Performance Evaluation 79 (2014) 182–197

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Estimating life-time distribution by observing population continuously[☆]

Hanhua Feng[a,*], Parijat Dube[b], Li Zhang[b]

[a] Infinio Systems, Inc., United States

[b] IBM T. J. Watson Research Center, United States

ARTICLE INFO

Article history:
Available online 10 July 2014

Keywords:
Lifetime distribution
Non-parametric estimation
$G/GI/\infty$
Streaming algorithm

ABSTRACT

Often in a real world system with a fairly large population, members are not individually traceable for various reasons. As a result, the relationship between a member's behavior and the system's behavior is quite hard to understand. In this paper, we focus on a fundamental problem in such a system: the relationship between its population size and the life-time distribution for its members. We answer two questions:

(A) If the life-time distribution is known and the times when members join are observable, how do we best estimate the population size?
(B) If the population size can be observed accurately, how do we estimate the unknown life-time distribution for members?

In the paper we focus on (B), using the results of (A) as a basis.

We model the system as a $G/GI/\infty$ queue with incomplete information, where jobs, once entering the queue, are no longer tracked. With this model, the population size is the number of jobs in the queue and the life times of members are the service times of the jobs. The problem (A) is to estimate the number of jobs in the queue, with known arrival times and a known service-time distribution. We show that, in terms of mean square error, the best deterministic estimator for the (stochastic) number of jobs in the system can be constructed using the survival function of the service-time distribution. The problem (B) is to estimate the unknown service-time distribution in a $G/GI/\infty$ queue where the number of jobs are observable. We demonstrate that the service-time distribution can be inferred indirectly from continuous observations of the number of jobs in the queue, and then propose a few easy-to-implement algorithms. Using only a limited amount of memory, these on-line, streaming algorithms continuously refine their results which, over time, converge to the true service-time distribution.

1. Introduction

Large populations are characteristic of many real-world systems at different temporal and spatial dimensions. Examples include viewer population in a big stadium during a major sporting event, customer population in a big mall or at a retail shop, viewer population during prime-time slots on television, and user population at a popular web-site. Such large

☆ This work is based on an extended abstract by the same authors, in *SIGMETRICS Performance Evaluation Review* 39(3): 40–42, 2011.

* Corresponding author.
E-mail addresses: hanhua@cs.columbia.edu (H. Feng), pdube@us.ibm.com (P. Dube), zhangli@us.ibm.com (L. Zhang).

http://dx.doi.org/10.1016/j.peva.2014.07.012
0166-5316/

population systems are characterized by dynamically changing population size due to frequent incoming and outgoing customers. Knowledge of system parameters like sojourn time of customers (jobs), population size, and waiting time in such systems, is important for planning operations and designing business strategies and policies targeting the system population. For example, duration of advertisements displayed on billboards in stadiums, placement of consumer goods in shops, embedding of commercial breaks on television can all be tied to the sojourn time (life time) of customers in the system. The knowledge of the number of customers is a measure of the system popularity like "hotness" of web-sites and target rating points of TV shows and is useful for revenue management during advertisement slot bidding.

Estimation of such system parameters usually requires monitoring the behavior of individual customers. An important problem with such large-scale systems is that often it is infeasible to track each individual either due to cost and scalability constraints associated with the monitoring system or due to other concerns like anonymity and privacy, etc. However aggregate statistics are much easier to obtain; for example, the total number of customers at any time is known by just counting the total arrivals and departures. Without the need for tracking each individual, estimation techniques based on easy-to-observe aggregate measures are therefore much desired for large population systems. For example, such techniques can be used for improving revenue by devising advertising strategies which exploits the population size in any system with large population with continuous arrivals and departures like stadiums, road-traffic, and mobile zones. Earlier works adopt this position, e.g., for road traffic [1] and for industrial production [2].

In this paper, we model such a large population system as a $G/GI/\infty$ queue, and study non-parametric estimation of different statistics of the queue when only partial information is available. While we make no assumptions on the arrival process, the service times are assumed to be independent and identically distributed (IID) random variables. The stochastic process $N(t)$ (number of busy servers or queue length) 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. Note that, since we are considering an infinite-server system, the sojourn time of a job in the system is same as its service time.

Specifically, we pose two estimation questions for $G/GI/\infty$ type queue:

- Given the arrival process and the service time distribution of jobs, what is the best deterministic estimator for $N(t)$?
- Given only $N(t)$, what is the best estimator for service time distribution?

Apart from answering the theoretical questions on the efficiency of estimators for $N(t)$ and service-time distribution, an important contribution of our work is the development of simple, practical algorithms for the real-time estimation of the service-time distribution in $G/GI/\infty$. These on-line, streaming algorithms operate on bounded state space with each new iteration requiring a fixed amount of computational time.

The paper is organized as follows. Section 2 deals with the problem of estimating the number of jobs in a $G/GI/\infty$ system. In Section 3, the inverse problem of estimating the service-time distribution in a $G/GI/\infty$ system is studied and efficient on-line streaming algorithms are developed and analyzed. Numerical results on the algorithms are presented in Section 4. Section 5 reviews the literature on parametric and non-parametric estimation in queues. Finally, we conclude in Section 6.

2. Estimating the number of jobs in a $G/GI/\infty$

In this section, we consider the problem of estimating the number of jobs in a $G/GI/\infty$ system, where exact arrival times are known. The service times are independent, identically distributed (IID) random variables denoted by X whose Cumulative Distribution Function (CDF) is denoted by $F(x) := \Pr[X \le x]$. We assume that this distribution is known *a priori*, although the exact service times of individual jobs are not.

2.1. Preliminaries

Let $I_X(t)$, a stochastic process solely determined by the non-negative random variable X, be the indicator function which equals 1 for $0 \le t < X$ and 0 otherwise. Then, for a job i with service time X_i arriving at time τ_i, the binary-valued stochastic process $I_{X_i}(t-\tau_i)$ indicates if the job X_i is in the system at time t. The expectation of $I_X(t)$ is given by:

$$\mathbb{E}[I_X(t)] = \left\{\begin{matrix} \Pr[X > t] = 1 - F(t), & (t \ge 0) \\ 0, & (t < 0) \end{matrix}\right\} =: \bar{F}(t),$$

where $\bar{F}(t)$ is called the *survival function* of X. Note that the survival function $\bar{F}(t)$ defined here is slightly different from usual: for $t < 0$, we set $\bar{F}(t)$ to be 0, instead of $1 - F(t)$ which is equal to 1 for a negative t.

The stochastic process $I_X(t)$ can be estimated by a deterministic real-valued function $\phi(t)$, and it is easy to show that, with $\phi = \bar{F}$ the expected square estimation error is minimized (among the class of all real-valued functions) for every t:

$$\min_{\phi} \mathbb{E}\left[(I_X(t) - \phi(t))^2\right] = \mathbb{E}\left[\left(I_X(t) - \bar{F}(t)\right)^2\right] \equiv \sigma^2[I_X(t)] = \bar{F}(t) - \bar{F}^2(t), \tag{1}$$

where $\sigma^2[Y]$ denotes the variance of a random variable Y.