## Data-Driven Fitting of the M/G/1 Queue

Dieleman, Nanne; Heidergott, Bernd; Peng, Yijie

*published in*
2019 16th International Conference on Service Systems and Service Management (ICSSSM)
2019

*DOI (link to publisher)*
10.1109/ICSSSM.2019.8887609

*document version*
Peer reviewed version

*document license*
Article 25fa Dutch Copyright Act

*citation for published version (APA)*
Dieleman, N., Heidergott, B., & Peng, Y. (2019). Data-Driven Fitting of the M/G/1 Queue. In *2019 16th International Conference on Service Systems and Service Management (ICSSSM): [Proceedings]* Article 8887609 Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/ICSSSM.2019.8887609

# Data-Driven Fitting of the M/G/1 Queue

1st Nanne Dieleman
*Department of Econometrics and Operations Research*
*Vrije Universiteit Amsterdam, The Netherlands*
email: n.a.dieleman@vu.nl

2nd Bernd Heidergott
*Department of Econometrics and Operations Research*
*Vrije Universiteit Amsterdam, The Netherlands*
email: b.f.heidergott@vu.nl

3rd Yijie Peng
*Department of Industrial Engineering and Management*
*Peking University, China*
email: pengyijie@pku.edu.cn

***Abstract*—In this paper we derive a MLE for waiting times in an G/G/1 queue. Numerical examples illustrate the application of our estimator to fitting a M/G/1 model to observed waiting times.**

***Index Terms*—queuing, gradient estimation, MLE**

## I. Introduction

Maximum likelihood estimation (MLE) is a well known technique for fitting probabilistic models to data, see for more details [4]. The basic philosophy of MLE is that one considers a parametric family of densities, denoted by $f_\theta(x)$, where $\theta$ denotes a design parameter. Given observation $X$ of a random phenomenon one finds the value for $\theta$ that maximizes $f_\theta(X)$, i.e., that maximizes the rate at which the actual observation $X$ actually occurs. For MLE one requires an analytically tractable density. This limits the applicability of the MLE technique to service systems, such as, waiting models from queueing theory, as these models typically have no closed-form solution for the underlying density. In [5], [6], Peng at al. derived a simulation-based estimator for estimating densities and their $\theta$-derivatives from observed data given that the density of modeled phenomenon $X$ is continuous and the input random variabels have support on the whole space. Waiting times in service systems have a point mass in zero and the input random variables including interarrival and service times have distributions supported on $[0, \infty)$. In this paper we provide an MLE that addresses the particular needs of service centers. While we state the results for the general setting, application of the MLE focuses on the case where interarrival times are exponentially distributed and services have support $[0, \infty)$. Moreover, we are only interested in the waiting times themselves rather than some waiting time-related cost.

With the approach put forward in this paper, an MLE-based output fitting of a waiting time model is made available. By "output fitting" we mean that a postulated model is directly fitted to output data, i.e., observed waiting times, opposed to "input fitting," where statistics like MLE are used to fit the postulated models for the input data such as interarrival times and service times. A classical ressult on input fitting is [2], where an MLE for the arrival and service rate of an M/M/1 queue provided based on the number of arrivals and service completions, together with the time spent in the empty state during a fixed time interval along with initial queue length. An MLE for the M/M/1 queue fitted to queue length data is given in [8].

The output fitting of a postulated model is motivated by the desire to have a realistic but still easy to handle model at hand for analysis and design studies. Indeed, with such an idealized and well-calibrated model at hand, questions can be answered, such as, what the effect of a faster server is on the waiting time. More importantly, comparing output fitting to input fitting gives the opportunity to analyze model mismatch. To see this, suppose that we have reason to believe that a single server can be described by an M/M/1 queue. The M/M/1 queue is only an idealized version of the real system. Comparing the estimated arrival rate and service rate obtained on a statistical analysis of the interarrival times and service times alone, to the rates obtained via output fitting, leads to understanding the model mismatch. Indeed, if the arrival rate measured via the interarrival times is $\lambda^{\text{in}}$, and the arrival rate estimated by our MLE via the actually observed waiting times $\lambda^{\text{out}}$, then $\lambda^{\text{in}} - \lambda^{\text{out}}$ indicates the model mismatch in modeling waiting times for this particular system by an M/M/1 queue. The rational lies in that if the postulated M/M/1 model were appropriate, then for a sufficiently large observation period of waiting times, $\lambda^{\text{in}}$ and $\lambda^{\text{out}}$ should be asymptotically equal.

The paper is organized as follows. In Section II the basic waiting time Markov chain model is introduced. The overall MLE formula for an observed sequence of waiting times in presented in Section III and numerical results are provided in Section IV. We conclude with pointing out directions of further research.

## II. The Markov Chain Model

Customers arrive at a service station according to a renewal point process. The inter-arrival times $\{I_n : n \in \mathbb{N}\}$ are independent and identically distributed (iid), with density $f_I(x)$ and $0 < E[I_n] < \infty$ and $\mathbb{P}(I_n = 0) = 0$. Customers are served in order of arrival, and consecutive service times are iid random variables $\{S_n(\theta) : n \in \mathbb{N}\}$ with density $f_S(x; \theta)$, for $\theta \in \Theta \subset \mathbb{R}$, and $0 < E[S_n(\theta)] < \infty$ and $\mathbb{P}(S_n(\theta) = 0) = 0$. Interarrival times and service times are assumed to be mutually independent. Consider the process of consecutive waiting times $\{X_n\}$, denoting the time that the corresponding customer has spent in at the service station from arrival to the beginning of service. The service system starts initially empty. Consecutive waiting times $Z_n$ follow the recursive equation [1]:

$$Z_n = \max\{0, Z_{n-1} + S_{n-1}(\theta) - I_n\},\ n \geq 2, \tag{1}$$

and $Z_1 = 0$, see [1].

Denote the transition kernel of the waiting time chain by $P_\theta$, i.e.,

$$P_\theta(B, z) = \mathbb{P}(Z_{n+1}(\theta) \in B | Z_n(\theta) = z)$$

for $x \geq 0$ and $B \subset (0, \infty)$ a (Borel) measurable set, or, more formally

$$P_\theta(B, z) = \int_0^\infty \left( \int_0^{s+z} 1_{\{x+z-a \in B\}} f_I(a) da \right) f_S(s) ds,$$

and otherwise, for $B \in [0, \infty)$ with $0 \in B$,

$$\begin{aligned} P_\theta(B, z) &= \int_0^\infty \left( \int_0^{s+z} 1_{\{z+s-a \in B\}} f_I(a) da \right) f_S(s) ds \\ &\quad + \int_0^\infty \left( \int_{s+z}^\infty f_I(s) ds \right) f_S(a) da. \end{aligned}$$

Inserting $B = (0, y]$ and differentiating with respect to $y$, we obtain as density of the continuous part of the transitin kernel on $(0, \infty)$

$$\begin{aligned} & f_\theta(y; z) \\ &= \mathbf{1}\{y > z\} \int_0^\infty \mathbf{1}\{s \geq y - z\} f_I(s - (y - z)) f_S(s) ds \\ &\qquad + \mathbf{1}\{y \leq z\} \int_0^\infty f_I(s - (y - z)) f_S(s) ds \\ &= \mathbb{E}[f_I(z + S - y) \mathbf{1}\{y - z \leq S\}], \end{aligned} \tag{2}$$

$y > 0$, (in words, an increase of wating time from $z$ to $y$ is only possible by a service time of at least $y - z$), and the point mass in 0 is given by

$$\begin{aligned} p_\theta(0, x) &= \int_0^\infty \left( \int_{s+x}^\infty f_I(s) ds \right) f_S(a) da \\ &= \mathbb{E}[\mathbf{1}\{x + S - I \leq 0\}]. \end{aligned} \tag{3}$$

## III. Deriving an MLE for the Arrival Rate

Taking derivatives with respect to $\theta = \lambda$, the arrival rate, gives,

$$\begin{aligned} \frac{\partial}{\partial \theta} f_\theta(y; z) &= \int_{\max(y-z,0)}^\infty \frac{\partial}{\partial \theta} f_I(z + s - y, 0) f_S(s) ds \\ &= \mathbb{E}\left[ \frac{\partial}{\partial \theta} f_I(z + S - y) \mathbf{1}\{y - z \leq S\} \right], \end{aligned} \tag{4}$$

for $y > 0$. Introducing the score function of the interarrival times

$$\mathrm{SF}_\theta(x) = \frac{\partial}{\partial \theta} \log(f_I(x)),$$

we have for the derivative of the point mass in 0

$$\begin{aligned} \frac{\partial}{\partial \theta} p_\theta(0, z) &= \int_0^\infty \left( \int_{s+z}^\infty \frac{\partial}{\partial \theta} f_I(s) ds \right) f_S(a) da \\ &= \int_0^\infty \left( \int_{s+z}^\infty \frac{\frac{\partial}{\partial \theta} f_I(s)}{f_I(s)} f_I(s) ds \right) f_S(a) da \\ &= \mathbb{E}\left[ \mathrm{SF}_\theta(I) \mathbf{1}\{I \geq S + z\} \right]. \end{aligned} \tag{5}$$

The straightforward log-likelihood for $\theta$ given observation $\{Z_0, Z_1, \ldots, Z_T\}$ is

$$\begin{aligned} L_T(\theta) &= \sum_{t=1}^T \Bigg( \bigg( \log(p_\theta(0, Z_{t-1}) \bigg) 1_{Z_t = 0} \\ &\qquad + \bigg( \log(f_\theta(Z_t; Z_{t-1}) \bigg) 1_{Z_t > 0} \Bigg). \end{aligned}$$

For the optimization part, we use the estimator

$$L_T(\theta) = \sum_{t=1}^T \left( \frac{\frac{\partial}{\partial \theta} p_\theta(0, Z_{t-1})}{p_\theta(0, Z_{t-1})} 1_{Z_t = 0} + \frac{\frac{\partial}{\partial \theta} f_\theta(Z_t; Z_{t-1})}{f_\theta(Z_t; Z_{t-1})} 1_{Z_t > 0} \right). \tag{6}$$

Inserting (2), (4), (3) and (5) into (6) yields

$$\begin{aligned} & L_T(\theta) = \\ & \sum_{t=1}^T \left( \frac{\sum_{m=1}^M \mathrm{SF}_\theta(I_t^m) \mathbf{1}\{Z_{t-1} + S_t^m \leq I_t^m\}}{\sum_{m=1}^M \mathbf{1}\{Z_{t-1} + S_t^m \leq I_t^m\}} 1_{Z_t = 0} \right. \\ & \left. + \frac{\sum_{m=1}^M \frac{\partial}{\partial \theta} f_I(Z_{t-1} + S_t^m - Z_t) \mathbf{1}\{Z_t - Z_{t-1} \leq S_t^m\}}{\sum_{m=1}^M f_I(Z_{t-1} + S_t^m - Z_t) \mathbf{1}\{Z_t - Z_{t-1} \leq S_t^m\}} 1_{Z_t > 0} \right), \end{aligned} \tag{7}$$

where $\{I_t^m, S_t^m\}$ is a collection of iid interarrival and service times. Note that in case of high utilization, so that the probability of $Z_t = 0$ is small, the above MLE reduces to the standard MLE for fitting the density to the observed interarrival times. For finding the value for $\theta$ that maximizes $L_T(\theta)$ we apply standard stochastic approximation, see [3].

For the M/G/1 queue with arrival rate $\theta$, we obtain

$$\frac{\partial}{\partial \theta} f_I(x) = (1 - \theta x) e^{-\theta x} \text{ and } \mathrm{SF}_\theta(I_t) = \frac{1}{\theta} - I_t.$$

The MLE estimator with respect to the arrival rate is displayed in the boxed equation (8).

$$\boxed{\begin{aligned} L_T(\theta) = \sum_{t=1}^T & \left( \frac{\sum_{m=1}^M \left(\frac{1}{\theta} - I_t^m\right) \mathbf{1}\{Z_{t-1} + S_m^t \leq I_m^t\}}{\sum_{m=1}^M \mathbf{1}\{Z_{t-1} + S_m^t \leq I_m^t\}} 1_{Z_t = 0} \right. \\ & \left. + \frac{\sum_{m=1}^M (1 - \theta(Z_{t-1} + S_m^t - Z_t)) e^{-\theta(Z_{t-1} + S_m^t - Z_t)} \mathbf{1}\{Z_t - Z_{t-1} \leq S_m^t\}}{\sum_{m=1}^M \theta e^{-\theta(Z_{t-1} + S_m^t - Z_t)} \mathbf{1}\{Z_t - Z_{t-1} \leq S_m^t\}} 1_{Z_t > 0} \right). \end{aligned}} \tag{8}$$

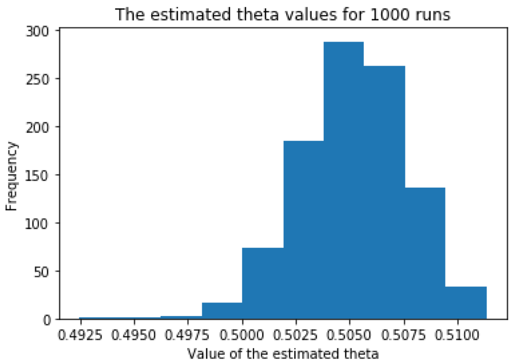

Fig. 1. Histogram of the 1000 estimated theta values for a dataset of 100 customers.

For finding the value for $\theta$ that maximizes $L_T(\theta)$ we apply standard stochastic approximation [3].

## IV. Numerical Results

For illustration purposes, we consider a single server queue with Poisson arrival process with rate $\theta$ and lognormal service times, and provide MLE's for fitting $\theta$ to observed waiting times. In particular, we choose the postulated model the true model (i.e., the model we actually sample the waiting times from) but we start the postulated model with a wrong $\theta$. So, MLE is carried out to trace the correct arrival rate for this model. Numerical examples will illustrate the performance of our estimators.

For the second set of numerical examples, we consider the following variation of the above model. The number of servers now varies between 1 and 2. We apply our MLE to fit a postulated single server model, where the MLE compensates for the intentional model mismatch. Numerical examples will illustrate the performance of our estimators.

### A. MLE for true Model

We consider a single server queue with Poisson arrival process with rate $\theta_0$ and lognormal service times. We simulate a sequence $(I_t^m, S_t^m, Z_t^m)$ of interarrival, service and waiting times satisfying the relation $Z_t^m = \max(Z_{t-1}^m + S_{t-1}^m - I_t^m, 0)$. For the numerical experiments, we set $\theta$ to $\theta_1 = 1.0 \neq \theta_0 = 0.4$ and apply MLE to trace the true value $\theta_0$.

We apply SA with a fixed epsilon of 5e-05. We start by considering $T$ small ($T = 100$), which is a realistic number of customers if the estimator is applied in practice. The service time is chosen in such a way that the load is roughly equal to $0.65$. We perform 1000 independent runs for the SA and plot in Figure 1 the corresponding histogram for the found optimal values for $\theta$. Table I presents the mean and standard deviation of the estimations. Assuming normality, the corresponding $95\%$ confidence interval is presented in Table I. The true value of $\theta$ does not lie in this confidence interval. The reason for this is that $T$ is rather small, and the MLE depends still too much on the observed waiting times.

| Measure | Value 100 customers | Value 1000 customers |
|---|---|---|
| Mean | 0.505 | 0.393 |
| Standard deviation | 0.00248 | 0.00194 |
| Lower bound CI | 0.500 | 0.389 |
| Upper bound CI | 0.510 | 0.396 |

TABLE I
Mean, standard deviation and the confidence interval for the dataset with 100 and 1000 customers.

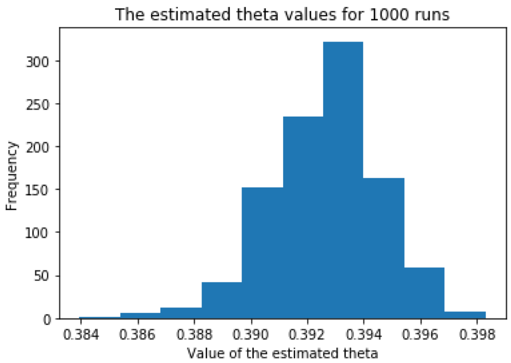

Fig. 2. Histogram of 1000 estimated $\theta$ values for a dataset of 1000 customers.

Therefore, we also consider the case for larger $T$ ($T = 1000$). Figure 2 presents the corresponding histogram. Table I reports the values for this case.

Table II presents the expected waiting time for the true$\theta$ and the two mean values of the estimated $\theta$-values. The expected waiting time of the case with $T = 1000$ lies close to the true expected waiting time, which is a consequence of the ergodicity of the system and the strong consistency of the MLE.

### B. MLE with Model Mismatch

We consider a single server queue with Poisson arrival process with rate $\theta_0$ and lognormal service times. However, for this system, we vary the number of servers between 1 and 2. More specifically, we run a discrete-eeevent (DES) simulation with two servers. Server 1 is always active. For server 2 the dynamic is a follows: whenever a customer leaves server 1 and there are at least two customers waiting in queue, server 2 becomes active with probability $p > 0$ and servers one customer. Once this service is finished, the server becomes inactive until started again via a customer leaving server 1. We simulate a sequence $(I_t^m, S_t^m, Z_t^m)$ of interarrival, service and waiting times but due to the model mismatch, the recursion $Z_t^m = \max(Z_{t-1}^m + S_{t-1}^m - I_t^m, 0)$ is not satisfied. The postulated waiting time is $g(X_t^m, Z_{t-1}) = \max(Z_{t-1}^m + S_{t-1}^m - I_t^m, 0)$, where $X_t^m = (S_{t-1}^m, I_t^m)$. Note

| | Expected waiting time |
|---|---|
| $\theta_0$ | 4.579 |
| $\hat{\theta}_{T=100}$ | 11.884 |
| $\hat{\theta}_{T=1000}$ | 4.337 |

TABLE II
The expected waiting time for the true $\theta$ and the two estimated $\theta$-values.

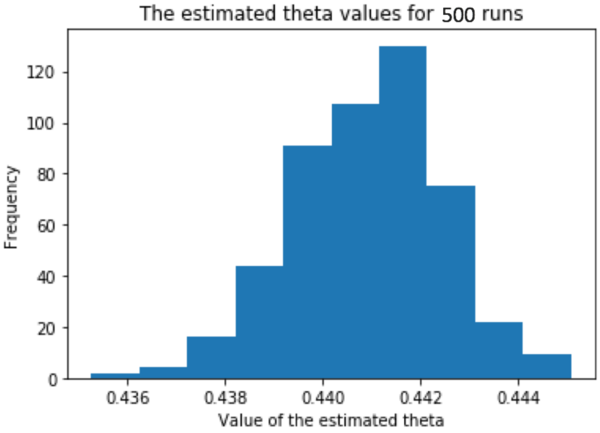

Fig. 3. Histogram of the 500 estimated $\theta$ values for a model mismatch dataset of 100 customers.

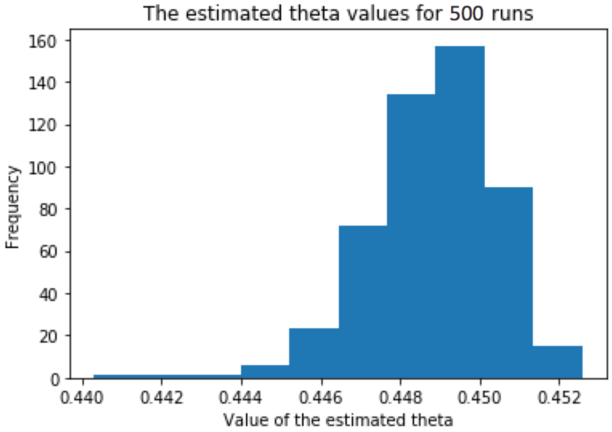

Fig. 4. Histogram of the 500 estimated $\theta$ values for a model mismatch dataset of 1000 customers.

that $p$ allows us to control the amount of model miss-match present in the model We now apply (8) to sensitivity of the arrival rate.

We take the true $\theta$ to be equal to $0.5$ and define the service parameters such that the load is roughly equal to$0.8$. A dataset is generated and the estimator is run for different values of the probability $p$. First of all, 500 independent SA runs are performed for the case that $p = 0.5$ and $T = 100$ as well as $T = 1000$. Figure 3 and 4 are the corresponding histograms. TableIII presents the absolute difference between the mean estimate and the true value of $\theta$ in the two cases.

Table IV presents the expected waiting time for the true $\theta$ and the mean of the two estimated $\theta$-values. As with the true model, the estimation of the larger dataset is closer to the true value.

| T-value | $\lvert\hat{\theta} - \theta_0\rvert$ |
|---|---|
| T = 100 | 0.0591 |
| T = 1000 | 0.0511 |

TABLE III
ABSOLUTE VALUE FOR THE DIFFERENCE BETWEEN THE MEAN ESTIMATE AND THE TRUE VALUE OF $\theta$.

| | Expected waiting time |
|---|---|
| $\theta_0$ | 3.883 |
| $\hat{\theta}_{T=100}$ | 2.931 |
| $\hat{\theta}_{T=1000}$ | 3.135 |

TABLE IV
THE EXPECTED WAITING TIME FOR THE TRUE OF $\theta$ AND THE TWO ESTIMATED VALUES.

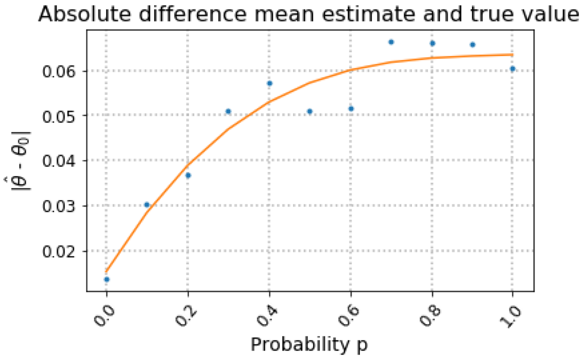

Fig. 5. The absolute difference between the true value of $\theta$ and the mean estimate for increasing values of $p$ together with a fitted third-degree polynomial.

Figure 5 shows the absolute difference between the true value of $\theta$ and the mean estimate for increasing values of $p$. When $p$ increases, server 2 is opened more often. This means that the model mismatch enlarges if $p$ increases. As expected, the absolute difference increases if the mismatch enlarges; see Figure 5. For example, if a deviation of at most 10 % of $\theta$ is deemed acceptable, than the M/G/1 model is only acceptable up to values of $p$ no larger than $0.38$.

## CONCLUSION

We established a MLE for fitting a M/G/1 model to waiting times. The estimator requires off-line simulation and the overall solution approach solves the maximum likelihood problem by applying stochastic approximation. Numerical examples illustrated the effect the size of the data set made available to the MLE has on the output fit. We discussed the application of our MLE approach to identifying model mismatch. Further research will be on improving the numerical performance of the estimator and on extending our results to Markovian queueing networks.

## REFERENCES

[1] S. Asmussen: *Applied Probability and Queues*, Springer, 2003.

[2] A. Clarke: Maximum likelihood estimates in a simple queue, *The Annals of Mathematical Statistics*, 28(4), p. 1036-1040, 1957.

[3] H. Kushner, G. Yin: *Stochastic Approximation and Recursive Algorithms and Applications,* Springer, 2003.

[4] R. Millar: *Maximum Likelihood Estimation and Inference*, Wiley, 2011.

[5] Y. Peng, M. Fu, B. Heidergott, H. Lam: Maximum Likelihood Estimation By Monte Carlo Simulation: Towards Data-Driven Stochastic Modeling, *Operations Research*, submitted in revision, 2019.

[6] Y. Peng, M. Fu, J. Q. Hu, B. Heidergott: A New Unbiased Stochastic Derivative Estimator for Discontinuous Sample Performances with Structural Parameters, *Operations Research*, 66, p. 487–499 2018. A new unbiased stochastic derivative estimator for discontinuous sample performances with structural parameters, Y. Peng, M. Fu, J. Hu, and B. Heidergott, *Operations Research*,

[7] R. Rubinstein, A. Shapiro: *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*, Wiley, New York, 1993.

[8] W. Wang, G. Casale, A. Kattepur, M. Nambiar: *Maximum likelihood estimation of closed queueing network demands from queue length data*, In Proceedings of the 7th ACM/SPEC on International Conference on Performance Engineering, p. 3-14, 2016.