Sojourn Time Estimation in an M/G/∞ Queue with Partial Information

Nafna Nelgabats ∗ Yuval Nov † Gideon Weiss ‡

Abstract

We propose an estimator for the CDF $G$ of the sojourn time in a steady-state M/G/∞ queuing system, when the available data consists of the arrival and departure epochs alone, without knowing which arrival corresponds to which departure. The estimator is a generalization of an estimator proposed by Brown (1970), and is based on a functional relationship between $G$ and the distribution of the time between a departure and the $r$th 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; semi-parametric estimation

2010 Mathematics Subject Classification: 90B22, 62M09

1 Introduction

An M/G/∞ queue is a system with actually no queueing: items arrive in a Poisson stream of rate $\lambda$ 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 no other than the M/G/∞ process; see the reviews by Chandrasekhar (1943)
and Kac (1959), and more recently by Bingham and Dunham (1997). Examples in medicine are
measurements such as the mobility of spermatozoa (Rothschild, 1953; Ruben, 1963; Lindley, 1956)
and the mobility of white blood cells (Brenner et al., 1978). Further examples include counting
animals in feeding grounds (Duffey and Watt, 1970), cars on a stretch of road (Brown, 1970), and
messages in transit in a communication system (Ayesta and Mandjes, 2009).

From the statistical point of view, an M/G/∞ system gives rise to a semi-parametric estimation
problem, in which one wishes to estimate both the infinite-dimensional CDF G and the single-
dimensional parameter λ. Given a complete realization of the system over a finite time interval,
one can easily estimate G non-parametrically 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, Gröbel 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 by 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 by Brown: One wishes 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 $i$th departure, $W_i$ the sojourn time ending with the $i$th 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. By conditioning on the number of arrivals in the time interval $(Y_i - W_i, Y_i)$, Brown derived the conditional CDF $H$ given $W_i$,

$$H_{Z|W}(z \mid w) = P(Z_i \leq z \mid W_i = w) = \begin{cases} 1 - e^{-\lambda z} & z < w \\ 1 & z \geq w. \end{cases}$$

(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 $G$, he obtained the functional relationship

$$H(z) = 1 - (1 - G(z))e^{-\lambda z},$$

which may be inverted to yield

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

By observing the arrival and departure times in a time interval spanning $n$ departure, one may construct an estimator $\lambda_n$ for $\lambda$, and also an “empirical” estimator $H_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Z_i \leq x\}}$ for $H$. Plugging $H_n$ and $\lambda_n$ for $H$ and $\lambda$ in equation (1) yields an estimator for $G$,

$$V_n(x) = 1 - (1 - H_n(x))e^{\lambda_n x}. \quad (2)$$
This estimator, however, is not always monotonically increasing, and to rectify this Brown defined
\[ G_n(x) = \sup_{y \leq x} V_n(y), \tag{3} \]
which was shown to converge uniformly and almost surely to \( G \), in the sense that
\[ P \left( \lim_{n \to \infty} \sup_x |G(x) - G_n(x)| \to 0 \right) = 1. \]

As Brown acknowledges, the estimator \( G_n \) in (3) is “clearly not the best estimator in any sense.” In particular, this \( G_n \) essentially estimates \( G \) only at the points \( Z_1, Z_2, \ldots, Z_n \), through \( G_n(Z_1), G_n(Z_2), \ldots, G_n(Z_n) \), whereas one would like to estimate \( G \) at the more “typical” points \( W_1, W_2, \ldots, W_n \) (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 inter-arrival times — the \( Z_i \)’s become small relative to the \( W_i \)’s, so inference regarding \( G(x) \) for large \( x \) becomes unreliable. See Figure 1.

![Figure 1: The true CDF \( G \) and a typical realization of Brown’s estimator \( G_n \) (equation 3), under moderate load (\( \rho = \) arrival rate/processing rate = 1) in the left panels, and under heavier load (\( \rho = 4 \)) in the right 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).]
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 $r$th latest arrival preceding it. The resulting estimator for $G$ coincides with Brown’s estimator when $r = 1$. By choosing the “right” $r$, one can estimate $G(x)$ more reliably than with Brown’s estimator, for a wider range of $x$ values.

2 Estimating $G$ via $r$th-latest arrivals, fixed $r$

Consider a steady-state M/G/$\infty$ queueing system with arrival rate $\lambda$ and sojourn time CDF $G$ having mean $m_G = \int_0^\infty x \, dG(x) < \infty$. Let $\rho$ 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 $i$th departure, $W_i$ the true (but unobservable) sojourn time corresponding to the $i$th departure, $Z_i^{(r)}$ the time between $Y_i$ and the $r$th nearest arrival to the left of $Y_i$, and $H^{(r)}$ the CDF of the $Z_i^{(r)}$. See Figure 2.

![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 $i$th departing item. Arrivals are denoted by × and departures by ●.](image)

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

$$P(Z_i^{(r)} \leq z \mid W_i = w) = \begin{cases} 
1 - \sum_{j=0}^{r-1} e^{-\lambda z}(\lambda z)^j/j! & z < w \\
1 - \sum_{j=0}^{r-2} e^{-\lambda z}(\lambda z)^j/j! & z \geq w.
\end{cases} \quad (4)$$

**Proof.** The event $\{Z_i^{(r)} \leq z\}$ is exactly the event “there were at least $r$ arrivals in the time interval $]0,

5
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 \( \lambda \), the probability of having at least \( r \) of them in the interval is \( 1 - \sum_{j=0}^{r-1} e^{-\lambda z} (\lambda z)^j / j! \). If \( z \geq w \), the arrival at \( Y_i - w \) is included in the interval, so to have a total of at least \( r \) arrivals, one needs 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 \( 1 - \sum_{j=0}^{r-2} e^{-\lambda z} (\lambda z)^j / j! \) (here and below, summation from 0 to \(-1\) in the case \( r = 1 \) is to be interpreted as the empty sum, that is, as zero).

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

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

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

\[
H^{(r)}(z) = P(Z_i^{(r)} \leq z) = \int_{0}^{\infty} P(Z_i^{(r)} \leq z | W_i = w) dG(w)
\]

\[
= \int_{0}^{z} \left( 1 - \sum_{j=0}^{r-2} e^{-\lambda z} (\lambda z)^j / j! \right) dG(w) + \int_{z}^{\infty} \left( 1 - \sum_{j=0}^{r-1} e^{-\lambda z} (\lambda z)^j / j! \right) dG(w)
\]

\[
= 1 - (1 - G(z)) e^{-\lambda z} (\lambda z)^{(r-1)} / (r-1)! - \sum_{j=0}^{r-2} e^{-\lambda z} (\lambda z)^j / j!.
\]

Inverting this relationship between \( H^{(r)}(z) \) and \( G(z) \), the proposition is proven. This equality is the counterpart of equation (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 events of the sequence \( \{Z_i^{(r)}\} \) has probability zero or one. 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 zero or one. 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 \( \sigma \)-field
of \( \{Z_i^{(r)}\} \). Since \( m_G < \infty \), there are finitely many customers at time \( Y_m \) (the time of the \( m \)th departure) with probability one, so that there exists a time \( t > Y_m \) in which all these customers have departed from the queue. All \( Z_i \) corresponding to departures subsequent to the \( r \)th 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 \( \sigma \)-field of \( \{Z_i^{(r)}\} \), and the proposition is proven.

Let \( H_n^{(r)}(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Z_i^{(r)} \leq x\}} \) be the empirical estimator of \( H^{(r)} \), and let \( \lambda_n \) be a strongly consistent estimator of \( \lambda \), say the usual maximum likelihood estimator (the inverse of the mean inter-arrival time). By plugging \( H_n^{(r)} \) and \( \lambda_n \) for \( H \) and \( \lambda \) in equation (5), we define the counterpart of \( V_n \) from equation 2,

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

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. One 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

\[
\tilde{G}_n^{(r)}(x) = \sup_{y \leq x} V_n^{(r)}(y).
\]

This function is a strongly and uniformly consistent estimator of \( G \), as proven 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

\[
P\left( \lim_{n \to \infty} \sup_x |G(x) - \tilde{G}_n^{(r)}(x)| \to 0 \right) = 1.
\]

**Proof.** The proof again follows the same steps as its counterpart in Brown (1970). Since the sequence \( \{Z_i^{(r)}\} \) is stationary and ergodic (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 non-monotonicity 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

\[
\begin{align*}
\text{minimize}_{g_1, \ldots, g_n} & \quad \sum_{i=1}^{n} (g_i - V_n^{(r)}(Z_i^{(r)}))^2 \\
\text{s.t.} & \quad g_i \leq g_j \text{ whenever } Z_i^{(r)} \leq Z_j^{(r)}.
\end{align*}
\tag{8}
\]

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 (Robertson \textit{et al.}, 1988). Further, the problem can be efficiently solved using the PAVA algorithm, as described, for example, in De Leeuw \textit{et al.} (2009).

Simulation results (as in Nelgabatz, 2012) indicate overwhelmingly that the isotonic estimator \( G_n^{(r)} \) defined via equations (7) and (8) is superior to the supremum estimator \( \tilde{G}_n^{(r)} \) defined in (6). Therefore, we use henceforth the isotonic estimator alone. While we believe that the isotonic estimator is consistent, it seems that proving so 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 one cannot use uniform convergence of \( V_n^{(r)} \) to \( G \) on compact sets.

As a criterion for judging the quality of the estimator, we use the mean \( L_{\infty} \) error, \( E(\sup_x |G_n^{(r)}(x) - G(x)|) \), which we estimate via simulation. This criterion is closely related to the statistic of the Kolmogorov–Smirnov test (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. Figure 3 shows the performance of the estimator as a function of the load \( \rho \) for three types of \( G \) — exponential, uniform, and Pareto — and for \( r = 1, \ldots, 10 \). The most prominent feature of Figure 3 is that for each fixed \( r \), the quality of the estimator varies greatly with \( \rho \), and that the different curves of \( E(\max_x |G_n^{(r)}(x) - G(x)|) \) for different \( r \) are very different from each other. As a result, different \( r \)'s are optimal for different values of \( \rho \), and the larger \( \rho \), the larger the optimal \( r \). This observation is in line with Figure 1, which showed that when \( \rho \) is large, using \( r = 1 \) yields a poor estimate.

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

3 Estimating \( G \) via \( r \)th-latest arrivals, adaptively chosen \( r \)

The values of the performance criterion \( E(\max_x |G_n^{(r)}(x) - G(x)|) \), as shown in Figure 3, are a rather complex function of \( G, \lambda, n, \) and \( r \) (the dependence on \( \lambda \) is suppressed in our notation, and the dependence on the observations \( Z_1, \ldots, Z_n \) is integrated out through the expectation operator). Of these factors, \( G \) and \( \lambda \) 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, one would use the optimal \( r \), which minimizes the performance criterion given the \( G, \lambda, \) and \( n \) at hand:

\[
r^* = \arg \min_r E(\max_x |G_n^{(r)}(x) - G(x)|).
\]
However, while \( n \) is always known in advance and \( \lambda \) 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 \( \lambda \) and \( G \), but rather, on the load \( \rho \), which summarizes the relationship between them. Recall that it was indeed the change in the load which rendered Brown’s original estimator unsatisfactory (see again Figure 1), and that \( \rho \) 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 \( \rho \) directly, we will calculate a statistic closely related to \( \rho \), use it to estimate \( \rho \), 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 (Ross, 2010), it is known that \( m_Q = \lambda m_G \), so that \( \rho = \lambda m_G = m_Q \). To estimate \( \rho \), one could use an estimate of \( m_Q \). If the queue-length process \( \{Q_t\} \) were observable over the sampling period \([0, T]\), one could use \( \hat{m}_Q = \frac{1}{T} \int_0^T Q_t \, dt \) for this purpose. However, under our assumptions, only the arrival and departure epochs are observable, so this approach may not be pursued. Still, one could use a similar approach: Let \( \{L_t, 0 \leq t \leq T\} \) 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, \( \{K_t, 0 \leq t \leq T\} \), via \( K_t = L_t - \min_{0 \leq s \leq T} L_s \), and let \( \hat{m}_K = \frac{1}{T} \int_0^T K_t \, dt \). The process \( \{K_t\} \) coincides with \( \{Q_t\} \) if the system becomes empty at least once during the sampling period, in which case also \( \hat{m}_K = \hat{m}_Q \); in general,

\[
K_t = Q_t - \min_{0 \leq s \leq T} Q_s \leq Q_t, \quad 0 \leq t \leq T
\]  

(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 \( E(\hat{m}_Q) = E(\frac{1}{T} \int_0^T Q_t \, dt) = \frac{1}{T} \int_0^T E(Q_t) \, dt = \frac{1}{T} \int_0^T m_Q \, dt = m_Q = \rho \)). From (10) it follows that \( E(\hat{m}_K) \leq \rho \). Although the difference in mean between \( \hat{m}_Q \) and \( \hat{m}_K \) increases with \( \rho \), \( \hat{m}_K \) still holds information on \( \hat{m}_Q \). See Figure 4. Let \( h \) be the function that relates \( \rho \) to \( \hat{m}_K \) (for fixed \( n \)) via \( h(\rho) = E(\hat{m}_K) \), and let \( h^{-1} \) be its inverse.

We now propose the following approach for selecting \( \hat{r}^* \), the estimate of the optimal \( r^* \) from equation (9). Given only \( n \), we use data such as those presented in the left panel of Figure 3, to
Figure 4: Mean queue length $m_Q$ and $E(\hat{m}_K)$ as a function of $\rho$, as estimated from simulation. $G$ is exponential and $n = 5000$.

<table>
<thead>
<tr>
<th>$c_0$</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$c_3$</th>
<th>$c_4$</th>
<th>$c_5$</th>
<th>$c_6$</th>
<th>$c_7$</th>
<th>$c_8$</th>
<th>$c_9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.8</td>
<td>2.8</td>
<td>4.5</td>
<td>6.0</td>
<td>7.3</td>
<td>8.1</td>
<td>9.1</td>
<td>9.8</td>
<td>10.7</td>
</tr>
</tbody>
</table>

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

partition $[0, \infty)$ into intervals separated by cutoff points $0 = c_0 < c_1 < c_2 < \ldots$, such that $r^* = i$ when $\rho \in [c_{i-1}, c_i]$ (i.e., when $\rho$ is in the $i$th 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 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 $E(\max_x |G_n^{(r)}(x) - G(x)|)$ using the optimal $r = r^*$ (i.e., the pointwise minimum of the curves from Figure 3), and the dashed curve consists of the $E(\max_x |G_n^{(r)}(x) - G(x)|)$ 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 $r$’s.

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 panel of Figure 5, $G$ is uniform and Pareto, respectively, but the $c_i$ and $\hat{h}^{-1}$ are exactly the ones 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 $r$’s at the same time. Recall that the input to the isotonic regression procedure was the collection of pairs $\{(Z_i^{(r)}, G_i^{(r)})\}$ for some $r$. One may consider several such collections of pairs corresponding to several $r$’s, and run the isotonic regression over their union. The dotted curves in Figure 5 are the results of such a procedure, when using combination of three $r$’s: 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 $\rho$, also improve upon those obtained using $r^*$.

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: Figure 6 shows again $E(\max_x |G_n^{(r)}(x) - G(x)|)$ as a function of $\rho$, 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 $n = 5000$ (Table 1). Again, the two curves are rather close to each other.
4 A note on simulation

All simulation were carried out using the R programming language (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 \( \rho \), the sojourn time distribution \( G \) was varied as follows. When \( G \) is \( \text{exp}(\eta) \) 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

\[
f(x) = 3\beta(\beta x + 1)^{-4}, \quad x \geq 0.
\]

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 scale parameter \( \beta \), 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 propose an estimator for the CDF \( G \) of the sojourn time in an M/G/\( \infty \) 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 that of Brown (1970); indeed, our proposed method is a generalization of Brown’s, aimed at correcting 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 uses the most recent previous arrival, whereas our generalization uses the \( r \)th 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 inter-arrival time), so when \( \lambda \) is large, we obtain information about \( G(x) \) mostly for small \( x \)’s in the support of \( G \). Our method corrects for that, since our \( r \)-delayed observations are typically around \( r/\lambda \), 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 \( \rho \)).

We note, however, that when \( \rho \) 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 \( \rho \) needs to be below 6 to produce reasonable estimates. The reason for this is that when one observes only the arrival and departure epochs, without their pairing, large values of \( \rho \) correspond to a very high noise level. One might say that the values of the sojourn times about which we wish to learn get almost completely hidden by the noise when \( \rho \) 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 one can obtain lower bounds on these rates of convergence, and see how they compare to what our estimator achieves.

Some of the works reviewed in the introduction estimated \( G \) via the busy-period process \( \{I(Q_t > 0)\} \) (Bingham and Pitts, 1999; Hall and Park, 2004; 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 $\rho$ with which our proposed method performs well. The expected length of a busy period of an M/G/$\infty$ queue is $\frac{1}{\lambda}(e^\rho - 1)$, so for $\rho = 6$, the average busy period has $\sim 400$ arrivals and departures, and hence $n = 5000$ corresponds to only $\sim 12$ 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, one does not know of course whether $G$ is exponential or not, 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 PDF’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.

Our criterion for judging the quality of the estimators $\hat{G}_n^r$ was $E(\sup_x |G_n^r(x) - G(x)|)$. One may replace it with alternative criteria: these could be other functionals of the entire distribution $G$, such as $E[\int (G_n^r(x) - G(x))^2 \, dx]$, or criteria based on $G(a)$ for some fixed $a$ of interest, such as the MSE $E[(G_n^r(a) - G(a))^2]$. In either case, the approach we outlined can be used with the obvious modifications.

References


