Heavy Traffic In Queues

William Thomas

October 31, 2018

Abstract

A GI/GI/1 queue describes a queue with a single server, under the assumption that the inter-arrival times of customers can be modelled by some arbitrary distribution, the service times can be modelled by some other arbitrary distribution and both the inter-arrival times and the service times are independent. Within this report, we examine the ideas proposed by Lindley (1952), who explored the stationary distribution of such a queue, and Kingman (1961), who explored how GI/GI/1 queues behave under heavy traffic. We then proceed to discuss the role of diffusion approximations in queuing theory and consider Heyman's diffusion approximation to the length of a queue under heavy traffic (Heyman, 1975).

1 Introduction

Queuing theory has long been of interest, with applications in telecommunications, traffic management, scheduling and industrial engineering, to name but a few. However, while some queuing models based on exponential inter-arrival times and exponential serving times, such as the M/M/1 or M/M/s queue, are more open to analysis and the derivation of theoretical results, queuing models based on arbitrary distributions are more complicated to analyse. Lindley (1952) presented a way in which these queues could be analysed, relating both their evolution and equilibrium distribution to random walks, thus making this area of queuing theory more accessible for further research. Kingman (1961) continued to build on this area, examining how queues behave when we have *heavy traffic*, that is, when customers arrive almost as quickly as they are served. This opened up a whole new area of asymptotic results for modelling queues.

Within this report, we aim to outline the general framework of a GI/GI/1 queue, before moving on to describe Lindley's results relating to the GI/GI/1 queue. These include a recurrence relation for customer waiting times, a relationship between queues and random walks and conditions for the existence of an equilibrium distribution for the waiting times. We will then move on to consider queues under heavy traffic, examining Kingman's approximation to the waiting time distribution and Heyman's diffusion approximation for the length of a queue (Heyman, 1975).

2 An Introduction to GI/GI/1 Queues

A typical queue consists of two main components: a set of customers, who arrive and join a queue according to some distribution, and a set of servers, who attend to customers in the queue for a period of time according to some other distribution. In particular, a GI/GI/1 queue is a queuing model based on the following setup:

- The intervals between customer arrivals are independent and follow some arbitrary distribution.
- The service times for each customer are independent and follow some (different) arbitrary distribution.
- A single server attends to customers according to the order in which they arrive.

Through the use of arbitrary distributions, we have the ability to model a wide variety of queues, but as a result, it becomes more difficult to analyse and derive results based on such queues. However, we will see later that under certain conditions, the choice of these arbitrary distributions becomes less important.

More formally, let t_r denote the time between the r^{th} and $(r+1)^{st}$ arrival and assume that the t_r are independent random variables with identical distributions and finite mean. If a customer arrives when the server is free, they are attended to immediately for some service time s_r. As before, the s_r are independent random variables with identical distributions and finite mean. Otherwise, the server is engaged and the customer must wait in a queue for some time w_r before being served. This is referred to as the *waiting time* for the r^{th} customer. We assume for now that $w_1 = 0$ and also that the sequences $\{t_r\}$ and $\{s_r\}$, $r = 1, 2, ...$ are independent of one another. With this framework for GI/GI/1 queues, we can now present the results put forward by Lindley (1952).

3 Lindley's Equation and Random Walk Duality

Let $u_r = s_r - t_r$ denote the difference between the service time for the r^{th} customer and the time between the r^{th} and the $(r+1)^{st}$ arrival. From the setup of a GI/GI/1 queue, Lindley (1952) recognised the following relationship between the waiting times, in what has come to be known as *Lindley's equation*:

$$w_{r+1} = \begin{cases} w_r + u_r & \text{if } w_r + u_r > 0 \\ 0 & \text{if } w_r + u_r \leq 0. \end{cases} \tag{1}$$

Intuitively, this equation is derived from the possible relationships between the waiting times, arrival times and service times. If $t_r < w_r + s_r$, the time it takes for the $(r+1)^{st}$ customer to arrive is shorter than the time it takes for the r^{th} customer to wait and be served. We then have the relationship $w_{r+1} = w_r + s_r - t_r$, which corresponds to the first case in Lindley's equation. This is because the r^{th} customer is still waiting or still being served, so the $(r+1)^{st}$ must wait w_{r+1} before being served. The second case arises when $t_r \geq w_r + s_r$, which means that the $(r+1)^{st}$ customer can be served immediately and hence $w_{r+1} = 0$.

With Lindley's equation, we are able to completely describe the evolution of the waiting times for a GI/GI/1 queue. Equivalently, Lindley's equation can be used to describe a one dimensional random walk with an impenetrable barrier at 0. This is a random walk that cannot become negative and is instead held at 0 until it becomes positive again. In this situation, we make jumps of amount u_r at discrete times and w_{r+1} is the distance from the origin after the r^{th} jump. An example random walk which evolves according to Lindley's equation is shown in Figure 1, which also demonstrates the difference between a normal random walk and a random walk with an impenetrable barrier.

4 Queues in Equilibrium

We are often interested in what happens to a queue as the number of customers approaches infinity. This behaviour is governed by the equilibrium distribution of the waiting times, which Lindley (1952) shows must satisfy *Lindley's integral equation*:

$$F(x) = \int_{u \leq x} F(x-u) dG(u) = \int_{y \geq 0} F(y) dG(x-y), \tag{2}$$

where $F(.)$ denotes the stationary cumulative distribution function of the waiting times and $G(.)$ denotes the cumulative distribution function of any u_r. While it is possible to solve this equation, typically via the Wiener-Hopf method (Smithies, 1940), we can learn about the equilibrium distribution by simply making use of the duality between queues and random walks.

Let S_k denote the position of a random walk at a time k, such that $\{S_k\}$ forms the random walk corresponding to our GI/GI/1 queue. Asmussen (2003) demonstrates that the waiting time for the n^{th} customer, W_n, satisfies

$$W_n = \max\{W_0 + S_n, S_n - S_1, ..., S_n - S_{n-1}, 0\}. \tag{3}$$

If we then define $M_n = \max_{0 \leq k \leq n} S_k$ and $M = \max_{0 \leq k < \infty} S_k$, we are able to obtain a relationship between the distribution of the waiting times and the maximum of the corresponding random walk:

$$W_n \stackrel{d}{=} \max\{W_0 + S_n, M_{n-1}\}. \tag{4}$$

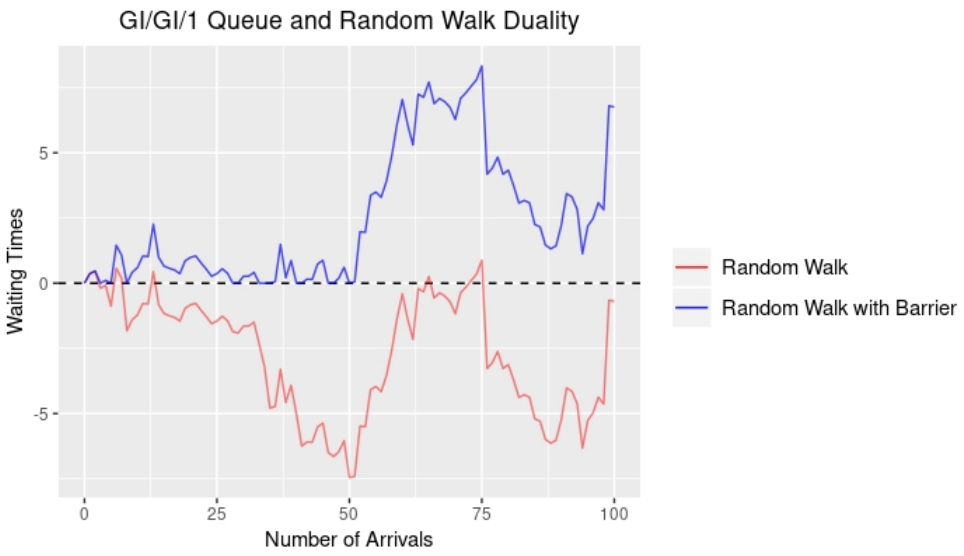

Figure 1: The red line represents a normal random walk with no impenetrable barrier, while the blue line represents a random walk with an impenetrable barrier. Both walks are simulated using i.i.d. exponential inter-arrival times and i.i.d. exponential service times. The barrier is given by the dashed black line.

That is, the distribution of the n^{th} waiting time in the GI/GI/1 queue is identically equal to the distribution of the maximum of the corresponding random walk at time n. But how can we use this information to obtain the equilibrium distribution?

A stationary distribution for the waiting times exists only under certain conditions which depend on the variables u_r. Note that because of our assumptions on t_r and s_r, the u_r are also independent and identically distributed random variables with finite mean. Lindley (1952) showed that a necessary and sufficient condition for the existence of an equilibrium distribution is that $\mathbb{E}[u_r] < 0$ or $u_r = 0$. In all other cases, the cumulative distribution function can be shown to go to 0 for all inputs. When the condition $\mathbb{E}[u_r] < 0$ is satisfied, it can then be shown that $W_n \xrightarrow{d} M$, as well as in total variation. With these results, we have found the equilibrium distribution of the waiting times, which is the unique distribution satisfying Lindley's integral equation (Asmussen, 2003).

5 Queues in Heavy Traffic

The *traffic intensity* for a queue is given by the ratio between the inter-arrival rates and the service rates, $\rho = \mathbb{E}[s_r]/\mathbb{E}[t_r]$. Using Kingman's definition, a queue is said to be under heavy traffic when the traffic intensity approaches 1 from below (Kingman, 1962). In this situation, we reach achieve equilibrium much slower than we would otherwise (Lindley, 1952).

Kingman (1962) demonstrates that under heavy traffic, the waiting times are in fact exponentially distributed with mean

$$\mathbb{E}[w] \approx \frac{Var(u_r)}{-2\mathbb{E}[u_r]} = \frac{\sigma_{t_r}^2 + \sigma_{s_r}^2}{2(\mathbb{E}[t_r] - \mathbb{E}[s_r])}, \tag{5}$$

where $\sigma_{t_r}^2$ and $\sigma_{s_r}^2$ denote the variances of the inter-arrival times and the service times respectively. In fact, Kingman (1962) shows that when the $\{u_r\}$ are independent, this is an upper bound for the mean of the waiting times. Under this approximation, the stationary cumulative distribution function of the waiting times is given by

$$F(w) \approx 1 - \exp\left\{\frac{2(\mathbb{E}[t_r] - \mathbb{E}[s_r])w}{\sigma_{t_r}^2 + \sigma_{s_r}^2}\right\}. \tag{6}$$

What this tells us is that under heavy traffic, the distributions of the inter-arrival times and the service times no longer play a significant role in determining the waiting time distribution.

We are also able to approximate the length of the queue under heavy traffic, that is, the number of customers in the queue at any one time. One such approximation is given by Heyman (1975), who proposes a diffusion model approximation. The intuition behind such an approach is that under heavy traffic flow, queues can behave more like a fluid, with the evolution of the queue resembling a continuous flow. Using a diffusion, we can both approximate the length of a queue and its waiting time distribution (Sani and Daman, 2014). It is easy to see this in the latter case, where we can consider a reflected Brownian motion with an impenetrable barrier at 0. Brownian motion can be viewed as the limit of a random walk as the times between jumps becomes infinitesimally small, so in combination with the duality between queues and random walks, we can begin to see some justification for the use of a diffusion approximation under heavy traffic.

If instead, we wish to approximate the queue size at time t, $N(t)$, by a diffusion process $\tilde{N}(t)$, Heyman (1975) suggests taking

$$\mathbb{E}[\tilde{N}(t)] = \lambda - \mu \tag{7}$$

$$Var(\tilde{N}(t)) = \lambda^3 \sigma_{t_r}^2 + \mu^3 \sigma_{s_r}^2. \tag{8}$$

where $\lambda = \mathbb{E}[t_r]^{-1}$ and $\mu = \mathbb{E}[s_r]^{-1}$. To see why this approximation is particularly appropriate, we must first consider the limiting behaviour of a diffusion process. Heyman (1975) shows that given a diffusion process with infinitesimal mean a and infinitesimal variance b, as $t \to \infty$ the cumulative distribution function of the process is exponential with mean $b/2a$. It then suffices to apply Little's Law (Little, 1961) to the limiting mean of the diffusion process. If we take

- L = average number of customers in the queue,
- λ = average arrival rate,
- W = average waiting time in the queue,

then Little's Law states that $L = \lambda W$. Indeed, in the limiting case as $t \to \infty$, an application of Little's Law to the mean of the diffusion process yields Kingsman's heavy traffic approximation. Clearly this lends some support to the use of a diffusion process approximation, but equally there are many other possible approximations that we could consider.

6 Discussion

Within this report, we have introduced several notable results in the analysis of GI/GI/1 queues. Lindley's equation allows us to completely describe how the queue waiting times evolve over time, while the notion of random walk duality means that we can describe every queue in the form of a random walk. This opens up queuing theory to existing analysis on random walks and enabled us to describe some distributional relationships between the queue waiting times and the maximum of a random walk. Under certain conditions, we can then describe the unique equilibrium distribution of the waiting times.

When we move to consider a heavy traffic scenario, we then become interested in approximating the equilibrium waiting time distribution. Using Kingsman's approximation, we are able to relate the waiting times to an exponential distribution, independently of the arrival and service time distributions. Under heavy traffic, we can also begin to view the evolution of the queue as a continuous flow which opens up the possibility of a diffusion approximation to the queue length.

The study and approximation of queues under heavy traffic is now a prominent research area and has moved towards ever more complicated queues, with multiple servers, several arrival channels or even broadening the notion of heavy traffic to include the case where $\rho \to 1$ from above (Iglehart and Whitt, 1970). However, while Lindley can be said to have developed some of the early results for GI/GI/1 queues and introduced the relationships between queues and random walks, it was the work of Kingsman that really began to push queuing theory towards more asymptotic results and approximations.

References

S. Asmussen. *Applied probability and queues (Second Edition)*. Springer, 2003.

D. P. Heyman. A diffusion model approximation for the GI/G/1 queue in heavy traffic. *The Bell System Technical Journal*, 54, 1975.

D. Iglehart and W. Whitt. Multiple channel queues in heavy traffic I. *Advances in Applied Probability*, 2: 150–177, 1970.

J. F. C. Kingman. The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society*, 57(4):902–904, 1961. doi: 10.1017/S0305004100036094.

J. F. C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2):383–392, 1962.

D. V. Lindley. The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(2):277–289, 1952. doi: 10.1017/S0305004100027638.

John D. C. Little. A proof for the queuing formula: $L = \lambda W$. *Oper. Res.*, 9(3):383–387, June 1961. ISSN 0030-364X. doi: 10.1287/opre.9.3.383. URL `http://dx.doi.org/10.1287/opre.9.3.383`.

S. Sani and O. A. Daman. Mathematical modelling in heavy traffic queuing systems. *American Journal of Operations Research*, 4:340–350, 2014.

F. Smithies. Singular integral equations. *Proceedings of the London Mathematical Society*, s2-46:409–466, 1940.