J. Appl. Prob. **50,** 1044–1056 (2013)
Printed in England
© *Applied Probability Trust* 2013

SOJOURN TIME ESTIMATION IN AN M/G/∞ QUEUE WITH PARTIAL INFORMATION

NAFNA BLANGHAPS,*
YUVAL NOV * ** AND
GIDEON WEISS,* *University of Haifa*

Abstract

We propose an estimator for the cumulative distribution function G of the sojourn time in a steady-state M/G/∞ queueing system, when the available data consists of the arrival and departure epochs alone, without knowing which arrival corresponds to which departure. The estimator generalizes an estimator proposed in Brown (1970), and is based on a functional relationship between G and the distribution function of the time between a departure and the rth latest arrival preceding it. The estimator is shown to outperform Brown's estimator, especially when the system is heavily loaded.

Keywords: M/G/∞; sojourn time estimation; Smoluchowski process; semiparametric estimation

2010 Mathematics Subject Classification: Primary 62M09
Secondary 90B22

1. Introduction

An M/G/∞ queue is a system with no actual queueing: items arrive in a Poisson stream at rate λ and are processed immediately, as there is no limit to the number of servers: so their sojourn times consist of their processing times alone. The sojourn times are assumed to have a common cumulative distribution function (CDF) G, and to be independent of each other and of the arrival process.

While we formulate our analysis in queueing-theoretic terms, the study of such systems antedates any queueing studies. One is often interested in estimating sojourn times in a changing population, in which items arrive in a Poisson stream and stay for a random time. An early example in physics is the Einstein–Smoluchowski theory, as developed by Smoluchowski (1906), (1916), which explains the Brownian motion of colloidal particles in suspension and was used to measure the diffusion coefficient. The counting process of these particles came to be known as the Smoluchowski process, which is none other than the M/G/∞ process; see the reviews Chandrasekhar (1943) and Kac (1959), and, more recently, Bingham and Dunham (1997). Examples in medicine are measurements such as the mobility of spermatozoa (see Rothschild (1953), Ruben (1963), and Lindley (1956)) and the mobility of white blood cells (see Brenner *et al.* (1978)). Further examples include counting animals in feeding grounds (see Duffey and Watt (1971)), cars on a stretch of road (see Brown (1970)), and messages in transit in a communication system (see Ayesta and Mandjes (2009)).

From the statistical point of view, an M/G/∞ system gives rise to a semiparametric estimation problem, in which one wishes to estimate both the infinite-dimensional CDF G and the single-

Received 29 March 2012; revision received 25 December 2012.

* Postal address: Department of Statistics, University of Haifa, Haifa, 31905, Israel.
** Email address: yuval@stat.haifa.ac.il

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

dimensional parameter λ. Given a complete realization of the system over a finite time interval, one can easily estimate G nonparametrically from direct observations of the sojourn times, and estimate λ from the arrival process. However, in an M/G/∞ system it is often difficult or impossible to keep track of each item from arrival to departure, and, thus, to observe its exact sojourn time. One may distinguish between three types of partial information that might be available in such a setting: (i) the queue-length process $\{Q_t\}$ (from which the arrival and departure epochs can be observed, but not their matching, and, hence, also not the sojourn times), (ii) the 'busy-period' process $\{I_{(Q_t>0)}\}$, which indicates only whether the system is empty or not, and (iii) the arrival and departure epochs alone (again, without their matching). Note that data of type (i) are more informative than those of types (ii) and (iii), in the sense that data of both types (ii) and (iii) can be reconstructed from data of type (i), but not vice versa. The two types of data (ii) and (iii) are not ordered in their informativeness, in the sense that neither can be reconstructed from the other.

The study of queueing systems with such partial information dates back at least to Parzen (1962). Later works focused more on statistical inference. Brown (1970) studied an M/G/∞ system with information of type (iii), and proposed a method for estimating G; more on this method below. Pickands and Stine (1997) derived two methods for estimating G and λ in a discrete-time M/G/∞ queue, under information of type (i). Bingham and Pitts (1999) studied the same problem, and also considered estimation under information of type (ii). Hall and Park (2004), and later Park (2007), proposed a kernel-based deconvolution estimator for the density of the processing time, under information of type (ii). In a recent related work, Grübel and Wegener (2011) derived a method for matching arrivals and departures in an M/G/∞ queue, when the distribution G is known and the available data are the order statistics of the arrivals and departures.

In this work we generalize the estimator proposed in Brown (1970) for the CDF G of the sojourn time in a steady-state M/G/∞ queue, under information of type (iii), i.e. when the arrival and departure epochs are observed without knowing which arrival corresponds to which departure. A concrete example for a situation in which such an estimator is required was given in Brown (1970): one may wish to estimate the distribution of time vehicles spend on a highway, given only the entrance and exit times.

We first summarize Brown's result. Let Y_i be the time of the ith departure, W_i the sojourn time ending with the ith departure (assumed to have a finite mean), Z_i the time between Y_i and the latest arrival preceding it, and H the CDF of Z_i. Note that, under the assumed available information, namely type (iii), the W_i are not observable, whereas the Z_i are observable. By conditioning on the number of arrivals in the time interval $(Y_i - W_i, Y_i)$, Brown derived the conditional CDF H given W_i as follows:

$$H_{Z\mid W}(z \mid w) = \mathbb{P}(Z_i \le z \mid W_i = w) = \begin{cases} 1 - \mathrm{e}^{-\lambda z}, & z < w, \\ 1, & z \ge w. \end{cases}$$

(Brown's conditioning on the number of arrivals is not necessary: see our proof of Proposition 1, below, for an alternative approach for proving a more general version of the last equality.) Integrating over w with respect to the sojourn time distribution G, Brown obtained the functional relationship $H(z) = 1 - (1 - G(z))\mathrm{e}^{-\lambda z}$, which may be inverted to yield

$$G(x) = 1 - (1 - H(x))\mathrm{e}^{\lambda x}. \tag{1}$$

By observing the arrival and departure times in a time interval spanning n departures, one may construct an estimator λ_n for λ, and also an 'empirical' estimator $H_n(x) = (1/n)\sum_{i=1}^{n} \mathbf{1}_{\{Z_i \le x\}}$

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

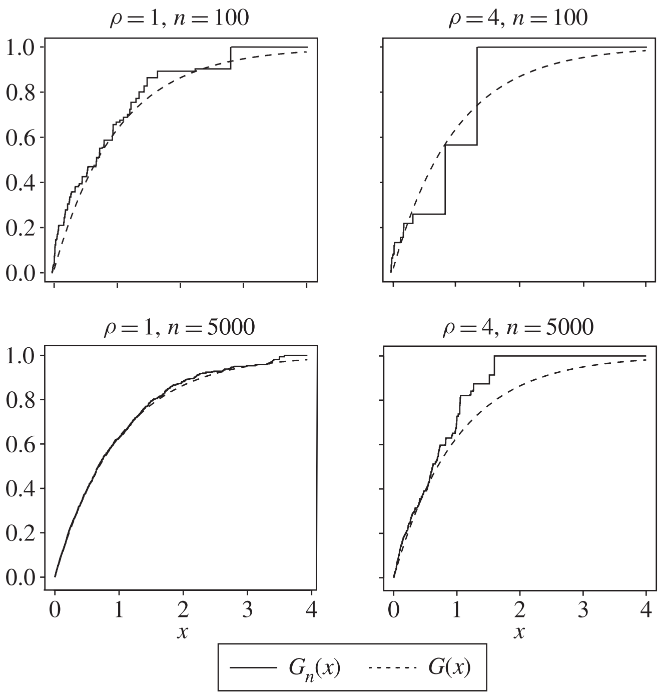

FIGURE 1: The true CDF G and a typical realization of Brown's estimator G_n, Equation (3), under moderate load (ρ = arrival rate/processing rate = 1) in the left-hand panels, and under heavier load ($\rho = 4$) in the right-hand panels. Note how the estimates $G_n(x)$ for large values of x are less accurate when the load is higher. Here, G is exponential with rate 1 and the sample size is $n = 100$ (*upper panels*) or $n = 5000$ (*lower panels*).

for H. Substituting H_n and λ_n for H and λ into (1) yields the following estimator for G:

$$V_n(x) = 1 - (1 - H_n(x))e^{\lambda_n x}. \tag{2}$$

This estimator, however, is not always monotonically increasing, and to rectify this Brown defined

$$G_n(x) = \sup_{y \le x} V_n(y), \tag{3}$$

which was shown to converge uniformly and almost surely to G, in the sense that

$$\mathbb{P}\Big(\lim_{n\to\infty} \sup_x |G(x) - G_n(x)| \to 0\Big) = 1.$$

As Brown acknowledged, the estimator G_n in (3) is 'clearly not the best estimator in any sense' (see Brown (1970), page 653). In particular, this G_n essentially estimates G only at the points $Z_1, Z_2, \dots, Z_n$, through $G_n(Z_1), G_n(Z_2), \dots, G_n(Z_n)$, whereas we would like to estimate G at the more 'typical' points $W_1, W_2, \dots, W_n$ (recall that G is the CDF of the W_i). Under heavy load—i.e. when the sojourn times W_i are large relative to the interarrival times—the Z_i become small relative to the W_i, so inference regarding $G(x)$ for large x becomes unreliable; see Figure 1.

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

In this work we generalize Brown's estimator by deriving a functional relationship between G and the CDF of the time between a departure and the rth latest arrival preceding it. The resulting estimator for G coincides with Brown's estimator when $r = 1$. By choosing the 'right' r, we can estimate $G(x)$ more reliably than with Brown's estimator, for a wider range of x values.

2. Estimating G via rth-latest arrivals for fixed r

Consider a steady-state M/G/∞ queueing system with arrival rate λ and sojourn time CDF G having mean $m_G = \int_0^\infty x \,\mathrm{d}G(x) < \infty$. Let ρ be the load on the system, defined as

$$\rho = \frac{\text{arrival rate}}{\text{processing rate}} = \frac{\lambda}{1/m_G} = \lambda m_G.$$

The arrivals and departures to and from the system are observed along the time interval $[0, T]$, which includes n departures. Let Y_i be the time of the ith departure, W_i the true (but unobservable) sojourn time corresponding to the ith departure, $Z_i^{(r)}$ the time between Y_i and the rth nearest arrival to the left of Y_i, and $H^{(r)}$ the CDF of the $Z_i^{(r)}$; see Figure 2.

Proposition 1. *The conditional CDF of $Z_i^{(r)}$ given $W_i = w$ is*

$$\mathbb{P}(Z_i^{(r)} \le z \mid W_i = w) = \begin{cases} 1 - \displaystyle\sum_{j=0}^{r-1} \frac{\mathrm{e}^{-\lambda z}(\lambda z)^j}{j!}, & z < w, \\ 1 - \displaystyle\sum_{j=0}^{r-2} \frac{\mathrm{e}^{-\lambda z}(\lambda z)^j}{j!}, & z \ge w. \end{cases} \tag{4}$$

Proof. The event $\{Z_i^{(r)} \le z\}$ is exactly the event 'there were at least r arrivals in the time interval $[Y_i - z, Y_i)$'. Since we condition on the event $\{W_i = w\}$, we distinguish between two cases. If $z < w$, the (known) arrival at $Y_i - w$ is not included in the interval $(Y_i - z, Y_i)$, and since the arrivals form a Poisson process with rate λ, the probability of having at least r of them in the interval is $1 - \sum_{j=0}^{r-1} \mathrm{e}^{-\lambda z}(\lambda z)^j/j!$. If $z \ge w$, the arrival at $Y_i - w$ is included in the interval, so to have a total of at least r arrivals, we need at least $r - 1$ additional arrivals in the interval; by the lack-of-memory property of the Poisson process, these additional arrivals still constitute a Poisson process, so the probability of having at least $r - 1$ of them in that interval is $1 - \sum_{j=0}^{r-2} \mathrm{e}^{-\lambda z}(\lambda z)^j/j!$ (here and below, summation from 0 to -1 in the case $r = 1$ is to be interpreted as the empty sum, that is, as 0).

Proposition 2. *For each r, the following functional relationship between G and $H^{(r)}$ holds:*

$$G(x) = 1 - \frac{1 - H^{(r)}(x) - \sum_{j=0}^{r-2} \mathrm{e}^{-\lambda x}(\lambda x)^j/j!}{\mathrm{e}^{-\lambda x}(\lambda x)^{r-1}/(r-1)!}. \tag{5}$$

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

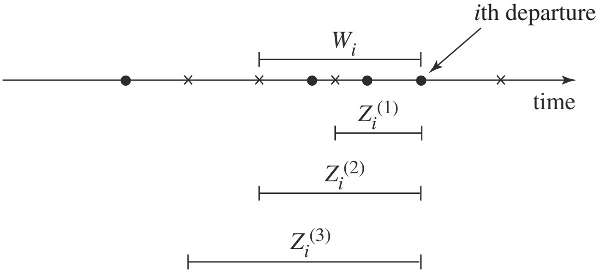

FIGURE 2: The random variables $Z_i^{(1)}$, $Z_i^{(2)}$, $Z_i^{(3)}$, and W_i in a realization in which there was exactly one arrival during the processing of the ith departing item. Arrivals are denoted by crosses and departures are denoted by filled circles.

Proof. Using the law of total probability, we integrate (4) over w with respect to the sojourn time distribution G to derive the unconditional CDF of $Z_i^{(r)}$:

$$\begin{aligned} H^{(r)}(z) &= \mathbb{P}(Z_i^{(r)} \le z) \\ &= \int_0^\infty \mathbb{P}(Z_i^{(r)} \le z \mid W_i = w)\,\mathrm{d}G(w) \\ &= \int_0^z \left(1 - \sum_{j=0}^{r-2} \frac{\mathrm{e}^{-\lambda z}(\lambda z)^j}{j!}\right) \mathrm{d}G(w) + \int_z^\infty \left(1 - \sum_{j=0}^{r-1} \frac{\mathrm{e}^{-\lambda z}(\lambda z)^j}{j!}\right) \mathrm{d}G(w) \\ &= 1 - \frac{(1-G(z))\mathrm{e}^{-\lambda z}(\lambda z)^{(r-1)}}{(r-1)!} - \sum_{j=0}^{r-2} \frac{\mathrm{e}^{-\lambda z}(\lambda z)^j}{j!}. \end{aligned}$$

Inverting this relationship between $H^{(r)}(z)$ and $G(z)$, the proposition is proved. This equality is the counterpart of (1).

Proposition 3. *For each r, the sequence $Z_1^{(r)}, Z_2^{(r)}, \ldots$ is stationary and ergodic.*

Proof. The proof follows the same steps as its counterpart in Brown (1970). Stationarity is immediately established from the assumption that the queue is in steady state. To prove ergodicity, it needs to be shown that every invariant event of the sequence $\{Z_i^{(r)}\}$ has probability 0 or 1. By Proposition 6.32 of Breiman (1992), every invariant event of $\{Z_i^{(r)}\}$ is a tail event, so it is sufficient to show that all tail events of $\{Z_i^{(r)}\}$ have probability 0 or 1. This will follow if we show that, for any fixed m, the random variables $Z_1^{(r)}, \ldots, Z_m^{(r)}$ are independent of the tail σ-field of $\{Z_i^{(r)}\}$. Since $m_G < \infty$, there are finitely many customers at time Y_m (the time of the mth departure) with probability 1, so that there exists a time $t > Y_m$ by which all these customers have departed from the queue. All Z_i corresponding to departures subsequent to the rth arrival after t are no longer a function of $Z_1^{(r)}, \ldots, Z_m^{(r)}$. Thus, $Z_1^{(r)}, \ldots, Z_m^{(r)}$ are independent of the tail σ-field of $\{Z_i^{(r)}\}$, and the proposition is proved.

Let $H_n^{(r)}(x) = (1/n)\sum_{i=1}^n \mathbf{1}_{\{Z_i^{(r)} \le x\}}$ be the empirical estimator of $H^{(r)}$, and let λ_n be a strongly consistent estimator of λ, say the usual maximum likelihood estimator (the inverse of the mean interarrival time). By substituting $H_n^{(r)}$ and λ_n for H and λ into (5), we define the counterpart of V_n from (2) as follows:

$$V_n^{(r)}(x) = 1 - \frac{1 - H_n^{(r)}(x) - \sum_{j=0}^{r-2} \mathrm{e}^{-\lambda_n x}(\lambda_n x)^j/j!}{\mathrm{e}^{-\lambda_n x}(\lambda_n x)^{r-1}/(r-1)!}.$$

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

The function $V_n^{(r)}$ still cannot serve as a reasonable estimator of G for two reasons: it may assume values outside the interval [0, 1] (this is true only for $r \geq 2$), and, in general, it is not monotonically increasing. We can deal with values outside [0, 1] either by discarding them, or by replacing values above 1 with 1, and values below 0 with 0. Simulation results (as in Nelgabatz (2012)) seem to indicate that the first approach yields slightly better results, so we adopt it henceforth.

One way to build a monotonically increasing function from $V_n^{(r)}$ is to follow Brown and define

$$\tilde{G}_n^{(r)}(x) = \sup_{y \leq x} V_n^{(r)}(y). \tag{6}$$

This function is a strongly and uniformly consistent estimator of G, as proved in the following theorem.

Theorem 1. *For each r, the estimator $\tilde{G}_n^{(r)}$ in (6) converges to G uniformly and almost surely as $n \to \infty$, in the sense that*

$$\mathbb{P}\Big(\lim_{n\to\infty} \sup_x |G(x) - \tilde{G}_n^{(r)}(x)| \to 0\Big) = 1.$$

Proof. The proof again follows the same steps as its counterpart in Brown (1970). Since the sequence $\{Z_i^{(r)}\}$ is stationary and ergodic (see Proposition 3), both $H_n^{(r)}(x)$ and $H_n^{(r)}(x-)$ converge to $H^{(r)}(x)$ and $H^{(r)}(x-)$, respectively, almost surely for all x, and because $H_n^{(r)}$ and $H^{(r)}$ are monotone and bounded the convergence is uniform. This implies that $V_n^{(r)}$ converges to G uniformly on compact intervals; hence, $\sup_{y\leq x} V_n^{(r)}(y)$ converges to $\sup_{y \leq x} G(y) = G(x)$, i.e. $\tilde{G}_n^{(r)}(x)$ converges to $G(x)$ for every x, and similarly for $x-$. This implies by monotonicity and boundedness that $\tilde{G}_n^{(r)}$ converges to G uniformly, almost surely.

An alternative way of eliminating the nonmonotonicity problem is to use isotonic regression, i.e. replace the $V_n^{(r)}(Z_i^{(r)})$ values with the closest possible values (under a quadratic loss function) that are monotonically increasing. More specifically, we let $G_n^{(r)}$ be a piecewise-constant function, with possible jumps at the points $Z_i^{(r)}$, such that

$$G_n^{(r)}(Z_i^{(r)}) = g_i, \tag{7}$$

where the g_i are the solution of

$$\operatorname*{minimize}_{g_1,\dots,g_n} \sum_{i=1}^{n} (g_i - V_n^{(r)}(Z_i^{(r)}))^2, \quad \text{such that } g_i \leq g_j \text{ whenever } Z_i^{(r)} \leq Z_j^{(r)}. \tag{8}$$

This quadratic programming problem has a convex objective function and its linear constraints define a convex, closed feasibility region; hence, the problem always possesses a unique solution (see Robertson *et al.* (1988)). Furthermore, the problem can be efficiently solved using the PAVA algorithm, as described, for example, in De Leeuw *et al.* (2009).

Simulation results (as in Nelgabatz (2012)) indicate overwhelmingly that the isotonic estimator $G_n^{(r)}$ defined via (7) and (8) is superior to the supremum estimator $\tilde{G}_n^{(r)}$ defined in (6). Therefore, we henceforth use the isotonic estimator alone. While we believe that the isotonic estimator is consistent, it seems that proving this is far from simple. The difficulty is that whereas the supremum estimator of $G(x)$ depends only on observations in the compact interval $[0, x]$, the isotonic estimator of $G(x)$ depends on all the observations in the range $[0, \infty)$, so we cannot use uniform convergence of $V_n^{(r)}$ to G on compact sets.

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

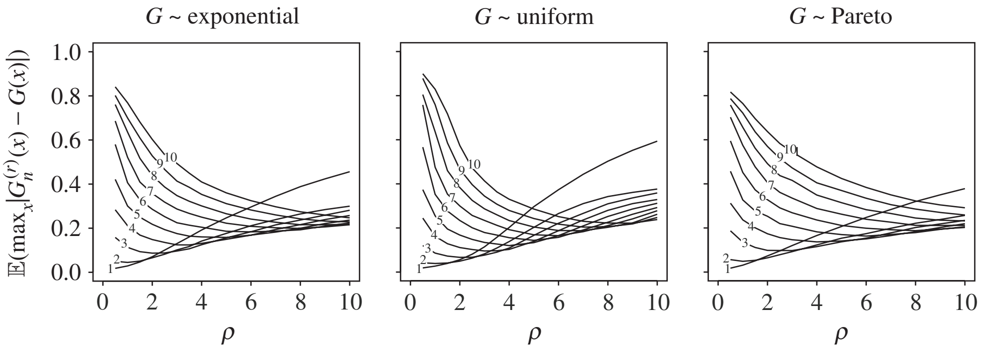

FIGURE 3: $\mathbb{E}(\sup_x |G_n^{(r)}(x) - G(x)|)$ as a function of the load ρ, as estimated by simulation. The ten curves in each panel correspond to $r = 1, \dots, 10$. The processing time distribution G is exponential (*left panel*), uniform (*middle panel*), and Pareto (*right panel*). The sample size $n = 5000$. See Section 4 for simulation details.

As a criterion for judging the quality of the estimator, we use the mean L_∞ error,

$$\mathbb{E}(\sup_x |G_n^{(r)}(x) - G(x)|),$$

which we estimate via simulation. This criterion is closely related to the statistic of the Kolmogorov–Smirnov test (see Massey (1951)), which is probably the most widely used statistical test for comparing two distributions. This is also a highly stringent criterion, and indeed, as shown below, the sample size n has to be rather large in order to obtain reasonable estimates in some settings. In Figure 3 we show the performance of the estimator as a function of the load ρ for three types of G—exponential, uniform, and Pareto—and for $r = 1, \dots, 10$. The most prominent feature of Figure 3 is that, for each fixed r, the quality of the estimator varies greatly with ρ, and that the different curves of $\mathbb{E}(\max_x |G_n^{(r)}(x) - G(x)|)$ for different r are very different from each other. As a result, different rs are optimal for different values of ρ, and the larger ρ, the larger the optimal r. This observation is in line with Figure 1, which showed that, when ρ is large, using $r = 1$ yields a poor estimate.

3. Estimating *G* via *r*th-latest arrivals for adaptively chosen *r*

The values of the performance criterion $\mathbb{E}(\max_x |G_n^{(r)}(x) - G(x)|)$, as shown in Figure 3, are a rather complex function of G, λ, n, and r (the dependence on λ is suppressed in our notation, and the dependence on the observations $Z_1, \dots, Z_n$ is integrated out through the expectation operator). Of these factors, G and λ are exogenously given, n is dictated by the data collection constraints, and so r is the only factor that may be chosen at the statistical analysis stage. Ideally, we would use the optimal r, which minimizes the performance criterion given the G, λ, and n at hand:

$$r^* = \arg\min_r \mathbb{E}\Big(\max_x |G_n^{(r)}(x) - G(x)|\Big). \tag{9}$$

However, while n is always known in advance and λ may be easily and accurately estimated from the data, it is the unknown G which complicates the choice of the optimal r (indeed, the very purpose of this work is to estimate G).

A simpler approach is to base the choice of r not on estimates of λ and G, but rather, on the load ρ, which summarizes the relationship between them. Recall that it was indeed the change

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

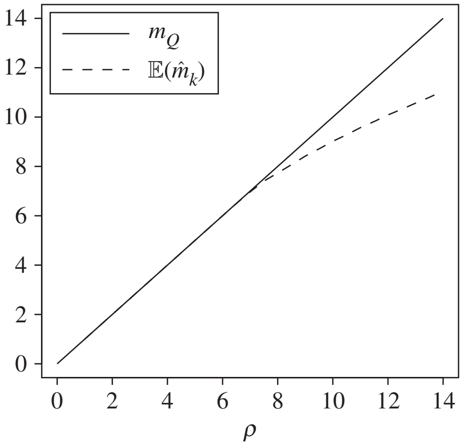

FIGURE 4: Mean queue length m_Q and $\mathbb{E}(\hat{m}_K)$ as a function of ρ, as estimated from simulation. The distribution G is exponential and $n = 5000$.

in the load which rendered Brown's original estimator unsatisfactory (see again Figure 1), and that ρ is the horizontal axis of the graphs in Figure 3, which allowed us to study how the performance of the estimator varies with r. While it is difficult to estimate ρ directly, we will calculate a statistic closely related to ρ, use it to estimate ρ, and only then choose r, as explained below.

Let m_Q be the mean queue length (recall that the system is assumed to be in steady state, so that this mean is constant in time). From Little's formula (see Ross (2010)), it is known that $m_Q = \lambda m_G$, so $\rho = \lambda m_G = m_Q$. To estimate ρ, then, we could use an estimate of m_Q. If the queue-length process $\{Q_t\}$ were observable over the sampling period $[0, T]$, we could use $\hat{m}_Q = (1/T)\int_0^T Q_t \, \mathrm{d}t$ for this purpose. However, under our assumptions, only the arrival and departure epochs are observable, so this approach may not be pursued. Still, we could use a similar approach. Let $\{L_t, \ 0 \le t \le T\}$ be a process that starts at $L_0 = 0$, and then increases by 1 with each arrival and decreases by 1 with each departure; this process can be constructed from the observed data. Next, define another process, $\{K_t, \ 0 \le t \le T\}$, via $K_t = L_t - \min_{0 \le s \le T} L_s$, and let $\hat{m}_K = (1/T)\int_0^T K_t \, \mathrm{d}t$. The process $\{K_t\}$ coincides with $\{Q_t\}$ if the system becomes empty at least once during the sampling period, in which case $\hat{m}_K = \hat{m}_Q$ also; in general,

$$K_t = Q_t - \min_{0 \le s \le T} Q_s \le Q_t, \qquad 0 \le t \le T \tag{10}$$

(the larger the sample size n, the higher the probability of $\{K_t\}$ coinciding with $\{Q_t\}$). As the load increases, $\hat{m}_Q$ grows in mean linearly (this is because $\hat{m}_Q$ is an unbiased estimator of m_Q, so $\mathbb{E}(\hat{m}_Q) = \mathbb{E}((1/T)\int_0^T Q_t \, \mathrm{d}t) = (1/T)\int_0^T \mathbb{E}(Q_t) \, \mathrm{d}t = (1/T)\int_0^T m_Q \, \mathrm{d}t = m_Q = \rho$). From (10), it follows that $\mathbb{E}(\hat{m}_K) \le \rho$. Although the difference in mean between $\hat{m}_Q$ and $\hat{m}_K$ increases with ρ, $\hat{m}_K$ still holds information on $\hat{m}_Q$; see Figure 4. Let h be the function that relates ρ to $\hat{m}_K$ (for fixed n) via $h(\rho) = \mathbb{E}(\hat{m}_K)$, and let h^{-1} be its inverse.

We now propose the following approach for selecting $\hat{r}^*$, the estimate of the optimal r^* from (9). Given only n, we use data such as those presented in the left-hand panel of Figure 3, to partition $[0, \infty)$ into intervals separated by cutoff points $0 = c_0 < c_1 < c_2 < \cdots$, such that $r^* = i$ when $\rho \in [c_{i-1}, c_i]$ (i.e. when ρ is in the ith interval, the optimal r is i); see Table 1. Using data such as those presented in Figure 4, we estimate h^{-1} by $\hat{h}^{-1}$. These are

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

TABLE 1: The first ten cutoff points c_i used to estimate r^* for a sample size $n = 5000$.

c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
0	1.8	2.8	4.5	6.0	7.3	8.1	9.1	9.8	10.7

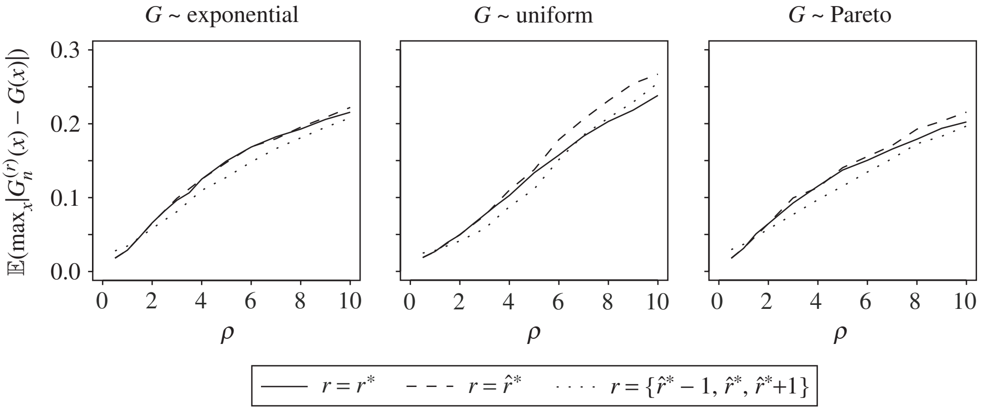

FIGURE 5: $\mathbb{E}(\sup_x |G_n^{(r)}(x) - G(x)|)$ as a function of the load ρ when r is the optimal r^*, the adaptively chosen $\hat{r}^*$, and a combination of $\hat{r}^* - 1$, $\hat{r}^*$, and $\hat{r}^* + 1$. The simulation parameters are as in Figure 3.

two preliminary steps that are done once (for each n), and are accomplished using simulation results only; no real data sample is used up to this point. Next, given the arrival and departure epoch data, we first compute $\hat{m}_K$ as explained above, and then let $\hat{r}^* = i$ for the i satisfying $h^{-1}(\hat{m}_K) \in [c_{i-1}, c_i]$. The final estimator of G is $G_n^{\hat{r}^*}$.

The left panel of Figure 5 shows how the estimates based on this procedure perform when G is exponential. The solid curve is the value of $\mathbb{E}(\max_x |G_n^{(r)}(x) - G(x)|)$ using the optimal $r = r^*$ (i.e. the pointwise minimum of the curves from Figure 3), and the dashed curve consists of the $\mathbb{E}(\max_x |G_n^{(r)}(x) - G(x)|)$ values obtained using the adaptively chosen $r = \hat{r}^*$, as described above. The two curves follow each other very closely, suggesting that the adaptive procedure succeeds, at least in the exponential case, to select good rs.

The cutoff points c_i and the function $\hat{h}^{-1}$ were estimated while assuming an exponential G, yet they produce reasonable estimates also when G departs considerably from exponentiality. In the middle and right panels of Figure 5, G is uniform and Pareto, respectively, but the c_i and $\hat{h}^{-1}$ are exactly those used in the left panel, in which G was exponential. The estimation results using the adaptively chosen $\hat{r}^*$ (dashed curves) still do not depart considerably from those obtained using the optimal r^* (solid curve) for the true G (uniform or Pareto).

The estimator may be further improved, as shown by the third, dotted curve appearing in each of the panels of Figure 5. This curve corresponds to an estimator that uses several rs at the same time. Recall that the input to the isotonic regression procedure was the set of pairs $\{Z_i^{(r)}, G_n^{(r)}(Z_i^{(r)})\}$ for some r. One may consider several such sets of pairs corresponding to several rs, and run the isotonic regression over their union. The dotted curves in Figure 5 are the results of such a procedure, when using a combination of three rs: the $\hat{r}^*$ selected using the adaptive procedure described above, and, in addition, $\hat{r}^* + 1$ and $\hat{r}^* - 1$ (the latter omitted when $\hat{r}^* = 1$). The resulting estimates indeed improve upon those obtained using $\hat{r}^*$ alone, and, for many values of ρ, also improve upon those obtained using r^*.

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

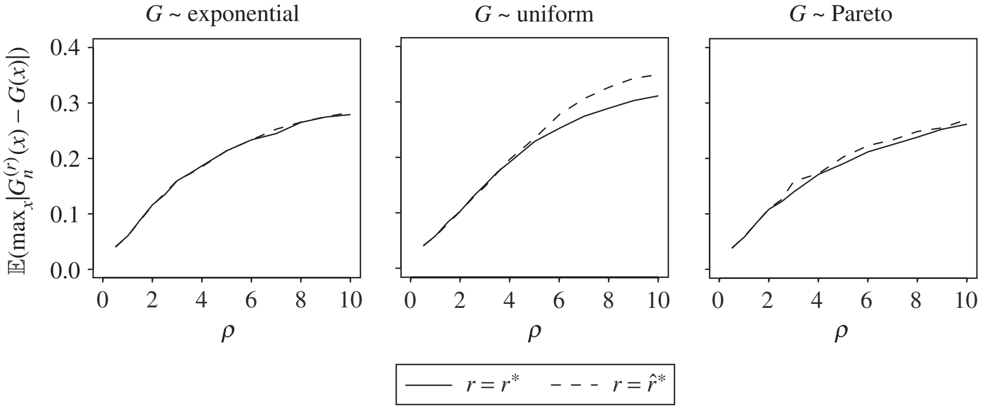

FIGURE 6: $\mathbb{E}(\sup_x |G_n^{(r)}(x) - G(x)|)$ as a function of the load ρ for a sample size $n = 1000$, where the adaptively chosen $\hat{r}^*$ values are based on the cutoff points $\{c_i\}$ derived for $n = 5000$ (see Table 1).

In principle, the cutoff points $\{c_i\}$ need to be estimated anew for each sample size n, at the preliminary simulation stage. However, the estimation results seem to depend rather weakly on the choice of n at that stage: in Figure 6 we show again $\mathbb{E}(\max_x |G_n^{(r)}(x) - G(x)|)$ as a function of ρ, this time for $n = 1000$; in each of the three panels, the solid curve is based on the optimal $r = r^*$ for this new n, but the dashed curve corresponds to the adaptively chosen $r = \hat{r}^*$, based on the $\{c_i\}$ values derived for $n = 5000$ (see Table 1). Again, the two curves are rather close to each other.

4. A note on simulation

All simulations were carried out using the R programming language (see www.r-project.org). Each estimated point is based on 1000 simulation replications.

The arrival rate was kept constant at $\lambda = 1$ in all cases. To vary the load ρ, the sojourn time distribution G was varied as follows. When G is $\exp(\eta)$ or $U(0, b)$, we set $\eta = 1/\rho$ or $b = 2\rho$, respectively, to induce the desired load. For a Pareto G, we used a distribution with the density

$$f(x) = 3\beta(\beta x + 1)^{-4}, \qquad x \geq 0.$$

This is a slightly modified version of the 'plain' Pareto distribution having only first and second finite moments (i.e. with density $f(x) = 3x^{-4}$, $x \geq 1$): it is shifted by a unit to the left, so that its support is $[0, \infty)$, and is further parameterized by the scale parameter β, so that its mean is $1/2\beta$. Thus, we set $\beta = 1/(2\rho)$ to induce the desired load.

To let the system approach steady state, the queue was started empty, and then run for 100 arrivals before sampling began.

Further simulation results and statistical analysis of the performance of our estimators in the estimation of $G(x)$ at individual points x are given in Nelgabatz (2012).

5. Discussion

In this paper we proposed an estimator for the CDF G of the sojourn time in an M/G/∞ queueing system, when only the arrival and departure epochs are known, without their pairing. To the best of our knowledge, the only other published work that deals with this exact problem is Brown (1970); indeed, our proposed method is a generalization of Brown's, aimed at correcting

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

one of its shortcomings. The data for Brown's method, as well as for ours, are the time intervals between departures and previous arrivals; Brown used the most recent previous arrival, whereas our generalization uses the rth most recent arrival. The intuition behind our method is the following. An observed time interval of length x provides information about the sojourn time distribution at the point x, that is, it helps in estimating $G(x)$. In Brown's method, the observations are typically around $1/\lambda$ (the average interarrival time), so when λ is large, we obtain information about $G(x)$ mostly for small values of x in the support of G. Our method corrects for that, since our r-delayed observations are typically around r/λ, and so by choosing the 'right' r, we can estimate $G(x)$ for a wider and more representative range of x values. As can be seen from Figures 3 and 5, our method improves greatly upon Brown's estimate for a vast range of parameter regimes (namely, large ρ).

We note, however, that when ρ is too large, the improved method we propose still produces unreliable estimates of G—say, with expected maximal deviation of more than 0.2. For the specific sample size $n = 5000$ which we studied, it seems that ρ needs to be below 6 to produce reasonable estimates. The reason for this is that when we observe only the arrival and departure epochs, without their pairing, large values of ρ correspond to a very high noise level. We might say that the values of the sojourn times about which we wish to learn get almost completely hidden by the noise when ρ is large. While our estimator seems to be consistent, we conjecture that the rate of convergence for the metric of expected maximal deviation is slow. An interesting question is whether we can obtain lower bounds on these rates of convergence, and see how they compare to what our estimator achieves.

Some of the literature reviewed in Section 1 estimated G via the busy-period process $\{I_{(Q_t>0)}\}$ (see Bingham and Pitts (1999), Hall and Park (2004), and Park (2007)). A necessary condition for obtaining even the crudest estimate from such methods is to have more than one busy period in the sampling period; to obtain a reasonable estimate, quite a few busy periods are required. Recall that the event 'there was more than one busy period' is contained in the event 'the system became empty at least once', which is exactly the event in which $\hat{m}_K = \hat{m}_Q$, and, thus, corresponds to a relatively low ρ with which our proposed method performs well. The expected length of a busy period of an M/G/∞ queue is $(1/\lambda)(\mathrm{e}^{\rho} - 1)$, so, for $\rho = 6$, the average busy period has approximately 400 arrivals and departures, and, hence, $n = 5000$ corresponds to approximately only 12 busy periods. Thus, our method still yields reasonable estimates in cases where the busy-period approach is almost completely uninformative.

The approach we propose for selecting r, and as a result, for estimating G, is based on a rule defined while assuming that G is exponential. In reality, we do not know of course in advance whether G is exponential or not (recall that estimating G is the very purpose of the procedure), but as shown in Figure 5, the exponential-based rule yields good estimates also when G is uniform or Pareto. We chose these two distributions as additional test cases because they are very different from the exponential distribution: while the exponential distribution has $[0, \infty)$ as its support and a light tail, the uniform distribution has a finite support, and the Pareto distribution has a heavy tail (we used a Pareto distribution with only two finite moments, i.e. one whose probability distribution function's tail is $O(x^{-4})$). While it is probably possible to devise 'pathological' CDFs G under which the exponential-based rule yields poor estimates, we believe that the three types of G we considered span a wide enough range of reasonable processing time distributions. Notwithstanding the last remark, it might be both theoretically and practically useful to develop a preliminary procedure for estimating the tail properties of G (perhaps by deriving the rate of convergence of the estimator), and incorporate the results of this step into the adaptive procedure.

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

Our criterion for judging the quality of the estimators $G_n^{(r)}$ was $\mathbb{E}(\sup_x |G_n^{(r)}(x) - G(x)|)$. We may replace it with alternative criteria: these could be other functionals of the entire distribution G, such as $\mathbb{E}[\int (G_n^{(r)}(x) - G(x))^2 \, \mathrm{d}x]$, or criteria focusing on $G(x)$ for some fixed x of interest, such as the mean-squared error $\mathbb{E}[(G_n^{(r)}(x) - G(x))^2]$. In the latter case, the optimal r^* should depend also on x, with a larger r required for larger values of x. For example, as can be seen from Figure 1, the estimates $G_n^{(1)}(x)$ are reasonable for small x (around typical values of the observations $Z_i^{(1)}$), but become less reliable as x increases; by using larger r, we can improve the estimation of $G(x)$ for such values of x. This means, however, that a separate table of the type of Table 1 will have to be constructed for each x, and in particular, that the optimal r^* given the load is not the same for all criteria.

Our analysis assumed that the system is in steady state. However, since the samples we are considering are rather large, and since the arrival process is stationary from its beginning and the departure process approaches stationarity very quickly, this assumption seems not to be crucial. Simulation of the process starting from an empty system yielded virtually identical results to those based on simulation with the warm-up period (see Section 4).

Acknowledgements

We thank an anonymous referee for constructive comments which significantly improved this paper. The research of Gideon Weiss was supported in part by the Israel Science Foundation grant 711/09.

References

Ayesta, U. and Mandjes, M. (2009). Bandwidth sharing networks under a diffusion scaling. *Ann. Operat. Res.* **170,** 41–58.

Bingham, N. H. and Dunham, B. (1997). Estimating diffusion coefficients from count data: Einstein–Smoluchowski theory revisited. *Ann. Inst. Statist. Math.* **49,** 667–679.

Bingham, N. H. and Pitts, S. M. (1999). Non-parametric estimation for the $M/G/\infty$ queue. *Ann. Inst. Statist. Math.* **51,** 71–97.

Breiman, L. (1992). *Probability*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Brenner, S. L., Nossal, R. J. and Weiss, G. H. (1978). Number fluctuation analysis of random locomotion. Statistics of a Smoluchowski process. *J. Statist. Phys.* **18,** 1–18.

Brown, M. (1970). An $M/G/\infty$ estimation problem. *Ann. Math. Statist.* **41,** 651–654.

Chandrasekhar, S. (1943). Stochastic processes in physics and astronomy. *Rev. Modern Phys.* **15,** 1–89.

De Leeuw, J., Hornik, K. and Mair, P. (2009). Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *J. Statist. Software* **32**.

Duffey, E. and Watt, A. S. (eds) (1971). *The Scientific Management of Animal and Plant Communities for Conservation*. Blackwell, Oxford.

Grübel, R. and Wegener, H. (2011). Matchmaking and testing for exponentiality in the M/G/∞ queue. *J. Appl. Prob.* **48,** 131–144.

Hall, P. and Park, J. (2004). Nonparametric inference about service time distribution from indirect measurements. *J. R. Statist. Soc. B*, **66,** 861–875.

Kac, M. (1959). *Probability and Related Topics in Physical Sciences*. Interscience, London.

Lindley, D. V. (1956). The estimation of velocity distributions from counts. In *Proceedings of the International Congress of Mathematicians, 1954, Amsterdam*, Noordhoff, Groningen, pp. 427–444.

Massey, F. J., Jr. (1951). The Kolmogorov–Smirnov test for goodness of fit. *J. Amer. Statist. Assoc.* **46,** 68–78.

Nelgabatz, N. (2012). Estimation for the service time distribution in an M/G/∞ system with partial information. Master's Thesis, University of Haifa.

Park, J. (2007). On the choice of an auxiliary function in the $M/G/\infty$ estimation. *Comput. Statist. Data Anal.* **51,** 5477–5482.

Parzen, E. (1962). *Stochastic Processes*. Holden-Day, San Fransisco, CA.

Pickands, J., III and Stine, R. A. (1997). Estimation for an $M/G/\infty$ queue with incomplete information. *Biometrika* **84,** 295–308.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. John Wiley,

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press

Chichester.

Ross, S. M. (2010). *Introduction to Probability Models*, 10th edn. Academic Press, Amsterdam.

Rothschild, V. (1953). A new method of measuring the activity of spermatozoa. *J. Experimental Biol.* **30,** 178–199.

Ruben, H. (1963). The estimation of a fundamental interaction parameter in an emigration–immigration process. *Ann. Math. Statist.* **34,** 238–259.

Smoluchowski, M. (1906). Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. *Ann. Physik* **326,** 756–780.

Smoluchowski, M. (1916). Drei Vorträge über Diffusion, Brownschen Bewegung und Koagulation von Kolloidteilchen. *Physik. Z.* **17,** 557–585.

https://doi.org/10.1239/jap/1389370098 Published online by Cambridge University Press