Estimating Customer Impatience in a Service System With Unobserved Balking

Inoue, Y.; Ravner, Liron; Mandjes, M.

DOI
10.1287/stsy.2022.0101

Publication date
2023

Document Version
Final published version

Published in
Stochastic Systems

License
CC BY

Link to publication

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Estimating Customer Impatience in a Service System With Unobserved Balking

Yoshiaki Inoue, Liron Ravner, Michel Mandjes

1. Introduction

In many service systems, potential customers are provided with real-time information on the delay they will experience (Ibrahim 2018, Dong et al. 2019). This information, typically in terms of anticipated delay or current workload, is used by the potential customers to decide whether to join the system or leave. In general, one could view balking as an implicit form of admission control; the load on the service system does not explode as a consequence of the fact that during busy periods some of the potential customers decide to leave, thus mitigating the overall congestion. This means that the service system is in effect self-regulating even if the potential service demand is unknown to the administrator. To make sound decisions in designing the service system, however, it is crucial to have knowledge of the volume of the potential demand and the mechanism based on which potential customers decide to join or to leave. With this knowledge, the service provider can, for example, determine the optimal admission price or decide whether it is economically viable to increase the service capacity.

This paper is motivated by the fact that it is challenging to extract this knowledge from data when balking customers who decide not to join are not observed, and remarkably, we manage to devise a procedure to estimate the underlying patience and arrival rate parameters. The model is a multiserver queue with a Poisson stream of customers, enabling evaluation of the corresponding likelihood function of the state-dependent effective arrival process. We establish strong consistency of the MLE and derive the asymptotic distribution of the estimation error. Several applications and extensions of the method are discussed. The performance is further assessed through a series of numerical experiments. By fitting parameters of hyperexponential and generalized hyperexponential distributions, our method provides a robust estimation framework for any continuous patience-level distribution.
space at a given time. This form of data censorship occurs also in systems that provide real-time delay data via electronic means such as dedicated apps for the service. Some examples are expected travel times in transportation services (e.g., Google Maps and Uber) and expected delivery times in food delivery services (e.g., UberEats in the United States or Deliveroo in Europe). Note that expected delay is often listed for the possible services (e.g., estimated arrival time of a taxi or delivery times for a list of Italian restaurants) even if a user does not indicate the specific service he or she is interested in, and therefore, the system cannot know whether a user balked from the service because of the delay information.

In this paper, we model such service systems with unobservable balking as an $M/G/s + H$ queue, where the last symbol stands for the cumulative distribution function of the customer patience. More specifically, we consider a first-come, first-served (FCFS) queue with $s (s \geq 1)$ servers and potential customers arriving according to a Poisson process with rate $\lambda > 0$. Potential customers have independent and identically distributed (i.i.d.) patience levels that are distributed as the nonnegative random variable $Y \sim H$ and bring i.i.d. service requirements that are distributed as the nonnegative random variable $B \sim G$. Customer $i$ joins the system if $Y_i$ is greater or equal to the virtual waiting time at the moment of his or her arrival and otherwise balks without being observed. Note that although customers are homogeneous in a statistical sense, they have individual (thus heterogeneous) patience levels, so this model can clearly represent the types of customers that will effectively join the system.

We set up an estimation problem by parametrizing the patience distribution as $H_\theta (\theta \in \Theta)$, where $\Theta$ denotes a compact set satisfying some regularity conditions. The goal of this work is to estimate the pair $(\lambda, \theta)$ based on observation of the effective queueing process, which is constructed from records of arrival and departure times not including balking customers. As some of the customers balk, the corresponding effective interarrival times are not the usual i.i.d. exponentially distributed random variables. The effective arrival process is in fact a nonhomogeneous Poisson process whose rate depends on the virtual waiting time process. In particular, the effective arrival rate to the system when the virtual waiting time $v \geq 0$ is $\lambda(1 - H(v))$. Note that the marginal effective arrival process is not a nonhomogeneous Poisson process with respect to time. Therefore, the dependence structure of the observations of interarrival times is directly determined by that of the virtual waiting time process. The workload-dependent arrival dynamics are utilized to derive a maximum-likelihood estimator (MLE) based on the effective interarrival observations. Specifically, the density of an interarrival time $A$ conditional on initial virtual waiting time $v \geq 0$ is given by

$$f_{A|v}(t) = \lambda (1 - H(v - t)) \exp \left(-\lambda \int_0^t (1 - H(v - u)) \, du \right), \quad t \geq 0.$$  

The Markovian dynamics of the queue further imply that the likelihood of the interarrival times conditional on the waiting times and job sizes is given by a product of densities with the above form for any sample of observations.

We further study the asymptotic properties of the proposed estimator. Importantly, the asymptotic performance of the estimator is not given by standard results due to the intricate underlying dependence structure. Indeed, so as to prove consistency and asymptotic normality, a subtle reasoning is required to make sure that the situation at hand fulfills specific regularity conditions. It is further shown that for homogeneous (deterministic) patience levels, that is, $P(Y = \theta) = 1$ for some $\theta > 0$, the asymptotic distribution of the errors is not normal but rather exponential (with rate $n$ and not $\sqrt{n}$). The main contribution of this paper is the development of procedures for consistent estimation of the total arrival rate (that is, corresponding to customers joining the system and balking customers) and the patience-level distribution, observing the queue-length process only. Hence, somewhat counterintuitively, we can estimate load that has never been observed.

The main purpose of this paper is to lay foundations for statistical inference for congested service systems with unobserved balking. Although the queueing model considered in this paper is of course not a perfect replica of any specific real system, it does provide a methodological statistical approach for service systems where unobserved balking due to congestion is a key feature. Furthermore, the framework presented here can be extended to capture additional features such as noisy waiting time information, as will be discussed in the concluding remarks.

It is also important to stress that the framework we develop in this paper is essentially different from the extensively studied problem of estimating patience parameters from observed abandonment data (see Brown et al. 2005, Mandelbaum and Zelty 2013). If balking customers are observed, then the data are comprised of waiting times and corresponding indicator variables stating whether a customer joined or balked. Specifically, when it is observed that a customer $i$ joined (resp. balked) given the waiting time $W_i$, the observation is of the form $Y \geq W_i$ (resp. $Y < W_i$). These data can be directly applied to construct an estimator for $\theta$, for example, by applying an MLE method. If abandonment times are directly observed, then even more information is available, namely the realization of $Y_i$ for every abandoning customer, and $Y \geq W_i$ for customers who obtained service. Note that in
the above scenarios estimation of the total arrival rate is straightforward because all arrivals are observed. However, when balking is unobserved, the only information available is that of interarrival and waiting times, and therefore, indirect estimation methods are called for. Of course, there may be applications where some of the balking or reneging customers are observed, whereas others are not. For such systems, it may be reasonable to adopt a hybrid approach that uses classical patience estimation methods together with those presented in this paper. For the sake of brevity, we assume throughout that balking happens immediately and is not observed by the system operator so that no data regarding the balking customers is available for estimation purposes.

1.1. Contributions

We proceed by detailing our paper’s contributions. The focus is on developing estimators with provable performance guarantees. Notably, setting up our estimator is straightforward in the sense that it relies on a closed-form expression for the likelihood and does not require any queueing-theoretic analysis. As a result, even for models for which stationary performance analysis is involved (or intractable), parameter inference can be done in a relatively direct manner. Our contributions relate to (i) our estimator and its asymptotic properties, (ii) results and experiments for specific patience-level distributions, and (iii) extensions and ramifications.

- Framing the system described above as a queueing model with impatient customers, we can evaluate the corresponding likelihood function (in terms of the observed quantities). This means that, in a parametric context, we can estimate the unknown parameters— that is, λ and the parameter(s) θ corresponding to the patience level Y — relying on a maximum-likelihood procedure. The estimation method has several attractive features, the most prominent one being that it does not require simultaneous estimation of the service requirement distribution or even making any distributional assumption about it. In addition, our methodology uses the fundamental queueing dynamics and does not, as is common in the literature, rely on any fluid and/or diffusion approximations.
  - In the case of a continuous patience-level distribution, we prove, under appropriate regularity conditions, strong consistency and asymptotic normality of the resulting estimation errors (scaled by \(\sqrt{n}\)). The proof of the asymptotic normality relies on an application of a suitable version of the martingale central limit theorem.
  - In the case that the patience level Y is constant (i.e., \(Y = \theta\) for some \(\theta\)), more refined results can be obtained. Most notably, we show that the estimation error (scaled by \(n\)) converges to an exponential distribution with a parameter that depends on the stationary workload density at \(\theta\) and the loss probability of customers.
  - A number of other special cases are dealt with, covering exponentially and (generalized) hyperexponentially distributed patience levels; we present closed-form expressions for the asymptotic variance of the estimation error. The class of generalized hyperexponentially distributions is highly relevant for various reasons, most importantly because these can be used to approximate any positive continuous distribution arbitrarily well (Botta and Harris 1986). In addition, in the call center literature (Roubos and Jouini 2013), it has been observed that such distributions yield a good fit for observed customer patience in terms of the time until abandonment. We provide several numerical examples that illustrate the robustness of our methodology. Indeed, our experiments show that even if the patience-level distribution is misspecified (i.e., is not generalized as hyperexponential itself), the distribution function of the estimated generalized hyperexponential distribution is still remarkably close to the true one.
  - We also discuss several ramifications and model extensions that can be handled using our framework. These include extending the framework to a system with noisy delay information, estimating the proportion of impatient customers in the population, and estimating specific utility parameters such as service value and waiting time sensitivity.

1.2. Related literature

There is a substantial body of literature on queues with impatient customers. Without attempting to provide an exhaustive overview, we include a brief account of the various research lines. Naor’s influential paper (Naor 1969) can be seen as a pioneering contribution to the field of “behavioral queues.” It presented a powerful stylized model for a queueing resource in which potential customers decide to join or to balk based on delay information. Since then, performance and economic analysis of queueing systems with balking customers has been studied extensively. In particular, the effect of providing workload information to strategic customers was investigated in, for example, Guo and Zipkin (2007, 2008). Many other model variations are reviewed in Hassin and Haviv (2003) and Hassin (2016). More background on applied probability and queueing systems in a general context is provided in, for example, Assmussen (2003) and Debacke and Mandjes (2015).

For various specific types of queueing systems with impatient customers, explicit performance results are known; see, for example, Boxma et al. (2010, 2011), Liu and Kulkarni (2006, 2008), and Mandelbaum and Shimkin (2000, 2004). Importantly, however, the estimation techniques presented in our paper do not require knowledge.
of, for example, the queue’s stationary distribution. It is also worth stressing that despite the fact that queues with impatient customers are well studied, to the best of our knowledge, hardly any workload-based estimation techniques have been developed. A notable exception is Gorst-Rasmussen and Hansen (2009), featuring an asymptotic analysis of statistical inference of waiting times truncated by abandonments; our setting, however, is crucially different from the one in Gorst-Rasmussen and Hansen (2009), because in the latter, abandoning customers and their sojourn times are observed.

There are various papers on estimation problems in the setting where all customers join but leave when getting impatient; examples are Courcoubetis et al. (1995), Aksın et al. (2013), Mandelbaum and Zeltyn (2013), and Yefenof et al. (2022). Importantly, this setting is crucially different from ours, in which the balking customers are not observed. In this branch of the literature, the challenge lies in the fact that for customers who are served, we have just a lower bound on their impatience time, making this a “right-censored” estimation problem.

Another related research direction concerns the so-called queue inference engine, a setting in which one does not observe the arrival process of the customers but rather the ordered service entry and service completion times (so-called “transactional data”); see, for example, Larson (1990) and Daily and Servi (1998). One comes across this situation when considering, for example, automatic teller machines, where the queue of customers waiting for the machine (and in particular their arrival times) is not observable, but the transactional data are recorded. In Daley and Servi (1998), a multiserver queue is considered, but with the specific feature that balking is assumed to occur when all servers are busy upon arrival, with a fixed probability independent of the queue length, making the underlying queuing dynamics fundamentally different from ours.

Strictly speaking, Negahban (2019, 2022) did not deal with transactional data; it is assumed that the interarrival times corresponding to the nonbalking customers are observed (so that its scope is somewhat similar to ours). At the methodological level, however, Negahban (2019, 2022) crucially differed from our approach; the proposed estimation procedure relies on discrete-event simulation rather than a rigorously backed maximum-likelihood procedure.

Our work fits in the larger branch of the literature on inference for queues. This domain focuses on estimation problems where the queue is observed (for instance periodically), with the objective to estimate its input parameters. We refer to Asanjarani et al. (2021) for a broad recent overview of this area. Examples are the procedures for the M/G/1 queue, as presented in Hansen and Pitts (2006) and den Boer and Mandjes (2017). We also refer to the Poisson-probing based approach (Ravner et al. 2019) for Levy-driven queues (Dębicki and Mandjes 2015) and the general framework presented in Duffie and Glynn (2004), as well as the corresponding hypothesis-testing problem (Mandjes and Ravner 2020). Some approaches rely on exploiting knowledge of the queue’s tail probabilities (Courcoubetis et al. 1995, Glynn and Zeevi 2000, Duffy and Meyn 2011). A related line of research concerns the detection of stability or instability of queueing networks; see the Monte Carlo-based methods of Leahu et al. (2017) and Mandjes et al. (2017).

The main conclusion from the above is that, despite the mature literature on queues with impatient customers and related estimation problems, there is a lack of methodologies that can deal with the situation in which balking customers are not observed.

1.3. Paper Organization
This paper is organized as follows. In Section 2 we describe the model featuring in this paper and introduce the notation used. In addition, we present a number of known results on queues with customer impatience that are relied on later. Then, in Section 3, we construct and analyze our maximum-likelihood estimator. For continuous patience-level distributions, we prove strong consistency and asymptotic normality of the estimation error. The case of deterministic patience is treated separately, leading to a similar conclusion regarding strong consistency, but with the notable difference that in this case the limiting distribution of the estimation error is exponential. Section 4 provides detailed analysis of the MLE for exponential, hyperexponential, and generalized hyperexponential patience distributions, the latter cases being relevant because they can be used to approximate the distributions of arbitrary nonnegative random variables. This includes explicit analysis of the likelihood function and its corresponding asymptotic variance as well as simulation experiments assessing the performance of the estimators. Section 5 illustrates how the method can be utilized for several applications. In Section 6, we provide concluding remarks and directions for followup research, including a brief discussion of a nonparametric estimation procedure.

2. Model and Preliminaries
In this section, we provide a detailed model description, introduce the notation used throughout this paper, state the objective of the paper, and present preliminaries.
2.1. Model Description

We consider a service system with \( s \) servers (\( s \in \mathbb{N} \)). Potential customers arrive at the service system according to a Poisson process with rate \( \lambda \). Each customer has a service requirement with cumulative distribution function (cdf) \( G(\cdot) \) and a patience level with cdf \( H(\cdot) \). We define \( B \) and \( Y \) as the generic random variables corresponding to service requirements and impatience levels, respectively. The interarrival times \( (T_i)_{i=1,2,\ldots} \), the service requirements \( (B_i)_{i=1,2,\ldots} \), and patience levels \( (Y_i)_{i=1,2,\ldots} \) are independent sequences of i.i.d. distributed random variables. Customer \( i \) joins the queue if \( Y_i \) is smaller than or equal to the waiting time he or she will experience. More specifically, let \( V(t) \) (for \( t \geq 0 \)) denote the virtual waiting time at time \( t \), which is defined as the time it would take before at least one server becomes idle if no customers have joined the system after time \( t \):

\[
V(t) = \inf \{ u \geq 0 : Q(t) - (N(t+u) - N(t)) \leq s - 1 \},
\]

where \( Q(t) \) denotes the number of customers in the system (including those in service) at time \( t \) and \( N(t) \) (for \( t \geq 0 \)) denotes the total number of customers finishing their service in a time interval \((0,t]\). For the single-server case (i.e., if \( s = 1 \)), the virtual waiting time equals the workload in system, that is, the sum of remaining service requirements of customers. Note that the virtual waiting time in the multiserver case \((s = 2,3,\ldots)\) does not coincide with the workload in system, unlike in the single-server case:

(i) When \( i \in \{0,1,\ldots,s-1\} \) servers are busy, the virtual waiting time equals zero, whereas the workload equals some positive value and decreases with slope \( i \) as time passes without effective arrivals.

(ii) When all \( s \) servers are busy, the virtual waiting time and the workload decrease with slope 1 and \( s \), respectively, as time passes without effective arrivals.

Figure 1 illustrates an example of the realization of the virtual waiting time process for \( s = 2 \). The actual waiting time of a customer arriving at time \( t \) equals \( V(t) \) so that a customer balks if the virtual waiting time just before the arrival instant exceeds the customer’s patience threshold. The resulting system can be considered as an M/G/s queue with impatient customers.

Because some customers will balk, there will be a difference between the interarrival times of the customers (covering both the balking and nonbalking customers) on one hand, and the effective interarrival times (covering just the nonbalking customers) on the other hand. Throughout this paper, the sequence of effective interarrival times is denoted by \( (A_i)_{i=1,2,\ldots} \). Importantly, observe that the random variables \( A_i \) are generally neither identically distributed nor independent. We can reconstruct them from the sequences \( (T_i)_{i=1,2,\ldots} \), \( (B_i)_{i=1,2,\ldots} \), and \( (Y_i)_{i=1,2,\ldots} \) in the following manner. For the following construction, we let a nonbalking customer arrive at time \( t = 0 \), as will be further motivated in Remark 1. Denoting the virtual waiting time process at time \( t \) by \( V(t) \), we have that

\[
j_1 := \inf \left\{ j \geq 1 : V\left( \sum_{k=1}^{j} T_k \right) \leq Y_j \right\}.
\]

Figure 1. An Example of the Virtual Waiting Time Process \((s = 2)\)
Figure 2. The Virtual Waiting Time $V(t)$ During a Busy Period for a Single-Server System and Deterministic Patience Levels

![Diagram of the Virtual Waiting Time]

Notes. The dotted red lines mark arrival instants of new customers. Arrivals 3 and 5 observe a virtual waiting time level above $\theta$ and immediately balk, and thus the virtual waiting time continues to deplete as if there was no arrival. Although six customers have arrived during the busy period, only four effective interarrival times and service requirements are observed, namely $(A_1, X_1), (A_2, X_2), (A_3, X_3), (A_4, X_4)$, being equal to $(T_1, B_1), (T_2, B_2), (T_3 + T_4, B_3), (T_5 + T_6, B_6))$.

Similarly, $A_i = \sum^{j_i}_{k=j_{i-1}+1} T_k$, where

$$j_i := \inf \left\{ j \geq j_{i-1} + 1 : V \left( \sum^{j}_{k=1} T_k \right) \leq Y_j \right\}.$$ 

Observe that $j_i$ has the interpretation of being the index of the $i$-th nonbalking customer. Let $\hat{A}_i := \sum^{j}_{k=1} A_k$ denote the $i$-th effective arrival time. We denote the upward-jump size in the virtual waiting time caused by the $i$-th joining customer by $X_i := V(\hat{A}_i) - V(\hat{A}_{i-1})$. In the single-server case ($s=1$), we have $X_i = B_{j_i}$. In the multiserver case ($s \geq 2$), on the other hand, $X_i$ takes a complicated form, depending on both the queue-length and the residual service times of customers in service seen on the arrival instant. For a single-server system ($s=1$), Figure 2 illustrates the workload process during a single busy period for the case of deterministic patience, that is, $Y = \theta$ for some $\theta > 0$. The building blocks of our framework, viz. the interarrival and service requirements and their effective counterparts, are highlighted in the figure. It is stressed that the distribution of the effective interarrival times can significantly deviate from the (exponential) distribution of the interarrival times. Figure 3 illustrates this effect for two different patience-level distributions. In addition, as mentioned, the effective interarrival times are not independent, as opposed to the interarrival times.

We focus on the case that the patience-level distribution is determined by a parameter $\theta$ that is a vector in $\mathbb{R}^p$ for some (known) $p \in \mathbb{N}$, entailing the estimation problem is parametric. In addition, the arrival rate $\lambda$ is to be estimated, which is evidently of a parametric nature as well. Alternatively, $\theta$ could be a function from some more general space, rendering the estimation problem nonparametric; in Section 6, we briefly discuss an approach that can be used in this setting. Throughout this paper, we assume that we observe the full queue-length process. This in particular means that we observe the arrival and departure epochs of the nonbalking customers (and that we do not observe the balking customers at all). The objective is to use this information to somehow learn the true arrival rate and the parameters of the patience-level distribution. More concretely, we wish to devise statistical procedures for estimating the arrival rate $\lambda$ and the distribution of the patience $Y$ (both corresponding to non-observable quantities) with provable performance properties.

2.2. Stationary Distribution

We assume that the mean service time $E[B]$ is finite and that, with $\rho := \lambda E[B]/\theta$ denoting the traffic intensity,

$$\rho \lim_{x \to \infty} (1 - H(x)) < 1.$$  (2)

Under these assumptions, we have that the queue is stable and that a stationary queue-length $Q := Q(\infty)$ and a stationary virtual waiting time $V := V(\infty)$ exist (see Baccelli and Hebuterne 1981 and Baccelli et al. 1984). Note
that \( \lim_{t \to \infty} H(x) < 1 \) holds only if there are customers with infinite patience, that is, \( P(Y = \infty) > 0 \). If \( P(Y < \infty) = 1 \), on the other hand, then we have \( \lim_{t \to \infty} H(x) = 1 \), so that the system is stable irrespective of the value of the traffic intensity \( \rho \). As we will see, the MLE method to be developed relies only on the transient dynamics of the system so that it can be applied even if the stability condition (2) is not satisfied. On the other hand, we will utilize the system stability to discuss the asymptotic properties of the proposed estimator.

For the M/G/s queue with impatient customers, exact expressions for the distributions of the stationary queue-length \( Q \) and virtual waiting time \( V \) are not known in the literature. However, we can derive the following relations they satisfy in a similar way to the single-server case (Kovalenko 1961). Let \( q_n := P(Q = n) \ (n = 0, 1, \ldots) \) denote the probability mass function of the stationary queue length. The stationary virtual waiting time distribution has probability mass \( \pi_0 := P(V = 0) = \sum_{n=0}^{\infty} q_n \) at zero, and it is absolutely continuous on \((0, \infty)\). Let \( X|Y \) denote a generic random variable for the stationary upward-jump size \( X \) conditioned that it takes a positive value and that the immediately preceding virtual waiting time equals \( y \):

\[
P(X|Y < x) = \lim_{t \to \infty} P(V(t) - V(t-) \leq x | V(t-) = y, V(t) > y).
\]

With the level-crossing argument, we can verify that the density function \( v(x) \) of the stationary virtual waiting time satisfies the following integral equation (cf. Boxma et al. 2011):

\[
v(x) = \lambda q_{n-1} \tilde{f}_0(x) + \lambda \int_0^x v(y) P(Y \geq y) \tilde{f}_1(y - x) \, dy, \quad x \geq 0,
\]

where \( \tilde{f}_1(x) := P(X|Y > x) \) denotes the complementary cdf of the conditional upward-jump size \( X|Y \).

Owing to the PASTA property, the virtual waiting time seen by an arriving customer has the same distribution as the stationary virtual waiting time \( V \). The stationary waiting time \( W \) of a nonbalking customer is thus given by a conditional random variable \( V|V \leq Y \). We can thus identify \( P(W = 0) \) and the probability density function (pdf) \( w(\cdot) \) of \( W \) in terms of the stationary virtual waiting time distribution \( v(\cdot) \):

\[
P(W = 0) = \frac{\pi_0}{1 - P_t}, \quad w(x) = \frac{v(x) P(Y \geq x)}{1 - P_t}, \quad x \geq 0,
\]

where \( P_t \) denotes the loss probability given by

\[
P_t = \int_0^\infty v(y) P(Y < y) \, dy.
\]
For \( s = 1 \), (3) simplifies to

\[
v(x) = \lambda \pi_0 \mathcal{G}(x) + \lambda \int_0^x v(y) P(Y \geq y) \mathcal{G}(x-y) dy, \quad x \geq 0,
\]

where \( \mathcal{G}(x) = 1 - G(x) \) denotes the complementary cdf of service requirements. A solution of the integral Equation (6) is, as can be found in Baccelli et al. (1984) and Inoue and Takine (2015), given in terms of the pdf \( g_e(x) = \mathcal{G}(x)/E[X] \) of the equilibrium distribution (i.e., the residual lifetime distribution) corresponding to the service requirements:

\[
v(x) = \pi_0 \sum_{n=1}^{\infty} u_n(x), \quad x \geq 0,
\]

\[
\pi_0 = \left( \sum_{n=1}^{\infty} u_n(x) dx \right)^{-1},
\]

\[
\begin{align*}
u_1(x) &= \rho g_e(x), \\
u_n(x) &= \rho \int_0^x u_{n-1}(y) P(Y \geq y) g_e(x-y) dy, \quad n = 2, 3, \ldots.
\end{align*}
\]

In some special cases, the expression (7) for the virtual waiting time density \( v(\cdot) \) can be further simplified, as shown in the examples below.

**Example 1** (Constant Patience Levels). We consider the case that the patience levels take a constant value \( \theta_0 \) (for some \( \theta_0 > 0 \)). Let \( g_e^{(n)}(\cdot) \) (for \( x \geq 0, n = 1, 2, \ldots \)) denote the \( n \)-fold convolution of the pdf \( g_e(\cdot) \) of the equilibrium distribution of the service times and \( G_e^{(n)}(x) \) the corresponding cdf. Also, define \( g_e^{(n)}(\cdot) \) by

\[
g_e^{(n)}(\cdot) := \left\{ \begin{array}{ll}
g_e^{(n)}(x)/G_e^{(n)}(\theta_0), & x \leq \theta_0, \\
0, & x > \theta_0.
\end{array} \right.
\]

We then have, by de Kok and Tijms (1985) and Inoue and Takine (2015), with as usual \( \ast \) denoting the convolution operator,

\[
\begin{align*}
\pi_0 &= \left( 1 + \rho + \rho \sum_{n=1}^{\infty} \rho^n G_e^{(n)}(\theta_0) \right)^{-1}, \\
v(x) &= \begin{cases}
\pi_0 \rho g_e(x) + \pi_0 \sum_{n=2}^{\infty} \rho^n G_e^{(n-1)}(\theta_0) \ast g_e^{(n-1)}(x), & 0 \leq x \leq \theta_0, \\
\pi_0 \rho g_e(x) + \pi_0 \sum_{n=1}^{\infty} \rho^n G_e^{(n)}(\theta_0), & x > \theta_0.
\end{cases}
\end{align*}
\]

**Example 2** (Exponential Service Times). If the service times follow an exponential distribution with mean \( 1/\mu \), we have, by Stanford (1979) and Inoue and Takine (2015), that

\[
\begin{align*}
\pi_0 &= \left( 1 + \lambda \int_0^{\infty} \exp \left[ -\mu x + \lambda \int_0^x P(Y \geq y) dy \right] dx \right)^{-1}, \\
v(x) &= \pi_0 \lambda \exp \left[ -\mu x + \lambda \int_0^x P(Y \geq y) dy \right], \quad x \geq 0.
\end{align*}
\]

Observe that if the patience distribution is also exponential with rate \( \theta \), then the expressions can be simplified by substituting \( \int_0^x P(Y \geq y) dy = \exp(-\theta x) \).

### 3. Parametric Estimation Procedure

In the setting considered, we focus on the first nonbalking \( n + 1 \) customers. We set the time origin \( t = 0 \) to the time instant that a nonbalking customer joins the system. We record the effective arrival times \( \hat{A}_1, \hat{A}_2, \ldots, \hat{A}_n \) and departure times. From this information, we can fully reconstruct the virtual waiting time process using (1); in
particular, we obtain the sequence of virtual waiting times \( V(0), V(\hat{A}_1), \ldots, V(\hat{A}_n) \) observed by the nonbalking customers and the sizes of upward jumps \( X_0, X_1, \ldots, X_n \) caused by them. In this section, the goal is to estimate in a parametric context the arrival rate and patience-level distribution from the data.

The crucial observation is that we can construct a likelihood function of the patience-level distribution using an observed sample of effective interarrival times \( A_1, A_2, \ldots, A_n \), waiting times \( W_0 = V(0), W_1 = V(\hat{A}_1), W_2 = V(\hat{A}_2), \ldots, W_n = V(\hat{A}_n) \), and upward-jump sizes \( X_0, X_1, \ldots, X_n \), which can be reconstructed from the observed sequences mentioned in the previous paragraph. We work under the natural assumption that the observed system is in stationarity, so that \( W_0 \) is the stationary waiting time of a nonbalking customer and \( X_0 \) is a stationary upward-jump size.

**Remark 1.** To make the sequence of waiting times \( W_i, W_2, \ldots, W_n \) stationary, we should let the time origin coincide with an arrival instant of a nonbalking customer. The validity of this statement can be argued as follows.

We first point out that the sequence of waiting times is in general not a Markov process; only for \( s = 1 \) it is. We obtain a Markov process when considering the vector of residual service times at effective arrival instants. The waiting times and upward jump sizes are in fact measurable functions of this process. Therefore, in what follows, we let the underlying Markov process of residuals be stationary.

First, it is observed that assuming the virtual waiting time at time \( t = 0 \) being stationary does not imply that the waiting time \( W_1 \) of the first customer follows the stationary distribution \( W \). To see this, consider (for simplicity) an ordinary \( M/G/s \) queue (that is, without customer impatience). If the virtual waiting time at time \( t = 0 \) follows the stationary virtual waiting time distribution \( V \), then the first customer arriving after time \( t = 0 \) finds an idle server with probability

\[
P(W_1 = 0) = P(V = 0) + \int_0^\infty e^{-\lambda x}v(x)dx > P(V = 0).
\]

Because the stationary waiting time \( W \) has the same distribution as the stationary virtual waiting time \( V \) in the ordinary \( M/G/s \) queue (due to PASTA), we find the inequality

\[
P(W_1 = 0) > P(V = 0) = P(W = 0).
\]

This shows that the waiting time \( W_1 \) of the first-arriving customer is biased in that it is not distributed as \( W \).

Intuitively, the bias is a consequence of the fact that a customer arriving after a long interarrival time is more likely chosen as the first arriving customer in our experiment (cf. the well-known inspection paradox in renewal theory), and such a customer tends to observe the system less congested than time average. It is easily seen that letting time \( t = 0 \) correspond to the arrival instant of a nonbalking customer makes the sequence of waiting times \( W_1, W_2, \ldots, W_n \) stationary.

Let \( \bar{H}(x) := 1 - H(x) \) denote the complementary cdf of patience levels. Also, we denote by \( \bar{H}(x) \) the probability of a customer joining when the virtual waiting time equals \( x \geq 0 \):

\[
\bar{H}(x) := P(Y \geq x) = \bar{H}(x) + P(Y = x).
\]

Note that \( \bar{H}(x) = \bar{H}(x) = P(Y > x) \) for continuous patience distributions.

We next describe the construction of the likelihood function. We first recall that the effective interarrival times \( A_1, A_2, \ldots, A_n \) are not exponentially distributed because of the fact that between two effective arrival intervals there may have been arriving customers who observed a virtual waiting time level exceeding their patience level (as depicted in Figure 3). Despite this, we can still characterize the effective interarrival time distribution in terms of the observed quantities. To see this, suppose that, for some \( t \geq 0 \), there have been no effective arrivals in the interval \( \hat{A}_{i-1}, \hat{A}_{i-1} + t \). An arrival in \( \hat{A}_{i-1} + t, \hat{A}_{i-1} + t + \Delta t \) then occurs with probability \( \lambda \Delta t + o(\Delta t) \), as \( \Delta t \downarrow 0 \). The corresponding customer joins the system (i.e., becomes effective) with probability \( \bar{H}(V(\hat{A}_{i-1}) - t) \) because (i) the virtual waiting time seen on the arrival equals \( \max(0, V(\hat{A}_{i-1}) - t) \) and (ii) \( \bar{H}(x) = 1 \) for \( x < 0 \). Therefore, the occurrence of the next (i.e., the \( i \)-th) effective arrival follows a time-inhomogeneous Poisson process with time-dependent intensity \( \lambda \bar{H}(V(\hat{A}_{i-1}) - t) \). Noting that \( V(\hat{A}_{i-1}) = W_{i-1} + X_{i-1} \) (for \( i = 1, 2, \ldots \)), we thus conclude that

\[
P(A_i > t | W_{i-1} + X_{i-1} = v) = \exp\left(-\lambda \int_0^t \bar{H}(v - u)du\right).
\]

This representation of the distribution of the effective interarrival times facilitates the evaluation of the likelihood. The (conditional) likelihood function for \( A := (A_1, A_2, \ldots, A_n) \) given \( W := (W_1, W_2, \ldots, W_n) \) and \( X := (X_1, X_2, \ldots, X_n) \)
is given by the product of the conditional densities
\[ f_{A_i|v}(t) = \frac{d}{dt} P(A_i > t | W_{i-1} + X_{i-1} = v) \]
\[ = \lambda H(v - t) \exp\left(-\lambda \int_0^t H(v - u) du\right). \]  
(11)

This yields the conditional likelihood function
\[ L_n(H; A, W | X) = \prod_{i=1}^n \lambda H(W_{i-1} + X_{i-1} - A_i) \exp\left(-\lambda \int_0^{A_i} H(W_{i-1} + X_{i-1} - u) du\right) \]
\[ = \lambda^n \prod_{i=1}^n H(W_i) \exp\left(-\lambda \int_0^{A_i} H(W_{i-1} + X_{i-1} - u) du\right). \]  
(12)

where we used a Lindley-type recursion, generalized to the multiserver queue:
\[ W_k = \max\{W_{k-1} + X_{i-1} - A_k, 0\}, \quad k = 1, 2, \ldots \]  
(13)

In the rest of this section, we assume that the patience-level distribution is characterized by a finite-dimensional vector of parameters \( \theta \in \Theta \subseteq \mathbb{R}^d \). For \( \theta \in \Theta \), let \( H_\theta(\cdot) \), \( \overline{H}_\theta(\cdot) \), and \( \tilde{H}_\theta(\cdot) \) denote \( H(\cdot) \), \( \overline{H}(\cdot) \), and \( \tilde{H}(\cdot) \) given parameter \( \theta \), respectively. Let any \( \tilde{H}_\theta(\cdot) \), with \( \theta \in \Theta \), be identifiable in the conventional Kullback-Leibler sense. The maximum-likelihood estimator (MLE) of \( (\lambda, \theta) \) is then given by
\[ (\hat{\lambda}_n, \hat{\theta}_n) \in \arg \max_{(\lambda, \theta) \in \mathbb{R}_+ \times \Theta} L_n(\lambda, \theta; A, W | X), \]
where
\[ L_n(\lambda, \theta; A, W | X) := \lambda^n \prod_{i=1}^n \tilde{H}_\theta(W_i) \exp\left(-\lambda \int_0^{A_i} \tilde{H}_\theta(W_{i-1} + X_{i-1} - u) du\right). \]  
(14)

In the sequel, we provide an asymptotic analysis describing the performance of this MLE. We first focus, in Sections 3.1 and 3.2, on the parameters pertaining to the patience only; that is, the estimation does not cover the arrival rate \( \lambda \). The object of study is, in self-evident notation, for \( \lambda > 0 \) given,
\[ \hat{\theta}_n \in \arg \max_{\theta \in \Theta} L_n(\theta; A, W | X). \]
Extending this to a procedure to also include estimation of the arrival rate is relatively straightforward; we get back to it in Section 3.3.

**Remark 2.** Observe that \( L \) is a conditional likelihood and not a full likelihood. This is due to the fact that the upward jump-sizes \( X_i \) have an elaborate distribution both marginally and jointly with the waiting times. The exception is the single-server case where \( X_i \sim B \) and is independent of the previous waiting and arrival times. Hence, for \( s = 1 \), the maximization of \( L_n(H; A, W | X) \) with respect to them is equivalent to that of the unconditional likelihood function for \( (A, W, X) \). It is noted that the procedure to estimate \( \theta \) does not require the simultaneous estimation of the service requirement or upward-jump size distributions.

In more detail, the remainder of this section is organized as follows. First, we establish (in Section 3.1) conditions for strong consistency and asymptotic normality of our estimator of the patience parameters, with the errors scaled by \( \sqrt{n} \), focusing on continuous patience-level distributions. We then consider (in Section 3.2) the case of deterministic patience and establish consistency (independently) for this case as well. It is noted that in the latter case the asymptotic errors are not normally but rather exponentially distributed, with the errors scaled by \( n \). This is due to the fact that for deterministic patience the MLE is obtained on the boundary of the sample data, much like in the well-known case of estimating the parameter \( \theta \) of a uniform distribution on \([0, \theta]\). As mentioned, Section 3.3 discusses the estimation of the arrival rate.

### 3.1. Continuous Patience-Level Distribution
This subsection covers the asymptotic performance of the MLE for the case that the patience-level distribution is continuous and parametric. Let \( \theta_0 \) denote the true parameter. Throughout the following analysis, the underlying
probability measure is $P_{\theta_0}$, that is, the probability measure corresponding to the true patience-level distribution. For the asymptotic results, we make the following assumptions, which will be discussed in Section 3.1.1.

**Assumption.** The following assumptions are imposed:

(A1) The parameter space $\Theta \subset \mathbb{R}^p$ is a compact set such that the true parameter lies in the interior of the set, that is, $\theta_0 \in \Theta^0$.

(A2) The observation period commences at $t = 0$, which corresponds to the stationary arrival instant of a nonbalking customer, and thus the sequence of waiting times $W_1, W_2, \ldots, W_n$ is stationary.

(A3) Let $\mathcal{H}_{\inf}(x) := \inf_{\theta \in \Theta} H_{\theta}(x)$. Depending on whether the right endpoint

$$h_{\sup} := \inf \{ x \geq 0 : \mathcal{H}_{\inf}(x) = 0 \}$$

takes a finite value or not, we assume one of the following properties: (i) If $h_{\sup} = \infty$, then there exists a positive nondecreasing function $f(\cdot)$ such that $\lim_{x \to \infty} f(x) \in (0, \infty]$ and for some constants $c_1, c_2 \in [0, \infty)$,

$$\lim_{x \to \infty} \frac{e^{f(x)}}{\mathcal{H}_{\inf}(x)} = c_1, \quad \lim_{x \to \infty} f(x) \mathcal{H}_{\theta_0}(x) = c_2.$$

(ii) If $h_{\sup} < \infty$, then there exists a positive nondecreasing function $f(\cdot)$ such that $\lim_{x \to h_{\sup}} f(x) = (0, \infty]$ and for some constants $c_1, c_2 \in [0, \infty)$,

$$\lim_{x \to h_{\sup}} \frac{e^{f(x)}}{\mathcal{H}_{\inf}(x)} = c_1, \quad \lim_{x \to h_{\sup}} f(x) \mathcal{H}_{\theta_0}(x) = c_2.$$

(A4) The gradient vector and Hessian matrix of $H_0$ are continuous with respect to $\theta$. With $\Psi_1(\theta)$ denoting the Hessian matrix of the log-likelihood corresponding to a single waiting time (in stationarity), $\mathbb{E}\Psi_1(\theta)$ has finite elements in all coordinates for any $\theta \in \Theta$.

(A5) The collection of functions $\{H_0(\theta) : \theta \in \Theta\}$ is equicontinuous; for any $x \geq 0$ and $\varepsilon > 0$ there exists a $\delta > 0$ such that $|H_0(\theta) - H_0(y)| < \varepsilon$ for any $y$ such that $|x - y| < \delta$ uniformly over all $\theta \in \Theta$.

We now state the main results of this subsection. In the sequel, $N(\mu, \Sigma)$ denotes a normally distributed random variable with mean vector $\mu$ and covariance matrix $\Sigma$.

**Theorem 1.** If Assumptions (A1)–(A3) and (A5) are satisfied, then, as $n \to \infty$,

$$\hat{\theta}_n \to_{\text{as}} \theta_0.$$  \hspace{1cm} (15)

**Theorem 2.** If Assumptions (A1)–(A5) hold, then, because $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0, I^{-1}(\theta_0)),$$  \hspace{1cm} (16)

where $I(\theta_0) := -\mathbb{E}\Psi_1(\theta_0)$, and $I^{-1}(\theta_0)$ denotes the inverse of $I(\theta_0)$.

The proofs of the above theorems are provided in Section 3.1.2 after the detailed discussion of (A1)–(A5) that we give in Section 3.1.1.

### 3.1.1. Discussion of Assumptions

We next provide more background on the assumptions imposed. In addition, we discuss their possible relaxation.

1. **(1) Assumption (A1) is natural, because it requires that the parameter space is big enough so that it contains the true parameter. In practice, the parameter space can be adjusted on the fly, for example, if for large $n$ we obtain boundary solutions for the first-order conditions.**

2. **(2) Assumption (A2) facilitates the use of known results on the convergence of stationary dependent sequences. Note that, as long as a stationary virtual waiting time distribution exists (for which we have given the condition as Equation (2) in Section 2), the same results should hold without making this assumption because of the regenerative nature of the process. However, without this assumption, the conditions for both consistency and asymptotic normality are harder to verify. In particular, the crucial step for consistency is the uniform convergence of the log-likelihood, established in Lemma 1 below. More elaborate conditions for uniform convergence are detailed in, for example, Ranga Rao (1962), Andrews (1987), and Pötscher and Prucha (1989).**

3. **(3) If there are patient customers who do not balk regardless of the waiting time, that is, $\lim_{x \to \infty} \mathcal{H}_{\theta_0}(x) > 0$, then Assumption (A3) requires that the parameter space $\Theta$ is chosen so that the existence of patient customers...**
(lim_{x \to \infty} H_{\theta}(x) > 0) is assumed for all \( \theta \in \Theta \), which is a reasonable assumption in modeling a service system where both patient and impatient customers exist. In the case of \( \lim_{x \to \infty} H_{\theta}(x) = 0 \), on the other hand, Assumption (A3) requires that the decay rate of \( H_{\theta}(x) \) does not vary too strongly among the distributions \( H_{\theta} \) with \( \theta \) in the parameter space \( \Theta \). In practice, this assumption is seldom violated. For example, if the true patience-level distribution decays exponentially, that is, \( \lim_{x \to \infty} e^{x} H_{\theta}(x) = c \) for some \( v > 0 \) and \( c > 0 \), then (A3) is satisfied if the infimum tail function \( H_{\inf}(x) \) does not decay faster than the doubly exponential function \( e^{-e^{-x}} \) (which evidently decays exceptionally fast). As another example, supposing that the true patience-level distribution obeys a power law, that is, \( \lim_{x \to \infty} x^{1} H_{\theta}(x) = c \) for some \( k > 0 \) and \( c > 0 \), then (A3) is satisfied if \( H_{\inf}(x) \) does not decay faster than \( e^{-x^{k}} \).

(4) Assumption (A4) enables the construction of a standard martingale CLT for the asymptotic distribution of the estimation error of the MLE. It is not a necessary condition, but in cases where the assumption is not satisfied, one is typically required to apply ad hoc analysis to derive an asymptotic distribution. In Section 3.2, we show that for a deterministic patience level that does not satisfy (A4), the asymptotic distribution of the error is exponential and therefore, one could see replacing (A5) with (A4) and verifying additional measurability conditions on \( \inf_{\theta \in \Theta} H_{\theta}(x) \) for a neighborhood \( \mathcal{B}_{\theta} \) of any \( \theta \in \Theta \) (see Andrews (1987), cor. 2).

### 3.1.2. Proofs.

From (14), first note that we can express the log-likelihood \( L_{n}(\theta; A, W | X) := \log L_{n}(\theta; A, W | X) \) by

\[
n \log \lambda + \sum_{i=1}^{n} \left( \log H_{\theta}(W_{i}) - \lambda \int_{0}^{A_{i}} H_{\theta}(W_{i-1} + X_{i-1} - u) du \right) \mathbb{1}_{[H_{\theta}(W_{i}) > 0]} .
\]

(17)

Observe that we replaced \( H_{\theta}(\cdot) \) with \( \tilde{H}_{\theta}(\cdot) \) because \( H_{\theta}(\cdot) \) is continuous.

With \( W := \lim_{t \to \infty} W_{t} \) existing, because \( n \to \infty \), the continuous mapping theorem yields

\[
\frac{1}{n} \ell_{n}(\theta; A, W | X) \to_{\text{as}} E[\ell(\theta; A_{1}, W_{0}, W_{1} | X_{0})],
\]

(18)

where \( A_{1} \) is the effective interarrival time when the initial virtual waiting time is \( W_{0} + X_{0} \) and \( W_{0} = d \) \( W \), that is, the stationary waiting time. From now on, we use the compact notations \( \ell_{n}(\theta) := \ell_{n}(\theta; A, W | X) \) and \( \ell(\theta) := E[\ell(\theta; A_{1}, W_{1}, W_{0} | X_{0})] \). The density of the effective interarrival times (11) is uniquely determined by the function \( H(\cdot) \), and thus so is the likelihood (14). This is because the function

\[
H(\cdot) := \left\{ H(\cdot) := \int_{[0,\infty]} \nu(\cdot) : x \geq 0 \right\}
\]

is uniquely determined by \( \theta \) as a consequence of the fact that for every \( \theta \in \Theta \) the measure \( \nu_{\theta}(\cdot) \) corresponds to a different distribution (that is, in the almost-everywhere sense). As a consequence, \( \{H(\cdot) : \theta \in \Theta\} \) is a parametric collection of distribution functions such that there is no pair \( \theta_{1}, \theta_{2} \in \Theta \) for which \( H(\cdot) = H(\cdot) \) almost everywhere. This entails that the model is identifiable in the Kullback-Leibler sense (see, e.g., Ferguson (1996), chapter 17), and hence,

\[
\ell(\theta) - \ell(\theta_{0}) < 0, \quad \forall \theta \neq \theta_{0}.
\]

(19)

The key step in the proof of Theorem 1 is establishing a uniform version of (18); then, strong consistency follows by the methodology of Wald (Ferguson (1996), chapters 16 and 17). The virtual waiting time observations are not independent, but by (A2) they are stationary so that we can apply a uniform law for stationary sequences that is commonly used in the econometrics literature. We rely specifically on a theorem taken from Ranga Rao (1962) and its extensions, in particular the ones developed in Andrews (1987) and Pötscher and Prucha (1989). The uniform convergence is established in Lemma 1 (proven in the Appendix).
Lemma 1. If Assumptions (A1)–(A3) and (A5) hold, then, because \( n \to \infty \),
\[
\sup_{\theta \in \Theta} \left| \frac{1}{n} \ell_n(\theta) - \ell(\theta) \right| \to_{as} 0.
\]  
(20)

Proof of Theorem 1. Observe that the strong consistency follows from the identifiability property (19) and from Lemma 1. Note in particular that the proof of Ferguson (1996) (theorem 17) does not rely on i.i.d. observations once uniform convergence has been established, and so the same steps can be applied here. In particular, a sufficient condition (relying on uniform convergence for strong consistency when observations are dependent) can also be found in Heijmans and Magnus (1986); for every \( \theta \neq \theta_0 \), there exists a neighborhood \( \mathcal{B}_\delta(\theta) \) such that
\[
\lim_{n \to \infty} \sup_{\gamma \in \mathcal{B}_\delta(\theta)} \left( \frac{1}{n} \ell_n(\gamma) - \frac{1}{n} \ell_n(\theta_0) \right) < 0,
\]  
(21)
almost surely. This follows directly from (19) and (20). □

We continue by proving asymptotic normality that was stated in Theorem 2. This amounts to showing that the estimation error \( \sqrt{n}(\hat{\theta}_n - \theta_0) \) is asymptotically normal because \( n \to \infty \), assuming that (A1)–(A5) are satisfied. To this end, we apply the well-known delta method and an appropriate version of the martingale CLT (see, e.g., Heyde 1997, theorem 12.6).

Proof of Theorem 2. For any given \( v \), we let \( \nabla \hat{H}_\theta(v) \in \mathbb{R}^p \) and \( \nabla^2 \hat{H}_\theta(v) \in \mathbb{R}^{p \times p} \) denote the gradient and Hessian, respectively, of \( \hat{H}_\theta \). These are both continuous with respect to any coordinate \( \theta_k \) as a consequence of (A4). Thus, both the gradient and the Hessian of the log-likelihood are continuous functions. These are given by \( \nabla \ell_n(\theta) := \hat{\ell}_n(\theta) \) and \( \nabla^2 \ell_n(\theta) := \sum_{i=1}^n \psi_i(\theta) \), respectively, where, for \( k = 1, \ldots, p \),
\[
(\hat{\ell}(\theta)(k)) = \sum_{i=1}^n \left( \frac{\nabla_k \hat{H}_\theta(W_i)}{\hat{H}_\theta(W_i)} - \lambda \int_0^{\hat{\lambda}_i} \nabla_k \hat{H}_\theta(W_{i-1} + X_{i-1} - u) du \right) 1_{\{\hat{H}_\theta(W_i) > 0\}},
\]  
(22)
and, for \( k, l = 1, \ldots, p \),
\[
(\psi_i(\theta)(k)) = \left( \frac{\hat{H}_\theta(W_i) \nabla^2_k \hat{H}_\theta(W_i) - \nabla_k \hat{H}_\theta(W_i) \nabla_l \hat{H}_\theta(W_i)}{\hat{H}_\theta(W_i)^2} - \lambda \int_0^{\hat{\lambda}_i} \nabla^2_k \hat{H}_\theta(W_{i-1} + X_{i-1} - u) du \right) 1_{\{\hat{H}_\theta(W_i) > 0\}}.
\]  
(23)

Following the lines of the standard delta method (van der Vaart 1998, chapter 3), we consider the expansion of \( \ell_n(\cdot) \) at the MLE \( \hat{\theta}_n \) around the true parameter:
\[
\frac{1}{n} \ell_n(\hat{\theta}_n) = \frac{1}{n} \ell_n(\theta_0) + (\hat{\theta}_n - \theta_0) \int_0^1 \frac{1}{n} \sum_{i=1}^n \psi_i(\alpha \hat{\theta}_n + (1 - \alpha) \theta_0) d\alpha.
\]  
(24)
If \( \theta_0 \in \Theta \), then because of the strong consistency that we found in Theorem 1, we have, because \( n \to \infty \), that
\[
\hat{\theta}_n \to_{as} \theta_0.
\]
Furthermore, because \( \theta_0 \in \Theta \) by assumption (A1), the smoothness assumption (A4) implies that, as \( n \) grows large, \( \ell_n(\theta) \) converges to a concave function, and the MLE is given by a sequence of roots \( \hat{\theta}_n \) satisfying \( \ell_n(\hat{\theta}_n) = 0 \); that is, from some \( N \) on there is no boundary solution for any \( n \geq N \) with probability one. Therefore, (24) yields, with
\[
B_n := -\int_0^1 \frac{1}{n} \sum_{i=1}^n \psi_i(\alpha \hat{\theta}_n + (1 - \alpha) \theta_0) d\alpha,
\]
that

\[ \lim_{n \to \infty} \sqrt{n}(\hat{\theta}_n - \theta_0)B_n \overset{d}{=} \lim_{n \to \infty} \frac{1}{\sqrt{n}} \hat{\ell}_n(\theta_0), \]  

(25)

that is, both sides of (25) converge to the same distribution (if the limits exist). In Lemma 2, we apply the stationarity of \( W_i \) in (A2) to show that \( B_n \overset{d}{=} I(\theta_0) \), where \( I(\theta_0) = -E\Psi_i(\theta_0) \). Furthermore, \( \ell_n(\theta_0)/\sqrt{n} \) is shown to satisfy a martingale CLT with asymptotic variance \( I(\theta_0) \). Because \( E\ell_i(\theta_0) = 0 \), Assumption (A2) implies that the asymptotic variance equals the stationary covariance of the gradient

\[ I(\theta_0) = EV^2(\ell_1(\theta_0)) = E[\nabla(\ell_1(\theta_0)\nabla(\ell_1(\theta_0))^\top], \]  

(26)

and hence, \( I(\theta_0) \) is a positive definite matrix (see Ferguson 1996, chapter 18, for more details). Assumption (A4) further demands that the elements of \( I(\theta_0) \) are all finite, and then combining the above and applying Slutsky’s theorem to (25) yields Theorem 2.

We are thus left with showing Lemma 2 below; its proof is given in the Appendix.

**Lemma 2.** If (A1)–(A5) hold and \( I(\theta_0) = -E\Psi_1(\theta_0) \) is a positive definite matrix with finite elements, then (a) as \( n \to \infty \), \( B_n \overset{d}{=} I(\theta_0) \), and (b) as \( n \to \infty \),

\[ \frac{1}{\sqrt{n}} \hat{\ell}_n(\theta_0) \overset{d}{\to} N(0, I(\theta_0)). \]  

(27)

### 3.2. Constant Patience-Level MLE

We next consider the case where all customers have the same patience level \( \theta_0 \), that is, \( H(y) = 1_{(y \geq \theta_0)} \); observe that in this case the stability condition (2) is always satisfied, but the continuity assumptions of Section 3.1 do not apply. As we will see, in this case the properties of the MLE are markedly different from those identified in Section 3.1.

In this setting, the likelihood function (14) reduces to, with the event \( \mathcal{E}_i \) denoting \( \{W_{i-1} + X_{i-1} \leq \theta\} \) and \( \mathcal{E}_i^c \) its complement,

\[ L(\theta; A, W|X) = \lambda^n \prod_{i=1}^{n} 1_{[W_i \leq \theta]} \left( 1_{[\mathcal{E}_i]} e^{-\lambda \int_{0}^{W_{i-1}} du} + 1_{[\mathcal{E}_i^c]} e^{-\lambda \int_{W_{i-1} + X_{i-1}}^{\theta} du} \right) \]

\[ = \lambda^n \prod_{i=1}^{n} 1_{[W_i \leq \theta]} \left( 1_{[\mathcal{E}_i]} e^{-\lambda W_i} + 1_{[\mathcal{E}_i^c]} e^{-\lambda(W_i + X_i - \theta)} \right). \]

Observe that \( L(\theta; A, W|X) \) (i) equals zero for \( \theta \in [0, \max_{i=1,2,\ldots,n} W_i] \), (ii) has an upward discontinuity at \( \theta = \max_{i=1,2,\ldots,n} W_i \) and decreases for

\[ \theta \in \left[ \max_{i=1,2,\ldots,n} W_i, \max_{i=1,2,\ldots,n} \{W_i + X_i\} \right], \]

and (iii) takes a constant value for \( \theta \geq \max_{i=1,2,\ldots,n} \{W_i + X_i\} \). We thus find the intuitively appealing property that the MLE is given by

\[ \hat{\theta}_n = \max_{i=1,2,\ldots,n} W_i, \]

(28)

that is, the maximum virtual waiting time at jump times. Note that this estimator resembles the MLE of the parameter \( \theta \) when the observations are uniformly distributed on \([0, \theta]\).

Although the estimator \( \hat{\theta}_n \) is clearly biased (which follows from \( E_{\theta_0}[\hat{\theta}_n] < \theta_0 \) for all \( n \)), we show that it converges almost surely to \( \theta_0 \) as \( n \to \infty \). In addition, we prove that the estimation error scaled at rate \( n \) converges to an exponential random variable.

**Theorem 3.** Because \( n \to \infty \),

\[ \hat{\theta}_n \overset{a.s.}{\to} \theta_0, \]

(29)
and

$$P(n(\theta_0 - \hat{\theta}_n) \geq x) \rightarrow e^{-v(\theta_0)x/(1-P_{\lambda})},$$

(30)

where $P_{\lambda}$ and $v(\theta_0)$ denote the stationary loss probability and the virtual waiting time density at level $\theta_0$. In addition, the asymptotic variance of the estimation error agrees with the variance of the limiting exponential distribution and is given by

$$\lim_{n \to \infty} n^2 \text{Var}[\theta_0 - \hat{\theta}_n] = \lim_{n \to \infty} \left\{ \mathbb{E}\left[\{n(\theta_0 - \hat{\theta}_n)^2 \} - \left(\mathbb{E}[n(\theta_0 - \hat{\theta}_n)]\right)^2 \right] \right\}$$

(31)

$$= \left(1 - P_{\lambda}\right)^2 \frac{\mathbb{E}[v(\theta_0)^2]}{\mathbb{E}[v(\theta_0)]}.$$  

(32)

The proof of Theorem 3, which is given in the Appendix, constructs lower and upper bounds for the MLE and establishes that they both converge to $\theta_0$. The same bounds are also used to characterize the asymptotic distribution of the estimation error.

### 3.3. Estimating the Arrival Rate

Where in the preceding subsections we focused on estimating the patience-level distribution (in a parametric context) for a given value of the arrival rate $\lambda$, we now discuss the estimation of $\lambda$. Let $\lambda_0$ denote the true value of $\lambda$. There are several ways to estimate it, the most basic one relying on the idle period observations. Denote the duration of the $k$-th idle period by $I_k$, where in the multiserver setting ($s > 1$) an idle period refers to time intervals such that at least one server is idle. Denote by $E_k$ the total number of effective arrivals during idle period $k = 1, 2, \ldots$ (meaning, for instance, $E_k = 1$ for all $k = 1, 2, \ldots$ when $s = 1$). Let $C_n$ denote the number of idle periods observed up until (and including) the $n$-th effective arrival. Then we propose the estimator

$$\hat{\lambda}_n = \frac{\sum_{k=1}^{C_n} E_k}{\sum_{k=1}^{C_n} I_k}.$$  

(33)

The rationale behind this estimator is that there is no balking during idle periods because the virtual waiting time is zero. Hence, the arrival process during these idle times is homogeneous Poisson with rate $\lambda$. Therefore, the estimator $\hat{\lambda}_n$ is a standard MLE of the rate parameter of an exponentially distributed random variable and satisfies all desired asymptotic properties.

Of course, the above procedure does not exploit a substantial amount of potentially useful data that is collected during busy periods. For the case of a deterministic patience level $\theta_0$, the above estimator is easily improved upon. Recall that $\hat{\theta}_n$ was defined as $\max\{W_1, W_2, \ldots, W_n\}$, with the immediate consequence that $\hat{\theta}_n \leq \theta_0$. Supposing we observe an arrival such that immediately after this arrival the virtual waiting time level is still below the current value of the estimator, then the next arrival is an effective arrival and hence, occurs after an exponentially distributed random variable with rate $\lambda$. In this way, we generate more observations, thus allowing us to estimate $\lambda$ with a better precision. Observe that the new observations and the idle times form an i.i.d. sequence.

For the case of a continuous patience-level distribution, one may use (33) or, alternatively, set up a joint MLE for $\lambda$ and $\theta$. The latter has clear advantages, in particular for small samples or heavily loaded systems. Because the log-likelihood function (17) is a smooth and concave function with respect to $\lambda$, the MLE of the arrival rate for any estimator $\hat{\theta}_n$ of $\theta$ is

$$\hat{\lambda}_n(\hat{\theta}_n) = \hat{\theta}_n \left\{ \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{\hat{\theta}_n} \mathbb{P}_{\phi_u}(W_{t-1} + X_{t-1} - u)du \mathbf{1}_{\{\phi_u(W_t) > 0\}} \right\}^{-1}. $$

(34)

The asymptotic results of Section 3.1 can, therefore, be extended in a straightforward manner so as to cover the joint estimation of $(\lambda, \theta)$. We demonstrate this in Section 4.1, where we detail a procedure for jointly estimating the arrival rate and the patience parameter for the case of an exponentially distributed patience level.
4. Exponential and Generalized Hyperexponential Patience

In this section, we discuss a robust and practical approach for estimating continuous patience distributions. In our approach, this is achieved by fitting the MLE of a generalized hyperexponential (GHE) distribution. This approach is attractive because the class of GHE distributions is known to be dense in the space of nonnegative continuous distributions (see, for instance, Botta and Harris 1986), which, in practical terms, means that any non-negative continuous distribution can be approximated arbitrarily closely by a GHE distribution. In our simulation-based experiments, highly accurate estimates are obtained, even when the baseline patience distribution itself is not GHE.

We first provide a detailed analysis of the joint MLE for the arrival rate and the single parameter of an exponential patience distribution to then move to the cases of hyperexponential and generalized hyperexponential patience distributions. We also present a heuristic search method that fits an approximate GHE distribution for any continuous patience distribution.

4.1. Exponentially Distributed Patience

Suppose that the arrival rate $\lambda_0$ is unknown and that the patience-level distribution is exponential, that is, $H_0(x) = 1 - e^{-\theta_0 x}$ for an unknown $\theta_0$ (and $x \geq 0$). Because of (17), it can be verified, by distinguishing between the cases $W_0 + X_0 \geq A_1$ and $W_0 + X_0 < A_1$, that the log-likelihood for a single observation is

$$
\ell_1(\lambda, \theta) = \log \lambda - \theta W_1 - \lambda \int_0^{A_1} e^{-\theta (W_0 + X_0 - u)} du
$$

$$
= \log \lambda - \theta W_1 - \lambda \left( e^{-\theta W_1} - e^{-\theta (W_0 + X_0)} + \theta (A_1 - W_0 - X_0)^+ \right);
$$

here the Lindley recursion (13) has been used.

Our objective is to analyze the MLE of both $\lambda$ and $\theta$ for a sample of size $n$, which we denote by $(\hat{\lambda}_n, \hat{\theta}_n)$. We assume that (A1) and (A2) hold, that is, compact parameter space and stationary $W_0$. In addition, (A3) holds because $H_0$ is exponential for all $\theta \in \Theta$. The log-likelihood $\ell_1(\lambda, \theta)$ is Lipschitz continuous (with respect to the observations), and so (A5) holds. We thus have that Theorem 1 holds and the MLE for $(\lambda, \theta)$ is strongly consistent.

Clearly, this is a smooth function. We proceed by computing the gradient and Hessian with respect to $(\theta, \lambda)$. It takes some elementary calculus to verify that gradient is given by

$$
\nabla(\ell_1(\lambda, \theta))_\lambda = \frac{1}{\lambda} \left( e^{-\theta W_1} - e^{-\theta (W_0 + X_0)} + \theta (A_1 - W_0 - X_0)^+ \right),
$$

(35)

$$
\nabla(\ell_1(\lambda, \theta))_\theta = -W_1 - \lambda \left[ (\theta (W_0 + X_0) + 1)e^{-\theta (W_0 + X_0)} - (\theta W_1 + 1)e^{-\theta W_1} \right].
$$

(36)

For any given value of $\theta$, $\ell_1(\lambda, \theta)$ is a concave function in $\lambda$, and therefore, the MLE of (34) is given by

$$
\hat{\lambda}_n(\theta) = \left( \frac{1}{n} \sum_{i=1}^{n} \left( e^{-\theta W_i} - e^{-\theta (W_{i-1} + X_{i-1})} + \theta (A_i - W_{i-1} - X_{i-1})^+ \right) \right)^{-1} \theta.
$$

(37)

The MLE $\hat{\theta}_n$ now follows by maximizing $\frac{1}{n} \sum_{i=1}^{n} \ell_1(\theta, \hat{\lambda}_n(\theta))$ for $\theta \in \Theta$. The optimizing $\theta$ is obtained either on the boundary of the parameter space $\Theta$ or by solving the first-order condition $\nabla(\ell_1(\theta, \hat{\lambda}_n(\theta)))_\theta = 0$. Because $n \to \infty$, we are guaranteed to find an interior solution as long as $\theta_0 \in \Theta^*$.

Taking second derivatives, we find that the entries of the Hessian matrix are given by

$$
\nabla^2(\ell_1(\lambda, \theta))_{\lambda, \lambda} = -\frac{1}{\lambda^2}.
$$

$$
\nabla^2(\ell_1(\lambda, \theta))_{\lambda, \theta} = \nabla^2(\ell_1(\lambda, \theta))_{\theta, \lambda} = -\frac{1}{\theta^2} \left[ (\theta (W_0 + X_0) + 1)e^{-\theta (W_0 + X_0)} - (\theta W_1 + 1)e^{-\theta W_1} \right].
$$

$$
\nabla^2(\ell_1(\lambda, \theta))_{\theta, \theta} = -\frac{1}{\theta^3} \left[ (1 + (\theta W_1 + 1)^2)e^{-\theta W_1} - (1 + (\theta (W_0 + X_0) + 1)^2)e^{-\theta (W_0 + X_0)} \right].
$$
Table 1. Confidence Intervals for the MLEs ($\hat{\theta}_n, \hat{\lambda}_n$) as Well as the Arrival Rate Estimated from Idle Periods $\tilde{\lambda}_n$ for Different Confidence Levels

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Load 80%</th>
<th>Load 90%</th>
<th>Load 95%</th>
<th>Load 99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\theta}_n$</td>
<td>[0.423,0.592]</td>
<td>[0.403,0.620]</td>
<td>[0.386,0.645]</td>
<td>[0.352,0.691]</td>
</tr>
<tr>
<td>$\hat{\lambda}_n$</td>
<td>[0.454,0.455]</td>
<td>[0.441,0.571]</td>
<td>[0.431,0.586]</td>
<td>[0.407,0.612]</td>
</tr>
<tr>
<td>$\tilde{\lambda}_n$</td>
<td>[0.569,0.536]</td>
<td>[0.460,0.546]</td>
<td>[0.453,0.555]</td>
<td>[0.440,0.573]</td>
</tr>
<tr>
<td>$\hat{\lambda}_n$</td>
<td>[0.958,1.048]</td>
<td>[0.948,1.061]</td>
<td>[0.937,1.074]</td>
<td>[0.918,1.099]</td>
</tr>
<tr>
<td>$\tilde{\lambda}_n$</td>
<td>[0.955,1.053]</td>
<td>[0.944,1.067]</td>
<td>[0.933,1.081]</td>
<td>[0.912,1.104]</td>
</tr>
<tr>
<td>$\hat{\lambda}_n$</td>
<td>[0.951,1.061]</td>
<td>[0.936,1.079]</td>
<td>[0.924,1.096]</td>
<td>[0.901,1.128]</td>
</tr>
<tr>
<td>$\tilde{\lambda}_n$</td>
<td>[0.953,1.052]</td>
<td>[0.941,1.069]</td>
<td>[0.929,1.082]</td>
<td>[0.909,1.110]</td>
</tr>
<tr>
<td>$\hat{\lambda}_n$</td>
<td>[0.944,1.063]</td>
<td>[0.928,1.083]</td>
<td>[0.915,1.101]</td>
<td>[0.890,1.134]</td>
</tr>
<tr>
<td>$\tilde{\lambda}_n$</td>
<td>[0.928,1.085]</td>
<td>[0.908,1.110]</td>
<td>[0.892,1.134]</td>
<td>[0.862,1.182]</td>
</tr>
</tbody>
</table>

Notes. A total of $n = 1000$ waiting observations were generated, $M = 100000$ times for each parameter setting. The parameter of the exponential patience parameter was $\theta = 0.5$, the arrival rate was $\lambda = 1$, and the service requirements followed a gamma distribution with parameters $(\eta, \mu)$ with $\mu = 1$ and $\eta \in \{0.5, 1, 2\}$.

Because $w^k e^{-w\theta}$ is a bounded function on $w \in [0, \infty)$ for any $\theta > 0$ and $k \in \mathbb{N}$, we conclude that $-\text{EV}^2(\ell_1(\lambda_0, \theta_0)) < \infty$, and therefore, (A4) applies, which implies that Theorem 2 holds. In particular, because $n \to \infty$,

$$\sqrt{n} \left( \hat{\lambda}_n - \lambda_0 \right) \to_d N(0, I(\lambda_0, \theta_0)),$$

and the asymptotic covariance is given by the inverse of the Hessian,

$$I(\lambda_0, \theta_0) = -\text{EV}^2(\ell_1(\lambda_0, \theta_0)).$$

We performed numerical experiments to assess the performance of the estimation procedure. Table 1 presents approximated confidence intervals for the maximum-likelihood estimators ($\hat{\lambda}_n, \hat{\theta}_n$); in addition, we also evaluated the estimator of the arrival rate based on idle periods, denoted by $\tilde{\lambda}_n$, as discussed in Section 3.3. The confidence intervals are evaluated for three congestion levels, namely $\rho = \lambda \mathbb{E} \in \{0.5, 1, 2\}$. In all experiments, the sample size (of observed waiting times) was $n = 1000$. Evidently, however, the heavier the system load, the fewer the number of idle periods; on average, $C_n = C_{1000}$ equals 665, 457, and 267 for the three congestion levels.

The numerical output is summarized in Table 1. A first observation is that the MLE for $\hat{\theta}_n$ is more accurate as the system load increases. This is because the patience threshold is reached more often, and therefore, the data are more informative. Furthermore, for a lightly loaded system, the MLE $\hat{\lambda}_n$ and idle-period based estimator $\tilde{\lambda}_n$ yield similar (accurate) results, as opposed to the high-load regime, in which the two estimators behave quite differently:

- As a consequence of the fact that idle periods are observed considerably less frequently for a high load, $\tilde{\lambda}_n$ becomes substantially less accurate.
- The accuracy of $\hat{\lambda}_n$, however, is only slightly reduced in the high-load regime. This is potentially because of the better accuracy of the estimation of $\theta$, being jointly estimated with $\lambda$ in the MLE procedure.

In the most heavily loaded example ($\rho = 2$, that is), the 95% confidence interval of the MLE is very similar to the 80% confidence interval of the idle period-based estimator. We conclude that even though the observations of the waiting times and the effective interarrival observations are highly dependent, the MLE $\hat{\lambda}_n$ provides substantially better confidence intervals than the idle period-based estimator $\tilde{\lambda}_n$.

Table 2 presents empirical confidence intervals for the maximum-likelihood estimators of $(\lambda, \theta)$ for a multi-server system with $s = 5$. The confidence intervals are wider than in the single-server case, even though the sample size was taken twice as big, being indicative of the fact that the variance of the estimation error in a system with multiple servers is higher. As in the single-server case, lower system load increases the variance of the estimation error for the patience parameter $\theta$. This effect is even stronger in the multiserver setting because of the fact that balkling occurs only when all five servers are working, and such a state is not frequently observed. For higher load, the estimation of the patience parameter is much more accurate, as was the case for the single-server system. For the arrival rate, we observe high accuracy for all load levels.

4.2. (Generalized) Hyperexponential Patience

This subsection focuses on the case of the patience-level distribution being generalized hyperexponential (GHE), meaning that the corresponding cdf can be written as a mixture of exponential cdfs with weights that sum to 1.
where the weights are assumed to be positive. This case is particularly relevant because, as argued in Roubos and Jouini (2013), this distribution has a good empirical fit-to-patience data. Another motivation for considering this distribution lies in the known fact that the cdf of any continuous positive random variable can be approximated arbitrarily accurately (in terms of a suitably defined metric) by a GHE cdf (Botta and Harris 1986).

For convenience, we now assume the arrival rate \( \lambda \) to be known, but it is noted that it can be estimated in a similar manner as in the exponential case discussed in Section 4.1. Suppose the degree of the GHE distribution is \( p \in \mathbb{N} \):

\[
\Pi_\theta(x) = \sum_{k=1}^{p} \alpha_k e^{-\beta_k x}, x \geq 0,
\]

where \( \theta = (\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{2p} \), \( \sum_{k=1}^{p} \alpha_k = 1 \) (where, importantly, we do not assume the \( \alpha_k \) to be positive) and \( \beta_k > 0 \) for \( k = 1, \ldots, p \). Without loss of generality, we assume that \( \alpha_1 > \alpha_2 > \cdots > \alpha_p \). Denote the coordinates of the true parameter \( \theta_0 \) by \( (\alpha_{0k}, \beta_{0k}) \) for \( k = 1, \ldots, p \). It means that we are to identify the \( 2p \)-dimensional parameter vector

\[
\theta_0 = (\alpha_{01}, \ldots, \alpha_{0p}, \beta_{01}, \ldots, \beta_{0p}).
\]

Similarly to the previous example of the exponential distribution, Assumptions (A1)–(A3) and (A5) are satisfied when assuming stationary waiting times and a compact parameter space \( \Theta \). Therefore, the MLE is strongly consistent by Theorem 1. One should be cautious when fitting a GHE distribution because of further conditions to be imposed on the parameters to make sure that \( \Pi_\theta(\cdot) \) is a proper cdf. Even though convergence to the true parameters of the distribution is eventually guaranteed, for finite samples the estimated parameters may not yield a proper distribution. For an in-depth discussion on these conditions and the identifiability of GHE distributions, we refer to Botta and Harris (1986) (Section 3).

Using the representation (17), the log-likelihood pertaining to a single observation is calculated in a similar manner as in the exponential case, yielding

\[
\ell_1(\theta) = \log \lambda + \log \left( \sum_{k=1}^{p} \alpha_k e^{-\beta_k W_i} \right) - \frac{\lambda}{\beta_j} \left( e^{-\beta_j W_i} - e^{-\beta_j(W_0 + X_0)} \right) - \lambda(A_1 - W_0 - X_0)^+.
\]

As follows with some elementary algebra, the gradient is then given by, for \( j = 1, \ldots, p \),

\[
\nabla(\ell_1(\theta))_{\alpha_j} = \frac{e^{-\beta_j W_i}}{\sum_{k=1}^{p} \alpha_k e^{-\beta_k W_i}} - \frac{\lambda}{\beta_j} \left( e^{-\beta_j W_i} - e^{-\beta_j(W_0 + X_0)} \right),
\]

\[
\nabla(\ell_1(\theta))_{\beta_j} = -\frac{W_i \alpha_j e^{-\beta_j W_i}}{\sum_{k=1}^{p} \alpha_k e^{-\beta_k W_i}} - \frac{\lambda A_1}{\beta_j} \left( [\beta_j (W_0 + X_0) + 1] e^{-\beta_j(W_0 + X_0)} - (\beta_j W_i + 1) e^{-\beta_j W_i} \right).
\]

The Hessian can be derived by computing the matrix of second derivatives. For the evaluation of the asymptotic covariance, a convenient alternative is to apply (26), that is,

\[
I(\theta_0) = -\mathbb{E} \nabla^2(\ell_1(\theta_0)) = \mathbb{E}[\nabla(\ell_1(\theta_0)) \nabla(\ell_1(\theta_0))^\top],
\]
so as to avoid the symbolic evaluation of the matrix of second derivatives. The entries of \( I(\theta_0) \) are finite because these are combinations of terms of the following types:

- Products of polynomials (of degree at most 2) and exponentials. For example, we come across a term that is, up to a multiplicative constant, \( W_0^2 e^{-\beta_0 W_0} \). Observe that the mapping \( x \mapsto x^p e^{-x} \) is bounded for \( x \geq 0 \).
- Ratios of the form, for example, up to a multiplicative constant

\[
\frac{W_1 e^{-\beta_1 W_1}}{\sum_{k=1}^{p} \alpha_k e^{-\beta_k W_1}} \leq W_1.
\]

In addition, it is verified that if the patience-level distribution is light-tailed (which is the case for the GHE distribution), then the stationary waiting time \( W \) is also light-tailed and has finite moments. Indeed, note that (3) implies

\[ v(x) \leq \lambda \pi_0 + \lambda \int_0^\infty v(y)\overline{H}(y)\,dy = \lambda (1 - F_1), \]

so that we have from (4) that the density of \( W \) is bounded as \( w(x) \leq \lambda \overline{H}(x) \); that is, if the patience level distribution is light-tailed, so is \( W \). As a consequence, from the above and Theorem 2, the estimation errors (that is, scaled by \( \sqrt{n} \)) converge to a multivariate normal distribution.

Although there are theoretical guarantees for the asymptotic performance of the MLE, computation of the MLE is not straightforward, even for small parameter spaces such as \( p = 3 \). It requires maximizing \( \ell_n(\theta)/n \), that is, solving a nonlinear and nonconcave program in \( p \times (p - 1) \) variables, with both equality and inequality constraints. This is computationally highly challenging, and standard optimization methods may lead to local maxima. In particular, observe that \( \nabla (\ell_1(\beta)) \) is very small for large values of \( \beta \), which implies that \( \ell_n(\theta)/n \) displays very “flat” behavior for large values of these \( \beta \). Our experiments revealed that a direct implementation using standard optimization packages often led to points that were even not local maxima. As a consequence, we decided to write dedicated code to compute the MLE.

In light of the inherent complexity of maximizing the likelihood function, in our numerical experiments we have applied the following nested two-step heuristic optimization method.

(a) For each vector of \( \alpha = (\alpha_1, \ldots, \alpha_p) \), the objective function \( \ell_n(\theta)/n \) was maximized with respect to \( \beta = (\beta_1, \ldots, \beta_p) \) using a conventional coordinate descent algorithm. This step was typically fast and accurate; as for given \( \alpha \), the objective function behaves nicely. Now we have reduced the problem to an optimization over \( \alpha \).

(b) Then, a standard L-BFGS quasi-Newton method (see e.g., Byrd et al. 1995) is applied to \( \ell_n(\theta)/n \), so as to search for the optimal vector \( \alpha \), with \( \beta \) being parameterized by \( \alpha \). The optimization is carried out with the following constraints that ensure that the parameters yield a proper distribution (Botta and Harris 1986):

\[
\left\{ \alpha : \alpha_1 > \alpha_2 > \cdots > \alpha_p, \sum_{i=1}^{p} \alpha_i = 1, \sum_{i=1}^{p} \alpha_i \beta_i > 0 \right\}.
\]

Note that there is no firm guarantee that this heuristic method converges to the optimal parameters. To overcome this, the search for the optimizing vector \( \alpha \) was conducted multiple times for different initial values. The resulting method turned out to be time-consuming, especially for a relatively large sample size \( n \) and/or a relatively large dimension \( p \) of the parameter space. However, it typically returns considerably more robust results than off-the-shelf optimization routines (that in addition tended to converge very slowly).

Table 3 presents the marginal confidence intervals obtained by the normal approximation for an example of a HE cdf \( H_p(\cdot) \) with \( p = 2 \). Observe that the variance of the estimation error is large, even for a substantial sample size. The accuracy is much higher for the low rate of \( \beta_1 = 0.25 \) than the higher rate of \( \beta_2 = 1 \). This may be explained by the fact that the likelihood function is almost flat for high values of \( \beta_2 \).

The results presented in Table 3 show that the estimates are concentrated around the true values. The variance of the estimation error, however, is high, even more so given the large number of observations used. It should be realized, though, that HE distributions are not always easily distinguishable, in the sense that seemingly different HE distributions may lead to a very similar cdf. Clearly, from a practical standpoint, the main question is not whether the correct parameters are recovered but rather whether the cdf corresponding to the estimated parameters provides a good fit with the true cdf. In this respect, the main conclusion of our experiments is that our MLE procedure performs remarkably well. For example, in Figure 4, we illustrate this by
In the example is \( \lambda = 1 \), and the service requirements followed a gamma distribution with parameters \((3, 2)\). The motivation behind this (i.e., in misspecified) HE distribution, rather than a GHE distribution, by computing the MLE. The motivation behind this choice is that, when working with a GHE distribution, there is the additional complication that the estimated parameters may not yield a proper cdf. More specifically, the space of valid parameters is hard to characterize, let alone to be coded in terms of constraints of an optimization problem. Nevertheless, we would still like to exploit the extra versatility that the GHE class offers so as to improve the fit (when compared with the HE MLE). This is especially relevant for scenarios in which the model is misspecified, bearing in mind that, as mentioned earlier, in principle any distribution can be arbitrarily accurately approximated by a GHE distribution.

With the above considerations in mind, we implemented the following (seemingly naïve) heuristic model selection method. We generate various random weight vectors \( \alpha \) (equipped with a random size \( p \in \mathbb{N} \)). Then, for each of them, we optimize the likelihood function over \( \beta \) using only step (a) above. The best model is then chosen by comparing the various combinations of \( \alpha \) and \( p \), relying on the Akaike Information Criterion (AIC). The AIC encompasses both the log-likelihood and, in order to avoid overfitting, a penalty for the number of parameters (Burnham and Anderson 2002). Because step (a) can be performed highly efficiently, a main advantage of this heuristic is that it works very fast, even for bigger values of \( p \), thus providing us with a

| Table 3. Confidence Intervals for the MLEs \( (\hat{\alpha}_1, \hat{\beta}_1, \hat{\beta}_2) \) Based on a Normal Approximation for a Sample Size of \( n = 100000 \) for Different Confidence Levels |

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Load</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\alpha}_{1n} )</td>
<td>( \rho = 0.5 )</td>
<td>([0.500,0.908])</td>
<td>([0.500,0.967])</td>
<td>([0.500,1.000])</td>
<td>([0.500,1.000])</td>
</tr>
<tr>
<td></td>
<td>( \rho = 1 )</td>
<td>([0.609,0.791])</td>
<td>([0.583,0.817])</td>
<td>([0.560,0.840])</td>
<td>([0.516,0.884])</td>
</tr>
<tr>
<td>( \hat{\beta}_{1n} )</td>
<td>( \rho = 0.5 )</td>
<td>([0.183,0.317])</td>
<td>([0.164,0.336])</td>
<td>([0.147,0.353])</td>
<td>([0.115,0.385])</td>
</tr>
<tr>
<td></td>
<td>( \rho = 1 )</td>
<td>([0.223,0.277])</td>
<td>([0.216,0.284])</td>
<td>([0.209,0.291])</td>
<td>([0.196,0.304])</td>
</tr>
<tr>
<td>( \hat{\beta}_{2n} )</td>
<td>( \rho = 0.5 )</td>
<td>([0.240,0.260])</td>
<td>([0.237,0.262])</td>
<td>([0.235,0.265])</td>
<td>([0.230,0.270])</td>
</tr>
<tr>
<td></td>
<td>( \rho = 1 )</td>
<td>([0.768,1.232])</td>
<td>([0.702,1.298])</td>
<td>([0.639,1.355])</td>
<td>([0.620,0.780])</td>
</tr>
<tr>
<td>( \hat{\beta}_{4n} )</td>
<td>( \rho = 0.5 )</td>
<td>([0.870,1.130])</td>
<td>([0.833,1.167])</td>
<td>([0.801,1.199])</td>
<td>([0.739,1.261])</td>
</tr>
<tr>
<td></td>
<td>( \rho = 2 )</td>
<td>([0.500,0.908])</td>
<td>([0.500,0.967])</td>
<td>([0.500,1.000])</td>
<td>([0.500,1.000])</td>
</tr>
</tbody>
</table>

Note. The true values of the parameters \( (\alpha_1, \alpha_2, \beta_1, \beta_2) \) of the HE patience were \((0.7, 0.3, 0.25, 1)\), the arrival rate was \( \lambda = 1 \), and the service requirements followed a gamma distribution with parameters \((\eta, \mu)\) with \( \mu = 1 \) and \( \eta \in \{0.5, 1, 2\} \).

The fitted functions are fitted to the data by computing the MLE. The motivation behind this choice is that, when working with a GHE distribution, there is the additional complication that the estimated parameters may not yield a proper cdf. More specifically, the space of valid parameters is hard to characterize, let alone to be coded in terms of constraints of an optimization problem. Nevertheless, we would still like to exploit the extra versatility that the GHE class offers so as to improve the fit (when compared with the HE MLE). This is especially relevant for scenarios in which the model is misspecified, bearing in mind that, as mentioned earlier, in principle any distribution can be arbitrarily accurately approximated by a GHE distribution.

With the above considerations in mind, we implemented the following (seemingly naïve) heuristic model selection method. We generate various random weight vectors \( \alpha \) (equipped with a random size \( p \in \mathbb{N} \)). Then, for each of them, we optimize the likelihood function over \( \beta \) using only step (a) above. The best model is then chosen by comparing the various combinations of \( \alpha \) and \( p \), relying on the Akaike Information Criterion (AIC). The AIC encompasses both the log-likelihood and, in order to avoid overfitting, a penalty for the number of parameters (Burnham and Anderson 2002). Because step (a) can be performed highly efficiently, a main advantage of this heuristic is that it works very fast, even for bigger values of \( p \), thus providing us with a

Figure 4. Tail Distribution of HE Customer Patience Compared with the Estimated Counterparts

Notes. The true values of the parameters \( (\alpha_1, \alpha_2, \beta_1, \beta_2) \) are given by \((0.7, 0.3, 0.25, 1)\). The fitted distributions \( H_{\hat{\beta}_{it}}(t) \), with \( i = 1, 2, 3, 4 \), are based on independent samples of \( n = 100000 \) observations with the HE MLE for \( p = 2 \). The MLE parameters of the fitted functions are \( \hat{\beta}_{11} = (0.529, 0.471, 0.193, 0.736) \), \( \hat{\beta}_{12} = (0.884, 0.116, 0.297, 1.969) \), \( \hat{\beta}_{13} = (0.75, 0.25, 0.269, 1.031) \), and \( \hat{\beta}_{14} = (0.573, 0.427, 0.212, 0.746) \). The arrival rate in the example is \( \lambda = 1 \), and the service requirements are gamma distributed with parameters \((3, 2)\).
technique to fit general continuous distributions. Recall from Botta and Harris (1986) that if we order the components of the weight vector (i.e., $\alpha_1 > \alpha_2 > \cdots > \alpha_p$), then a sufficient condition for $(\alpha, \beta)$ to yield a cdf is that $\alpha_1 > 0$ and $\sum_{i=1}^{p} \alpha_i \beta_i > 0$; we use this principle to select feasible solutions produced by the above heuristic.

Figure 5, (a) and (c), illustrates the estimated survival function of using the HE MLE with varying weights for three examples of patience-level distributions. In the first example, the patience-level distribution is indeed HE with $p=4$, whereas the second and third examples are misspecified (corresponding to a lognormal and gamma patience-level distribution, respectively). For each distribution, the MLE was computed using $p=1, 2, 4$ (where $p=3$ has been left out because it is barely distinguishable from $p=4$).

• As was the case in the setting of Figure 4, the experiments corresponding to the HE patience-level distribution show that, even for $p=4$ and as many as $n=10000$ waiting-time observations, the MLE does not accurately capture the true parameters. Nevertheless, the fit in terms of the cdf, as displayed in Figure 5(a), is remarkably good. We in addition performed the GHE fitting heuristic, which also provides a highly accurate fit and is considerably faster to compute than the MLE for $p=4$.

• The case of lognormal patience, as illustrated by Figure 5(b), shows that the fit is quite good for all estimators, except for $p=1$ (i.e., an exponential distribution).

• Figure 5(c) presents the fitted distributions for gamma patience. The fit of the HE MLE is decent, but importantly, in this case the GHE heuristic performs considerably better than the HE MLE (for $p=1, 2, 4$). The GHE heuristic selects a model with $p=10$ weights.

The above observations are replicated for a multiserver system with $s=10$. Figure 6, (a) and (b), plots the true and estimated patience distribution for lognormal and gamma distributions, respectively, using the misspecified hyperexponential MLE. As in the single-server case, the GHE approximation yields an accurate fit.
5. Applications

In this section, we discuss a series of applications in which our estimation methodology can be directly used.

5.1. Unknown Proportion of Impatient Customers

In this subsection, we consider a system where an unknown proportion \( \theta \in (0, 1) \) of the arriving customers has a known deterministic patience threshold \( w > 0 \). The other customers are patient and always join the queue. Suppose one wishes to estimate the fraction \( \theta \) from data.

To this end, first observe that

\[
H_{\theta}(x) = \begin{cases} \theta x & \text{if } x \leq w \\ 1 & \text{if } x > w \end{cases}.
\]

Assumption (A4) is not satisfied because \( H_{\theta}(x) \) is not continuous (in \( x \)) at \( x = w \). However, with respect to \( \theta \), we do have that \( H_{\theta}(x) \) is continuous. In addition, \( \ell_{\theta}(\theta) \) is concave, and therefore, the MLE is given by the first-order condition \( \ell_{\theta}'(\hat{\theta}) = 0 \). In this case, we can apply results from Heijmans and Magnus (1986) to establish strong consistency. In particular, the smoothness of \( \ell_{\theta}'(\theta) \) enables direct verification of the sufficient condition (21). Asymptotic normality then follows by verifying the conditions of Theorem 2 directly.

5.2. Noisy Delay Messages

Suppose that the patience threshold is a constant \( \theta \), but the customers do not observe their exact waiting time but rather some noisy estimate \( W_e = e(W) \) for some random function \( e \) such that \( E[W_e | W] = W \). We assume that customers join based on this “perturbed delay” \( W_e \) that is, if \( W_e \leq \theta \). Suppose we are in the context that the parameters underlying the noise distribution are known, but the threshold \( \theta \) is not.

The probability of joining at virtual waiting time level \( W \) can be computed given the specific noise distribution. A few examples are as follows:

- **Additive perturbations.** In this case, \( W_e = W + e \). One could, for instance, consider normally distributed perturbations: \( e \sim N(0, \sigma^2) \), independent of \( W_e \) with \( \sigma > 0 \). Let customers facing \( W_e < \theta \) join the system. We thus have, with \( \Phi(\cdot) \) the cdf of the standard normal distribution,

\[
\Phi(\theta - v) = \Phi\left(\frac{\theta - v}{\sigma}\right).
\]

The asymptotic properties of Section 3.1 hold in this case because \( \Phi(\cdot) \) satisfies the regularity conditions.
• Multiplicative perturbations. Now, \( W_c = \mathbb{W}_G \), with \( G \) nonnegative unit-mean and independent of \( W \). In this case,

\[
H_\theta(v) = P(W_c \leq \theta | W = v) = P(G \leq \frac{\theta}{v}).
\]

Now, if the random variable \( G \) is such that the regularity conditions of Section 3.1 are met, then it follows that the asymptotic properties hold in this case as well (with an asymptotic variance that can be expressed in terms of \( G \)).

5.3. Admission Pricing

Our estimation procedure can be exploited in the context of various problems rooted in operations research. Evidently, when having estimates of the arrival rate and the patience-level distribution, one could consider the option of increasing the service rate so as to potentially raise profits. Thus, the operational decision to be made is whether the increased revenues outweigh the cost of speeding up the service rate. In this subsection, we consider another operations research-related problem.

Suppose that the queue has an admission price of \( p \) and that customers are homogenous with a utility function as featured in the Naor (1969) model. More concretely, let there be constants \( r \) and \( c \) such that a customer will join the queue only if the virtual waiting time \( w \) upon the customer’s arrival satisfies

\[
r - p - cw \geq 0,
\]

or, alternatively, \( w \) is smaller than the threshold value \( (r - p)/c \). Now observe that for any fixed price \( p \) the threshold \( \hat{\theta}(p) = (r - p)/c \) can be estimated using the MLE procedure presented in Section 3.2. If one of the cost function parameters \( r \) and \( c \) is known, then the other parameter can be estimated directly. If both are unknown, then their estimation can be performed by an exploration procedure; set two prices, say, \( p_1 \) and \( p_2 \), and observe the system for each price. Supposing each of these two experiments is done with \( n \) clients, we let \( \hat{\theta}_n(p_i) \) be the MLE for price \( i = 1, 2 \). Then the estimators for the cost function parameters \( r \) and \( c \) are given by, respectively,

\[
\hat{r}_n = \frac{\hat{\theta}_n(p_1)p_2 - \hat{\theta}_n(p_2)p_1}{\hat{\theta}_n(p_1) - \hat{\theta}_n(p_2)}, \quad \hat{c}_n = \frac{\hat{r}_n - p_1}{\hat{\theta}_n(p_1)}.
\]

Theorem 3, in combination with the continuous mapping theorem, implies that, because \( n \to \infty \), both \( \hat{r}_n \to_{as} r \) and \( \hat{c}_n \to_{as} c \). Furthermore, for large \( n \), confidence intervals for the stationary average revenue per unit of time can be approximated using (30). In particular, observing that the loss probability depends on the price, in self-evident notation,

\[
n(\theta(p_i)) - \hat{\theta}_n(p_i)) \to_{d} \text{Exp}\left[\frac{v_i(\theta(p_i))}{1 - P_i(p_i)}\right], \quad i = 1, 2,
\]

because \( n \to \infty \). This opens the possibility of approximating, for \( n \) large, the (joint) distribution of \( n(\hat{r}_n - r_0, \hat{c}_n - c_0) \). Note that the loss probability \( P_i \) needs to be estimated as well. For the single-server case, this can be done by estimating the idle probability upon arrival for a given price and using (5).

6. Concluding Remarks

This paper has considered a service system in which clients potentially balk based on the virtual waiting time level they face at arrival. The main objective concerned the development of a framework for estimating the arrival rate and patience-level distribution. Our approach resolves the complication that in our setup only nonbalking clients are observed. Distinguishing between the case of a continuous patience-level distribution and constant patience, we developed MLE estimators and quantified their asymptotic properties. Through a sequence of examples and ramifications, we have illustrated the performance and broad applicability of our findings.

An important next step could concern the extension to a nonstationary arrival process. For example, one may assume that the potential arrival process is a nonhomogeneous Poisson process with an arrival rate function that depends on time, that is, \( \{\lambda(t) : t \geq 0\} \). This means that when the virtual waiting time is \( v \), the effective arrival rate at time \( t \) is given by \( \lambda(t)(1 - H(v)) \), and the likelihood function in Section 2 can be updated accordingly. If the arrival rate function is known, then our estimation procedure for \( \theta \) essentially carries over, including its performance guarantees. If the arrival rate function is unknown and parametric assumptions about it are made, then the joint estimation of the arrival rate and patience parameters is possible. Furthermore, in the practically relevant case that the time-dependent arrival rate is cyclic (i.e, \( \lambda(t) = \lambda(t + s) \) for all \( t \geq 0 \) and some cycle length \( s \geq 0 \); think of daily or weekly patterns), then the queueing process will still have regenerative dynamics that can be estimated using the MLE technique described in this paper.
exploited for asymptotic analysis. It is thus anticipated that the framework laid out in this paper can serve as a basis for developing estimation techniques for a wide class of nonstationary models.

Although we have developed estimation procedures assuming the independence between the customers’ service requirements and patience levels, one could consider an extension that allows dependence. Such an extension could, for example, model the situation in which customers with larger service requirements can be assumed to have more patience. Multiclass queueing models can be used to capture such dependence in service requirements and patience levels (Sakuma and Takeine 2017). More specifically, suppose that there are two customer classes 1 and 2, where class \( k \) \((k = 1, 2)\) customers arrive with rate \( \lambda_k \) and have service requirements (resp. patience levels), i.i.d. according to a class-dependent cdf \( G_k(·) \) (resp. \( H_k(·) \)). It is natural to assume that the customer class is unobservable by the estimator, thus yielding a sequence of dependent service requirements and patience levels. In this setting, we have to estimate the service-requirement cdf \( G_k(·) \) jointly with the patience-level cdf \( H_k(·) \) and the arrival rate \( \lambda_i \) to compute the likelihood function corresponding to an observed sequence \((A, W, X)\).

Another possible direction for future work concerns the situation in which each customer decides to balk based on his or her sojourn time, that is, the virtual waiting time seen by this customer just after (instead of before) his or her arrival. If the customer precisely knows his or her service requirement, this problem is easily reduced to the one considered in this paper. The other obvious option is that the customer balks if the virtual waiting time just before his or her arrival increased by a “guess” of his or her service requirement exceeds the customer’s patience threshold.

An alternative to the approach followed in the paper would be to pursue nonparametric estimation, cf. the results in Hansen and Pitts (2006) for the conventional case without balking. In the single-server case, it may be of help that we can write, with \( W(x) \) and \( D(x) \) denoting the cdf of the waiting times and sojourn times of non-balking customers, respectively,

\[
H(x) = \frac{w(x)}{\lambda(W(x) - D(x))}
\]

if \( w(x) > 0 \) and 0 else; the validity of (38) follows from combining (4) with a level-crossing identity. Whereas it is clear how to estimate \( W(x) \) and \( D(x) \) using the empirical distribution, estimation of \( w(x) \) is more challenging; kernel-based techniques may turn out useful in this context.

A final issue to consider is that in many systems customers do not observe exact waiting times but rather queue lengths. Specifically, this can be modeled as an \( M/G/s + H \) system with customers that join or balk after having observed the number of customers in the system. In this case, the patience level \( Y \) is a discrete random variable indicating at what queue length a customer is willing to join. Therefore, the interarrival process at any given queue length \( q \geq 0 \) is a Poisson process with rate \( \lambda_q = \lambda H(q) \). As was done in this paper, and MLE for the arrival rates (and the corresponding parameter \( \theta \)) can be derived from a sample of interarrival times and the respective queue lengths.

Acknowledgments

The authors are grateful to the associate editor and two referees for their helpful comments and feedback. The authors also thank Onno Boxma (Eindhoven University of Technology, The Netherlands), Avi Mandelbaum (Technion, Haifa, Israel), Ran Snitkovsky (Columbia University & CUHK Shenzhen), and Galit Yom-Tov (Technion, Haifa, Israel) for providing useful feedback.

Appendix: Proofs

Proof of Lemma 1. Assuming (A1)–(A3) and (A5), we verify the conditions for (20) as given in Ranga Rao (1962), (theorem 6.4). These are as follows: (i) \( \Theta \) is compact, (ii) \((A_n, W_n, X_n)_{n \geq 0}\) is a stationary ergodic sequence, (iii) the function \( \frac{1}{n} \ell_n(\theta) \) converges almost surely pointwise to \( \ell(\theta) \), (iv) the set \( \{\ell(\theta) : \theta \in \Theta\} \) is equicontinuous with respect to \((A, X, W)\), and (v) there exists a function \( K : [0, \infty)^3 \to \mathbb{R} \) such that

\[
\ell_1(\theta; a, z, y, x) \leq K(a, z, y, x), \quad \forall \theta \in \Theta; \quad \mathbb{E}[K(A_1, W_0, W_1, X_0)] < \infty.
\]  

(A.1)

Condition (i) is assumed in (A1). Assumption (A2) states that \( W_n \) is a stationary ergodic sequence, which also implies that the same is true for \( A_n \) and \( X_n \), thus yielding the validity of (ii). Note that conditional on \( W = z \) the distribution of \( A_i \) is determined by \( z \) and the external arrival process. Moreover, the upward-jump size is determined by the new service requirement together with the vector of residual service requirements in the system, which are assumed to be stationary and ergodic. Condition (iii) is (18). The equicontinuity assumption (A5) on \( H_0 \) implies that the same holds for the set of log-likelihood functions because both the log and the integral terms in (17) are continuous functions with respect to the observations \((A, X, W)\) so that we have established (iv). In the remainder of the proof, we verify (v).
Recall, from (A3), the definition \( \Pi_{\inf}(y) = \inf_{\theta \in \Theta} \Pi(y) \) for any \( y \geq 0 \). Because \( \Theta \) is compact and all functions \( H_\theta \) are continuous, we have that \( \Pi_{\inf}(\cdot) \) is a continuous monotone nonincreasing function such that \( \Pi_{\inf}(0) \in (0,1] \). Defining

\[
K(a,y) := |\log a| + |\log \Pi_{\inf}(y)| 1_{(\Pi_{\inf}(y) > 0)} + \lambda a,
\]

by recalling that \( H_\theta(y) \leq 1 \) for any \( y \geq 0 \), we conclude from (17) and the triangle inequality that

\[
|\ell_\ell(\theta; a, z, y, x)| \leq K(a,y), \quad \forall \theta, a, z, y, x \geq 0.
\]

Clearly, \( EA_1 < \infty \), and so to complete the proof we need to show that \( E[\log \Pi_{\inf}(W_1)] \leq \infty \). Recall that \( W_1 \sim W_0 \) is the stationary virtual waiting time seen by an effective arrival, whose pdf \( w(\cdot) \) is given by (cf. (4))

\[
w(y) = \frac{v(y)\Pi_{\inf}(y)}{1-P_v}.
\]

As mentioned in Section 3.1.1, (A5) ensures that the support of the distribution \( H_\theta \) does not depend on \( \theta \). The above equation thus implies \( W_0 < h_{\sup} \) with probability one. 

Note that we have

\[
E[\log \Pi_{\inf}(W_0)] = P(W_0 = 0)(-\log (\Pi_{\inf}(0))) + P(W_0 > 0)E[-\log (\Pi_{\inf}(W_0))|W_0 > 0],
\]

where the first term on the right-hand side is always finite. We thus provide a proof of \( E[-\log (\Pi_{\inf}(W_0))|W_0 > 0] < \infty \) below.

Regardless of whether \( h_{\sup} < \infty \) or \( h_{\sup} = \infty \), it follows from (A3) that for any \( \epsilon \in (0,\infty) \), there exist \( 0 < y_1^* < h_{\sup} \) and \( 0 < y_2^* < h_{\sup} \) such that

\[
\frac{e^{-f(y)}}{\Pi_{\inf}(y)} < c_1 + \epsilon, \quad y \geq y_1^*, \quad 0 < f(y)\Pi_{\inf}(y) < c_2 + \epsilon, \quad y \geq y_2^*.
\]

Using these inequalities, with \( y^* := \max\{y_1^*, y_2^*\} \), we obtain

\[
(1-P_v)E[-\log (\Pi_{\inf}(W_0))|W_0 > 0]
\]

\[
= (1-P_v) \int_{y_1^*}^{h_{\sup}} (-\log (\Pi_{\inf}(y)))w(y)dy
\]

\[
= \int_{y_1^*}^{h_{\sup}} (-\log (\Pi_{\inf}(y)))\Pi_{\inf}(y)v(y)dy
\]

\[
< \int_{y_1^*}^{h_{\sup}} (-\log (\Pi_{\inf}(y)))\Pi_{\inf}(y)v(y)dy + \int_{y_1^*}^{y_2^*} \left(-\log \left(\frac{e^{-f(y)}}{c_1 + \epsilon}\right)\right)\frac{c_2 + \epsilon}{f(y)} \cdot v(y)dy
\]

\[
\leq -\log (\Pi_{\inf}(y^*)) \int_{y_1^*}^{y_2^*} v(y)dy + \int_{y_1^*}^{y_2^*} (f(y) + |\log (c_1 + \epsilon)|) \frac{c_2 + \epsilon}{f(y)} \cdot v(y)dy
\]

\[
\leq -\log (\Pi_{\inf}(y^*)) + (c_2 + \epsilon) \int_{y_1^*}^{y_2^*} v(y)dy + \left|\log (c_1 + \epsilon)\right| \frac{c_2 + \epsilon}{f(y^*)} \int_{y_1^*}^{y_2^*} v(y)dy
\]

\[
\leq -\log (\Pi_{\inf}(y^*)) + (c_2 + \epsilon) \left(1 + \frac{|\log (c_1 + \epsilon)|}{f(y^*)}\right) < \infty,
\]

which implies \( E[-\log (\Pi_{\inf}(W_0))|W_0 > 0] < \infty \). □

**Proof of Lemma 2.** (a) With Lemma 1 and Theorem 1 at our disposal, the proof essentially follows standard arguments. For instance, see the proof of Ferguson (1996) (Theorem 18) for i.i.d. observations or the proof of Ravner et al. (2019) (Lemma 14) that applies a martingale CLT for an MLE based on dependent workload observations in the M/G/1-queue context.

(b) We construct a martingale CLT for stationary sequences (Billingsley 1999, Theorem 18.3) and verify that the corresponding conditions are satisfied. Let \( Z_i := \ell_i(\theta_0) - \ell_{i-1}(\theta_0) \) for \( i \in \{1,2,\ldots\} \), where \( \ell_0(\theta_0) := 0 \). The identifiability condition (19) implies that the smooth function \( \ell(\theta) \) is maximized at \( \theta_0 \) and hence, \( \ell(\theta_0) := \nabla \ell(\theta_0) = 0 \), that is, all coordinates equal zero. Therefore, if \( W_{i-1} \) is stationary, we have that

\[
E[(Z_i|W_{i-1}) = E[(\ell_1(\theta_0) - \ell_0(\theta_0)|W_0) = (\ell(\theta_0))_k = 0, \quad \forall k \in \{1,\ldots,p\},
\]
and
\[ Z_m := K_n^{-1} Z_i, \ i \in \{1, \ldots, n\}, \]
is a martingale difference, where \( K_n = \text{Var}[\hat{\ell}_n(\theta_0)] \) is the Cholesky decomposition of the covariance matrix of \( \hat{\ell}_n(\theta_0) \). Furthermore,
\[ \frac{1}{n} \text{Var}[\hat{\ell}_n(\theta_0)] = \frac{1}{n} \text{Var} \left[ \frac{1}{n} \hat{\ell}_n(\theta_0) \right] \to_{\text{as}} - E \Psi'_1(\theta_0) = I(\theta_0), \]
because \( n \to \infty \), and because \( I(\theta_0) < \infty \), by (A4) we have that \( \lim_{n \to \infty} (K_n) = \infty \) for all \( k = 1, \ldots, p \), and in addition,
\[ \lim_{n \to \infty} K_n = K, \]
where \( K \) is defined through \( KK^\top = I(\theta_0) \). Therefore by Heyde (1997), (theorem 12.6; which is the multi-dimensional counterpart of Billingsley 1999, theorem 18.1) we conclude that \( \Sigma_n Z_m \) converges to a standard normal random variable, and thus
\[ \frac{1}{\sqrt{n}} \hat{\ell}_n(\theta_0) = \frac{1}{\sqrt{n}} K_n \sum_{m=1}^{n} Z_m \to_{d} KN(0, 1), \]
which is equivalent to (27). \( \square \)

**Proof of Theorem 3.** This proof consists of a lower bound and an upper bound.

\( \triangleright \) **Lower bound.** Note that for \( \theta_0 > W_{k-1} \), we have
\[ A_k \sim \begin{cases} E_{1,kx}, & W_{k-1} + X_{k-1} \leq \theta_0, \\ W_{k-1} + X_{k-1} - \theta_0 + E_{1,kx}, & W_{k-1} + X_{k-1} > \theta_0, \end{cases} \tag{A.2} \]
where \( (E_{1,k}) \) denote a sequence of i.i.d. exponentially distributed random variables with parameter \( \lambda \). Let \( \mathcal{N}_n^+ \subset \{1, 2, \ldots, n\} \) denote the set of indices of observed virtual waiting times such that
\[ \mathcal{N}_n^+ = \{ k \in \{1, 2, \ldots, n\} : W_{k-1} + X_{k-1} \geq \theta_0 \}. \]
We see from (13) and (A.2) that given \( k \in \mathcal{N}_n^+ \), the observed virtual waiting time \( W_k \) is stochastically identical to \( \max (0, \theta_0 - X_k) \), where \( (X_k)_{k=0}^\infty \) denote a sequence of i.i.d. random variables that are exponentially distributed with parameter \( \lambda \). Therefore, \( (W_k)_{k \in \mathcal{N}_n^+} \) are i.i.d. with
\[ P(W_k \leq x | k \in \mathcal{N}_n^+) = e^{-\lambda (\theta_0 - x)}, \quad x \in [0, \theta_0], \tag{A.3} \]
so that
\[ P \left( \max_{i \in \mathcal{N}_n^+} W_i \leq x \right) = E \left[ \prod_{i \in \mathcal{N}_n^+} 1_{\{W_i \leq x\}} \right] = E \left[ e^{-\lambda (\theta_0 - x) |\mathcal{N}_n^+|} \right], \tag{A.4} \]
where \( |\mathcal{N}_n^+| \) denotes the number of elements in \( \mathcal{N}_n^+ \).

Obviously, we have \( \theta_0 \geq \theta_n \) and
\[ \hat{\theta}_n \geq \max_{i \in \mathcal{N}_n^+} W_i, \quad \text{a.s.} \tag{A.5} \]
Let, as before, \( c(\cdot) \) and \( w(\cdot) \) denote the pdfs of the stationary virtual waiting time and the stationary waiting time of non-balking customers, and let \( P_l \) denote the stationary loss probability. From the ergodicity, we have
\[ \frac{|\mathcal{N}_n^+|}{n} \to_{\text{as}} \frac{\bar{q}_{k-1}}{1 - P_l} \int_0^\theta c(\theta) \, dy = \frac{\bar{v}(\theta_0)}{\lambda(1 - P_l)}, \tag{A.6} \]
because \( n \to \infty \), where the second equality follows from (3) and (4).

Because (A.6) implies \( |\mathcal{N}_n^+| \to_{\text{as}} \infty \), we have, for any \( \epsilon > 0 \), using (A.4) and (A.5),
\[ P \left( |\hat{\theta}_n - \theta_0| > \epsilon \right) = P \left( \hat{\theta}_n < \theta_0 - \epsilon \right) + P \left( \hat{\theta}_n > \theta_0 + \epsilon \right) \leq P \left( \max_{i \in \mathcal{N}_n^+} W_i \leq \theta_0 - \epsilon \right) + P \left( \max_{i \in \mathcal{N}_n^+} W_i \geq \theta_0 + \epsilon \right) = E \left[ e^{-\lambda |\mathcal{N}_n^+|} \right] \to 0 \]
because \( n \to \infty \). Therefore, we have that \( \hat{\theta}_n \to_{\text{p}} \theta_0 \) because \( n \to \infty \). Convergence in probability implies that there exists a subsequence \( (\hat{\theta}_{n_m})_{m=1}^\infty \) such that \( \hat{\theta}_{n_m} \to_{\text{a.s.}} \theta_0 \) as \( m \to \infty \), and because \( \hat{\theta}_n \) is a monotone, nondecreasing sequence, we conclude (29), that is, the MLE is strongly consistent.
Furthermore, it follows from (A.4) and (A.6) that
\[ P\left(n \left\{ \theta_0 - \max_{i \in \mathcal{N}_n^*} W_i \right\} \geq x \right) = P\left( \max_{i \in \mathcal{N}_n^*} W_i \leq \theta_0 - \frac{x}{n} \right) = \begin{cases} E \left[ e^{-\lambda x \left| \mathcal{N}_n^* \right| / n} \right], & 0 \leq x < n\theta_0, \\ 0, & x \geq n\theta_0, \end{cases} \]
so that for each \( x \geq 0 \), we obtain from the continuous mapping theorem,
\[ \lim_{n \to \infty} P\left(n \left\{ \theta_0 - \max_{i \in \mathcal{N}_n^*} W_i \right\} \geq x \right) = e^{-\varphi_0(0)/(1-P_t)}. \]
\[ \triangleright \text{Upper bound.} \]
Let \( \mathcal{N}_n^* \subset \{1, 2, \ldots, n\} \) denote the set of indices of observed virtual waiting times such that
\[ \mathcal{N}_n^* = \left\{ k \in \{1, 2, \ldots, n\}; W_{k-1} + X_{k-1} > \max_{i \in \{1, 2, \ldots, k-1\}} W_i \right\}. \]
By definition, we have
\[ \hat{\theta}_n = \max_{i \in \mathcal{N}_n^*} W_i. \]
For \( k = 1, 2, \ldots, n \), we define
\[ Z_k := \max\{0, \theta_0 - [A_k - \max\{0, W_{k-1} + X_{k-1} - \theta_0\}]\} \]
\[ = \begin{cases} \max\{0, \theta_0 - A_k\}, & W_{k-1} + X_{k-1} \leq \theta_0, \\ \max\{0, W_{k-1} + X_{k-1} - A_k\}, & W_{k-1} + X_{k-1} > \theta_0. \end{cases} \]
From (13) and (A.2), it follows that \( W_i \leq Z_i \) a.s. for \( i = 0, 1, \ldots, n \). We then have from (A.9),
\[ \hat{\theta}_n \leq \max_{i \in \mathcal{N}_n^*} Z_i \quad \text{a.s.} \]
Furthermore, \( Z_k \) is stochastically identical to \( \max\{0, \theta_0 - E_{i,k}\} \), where \( (E_{i,k})_{k=1,2,\ldots} \) are i.i.d. exponential random variables with parameter \( \lambda \), as defined above. \( \{Z_k\}_{k \in \mathcal{N}_n^*} \) thus forms a sequence of i.i.d. nonnegative random variables with the same cdf as (A.3) so that (cf. (A.4))
\[ P\left( \max_{i \in \mathcal{N}_n^*} Z_i \leq x \right) = \mathbb{E}\left[ e^{-\lambda (0-x) \left| \mathcal{N}_n^* \right| / n} \right], \]
which implies (cf. (A.7))
\[ P\left(n \left\{ \theta_0 - \max_{i \in \mathcal{N}_n^*} Z_i \right\} \geq x \right) = \begin{cases} E \left[ e^{-\lambda x \left| \mathcal{N}_n^* \right| / n} \right], & 0 \leq x < n\theta_0, \\ 0, & x \geq n\theta_0. \end{cases} \]
Let \( \varphi_{0_0}(x) (x \in (0, \theta_0)) \) denote the stationary probability that the virtual waiting time just after an acceptance of a customer takes a value in \( (\theta_0 - x, \theta_0) \):
\[ \varphi_{0_0}(x) = \pi_0 P(\theta_0 - x < X_{i0} < \theta_0) + \int_0^{\theta_0} v(y) P(\theta_0 - y - x < X_{i0} < \theta_0 - y) dy. \]
Because we have \( \int_0^{\theta_0} v(y) dy < \infty \), it follows from the dominated convergence theorem that
\[ \lim_{x \to 0^+} \varphi_{0_0}(x) = 0, \]
so that we have
\[ \forall \delta > 0, \exists \delta_0 > 0: \varphi_{0_0}(\delta_0) < \frac{\delta}{4}. \]
Now define \( \mathcal{N}_n^{**} := \mathcal{N}_n^* \setminus \mathcal{N}_n^* \), which concretely means that
\[ \mathcal{N}_n^{**} = \left\{ k \in \{1, 2, \ldots, n\}; \max_{i=1,2,\ldots,k-1} W_i < W_{k-1} + X_{k-1} < \theta_0 \right\}. \]
Because \( \mathcal{N}_n^* \) and \( \mathcal{N}_n^{**} \) are two disjoint sets whose union is \( \mathcal{N}_n^{**} \), we have that
\[ |\mathcal{N}_n^{**}| = |\mathcal{N}_n^*| + |\mathcal{N}_n^{**}|. \]
Note that (29) implies
\[ \exists n_0 \in \{0, 1, \ldots\}, \ P(\hat{\theta}_n < \theta_0 - d_0) < \delta \frac{n}{4} \quad \text{for } n = n_0, n_0 + 1, \ldots \]  \hfill (A.14)

Because \( \hat{\theta}_n \) is nondecreasing in \( n \) a.s., we have for \( n = n_0, n_0 + 1, \ldots \),
\[ E(\mid N_{n}^{*}\mid) = E(\mid N_{n_0}^{*} \mid + \mid N_{n}^{*}\mid \mid N_{n_0}^{*}) \]
\[ = E(\mid N_{n_0}^{*}\mid) + P(\hat{\theta}_{n_0} < \theta_0 - d_0) \cdot E(\mid N_{n}^{*}\mid \mid \hat{\theta}_{n_0} < \theta_0 - d_0) \]
\[ + P(\hat{\theta}_{n_0} \geq \theta_0 - d_0) \cdot E(\mid N_{n}^{*}\mid \mid \hat{\theta}_{n_0} \geq \theta_0 - d_0) \]
\[ \leq E(\mid N_{n}^{*}\mid) + P(\hat{\theta}_{n_0} < \theta_0 - d_0) \cdot (n - n_0 + 1) \]
\[ + P(\hat{\theta}_{n_0} \geq \theta_0 - d_0) \cdot E(\mid \{k \in \{n_0, n_0 + 1, \ldots, n\}; d_0 < W_{k-1} + X_{k-1} < \theta_0) \mid) \]

It then follows from (A.12) and (A.14) that
\[ E\left( \frac{\mid N_{n}^{*}\mid}{n} \right) \leq \frac{\delta}{4} \cdot \frac{n - n_0 + 1}{n} \]
\[ + \frac{n - n_0 + 1}{n} \cdot E\left( \mid \{k \in \{n_0, n_0 + 1, \ldots, n\}; d_0 < W_{k-1} + X_{k-1} < \theta_0) \mid \right) \]
\[ \rightarrow \frac{\delta}{4} + \phi_{\theta_0}(d_0) \]
because \( n \rightarrow \infty \), which is majorized by \( \delta/2 \). We conclude that we have established that \( E(\mid N_{n}^{*}\mid/n) \leq \delta \) for sufficiently large \( n \). Because \( \delta > 0 \) is arbitrary, we find
\[ \lim_{n \to \infty} E\left( \frac{\mid N_{n}^{*}\mid}{n} \right) = 0. \]  \hfill (A.15)

Using the Markov Inequality, we have for any \( \epsilon > 0 \),
\[ P\left( \frac{\mid N_{n}^{*}\mid}{n} > \epsilon \right) \leq \frac{1}{\epsilon} \cdot E\left( \frac{\mid N_{n}^{*}\mid}{n} \right), \]
so that (A.15) immediately implies that \( \mid N_{n}^{*}\mid/n \rightarrow 0 \) because \( n \rightarrow \infty \). Upon combining the above, we have from (A.6) and (A.13),
\[ \frac{\mid N_{n}^{*}\mid}{n} \rightarrow_{P} \frac{V(\theta_0)}{\lambda(1 - P)} \quad (n \to \infty). \]  \hfill (A.16)

Therefore, we obtain from (A.11),
\[ \lim_{n \to \infty} P\left( n\left( \theta_0 - \max_{i \in N_{n}^{*}} Z_i \right) > x \right) = e^{-\lambda\theta_0 x/(1-P)}. \]  \hfill (A.17)

Finally, noting that (A.5) and (A.10) imply
\[ P\left( n\left( \theta_0 - \max_{i \in N_{n}^{*}} Z_i \right) \geq x \right) \leq P\left( n(\theta_0 - \hat{\theta}_n) \geq x \right) \leq P\left( n\left( \theta_0 - \max_{i \in N_{n}^{*}} W_i \right) \geq x \right) \]
we have from (A.8) and (A.17),
\[ \lim_{n \to \infty} P\left( n(\theta_0 - \hat{\theta}_n) \geq x \right) = e^{-\lambda\theta_0 x/(1-P)}. \]

Also, (31), (A.5), and (A.10) yield the following bounds for the scaled variance of the estimation error:
\[ n^2 \text{Var}(\theta_0 - \hat{\theta}_n) \geq E\left[ \left( n\left( \theta_0 - \max_{i \in N_{n}^{*}} Z_i \right) \right)^2 \right] - \frac{1}{n} \cdot E\left( \left( n\left( \theta_0 - \max_{i \in N_{n}^{*}} W_i \right) \right)^2 \right), \]
\[ n^2 \text{Var}(\theta_0 - \hat{\theta}_n) \leq E\left[ \left( n\left( \theta_0 - \max_{i \in N_{n}^{*}} Z_i \right) \right)^2 \right] - \frac{1}{n} \cdot E\left( \left( n\left( \theta_0 - \max_{i \in N_{n}^{*}} W_i \right) \right)^2 \right). \]
We can (i) explicitly calculate the right-hand sides of these inequalities using (A.7) and (A.11) and (ii) verify (with (A.6) and (A.16)) that their limits (because $n \to \infty$) coincide and are given by $(1 - P_1)/\varphi(0)\theta$; we omit the details of these straightforward calculations.

References


