $\tilde{\xi}_x \in (x-2, x)$. Hence, there exists $\tilde{\xi}_x \in (x-2, x)$ such that $f(\tilde{\xi}_x)$, $|c\mu + \tilde{\xi}_x| f'(\tilde{\xi}_x)$, and $f''(\tilde{\xi}_x)$ all converge to zero as $x \to \infty$. Put $\kappa = \tilde{\xi}_\tau \to \infty$ and let $\tilde{f}_\gamma$ be the twice continuously differentiable function on $[0, \infty)$ such that $\tilde{f}_\gamma(x) = f(x)$ for $x \leqslant \kappa$, and $\tilde{f}_\gamma(\cdot)$ is defined on $[\kappa, \kappa + 1]$ through a smooth spline so that

$$\sup_{\kappa \leqslant x \leqslant \kappa+1} \bigl|\tilde{f}_\gamma^{(k)}(x)\bigr| \to 0 \quad (k = 0, 2), \tag{A.6}$$

$$\sup_{\kappa \leqslant x \leqslant \kappa+1} \bigl|(c\mu + x)\tilde{f}_\gamma'(x)\bigr| \to 0. \tag{A.7}$$

The spline, $s_\gamma(x)$, is defined on $[\kappa, \kappa + 1]$ as follows:

$$\begin{aligned} s_\gamma(x) = f(\kappa) + f'(\kappa)(x - \kappa) + \bigl(-3f(\kappa) - 2f'(\kappa)\bigr)(x - \kappa)^2 \\ + \bigl(2f(\kappa) + f'(\kappa)\bigr)(x - \kappa)^3. \end{aligned}$$

Finally, put $f_\gamma(x) = \tilde{f}_\gamma(\gamma^{1/2}x) + \gamma^{1/2}h(\gamma^{1/2}x)$, where h is twice continuously differentiable, has compact support, and satisfies $h'(0) = f''(0)/2$. Clearly, (A.1) is satisfied for this choice of $(f_\gamma\colon \gamma > 0)$.

As for (A.2), observe that

$$\begin{aligned} &\sup_{k \geqslant 0} \bigl|(A_\gamma f_\gamma)(k) - (Af)\bigl(\gamma^{1/2}k\bigr)\bigr| \\ &\quad \leqslant \bigl|(A_\gamma f_\gamma)(x) - (Af)(0)\bigr| + \sup_{1 \leqslant k \leqslant \lfloor \kappa\gamma^{-1/2} \rfloor} \bigl|(A_\gamma f_\gamma)(k) - (Af)\bigl(\gamma^{1/2}k\bigr)\bigr| \\ &\qquad + \sup_{\kappa \leqslant x \leqslant \kappa+1} \bigl|\bigl(A_\gamma \tilde{f}_\gamma\bigr)(k)\bigr| + \sup_{x \leqslant \kappa} \bigl|(Af)(x)\bigr|. \end{aligned} \tag{A.8}$$

The fourth term on the right-hand side clearly converges to zero (because $x \to \infty$), while the first term can be written as

$$\begin{aligned} &\frac{\rho\mu}{\gamma}\bigl(f\bigl(\gamma^{1/2}\bigr) - f(0) + \gamma^{1/2}\bigl(h(\gamma) - h(0)\bigr) - \mu f''(0) \\ &\quad = \rho\mu\gamma^{-1/2} f'(0) + \frac{\rho\mu}{2} f''(0) + \rho\mu h'(0) + \mathrm{o}(1) - \mu f''(0) \\ &\quad \to 0 \end{aligned}$$

as $\gamma \downarrow 0$, because $f'(0) = 0$ and $h'(0) = f''(0)/2$. For the second and third terms, note that

$$\begin{aligned}(A_\gamma f_\gamma)(k) = {} & \bigl(\mu(\rho - 1)\gamma^{-1/2} - (k-1)\gamma^{1/2}\bigr) f'_\gamma\bigl(\gamma^{1/2}k\bigr) \\ & + \mu\biggl(\frac{\rho+1}{2} + \gamma(k-1)\biggr) f''_\gamma\bigl(\xi_\gamma(k)\bigr) + \mathrm{o}(1)\end{aligned}$$

where $\xi_\gamma(k)$ lies in $\gamma^{1/2}(k-1, k+1)$. Use of (A.3)–(A.7) then establishes that the second and third terms in (A.8) converge to zero, completing the proof of (A.2).

The proofs of parts 2 and 3 of theorem 1 are very similar to that of part 1, and are omitted.

Parts 4 and 5. Since parts 4 and 5 of theorem 1 involve a limit process on $\Re$ (as opposed to the half line), we could appeal to the results of Stone [21] (which can be used when the state space of the limiting diffusion process is the real line) to establish these results. This is the methodology used in [14] and part 4 of theorem 1 is a minor variant of his result. However, in an effort to illustrate semigroup methodology, we outline these parts of the proof.

Here the generator of the limit process Y is:

$$A = -x\frac{\mathrm{d}}{\mathrm{d}x} + \lambda\frac{\mathrm{d}^2}{\mathrm{d}x^2}$$

and $\{(f, Af)\colon\ f \in \widehat{C}(-\infty,\infty) \cap C^2(-\infty,\infty),\ Af \in \widehat{C}(-\infty,\infty)\}$ generates a Feller semigroup on $\widehat{C}(-\infty,\infty)$. We represent $\widetilde{Q}_\gamma(\cdot) = \gamma^{1/2}(Q_\gamma(\cdot/\gamma) - (\lambda-\mu)/\gamma) = (\eta \circ X_\gamma)(k)$, where $\eta_\gamma(k) = \gamma^{1/2}(k - (\lambda-\mu)/\gamma)$ and X_γ is a CTMC on $\mathbb{Z}^+$ with birth rate $\lambda_n = \rho\mu/\gamma$ $(n \geqslant 0)$ and death rates $\mu_n = (\mu + (n-1)\gamma)/\gamma$ $(n \geqslant 1)$.

For (f, Af) as above, we need to prove that there exists $(f_\gamma\colon\ \gamma > 0)$ satisfying

$$\sup_{k\geqslant 0}\left| f_\gamma(k) - f\biggl(\gamma^{1/2}\biggl(k - \frac{\lambda-\mu}{\gamma}\biggr)\biggr)\right| \to 0 \tag{A.9}$$

and

$$\sup_{k\geqslant 0}\left| (A_\gamma f_\gamma)(k) - (Af)\biggl(\gamma^{1/2}\biggl(k - \frac{\lambda-\mu}{\gamma}\biggr)\biggr)\right| \to 0 \tag{A.10}$$

as $\gamma \downarrow 0$, where A_γ is the generator of X_γ. To construct f_γ, we select $\tilde{f}_\gamma$ (as in the proof of theorem 1) so that $\tilde{f}_\gamma$ agrees with f on $(-\kappa, \kappa)$ (with $\kappa \to \infty$ sufficiently slowly) and vanishes outside $(-\kappa-1, \kappa+1)$, and then put $f_\gamma(k) = \tilde{f}_\gamma(\gamma^{1/2}(k - (\lambda-\mu)/\gamma))$.

Because of the compact support of $\tilde{f}_\gamma$, (A.10) reduces to showing that

$$\sup_{|j|\leqslant(\kappa+1)\gamma^{-1/2}} \bigl|(A_\gamma f_\gamma)(j+l) - (Af)\bigl(\gamma^{1/2}(j+\Delta)\bigr)\bigr| \to 0$$

as $\gamma \downarrow 0$, where $l = \lfloor(\lambda-\mu)/\gamma\rfloor$ and $\Delta = \lfloor l - (\lambda-\mu)/\gamma\rfloor$. But

$$(A_\gamma f_\gamma)(j+l) = -(\Delta + j)\gamma^{1/2}\tilde{f}'_\gamma\bigl(\gamma^{1/2}(j+\Delta)\bigr) + \bigl(\lambda + (j+\Delta)\gamma\bigr)\tilde{f}''_\gamma\bigl(\xi_\gamma(j)\bigr)$$

and

$$(Af)\big(\gamma^{1/2}(j+\Delta)\big) = -(\Delta+j)\gamma^{1/2}f'\big(\gamma^{1/2}(j+\Delta)\big) + \lambda f''\big(\gamma^{1/2}(j+\Delta)\big)$$

where $\xi_\gamma(j)$ lies within $\gamma^{1/2}$ of $\gamma^{1/2}(j+\Delta)$. A straightforward argument then yields (A.9) and (A.10).

The proof of part 5 follows a similar pattern to that of part 4; its proof is therefore omitted. □

Proof of theorem 2. We prove part (i) for $p = 1$; the other cases follow an identical style of argument.

The triangle inequality for metrics implies that it is sufficient to prove that $\gamma^{1/2}X_\gamma(\cdot/\gamma) \Rightarrow X(\cdot)$ in $D[0,\infty)$ as $\gamma \downarrow 0$. Note that $\gamma^{1/2}X_\gamma(\cdot/\gamma)$ can be represented as $(\eta_\gamma \circ \widehat{X}_\gamma)(\cdot)$, where $\widehat{X}_\gamma(\cdot) = X_\gamma(\cdot/\gamma)$ and $\eta_\gamma(x) = \gamma^{1/2}x$. The process $\widehat{X}_\gamma$ has generator $\widehat{A}_\gamma = A/\gamma$, where A is defined as in the proof of theorem 1. Furthermore, the identical collection $\{(f, Af)\}$ (as in theorem 1) generates a Feller semigroup on $\widehat{C}(0,\infty)$ for X.

We must show that for each pair (f, Af), there exists $(f_\gamma, \widehat{A}_\gamma f_\gamma)$ in the collection such that

$$\sup_{x\geqslant 0}\big|f_\gamma(x) - f(\gamma^{1/2}x)\big| \to 0$$

and

$$\sup_{x\geqslant 0}\big|\big(\widehat{A}_\gamma f_\gamma\big)(x) - (Af)\big(\gamma^{1/2}x\big)\big| \to 0$$

as $\gamma \downarrow 0$. As in the proof of theorem 1, we select $\tilde{f}_\gamma$ so that $\tilde{f}_\gamma$ is twice continuously differentiable, agrees with f on $[0,\kappa]$ (with $\kappa \to \infty$ sufficiently slowly), and vanishes on $[\kappa+1,\infty)$. Set $f_\gamma(x) = \tilde{f}_\gamma(\gamma^{1/2}x)$. The key calculation is that

$$\big(\tilde{A}_\gamma f_\gamma\big)(x) = \big((\rho-1)\mu - \gamma(x-1)\big)\gamma^{-1/2}f'\big(\gamma^{1/2}x\big) + 2\rho\mu f''\big(\gamma^{1/2}x\big)$$

and

$$(Af)\big(\gamma^{1/2}x\big) = -\big(c\mu + \gamma^{1/2}x\big)f'\big(\gamma^{1/2}x\big) + \mu f''\big(\gamma^{1/2}x\big)$$

for $x \leqslant \kappa\gamma^{-1/2}$. The rest of the details can easily be filled in. □

The proof of theorem 3 follows a similar pattern to theorem 2; its proof is therefore omitted.

References

[1] J. Abate and W. Whitt, Numerical inversion of Laplace transforms of probability distributions, ORSA J. Computing 7 (1995) 36–43.

[2] C.J. Ancker and A. Gafarian, Queueing with impatient customers who leave at random, J. Industr. Engrg. 13 (1962) 84–90.

[3] P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1999).

[4] A. Birman and Y. Kogan, Asymptotic evaluation of closed queueing networks with many stations, Stochastic Models 8 (1992) 543–563.

[5] S. Browne and W. Whitt, Piecewise-linear diffusion processes, in: *Advances in Queueing: Theory, Methods, and Open Problems*, ed. J. Dshalalow (CRC Press, Boca Raton, FL, 1995) pp. 463–480.

[6] E. Coffman, A. Puhalskii, M. Reiman and P. Wright, Processor-shared buffers with reneging, Performance Evaluation 19 (1994) 25–46.

[7] D. Duffie, J. Pan and K. Singleton, Transform analysis and asset pricing for affine jump-diffusions, Econometrica 6 (2000) 1343–1376.

[8] P. Echeverria, A criterion for invariant measures of Markov processes, Z. Wahrsch. Verw. Gebiete 61 (1982) 1–16.

[9] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence* (Wiley, New York, 1986).

[10] P. Fleming, A. Stolyar and B. Simon, Heavy traffic limit for a mobile phone system loss model, in: *Proc. of 2nd Internat. Conf. on Telecommunication Systems Mod. and Analysis*, Nashville, TN, 1994.

[11] O. Garnett, A. Mandelbaum and M. Reiman, Designing a call center with impatient customers, Preprint (2001).

[12] P. Glynn, Strong approximations in queueing theory, in: *Asymptotic Methods in Probability and Statistics*, ed. B. Szyszkowitcz (Elsevier, Amsterdam, 1998) pp. 133–150.

[13] J.M. Harrison, *Brownian Motion and Stochastic Flow Systems* (Wiley, New York, 1985).

[14] D. Iglehart, Limit diffusion approximations for the many-server queue and the repairman problem, J. Appl. Probab. 2 (1965) 429–441.

[15] D. Iglehart and W. Whitt, Multiple channel queues in heavy traffic I, Adv. in Appl. Probab. 2 (1970) 150–177.

[16] P. Lions and A. Sznitman, Stochastic differential equations with reflecting boundary conditions, Comm. Pure Appl. Math. 37 (1984) 511–537.

[17] A. Mandelbaum and G. Pats, State-dependent queues: Approximations and applications, in: *Stochastic Networks* (Springer, Berlin, 1995) pp. 239–282.

[18] D. Mitra and J.A. Morrison, Erlang capacity and uniform approximations for shared unbuffered resources, IEEE/ACM Trans. Networking 1 (1993) 664–667.

[19] C. Palm, Etude des délais d'attente, Ericsson Technics 5 (1937) 37–56.

[20] R. Srikant and W. Whitt, Simulation run lengths to estimate blocking probabilities, ACM Trans. Modeling Comput. Simulation 6 (1996) 7–52.

[21] C.J. Stone, Limit theorems for birth and death processes and diffusion processes, Ph.D. thesis, Department of Statistics, Stanford University (1961).

[22] R. Syski, *Introduction to Congestion Theory in Telephone Systems* (Oliver and Boyd, Edinborough, 1960).

[23] A. Ward and P. Glynn, A diffusion approximation for a GI/G/1 queue with reneging, Working paper (2002).

[24] A. Ward and P. Glynn, Properties of the reflected Ornstein–Uhlenbeck process, Working paper (2002).

[25] W. Whitt, Heavy traffic limit theorems for queues: A survey, in: *Lecture Notes in Economics and Mathematical Systems*, Vol. 98 (Springer, Berlin, 1974) pp. 307–350.

[26] W. Whitt, Improving service by informing customers about anticipated delays, Managm. Sci. 45(2) (1999) 192–207.

[27] R.W. Wolff, *Stochastic Modeling and the Theory of Queues* (Prentice-Hall, Englewood Cliffs, NJ, 1989).