From Queueing Theory and its Applications
Liber Amicorum for J. W. Cohen
Edited by O. J. Boxma and R. Syski
North-Holland, Amsterdam, 1988, pp. 149-191

Approximations for the M/M/1 Busy-Period Distribution

Joseph Abate
AT&T Bell Laboratories
184 Liberty Corner Road, Warren, NJ 07060, USA

Ward Whitt
AT&T Bell Laboratories
600 Mountain Avenue, Murray Hill, NJ 07974, USA

This paper develops and evaluates relatively simple closed-form approximations for the M/M/1 busy-period cdf (cumulative distribution function). The first-order effect of the traffic intensity is identified and isolated by scaling. The principal approximations considered for the scaled busy-period cdf are: hyperexponential (H_2) approximations (mixtures of two exponentials) obtained by matching moments and derivatives of the cdf at the origin, asymptotic expansions as $t \to \infty$ and related approximations, approximants based on Widder's formula for Laplace transform inversion, and inverse Gaussian distributions. For small times, the hyperexponential approximation obtained by matching three derivatives at the origin performs best among these candidates. For larger times, an asymptotic normal approximation and an inverse Gaussian distribution perform remarkably well. Inverse Gaussian distributions arise naturally via diffusion approximations, but significantly better inverse Gaussian approximations can be obtained by making additional refinements, e.g., by matching moments or by matching the asymptotic exponential rate as $t \to \infty$. Overall, the results provide a better understanding of the M/M/1 model and a basis for developing approximations for busy-period cdf's in more general models and other quantities of interest in the M/M/1 model.

1. Introduction and summary

In the preface to the first edition of his fundamental book, *The Single Server Queue,* Cohen [15] describes the focus:

> 'The present book concentrates on the most basic model of queueing theory, i.e., the single server model. Its aim is twofold. Firstly a description of those mathematical techniques which have been proved to be the most fruitful for the investigation of queueing models, and secondly, an extensive analysis of the single server queue and its most important variants.'

Thus, the single server queue is of great interest for its own sake, as a model representing many queueing phenomena, as well as a means to illustrate useful techniques that can be applied to other stochastic models.

For these same reasons, we have also been studying the single-server queue

[1-6]. Our goal is to obtain simple approximations and a better understanding of the basic queueing models in order to facilitate practical engineering applications. In particular, our research has been directed toward developing simple approximations describing the transient behavior of the standard GI/G/1 model. We are trying to contribute theoretical insight to empirical investigations such as have been made by ODONI and ROTH [39]. Of course, we want to obtain good numbers, but more than numbers we want relatively simple formulas. The simple formulas help communicate understanding and help do further analysis, e.g., optimization and description of more complex models.

We also intend to apply our methods and results to analyze queues with time-dependent parameters, as in CLARK [14] and DUDA [20,21], but so far we have concentrated on the standard GI/G/1 model; we have been studying the transient behavior of a stationary model with general initial conditions. In fact, much of our work so far, including this paper, focuses on the most elementary M/M/1 special case. We hope to gain from such a narrow focus, just as Cohen does with *The Single Server Queue* (although some may think we are carrying a good thing to an extreme).

1.1. Simple approximations and asymptotic analysis

Developing simple approximations and a practical understanding is an important activity for the mathematician as well as the engineer. The possible benefits of mathematical insight applied in this direction are well illustrated by the approximations for the M/G/s queue developed by BOXMA, COHEN and HUFFELS [12].

In our analysis of transient behavior, a major role is played by asymptotic analysis. In particular, we exploit diffusion approximations resulting from heavy-traffic limit theorems [29]. The heavy-traffic limit theorems provide a scaling of time that is very useful. With appropriate scaling of space and time, the M/M/1 queue-length process approaches a nondegenerate limit as the traffic intensity ρ approaches 1. The limit at 1 is of course regulated or reflecting Brownian motion (RBM). With this scaling, we can thus regard RBM as the special M/M/1 model with $\rho=1$. We can thus obtain insight about RBM and M/M/1 from each other, as was done by COHEN and HOOGHIEMSTRA [16].

We are also interested in limit theorems showing how the time-dependent distribution approaches steady state, i.e., the theory of relaxation times as analyzed extensively by COHEN [15]. In agreement with NEWELL [38], we find that the first-order effect of the relaxation time is captured by the time scaling associated with the diffusion approximation. In agreement with ODONI and ROTH [39], we also find that the approximations following naturally from the relaxation-time limit theorems are not especially accurate. Of course, they do describe the important first-order effects, but they can be significantly improved. It appears that the region where the limit theorems related to the relaxation time are relatively accurate (within 10%) is usually considerably beyond the region of primary practical interest; see Table 3 of [1], Table 5 of [3], and Table 10[1] here. Fortunately (for applying steady-state results), the

1. All tables of this paper have been collected in Appendix A.

M/M/1 processes tend to be closer to their steady-state limits than the relaxation times predict at times of primary interest. (This summary judgement requires some qualification: It is correct for the quantities we have considered, such as the mean queue length starting empty or starting with relatively few customers, but it is not correct for the mean queue length starting with a great number of customers; then the approach to steady state is essentially linear.)

We not only investigate the quality of approximations of the transient behavior via relaxation-time limit theorems, but we also develop new approximations that are easy to apply and understand, and that are reasonably accurate. For example, mixtures of two exponentials (hyperexponential or H_2 distributions), which have three parameters, have proven to be very useful in our previous work [1-5]. Many M/M/1 quantities of interest turn out to be completely monotone (a general mixture of exponentials), which provides theoretical support for the H_2 approximations. Moreover, in regions of interest often one exponential dominates, so that we obtain a simple exponential approximation there. Then we have a very simple formula describing the time-dependent behavior of the queue.

1.2. The M/M/1 busy-period distribution

We continue this line of investigation here by investigating approximations for the M/M/1 busy-period distribution. It is well-known that the busy period plays a fundamental role in both the equilibrium distribution (e.g., via regenerative structure) and the transient behavior; see [4], [15] and [28]. In this paper, we consider several different approximations for the M/M/1 busy-period distribution and compare their accuracy. It turns out that the H_2 approximation obtained by matching three moments does not work as well for the busy-period distribution as it does for other M/M/1 quantities, so that we end up proposing different approximations. A specific approximation is proposed in Section 7, but additional insight is gained from seeing how all the candidates fare. The entire analysis illustrates, as others have discovered before, that the busy-period distribution is rather strange and difficult.

As indicated above, our motivation is not only to describe the M/M/1 model itself, but also to develop results and techniques applicable to more general models. In a sequel to this paper we apply the results here to develop approximations for the GI/G/1 busy-period distribution. We briefly discuss approximations for the GI/G/1 busy-period distribution in Section 8, but all numerical comparisons here are restricted to the M/M/1 model.

1.3. What is the busy-period distribution like?

Before discussing any formal results, it seems worthwhile to give a brief intuitive discussion of the behavior of the GI/G/1 busy-period distribution, because it has somewhat peculiar properties that are easy to understand after some thought. ***The main observation is that the behavior for small time values is***

very different from the behavior for large time values. The entire paper elaborates on this theme.

For small time values, the GI/G/1 busy-period distribution is quite sensitive to the specific interarrival-time and service-time distributions. For example, the probability that the GI/G/1 busy period coincides with one service time is $P(v<u)$ where v is a service time and u is an interarrival time. For very small time values, the busy-period cdf (cumulative distribution function) $B(t)$ is usually approximated very well by

$$B(t) \approx P(v \leq t, v<u), \tag{1.1}$$

which obviously depends strongly on the specific distributions for u and v. Moreover, for small time values the busy-period cdf does not depend critically on ρ. There is no dramatic change as ρ approaches 1 from below or even as ρ becomes greater than 1. Even if $\rho>1$, the busy period may be fairly short.

On the other hand, for large time values the busy-period distribution is primarily determined by a large number of convolutions of the interarrival-time and service-time distributions, so that central-limit-theorem behavior begins to play a central role. Thus, for large time values the busy-period distribution primarily depends on the interarrival-time and service-time distributions only through their first two moments, and approximations by RBM and other related quantities (normal and inverse Gaussian distributions) should be good. Moreover, for large time values the busy-period cdf obviously depends critically on ρ. As ρ approaches 1, the busy period can have a very fat tail. If $\rho>1$, then the busy period may not terminate at all.

The point we wish to emphasize is that the small-time and large-time behavior are quite different. Thus it should come as no surprise that we obtain useful results, both theorems and approximations, by treating these different regions separately. For example, we can describe the heavy-traffic behavior as $\rho \to 1$ in each of the separate time regions quite nicely, as we show below in Sections 2.6 and 2.7.

1.4. Literature review

There obviously is an enormous body of related literature, including the fundamental papers by KENDALL [33], BAILEY [8,9] and KARLIN and MCGREGOR [32]. For the most part, we refer to COHEN [15] for a map of the known world. With regard to previous work on *approximations* of busy-period distributions in the single-server queue, we cite RICE [40], RIORDAN [42, pp. 106-109], KOSTEN [34, pp. 42-45], HEYMAN [26] and DUDA [20,21]. We will discuss these previous approximations as they relate to our work.

1.5. Organization of this paper

We review supporting theory for the approximations in Section 2 and discuss the main approximations in Sections 3-7. We discuss hyperexponential approximations in Section 3, approximations related to asymptotic expansions in Section 4, approximations based on Widder's formula for Laplace transform inversion in Section 5, and inverse Gaussian approximations in Section 6. We

present a final composite approximation in Section 7. We briefly discuss approximations for the GI/G/1 busy-period distribution in Section 8, and related approximations for the M/M/1 expected cumulative idle time up to time t in Section 9. Finally, we present conclusions in Section 10.

2. Background

In this section we review supporting theory for our approximations. We draw on our previous M/M/1 papers [3-6], but we also make connections to the extensive literature.

2.1. The M/M/1 model with time scaling

Let $Q(t)$ represent the queue length (including the customer in service, if any) at time t in the M/M/1 model. Without loss of generality, let the service rate be 1, so that the arrival rate coincides with the traffic intensity ρ. Assume that $\rho<1$, so that the system is stable with $Q(t)$ converging in distribution to $Q(\infty)$ as $t\to\infty$, where $P(Q(\infty)=k)=(1-\rho)\rho^k$, $k\geq 0$.

As in [3-6], we further scale time by $(2\theta^2)^{-1}=2/(1-\rho)^2$; i.e., we consider $Q(2t/(1-\rho)^2)$ and let $P_{ij}(t)$ be the time-scaled transition function

$$P_{ij}(t) = P(Q(2t/(1-\rho)^2)=j \mid Q(0)=i). \tag{2.1}$$

The time scaling in (2.1) is a very important part of the story. Our goal is to describe the queue length process $Q(t)$, its transition function $P_{ij}(t)$ and the associated busy-period distribution. It is significant that the first-order effect of the single parameter ρ is captured by this time scaling. As discussed in Section 2.2 of [3], the time-scaling captures the heavy-traffic behavior as $\rho\to 1$. In particular, the family of processes $\{2^{-1}(1-\rho)Q(2t/(1-\rho)^2): t\geq 0\}$ indexed by ρ converges to canonical regulated or reflecting Brownian motion (RBM), having drift coefficient -1 and diffusion coefficient 1, as $\rho\to 1$. Thus, with the time scaling, we can treat RBM as the nondegenerate M/M/1 queue with $\rho=1$. With the time scaling the M/M/1 processes for different ρ tend to fluctuate in about the same time scale; we remove the dominant ρ effect, so that we can more easily compare M/M/1 processes with different ρ. A simple practical consequence is that we can more meaningfully compare M/M/1 quantities for all possible traffic intensities ρ, $0\leq\rho\leq 1$, in the same table; see Table 1. In contrast, without the scaling, the interpretation of a given time t, depends more on ρ, so that one table including several different ρ is less meaningful; e.g., see Heyman [26]. (In fact, for the M/M/1 busy-period distribution, additional scaling is appropriate, as we will indicate in Section 2.7 below.)

For more general models, such as GI/G/1 or GI/G/m, it is useful to scale time so that *canonical* RBM also appears as the limit as $\rho\to 1$. As indicated in Section 2 of [1], it is easy to convert RBM with negative drift μ and variance coefficient ν to canonical RBM, so that it is easy to obtain the appropriate scaling; see Section 8. With such a scaling, we identify and isolate the first-order effect of the variability (as it differs from M/M/1) as well as the traffic intensity.

2.2. The busy-period distribution and related quantities

Here we are primarily interested in the busy-period distribution. Let $B(t)$ be the time-scaled busy-period cdf (cumulative distribution function); let $B^c(t) \equiv 1 - B(t)$ be the complementary busy-period cdf; and let $b(t)$ be the density. Let $\hat{B}^c(s)$ and $\hat{b}(s)$ be the associated Laplace transforms. From KENDALL [33] or BAILEY [8,9], we know that

$$\begin{aligned} \hat{b}(s) &\equiv \int_0^\infty e^{-st} b(t)dt = z_1(s) = [1-\theta+\theta^2 s - \theta\Psi(s)]/\rho, \\ \hat{B}^c(s) &= 2\theta/[1+\theta s+\Psi(s)], \end{aligned} \tag{2.2}$$

where

$$\begin{aligned} &\theta = (1-\rho)/2, && \Psi(s) = [1+2(1-\theta)s+(\theta s)^2]^{1/2}, \\ &r_1(s) = \Psi+(1-\theta s), && r_2(s) = \Psi-(1-\theta s), \\ &\rho z_1 = 1-\theta r_1, && \rho z_2 = 1+\theta r_2, \\ &r_1 r_2 = 2s, \quad \rho z_1 z_2 = 1 \text{ and } && \rho(1-z_1)(z_2-1) = 2\theta^2 s. \end{aligned} \tag{2.3}$$

The functions $z_1 \equiv z_1(s)$ and $z_2 \equiv z_2(s)$ are the two roots of the basic quadratic equation $\rho z^2 - (1+\rho+2\theta^2 s)z + 1 = 0$. The Laplace transform $\hat{b}(s)$ is derived by KENDALL [33] by making connections to branching processes and by BAILEY [8,9] by considering a modified M/M/1 queue length process that is absorbed when it reaches the origin. As indicated in the proof of Theorem 3.1 of [4], we can also obtain $\hat{b}(s)$ directly from $\hat{P}_{10}(s)$ and $\hat{P}_{00}(s)$ using first principles, where $\hat{P}_{ij}(s)$ is the Laplace transform of $P_{ij}(t)$ in (2.1); in particular,

$$\hat{b}(s) = \hat{P}_{10}(s)/\hat{P}_{00}(s). \tag{2.4}$$

Indeed, the time-domain version of (2.4) is the starting point for Cohen's treatment in [15]; see (2.30) on p. 187.

Additional properties of the M/M/1 busy-period distribution appear in [3-5]. For example, Corollary 4.2.3 in [4] establishes an interesting connection between the probability of emptiness $P_{00}(t)$ and $B(t)$, namely,

$$P_{00}(t) = 1 - \rho B(t), \quad t \geq 0, \tag{2.5}$$

which quickly yields expressions for the expected cumulative idle time in $[0, t]$, say $EI_0(t)$; the expected workload in the system at time t, $EW_0(t)$; and the expected queue length $EQ_0(t)$; all given that $Q(0) = 0$ (this condition being indicated by the subscript), namely,

$$\begin{aligned} EI_0(t) &= \int_0^t P_{00}(s)ds = \int_0^t [1-\rho B(u)]du, \\ EQ_0(t) &= EW_0(t) = \rho t - t + EI_0(t); \end{aligned} \tag{2.6}$$

see Corollary 4.2.6, the alternate proof of Corollary 5.2.1, and Theorem 8.1 of [4]. (In fact, (2.5) is an immediate consequence of the relation $\rho z_1 = 1 - \theta r_1$ in (2.3), because $\hat{P}_{00}(s) = \theta r_1/s$ and $\hat{b}(s) = z_1$ by (2.6) of [4] and (2.2) above.) As

indicated in Remark 4.1 after Corollary 4.2.3 in [4], (2.5) evidently has a long history, but it does not seem to have been given enough emphasis.

Relations (2.5) and (2.6) provide additional motivation for this paper because they show that other M/M/1 quantities of interest can be expressed in terms of the busy-period distribution. (For more on this, see [4-6] and Section 2.3 below.) Our approximations for $B(t)$ automatically yield associated approximations for these other M/M/1 quantities. We discuss some of these related approximations in Section 9.

In this regard, it is significant that the busy-period distribution coincides with the equilibrium waiting-time distribution in an M/M/1 queue with the last-come first-served (LCFS) discipline, as was first noted by RIORDAN [41]. This LCFS waiting-time distribution is useful for its own sake, but also because it provides an upper bound on the waiting-time distribution for a large class of service disciplines, as noted by VAULOT [47]; see p. 416 of HEYMAN and SOBEL [27]. Not only do the results here apply to this M/M/1-LCFS waiting-time distribution, but previous descriptions of the M/M/1-LCFS waiting-time distributions in VAULOT [47], RIORDAN [41] and pp. 106-109 of [42], and pp. 42-45 of KOSTEN [34] also apply here. Indeed, this previous work seems most closely related to our interest in relatively simple closed-form approximations. Vaulot, Riordan and Kosten discovered exact representations that are convenient for generating numbers as well as relatively simple closed-form approximations.

2.3. The stationary-excess relations

An important discovery in [3-5] is a connection between various M/M/1 quantities and the busy-period cdf $B(t)$ via the stationary-excess operator [51]. For any cdf $G(t)$ on the positive real line with mean m_1, the associated stationary-excess (or equilibrium-residual-life) cdf $G_e(t)$ is defined by

$$G_e(t) = m_1^{-1} \int_0^t [1-G(u)]du, \quad t \geq 0. \tag{2.7}$$

In [3] we focused on moment cdf's which are the (time-scaled) moments $E[Q(t)^k]$ under the condition that $Q(0)=0$, divided by their steady-state limits. Corollary 3.1.3 of [3] and Corollary 5.2.1 of [4] show that the first-moment cdf, denoted by $H_1(t)$, coincides with $B_e(t)$, the stationary-excess cdf associated with the busy-period cdf $B(t)$. In fact, this is just a restatement of the second relation in (2.6).

Theorem 1 of [5] shows that the correlation function of the *stationary* (time scaled) queue length process, denoted by $c_q(t)$, in turn coincides with $H_{1e}^{c}(t)$, the complementary stationary-excess cdf associated with the first-moment cdf $H_1(t)$. Thus, $c_q(t)$ is obtained from the busy-period $B(t)$ by applying the stationary-excess operator in (2.7) *twice.*

These stationary-excess relations have important implications for both theory and approximations. For example, the spectral representation and the asymptotic behavior for $B(t)$ thus easily extend to $H_1(t)$ and $c_q(t)$; see

Sections 2.8 and 4. Since the stationary-excess operator tends to be a smoothing operator [51], the functions $B^c(t)$, $H_1^c(t)$ and $c_q(t)$ tend to be successively better behaved. In particular, the stationary-excess relations help explain why it is harder to obtain good approximations for $B^c(t)$ than it is for $B_e^c(t)=H_1^c(t)$ or $c_q(t)$. DELBROUCK [18] and ERLANDER [22] previously considered $B_e(t)$ and noted that it has nice properties.

2.4. Numerical inversion of the Laplace transform

As first shown by KENDALL [33], the Laplace transform $\hat{b}(s)$ in (2.2) can be inverted (using pair 556.1 from CAMPBELL and FOSTER [13]) to obtain

$$b(t) = \frac{1}{t\sqrt{\rho}} e^{-t/\tau}[e^{-\nu} I_1(\nu)], \quad t \geq 0, \tag{2.8}$$

where $I_1(\nu)$ is a modified Bessel function of the first kind, e.g., p. 377 of ABRAMOWITZ and STEGUN [7],

$$\tau = \frac{(1+\sqrt{\rho})^2}{2} \text{ and } \nu = t\theta^{-2}\sqrt{\rho}. \tag{2.9}$$

The parameter τ in (2.9) is the *time-scaled relaxation time* which describes the asymptotic behavior of (2.8) as $t \to \infty$. (Recall that $f(\nu) \sim g(\nu)$ means that $f(\nu)/g(\nu) \to 1$ as $\nu \to \infty$. Use 9.7.1 on p. 377 of [7] to see that $e^{-\nu} I_1(\nu) \sim (2\pi\nu)^{-1/2}$ as $\nu \to \infty$, so that $e^{-\nu} I_1(\nu) \sim \sqrt{t}$ as $t \to \infty$; i.e., $b(t) \sim Kt^{-3/2} e^{-t/\tau}$, so that $e^{-t/\tau}$ is the dominant term in (2.8) as $t \to \infty$; see Section 4. The unscaled relaxation time is $(1-\sqrt{\rho})^{-2}$, as given on p. 180 of COHEN [15]; τ is the time-scaled relaxation time, because $(1-\rho)^2 = (1-\sqrt{\rho})^2(1+\sqrt{\rho})^2$. Note that the dominant portion of the unscaled relaxation time for high ρ is in the time scaling.)

Asymptotic results related to the relaxation time can be used to generate approximations, but as in [1,3], we show that the natural approximations generated from these limits do not perform well for times of primarily practical interest. However, in Section 4 we show that appropriate refinements of the asymptotic results do perform well for times of practical interest.

We focus on the complementary cdf $B^c(t)$ instead of the density $b(t)$. Unfortunately, $\hat{B}^c(s)$ in (2.2) is not as easily inverted as $\hat{b}(s)$, so to obtain numerical values of $B^c(t)$ we numerically invert the Laplace transform $\hat{B}^c(s)$ in (2.2). Indeed, numerical transform inversion is not an unreasonable way to obtain numerical values for $b(t)$ in (2.8).

Given that so many queueing results are expressed in terms of Laplace transforms, it is surprising that relatively little attention has been given to numerical Laplace transform inversion in the applied probability literature. For example, queueing textbooks do not provide practical guidelines and procedures for numerically inverting Laplace transforms. However, as we indicated in Section 4.4 of [1], there are numerous techniques for the numerical inversion of Laplace transforms. As in [1], we use the Gaver-Stehfest procedure [23,43]. It yields good (but not exceptional) accuracy on a small computer (a personal computer with BASIC) with little programming effort. The

Gaver-Stehfest procedure was also employed and compared to other procedures by NANCE ET AL. [37]. For still other procedures plus comparisons, see DAVIES and MARTIN [17].

Our goal in numerical transform inversion is only to obtain limited accuracy. We are primarily interested in relatively simple analytical approximations that are convenient for practical engineering applications. If we actually wanted great numerical accuracy, we would use a different inversion technique. In fact, the best way to get numerical results is not to use transforms at all, but to exploit integral representations, as discussed in [6] and Section 2.8 below.

Numerical values of the complementary time-scaled busy-period cdf $B^c(t)$ obtained by applying the Gaver-Stehfest procedure are given in Table 1. (These values were checked by also applying an independent numerical inversion technique of JAGERMAN [30,31].) Table 1 shows that we have approximately the right time scale for times when $0.001 \leqslant B^c(t) \leqslant 0.2$, which we call the second regime [1]. (The second regime is the time interval where the process reaches steady state for practical purposes.) For example, consider $t=3.00$ where $B^c(t)=0.00378$ with $\rho=0.50$. *With the time scaling,* as ρ changes, the value of $B^c(3)$ changes quite slowly, until ρ gets very high. (This unusual behavior as $\rho \to 1$ is very different from our experience in [3] and [5]; we will discuss this anomaly further in Section 2.7.)

On the other hand, for very small times (which we call the first regime), $B^c(t)$ changes more slowly *when we omit the time scaling.* In Table 1 we remove the time scaling by considering the times $\theta^2/2$ and θ^2. The fact that the time scaling helps for larger times, but not for smaller times, reflects the special nature of the busy-period distribution discussed in Section 1.3.

2.5. Moments

While the M/M/1 busy-period distribution is somewhat inaccessible, its moments are readily available. Theorem 3.2 of [4] gives a basic recursion due to RIORDAN [41,42]: Let m_k be the k-th moment of the time-scaled busy-period distribution; then

$$m_{k+2} = (2k+1)(1-\theta)m_{k+1} - (k^2-1)\theta^2 m_k \tag{2.10}$$

for $m_0=1$ and $m_1=\theta$. (This result appears in a description of the M/M/1-LCFS waiting-time distribution.) Moreover, an explicit formula for the moments is given on p. 232 of TAKÁCS [44], namely,

$$m_{k+2} = \sum_{i=0}^{k} \frac{(k+i+2)!\,k!\,\theta^{k+1-i}(1-2\theta)^i}{(i+1)!\,2^{i+1}\,i!\,(k-i)!}. \tag{2.11}$$

The first five moments are

$$\begin{aligned} &m_1 = m_2 = \theta, \quad m_3 = 3\theta(1-\theta), \quad m_4 = 15\theta[1-2\theta+\tfrac{4}{5}\theta^2], \\ &m_5 = 105\theta[1-3\theta+\tfrac{18}{7}\theta^2-\tfrac{4}{7}\theta^3]. \end{aligned} \tag{2.12}$$

From these moments we can obtain important insight into the busy-period distribution, especially its large-time behavior, and how it behaves as a function of $\theta \equiv (1-\rho)/2$ or ρ. First, the squared coefficient of variation (variance divided by the square of the mean) is

$$c_B^2 \equiv \frac{m_2 - m_1^2}{m_1^2} = \frac{\theta - \theta^2}{\theta^2} = \theta^{-1} - 1 = \frac{1+\rho}{1-\rho}, \tag{2.13}$$

which is independent of the time scale. Formula (2.13) shows that the distribution becomes highly variable as $\rho \to 1$; $c_B^2 \to \infty$ as $\rho \to 1$.

Moreover, the skewness and kurtosis also explode as $\rho \to 1$; i.e.,

$$\begin{aligned} \frac{m_3}{m_2^{3/2}} &= \frac{3\theta(1-\theta)}{\theta^{3/2}} = \frac{3(1-\theta)}{\sqrt{\theta}} = \frac{3\sqrt{2}(1+\rho)}{\sqrt{1-\rho}} \\ \frac{m_4}{m_2^2} &= \frac{15\theta[1-2\theta+\frac{4}{5}\theta^2]}{\theta^2} = \frac{15[1-2\theta+\frac{4}{5}\theta^2]}{\theta} = \frac{30\rho}{1-\rho} + 6(1-\rho). \end{aligned} \tag{2.14}$$

2.6. A power-series representation

In Theorem 9 of [5] we obtained a power-series representation for the complementary busy-period cdf in terms of the moments described in Section 2.5. In particular,

$$B^c(t) = 1 + \sum_{k=0}^{\infty} (-1)^{k+1} \frac{m_{k+2} t^{k+1}}{(k+2)!(k+1)!\theta^{2k+3}}. \tag{2.15}$$

We can use a few terms from (2.15) to get a good approximation for $B^c(t)$ for small t. However, the times t must be *relatively small before time scaling.* With $t = \alpha\theta^2$ (to undo the time scaling), the first three terms from (2.15) yield

$$\begin{aligned} B^c(\alpha\theta^2) &\approx 1 - \frac{m_2\alpha}{2\theta} + \frac{m_3\alpha^2}{12\theta} - \frac{m_4\alpha^3}{144\theta} \\ &\approx 1 - \frac{\alpha}{2} + \frac{(1+\rho)\alpha^2}{8} - \frac{(1+3\rho+\rho^2)\alpha^3}{48}. \end{aligned} \tag{2.16}$$

For example, for $\alpha = 1/2$, (2.16) becomes

$$B^c(\theta^2/2) \approx 0.779 + 0.023\rho - 0.0026\rho^2. \tag{2.17}$$

From Table 1, we see that (2.17) is very accurate, having a maximum relative percent error of about 0.1%. Similarly, when $\alpha = 1$, we obtain

$$B^c(\theta^2) \approx 0.604 + 0.063\rho - 0.0208\rho^2, \tag{2.18}$$

which yields a maximum relative percent error of about 1%.

In summary, (2.15) provides an alternative way to obtain numerical results and relatively simple approximations such as (2.16). However, the simple approximation (2.16) is only valid for relatively small times (small in the original time scaling).

2.7. Light and heavy traffic

Further insight into the busy-period distribution can be obtained by considering the limits as $\rho \to 0$ and $\rho \to 1$. As in Corollary 5.2.2(b) of [4], we can work with the Laplace transform to obtain the light-traffic limit. In particular, by the argument in [4],

$$\lim_{\rho \to 0} \hat{b}(s) = \frac{2}{2+s}; \tag{2.19}$$

so that

$$\lim_{\rho \to 0} B(t) = 1 - e^{-2t}, \quad t \geq 0. \tag{2.20}$$

In other words, the busy-period distribution converges (weakly) to a simple exponential with mean 1/2 as $\rho \to 0$. Indeed, this is obvious because the busy period obviously approaches a single service time as $\rho \to 0$. (The 2 appears because of the time scaling.) Moreover, this limit could be anticipated (and deduced) from the limit for the first-moment cdf $H_1(t)$ established in Corollary 5.2.2(b) of [4] plus the stationary-excess relation discussed in Section 2.2. (Empirically, note that the $\rho = 0$ column of Table 1 here coincides with the $\rho = 0$ column of Table 1 in [3].)

The heavy-traffic limiting behavior is more complicated. Unlike $H_1(t)$ in Corollary 5.2.2(a) of [4], $B^c(t)$ does *not* converge to a proper limit as $\rho \to 1$. However, Theorem 3.5 of [4] describes the limit as $\rho \to 1$. It is significant that, even with the time scaling, it involves an additional normalization by $\theta \equiv (1-\rho)/2$ as $\rho \to 1$. The limits for the normalized complementary cdf $B^c(t)$ and density $b(t)$ are

$$\begin{aligned} &\lim_{\rho \to 1} 2(1-\rho)^{-1} B^c(t) = \tilde{h}_1(t) = 2t^{-1/2}\phi(t^{1/2}) - 2[1 - \Phi(t^{1/2})], \quad t \geq 0, \\ &\lim_{\rho \to 1} 2(1-\rho)^{-1} b(t) = (2\pi t^3)^{-1/2} e^{-t/2}, \quad t \geq 0, \end{aligned} \tag{2.21}$$

where $\Phi(t)$ is the cdf of a standard normal cdf (with mean 0 and variance 1) and $\phi(t)$ is its density. The limits in (2.21) are the density $\tilde{h}_1(t)$ of the first-moment cdf $\tilde{H}_1(t)$ for RBM (M/M/1 with $\rho = 1$) as given in (4.4) of [1] and its derivative $\tilde{h}_1'(t)$. (We use the notation $\sim$ to designate the RBM case with $\rho = 1$.) This is not too surprising, once we recognize that $\theta^{-1} B^c(t)$ on the left side of (2.21) is just the density of the stationary-excess cdf $B_e(t) \equiv H_1(t)$ for given ρ, as indicated in Section 2.3 above. Thus, (2.21) represents convergence of densities (i.e., $h_1(t) \to \tilde{h}_1(t)$ as $\rho \to 1$, a local limit theorem) and their derivatives, paralleling the convergence of distributions as $\rho \to 1$ established in Corollary 5.2.2(a) of [4].

By Theorem 1.3 of [1], the limit $\tilde{h}_1(t)$ in (2.21) also coincides with an exponential mixture of inverse Gaussian densities, i.e.,

$$\tilde{h}_1(t) = \int_0^\infty 2e^{-2x} f(t;x,0)\,dx, \tag{2.22}$$

where $f(t;x,0)$ is the density of an inverse Gaussian cdf $F(t;x,0)$, with

$$f(t;x,0) = \frac{x}{\sqrt{2\pi t^3}}\exp\left[-\frac{(x-t)^2}{2t}\right], \quad t \geq 0,$$
$$F(t;x,0) = \Phi\left[\frac{t-x}{\sqrt{t}}\right] + e^{2x}\Phi\left[\frac{-t-x}{\sqrt{t}}\right], \quad t \geq 0, \tag{2.23}$$

as in (1.5) and (1.6) of [1]; $f(t;x,0)$ is the density of the first-passage time from x to 0 for (canonical) RBM. Inverse Gaussian distributions arise naturally in diffusion approximations for first-passage-time distributions in queueing; e.g., see HEYMAN [26] and DUDA [20,21]. We will be considering the inverse Gaussian distribution further in Section 6.

The extra normalization by θ in (2.21) means that

$$B^c(t) \approx \frac{(1-\rho)}{2}\tilde{h}_1(t), \tag{2.24}$$

which converges to 0 as $\rho \to 1$, as can be seen by looking at the $\rho = 0.95$ and $\rho = 0.99$ columns of Table 1. A better scaling for the busy-period distribution is thus $\theta^{-1}B^c(t)$. (We are including the time scaling.) Note that $\theta^{-1}B^c(t)$ has proper limits both as $\rho \to 0$ and as $\rho \to 1$, just as $H_1(t)$ and $c_q(t)$ do after the time scaling alone. Part I of Table 2 displays the numerical values of $\theta^{-1}B^c(t)$ obtained by multiplying the values in Table 1 by θ^{-1}. Note that this doubly-scaled function is indeed even less sensitive to ρ. For example, for $t = 1$, $\theta^{-1}B^c(t)$ decreases from 0.27 to 0.16 as ρ increases from 0 to 1. The remaining adjustment is not uniform over t though. In Table 2, $\theta^{-1}B^c(t)$ is increasing and then decreasing for $t = 0.10$ and 0.25, strictly decreasing for $0.50 \leq t \leq 1.50$, and then strictly increasing for $t \geq 2.00$.

To a large extent, the behavior of $B^c(t)$ as a function of ρ is captured by the two scalings. What is left is shown in Part I of Table 2; the function $\theta^{-1}B^c(t)$ moves from an exponential at $\rho = 0$ to $\tilde{h}_1(t)$ at $\rho = 1$. A simple light-and-heavy-traffic-interpolation approximation based on this is the convex combination

$$\theta^{-1}B^c(t) \approx (1-w)2e^{-2t} + w\tilde{h}_1(t), \quad t \geq 0, \tag{2.25}$$

where $w \equiv w(\rho)$ is an increasing function of ρ such that $0 \leq w \leq 1$. For approximation (2.25) to perform reasonably well, it helps for $\theta^{-1}B^c(t)$ to be monotone. As we observed above, in Table 2 this is not true for $t \leq 0.25$, but it is true for $t \geq 0.5$. From Table 2, it is easy to see that (2.25) should perform quite well for $t \geq 0.25$. Moreover, from (2.12) here plus Corollary 1.3.4 of [1], we see that (2.25) with $w(\rho) = \rho$ yields the correct first three moments of $B(t)$. (It does not yield the correct fourth moment.) On the other hand, for very small t, (2.25) can not be good; e.g., $B^c(0) = 1$ and $\tilde{h}_1(0) = \infty$. Nevertheless, (2.25) can be a pretty good approximation for $t \geq 0.25$. Experiments with a few weighting functions suggested the weight function $w(\rho) = \rho^x$ for $x = 0.75$; this is displayed in Part II of Table 2. The performance is pretty good for t neither too small nor too large, e.g., for $0.50 \leq t \leq 4.00$. Having $x < 1$ in this weighting function is consistent with the observation that the heavy-traffic limit

($\rho=1$) is more descriptive for ρ near 1 than the light traffic limit ($\rho=0$) is for ρ near 0. (Of course, the nice behavior near $\rho=1$ is achieved by the double scaling.) Approximation (2.25) is not very satisfying because, beyond the double scaling, it is ad hoc and because it evidently deteriorates for both small and large t. Some of our other approximations will do better in both regards.

2.8. The spectral representation

It is significant that the busy-period distribution can be represented as a mixture of exponential distributions, i.e., $B^c(t)$ is completely monotone; see Theorem 3.3 of [3]. An *explicit* representation as a mixture of exponentials follows from the spectral representation for $B^c(t)$. The spectral representation for the M/M/1 busy-period distribution was first given in (6.4) of KARLIN and MCGREGOR [32]; another derivation is given in [6]. With our time scaling, the spectral representation for the busy-period distribution has a remarkably simple form (taken from [6]), namely,

$$B^c(t) = \int_{\tau_1}^{\tau_2} e^{-t/y} w(y)\,dy, \quad t \geq 0, \tag{2.26}$$

where

$$w(y) = \frac{\theta\sqrt{(y-\tau_1)(\tau_2-y)}}{\pi\rho y^2}, \ \tau_1 \leq y \leq \tau_2,$$
$$\tau_1 = \frac{1+\rho-2\sqrt{\rho}}{2} \quad \text{and} \quad \tau_2 = \frac{1+\rho+2\sqrt{\rho}}{2} = \frac{(1+\sqrt{\rho})^2}{2}; \tag{2.27}$$

i.e., the mixing distribution has a proper probability density $w(y)$, which is given in (2.27). The mixing density $w(y)$ in (2.27) is unimodal with mode

$$y_{\max} = \frac{3(1+\rho)-\sqrt{1+34\rho+\rho^2}}{4}. \tag{2.28}$$

The relaxation time appears as the upper limit of integration τ_2.

Note that (2.26) is an ideal representation for obtaining numerical results by numerical integration. Also the asymptotic behavior as $t\to\infty$ (to be discussed in Section 4) can easily be derived by applying Laplace's method to (2.26).

As noted in [6], corresponding simple spectral representations also hold for the complementary first-moment cdf $H_1^c(t)$ and the correlation function $c_q(t)$, by virtue of the stationary-excess relations discussed in Section 2.3; it is elementary to integrate with respect to t in (2.26). These spectral representations for $H_1(t)$ and $c_q(t)$ easily extend to RBM by taking the limit as $\rho\to 1$. The nice spectral representation for the correlation function of RBM was previously discovered by WOODSIDE ET AL. [53]. The resulting mixing distributions are all simple modifications of the beta distribution.

2.9. Connections to the unrestricted process

As indicated in Section 7 of [4] and as discussed by others before, one way to analyze the M/M/1 queue is to relate it to the unrestricted process on all the integers with transition function $Q_{ij}(t)$. For this process, a basic role is played by $Q_{00}(t)$ or the scaled version (with our time scaling)

$$\gamma_\rho(t) = \theta^{-1}Q_{00}(t) = \theta^{-1}e^{-t/\tau}e^{-\nu}I_0(\nu), \tag{2.29}$$

which is a bonafide probability density function; see (7.7) of [4]. From Corollary 7.2.2 of [4], we immediately obtain the representation

$$\theta^{-1}B^c(t) = \left[\frac{1+\rho}{2\rho}\right][2\gamma_\rho(t)-\gamma_{\rho e}(t)]+\frac{\theta^2}{\rho}\gamma_\rho'(t), \quad t\geq 0, \tag{2.30}$$

where $\gamma_{\rho e}(t)$ is the density of the associated stationary-excess cdf defined by (2.7). The heavy-traffic limit (2.21) follows directly from (2.30) as indicated in [4]. Moreover, all M/M/1 approximations could be generated from approximations for the basic quantity $\gamma_\rho(t)$ in (2.29) and its relatives $\gamma_{\rho e}(t)$ and $\gamma_\rho'(t)$.

2.10. A differential equation for the busy-period density

We conclude this section with a curiosity. We point out that the busy-period density $b(t)$ satisfies a second-order linear differential equation with monomial coefficients, namely,

$$\theta^2 t b''(t)+[2(1-\theta)t+3\theta^2]b'(t)+[t+3(1-\theta)]b(t) = 0, \tag{2.31}$$

with boundary conditions

$$b(0)=1/2\theta^2 \quad \text{and} \quad b'(0)=-(1-\theta)/2\theta^4. \tag{2.32}$$

This follows from the Bessel function representation (2.8) and the differential equation for Bessel functions in 9.6.1 on p. 374 of [7].

3. Hyperexponential approximations

In this section we consider hyperexponential (H_2) approximations for the M/M/1 complementary busy-period cdf $B^c(t)$. An H_2 distribution is a mixture of two exponentials; i.e., a complementary H_2 cdf is of the form

$$G^c(t) \equiv 1-G(t) = pe^{-\lambda_1 t}+(1-p)e^{-\lambda_2 t}, \quad t\geq 0, \tag{3.1}$$

for $0\leq p\leq 1$, $\lambda_1\geq 0$ and $\lambda_2\geq 0$, so that it is determined by the parameter triple (p,λ_1,λ_2).

3.1. Motivation and methodology

We are motivated to consider H_2 approximations because we already have had considerable success using them to approximately describe the transient behavior of the M/M/1 queue and regulated Brownian motion (RBM) [1-5]. In particular, the first and second moment of the M/M/1 queue length at time t starting at 0 and the correlation functions of several stationary M/M/1 processes are well approximated (for t neither too small nor too large) by H_2

approximations obtained by matching the first three moments; see Tables 3 and 5 of [1], Tables 1-2, 5-7 of [3] and Tables 1-2, 6-9 of [5]. The qualification on time is somewhat vague and unpleasant, but as in Section 1.6 of [1], we contend that it is useful to consider three regimes: small time, medium time and large time, with the middle regime being when most M/M/1 processes reach steady state for most practical purposes, e.g., when the moments are within 0.1-15.0 percent of their steady-state limits. We evaluate our approximations with regard to their performance in the different regimes. Approximations that perform reasonably well in at least one regime seem worth considering, but we are primarily interested in the first two regimes; the third regime usually involves exceptionally large times that are of little practical interest.

As with many of the previous M/M/1 quantities studied, there is a strong theoretical basis for using an H_2 approximation for the busy-period distribution, because its density is completely monotone; i.e., the busy-period distribution *is* a mixture of exponentials, as indicated in Section 2.8. However, an H_2 fit still involves an approximation because the busy-period distribution is not a mixture of *two* exponentials, but uncountably many exponentials. Nevertheless, we might expect to capture the rather distinct small-time and large-time behavior of the busy-period distribution via the two exponential components.

An H_2 approximation for the busy-period cdf by moment matching is also attractive because the busy-period moments are readily available, as indicated in Section 2.5. Moreover, there are relatively straightforward schemes for expressing the H_2 parameter triple (p,λ_1,λ_2) in terms of the moment triple (m_1,m_2,m_3), i.e., the first three moments; see Section 5.1 of [1], especially (5.7) there. In fact, the three-moment H_2 approximation for the complementary busy-period cdf was previously developed by RIORDAN [42, p. 108] in his analysis of the waiting-time distribution in the M/M/1-LCFS system. (Riordan claims that the H_2 approximation matches the fourth moment too, but this seems to be incorrect.)

Moment matching usually is pretty accurate in describing the cdf for relatively large times, but not for small times. For smaller times, it is natural to match derivatives at the origin. As interest moves from times very near the origin to larger times, it is natural to shift from a three-derivative (3D) fit to a two-derivative, one-moment (2D, 1M) fit, to a one-derivative, two-moment (1D, 2M) fit, and eventually to a three-moment (3M) fit. Working with moments and derivatives at the origin, there are thus four candidate H_2 approximations we might consider. (We could of course obtain better numerical results by considering mixtures of more than two exponentials, but we would then lose the desired simplicity. Remember that the goal is not primarily numbers.)

It is significant that we can obtain all four of these H_2 approximations by the single three-moment matching procedure mentioned above. We obtain this simplification in fitting because the cases involving one or more derivatives are mapped into the three-moment case by the stationary-excess operator, as described in Section 7 and 8 of [5]. For a cdf $G(t)$ with a density $g(t)$, suppose that m_k is the k-th moment and d_k is the k-th derivative of $G(t)$ at $t=0$

(assume they exist); e.g., $d_1 = g(0)$. The stationary-excess cdf $G_e(t)$ associated with $G(t)$ is defined in (2.7). The key property is that the k-th moment m_{ek} and the k-th derivative (at $t=0$) d_{ek} of $G_e(t)$ are expressed in terms of the associated quantities of $G(t)$ by

$$m_{ek} = m_{k+1}/(k+1)m_1, \quad d_{e1} = 1/m_1 \quad \text{and} \quad d_{e(k+1)} = -d_{e1}d_k. \tag{3.2}$$

As a consequence, specifying the first i moments and first j derivatives of $G(t)$ is equivalent to specifying the first $(i-1)$ moments and first $(j+1)$ derivatives of $G_e(t)$. Moreover, the stationary-excess cdf associated with an H_2 cdf is another H_2 cdf with the same exponential parameters. Consequently, we only need to perform three-moment H_2 fits.

3.2. The H_2 formulas

Six H_2 approximations for the complementary busy-period cdf $B^c(t) \equiv 1 - B(t)$ are displayed in Table 3; they come from Table 4 of [5]. In addition to the four H_2 approximations based on derivatives and moments, we also display the simple one-moment exponential fit and the two-moment bounding H_2 distribution consisting of an atom at 0 plus a single exponential (the special case of (3.1) with $\lambda_1 = 0$). In a certain sense, all two-moment H_2 fits fall between these last two special cases; see [50].

It is significant for interpretation and further application, that we obtain all six H_2 approximations as relatively simple closed-form expressions of the traffic intensity ρ (the single M/M/1 parameter). Purely numerical results could be helpful too, but they would not provide nearly as much beyond numerical transform inversion of (2.2) or numerical integration of (2.26).

3.3. Performance for larger times (second regime)

Of course, it remains to describe how these various H_2 approximations perform. For relatively large times (e.g., $t \geq 1$ in scaled time), the quality of the approximations improves as we replace derivatives with moments; in particular, for larger times the three-moment H_2 fit is substantially better than the others. Table 4 compares the two leading candidates, (3M) and (1D, 2M), with the exact values obtained by numerical transform inversion in the case $\rho = 0.70$. (Other approximations appearing there will be discussed later.) Tables 5 and 6 display the (2D, 1M) H_2 approximation, along with other approximations to be discussed later, for the cases $\rho = 0.25$ and $\rho = 0.75$. For times away from zero with time scaling, replacing moments with derivatives clearly does not help.

To help put the H_2 three-moment approximation in perspective, we also compare it to HEYMAN's [26] diffusion approximation for the M/M/1 busy-period distribution in Table 4. This table and other cases show that the new H_2 three-moment approximations offer order-of-magnitude improvements for times away from zero. Heyman's approximation is motivated by GAVER [24] and is related to our inverse Gaussian approximation in Section 6.

Table 4 shows that the three-moment H_2 approximation for the busy-period distribution performs pretty well, but it does not perform nearly as well as

previous three-moment H_2 approximations for other M/M/1 quantities; cf. Tables 1-2, 6-7 of [3] and Tables 1-2 of [5]. In particular, the three-moment H_2 approximation deteriorates significantly as ρ increases. This is demonstrated by Table 7, which compares the three-moment approximations to the exact values for the cases $\rho=0.25$, $\rho=0.95$ and $\rho=0.99$. The relative percent errors (100[approx.−exact]/exact) are also given there. From Tables 4 and 7, we see that the (3M) H_2 approximation performs best in the second regime; the performance is not good for times that are very small or very large. (Indeed, it does not have the correct exponential rate as $t\to\infty$.) When ρ gets very high, the (3M) H_2 approximation really does not perform well at all. This observation motivated us to look for explanations and other approximations.

3.4. Performance for smaller times (first regime)

Of course, in the first regime, i.e., for very small time values (e.g., without time scaling), the derivatives do help. With time scaling, we achieve this by considering times of order θ^2, which undoes the time scaling. As indicated by Table 8, for such smaller times, the H_2 approximations based on three derivatives or two derivatives plus one moment perform very well. Moreover, for these times, the (3D) H_2 approximation is uniformly better than the (2D, 1M) approximation. As in Section 3.3, the quality of the approximation degrades somewhat as ρ increases.

In this time range, the relevant comparison is with the power series representation in Section 2.6. However, the three-derivative H_2 approximation performs better than a few terms from the power series. The power-series approximation, keeping terms up to t^3, is (in unscaled time)

$$\begin{aligned} B^c(t) &\approx 1-\frac{m_2}{2\theta}\left[\frac{t}{\theta^2}\right]+\frac{m_3}{12\theta}\left[\frac{t}{\theta^2}\right]^2-\frac{m_4}{144\theta}\left[\frac{t}{\theta^2}\right]^3 \\ &\approx 1-\frac{1}{2}\left[\frac{t}{\theta^2}\right]+\frac{(1+\rho)}{8}\left[\frac{t}{\theta^2}\right]^2-\frac{(1+3\rho+\rho^2)}{48}\left[\frac{t}{\theta^2}\right]^3. \end{aligned} \tag{3.3}$$

Three terms from the power series representation perform well only for times less than θ^2. For very large t, the truncated power series obviously performs very poorly.

When the three-derivative H_2 approximation is expanded in a power series, the first three terms necessarily coincide with the first three terms of the power series representation. In addition, the three-derivative H_2 approximation has approximately the correct shape overall. We thus conclude that the three-derivative H_2 approximation dominates the power-series representation, so that we do not consider the power-series representation further for approximations.

Furthermore, for $\rho=0.75$ the relative percent error using (4.1) is 11%, 6% and 4% for the large times $t=8$, 10 and 12, respectively. (From Table 1, we see that $B^c(8)=0.000037$, so that $t=8$ is already quite a large time.) Thus, just as in Table 3 of [1] and Table 5 of [3], we find that the standard asymptotic theory related to the relaxation time is not very useful for generating good approximations. (Of course, relaxation times do give at least a rough indication.)

In Corollary 5.2.3 of [4] we produced another asymptotic expansion for $B^c(t)$ related to the heavy-traffic limit in (2.21), which we call the *asymptotic normal approximation*. In particular, in [4] we show that

$$\theta^{-1}B^c(t) \sim \frac{2}{\rho^3/4}\left[t^{-1/2}\phi(\sqrt{2t/\tau})-\sqrt{2/\tau}[1-\Phi(\sqrt{2t/\tau})]\right] \quad \text{as} \quad t\to\infty, \quad (4.3)$$

where again τ is given in (2.9), $\theta=(1-\rho)/2$, Φ is the standard normal cdf and ϕ is its density. Since $\tau\to 2$ as $\rho\to 1$, the right side of (4.3) approaches the heavy-traffic limit $\bar{h}_1(t)$ in (2.21) as $\rho\to 1$. Moreover, as shown in [4], the two asymptotic expansions for $B^c(t)$ in (4.1) and (4.3) are asymptotically equivalent as $t\to\infty$. In fact, (4.3) was previously developed for the M/M/1-LCFS waiting-time distribution; see p. 109 of RIORDAN [42] and p. 45 of KOSTEN [34].

It turns out that the asymptotic normal approximation for $\theta^{-1}B^c(t)$ in (4.3) performs significantly better than both the standard asymptotic expansion in (4.1) and the heavy-traffic limit $\bar{h}_1(t)$ in (2.21). First, as indicated above, it coincides with the standard asymptotic expansion as $t\to\infty$ and the heavy-traffic limit as $\rho\to 1$. However, (4.3) performs significantly better than the standard asymptotic expansion for $t<\infty$ and the heavy-traffic approximation for $\rho<1$.

Numerical comparisons with exact values are made for the asymptotic normal approximation in Tables 5, 6, 10 and 11. (Inverse Gaussian approximations will be discussed in Section 6.) Tables 5 and 6 make numerical comparisons for the cases $\rho=0.25$ and $\rho=0.75$, while Table 10 makes numerical comparisons for $\rho=0.50$. As should be expected, these tables show that the quality of the approximation improves as t increases for fixed ρ and as ρ increases for fixed t. Table 11 shows the values of t, as a function of ρ, where the relative percent error is less than 1%. For $\rho\geq 0.5$, the asymptotic normal approximation is excellent in both the second and third regimes. In these regimes, it clearly dominates all the hyperexponential approximations discussed in Section 3. On the other hand, the asymptotic normal approximation for $B^c(t)$ is not good for small t; e.g., it need not even be a cdf. (An obvious refinement to any approximation for $B^c(t)$ is to replace values greater than 1 by 1.)

It is apparent from Tables 5, 6 and 10 that the asymptotic normal approximation is an upper bound on $B^c(t)$. However, we have yet to prove this.

3.7. The spectral representation

Further insight into the performance of the H_2 approximations is gained by considering the spectral representation, i.e., the explicit representation as a mixture of exponentials given in (2.26). For example, typical values for the case $\rho=0.75$ are given in Table 9. Note that the mixing density $w(y)$ in (2.27) has a sharp peak near the mode at 0.012 and very little mass at values near the upper limit τ_2 (the relaxation time). In this case, there is very little mass above $y=0.50$. On the other hand, the three-moment H_2 fit in this case is

$$B^c(t) \approx 0.933e^{-t/0.07}+0.0670e^{-t/0.93}, \quad t\geq 0. \tag{3.6}$$

From (3.6), we see that the H_2 approximation cannot be asymptotically correct as $t\to\infty$ because its relaxation time 0.93 does not agree with the true value 1.74. On the other hand, 0.93 is in the region above 0.50 where $w(y)$ has little mass. The low mass near τ_2 in $w(y)$ in (2.27) is a clear danger signal for approximation.

In contrast, the mixing densities for the first-moment-function complementary cdf $H_1^c(t)$ and the correlation function $c_q(t)$ studied in [3] and [5] are $(y/\theta)w(y)$ and $(2y^2/\theta)w(y)$, respectively, for $w(y)$ in (2.27). (These are easily obtained from using the stationary-excess operator, as mentioned in Section 2.8.) These other mixing densities clearly have much more mass near the upper end point τ_2, as is shown in Table 9. Indeed, the mixing density for the correlation function $c_q(t)$ is a symmetric beta distribution about $(1+\rho)/2$. Consequently, we should expect the H_2 approximations to perform better for $H_1(t)$ and $c_q(t)$ than for $B^c(t)$, as they do.

4. Asymptotic expansions

In this section we consider approximations associated with the exact asymptotic behavior as $t\to\infty$. The standard asymptotic expansion (three terms) for the complementary busy-period cdf $B^c(t)$ is

$$B^c(t) \sim \tau\theta L(t,\rho)\left[1-\frac{3}{t}\left[\frac{\tau}{2}\right]a_1+\frac{15}{t^2}\left[\frac{\tau}{2}\right]^2 a_2+O\left[\frac{1}{t^3}\right]\right], \tag{4.1}$$

where $\theta=(1-\rho)/2$ as in (2.3), $\tau=(1+\sqrt{\rho})^2/2$ as in (2.9), $a_1=1+\epsilon$, $a_2=1+\epsilon-\epsilon^2/2$, $\epsilon=\theta^2/4\tau\sqrt{\rho}$,

$$L(t,\rho) = (2\pi\rho^{3/2}t^3)^{-1/2}e^{-t/\tau}, \quad t\geq 0, \tag{4.2}$$

and $f(t)\sim g(t)$ means that $f(t)/g(t)\to 1$ as $t\to\infty$. The expansion (4.1) can be obtained directly from the Laplace transform $\hat{B}^c(s)$ in (2.2) by applying Heaviside's theorem; p. 254 of Doetsch [19] or p. 165 of Gnedenko and Kovalenko [25]. Alternatively, it can be obtained from (2.8) using 9.7.1 of [7] plus integration by parts.

However, even with all three terms displayed in (4.1), we get quite poor numerical results for times of primary practical interest. For example, large relative percent errors at medium times ($2\leq t\leq 5$) are shown in Table 10 for the case $\rho=0.5$. (Other approximations yet to be discussed also appear there.)

Furthermore, for $\rho=0.75$ the relative percent error using (4.1) is 11%, 6% and 4% for the large times $t=8$, 10 and 12, respectively. (From Table 1, we see that $B^c(8)=0.000037$, so that $t=8$ is already quite a large time.) Thus, just as in Table 3 of [1] and Table 5 of [3], we find that the standard asymptotic theory related to the relaxation time is not very useful for generating good approximations. (Of course, relaxation times do give at least a rough indication.)

In Corollary 5.2.3 of [4] we produced another asymptotic expansion for $B^c(t)$ related to the heavy-traffic limit in (2.21), which we call the *asymptotic normal approximation.* In particular, in [4] we show that

$$\theta^{-1}B^c(t) \sim \frac{2}{\rho^3/4}\left\{t^{-1/2}\phi(\sqrt{2t/\tau})-\sqrt{2/\tau}[1-\Phi(\sqrt{2t/\tau})]\right\} \quad \text{as} \quad t\to\infty, \quad (4.3)$$

where again τ is given in (2.9), $\theta=(1-\rho)/2$, Φ is the standard normal cdf and ϕ is its density. Since $\tau\to 2$ as $\rho\to 1$, the right side of (4.3) approaches the heavy-traffic limit $\tilde{h}_1(t)$ in (2.21) as $\rho\to 1$. Moreover, as shown in [4], the two asymptotic expansions for $B^c(t)$ in (4.1) and (4.3) are asymptotically equivalent as $t\to\infty$. In fact, (4.3) was previously developed for the M/M/1-LCFS waiting-time distribution; see p. 109 of RIORDAN [42] and p. 45 of KOSTEN [34].

It turns out that the asymptotic normal approximation for $\theta^{-1}B^c(t)$ in (4.3) performs significantly better than both the standard asymptotic expansion in (4.1) and the heavy-traffic limit $\tilde{h}_1(t)$ in (2.21). First, as indicated above, it coincides with the standard asymptotic expansion as $t\to\infty$ and the heavy-traffic limit as $\rho\to 1$. However, (4.3) performs significantly better than the standard asymptotic expansion for $t<\infty$ and the heavy-traffic approximation for $\rho<1$.

Numerical comparisons with exact values are made for the asymptotic normal approximation in Tables 5, 6, 10 and 11. (Inverse Gaussian approximations will be discussed in Section 6.) Tables 5 and 6 make numerical comparisons for the cases $\rho=0.25$ and $\rho=0.75$, while Table 10 makes numerical comparisons for $\rho=0.50$. As should be expected, these tables show that the quality of the approximation improves as t increases for fixed ρ and as ρ increases for fixed t. Table 11 shows the values of t, as a function of ρ, where the relative percent error is less than 1%. For $\rho\geq 0.5$, the asymptotic normal approximation is excellent in both the second and third regimes. In these regimes, it clearly dominates all the hyperexponential approximations discussed in Section 3. On the other hand, the asymptotic normal approximation for $B^c(t)$ is not good for small t; e.g., it need not even be a cdf. (An obvious refinement to any approximation for $B^c(t)$ is to replace values greater than 1 by 1.)

It is apparent from Tables 5, 6 and 10 that the asymptotic normal approximation is an upper bound on $B^c(t)$. However, we have yet to prove this.

5. Approximants from Widder's formula

In the general theory of Laplace transforms, there are techniques for generating relatively simple analytic approximate inversions called *approximants*. For example, in Section 3.6 we noted that our three-moment hyperexponential approximation for $B^c(t)$ can be obtained as a Padé approximant. In this section we briefly consider simple approximants obtained from Widder's inversion formula ([52], [30] or Section 3.1 of [17]), which represents a function $f(t)$ as the limit of successive derivatives of its Laplace transform $\hat{f}(s)$, namely,

$$f(t) = \lim_{n\to\infty} \frac{(-1)^n}{n!}\left[\frac{n+1}{t}\right]^{n+1} \hat{f}^{(n)}([n+1]/t), \tag{5.1}$$

where $\hat{f}^{(n)}(s)$ is the n-th derivative of $\hat{f}(s)$.

Formula (5.1) is not attractive for direct numerical inversion because it involves repeated differentiation, which leads to numerical instabilities, but it can be used to generate approximants. According to Davies and Martin [17], ter Haar [45] was the first to do this (in 1951), proposing essentially the 0-th term from (5.1),

$$f_0(t) \equiv t^{-1}\hat{f}(t^{-1}), \quad t \geqslant 0. \tag{5.2}$$

Jagerman [30,31] also proposed the ter Haar approximant and made a significant enhancement. In particular, Jagerman [30] observed that for a log-convex function we can determine the relaxation time constant from the singularity in the transform, so that we can force the approximant to have an exponential term with this time constant. The resulting Jagerman approximant is

$$f_\tau(t) \equiv \frac{e^{-t/\tau}}{(1-t/\tau)} f_0\left[\frac{t}{1-t/\tau}\right], \quad t \geqslant 0, \tag{5.3}$$

where $f_0(t)$ is the ter Haar approximant in (5.2) and τ is the relaxation time constant.

The ter Haar approximant (5.2) and the Jagerman approximant (5.3) are particularly appealing candidates because, as Jagerman [30, 31] shows, they inherit many of the structural properties of the original function $f(t)$. Thus, these approximants are clearly in the spirit of the simple closed-form expressions we are seeking.

Since the busy-period complementary cdf $B^c(t)$ is completely monotone, it is log-convex. Hence, it is natural to consider both the ter Haar and Jagerman approximants. In our time scale, the ter Haar approximant for $B^c(t)$ is

$$B_0^c(t) = \frac{2\theta}{\theta + t + \sqrt{\theta^2 + 2(1-\theta)t + t^2}}, \quad t \geqslant 0, \tag{5.4}$$

and the Jagerman approximant is

$$B_a^c(t) = e^{-t/\tau}\frac{2\theta}{\theta + (1-\theta/\tau)t + \sqrt{\theta^2 + 2\sqrt{\rho}\,t}}, \quad t \geqslant 0, \tag{5.5}$$

where τ is the relaxation time in (2.9).

Even though the ter Haar approximant (5.4) inherits several properties of $B^c(t)$; e.g., $B_0^6(0)=1$, $B_0^6(\infty)=0$, and $B_0^6(t)$ is monotone and convex, the numerical results are not good, as can be seen from Table 4. Indeed it is obvious that (5.4) does not nearly decay quickly enough, because the exponential term is missing.

We might expect much better performance from (5.5) because it has the correct asymptotic exponential decay rate. Indeed, as $\rho\to 0$, (5.5) agrees with the exact result $B^c(t)=e^{-2t}$. However, (5.5) is not asymptotically correct as $t\to\infty$, i.e., we do *not* have $B^c(t)\sim B_a^c(t)$ as $t\to\infty$. As $t\to\infty$, $B_a^c(t)\sim Kt^{-1}e^{-t/\tau}$; there is a t in the denominator instead of the $t^{3/2}$ in (4.1) and (4.2), so $B_a^c(t)$ also does not decay quickly enough, as Table 4 shows. Indeed (5.5) is certainly much better than (5.4), but (5.5) typically becomes greater than the standard asymptotic expansion for $2\leq t\leq 3$ (with time scaling). Thus, neither approximant performs well in the second and third regimes. In fact, even though (5.5) is asymptotically exact as $\rho\to 0$, it does not even perform well in the second regime for very small ρ such as $\rho=0.1$. The reason is that the order of the two limits $t\to\infty$ and $\rho\to 0$ matters.

From Sections 3 and 4 here plus our previous work [1-5], we know that for times of primary interest the actual exponential rate is substantially greater than the asymptotic exponential rate. Hence, we are motivated to force a bigger rate, as arises in the H_2 approximations in Section 3. Unfortunately, however, the Jagerman enhancement in [30] permits any smaller rate, but not any larger rate, so that we cannot improve the approximation in that way.

One approach to the poor performance of these approximants is to replace them by higher-order terms from the sequence (5.1), but we do not pursue this goal. With higher order terms, Theorem 8 of [30] can also be applied in various ways to match moments, but of course the resulting approximant gets more complicated. It is good to be aware that improved approximants by this general approach are possible, though. Indeed, JAGERMAN [30,31] uses (5.1) as a basis for creating a full numerical inversion procedure. To avoid the troublesome differentiation in (5.1), Jagerman applies complex analysis to obtain a Fourier-series type inversion procedure from (5.1). Since such higher-order approximants are not closed-form, they are perhaps best compared to other numerical inversion schemes and numerical integration. From Tables III-VI of [31], it is apparent that the higher-order approximants for $B^c(t)$ do not get accurate very quickly. It appears that $B^c(t)$ is a relatively difficult function to approximate by this general procedure.

6. INVERSE GAUSSIAN APPROXIMATIONS

Inverse Gaussian approximations for the busy-period distribution arise naturally from diffusion approximations.

6.1. Heavy-traffic limits and RBM approximations

First, the inverse Gaussian (IG) cdf $F(t;x,0)$ in (2.23) is the cdf of the time required for canonical RBM to first reach 0 starting in x. Thus an IG cdf is a natural approximation for the cdf of a first passage time downward in a queue when we are using a diffusion (RBM) approximation for the queueing process. Indeed, an IG cdf arises as the heavy-traffic limit for queueing first passage times downward as $\rho \to 1$, as indicated in the proof of Corollary 3.4.1 of [4]. In particular, consider the time-scaled M/M/1 queue and let $T_n(\rho)$ be the first-passage time from state n to 0, as a function of ρ. (Of course, $T_n(\rho)$ is the sum of n independent busy periods.) Let $T_x(1)$ be the first-passage time from x to 0 in RBM (the M/M/1 queue with $\rho = 1$) with IG cdf $F(t;x,0)$ in (2.23). Let $\Rightarrow$ denote convergence in distribution, as in BILLINGSLEY [11]. Then the standard heavy-traffic limit for $T_n(\rho)$ as $\rho \to 1$, obtained by applying the continuous mapping theorem (Theorem 5.1 of [11]) with the usual heavy-traffic limit for the queue-length process [29], is

$$T_{[x\theta^{-1}]}(\rho) \Rightarrow T_x(1) \quad \text{as} \quad \rho \to 1, \tag{6.1}$$

where $[x]$ is the greatest integer less than or equal to x. It is significant that essentially the same result holds for GI/G/m queues and many other models, so that the limit (6.1) suggests IG approximations for more general systems.

A difficulty with (6.1) for our purposes is that the starting state in the M/M/1 system with traffic intensity ρ is $[x\theta^{-1}]$, where

$$[x\theta^{-1}] = [2x/(1-\rho)] \to \infty \text{ as } \rho \to 1. \tag{6.2}$$

However, we can ignore difficulty (6.2), and suppose that

$$T_{[x\theta^{-1}]}(\rho) \approx T_x(1) \tag{6.3}$$

for fixed $x > 0$ and $\rho < 1$. We then set $x\theta^{-1} = 1$ to make the left side equal to the desired $T_1(\rho)$; i.e., we approximate by

$$T_1(\rho) \approx T_\theta(1), \tag{6.4}$$

i.e., the first-passage time for canonical RBM starting in θ. We will come back to this later.

Instead of (6.1) and (6.4), we would probably prefer a limit directly for $T_1(\rho)$ as $\rho \to 1$. Fortunately, such a limit has been obtained, as indicated in Section 2.7, but the limit is not in the usual form, involving $[T_1(\rho) - a(\rho)]/b(\rho)$ for some functions of ρ, $a(\rho)$ and $b(\rho)$. In particular, (2.21)-(2.23) states that

$$\theta^{-1} B^c(t) = \frac{2}{1-\rho} P(T_1(\rho) > t) \to \tilde{h}_1(t) = \int_0^\infty 2e^{-2x} f(t;x,0)dx; \tag{6.5}$$

i.e., the limit is a mixture of IG *densities*, which seems to be not an especially convenient form.

6.2. Heuristic diffusion approximation

Diffusion process approximations can also be developed by 'direct fits,' without considering any heavy traffic limits; this is the approach of HEYMAN [26] and DUDA [20,21] for example. The idea is to apply some of the logic of the queue and some of the logic of the diffusion process. For example, Heyman works with the virtual waiting time process and approximates it by RBM. Since he works with the virtual waiting time process, the logic of the queue dictates that the initial level at the beginning of a busy period be the random amount determined by the service-time distribution. Thereafter, the process is assumed to evolve as RBM. Thus the approximation for the busy-period cdf is a mixture of IG cdf's, as in (6.5).

On the other hand, if we work with the queue-length process, as we do here, then with this procedure we would approximate the queue-length process by RBM but not randomize the initial position. We would thus approximate the busy-period cdf by the cdf of an IG distribution, without any randomization. Of course, with this procedure we still must specify the drift and diffusion coefficients of RBM and the initial state for the first passage time. The end result might then be (6.4). We mention that Heyman found that his approximation for the M/G/1 busy-period distribution worked best when the service-time distribution is deterministic, i.e., when no randomization was performed. As indicated in Table 4, Heyman's M/M/1 approximation does not perform very well. As discussed in [48], there seems to be a limit to what you can obtain by the diffusion logic alone. It is important to consider refinements based on additional properties such as moments.

6.3. Direct inverse Gaussian approximations

We propose using the heavy-traffic limit to motivate considering the IG cdf as an approximation for the busy-period cdf, but fitting the IG parameters directly. First, we note that there is a one-to-one correspondence between a two-parameter IG cdf and a two-parameter RBM representation. The first-passage time for RBM with drift coefficient $\mu<0$ and variance coefficient ν starting in state x has the complementary IG cdf

$$G^c(t) = \Phi^c\left[\frac{-\mu t - x}{\sqrt{\nu t}}\right] - e^{-\frac{2\mu x}{\nu}}\Phi^c\left[\frac{-\mu t + x}{\sqrt{\nu t}}\right], \quad t \geq 0. \tag{6.6}$$

However, there are only two basic parameters in (6.6). We can set $\mu = -1$ without loss of generality, and do, because the parameter triples (μ, ν, x) and $(-1, \nu/\mu^2, x/\mu)$ yield identical distributions in (6.6), as is easy to see. (We can also apply Proposition 2.1 of [1] for this.) The IG distribution with parameters $\mu = -1$, ν and x has moments

$$\begin{aligned} &m_1 = x, \ m_2 = x(x+\nu), \ m_3 = 3x\nu(x+\nu)+x^3, \\ &m_{n+1} = \sum_{i=0}^{n} \frac{(n+i)!x^{n+1-i}\nu^i}{(n-i)!2^i i!}; \end{aligned} \tag{6.7}$$

see p. 366 (12) of TWEEDIE [46].

A specific IG approximation is obtained by choosing x and ν in (6.6). We choose x and ν by matching moments to those of $B(t)$ in Section 2.4 and possibly by matching other properties. First, to match the mean, we set $x=\theta=(1-\rho)/2$. We then consider three cases of ν:

(i) $\nu=1$ (canonical RBM)

(ii) $\nu=1-\theta$ (all moments match up to order θ^3) (6.8)

(iii) $\nu=\dfrac{\tau}{2}\approx 1-\theta-\dfrac{\theta^2}{4}-\dfrac{\theta^3}{4}$ (proper exponential rate as $t\to\infty$).

The first case in (6.8) with $\nu=1$ is just the heavy-traffic approximation (6.4) involving canonical RBM derived from (6.1). The second case, $\nu=1-\theta$, is obtained by matching the second moment, $\theta(\theta+\nu)=\theta$. From (6.5), we see that in this second case

$$m_{n+1} = (2n-1)!!(\theta-(n-1)\theta^2)+O(\theta^3), \tag{6.9}$$

where $(2n-1)!!=(2n-1)(2n-3)\cdots(3)1$, which together with (2.11) implies that *all* moments in (6.7) match the moments of $B(t)$ in Section 2.5 up to order $O(\theta^3)$ when $\nu=1-\theta$. Indeed, we can obtain the whole IG approximation directly from the moments of the busy period in this way. In particular, we can represent each busy-period moment by a power series in θ and look for a distribution that matches the moment sequence obtained by keeping the leading terms (matching terms up to order θ^3). From Section 2.5 and (6.7), we find that the IG distribution with $\nu=1-\theta$ solves this asymptotic moment problem. We thus should expect case (ii) to perform very well for higher traffic intensities, which it does.

The third case in (6.8), $\nu=\tau/2$, is just what is needed to give the IG distribution the proper exponential rate, as we would expect from (4.1) and (4.3); i.e., from (6.6) and (4.1), we easily obtain that

$$G^c(t) \sim A(\theta)B^c(t) \quad \text{as} \quad t\to\infty, \tag{6.10}$$

where

$$A(\theta) = \frac{(\tau/2)^{-1/2}e^{2\theta/\tau}}{\rho^{-3/4}} \tag{6.11}$$

$$= \frac{\rho^{3/4}(1+\sqrt{\rho})^2 e^{2(1-\rho)/(1+\sqrt{\rho})^2}}{4} \to 1 \quad \text{as} \quad \rho\to 1.$$

The third case also matches all moments up to order $O(\theta^2)$.

In fact, the corresponding densities also have the almost asymptotic property (6.10). Indeed all the approximations for the complementary cdf $B^c(t)$ yield corresponding approximations for the density $b(t)$ by differentiating, but in general the quality of approximations can deteriorate drastically upon differentiation. However, the IG density $g(t)$ in case (iii) of (6.8) is also good for $b(t)$. Indeed,

$$g(t) \sim A(\theta)b(t) \quad \text{as} \quad t\to\infty, \tag{6.12}$$

for $A(\theta)$ in (6.11). In fact, we can make the IG approximations for $b(t)$ and $B^c(t)$ asymptotically correct by dividing $g(t)$ and $G^c(t)$ by $A(\theta)$ in (6.11). However, then $g(t)$ is no longer a probability density function and $G^c(t)$ is no longer a complementary cdf. This modification gives the best numbers for t not too small, though; see Section 6.5 below.

It is furthermore significant that all three IG cdf's are asymptotically correct in heavy traffic, in the strong sense that

$$\lim_{\theta \to 0} \theta^{-1}[B^c(t) - G^c(t)] = 0, \tag{6.13}$$

i.e., in addition to the heavy-traffic limit theorem for $B^c(t)$ in (2.21),

$$\lim_{\theta \to 0} \theta^{-1} G^c(t) = \tilde{h}_1(t). \tag{6.14}$$

Indeed, the limit is true for the density

$$\lim_{\theta \to 0} \theta^{-1} g(t) = \tilde{h}_1'(t) = -L(t, 1) \tag{6.15}$$

for $L(t, \rho)$ in (4.2). The limit in (6.15) is easily established by differentiating (6.6) and taking the limit.

We remark that we do not yet have any direct probabilistic explanation why we should change the variance coefficient ν to the second or third candidate in (6.8) in the context of RBM approximations for the queue.

6.4. Numerical comparisons

Numerical comparisons of the IG approximations with exact values for the M/M/1 complementary busy period cdf $B^c(t)$ and other approximations appear in Tables 5, 6, 10, 11 and 12. Table 10 compares the three IG approximations with the asymptotic approximations in Section 4 for the case $\rho = 0.50$. The relative percent errors are displayed in Table 10.

From Tables 4 and 10, we see that for the first IG approximation in (6.8) involving canonical RBM with $\nu = 1$ performs better than Heyman's diffusion approximation using the random initial conditions, but not quite as well as the three-moment H_2 approximation. More importantly, this first IG approximation is clearly dominated by the other two IG approximations with $\nu = 1 - \theta$ and $\nu = \tau/2$. The second and third IG approximations perform very well except for very small t. There is relatively little difference between these two approximations when ρ is not too small (e.g., $\rho \geq 0.50$) because

$$\tau/2 = (1 + \sqrt{\rho})^2/4 = (1 + \rho + 2\sqrt{\rho})/4 \quad \text{and} \quad 1 - \theta = (1 + \rho)/2. \tag{6.16}$$

The third IG approximation may be preferred because it is nearly asymptotically correct as $t \to \infty$; it is good for the third regime as well as the second regime.

Tables 5 and 6 show that the asymptotic normal approximation in (4.3) is better than the second IG approximation with $\nu = 1 - \theta$ for very large t, but the IG approximations are better for small t. Tables 5 and 6 also show that the quality of the IG approximation with $\nu = 1 - \theta$ improves with increasing ρ; it is

much better for $\rho=0.75$ than at $\rho=0.25$. Table 11 gives the range for which the relative percent error is less than one percent. For $\rho\geq 0.5$, the second and third IG approximations perform very well in the second regime; the third IG approximation also performs well in the third regime.

We have seen that the IG distribution is not a good approximation for very small times. This should not be surprising because the heavy-traffic limit (6.1) focuses on relatively large times. Indeed, the IG density $f(t;x,0)$ is 0 and increasing at $t=0$, while the true M/M/1 busy-period density is positive and strictly decreasing. However, the IG distribution does describe the busy-period distribution remarkably well when we consider larger times.

6.5. Density approximations

From (2.21) and (6.12), we should expect that the IG *density*, using (iii) in (6.8), is a good approximation for the M/M/1 busy period *density* $b(t)$ for t sufficiently large, and indeed it is.

However, we have just noted that $g(0)=0$, while $b(0)=1/2\theta^2$. In fact, all derivatives of g at 0 are 0. In contrast, $b'(0)=-(1-\theta)/2\theta^4$; see (2.32). Hence, for small t the IG approximation is very bad. It is natural to ask how large t must be for this bad initial match to disappear. This is partly answered by properties of the IG density. From p. 365 of TWEEDIE [46] we see that the IG density is unimodal with mode at (in our time scale)

$$t_{\text{mode}} = \theta\left(\left(1+\frac{9\tau^2}{16\theta^2}\right)^{1/2}-\frac{3\tau}{4\theta}\right) \approx \frac{2\theta^2}{3\tau} \approx \frac{\theta^2}{3}, \tag{6.17}$$

so that the initial period where the IG density has the wrong slope is brief. In fact, numerical evidence indicates that the IG density $g(t)$ is a good approximation for $t\geq 2\theta^2$, i.e., after about one service time. Comparisons between the IG density $g(t)$ and the busy-period density $b(t)$ for the cases $\rho=0.50$ and 0.75 appear in Tables 13 and 14.

The *relative percent error,*

$$\text{RE} \equiv \frac{g(t)-b(t)}{b(t)}\times 100, \tag{6.18}$$

is also given in Tables 13 and 14. Interestingly, the relative percent error quite rapidly approaches the *asymptotic relative percent error,*

$$\text{ARE} \equiv \lim_{t\to\infty}\text{RE} = 100(A(\theta)-1) \tag{6.19}$$

where $A(\theta)$ is given in (6.11). Values of the ARE appear in Table 15. The IG approximation in (iii) of (6.8) truly does exceptionally well as long as ρ and t are not too small.

7. A COMPOSITE APPROXIMATION

None of the approximations considered so far performs well for all t and ρ. However, it is easy to form a composite approximation that does. For this purpose, we suggest using the third inverse Gaussian (IG) approximation with

$\nu=\tau/2=(1+\rho+2\sqrt{\rho})/4$ from Section 6 when $t\geq t^*(\rho)$ and the three-derivative (3D) H_2 approximation from Section 3 when $t<t^*(\rho)$, where

$$t^*(\rho) = \frac{3\theta^2}{2\rho} = \frac{3(1-\rho)^2}{8\rho}. \tag{7.1}$$

From our previous analysis, we see that the composite IG-3D approximation is asymptotically exact as $\rho\to 0$, $\rho\to 1$, and as $t\to 0$. By (6.10)-(6.12), it is nearly asymptotically exact as $t\to\infty$, and can be modified to be asymptotically exact, as indicated in Section 6.

Numerical comparisons supporting the composite IG-3D approximation appear in Table 12. The relative percent error is given for the cases $\rho=0.25$, 0.50, 0.75 and 0.85. For $\rho\geq 0.50$, the relative percent error is evidently never greater than 2.8%.

8. Extensions to the GI/G/1 model

The principal approximations considered in this paper apply equally well to the M/G/1 queue with general service-time distribution, because the moments and derivatives of the M/G/1 busy-period cdf $B(t)$ are readily available. In particular, for the M/G/1 queue we propose the IG approximation in Section 6 based on a two-moment match for the second regime and the IG-3D composite in Section 7 overall.

With the identical time scaling, the first two moments of the M/G/1 busy period are $m_1=\theta$ and $m_2=\theta(c_s^2+1)/2$, where c_s^2 is the squared coefficient of variation (variance divided by the square of the mean) of the service-time distribution. The resulting IG approximation obtained from matching the first two moments is thus (6.6) with $x=\theta$, just as for M/M/1, and

$$\nu = \frac{c_s^2+1}{2} - \theta = \frac{c_s^2+\rho}{2}. \tag{8.1}$$

In fact, however, for the GI/G/1 queue we would scale time by $(c_a^2+c_s^2)/(1-\rho)^2$, where c_a^2 is the squared coefficient of variation of the interarrival time. This is the proper scaling to yield canonical RBM in the limit as $\rho\to 1$; see [29] and Section 2 of [1]. For the M/G/1 queue, with time scaled by $(1+c_s^2)/(1-\rho)^2$, the first two moments of the busy period again coincide, yielding

$$m_1 = m_2 = \frac{1-\rho}{1+c_s^2}. \tag{8.2}$$

With this time scaling and the same approximation procedure, (ii) in (6.8), we thus set $x=x(x+\nu)=m_1$ to obtain the IG approximation (6.6) with

$$x = \frac{1-\rho}{1+c_s^2} \quad \text{and} \quad \nu = 1-x. \tag{8.3}$$

For GI/G/1 with the time scaling $(c_a^2+c_s^2)/(1-\rho)^2$, we do not have the busy-period moments in such a convenient form. Nevertheless, it is natural to consider the IG approximation in (6.6) with

$$x = \frac{1-\rho}{c_a^2+c_s^2} \quad \text{and} \quad \nu = 1-x, \tag{8.4}$$

provided that $x<1$. Note that (8.4) reduces to (8.3) when $c_a^2=1$ and (8.3) reduces to the second M/M/1 case in (6.8) with $\nu=1-\theta$ when $c_s^2=1$. Also note that this IG approximation differs from what appears in (6.4) and DUDA [20,21] because $\nu<1$ in (8.4). Table 10 shows the improvement for the M/M/1 case. We intend to discuss approximations for the GI/G/1 busy-period distribution and make numerical comparisons in a subsequent paper.

9. Related quantities of interest

Approximations for the M/M/1 complementary busy-period cdf $B^c(t)$ immediately translate into approximations for the M/M/1-LCFS waiting-time cdf due to [41,42] and the M/M/1 probability of emptiness $P_{00}(t)$ due to (2.5). From the probability of emptiness, we then get the expected cumulative idle time $EI_0(t)$ and the mean queue length $EQ_0(t)$ via (2.6). Equivalently, from [3] we know that the normalized mean queue length $H_1(t)$ is just

$$H_1(t) = B_e(t) = (1-\rho)\int_0^t B^c(u)du, \quad t\geq 0. \tag{9.1}$$

Consequently, any approximation for $B^c(t)$ translates into an approximation for $H_1(t)$ through an integration. However, the relatively simple direct H_2 approximation for $H_1(t)$ developed in [3] still seems preferable. (Direct H_2 approximations are compared to H_2 approximations obtained from integration in [5].)

The approximation for $H_1(t)$ in [3] then can be used to obtain an approximation for $EI_0(t)$ via (2.6), namely,

$$EI_0(t) = (1-\rho)t + EQ_0(t) = (1-\rho)t + \frac{\rho}{1-\rho} + \frac{\rho}{1-\rho}H_1^c(t). \tag{9.2}$$

In fact, KUMAR and WONG [35] propose the linear portion of (9.2) as an approximation (and bound) for $EI_0(t)$; i.e., they use (9.2) with $H_1^c(t)\approx 0$, and show that it captures the essential behavior. This linear approximation can also be motivated by the fact (Theorem 9.3 of [4]), that $EI_0(t)$ coincides with a constant (the mean service time) multiplied by the renewal function associated with the busy-period distribution. Renewal functions are well-known to be asymptotically linear.

10. Conclusions

We have investigated several closed-form approximations for the M/M/1 complementary busy-period cdf $B^c(t)$. For larger times, the best approximations are the asymptotic normal approximation in (4.3) and the inverse Gaussian (IG) distribution in (6.6) with $\nu=1-\theta$ or $\nu=\tau/2$ (cases (ii) and (iii) in (6.8)). The light-and-heavy-traffic interpolation in (2.25) also performs well. For small times, the three-derivative (3D) hyperexponential (H_2) approximation in Table 3 performs very well. For very small times the power series

representation in Section 2.6 is also good. A single composite approximation based on (3D-H_2) and (IG) is proposed in Section 7.

The three-moment (3M) H_2 approximations, which worked so well in [1-5], do not perform as well for the M/M/1 busy period (Section 3). Simple diffusion approximations as proposed by HEYMAN [26] and DUDA [20,21] do not perform especially well, as shown in Table 4, but our proposed inverse Gaussian approximations build on these diffusion approximations. Approximants based on Widder's formula (Section 5) also do not perform very well here, but these are very general procedures, which in some circumstances are excellent.

The M/M/1 complementary busy-period cdf has proven to be difficult to approximate. As illustrated by the relatively poor performance of the (3M) H_2 approximations, it is not enough to match three moments. As illustrated by the poor performance of the three-term asymptotic expansion in (4.1), it is not nearly enough to capture the asymptotic behavior as $t \to \infty$. As illustrated by the relatively poor performance of unrefined diffusion approximations in [20,21,26], it is not enough to consider the standard heavy-traffic limits and diffusion approximations. However, by combining several of these properties, it is possible to obtain quite good approximations. The positive approximation results provide a basis for treating busy periods in more general models (Section 8) and other M/M/1 quantities (Section 9).

ACKNOWLEDGMENT

We thank D. L. Jagerman, A. Kumar and W. S. Wong for helpful discussions about their work and for suggesting that we consider approximants based on Widder's formula (Section 5).

REFERENCES

1. J. ABATE, W. WHITT (1987). Transient behavior of regulated Brownian motion, I: starting at the origin. *Adv. in Appl. Probab. 19,* 560-598.
2. J. ABATE, W. WHITT (1987). Transient behavior of regulated Brownian motion, II: non-zero initial conditions. *Adv. in Appl. Probab. 19,* 599-631.
3. J. ABATE, W. WHITT (1987). Transient behavior of the M/M/1 queue starting at the origin. *Queueing Systems 2,* 41-65.
4. J. ABATE, W. WHITT (1988). Transient behavior of the M/M/1 queue via Laplace transforms. *Adv. in Appl. Probab. 20,* to appear.
5. J. ABATE, W. WHITT (1988). The correlation functions of RBM and M/M/1. *Stochastic Models,* to appear.
6. J. ABATE, W. WHITT (1989). Simple spectral representations for the M/M/1 queue. *Queueing Systems,* to appear.
7. M. ABRAMOWITZ, I.A. STEGUN (eds.) (1982). *Handbook of Mathematical Functions,* Dover, New York.
8. N.T.J. BAILEY (1954). A continuous time treatment of a simple queue using generating functions. *J. Roy. Statist. Soc. B 16,* 288-291.
9. N.T.J. BAILEY (1957). Some further results in the non-equilibrium theory of a simple queue. *J. Roy. Statist. Soc. B 19,* 326-333.

10. G.A. BAKER, JR. (1975). *Essentials of Padé Approximants,* Academic Press, New York.
11. P. BILLINGSLEY (1968). *Convergence of Probability Measures,* Wiley, New York.
12. O.J. BOXMA, J.W. COHEN, N. HUFFELS (1979). Approximations of the mean waiting time in an M/G/s queueing system. *Oper. Res. 27,* 1115-1127.
13. G.A. CAMPBELL, R.M. FOSTER (1948). *Fourier Integrals for Practical Application,* Van Nostrand, New York.
14. G.M. CLARK (1981). Use of Polya distributions in approximate solution to nonstationary M/M/s queues. *Commun. ACM 24,* 206-217.
15. J.W. COHEN (1982). *The Single Server Queue,* 2nd ed., North-Holland, Amsterdam.
16. J.W. COHEN, G. HOOGHIEMSTRA (1981). Brownian excursion, the M/M/1 queue and their occupation times. *Math. Oper. Res. 6,* 608-629.
17. B. DAVIES, B. MARTIN (1979). Numerical inversion of the Laplace transform. *J. Comp. Phys. 33,* 1-32.
18. L.E.N. DELBROUCK (1976). Approximations for certain congestion functions in single server queueing systems. *Eighth Int. Teletraffic Congress, 233-1-5.*
19. G. DOETSCH (1974). *Introduction to the Theory and Application of the Laplace Transformation,* Springer-Verlag, New York.
20. A. DUDA (1983). Transient diffusion approximation for some queueing systems. *Perf. Eval. Review 12,* 118-128.
21. A. DUDA (1986). Diffusion approximations for time-dependent queueing systems. *IEEE J. Sel. Areas Commun. SAC-4,* 905-918.
22. S. ERLANDER (1965). The remaining busy period for a single server queue with Poisson input. *Oper. Res. 13,* 734-746.
23. D.P. GAVER, JR. (1966). Observing stochastic processes and approximate transform inversion. *Oper. Res. 14,* 444-459.
24. D.P. GAVER, JR. (1968). Diffusion approximation and models for certain congestion problems. *J. Appl. Probab. 5,* 607-623.
25. B.V. GNEDENKO, I.N. KOVALENKO (1968). *Introduction to Queueing Theory,* Israel Scientific Translations Ltd., Jerusalem.
26. D.P. HEYMAN (1974). An approximation for the busy period of the M/G/1 queue using a diffusion model. *J. Appl. Probab. 11,* 159-169.
27. D.P. HEYMAN, M.J. SOBEL (1982). *Stochastic Models in Operations Research,* Vol. I, McGraw Hill, New York.
28. P. HOKSTAD (1979). On the relationship of the transient behavior of a general queueing model to its idle and busy period distributions. *Math. Operationsforsch. Statist. Ser. Optimization 10,* 421-429.
29. D.L. IGLEHART, W. WHITT (1970). Multiple channel queues in heavy traffic, II: sequences, networks and batches. *Adv. in Appl. Probab. 2,* 355-369.
30. D.L. JAGERMAN (1978). An inversion technique for the Laplace transform with applications to approximation. *The Bell System Tech. J. 57,* 669-710.

31. D.L. JAGERMAN (1982). An inversion technique for the Laplace transform. *The Bell System Tech. J. 61,* 1995-2002.
32. S. KARLIN, J. MCGREGOR (1958). Many server queueing processes with Poisson input and exponential service times. *Pacific J. Math. 8,* 87-118.
33. D.G. KENDALL (1951). Some problems in the theory of queues. *J. Roy. Statist. Soc. B 13,* 151-171.
34. L. KOSTEN (1973). *Stochastic Theory of Service Systems,* Pergamon Press, New York.
35. A. KUMAR, W.S. WONG (1987). *Some Mean Value Formulas for the Transient M/M/1 Queue,* unpublished paper, AT&T Bell Laboratories, Holmdel, NJ.
36. Y.L. LUKE (1969). *The Special Functions and their Approximants,* Academic Press, New York.
37. R.E. NANCE, U.N. BHAT, B.G. CLAYBROOK (1972). Busy period analysis of a time-sharing system: transform inversion. *J. Assoc. Comput. Mach. 19,* 453-462.
38. G.F. NEWELL (1982). *Applications of Queueing Theory,* 2nd ed., Chapman and Hall, London.
39. A.R. ODONI, E. ROTH (1983). An empirical investigation of the transient behavior of stationary queueing systems. *Oper. Res. 31,* 432-455.
40. S.O. RICE (1962). Single server systems — II. Busy periods. *The Bell System Tech. J. 41,* 279-310.
41. J. RIORDAN (1961). Delays for last-come first-served service and the busy period. *The Bell System Tech. J. 40,* 785-793.
42. J. RIORDAN (1962). *Stochastic Service Systems,* Wiley, New York.
43. H. STEHFEST (1970). Algorithm 368. Numerical inversion of Laplace transforms. *Commun. ACM 13,* 47-49 [erratum *13,* 624].
44. L. TAKÁCS (1967). *Combinatorial Methods in the Theory of Stochastic Processes,* Wiley, New York.
45. D. TER HAAR (1951). An easy approximate method of determining the relaxation spectrum of a viscoelastic material. *J. Polymer Sci. 6,* 247-250.
46. M.C.K. TWEEDIE (1957). Statistical properties of inverse Gaussian distributions I. *Ann. Math. Statist. 28,* 362-377.
47. E. VAULOT (1954). Délais d'attente des appels téléphoniques dans l'ordre inverse de leur arrivée. *C. R. Acad. Sci.* (Paris) *238,* 1188-1189.
48. W. WHITT (1982). Refining diffusion approximations for queues. *Oper. Res. Letters 1,* 165-169.
49. W. WHITT (1984). On approximations for queues, I: extremal distributions. *AT&T Bell Lab. Tech. J. 63,* 115-138.
50. W. WHITT (1984). On approximations for queues, III: mixtures of exponential distributions. *AT&T Bell Lab. Tech. J. 63,* 163-175.
51. W. WHITT (1985). The renewal-process stationary-excess operator. *J. Appl. Probab. 22,* 156-167.
52. D.V. WIDDER (1934). The inversion of the Laplace integral and the

related moment problem. *Trans. Amer. Math. Soc. 36,* 107-200.
53. C.M. WOODSIDE, B. PAGUREK, G.F. NEWELL (1980). A diffusion approximation for correlation in queues. *J. Appl. Probab. 17,* 1033-1047.

Appendix A

time	values of the traffic intensity							
t	$\rho = 0.00$	$\rho = 0.10$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.85$	$\rho = 0.95$	$\rho = 0.99$
$\theta^2/2$	.7788	.7812	.7848	.7905	.7961	.7983	.8004	.8012
θ^2	.6065	.6140	.6249	.6422	.6584	.6646	.6707	.6731
0.10	.8187	.7836	.7116	.5158	.2394	.1356	.0425	.00829
0.25	.6065	.5495	.4501	.2676	.1143	.0646	.0203	.00398
0.50	.3679	.3128	.2344	.1303	.0563	.0321	.0102	.00201
0.75	.2231	.1837	.1348	.0764	.0340	.0196	.00628	.00124
1.00	.1353	.1109	.0827	.0489	.0224	.0130	.00422	.00084
1.50	.0498	.0432	.0352	.0229	.0112	.00664	.00219	.00044
2.00	.0183	.0180	.0165	.0119	.00619	.00374	.00125	.00025
3.00	.00248	.00358	.00423	.00378	.00221	.00139	.000480	.000097
4.00	.000335	.000796	.00122	.00135	.000888	.000577	.000206	.000042
5.00	.000045	.000189	.000380	.000514	.000381	.000256	.000094	.000020
6.00	.000006	.000047	.000123	.000205	.000171	.000119	.000045	.000009
8.00	.000000	.000003	.000014	.000036	.000037	.000028	.000011	.000002

TABLE 1. *Numerical values of the time-scaled M/M/1 complementary busy-period cdf $B^c(t)$ obtained by Laplace transform inversion of (2.2).*

Part I: exact values of $\theta^{-1} B^c(t)$								
time	values of the traffic intensity							
t	$\rho = 0.00$	$\rho = 0.10$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.85$	$\rho = 0.95$	$\rho = 0.99$
0.10	1.64	1.74	1.90	2.06	1.92	1.81	1.70	1.66
0.25	1.21	1.22	1.20	1.07	0.91	0.86	0.81	0.80
0.50	0.74	0.70	0.63	0.52	0.45	0.43	0.41	0.40
0.75	0.44	0.41	0.36	0.31	0.27	0.26	0.25	0.25
1.00	0.27	0.25	0.22	0.20	0.18	0.17	0.17	0.16
1.50	0.100	0.096	0.094	0.092	0.089	0.089	0.088	0.088
2.00	0.036	0.040	0.044	0.048	0.050	0.050	0.050	0.050
3.00	0.0049	0.008	0.011	0.015	0.018	0.019	0.019	0.020
4.00	0.00067	0.002	0.003	0.005	0.007	0.008	0.008	0.008
5.00	0.00009	0.00044	0.0010	0.0020	0.0030	0.0035	0.004	0.004
6.00	0.000012	0.00011	0.00035	0.0008	0.0014	0.0016	0.002	0.002
8.00	0.00000	0.000007	0.00004	0.00014	0.0003	0.0004	0.0005	0.0005

Part II: approximate values of $\theta^{-1} B^c(t)$					
time	values of the traffic intensity				
t	$\rho = 0.10$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.95$
0.10	1.64	1.64	1.64	1.55	1.65
0.25	1.14	1.06	0.96	0.87	0.81
0.50	0.68	0.62	0.54	0.46	0.41
0.75	0.41	0.38	0.33	0.29	0.25
1.00	0.25	0.23	0.21	0.19	0.17
1.50	0.097	0.095	0.092	0.089	0.088
2.00	0.039	0.041	0.045	0.048	0.050
3.00	0.0075	0.010	0.014	0.017	0.019
4.00	0.0021	0.0034	0.0053	0.0070	0.0082
5.00	0.00078	0.0014	0.0024	0.0032	0.0038
6.00	0.00035	0.00068	0.0011	0.0015	0.0018
8.00	0.000087	0.00017	0.00029	0.00039	0.00047

TABLE 2. *Part I: numerical values of the doubly-scaled M/M/1 complementary busy-period cdf, $\theta^{-1} B^c(t)$, obtained by multiplying the values in Table 1 by $\theta^{-1} = 2/(1-\rho)$;*
Part II: light-and-heavy-traffic interpolation $(1-w)2e^{-2t} + w\tilde{h}_1(t)$ in (2.25) with $w(\rho) = \rho^x$ for $x = 0.75$.

type of approximation	$B^c(t) \approx pe^{-\lambda_1 t} + (1-p)e^{-\lambda_2 t}$		
	p	λ_1^{-1}	λ_2^{-1}
One-moment exponential fit	0	—	$\theta = \frac{1-\rho}{2}$
Two-moment bounding H_2 (atom at 0 plus exponential)	$(1-2\theta) = \rho$	∞ $(\lambda_1 = 0)$	$\frac{m_1(1+c^2)}{2} = \frac{1}{2}$
Three derivatives $g''(0), g'(0), g(0)$	$\frac{1-\sqrt{\rho/(\rho+4)}}{2}$	$\frac{2\theta^2 p}{1-p}$	$\frac{2\theta^2(1-p)}{p}$
One moment and two derivatives $g'(0), g(0), m_1$	$\frac{1}{2}$	$\theta(1-\sqrt{\rho})$	$\theta(1+\sqrt{\rho})$
Two moments and a derivative $g(0), m_1, m_2$	$\frac{1+\sqrt{\rho/(4-3\rho)}}{2}$	$\frac{\theta(1-p)}{p}$	$\frac{\theta p}{1-p}$
Three moments m_1, m_2, m_3	λ_2^{-1}	$\frac{1-\sqrt{\rho}}{2}$	$\frac{1+\sqrt{\rho}}{2}$

TABLE 3. *Hyperexponential (H_2) approximations for the time-scaled M/M/1 complementary busy-period cdf $B^c(t)$.*

time t	exact transform inversion	H_2 Fits: three moments	H_2 Fits: two moments and one deriv.	HEYMAN's [26] diffusion approx.	approximants from Widder's formula (5.1): TER HAAR [45]	approximants from Widder's formula (5.1): JAGERMAN [30]
.09	.317	.38	.24	.30	.45	.44
.18	.188	.17	.15	.23	.32	.31
.27	.132	.095	.13	.19	.26	.24
.36	.0994	.066	.11	.17	.22	.19
.45	.0782	.054	.094	.15	.19	.16
.90	.0321	.031	.045	.12	.12	.08
1.00	.0273	.027	.038		.11	.07
1.25	.0189	.021	.026		.09	.05
1.50	.0135	.016	.017		.08	.04
2.00	.0074	.0093	.0075		.06	.02
3.00	.0026	.0031	.0015		.04	.01
4.00	.0010	.0011	.0003		.03	.004
5.00	.0004	.0004	.0001			.002
6.00	.0002	.0001	.0000			.001

TABLE 4. *A comparison of five approximations of the time-scaled M/M/1 complementary busy-period cdf* $B^c(t)$ *with exact values obtained from Laplace transform inversion: the case of* $\rho = 0.70$.

time t	exact transform inversion	H_2 fit one moment, two derivatives	inverse Gaussian (6.6), $\nu = 1-\theta$	asymptotic normal (4.3)
$\theta^2/4 = 0.035$	.884	.884	.98	3.2
.05	.840	.840	.94	2.5
.10	.712	.712	.77	1.5
.25	.450	.452	.435	.64
.50	.234	.240	.216	.28
.75	.135	.141	.125	.15
1.00	.0827	.0869	.0785	.090
1.50	.0352	.0349	.0350	.0372
2.00	.0165	.014	.0172	.0172
3.00	.00424	.0024	.00486	.00436
4.00	.00123	.0004	.00154	.00125
6.00	.000125	.00005	.000187	.000125

TABLE 5. *A comparison of approximations of the time-scaled M/M/1 complementary busy-period cdf $B^c(t)$ with exact values obtained by Laplace transform inversion: the case of $\rho = 0.25$.*

time t	exact transform inversion	H_2 fit one moment, two derivatives	inverse Gaussian (6.6), $\nu = 1-\theta$	asymptotic normal (4.3)
$\theta^2/4 = 0.0039$	.888	.888	.96	1.8
.05	.378	.43	.372	.40
.10	.239	.33	.236	.25
.25	.114	.17	.113	.116
.50	.0563	.059	.0560	.0568
.75	.0340	.020	.0338	.0342
1.0	.0224	.0069	.0224	.0225
1.5	.0112	.0008	.0112	.0112
2.0	.00619		.00620	.00620
3.0	.00221		.00222	.00222
4.0	.000887		.000895	.000889
6.0	.000171		.000173	.000171

TABLE 6. *Same comparison as Table 5: the case of $\rho = 0.75$.*

time t	$\rho = 0.25$			$\rho = 0.95$			$\rho = 0.99$		
	exact transform inversion	three-moment H_2 fit	relative percent error	exact transform inversion	three-moment H_2 fit	relative percent error	exact transform inversion	three-moment H_2 fit	relative percent error
0.10	0.712	0.722	+ 1.4	0.0425	0.0118	—72.1	0.0083	0.0023	—72.3
0.25	0.450	0.455	+ 1.1	0.0203	0.0098	—50.0	0.0040	0.0020	—50.0
0.50	0.234	0.230	— 1.7	0.0102	0.0076	—20.0	0.0020	0.0015	—25.0
0.75	0.135	0.129	— 4.5	0.0063	0.0059	— 6.3	0.0012	0.0012	— 4.8
1.00	0.0827	0.0796	— 3.7	0.0042	0.0046	+ 9.5	0.0008	0.00092	+ 9.5
1.50	0.0352	0.0357	+ 1.4	0.00219	0.0028	+27.3	0.00044	0.00056	+27.3
2.00	0.0165	0.0176	+ 6.7	0.00129	0.0017	+30.8	0.00025	0.00034	+36.0
3.00	0.0042	0.0046	+ 8.0	0.00048	0.00061	+27.1	0.00010	0.00012	+27.8
4.00	0.00123	0.00121	— 1.6	0.00021	0.00022	+ 4.8	0.00004	0.000045	+ 7.1
5.00	0.00038	0.00032	—16.8	0.00009	0.00008	—14.9	0.00002	0.000017	—15.0
6.00	0.00013	0.000084	—38.5	0.00005	0.00002	—35.6	0.00001	0.000006	—35.8
8.00	0.000014	0.0000058	—57.2	0.000011	0.000004	—63.7	0.000002	0.0000008	—66.7

TABLE 7. *A comparison of the three-moment H_2 approximation of the time-scaled M/M/1 complementary busy-period cdf $B^c(t)$ with exact values obtained from Laplace transform inversion: the cases of $\rho = 0.25$, 0.95 and 0.99.*

time t	$\rho = .25$		$\rho = .50$		$\rho = .75$	
	(2D, 1M)	(3D)	(2D, 1M)	(3D)	(2D, 1M)	(3D)
.5 θ^2	+0.0	—0.0	+ 0.0	—0.0	+ 0.1	—0.0
1.0 θ^2	+0.1	—0.0	+ 0.1	—0.0	+ 0.9	—0.1
1.5 θ^2	+0.4	—0.1	+ 1.1	—0.2	+ 2.7	—0.4
2.0 θ^2	+0.7	—0.2	+ 2.4	—0.7	+ 5.1	—1.2
2.5 θ^2	+1.2	—0.4	+ 4.1	—1.5	+ 8.3	—2.5
3.0 θ^2	+1.8	—1.1	+ 5.9	—2.4	+12.0	—4.3
4.0 θ^2	+3.0	—2.5	+10.1	—6.3	+19.8	—9.4

TABLE 8. *The small-time performance of two hyperexponential fits: two derivatives and one moment (2D, 1M) and three derivatives (3D). The values displayed are relative percent error, i.e., 100× (Approx. − Exact)/Exact.*

time y	mixing density $w(y)$		
	in $B^c(t)$ (2.27)	in $H_1^c(t)$ [3]	in $c_q(t)$ [5]
$\tau_1 = 0.009$	0.00	0.00	0.00
0.010	22.1	1.77	0.04
$y_{max} = 0.012$	26.6	2.55	0.06
0.015	24.0	2.88	0.09
0.030	11.2	2.69	0.16
0.100	2.05	1.64	0.33
0.50	0.17	0.68	0.68
1.00	0.05	0.40	0.80
1.50	0.01	0.12	3.60
$\tau_2 = 1.74$	0.00	0.00	0.0

TABLE 9. *Values of the mixing density $w(y)$ in (2.8) in the spectral representation of the busy-period complementary cdf $B^c(t)$ for the case $\rho = 0.75$. Also given are the corresponding mixing densities for the first-moment function complementary cdf $H_1^c(t)$ and the correlation function $c_q(t)$.*

time t	inverse Gaussian approximations (6.6) and (6.8)			asymptotic normal (4.3)	standard asymptotic expansion (3 terms) (4.1)
	$\nu = 1$	$\nu = 1-\theta$	$\nu = \tau/2$		
0.016	+6.2	+9.6	+9.8	+149.	
0.05	—4.3	+5.0	+6.0	+51.	
0.10	—10.7	—0.8	+0.3	+24.	
0.25	—11.2	—3.7	—2.7	+8.2	
0.50	—5.1	—2.3	—2.7	+4.3	
0.75	+1.0	—2.1	—2.5	+3.0	
1.00	+7.2	—1.4	—2.5	+2.3	
1.50	+21.	+0.0	—2.2	+1.7	
2.00	+34.	+1.7	—2.2	+1.3	+252.
3.00	+63.	+3.4	—2.1	+0.8	+84.
4.00		+5.2	—2.0	+0.7	+39.
5.00		+8.0	—1.9	+0.6	+21.

TABLE 10. *A comparison of five approximations of the time-scaled M/M/1 complementary busy-period cdf $B^c(t)$ for the case $\rho = 0.50$. The displayed values are relative percent error, 100 (Approx. − Exact)/Exact.*

traffic intensity ρ	inverse Gaussian ($\nu = 1-\theta$)	asymptotic normal
0.25	1.4 - 1.6	5.5 - ∞
0.50	1.1 - 1.9	2.5 - ∞
0.75	0.18 - 4.4	0.45 - ∞
0.85	0.07 - 12	0.11 - ∞
0.95	0.01 - ∞	0.01 - ∞

TABLE 11. *Time range where the relative percent error is less than one percent for the asymptotic normal approximation in (4.3) and the inverse Gaussian approximation in (6.6) with $x = \theta$ and $\nu = 1-\theta$.*

time t	traffic intensity							
	$\rho = .25$		$\rho = .50$		$\rho = .75$		$\rho = .85$	
	3D	IG	3D	IG	3D	IG	3D	IG
1 θ^2	—0.0		—0.0		—0.1		—0.1	
2 θ^2	—0.2		—0.7		—1.2	—0.8	—1.5	—0.7
3 θ^2	—1.1		—2.4	—2.4		—1.2		—1.0
4 θ^2	—2.5			—2.7		—1.2		—1.0
5 θ^2	—4.7			—2.8		—1.2		—1.0
6 θ^2	—8.1	—8.3		—2.8		—1.1		—0.8
1.0		—8.5		—2.5		—0.4		—0.2
1.5		—8.5		—2.2		—0.4		—0.2
2.0		—8.5		—2.2		—0.4		—0.1
3.0		—8.7		—2.1		—0.4		—0.1
4.0		—8.6		—2.0		—0.3		—0.1
5.0		—8.4		—1.9		—0.3		—0.1

TABLE 12. *Relative percent errors, 100 (Approx. − Exact)/Exact, for the composite IG-3D approximation in Section 7 and its inverse Gaussian (IG) and three-derivative (3D) H_2 approximation components.*

time t	busy-period density $b(t)$	IG density $g(t)$	relative percent error RE in (6.18)
0.00	128.0	0.00	
0.01	19.0	25.0	
0.02	12.1	13.8	
0.03	8.28	8.65	
0.04	5.98	6.02	+0.8
0.05	4.52	4.48	—0.8
0.10	1.71	1.68	—1.7
0.50	0.1295	0.1286	—0.7
1.00	0.0346	0.0344	—0.5
2.00	0.00691	0.00689	—0.4
4.00	0.000776	0.000773	—0.4
6.00	0.0001340	0.0001336	—0.3
8.00	0.0000276	0.0000275	—0.3

TABLE 13. *A comparison between the inverse Gaussian (IG) density $g(t)$ using (iii) in (6.8) and the M/M/1 busy-period density $b(t)$ in the case $\rho = 0.75$.*

time t	busy-period density $b(t)$	IG density $g(t)$	relative percent error RE in (6.18)
0.00	8.000	0.000	
0.01	7.106	2.243	
0.02	6.333	6.726	
0.03	5.662	7.432	
0.04	5.078	6.852	
0.06	4.123	5.261	
0.08	3.388	4.030	
0.10	2.816	3.161	+12.4
0.20	1.301	1.295	—0.4
0.40	0.458	0.444	—3.0
0.80	0.225	0.219	—2.9
1.0	0.0816	0.0794	—2.6
2.0	0.0148	0.0144	—2.3
3.0	0.00407	0.00399	—2.1
4.0	0.00134	0.00131	—2.0
5.0	0.000482	0.000473	—2.0
6.0	0.000185	0.000181	—2.0
7.0	0.0000739	0.0000724	—2.0
8.0	0.0000305	0.0000299	—1.9
9.0	0.0000129	0.0000126	—1.9

TABLE 14. *A comparison between the inverse Gaussian density $g(t)$ using (iii) in (6.8) and the M/M/1 busy-period density $b(t)$ in the case $\rho = 0.50$.*

traffic intensity ρ	asymptotic relative percent error ARE in (6.19)
0.10	—23.6
0.15	—15.9
0.20	—11.3
0.25	—8.2
0.30	—6.0
0.35	—4.48
0.40	—3.33
0.45	—2.47
0.50	—1.82
0.55	—1.32
0.60	—0.947
0.65	—0.660
0.70	—0.443
0.75	—0.283
0.80	—0.167
0.85	—0.087
0.90	—0.0359
0.95	—0.0084

TABLE 15. *The asymptotic relative percent error for the inverse Gaussian (IG) density approximation, using (iii) in (6.8), of the M/M/1 busy-period density.*