On Estimation Problems for the G/G/∞ Queue

Hanhua Feng, Parijat Dube, Li Zhang
IBM T. J. Watson Research Center
Yorktown Heights, NY, 10598, USA.
{hanhfeng, pdube, zhangli}@us.ibm.com

ABSTRACT
We consider estimation problems in G/G/∞ queue under incomplete information. Specifically, we are interested in scenarios where it is infeasible to track each individual job in the system and only aggregate statistics are known or observable. We first show that the minimum expected square estimator for the queue length process can be written as the sum of the minimum square estimators for an indicator function associated with each job, which is simply the survival function of the service time variable for each job. We also obtain tight lower and upper bounds on the time average of the square estimation error. Next we look at the inverse problem of estimating the service time distribution when the observed process is only the queue length process. We develop an on-line stochastic-optimization based estimation algorithm for the service time distribution and study its convergence under different parameter settings.

Keywords
G/G/∞ queue, estimation, mean square estimation error, stochastic optimization.

1. INTRODUCTION
The count process N(t) (number of busy servers or queue length process) of any infinite server queue can be constructed by just knowing the arrival process {A_i} and the service time process {X_i} of each job. The arrival process on the other hand can just be constructed by knowing N(t), as the jump epochs in N(t) correspond to arrival times of jobs. We study non-parametric estimation of statistics of different processes in a G/G/∞ queue when only partial information is available. While we make no assumptions on the arrival process, the service times are assumed to be i.i.d. Specifically, we pose two estimation questions for G/G/∞ type queue:
• Given the arrival process and the service time distribution of jobs, what is the best estimator for N(t)?
• Given only N(t), what is the best estimator for service time distribution?

Large population systems are characteristic of many real-world systems at different temporal and spatial dimensions. An important concern with such large-scale systems is that often it is infeasible to track each individual either due to scalability of the monitoring system or due to other concerns like privacy etc. However aggregate statistics are easier to obtain like the count process N(t) which at any time is known by just knowing the total arrivals and total departures without tracking arrival and departure time for each individual. Earlier works adopt this position, e.g., for road traffic [2] and for industrial production [6]. For a single or multiple servers with Poisson arrivals, Larson [5] considered the problem of inferring the mean waiting time and mean queue length from recorded service starting and ending times of all jobs. Pickands and Stine [7] considered the problem of estimating arrival rate and service time distribution of a discrete time M/G/∞ queue using only the count process information. Bingham and Pitts [1] applied the theory developed in Grubel and Pitts [4] for estimating the service time distribution in M/G/∞ queue under three different settings for the observed process. Heavily exploiting the property of a Poisson arrival process, these existing works are not directly applicable to the G/G/∞ scenario.

2. ESTIMATOR FOR COUNT PROCESS
We first consider the problem of finding the best estimator for the count process given the arrival process and the service time distribution. Consider a G/G/∞ queue where the service time X is i.i.d. with cumulative distribution function F(x). Consider a single job arriving at time 0. Let I_X(t) be the indicator random variable which is 1 for 0 ≤ X < t and 0 otherwise. Basically \{I_X(t)\} is a stochastic process taking binary values, 1 when the job is in the queue and 0 when the job is not in the queue. Let us use a deterministic real-valued function φ(t) as an estimator for \{I_X(t)\}. The mean of \{I_X(t)\} is given by:
\[ \mathbb{E}[I_X(t)] = \begin{cases} \Pr[X > t], & (t \geq 0) \\ 0, & (t < 0) \end{cases} =: \bar{F}(t), \]
where \bar{F}(t) is the survival function of X, defined differently than the Complementary Cumulative Distribution Function (CCDF) of X for t < 0. Thus knowing the survival function one can use \bar{F}(t) as an estimator for I_X(t), \text{∀} t. It can be easily shown that \( \hat{\phi} = \bar{F} \) minimizes the expected square estimation error for I_X(t) among class of all real-valued functions. Further when \( \hat{\phi} \equiv \bar{F} \), the expected square estimation error \( \mathbb{E}[(I_X(t) - \bar{F}(t))^2] \) equals \( \bar{F}(t) - \bar{F}(t)^2 \) which is same as the variance of I_X(t).

2.1 A Deterministic Function as Estimator for Count Process
Let \( a_i \) be the arrival time of job \( i \), and \( X_i \) be its service time. Then the number of jobs in the system can be written
\[ N(t) := \sum_{i \in \mathcal{I}} I_X_i(t - a_i), \]
where \( \mathcal{I} \) is the set of job indices, which can be either the set of integers or its subset. Since \( X_i \) are i.i.d., we use a deterministic \( \hat{\phi}() \) to estimate \( I_X_i(\cdot) \) for all \( i \in \mathcal{I} \), we have
\[ \hat{N}(t) := \sum_{i \in \mathcal{I}} \hat{\phi}(t - a_i). \]
Let \( \mathcal{E}(t) \) denote the expected square estimation error, i.e.,
\[ \mathcal{E}(t) = \mathbb{E}\left[ (N(t) - \hat{N}(t))^2 \right]. \]

Lemma 1. For any deterministic function \( \hat{\phi}(t) \), \( \mathcal{E}(t) \) is minimized when \( \hat{\phi} \equiv \bar{F} \) and the minimum \( \mathcal{E}(t) \) is simply the...
sum of the variances of individual $I_{X_i}(t-a_i), i \in I$ i.e.,

$$E_{\text{var}}(t) := \sum_{i \in I} \left( \tilde{F}(t-a_i) - \tilde{F}^2(t-a_i) \right).$$

(1)

Proof. From the definition of $N(t), \tilde{N}(t)$ and $E(t)$ we have:

$$E(t) = E \left[ \left( \sum_{i \in I} (I_{X_i}(t-a_i) - \phi(t-a_i)) \right)^2 \right]$$

$$= \sum_{i \in I} E \left[ (I_{X_i}(t-a_i) - \phi(t-a_i))^2 \right]$$

$$+ \sum_{i \neq j \in I} \sum_{j \neq i \in I} E \left[ (I_{X_i}(t-a_i) - \phi(t-a_i))(I_{X_j}(t-a_j) - \phi(t-a_j)) \right].$$

As $I_{X_i}(t-a_i)$ is independent of $I_{X_j}(t-a_j)$ for $i \neq j$, we get

$$E(t) = \sum_{i \in I} E \left[ (I_{X_i}(t-a_i) - \phi(t-a_i))^2 \right]$$

$$+ \sum_{i \neq j \in I} \sum_{j \neq i \in I} E \left[ (I_{X_i}(t-a_i) - \phi(t-a_i))(I_{X_j}(t-a_j) - \phi(t-a_j)) \right].$$

(2)

Clearly, this expected square error is minimized when $\tilde{F} = \phi$, in which case we get (1). □

## 2.2 Time average of the square error

The time average of the square error is defined to be

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T \left( N(t) - \tilde{N}(t) \right)^2 dt = E \left[ \lim_{T \to \infty} \frac{1}{T} \int_0^T \left( N(t) - \tilde{N}(t) \right)^2 dt \right]$$

for a stationary queue. Let us assume that $0 \leq a_i \leq T$. We assume that the number of arrivals in any finite interval is finite and the service time distribution has finite first moment, i.e., $\int_0^\infty \tilde{F}(t)dt < \infty$. Then,

$$\frac{1}{T} \int_0^T \left( N(t) - \tilde{N}(t) \right)^2 dt = \frac{1}{T} \int_0^T E(t) dt$$

$$= \frac{1}{T} \sum_{i \in I} \int_0^T (\tilde{F}(t-a_i) - \tilde{F}^2(t-a_i)) dt$$

$$= \frac{1}{T} \sum_{i \in I} \int_0^{T-a_i} (\tilde{F}(t) - \tilde{F}^2(t)) dt,$$

where $|I|$ is the number of arrivals within $[0,T]$. As $T \to \infty$, it approaches to

$$\lim_{T \to \infty} \frac{|I|}{T} \int_0^\infty (\tilde{F}(t) - \tilde{F}^2(t)) dt = \lambda \int_0^\infty (\tilde{F}(t) - \tilde{F}^2(t)) dt$$

Let the long-term time average of number of arrivals be $\lambda$ and define $\rho = \lambda \tilde{E}[X]$. Since $\int_0^\infty \tilde{F}(t)dt = \tilde{E}[X]$ and $0 \leq \tilde{F}^2(t) \leq \tilde{F}(t)$: we have

$$0 \leq \lim_{T \to \infty} \frac{1}{T} \int_0^T \left( N(t) - \tilde{N}(t) \right)^2 dt$$

$$= \lambda \int_0^\infty \left( \tilde{F}(t) - \tilde{F}^2(t) \right) dt \leq \lambda \tilde{E}[X] = \rho$$

Both bounds are tight; when $X$ is of deterministic distribution, the error is zero. We can use a $D_2$ distribution [3] to show that the upper bound is also tight. In addition, we can construct two series $D_2$ distributions of any given mean and variance, such that one series approaches to the lower bound and the other approaches to the upper bound. It is easy to deduce that if $X$ is exponentially distributed, the average square error is $\rho^2/2$. In other words, we can see that the average error is tightly bounded above by $\sqrt{\rho}$ and below by 0, and exponential service time gives $\sqrt{\rho}/2$.

## 3. ESTIMATION OF THE SERVICE TIME DISTRIBUTION

Observe that the R.H.S in (2) is a quadratic functional of $\phi$ minimized only at $\phi = \tilde{F}$. Discretizing the arrival and service times, and assuming $\phi$ has a finite support, we can represent $\phi$ using a high-dimensional vector. Therefore, we can use convex optimization techniques to find the minimal solution of $\phi$ using $E$ as the objective function, if we can compute $E$ for each given $\phi$ easily. The process $N$ is fixed if the arrival process is given; however, $N$ is a stochastic process depending on both the arrival process and the service times. Let $\phi_k = \phi(kT)$ be the vector entry at discrete time $k$, where $T_k$ is the length of a unit sample interval.

### 3.1 Off-Line Estimation Algorithm

We can design a simple gradient descent method algorithm for offline estimation.

1. Choose a step size $\alpha$ and a positive number $n$.
2. Choose an initial $\phi^{(0)} = \phi_0$.
3. At iterate $j$, first generate a fixed finite arrival sequence $\{a_1, a_2, \ldots, a_n\}$ and use this same arrival sequence to repeatedly run the queue with random service times. $\tilde{E}[N(t)]$ is then approximated by averaging observed $N(t)$ in these repititions. Then the elements of the gradient $\nabla \tilde{E}[N(t)]$ for each discrete time $t$ is given by

$$\frac{\partial \tilde{E}[N(t)]}{\partial \phi_k} = 2 \left( \tilde{N}(t) - \tilde{E}[N(t)] \right) \frac{\partial \tilde{N}(t)}{\partial \phi_k}.$$

Note that $\tilde{N}(t) = \sum_i \phi_i(t-a_i)$. Therefore, $\partial \tilde{N}(t)/\partial \phi_k = 1$ only if $kT = (t-a_i)$ for some $i \in I$, or zero otherwise.

4. Update $\phi^{(j+1)} = \phi^{(j)} - \alpha \nabla \tilde{E}[N(t)]$ and go back to step 3.

The problem of the off-line problem is that we cannot restart the queue and provide a fixed arrival process in a real system. In fact, usually we have no control over the arrival process. What we get is only a single sample path of the arrival process and the service times. Hence in the next section we provide an on-line stochastic optimization algorithm to estimate $\tilde{F}$.

### 3.2 On-Line Estimation Algorithm

The on-line algorithm is based on stochastic gradient descent algorithm. For on-line estimation we need to assume that the arrival process is a stationary process.

We assign a fixed $\phi$ to each job. As a new job comes in, it gets the most updated $\phi$. For every $m$ arrivals, we compute the gradient using the formula as in the off-line algorithm as follows:

1. First choose an initial $\phi = \phi_0$.
2. For each discrete time $t$:
   
   (a) Observe $N(t)$.
   
   (b) Compute $\tilde{N}(t) = \sum_i \phi_i(t-a_i)$. Since we assume $\phi$ has a finite support, say $[0, s]$, then we need only to add all jobs with $a_i + s \geq t$.
   
   (c) Compute $\Delta N(t) = \tilde{N}(t) - N(t)$. Set $\partial \tilde{E}[N(t)]/\partial \phi_k = 2\Delta N(t)$, $\forall k$, if $kT = (t-a_i)$ for some $i \in I$.
   
   3. $t \leftarrow t + 1$.
   
   4. If $(jm+1)$-th arrival comes, $\forall j$, update $\phi: \phi \leftarrow \phi - \alpha \nabla \tilde{E}$. 


5. (Optional) Constrain $\phi$ so that it is a decreasing function with value between 0 and 1.
6. Output an average value of $\phi$ as an approximation of $\hat{F}$ (we can use the average of entire history of $\phi$, or using other averaging method, e.g., exponential moving average). (Project $\phi$ into the space of decreasing functions between 0 and 1 before output it if Step 5 is not done.)
7. Go to step 2.

In step 5, optionally we can limit $\phi$ to be a valid CCDF function for $t \geq 0$ by projecting $\phi$ to the space of $\hat{F}$ in every iteration. However, even without the projection, $\phi$ approaches to $\hat{F}$ in the general functional space; therefore, we can do the projection before estimating an approximation of $\hat{F}$ in Step 7. We will show that the algorithm gives a better estimation of $\hat{F}$ if we don’t impose the constraints in Step 5 but instead use it only for output.

### 3.2.1 Simulation Study of On-Line Algorithm

![Figure 1: Error with different step sizes for exponential service time distribution case.](image1)

![Figure 2: Estimates of $\phi$ at different iterations converge to the (unknown) exponential service time distribution.](image2)

![Figure 3: Error with different step sizes for uniform service time distribution case.](image3)

![Figure 4: Estimates of $\phi$ at different iterations converge to the (unknown) uniform service time distribution.](image4)

Figures 1 and 2 show the simulation results for exponential service times with mean 10. We set $T_\alpha = 0.01$, $m = 2$ and $s = 100$. The arrival process is a renewal process with i.i.d. inter-arrival time uniformly distributed between 0 and 2 (with mean of 1). Therefore, the expected load will be 10. Figure 1 shows the convergence of square root of time average of $\mathcal{E}$ with different step sizes ($\alpha = 0.001$, 0.01, 0.02). For $\alpha = 0.001$ we show both, optimization with constraints ($0 \leq \phi \leq 1$ and $\phi_k \geq \phi_{k+1}$; i.e., on-line algorithm with Step 5), and without constraints. For $\alpha = 0.01$ and $\alpha = 0.02$, we show only unconstrained case. We can see that with smaller steps, the errors converge to a closer neighborhood of the optimal expected error due to our stochastic setting. Figure 2 shows the estimated $\phi$ from the optimization algorithm, averaged over time, at $t = a_{16261}$, $t = a_{16340}$, and $t = a_{1650}$, for $\alpha = 0.01$ and with unconstrained case. We can see a gradual convergence of the average $\phi$ towards the expected uniform distribution. At $t = a_{1650}$ (not drawn in the figure), the average $\phi$ is almost indistinguishable from the CCDF of the exponential distribution.

### 4. REFERENCES