Predicting Response Times in Processor-Sharing Queues

Amy R. Ward
Department of Management Science and Engineering
Stanford University
Stanford, CA 94305-4022

Ward Whitt
AT&T Labs
Shannon Laboratory
180 Park Avenue
Florham Park, NJ 07932-0971

(To appear in *Proceedings of the Fields Institute Conference on Communication Networks*, P. W. Glynn, D. J. MacDonald and S. J. Turner (eds.), 2000)

1991 *Mathematics Subject Classification.* Primary 54C40, 14E20;
Secondary 46E25, 20C20.

This work was done at AT&T Labs as part of the AT&T Labs Fellowship Program.

Abstract. We investigate the possibility of reliably predicting response times in real time (e.g., at the time a job arrives) in the $M/G/1$ processor-sharing queue. The proposed prediction is the conditional mean given current state information. We consider several forms of state information, always including the remaining service requirement of the job of interest and the number of other jobs in the system. We consider three cases for the other jobs' service requirements: First, we assume that all the remaining service requirements are known; second, we assume that the amount of completed work of each customer in service is known; and third, we assume that nothing more is known. We thus are able to study the value of different kinds of information. We calculate the conditional mean and variance of the response time, given the state information, by numerically inverting Laplace transforms. We evaluate the reliability by looking at the ratio of the standard deviation to the mean. This ratio tends to decrease as the remaining service requirements or the number of jobs in service increase. We establish this property theoretically by proving laws of large numbers and central limit theorem refinements.

Contents

Part 1. This is a Part Title Sample 1

Chapter 1. Predicting Response Times in Processor-Sharing Queues 3

1.1. Introduction 3

1.2. Basic M/G/1/PS Theory 6

1.3. Stochastic Comparisons for Mean Response Times 11

1.4. Reliable Prediction When the Initial Work is Large 14

1.5. Only Conditioning on the Number of Jobs 18

1.6. Conditioning Upon Completed Work 22

1.7. Numerical Inversion Issues 26

1.8. Conclusions 29

Bibliography 31

Part 1

This is a Part Title Sample

CHAPTER 1

Predicting Response Times in Processor-Sharing Queues

1.1 Introduction

In this paper we investigate the possibility of reliably predicting remaining response times in a processor-sharing (PS) queue by exploiting system state information. The response time is the time until the job can complete service and depart. With the PS discipline, each job receives service at rate $1/n$ when there are n jobs in service. We consider a PS queue because it is often a good model for computer and communication systems. The PS service discipline is a good approximation for round robin (RR) disciplines used in computer operating systems. The PS discipline may also be a reasonable approximation for head-of-the-line (per-flow) fair-queueing service disciplines used in communication network routers when the number of packets from each flow in the router is suitably small; e.g. see Demers, Keshav and Shenker [**8**] and Parekh and Gallager [**22**]. We hope to be able to improve customer satisfaction by informing customers upon job arrival and thereafter about anticipated response times; see Hui and Tse [**15**], Katz, Larson and Larson [**17**], Taylor [**25**] and references therein. Predicted response times might also be used to help system managers determine when to add resources to increase the processing rate and thereby reduce congestion, again improving customer satisfaction.

The system state information used in the prediction always includes the service requirement of the job of interest (e.g., a new arrival), the number of other jobs in service at that time and the model for future arrivals; i.e., the arrival rate and service-requirement distribution in an $M/G/1/PS$ model. Since the service requirement of a new job may be unknown, we aim to describe the expected conditional response time as a function of the service-requirement. If the new arrival is to be informed, then the new arrival might be told the conditional expected response time as a function of the service requirement of that job. The new arrival might also be given some idea about the variability of this estimate, again as a function of the service requirement of the new job.

We consider three cases for the service requirements of other jobs: First, we assume that the remaining service requirements of all jobs are known; second, we assume that the amount of completed work for each job in service is known; and third, we assume that nothing is known beyond the number of other jobs. In the last two cases, we use the distributions of the remaining service.

The main question we ask is: Is reliable prediction possible? By reliable prediction, we mean two things: first, that the expected conditional response time given the system state information can be computed and, second, that the variability of the conditional response-time distribution is suitably small, i.e., the conditional response-time distribution clusters sufficiently closely about the mean, so that a point estimate such as the conditional mean is reasonably reliable. We focus on this second issue by calculating the ratio of the standard deviation to the mean.

Fortunately, useful expressions have been derived for the conditional distribution of response times given state information in $M/G/1/PS$ queues, but these expressions are quite complicated. Even the desired expected values are available only via Laplace transforms. Hence, a key component of our analysis is the numerical inversion of Laplace transforms, drawing on Abate and Whitt [**4**], [**5**]; see Abate, Choudhury and Whitt [**1**] for a review. Numerical transform inversion was previously applied to PS queues by Braband [**6**], but not for the purpose of real-time prediction. We show that the conditional mean can indeed be very efficiently computed by numerical transform inversion. We show that the standard deviation can also be efficiently computed and that the ratio of the standard deviation to the mean tends to be reasonably small, especially as the conditional mean grows. We also investigate how this ratio depends upon different information, revealing the value of different kinds of information.

This paper is a sequel to Whitt [**26**], [**27**], [**28**], which recently investigated the possibility and advantages of making reliable delay predictions in a queue with the first-come first-served (FCFS) service discipline. Prediction with the PS discipline is more difficult than with the FCFS discipline because, after conditioning on all available state information, the remaining delay may be fully known with the FCFS discipline, whereas with the PS discipline the remaining response times depend on uncertain future arrivals and their uncertain service requirements. Just as in the FCFS setting, with the PS discipline there tend to be two situations: first, when the conditional mean is smaller and, second, when the conditional mean is larger. When the conditional mean is smaller, the variability (as measured by the ratio of the standard deviation to the mean) tends to be larger, but that variability tends not to matter so much, since the response time is not long. In contrast, when the conditional mean is larger, the variability tends to be much smaller (relatively, as measured by the ratio of the standard deviation to the mean), so that when jobs face long response times reliable prediction is possible.

We emphasize that reliable prediction depends critically upon exploiting system state information. The steady-state response time distribution without conditioning tends to be approximately exponential, which we regard as not supporting reliable prediction. For example, the unconditional steady-state response-time (either conditioning or not conditioning upon the service requirement of the arrival) in an $M/G/1/PS$ queue is known to be asymptotically exponential in heavy traffic, i.e., as $\rho \to 1$ where ρ is the traffic intensity. In contrast, with conditioning, the conditional response-time distribution when the conditional mean is large (due to substantial work initially in the system) tends to be approximately normally distributed with a smaller ratio of standard deviation to mean. To illustrate this point, we consider a numerical example.

Example 1.1 `[exN11]` Consider the $M/M/1/PS$ queue in which the service-requirement cumulative distribution function (cdf) is exponential with mean 1. In Figure 1 we display the ratio of the standard deviation (SD) to the mean (M) for the steady-state conditional response time as a function of the arrival rate, which here coincides with the traffic intensity ρ. We assume that the arriving job has service requirement 2. The mean and variance of the steady-state response time in the $M/M/1/PS$ model were determined by Coffman, Muntz and Trotter [**7**]. Figure 1 shows that the ratio SD/M for the steady-state response time decreases to 1 as $\rho \to 1$, and stops there because it becomes infinite for $\rho \geq 1$. By this measure of variability, the steady-state distribution is always at least as variable as an exponential distribution. Consistent with the heavy-traffic limit to an exponential distribution, the ratio approaches 1 as $\rho \to 1$.

In Figure 1 we also display SD/M ratios when we condition on there being $n - 1$ other jobs each with remaining service requirement 1. We display the ratio SD/M as a function of ρ for the cases $n = 1$ and $n = 20$. The curves decrease as n increases, so that the worst case is $n = 1$, corresponding to no other jobs being in the system. Consistent with intuition, the ratio SD/M in the case $n = 1$ increases with ρ, but it remains less than 1 for all positive ρ. Indeed, in equation (1.41) below we prove in general (for any x) that the ratio SD/M is always less than 1 for $n = 1$. Moreover, as n increases the curve decreases, showing that the reliability of prediction improves significantly as n increases. In Theorem 1.9, we show that the ratio decreases by the factor $1/\sqrt{n}$ as n increases.

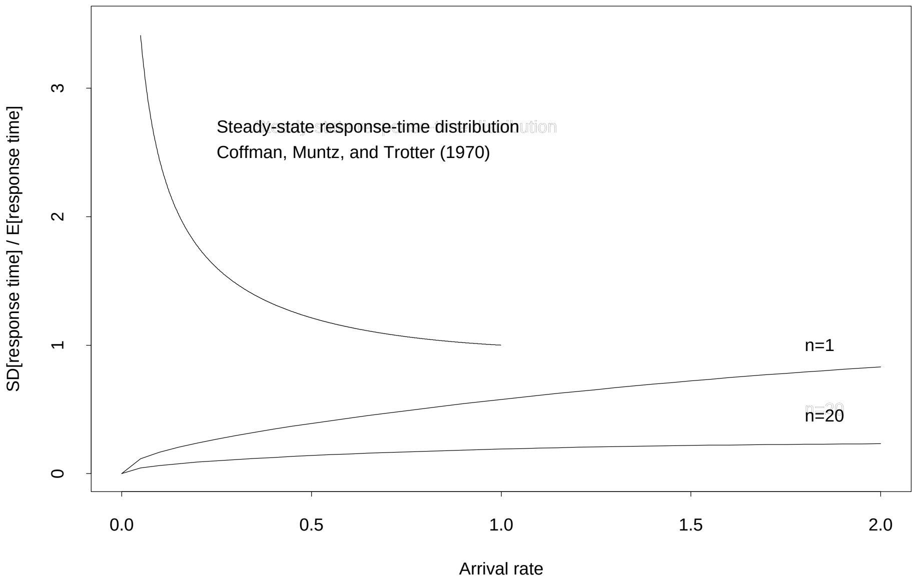

Figure 1 The ratio of the standard deviation to the mean for the response time as a function of the arrival rate: A comparison of the steady-state and conditional response-time distributions.

Here is how the rest of this paper is organized: In Section 2 we review the basic theory associated with the $M/G/1/PS$ queue. In particular, we review the Laplace transforms we use for numerical inversion and our other theoretical results. In Section 3 we make stochastic comparisons showing how the conditional mean response given the remaining service requirements of all jobs in service depends on the service-requirement cdf of future arrivals. We show that the conditional mean increases when the service-requirement cdf gets *less* variable in convex stochastic order. Consequently, the case of deterministic service times serves as an *upper* bound for any service-requirement cdf with the same mean. It is known that the unconditional steady-state mean is insensitive to the service-requirement cdf beyond its mean. Hence, it can be anticipated that the effect above is counteracted by a stochastic increase in the remaining service requirements of all jobs in service when the service-requirement cdf gets more variable, which we also establish.

In Section 4 we prove that the conditional response-time distribution given all remaining service requirements of jobs in service satisfies laws of large numbers and central limit theorems as the service requirement of the designated job increases or the number of other jobs increases. Our asymptotic results here are consistent with (and are largely contained in) previous asymptotic results by Grishechkin [**11**], [**12**], [**13**]. These asymptotic results show that the distribution clusters closely about the mean and tends to be normally distributed when the conditional mean is large.

In Sections 5 and 6 we consider the other two cases in which, first, we condition only upon the number of other jobs and, second, we condition upon the amounts of completed work. We show how to compute the mean and variance of the conditional response time by numerical transform inversion in these cases as well. In Section 7 we discuss technical issues arising in the numerical inversion. Finally, in Section 8 we draw conclusions.

1.2 Basic M/G/1/PS Theory

In this section we review the basic $M/G/1/PS$ theory. This theory was first developed by Kitaev and Yashkov [**20**] and Yashkov [**29**]; see Yashkov [**30**], [**31**]. A convenient reference is Ott [**21**], which contains extensions and complements. (See Grishechkin [**14**], Kitaev [**19**] and Zwart and Boxma [**32**] for other recent work on PS queues.) The $M/G/1/PS$ queue has one server, unlimited waiting space, the processor-sharing (PS) service discipline, a Poisson arrival process with rate λ and independent and identically distributed (iid) service requirements having a cdf G with k^{th} moment m_k. Throughout we assume that $m_1 < \infty$. We will sometimes also assume that $m_2 < \infty$ or $m_3 < \infty$ as well. Let g be the associated service-requirement probability density function (pdf), $G^c(t) \equiv 1 - G(t)$ the associated complementary cdf (ccdf), and G_e the associated stationary-excess (or equilibrium residual-life) cdf, i.e.,

$$G_e(t) = \frac{1}{m_1} \int_0^t G^c(u)du, \quad t \geq 0 . \tag{1.1}$$

We use the assumption that $m_1 < \infty$ to ensure that G_e is a proper cdf. The k^{th} moment of G_e is $m_{ek} = m_{k+1}/(k+1)m_1$. Thus, if $m_2 < \infty$, then $m_{e1} < \infty$, but in general we do not require it.

For the service-requirement pdf g (and similarly for any other pdf), let $\hat{g}$ be the associated Laplace transform (LT) of g and the Laplace-Stieltjes transform (LST) of the associated cdf G, i.e.,

$$\hat{g}(s) = \int_0^\infty e^{-st} g(t)dt = \int_0^\infty e^{-st} dG(t) , \tag{1.2}$$

with only the second definition applicable if the cdf G does not have a pdf.

With the PS discipline, the workload (unfinished work) process is the same as for the first-come first-served discipline. Let W be the steady-state workload cdf and $\hat{w}$ its LST, which is given by the classical Pollaczek-Khintchine formula

$$\hat{w}(s) = \int_0^\infty e^{-st} dW(t) = \frac{1-\rho}{1-\rho\hat{g}_e(s)} , \tag{1.3}$$

where $\hat{g}_e(s)$ is the LST of the stationary-excess cdf G_e in (1.1), which requires that $\rho \equiv \lambda m_1 < 1$. Even though we are not concerned with steady-state behavior, the LST $\hat{w}$ plays a role.

In fact, the key quantity related to W is the nondecreasing function (corresponding to a non-probability measure)

$$R(t) = \frac{W(t)}{1-\rho} = 1 + \sum_{n=1}^{\infty} \rho^n G_e^{n*}(t) \ , \tag{1.4}$$

where $G_e^{n*}(t)$ is the n-fold convolution of the cdf G_e in (1.1), with LST

$$\hat{r}(s) = \frac{\hat{w}(s)}{1-\rho} = \frac{1}{1-\rho\hat{g}_e(s)} = \sum_{n=0}^{\infty} \rho^n \hat{g}_e(s)^n \ . \tag{1.5}$$

Unlike W, $R(t)$ is well defined for all $\rho > 0$ and all $t > 0$. For $\rho \geq 1$, $R(t) \to \infty$ as $t \to \infty$, but that will not pose a problem. We establish the key supporting property here.

Lemma 1.2 `[le21]` *For all $\rho > 0$ and $t > 0$, and for all service-time cdf's G, there is an $n_0 \equiv n_0(G, \rho, t)$ such that $G_e^{n*}(t) < \rho^{-n}$ for $n \geq n_0$, so that $R(t) < \infty$ for R in (1.4).*

Proof We need to bound the probability $P(S_n \leq t)$ where S_n is the sum of n i.i.d. random variables distributed as G_e. We bound this probability above by a probability $P(S_n' \leq t)$ where S_n' is the sum of n i.i.d. scaled Bernoulli random variables each stochastically smaller than G_e. In particular, we choose ϵ and put $1 - p \equiv G_e(\epsilon)$ probability at 0 and $p \equiv 1 - G_e(\epsilon)$ probability at ϵ. For ρ given, where without loss of generality we assume that $\rho > 1$, choose ϵ sufficiently small so that $1 - p \equiv G_e(\epsilon) < \rho^{-1}$. This choice is always possible because $G_e(t)$ has the pdf $m^{-1}G^c(t)$, i.e., $G_e(t) \to 0$ as $t \to 0$. With the choices made so far,

$$P(S_n \leq t) \leq P(S_n' \leq t) \leq P(S_n'' \leq k)$$

for k an integer satisfying $k > t/\epsilon$, where S_n'' is a standard binomial random variable with parameters n and p. Thus, for all n sufficiently large (in the last step),

$$P(S_n'' \leq k) = \sum_{j=0}^{k} \binom{n}{j} (1-p)^{n-j} p^j \leq (k+1) \binom{n}{k} (1-p)^{n-k} < \rho^{-n} \ .$$

□

We are concerned with the response time $T(x)$ of a new arrival with service requirement x, but we want to consider $T(x)$ conditional on various types of state information. We will initially condition on the number n of jobs that are in service when this new arrival joins and the remaining service requirements of the $n-1$ other jobs in service. Let x_j denote the remaining service requirements of the j^{th} job already in service, $1 \leq j \leq n-1$. Without loss of generality, let the remaining service requirements be ordered, so that $x_0 \equiv 0 < x_1 < \cdots < x_{n-1}$. Let $T_0(x)$ denote the conditional response time given that there are no more arrivals. In a prediction, we may want to announce $T_0(x)$ as the minimum possible response time given current state information. It may be useful in addition to the conditional mean and other summary statistics such as the conditional standard deviation. Given x_j for all j, we can calculate $T_0(x)$. (It is deterministic.) In particular, given x_j, $1 \leq j \leq n-1$,

$$T_0(x) = \sum_{j=1}^{k} (n-j+1)(x_j - x_{j-1}) + (n-k)(x - x_k) \ , \quad x_k \leq x < x_{k+1} \ . \qquad (1.6)$$

It is important to note that the conditional distribution of $T(x)$ has an atom (positive probability mass) at $T_0(x)$, which corresponds to the event of no arrivals in the interval $[0, T_0(x)]$; i.e.,

$$P(T(x) = T_0(x) | x, n, x_j, 1 \leq j \leq n-1) = e^{-\lambda T_0(x)} \ . \qquad (1.7)$$

However, it is easy to see that, aside from the atom at $T_0(x)$, the conditional distribution of $T(x)$ has a pdf. This follows from the fact that the probability of m new arrivals in $[0, T_0(x)]$ has a Poisson probability and, conditional on there being m arrivals, their locations in $[0, T_0(x)]$ are distributed as m i.i.d. uniform random variables. These uniform distributions smooth out the conditional distribution of $T(x)$ giving it a pdf.

For the following discussion, let the remaining times $x_1, \ldots, x_{n-1}$ be unordered and let $x_n = x$. Let $T_n(x) \equiv [T(x)|n, x_1, \ldots, x_n]$ be a random variable with the conditional cdf. It is significant that the effects of the n different jobs initially in the system can be decomposed into separate independent components, i.e.,

$$T_n(x) \equiv [T(x)|n, x_1, \ldots, x_n] \overset{\mathrm{d}}{=} \sum_{j=1}^{n} T_j(x_j, x) \ , \qquad (1.8)$$

where $\overset{\mathrm{d}}{=}$ denotes equal in distribution and the n random variables $T_j(x_j, x)$ are mutually independent, distributed as a random variable $T(x_j \wedge x, x)$, with $x_j \wedge x = \min\{x_j, x\}$,

$$T(x, x) \overset{\mathrm{d}}{=} T_1(x) \equiv [T(x)|1, x] \qquad (1.9)$$

and, for $x_j < x$,

$$T_1(x) \overset{\mathrm{d}}{=} T(x_j, x) + T_1(x - x_j) \ , \qquad (1.10)$$

where the two random variables on the right are independent; see Ott [**21**].

We now give an intuitive explanation of formulas (1.8)–(1.10). The idea is to separate the effects of the n customers initially in the system. For this purpose, it is helpful to think of the PS discipline as the limit of the round robin (RR) discipline as the allocated quantum of service per round becomes negligibly small. Then the customers in the system are being served separately. Also, because of the M/G/1 model assumptions, the input during different round robin quanta are independent and identically distributed. Then $T(x_j, x)$ is the delay $x_j \wedge x$ imposed upon the designated customer with service requirement x by customer j plus the additional delays imposed upon the designated customer by arrivals when customer j is in service (before receiving $x_j \wedge x$ service) or when these new arrivals themselves are in service. A new arrival with service requirement S arriving when customer j has completed u units of service, $0 < u < x_j \wedge x$, will impose an additional $S \wedge (x - u)$ of delay upon the designated customer.

Thus equation (1.10) is intuitively clear: we can divide the time to reduce the service requirement of a single job by x into the time $T(x_j, x)$ required to achieve the first x_j reduction, including delays caused by all new arrivals during that period, and then adding the remaining time $T_1(x - x_j)$. Note that $T(x_j, x) = T(x_j \wedge x, x)$.

With this interpretation, it is immediate that $T(x,y)$ is stochastically increasing in y and, as $y \to \infty$, $T(x,y)$ approaches the first passage time $T_{x,0}$ for the $M/G/1$ workload to go from x to 0, which has Laplace transform

$$\hat{f}_{x_0}(s) \equiv Ee^{-sT_{x,0}} = e^{-x\zeta(s)} \ , \tag{1.11}$$

where

$$\zeta(s) = s + \lambda - \lambda\hat{b}(s) \ , \tag{1.12}$$

with $\hat{b}(s)$ being the busy-period transform, which satisfies the Kendall functional equation

$$\hat{b}(s) = \hat{g}(s + \lambda - \lambda\hat{b}(s)) \ ; \tag{1.13}$$

e.g., see Section 4 of Abate and Whitt [**3**].

To specify the Laplace transforms, let

$$\hat{t}_1(s|x) = Ee^{-sT_1(x)} \ . \tag{1.14}$$

From (1.10), we have

$$\hat{t}(s|x_j, x) \equiv E[e^{-sT(x_j,x)}] = \frac{\hat{t}_1(s|x)}{\hat{t}_1(s|x - x_j)} \quad \text{for} \quad x_j \le x \ . \tag{1.15}$$

From (1.8) and (1.15), we obtain

$$\begin{aligned} \texttt{[T6]}\hat{t}_n(s|x) &\equiv E[e^{-sT_n(x)}] \equiv E[e^{-sT(x)}|n, x_1, \ldots, x_n = x] \\ &= \prod_{j=1}^{n} \hat{t}(s|x_j \wedge x, x) = \prod_{j=1}^{n} \left(\frac{\hat{t}_1(s|x)}{\hat{t}_1(s|x - [x_j \wedge x])} \right) \ . \end{aligned} \tag{1.16}$$

Hence, to specify the conditional distribution of $T_n(x)$ and its Laplace transform $\hat{t}_n(s|x)$, it suffices to specify the distribution of $T_1(x)$ which in turn is characterized by its Laplace transform $\hat{t}_1(s|x)$ in (1.14). As in (1.16) of Ott [**21**], $\hat{t}_1(s|x)$ is characterized by the differential equation

$$\frac{\partial}{\partial x} \log \hat{t}_1(s|x) = -(s + \lambda) + \lambda G^c(x)\hat{t}_1(s|x) + \lambda \int_0^x \frac{\hat{t}_1(s,x)}{\hat{t}_1(s, x - y)} dG(y) \ . \tag{1.17}$$

We now turn to means and variances, again following Ott [**21**]. First,

$$ET_1(x) = B^{(1)}(x) \equiv \int_0^x R(u)du \tag{1.18}$$

for R in (1.4) and

$$Var\, T_1(x) = B^{(1)}(x)^2 - 2B^{(2)}(x) \ , \tag{1.19}$$

where

$$B^{(2)}(x) = \int_0^x B^{(1)}(x - u)dB^{(1)}(u) \ . \tag{1.20}$$

Second, by (1.10),

$$\begin{aligned} \texttt{[WC4]}ET(x_j \wedge x, x) &= ET_1(x) - ET_1(x - (x_j \wedge x)) \\ &= B^{(1)}(x) - B^{(1)}(x - (x_j \wedge x)) = \int_{x-(x_j \wedge x)}^{x} R(u)du \end{aligned} \tag{1.21}$$

and

$$\begin{aligned}\texttt{[WC5]}\mathit{Var}\,T(x_j \wedge x, x) &= \mathit{Var}\,T(x) - \mathit{Var}\,T(x - (x_j \wedge x)) \\ &= B^{(1)}(x)^2 - B^{(1)}(x - (x_j \wedge x))^2 \\ &\quad -2B^{(2)}(x) + 2B^{(2)}(x - (x_j \wedge x))\ . \end{aligned} \tag{1.22}$$

We now turn to the mean and variance of $T_n(x)$. Let $m_n(x) \equiv m(x, n, x_1, \ldots, x_{n-1})$ and $v_n(x) \equiv v(x, n, x_1, \ldots, x_{n-1})$ be the conditional mean and variance of $T(x)$, i.e.,

$$m_n(x) \equiv ET_n(x) \equiv E[T(x)|x, n, x_1, \ldots, x_{n-1}]\ . \tag{1.23}$$

By (1.8) and (1.18)–(1.22),

$$m_n(x) = nB^{(1)}(x) - \sum_{j=1}^{n-1} B^{(1)}(x - (x \wedge x_j)) \tag{1.24}$$

and

$$\begin{aligned}\texttt{[WA11]}v_n(x) &= \left[nB^{(1)}(x)^2 - \sum_{j=1}^{n-1} B^{(1)}(x - (x \wedge x_j))^2 \right] \\ &\quad -2 \left[nB^{(2)}(x) - \sum_{j=1}^{n-1} B^{(2)}(x - (x \wedge x_j)) \right]\ , \end{aligned} \tag{1.25}$$

where $B^{(1)}$ and $B^{(2)}$ are given in (1.18) and (1.20). Thus we can calculate the mean directly by numerically inverting its LT

$$\hat{m}_n(s) \equiv \int_0^\infty e^{-sx} m_n(x) dx = n\hat{B}^{(1)}(s) - \sum_{j=1}^{n-1} e^{-s(x \wedge x_j)} \hat{B}^{(1)}(s)\ , \tag{1.26}$$

where

$$\hat{B}^{(1)}(s) = \frac{\hat{r}(s)}{s^2} = \frac{1}{s^2(1 - \rho\hat{g}_e(s))}\ . \tag{1.27}$$

There is some loss of accuracy in doing a single inversion for $m_n(x)$, because $m_n(x)$ has a discontinuous derivative (see Section 7.2), but this usually should not be a serious problem.

The variance in (1.25) requires more steps. We can calculate the second term

$$v_{n2}(x) \equiv -2 \left[nB^{(2)}(x) - \sum_{j=1}^{n-1} B^{(2)}(x - (x \wedge x_j)) \right] \tag{1.28}$$

by numerically inverting its transform

$$\hat{v}_{n2}(s) \equiv \int_0^\infty e^{-st} v_{n2}(t) dt = -2 \left[n\hat{B}^{(2)}(s) - \sum_{j=1}^{n-1} e^{-sx_j} \hat{B}^{(2)}(s) \right]\ , \tag{1.29}$$

where $\hat{B}^{(2)}(s) = \hat{b}^{(2)}(s)/s$ with $\hat{b}^{(2)}(s) = \hat{b}^{(1)}(s)^2 = [sB^{(1)}(s)]^2$, so that $B^{(2)}(s) = sB^{(1)}(s)^2$, but we must calculate $B^{(1)}(x)^2$ and $B^{(1)}(x - x_j)^2$ by separately calculating $B^{(1)}(x)$ and $B^{(1)}(x - x_j)$, $1 \leq j \leq n-1$, by numerically inverting $\hat{B}^{(1)}(s)$ in (1.27). Hence $v_n(x)$ requires $n+1$ inversions, while $m_n(x)$ requires only one. We discuss technical issues involved in carrying out these inversions in Section 7.

Example 1.3 [exN21] If the service-requirement pdf g is exponential with mean 1, then $\hat{g}(s) = \hat{g}_e(s) = (1+s)^{-1}$, $\hat{B}^{(1)}(s) = (1+s)/s^2(1-\rho+s)$ and $\hat{B}^{(2)}(s) = (s^{-1} + 2s^{-2} + s^{-3})(1-\rho+s)^{-2}$, so that

$$\texttt{[OKB1]} m_1(x) = B^{(1)}(x) = \frac{x}{1-\rho} - \frac{\rho}{(1-\rho)^2}[1 - e^{-(1-\rho)x}] \ , \tag{1.30}$$

$$\begin{aligned}\texttt{[OKB2]} B^{(2)}(x) &= \frac{x^2}{2(1-\rho)^2} - \frac{2\rho x}{(1-\rho)^3} + \frac{(2+\rho)\rho}{(1-\rho)^4}(1 - e^{-(1-\rho)x}) \\ &\quad - \frac{\rho^2}{(1-\rho)^3} x e^{-(1-\rho)x} \ ,\end{aligned} \tag{1.31}$$

and

$$\begin{aligned}\texttt{[OKB3]} v_1(x) &= B^{(1)}(x)^2 - 2B^{(2)}(x) = \frac{2\rho x}{(1-\rho)^3} + \frac{2\rho(1+\rho)}{(1-\rho)^3} x e^{-(1-\rho)x} - \frac{\rho(4+\rho)}{(1-\rho)^4} \\ &\quad + \frac{4\rho}{(1-\rho)^4} e^{-(1-\rho)x} + \frac{\rho^2}{(1-\rho)^4} e^{-2(1-\rho)x} \ .\end{aligned} \tag{1.32}$$

As a check on equations (1.30)–(1.32), we note that equations (1.30) and (1.32) agree with equations (33) and (34) on p. 128 of Coffman et al. [**7**] for their case $n = 0$. (To get the mean response time, τ needs to be added to (33).) Using (1.10), (1.30) and (1.32), we can easily compute the moments of $T(x_j, x)$ for $x_j < x$, getting

$$ET(x_j, x) = m_1(x) - m_1(x - x_j) \to \frac{x_j}{1-\rho} \quad \text{as} \quad x \to \infty$$

and

$$Var\, T(x_j, x) = v_1(x) - v_1(x - x_j) \to \frac{2\rho x_j}{(1-\rho)^3} \quad \text{as} \quad x \to \infty \ .$$

1.3 Stochastic Comparisons for Mean Response Times

We now consider the impact of the service-requirement cdf on the expected conditional response times. To make these comparisons we use basic stochastic order relations; e.g., see Chapter 1 of Stoyan [**24**]. We say that one random variable X_1 with cdf G_1 is *stochastically less than or equal to* another random variable X_2 with cdf G_2, and write $X_1 \le_{st} X_2$ or $G_1 \le_{st} G_2$, if $Ef(X_1) \le Ef(X_2)$ for all nondecreasing real-valued functions f for which the expectations are well defined or, equivalently, if $G_1^c(t) \le G_2^c(t)$ for all t. We say that X_1 is less than or equal to X_2 in the *convex order*, and write $X_1 \le_c X_2$ or $G_1 \le_c G_2$, if $Ef(X_1) \le Ef(X_2)$ for all convex real-valued f for which the expectations are well defined. Since $f(x) = x$ and $f(x) = -x$ are both convex, we must have $EX_1 = EX_2$ when $X_1 \le_c X_2$. When $X_1 \le_c X_2$, we have $Var\, X_1 \le Var\, X_2$. The ordering $G_1 \le_c G_2$ holds if and only if

$$\int_x^\infty G_1^c(t)dt \le \int_x^\infty G_2^c(t)dt \quad \text{for all} \quad x \tag{1.33}$$

with

$$m_{11} = \int_0^\infty G_1^c(t)dt = \int_0^\infty G_2^c(t)dt = m_{21} \ ; \tag{1.34}$$

see Section 1.3 of Stoyan [**24**].

From the classical theory (e.g., Section 3.3 of Kelly [**18**]), we know that for any service-requirement cdf G, the steady-state number of jobs in service Q in the $M/G/1/PS$ model has the geometric distribution $P(Q = k) = (1-\rho)\rho^k$, $k \ge 0$, with mean $EQ = \rho/(1-\rho)$ and, given that there are n jobs in service in steady state, the n remaining service requirements are distributed as n iid random variables each with the stationary-excess cdf G_e. If G_1 and G_2 are two service-requirement cdf's ordered by $G_1 \leq_c G_2$, then the associated stationary-excess cdf's defined in (1.1) are stochastically ordered, i.e., $G_{1e} \leq_{st} G_{2e}$; this is easy to see from (1.1), (1.33) and (1.34). Hence, if G increases in convex order, the mean m_1 and thus ρ remain unchanged, but G_e increases stochastically, so that the steady-state distribution of the remaining service requirements increases stochastically.

We now show that there is an opposite effect on the conditional mean response times given fixed information.

Theorem 1.4 [thH1] *Consider two $M/G/1/PS$ systems with common arrival rate λ and service-requirement cdf's ordered by $G_1 \leq_c G_2$. Then*

$$ET_1(x,y) \geq ET_2(x,y) \quad \textit{for all} \quad x < y \tag{1.35}$$

and

$$E[T_1(x)|n, x, x_1, \ldots, x_{n-1}] \geq E[T_2(x)|n, x, x_1, \ldots, x_{n-1}] \tag{1.36}$$

for all $n, x, x_1, \ldots, x_{n-1}$.

Proof Since $G_1 \leq_c G_2$, $G_{1e} \leq_{st} G_{2e}$. Hence $G_{1e}(t) \geq G_{2e}(t)$ for all t, which in turn implies that $R_1(t) \geq R_2(t)$ for all t by (1.4). The first conclusion follows from the representation

$$ET(x,y) = B^{(1)}(y) - B^{(1)}(y-x) = \int_{y-x}^{y} R(t)dt \ . \tag{1.37}$$

The second conclusion follows because $[T(x)|n, x, x_1, \ldots, x_{n-1}]$ can be expressed as the sum of n independent random variables each distributed as $T(x,y)$ for appropriate (x,y), as given in (1.8). □

The steady-state mean of $T(x)$ without further conditioning is well known to be

$$ET(x) = \frac{x}{1-\rho} \ , \tag{1.38}$$

which clearly is independent of the service-requirement cdf beyond its mean. Thus the two effects of increased service-requirement variability described above exactly cancel out: The remaining service requirements tend to increase, while the conditional mean for any given set of service requirements tends to decrease, so that (1.38) holds.

The stochastic comparisons in Theorem 1.4 allow us to bound the conditional means $ET(x,y)$ and $ET_n(x)$ above by using less variable service-requirement cdf's. In particular, upper bounds for an arbitrary service-requirement cdf G with mean m_1 are obtained by using the deterministic (D) distribution assigning probability 1 to the mean m_1. Hence we can obtain conservative estimates even if we do not know the service-requirement cdf provided we use a deterministic distribution with atom (unit probability) on a value greater than or equal to the true mean.

Example 1.5 [exH1] Gamma distributions with common mean are convex ordered by their shape parameter; see Example 1.5.1(e) on p. 14 of Stoyan [**24**]. For example

$$\hat{g}_p(s) = \left(\frac{1}{1+s/p}\right)^p \tag{1.39}$$

is the Laplace transform of the gamma pdf with mean 1 and shape parameter p, which has variance $1/p$. If $p_1 < p_2$, then $G_{p_1} \geq_c G_{p_2}$. As $p \to \infty$, $\hat{g}_p(s) \to e^{-s}$, which is the Laplace transform of the deterministic distribution with mean 1. To illustrate, consider the four convex ordered cases

$$D \equiv \Gamma_\infty \leq_c \Gamma_5 \equiv E_5 \leq_c M \leq_c \Gamma_{0.2}$$

with variances 0, 0.2, 1 and 5. Plots of $E[T(x)|1,2]$ as a function of the arrival rate, for different service-requirement cdf's, are given in Figure 2 (i.e., $n = 1$, $x_1 = x = 2$). These are obtained by numerically inverting the LT in (1.26) for $n = 2$. We have an independent check on the numerical transform inversion algorithm by comparing results for $ET_1(x)$ with results displayed in Example 1.3 for the $M/M/1/PS$ model.

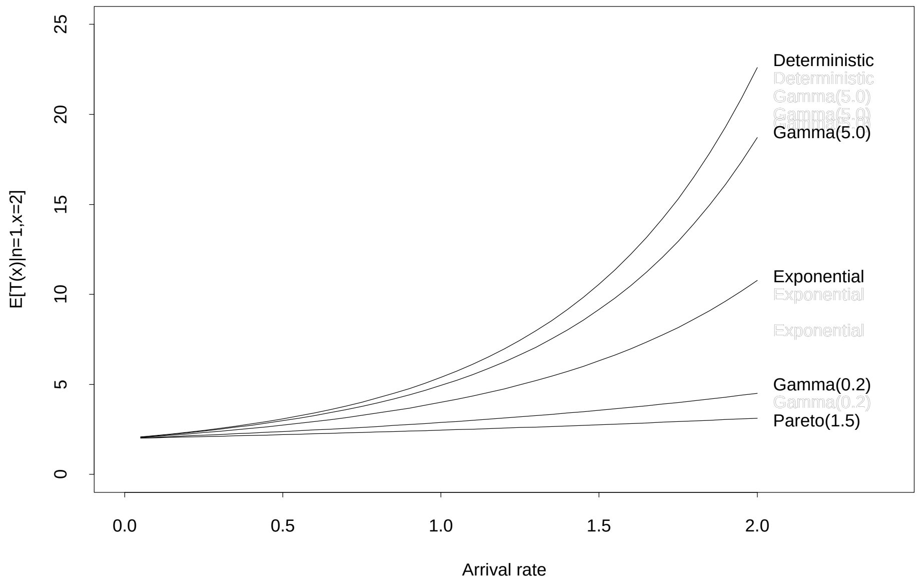

Figure 2 The mean $E[T(x)|1,2]$ as a function of the arrival rate for five different service-requirement cdf's.

Also displayed in Figure 2 are the conditional means for the Pareto service-requirement ccdf

$$G_q^c(t) = (1 + t/(q-1))^{-q} \ , \tag{1.40}$$

with parameter $q = 1.5$. This Pareto cdf has mean 1 for $q > 1$, second moment $m_2 = 2(q-1)/(q-2)$ for $q > 2$ and power tail with exponent q. As q decreases the cdf $G_q(t)$ gets more variable in the convex order. The Laplace transform $\hat{G}_q^c(s)$ of $G_q^c(t)$ in (1.40) which was used in (1.26) was calculated using continued fractions as in Abate and Whitt [**5**]; see especially Section 8 there.

The numerical results in Figure 2 suggest that the Pareto (1.5) distribution with mean 1 is more variable in convex order than the gamma (0.2) distribution with mean 1, but that is not true. This can be seen from plots of the ccdf's in Figure 3. For ccdf's with common mean, convex order is characterized by the ccdf's crossing once and only once. Figure 3 shows that the two Pareto ccdf's cross each other once, but both cross the gamma (0.2) ccdf twice.

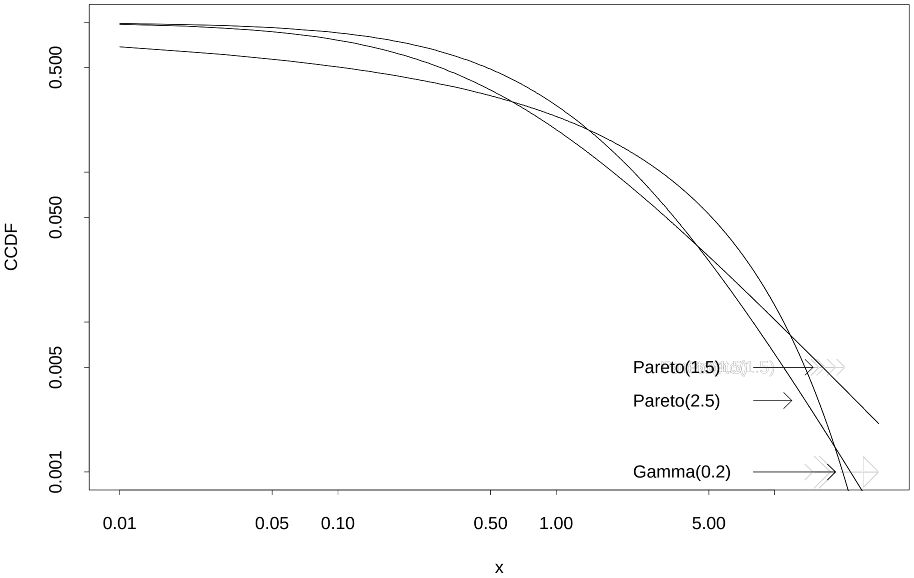

Figure 3 A plot of three ccdf's in log-log scale to show convex order between the two Pareto ccdf's and lack of it between these and the gamma ccdf.

Example 1.6 `[exH2]` We have not yet been able to establish an analog of Theorem 1.4 for the conditional variances. However, we have observed this ordering empirically. Indeed, we see such an ordering for the ratio SD/M. To illustrate, in Figure 4 we display the ratio SD/M for $[T(x)|1,2]$, the same case considered in Example 1.5.

1.4 Reliable Prediction When the Initial Work is Large

In this section we show that reliable prediction becomes possible when the initial work is large, i.e., when there is a large service requirement x of the arriving job or a large number n of other jobs in the system.

We first note that the squared coefficient of variation (SCV, variance divided by the square of the mean) of $T_1(x)$ for any x is bounded above by 1, the value for an exponential random variable. In particular, by (1.18) and (1.19),

$$Var\, T_1(x) = B^{(1)}(x)^2 - 2B^{(2)}(x) < B^{(1)}(x)^2 = ET_1(x)^2 \ . \tag{1.41}$$

However, we want the SCV of $T_n(x)$ to be much less than 1.

The SCV of $T_n(x)$ tends to decrease as we increase x or n because, by (1.8), $T_n(x)$ can be expressed as a sum of independent random variables. Indeed, this independence makes it possible to establish laws of large numbers (LLNs) and central limit theorem (CLT) refinements. To state the results, let $\Rightarrow$ denote convergence in distribution and let $N(m, \sigma^2)$ denote a normal random variable with mean m

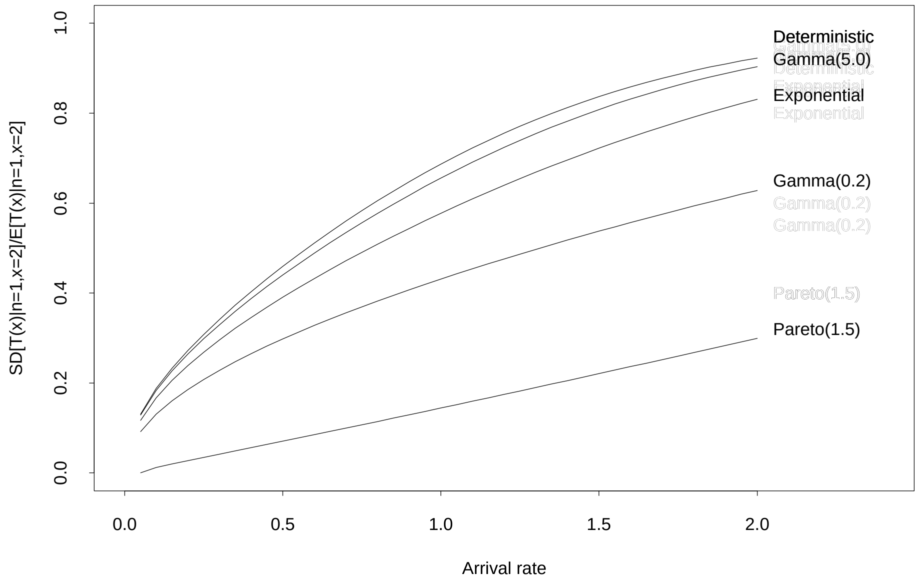

Figure 4 The ratio of the standard deviation to the mean of $[T(x)|1,2]$ as a function of the arrival rate for five different service-requirement cdf's.

and variance σ^2. Recall that convergence in distribution to a deterministic limit is equivalent to convergence in probability. (This initial asymptotic result was previously established by S. A. Grishechkin. The M/M/1/PS model is covered by [**11**], as reviewed by Theorem 6 of [**12**], while the M/G/1/PS model is covered by Theorem 1.1 of [**13**].)

Theorem 1.7 `[thJ1]` *If $x \to \infty$, then*

(a) $\frac{m_n(x)}{x} \to \frac{1}{1-\rho}$,

(b) $\frac{T_n(x)}{m_n(x)} \Rightarrow 1$.

(c) *If in addition $m_2 < \infty$, then*

$$x^{-1/2}[T_n(x) - x/(1-\rho)] \Rightarrow N\left(0, \frac{\rho m_2}{(1-\rho)^3}\right) \quad as \quad x \to \infty \ .$$

Proof Since parts (a) and (b) are consequence of (c) under the extra moment condition, we concentrate on establishing (c). Similar direct arguments apply to parts (a) and (b) without assuming that $m_2 < \infty$. It furthermore suffices to establish the result (c) for $T_1(x)$ because $T_n(x)$ is the sum of $T_1(x)$ and $(n-1)$ other independent random variables $T_j(x_j \wedge x, x)$, as shown in (1.16). These other variables are bounded above by (and approach as $x \to \infty$) $T_j(x_j, \infty)$, which we have noted is distributed as the first passage time $T_{x_j,0}$ with finite mean and variance, independent of x. Hence, these variables are asymptotically negligible in the scaling. We propose two different arguments to treat $T_1(x)$: The first is a high-level argument exploiting the Lindeberg-Feller CLT for independent non-identically distributed random variables on p. 262 of Feller [**10**], while the second is a direct asymptotic argument using the characteristic function of the normalized version of $T_1(x)$. First, for the high-level argument, we observe that $T_1(x)$ can be expressed as the sum of the $k+1$ independent random variables $T_1(1,x)$, $T_2(1,x-1),\ldots,T_{k+1}(1,x-k)$ where k is the largest integer such that $x \geq k+1$ $(x-k \geq 1)$: We simply write the transform as

$$\begin{aligned}\texttt{[J2]}\hat{t}_1(s|x) &= \frac{\hat{t}_1(s|x)}{\hat{t}_1(s|x-1)}\frac{\hat{t}_1(s|x-1)}{\hat{t}_1(s|x-2)}\frac{\hat{t}_1(s|x-2)}{\hat{t}_1(s|x-3)}\cdots\frac{\hat{t}_1(s|x-k)}{\hat{t}_1(s|x-k-1)} \\ &= \hat{t}(1,x)\hat{t}(1,x-1)\hat{t}(1,x-2)\cdots\hat{t}(1,x-k)\ . \qquad (1.42)\end{aligned}$$

Since the distributions of all the random variables $T_{j+1}(1,x-j)$ approach the distribution of T_{10} and are stochastically bounded above by this distribution, and since this distribution has finite mean and variance, we can easily verify the Lindeberg condition (4.15) on p. 262 of Feller [**10**]. The variance of $\hat{t}_1(s|x)$ is asymptotically $x\,Var(T_{10})$ to within an error negligible in the scaling. We obtain $Var(T_{10}) = m_2\rho/(1-\rho)^3$ from Theorem 7 of Abate and Whitt [**3**]. Now turning to the alternate transform proof, we use the Laplace transform of $T_1(x)$ in (1.17) to construct the characteristic function of $x^{-1/2}[T_1(x) - x/(1-\rho)]$; i.e., for real u,

$$\phi(u|x) \equiv Ee^{iu[x^{-1/2}(T_1(x)-x/(1-\rho))]} = \hat{t}_1\left(-\frac{iu}{\sqrt{x}}\middle|x\right)e^{-iu\sqrt{x}/(1-\rho)} \qquad (1.43)$$

for $\hat{t}_1(s|x)$ in (1.17). We then do Taylor series expansions of $\hat{t}_1\left(-\frac{iu}{\sqrt{x}}|x\right)$ and $\hat{t}\left(-\frac{iu}{\sqrt{x}}|z,y\right)$ about $\hat{t}_1(0|x)$ and $\hat{t}(0|z,y)$ appearing in two integrals on the right side of (1.43) (using (1.17)), exploiting the established formulas for the first two moments. This reasoning yields

$$\phi(u|x) \to e^{-\sigma^2 t^2/2} \quad \text{as} \quad x \to \infty\ , \qquad (1.44)$$

where $\sigma^2 = \rho m_2/(1-\rho)^3$, which establishes the desired result by the continuity theorem for characteristic functions; see p. 508 of Feller [**10**]. The second moment condition $m_2 < \infty$ is used to show that the integrated remainder terms in the Taylor series expansions are asymptotically negligible; e.g., we need $y^2G^2(y) \to 0$ as $y \to \infty$, which follows from (6.4) on p. 150 of Feller [**10**]. □

Example 1.8 `[exJ1]` To illustrate Theorem 1.7, we consider a numerical example. In Figures 5 and 6 we plot $m_1(x)/x$ and $v_1(x)/xm_2$ as functions of x for the several service-requirement cdf's (the gamma and Pareto cdf's considered in Figures 2 and 4). In each case we compare these curves to the limiting cases as $x \to \infty$, $(1-\rho)^{-1}$ for the mean and $\rho(1-\rho)^{-3}$ for the variance. We have normalized the mean and variance so that these limits should both be independent of x and the service-requirement cdf G. Consistent with Theorem 1.7(a), Figure 5 shows that $m_1(x)/x \to (1-\rho)^{-1} = 2$ as $x \to \infty$ in this example. Consistent with Theorem 1.7(c), the rate of convergence of $m_1(x)/x$ is slower for the more variable distributions. The rate is especially slow for Pareto (1.5) for which $m_2 = \infty$. Also consistent with Theorem 1.7(c), Figure 6 shows that $v_1(x)/xm_2 \to \rho/(1-\rho)^3 = 4$ as $x \to \infty$.

We now consider CLTs when n increases. For the special case of the M/M/1/PS system, asymptotic results as $n \to \infty$ were obtained previously by Grishechkin in [**11**] and Theorem 6 of [**12**]. For the more general M/G/1/PS model, we clearly need to impose some regularity conditions, because the remaining service requirement x_n of job n could approach 0 very rapidly as $n \to \infty$, so that the total service

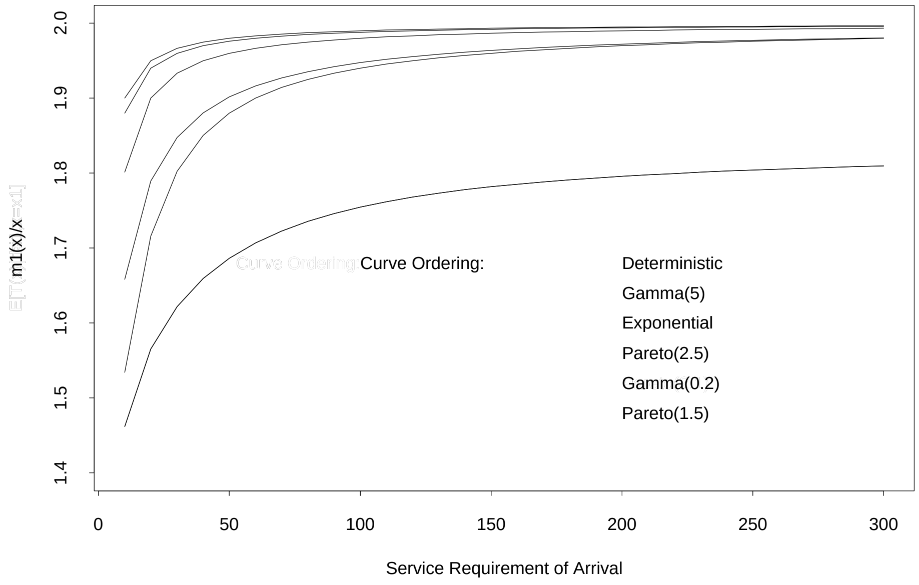

Figure 5 The normalized mean $m_1(x)/x$ as a function of the arrival service requirement x for several service-requirement distributions.

requirement of all n jobs is less than a single job; i.e., we could have $T_n(x) < T_2(x) \equiv [T(x)|x, 2, x_1]$ for $0 < x_1 < x$ and all n, under which we would not have a LLN or a CLT.

A simple regularity condition is to start with m jobs with remaining service requirements $x_1, \ldots, x_m$. Then we can consider n groups of jobs with these service requirements. Let $T_{n,m}(x)$ represent this scheme, which has nm other jobs, n each with remaining service times $x_1, \ldots, x_m$.

Theorem 1.9 `[thJ2]` *In the group framework above, if $n \to \infty$, then*

(a) $\frac{T_{n,m}(x)}{ET_{n,m}(x)} \to 1$,

(b) $\frac{T_{n,m}(x)}{nET_{1,m}(x)} \to 1$.

(c) *If in addition $m_2 < \infty$, then*

$$n^{-1/2}[T_{n,m}(x) - nET_{1,m}(x)] \Rightarrow N(0, Var\, T_{1,m}(x)) \quad as \quad n \to \infty \ .$$

Proof $T_{n,m}(x)$ is distributed as the sum of n i.i.d. random variables distributed as $T_{1,m}(x)$ plus a single random variable distributed as $T_1(x)$. Thus the classical LLN and CLT for i.i.d. random variables can be used. □

It is clearly possible to generalize Theorem 1.9. One way is to prohibit arbitrarily small service requirements.

Theorem 1.10 `[thJ3]` *Suppose that $m_2 < \infty$ and that there is an $\epsilon > 0$ and an $M < \infty$ so that $x_j > \epsilon$ for all but M of the remaining service requirements.*

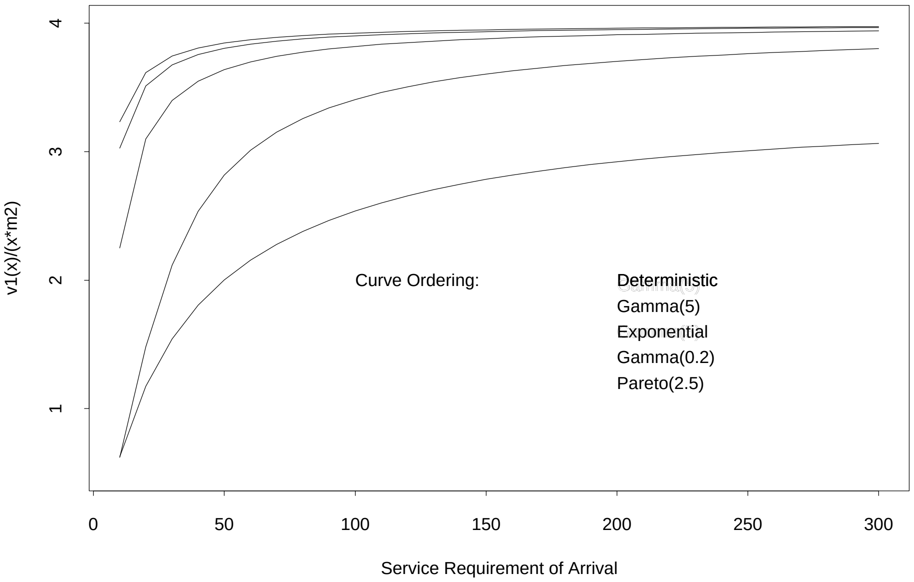

Figure 6 The normalized variance $v_1(x)/xm_2$ as a function of the arrival service requirement x for several service-requirement distributions.

Then

$$\frac{T_n(x) - ET_n(x)}{\sqrt{Var\,T_n(x)}} \Rightarrow N(0,1) \quad as \quad n \to \infty.$$

Proof The Lindeberg-Feller CLT for independent non-identically distributed random variables can be applied, just as in Theorem 1.7. The up-to M random variables with remaining service times less than ϵ are asymptotically negligible. The remaining $n - M$ variables are stochastically bounded between $T(\epsilon, x)$ and $T(x,x) = T_1(x)$, which allows us to verify condition (4.15) on p. 262 of Feller [**10**]. □

Example 1.11 [exJ2] We illustrate Theorems 1.7 and 1.9 by considering a numerical example involving $[T(x)|n, x_1, \ldots, x_n]$ in which $x = x_n = 2$ and $x_j = 1$ for $1 \leq j \leq n-1$. In Figure 7 we plot the ratio SD/M as a function of the arrival rate for several different values of n.

1.5 Only Conditioning on the Number of Jobs

In this section we consider the prediction problem when we condition only upon the service requirement x of the arriving job and the total number n of jobs in the system; i.e., we do not exploit either the remaining service requirements or the completed service requirements of the other $n-1$ jobs. However, we do exploit knowledge of the service-requirement cdf G. In this situation prediction is facilitated by the fact that the remaining $n-1$ service requirements of the other jobs, conditional on there being $n-1$ other jobs in the system in steady state, are

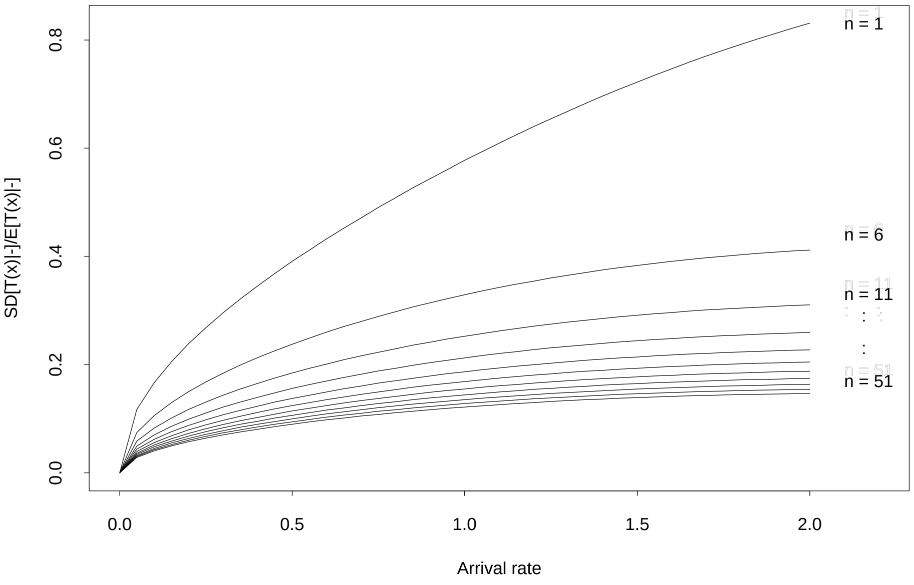

Figure 7 The ratio of the standard deviation to the mean for $[T(x)|n, x_1, \ldots, x_n]$ as a function of the arrival rate for several values of n, when $x = x_n = 2$ and $x_j = 1$, $1 \leq j \leq n-1$.

distributed as $n-1$ iid random variables with the service-requirement stationary-excess cdf G_e in (1.1). (In Section 3 we noted that this is part of the classical theory.) In this section we assume that $\rho < 1$, so that we can invoke this steady-state theory.

In the special case of the $M/M/1/PS$ model, the conditional response time given the number of jobs in service was treated by Coffman, Muntz and Trotter [**7**] and Sengupta and Jagerman [**23**]. for the $M/M/1/PS$ model, this conditional response time is unchanged by also conditioning on the amounts of completed service, because the remaining service again has an exponential distribution.

As before let $T_1(x)$, $m_1(x)$ and $v_1(x)$ be the random value, mean and variance of the conditional response time of a single job with service requirement x given that it is initially the only job. Let $T(x,n)$, $M_k(x,n)$ and $V(x,n)$ be the random value, k^{th} moment and variance of the conditional response time of this same job with service requirement x given that initially there are also $n-1$ other jobs in the system (without specifying remaining or completed service). Let $M_k(x)$ and $V(x)$ be the k^{th} moment and variance of the time required for one of the other jobs to complete the minimum of x and its service requirement. Because of the properties mentioned above, we clearly have

$$\texttt{[M1]}M_1(x,n) = m_1(x) + (n-1)M_1(x) \ , \tag{1.45}$$

$$\texttt{[M2]}V(x,n) = v_1(x) + (n-1)V(x) \ , \tag{1.46}$$

$$\texttt{[M3]}V(x) = M_2(x) - M_1(x)^2 \ , \tag{1.47}$$

We have already shown how to calculate $m_1(x)$ and $v_1(x)$ in (1.24) and (1.25). We now show how to compute all the remaining quantities by showing how to compute the two moments $M_1(x)$ and $M_2(x)$.

Theorem 1.12 `[thM1]` *The two moments $M_1(x)$ and $M_2(x)$ can be expressed as*

$$M_1(x) = B^{(1)}(x) - \int_0^x B^{(1)}(x-u)dG_e(u) \tag{1.48}$$

and

$$\begin{aligned} \texttt{[M5]}M_2(x) &= 2B^{(1)}(x)^2 - 2B^{(2)}(x) + 2\int_0^x B^{(2)}(x-u)dG_e(u) \\ &\quad -2B^{(1)}(x)\int_0^x B^{(1)}(x-u)dG_e(u) \ , \end{aligned} \tag{1.49}$$

where G_e, $B^{(1)}$ and $B^{(2)}$ are defined in Section 2. The mean in (1.48) has LT

$$\hat{M}_1(s) \equiv \int_0^\infty e^{-st}M_1(t)dt = \hat{B}^{(1)}(s)(1-\hat{g}_e(s)) \ , \tag{1.50}$$

while the convolution integrals in the third and fourth terms of (1.49) have LTs $s\hat{B}^{(1)}(s)^2\hat{g}_e(s)$ and $\hat{B}^{(1)}(s)\hat{g}_e(s)$.

Proof For (1.48), apply (1.24). For (1.49), apply (1.47), (1.24) and (1.25). □

From (1.45)–(1.47) and Theorem 1.12, we see that we can apply numerical transform inversion to calculate $M_1(x)$, $V(x)$, $M_1(x,n)$ and $V(x,n)$ for any x and n.

Example 1.13 `[exN51]` Continuing Example 1.3, let the service-requirement pdf g be exponential with mean 1. Then $\hat{M}_1(s) = 1/s(1-\rho+s)$, so that

$$M_1(x) = (1-\rho)^{-1}(1-e^{-(1-\rho)x}), \quad x \geq 0 \ , \tag{1.51}$$

and $M_1(x,n)$ is obtained by combining (1.45), (1.51) and (1.30). We see that $M_1(x,n)$ agrees with formula (33) of Coffman et al. [**7**] for all n. We have numerically verified the second-moment formula. To illustrate, Table 1 displays results for the mean $M_1(x,n)$ and variance $V(x,n)$ as a function of n for $x=1$ and $\rho=0.5$. These results were obtained from both [**7**] and numerical inversion. For the numerical inversion (see Section 7), setting the target discretization error bound at 10^{-8} and calculating A for B_1 as $-\log(10^{-8}/6x)$, we obtained accuracy to at least 10^{-7} for the mean and 10^{-6} for the variance.

n	mean	variance
1	1.213061	1.588668 $e-1$
2	2.000000	4.659456 $e-1$
4	3.573877	1.080103
6	5.147755	1.694261
8	6.721632	2.308418

Table 1 The mean $M_1(x,n)$ and variance $V(x,n)$ as a function of n for Example 1.13 with $\rho=0.5$ and $x=1$.

`[ta1]`

Given the iid structure, it is evident that we also have the following CLT, paralleling Theorem 1.9(c).

Theorem 1.14 `[thM2]` *When we condition only upon n and x,*

$$\frac{T(x,n) - nM_1(x)}{\sqrt{nV(x)}} \Rightarrow N(0,1) \quad \textit{as} \quad n \to \infty \ ,$$

where $M_1(x)$ and $V(x)$ are the mean and variance in (1.47)–(1.49).

In order to establish a CLT as $x \to \infty$, we first give conditions for $M_1(x)$ and $M_2(x)$ to converge to finite limits as $x \to \infty$.

Theorem 1.15 `[thM3]` *Consider the moments $M_1(x)$ and $M_2(x)$ in Theorem 1.12.*

(a) *If $m_2 < \infty$, then*

$$M_1(x) \to \frac{m_{e1}}{1-\rho} = \frac{m_2}{2m_1(1-\rho)} \quad \textit{as} \quad x \to \infty \ .$$

(b) *If $m_3 < \infty$, then*

$$M_2(x) \to \frac{m_{e2}}{(1-\rho)^2} = \frac{m_3}{3m_1(1-\rho)^2} \quad \textit{as} \quad x \to \infty \ .$$

Proof (a) Rewrite (1.48) as

$$M_1(x) = B^{(1)}(x)G_e^c(x) - \int_0^x [B^{(1)}(x) - B^{(1)}(x-u)]dG_e(u) \ .$$

Then note that

$$B^{(1)}(x)G_e^c(x) \le xG_e^c(x)/(1-\rho) \to 0 \quad \text{as} \quad x \to \infty$$

because $m_{e1} < \infty$, see (6.4) on p. 150 of Feller [**10**], while

$$B^{(1)}(x) - B^{(1)}(x-u) \to \frac{u}{1-\rho} \quad \text{as} \quad x \to \infty$$

and $B^{(1)}(x) - B^{(1)}(x-u) \le u/(1-\rho)$, so that by the dominated convergence theorem

$$\int_0^x [B^{(1)}(x) - B^{(1)}(x-u)]dG_e(u) \to (1-\rho)^{-1}\int_0^\infty u dG_e(u) = \frac{m_{e1}}{1-\rho} \quad \text{as} \quad x \to \infty \, .$$

(b) Rewrite (1.49) as

$$M_2(x) = 2B^{(1)}(x)^2 G_e^2(x) - 2B^{(2)}(x)G_e^c(x)$$
$$+2\int_0^x \left\{B^{(1)}(x)[B^{(1)}(x) - B^{(1)}(x-u)] - [B^{(2)}(x) - B^{(2)}(x-u)]\right\} dG_e(x) \ .$$

Then note that

$$2B^{(1)}(x)^2 G_e^c(x) \le 2x^2 G_e^c(x)/(1-\rho)^2 \to 0$$

and

$$2B^{(2)}(x)G_e^c(x) \le x^2 G_e^c(x)/(1-\rho)^2 \to 0$$

by the moment condition $m_3 < \infty$, using (6.4) on p. 150 of Feller [**10**] again. Also,

$$2\left\{B^{(1)}(x)[B^{(1)}(x) - B^{(1)}(x-u)] - [B^{(2)}(x) - B^{(2)}(x-u)]\right\}$$
$$\to \frac{2xu}{(1-\rho)^2} - \frac{2}{(1-\rho)^2}\left(\frac{x^2}{2} - \frac{(x-u)^2}{2}\right) = \frac{u^2}{(1-\rho)^2} \quad \text{as} \quad x \to \infty \ ,$$

so that indeed

$$M_2(x) \to m_{e2}/(1-\rho)^2$$

by the dominated convergence theorem. (We omit the demonstration of domination, which is tedious. We exploit the fact that W has finite first and second moments since we have assumed that $m_3 < \infty$.) □

We apply Theorem 1.15 to establish a CLT as $x \to \infty$.

Theorem 1.16 `[thM4]` *If $m_3 < \infty$, then*

$$x^{-1/2}[T(x,n) - x/(1-\rho)] \Rightarrow N\left(0, \frac{\rho m_2}{(1-\rho)^3}\right) \quad \textit{as} \quad x \to \infty \ .$$

Proof By Theorem 1.7, we have the CLT for $T_1(x)$. By Theorem 1.15, the $n-1$ other variables are asymptotically negligible in the scaling as $x \to \infty$. □

We close this section by deriving an expression for the Laplace transform $\hat{t}(s;x,n) \equiv Ee^{-sT(x,n)}$ in terms of $\hat{f}_1(s|x) = 1/\hat{t}_1(s|x)$. See Remark 1.1 of Ott [**21**] for previous use of this reciprocal of the transform $\hat{t}_1(s,x)$. Applying (1.16), we see that

$$\hat{t}(s;x,n) = \frac{[\int_0^x \hat{f}_1(s|x-y)dG_e(y) + G_e^c(x)]^{n-1}}{\hat{f}_1(s|x)^n} \ . \tag{1.52}$$

We can then express the numerator terms in (1.52) in terms of double transforms; i.e., let

$$\hat{f}_e(s,x) = \int_0^x \hat{f}_1(s|x-y)dG_e(y) + G_e^c(x) \tag{1.53}$$

and

$$f_1^*(s,\tau) \equiv \int_0^\infty e^{-\tau x} f_1(s|x)dx \ . \tag{1.54}$$

Then

$$f_e^*(s,\tau) \equiv \int_0^\infty e^{-\tau x}\hat{f}_e(s,x)dx = f_1^*(s,\tau)\hat{g}_e(\tau) + \frac{1-\hat{g}_e(\tau)}{\tau} \ . \tag{1.55}$$

From (1.11) of Ott [**21**],

$$f_1^*(s,\tau) = \frac{\tau - \lambda(1-\hat{g}(\tau))}{\tau(\tau - s - \lambda(1-\hat{g}(\tau))} \ . \tag{1.56}$$

1.6 Conditioning Upon Completed Work

In this section we make predictions based on the amount x of required work for the new arrival, the number n of jobs in the system and the amount of *completed* work for each of the $n-1$ other jobs already in the system, denoted by y_j, $1 \le j \le n-1$. Let $G(x|y)$ be the conditional cdf of remaining work given the completed work for one job, i.e.,

$$G(x|y) = \frac{G(x+y) - G(y)}{1 - G(y)}, \quad x > 0 \ . \tag{1.57}$$

Since $G(\cdot|y) = G = G_e$ when G is exponential, the results in this section agree with the results in Section 5 for the $M/M/1/PS$ model, but not otherwise. (The exponential distribution is the only continuous distribution for which $G_e = G$.)

We start by indicating how to calculate the mean and variance of the conditional response time. The reasoning is very similar to Section 5. Let $T_c(x,n)$, $M_{ck}(x,n)$ and $V_c(x,n)$ denote the random variable, k^{th} moment and variance, respectively, of the conditional response time of the job with service requirement x given that there are initially $n-1$ other jobs with a vector of completed amounts of service $(y_1, \ldots, y_{n-1})$. We suppress the vector $(y_1, \ldots, y_{n-1})$ in our notation, using the subscript c to indicate this form of information. Let $M_{yk}(x)$ and $V_y(x)$ be the k^{th} moment and variance of the time required for one of the other jobs to complete the minimum of x service and its own remaining service requirement given that it had previously completed an amount y of service. Paralleling (1.45)–(1.47), we have the following relations:

$$\begin{aligned} \texttt{[N1]}M_{c1}(x,n) &= m_1(x) + \sum_{j=1}^{n-1} M_{y_j 1}(x) \ , & (1.58) \\ \texttt{[N2]}V_c(x,n) &= v_1(x) + \sum_{j=1}^{n-1} V_{y_j}(x) \ , & (1.59) \\ \texttt{[N3]}V_y(x) &= M_{y2}(x) - M_{y1}(x)^2 \ . & (1.60) \end{aligned}$$

Since we have shown how to compute $m_1(x)$ and $v_1(x)$ in (1.24) and (1.25), it suffices to show how to compute $M_{y1}(x)$ and $M_{y2}(x)$ for any $x > 0$ and $y > 0$. The following is a direct analog of Theorem 1.12.

Theorem 1.17 `[thN1]` *The first two moments $M_{y1}(x)$ and $M_{y2}(x)$ can be expressed as*

$$M_{y1}(x) = B^{(1)}(x) - \int_0^x B^{(1)}(x-u)dG(u|y) \qquad (1.61)$$

and

$$\begin{aligned} \texttt{[N5]}M_{y2}(x) &= 2B^{(1)}(x)^2 - 2B^{(2)}(x) + 2\int_0^x B^{(2)}(x-u)dG(u|y) \\ &\quad -2B^{(1)}(x)\int_0^x B^{(1)}(x-u)dG(u|y) \ , \qquad (1.62) \end{aligned}$$

where $G(u|y)$ is the conditional remaining-service cdf in (1.57) and $B^{(1)}$ and $B^{(2)}$ are defined in Section 2. The mean in (1.61) has LT

$$\hat{M}_{y1}(s) \equiv \int_0^\infty e^{-st} M_{y1}(t)dt = \hat{B}^{(1)}(s)(1-\hat{g}_y(s)) \ , \qquad (1.63)$$

while the convolution integrals in the third and fourth terms of (1.62) have LTs $s\hat{B}^{(1)}(s)^2\hat{g}_y(s)$ and $\hat{B}^{(1)}(s)\hat{g}_y(s)$, where $\hat{g}_y(s)$ is the LST of the cdf $G(u|y)$.

We can exploit (1.58)–(1.60) and Theorem 1.17 to calculate the mean $M_{c1}(x,n)$ and the variance $V_c(x,n)$ by numerical transform inversion. However, now we need to know the LST $\hat{g}_y(s)$ of the conditional cdf $G(x|y)$ in (1.57) as well as the LST $\hat{g}(s)$ of the original service-requirement cdf G. Fortunately, for many families of distributions, the conditional cdf $G(x|y)$ often inherits the form of the original cdf $G(x)$; e.g., see Section 4 of Duffield and Whitt [**9**].

Just as in Sections 4 and 5, the variable $T_c(x,n)$ is the sum of n independent random variables, so that we can establish LLNs and CLTs under regularity conditions. As in Section 4, we have to account for the nonidentical distributions. Analogs of Theorems 1.7 and 1.9 follow by essentially the same reasoning. Hence, we see that $T_c(x,n)$ should be approximately normally distributed when either x or n is large (or both).

Example 1.18 [exL1] We now consider numerical examples to illustrate the consequences of using the three different kinds of information in Sections 2, 5 and 6. A useful numerical check is provided by the $M/M/1/PS$ model where the exponential cdf G has mean 1, because then the conditional-remaining-service-requirement cdf's are all exponential with mean 1 in Sections 5 and 6, so the formulas agree. We now consider information comparisons for the $M/H_2/1/PS$ model, having a hyperexponential (H_2) service-requirement cdf. In particular, the pdf is

$$g(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}, \quad x \ge 0 , \tag{1.64}$$

where $p = (1 + 1/\sqrt{3})/2$, $\lambda_1 = 2p$ and $\lambda_2 = 2(1-p)$, which yields mean 1, SCV $c_s^2 = 2$ and "balanced means", i.e., $p\lambda_1^{-1} = (1-p)\lambda_2^{-1}$. We assume that the arriving job has a service requirement of 10. Since that requirement is large compared to the mean (1), there is considerable opportunity for that customer's response time to be influenced by other customers.

We consider two cases: There is either 1 old customer or 10, making $n = 2$ and 11. We let the required remaining service time or the age be the same value x for all old customers, and we let x range in powers of 10 from 0.01 to 100. The expected values for the three forms of conditioning are displayed in Table 2. We see that the mean is lowest (highest) when the good news (bad news) corresponding to low (high) x is clearest, when x denotes the required remaining service requirement of each old job. When we condition on the ages, the conditional mean increases with x, as it should, because an H_2 distribution has decreasing failure rate (is DFR), meaning that the conditional remaining service requirement increases as the age increases. From Table 2, we see that the form of information can significantly influence the mean.

We display the corresponding variances in Table 3. We display the conditional variance when we condition on n only, and the variance ratios $Var(T(x)|Reqd.)/Var(T(x)|n)$ and $Var(T(x)|Ages)/Var(T(x)|n)$ in the other two cases. For smaller x, the conditional variances are ordered as expected, according to the utility of the information. However, that is not true for larger values of x. In those cases, the means are very different, so that the variance becomes somewhat a second-order factor. From Tables 2 and 3, it is evident that the form of information can significantly affect both the predicted value (the conditional mean) and the reliability of that prediction (the conditional variance or standard deviation).

Just as at the end of Section 5, we can obtain an expression for the Laplace transform of $T_c(x,n)$. Let $\hat{t}_c(s|x,n) \equiv \hat{t}_c(s|x,n,y_1,\ldots,y_{n-1})$ be the LST

$$\hat{t}_c(s|x,n) \equiv E[e^{-sT_c(x,n)}] , \tag{1.65}$$

where y_j are the amounts of completed work. Paralleling (1.52), we can write

$$\hat{t}_c(s|x,n) = \prod_{j=1}^{n-1} \frac{\left[\int_0^x \hat{f}_1(s|x-u)dG(u|y_j) + G^c(x|y_j)\right]}{\hat{f}_1(s|x)^n} . \tag{1.66}$$

We then can express the double transform of the numerator terms

$$\hat{f}_y(s|x) \equiv \int_0^x \hat{f}_1(s|x-u)dG(u|y_j) + G^c(x|y_j) \tag{1.67}$$

$n = 2$	Form of Conditioning		
x	required	ages	n only
0.01	17.2	19.1	20.0
0.10	17.4	19.1	20.0
1.00	19.1	19.9	20.0
10.00	34.4	21.6	20.0
100.00	34.4	21.6	20.0
$n = 11$	Form of Conditioning		
x	required	ages	n only
0.01	17.4	36.2	45.3
0.10	19.1	36.8	45.3
1.00	36.6	44.0	45.3
10.00	189.0	61.2	45.3
100.00	189.0	61.2	45.3

Table 2 Comparison of different forms of conditioning in Example 1.18: The expected conditional response time in an $M/H_2/1/PS$ model for a new arrival with service requirement 10 given $n-1$ old jobs ($n = 2$ and 11) each having required remaining service or age x, as a function of x.

[ta2]

$n = 2$	Variance	Variance Ratios	
x	given n only	required	ages
0.01	64.9	0.663	0.875
0.10		0.674	0.883
1.00		0.790	0.984
10.00		1.322	1.158
100.00		1.322	1.158
$n = 11$	Variance	Variance Ratios	
x	given n only	required	ages
0.01	262.8	0.167	0.691
0.10		0.196	0.712
1.00		0.482	0.960
10.00		1.796	1.389
100.00		1.796	1.389

Table 3 Comparison of different forms of conditioning in Example 1.18: The variance of the conditional response time of a new job given n only and the ratios of the other conditional variance to this one, again for a new arrival wtih service requirement 10.

[ta3]

as

$$f_y^*(s,\tau) \equiv \int_0^\infty e^{-\tau x} f_y(s|x) = f_1^*(s,\tau)\hat{g}_{y_j}(\tau) + \frac{1-\hat{g}_{y_j}(\tau)}{\tau} \tag{1.68}$$

where $f_1^*(s,\tau)$ is given in (1.56) and $\hat{g}_y(\tau)$ is the LST of the cdf $G(x|y)$ in (1.57).

1.7 Numerical Inversion Issues

In this section we discuss technical issues that arise when we perform numerical inversion of the transforms in Sections 2, 5 and 6.

1.7.1 Controlling the Discretization Error. When we use the Fourier-series method, the discretization error is controlled by exploiting the structure of the function to be calculated; e.g., see (11) and (12) of [**4**]. In particular, when the function is $f(x)$, the discretization error is

$$e_d \equiv e_d(x) = \sum_{k=1}^{\infty} e^{-kA} f((2k+1)x) \ . \tag{1.69}$$

From (1.69), we see that the discretization error is controlled by choosing A suitably large. However, we are usually constrained from choosing A too large because very large A introduces roundoff error; e.g., see Section 2.2 of [**1**]. If A needs to be increased significantly, then we also need to increase the roundoff-control parameter l (e.g., typically to $l = 2$ or 3 from 1), which increases the computation by approximately a factor of l. Working with standard double precision, it is usually reasonable to aim for discretization error of the order 10^{-8} (in the interval $10^{-6} - 10^{-10}$).

When we calculate cdf's or ccdf's we can use predetermined control parameters, because cdf's and ccdf's are always bounded by 1. However, this convenient property does not hold for the functions considered here. Hence, we need to bound the functions above by convenient functions we can analyze.

There are two cases to consider here: $\rho < 1$ and $\rho \geq 1$. For $\rho < 1$, we can easily establish convenient bounds. First, from (1.4), by bounding $W(x)$ above by 1,

$$R(x) \leq \frac{1}{1-\rho} \ . \tag{1.70}$$

Hence, from (1.18)–(1.20), (1.24) and (1.28),

$$\begin{aligned}
\texttt{[X2]}\, m_1(x) &= B^{(1)}(x) \leq \frac{x}{1-\rho} \ , &(1.71)\\
\texttt{[X3]}\, v_1(x) &\leq B^{(1)}(x)^2 \leq \frac{x^2}{(1-\rho)^2} \ , &(1.72)\\
\texttt{[X4]}\, B^{(2)}(x) &\leq \frac{x^2}{2(1-\rho)^2} \ , &(1.73)\\
\texttt{[X5]}\, m_n(x) &\leq nB^{(1)}(x) \leq \frac{nx}{1-\rho} \ , &(1.74)\\
\texttt{[X6]}\, |v_{2n}(x)| &\leq 2nB^{(2)}(x) \leq \frac{nx^2}{(1-\rho)^2} \ . &(1.75)
\end{aligned}$$

It is easy to see that similar bounds hold for the functions in Sections 5 and 6.

By (1.71)–(1.75), we need to consider functions of the form $c_1 x$ and $c_2 x^2$ for constants c_1 and c_2. The first case was explicitly worked out in (5.29) of [**2**] and applied to renewal functions in Section 13 of [**2**].

Given (1.69), if $|f(x)| \leq cx^k$, then the discretization error is bounded by

$$|e_d(x)| \leq \sum_{k=1}^{\infty} e^{-kA}(2k+1)cx^k \leq cx^k \frac{(3e^{-A} - e^{-2A})}{(1-e^{-A})^2} \approx 3cx^k e^{-A} \ . \tag{1.76}$$

If the desired discretization-error bound is ϵ (e.g., 10^{-8}), then we should set A at

$$A = -\log(\epsilon/3cx^k) \ . \tag{1.77}$$

For example, for $m_n(x)$ we combine (1.74) and (1.77) to obtain

$$A[m_n(x)] = -\log(\epsilon(1-\rho)/3nx) \ . \tag{1.78}$$

The behavior for $\rho \geq 1$ is more complicated, but simple bounds can be obtained from the inequality

$$T_n(x) \leq \int_0^x [n + A(t)]dt \ , \tag{1.79}$$

where $A(t)$ is the number of arrivals in $[0,t]$, which is obtained by assuming that there are no departures at all before the time that the job of interest with service requirement x departs. From (1.79), we get the simple bounds

$$m_n(x) \equiv ET_n(x) \leq \int_0^x [n + EA(t)]dt = nx + \frac{\lambda x^2}{2} \tag{1.80}$$

and

$$\begin{aligned} \texttt{[X14]}E[T_n(x)]^2 &\leq E\left[\left(\int_0^x [n+A(t)]dt\right)^2\right] \\ &\leq E[x^2(n+A(x))^2] = x^2(n^2 + 2n\lambda x + \lambda x + (\lambda x)^2) \end{aligned} \tag{1.81}$$

For example, the analog of (1.77) for (1.81) is

$$A = -\log(\epsilon/3[n^2x^2 + (2n+1)\lambda x^3 + \lambda^2 x^4]) \ . \tag{1.82}$$

These bounds in (1.80) and (1.81) for $\rho \geq 1$ make it natural to ask how the $M/G/1/PS$ queue grows when $\rho \geq 1$. This has been studied by Jean-Marie and Robert [**16**].

1.7.2 Smoothness. We note that the conditional mean $m_n(x)$ in (1.24) has a discontinuous derivative $\dot{m}_n(x)$ with jumps at the points x_j for $x_j < x$. The discontinuity can be seen from the terms e^{-sx_j} in (1.26), which correspond to atoms at x_j. In particular,

$$\dot{m}_n(x) = n\dot{B}^{(1)}(x) - \sum_{j=1}^{n-1} \dot{B}^{(1)}(x - (x \wedge x_j)) \ ,$$

where

$$\dot{B}^{(1)}(x) = R(x) \ .$$

Since $R(x - x_j) = 0$ for $x < x_j$,

$$\dot{B}^{(1)}(x - x_j) = 0 \quad \text{for} \quad x < x_j \ .$$

However, from (1.4),

$$R(x - x_j) \downarrow 1 \quad \text{as} \quad x \downarrow x_j \ ,$$

so that

$$\dot{B}^{(1)}(x - x_j) \downarrow 1 \quad \text{as} \quad x \downarrow x_j \ .$$

Similarly, $v_{n2}(x)$ and $B^{(2)}(x - x_j)$ in (1.28) have discontinuous second derivatives with jumps at x_j for $x_j < x$. (The convolution in (1.20) is a smoothness operator.)

These discontinuities degrade numerical accuracy to some extent. The problem for $m_n(x)$ tends not to be nearly as bad as if $m_n(x)$ itself had a discontinuity.

Similarly, the problem is even less serious when the discontinuity appears in the second derivative, as with $B^{(2)}(x)$.

If it is deemed necessary, we can smooth the functions, e.g., see Section 6 of [**2**]. A simple way to smooth is by approximating the atoms at x_j. In particular, we can replace e^{-sx_j} in (1.26) and (1.29) by

$$\hat{e}_n(s; x_j) \equiv (1 + sx_j/n)^n \tag{1.83}$$

for large n like 1024. (A power of 2 is convenient for computation.) Formula (1.83) corresponds to an Erlang E_n pdf with mean x_j and SCV $1/n$. As n increases, it approaches the point mass at x_j. The numerical accuracy can be checked by trying different n.

Another alternative is to work directly with $\hat{B}^{(1)}(s)$ and calculate $B^{(1)}(x)$ and $B^{(1)}(x - x_j)$ separately. To calculate $m_n(x)$ we then need n inversions instead of 1. Then the only discontinuities are at $x = 0$, which tends to cause no problem. To illustrate, we do a numerical example.

Example 1.19 [ex71] We consider the $M/M/1/PS$ model with mean service time 1 and arrival rate $\rho = 0.5$. We let the arrival see 4 jobs with $x_j = j$, $1 \leq j \leq 4$. We calculate $m_5(x)$ by both one inversion and five inversions. We provide an additional check by calculating the exact value using (1.24) and (1.30). In the inversion we set the target discretization error at 10^{-12}. Using (1.71) and (1.77), we let $A = -\log(10^{-12}/3(10)x)$ for one inversion and $A = -\log(10^{-12}/3(2)x)$ for five inversions. Table 4 displays some of the results, focusing on the behavior at and near the points of discontinuity of $\dot{m}_5(x)$, in particular, at $x = 2.0$ and $x = 4.0$. The absolute error with five separate inversions is consistently about 10^{-11}, while it ranges from less than 10^{-3} to less than 10^{-6} with one inversion, peaking at the integers j with $1 \leq j \leq 4$. These results show the consequences of lack of smoothness, but they indicate that for practical purposes, a single inversion usually should suffice.

	exact by	absolute error by inversion	
x	(1.24) and (1.30)	one inversion	five inversions
1.50	9.166	1.8 $e-07$	1.0 $e-11$
1.80	11.125	2.7 $e-07$	9.9 $e-12$
1.90	11.792	7.3 $e-07$	9.8 $e-12$
2.00	12.466	2.0 $e-04$	9.8 $e-12$
2.10	13.0.43	1.6 $e-06$	9.8 $e-12$
2.20	13.621	1.1 $e-06$	9.6 $e-12$
2.50	15.363	1.7 $e-07$	9.4 $e-12$
3.50	20.662	6.0 $e-07$	8.4 $e-12$
3.80	22.049	5.1 $e-07$	8.2 $e-12$
3.90	22.505	5.1 $e-07$	8.0 $e-12$
4.00	22.958	4.1 $e-04$	8.0 $e-12$
4.10	23.307	4.7 $e-06$	7.9 $e-12$
4.20	23.648	2.9 $e-06$	7.8 $e-12$

Table 4 A comparison of three methods for computing the conditional mean $m_5(x)$ in the $M/M/1/PS$ model in Example 1.19: one inversion, five separate inversions and exact using Example 1.3.

[ta4]

1.8 Conclusions

We have demonstrated that it is possible, through the use of numerical transform inversion, to calculate the conditional mean and variance of a customer's response time in a $M/G/1/PS$ queue, given various forms of system state information. We have also demonstrated that the conditional mean often leads to reliable predictions. In particular, when either the number of jobs in the system or the workload of the arriving job is large, the distribution of the conditional response time estimate clusters about its mean. We have developed supporting theory for this claim for 3 different forms of conditioning: conditioning on remaining service requirements (sections 3 and 4), conditioning on only the number of jobs in the system at the time of arrival (section 5), and conditioning on the completed work of the jobs in the system (section 6). We have given numerical examples demonstrating the possibility of computing prediction estimates in practice.

To compare the different forms of conditioning, we suggest quantifying the value of information by comparing the values of the mean and variance with and without the additional information. We gave a numerical example comparing three alternative forms of information in Example 1.18. In some cases, the conditional mean changes significantly when the information is clearly important because the single predicted value changes. Prediction is enhanced by some form of information, even when the mean does not change much, if the variance is reduced, because then the predicted value is more reliable. We typically find the greatest variance reduction when we condition on the remaining service requirements. We further note that the value of information is highly dependent upon the service-requirement cdf. For example, for an exponential service-requirement cdf, knowing the ages of the jobs in the system offers no additional information, whereas for the deterministic cdf, knowing the ages of the jobs in the system is equivalent to knowing the remaining service requirements of the jobs. Thus the value of system-state information depends both upon the workload of the arriving job and the service-requirement cdf. It remains to further explore how to appropriately capture the value of different forms of information.

Acknowledgment. We thank Sergei Grishechkin for telling us about his related work on limits for processor-sharing queues.

Bibliography

[1] J. Abate, G. L. Choudhury and W. Whitt, An introduction to numerical transform inversion and its application to probability models. Chapter 8 in *Computational Probability*, W. Grassman (ed.), Kluwer, Boston, 1999, 257–323.

[2] J. Abate and W. Whitt, The Fourier-series method for inverting transform of probability distributions, *Queueing Systems* **10** (1992), 5–88.

[3] J. Abate and W. Whitt, Transient behavior of the $M/G/1$ workload process. *Operations Res.* **42** (1994), 750–764.

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

[5] J. Abate and W. Whitt, Computing Laplace transforms for numerical inversion via continued fractions. *INFORMS J. Computing* **11** (1999), 394–405.

[6] J. Braband, Waiting time distributions for closed $M/M/N$ processor sharing queues. *Queueing Systems* **19** (1995), 331–344.

[7] E. G. Coffman, Jr., R. R. Muntz and H. Trotter, Waiting time distributions for processor-sharing systems, *JACM* **17** (1970), 123–130.

[8] A. Demers, S. Keshav and S. Shenker, Analysis and simulation of a fair queueing algorithm, *Proc. ACM Sigmetrics* '89, (1989) 3–12. *Journal of Internetworking: Research and Experience*, **1** (1990) 3–26.

[9] N. G. Duffield and W. Whitt, A source traffic model and its transient analysis for network control. *Stochastic Models* **14** (1998), 51–78.

[10] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. II, second ed., Wiley, New York, 1971.

[11] S. A. Grishechkin, Single-channel system with circular access or processor-sharing and branching processes. *Math. Notes of Acad. Sci. USSR* **44** (1988), 716–724.

[12] S. A. Grishechkin, On connection between the theory of branching processes and queueing theory. *Probability Theory and Mathematical Statistics, Proceedings of the Fifth Vilnius Conference*, vol. I, B. Grigelionis, Yu. V. Prohorov, V. V. Sanonov and V. Statulevičius (eds.), MOKSLAS, Vilnius, Lithuania, 1990, pp. 455–462.

[13] S. A. Grishechkin, Crump-Mode-Jagers branching processes as a method of investigating M/G/1 systems with processor sharing. *Theor. Probability Appl.* **36** (1991), 19–35.

[14] S. A. Grishechkin, On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes. *Adv. Appl. Prob.* **24** (1992), 653–698.

[15] M. K. Hui and D. K. Tse, What to Tell Customers in Waits of Different Lengths: An Integrative Model of Service Evaluation. *J. Marketing* **60** (1996) 81–90.

[16] A. Jean-Marie and P. Robert, On the transient behavior of the processor sharing queue, *Queueing Systems* **17** (1994), 129–136.

[17] K. L. Katz, B. M. Larson and R. C. Larson, Prescription for the Waiting-in-Line Blues: Entertain, Enlighten and Engage. *Sloane Management Review* **32** (1991) 44–53.

[18] F. P. Kelly, *Reversibility and Stochastic Networks*, Wiley, New York, 1979.

[19] M. Yu. Kitaev, The $M/G/1$ processor-sharing model: transient behavior. *Queueing Systems* **14** (1993), 239–274.

[20] M. Yu. Kitaev and S. F. Yashkov, Distribution of the conditional sojourn time in a system with division of time of servicing. *Engrg. Cybernetics* **16** (1978), 162–167.

[21] T. J. Ott, The sojourn-time distribution in the $M/G/1$ queue with processor sharing, *J. Appl. Prob.* **21** (1984), 360–378.

[22] A. K. Parekh and R. G. Gallager, A generalized processor sharing approach to flow control in integrated services networks: the single node case. *IEEE/ACM Trans. Networking* **1** (1993), 344–357.

[23] B. Sengupta and D. L. Jagerman, A conditional response time of the $M/M/1$ processor-sharing queue, *AT&T Tech. J.* **64** (1985), 409–421.

[24] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, Wiley, Chichester, 1983.

[25] S. Taylor, Waiting for Service: The Relationship Between Delays and Evaluations of Service. *J. Marketing* **58** (1994) 56–69.

[26] W. Whitt, Improving service by informing jobs about anticipated delays. *Management Sci.* **45** (1999), 192–207.

[27] W. Whitt, Predicting queueing delays, *Management Sci.* **45** (1999), 870–888.

[28] W. Whitt, Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Oper. Res. Letters* **24** (1999), 205–212.

[29] S. F. Yashkov, A derivation of response time distribution for an $M/G/1$ processor-sharing queue, *Problems of Control and Information Theory* **12** (1983), 133–148.

[30] S. F. Yashkov, Processor-sharing queues: some progress in analysis. *Queueing Systems* **2** (1987), 1–18.

[31] S. F. Yashkov, Mathematical problems in the theory of shared-processor systems, *Journal of Soviet Math.* **58** (1992), 101–147.

[32] A. P. Zwart and O. J. Boxma, Sojourn time asymptotics in the $M/G/1$ processor sharing queue. CWI, Amsterdam, 1999.