CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Heavy Traffic Approximations for Models Using the Brownian Motion

AUTHOR: Brianne Hamilton

SUPERVISOR: Professor Yiqiang Zhao

DATE: August 24, 2020

Heavy Traffic Approximations for Models using the Brownian Motion

Illustrated by the $M/M/1$ Queue

by

Brianne Hamilton

An honours project submitted to the School of Mathematics and Statistics in partial fulfilment of the requirements for the degree of

Bachelor's of Mathematics Honours

in

Statistics with a Concentration in Actuarial Science

Carleton University
August 25, 2020

Contents

1 Introduction

In this project we will be looking at heavy traffic approximations of queueing systems and see that this approximation allows us to observe that the number of customers in a queue can be modelled by a Brownian motion. The type of Brownian motion will depend on the process. In this project we will be focused on one dimensional queue's only with both normal service times where we will see that the process can be represented by a reflected Brownian motion. Then, later on in the project, we will see a special case of a one dimensional queue where the service times are longer for the first customer to an empty queue than they are for any customers arriving to an already busy queue. This process will be represented by a sticky Brownian motion. The properties of the heavy traffic approximations that give both the reflected and sticky Brownian motions will be illustrated using the $M/M/1$ queue. Using this illustration we will see the applications of the concepts to an actual queueing network. The $M/M/1$ queue is one of the simplest models and thus will allow us to see in detail the properties we discussed. We will see how these approximations work when attempting to find the number of customers in the queue at a given point in time.

To accomplish this we will first introduce the concepts of the $M/M/1$ queue to give a background to the type of queue that will be used to illustrate all of the properties. We will then do a review of the Brownian motion including the reflected and sticky Brownian motions. We will discuss why they are useful to model the number of customers in the queue and how to get from the random walk process that is the number of customers coming and going from a queue to the continuous process that is a Brownian motion. In the section following the review of the Brownian motion we will see the heavy traffic property and how that is obtained. This property allows us to approximate the number of customers in the queue when it is busy, or experiencing heavy traffic. Finally in the following two sections after we will illustrate the properties discussed throughout the previous sections using the $M/M/1$ queue and we will find the distributions for the number of customers in each of the cases. An in depth understanding of the heavy traffic property, reflected, and sticky Brownian motion is our goal for this project so in addition to going over these concepts, an illustration of them will further solidify our knowledge.

The random walk, that will be used when discussing the Brownian motion and its properties, was first introduced by Karl Pearson in 1905, many of the papers looked at for this project referenced his work on the Brownian motion. It can be shown that the number of customers in a queue at any time is a random walk as it goes up or down one with probability depending on the distribution of the arrivals and service times. The Markov chain is a type of random walk where the next state depends only on the current state, this is also true for the one-dimensional queuing models we'll be discussing as the number of customers that will be in the queue next depends only on the number that are currently in the queue and the probability of an arrival or service happening first. The Brownian

motion was then introduced as a limit of random walks. This concept was first introduced by Norbert Wiener, and as such the process is also sometimes referred to as a Wiener process. We will be using his work in this project as the foundation for the reflected and sticky Brownian motion. We will also be using multiple other sources for the Brownian motion information such as Dai[2], Harrison [7], and Chen [22] as well as others that are referenced throughout this project. These papers built on the concepts of a Brownian motion and used different queue models to represent the information. Currently in the works of a Brownian motion many multi-dimensional models are being studied by researchers such as Dai[1] to find their tail asymptotics. The topics covered in this project should be enough to give a solid understanding of the base concepts such that understanding more complicated papers like the one mentioned is attainable.

The reflected Brownian motion was first used to describe queueing systems that experiences heavy traffic that was established by Kingman in 1962 [13] and was studied further by Inglehart and Whitt in 1970 [10]. In this project we use references from both of their works to show that in queueing theory the reflected Brownian motion can be used to represent the number of customers in the queue. We will show this by imposing the heavy traffic conditions on our queue and we will see that this gives the same distribution as the reflected Brownian motion when we have regular service times. After showing the general idea and proof behind the reflected Brownian motion we will use the $M/M/1$ queue to show how the reflected Brownian motion will work in practice. The reflected Brownian motion is the obvious choice for modelling customers in a queue as the number can never go below 0, thus it is often used by current researchers when studying queueing theory and we will reference some of them throughout the project.

The sticky Brownian motion for a single server queue was studied by Finch in 1959 where the customer to an empty queue would have to wait a random amount of time before beginning their service. Then later, in 1962, Yeo [18] looked at the queue in the case that the server went away and came back for some time as well as the case of if the first customer had a longer but known distribution service time. In this project we focus on Yeo's work when the first customer to an empty queue has a longer service time than a customer arriving to an already busy system. Other papers look at to get a solid understanding of the sticky Brownian motion include works done by Whitt[18], and Harrison [7]. The sticky Brownian motion is also currently being used in research to again, mostly study multidimensional processes and their behaviours around the origin (i.e at 0).

The heavy traffic property was first introduced by Kingman in 1961 with his paper "The single server Queue in Heavy traffic" [14] and he furthered his research in 1962 with the paper "On Queue's in Heavy Traffic"[13] . Since then his work has been used as the foundation for modelling many different queue's in heavy traffic. In this project we use the concepts he introduced in our definition of the heavy traffic property and the illustration

through an $M/M/1$ queue. We also use the work by Hafflin and Whitt [6] who did more research on the types of systems that can apply the heavy traffic property to find the distributions of both the number of customers in the queue and the wait times. We also used a lot of the information in a presentation done by Shaikhet [16] where the heavy traffic approximation was made easily understandable by using a fluid diffusion approximation. A lot of research currently being done in both diffusion and heavy traffic approximations builds upon the work done by Kingman and we will be using his work as well as that of the researchers after him to attempt to understand the heavy traffic principles and how they can be applied to the queue's we wish to study in this project.

As we've mentioned briefly in the previous paragraphs, a lot of work is being done for both reflected and sticky Brownian motion models in multi dimensions. The heavy traffic property is used to model these multi dimensional cases as well which leads to a diffusion process. There are many papers that have calculated the parameters for multidimensional queues that either operate normally or stick when the system empties. They build on some of the basic principles that are discussed in this project. One such paper is "Multidimensional Sticky Brownian Motion: Approximation, dependance and Tail Asymptotics" by Dai and Zhao [1] which looks at the multidimensional sticky Brownian motion as a semimartingale reflecting Brownian motion in the orthant. In the paper they calculate the stationary distributions of the sticky Brownian motion as a diffusion approximation. In this project we will be more focused on the basics of these topics and compile them in a way such that most people would be able to also understand these concepts and then using this would be able to comprehend the more complex current research that is being done. Over the course of this project we will accomplish this through the definitions of each topic and then later with the illustration by the $M/M/1$ queue.

We start in the following section with an introduction to the $M/M/1$ queue which we will be using to illustrate the properties of the reflected and sticky Brownian motions later on. We then go into an explanation of the Brownian motion and the heavy traffic property which will be used in this illustration. By the end of this project we will have a thorough understanding of these concepts that we would need in order to comprehend the current research being done in the field.

2 The $M/M/1$ Queue

In this project we will be using the $M/M/1$ queue to illustrate the heavy traffic properties and how they can give us the reflected and sticky Brownian motion and we will see their applications to finding the number of customers in a queue. We will be illustrating all of these concepts using the $M/M/1$ queue. In doing this we show that it can be done for a

queue model and it allows us to see the properties of the reflected and sticky Brownian motion as well as heavy traffic property in a more concrete way. This will solidify our knowledge of the concepts and since the $MM/1$ queue is a quite basic model it will be simpler to show the properties and we can focus more on the Brownian motion rather than the actual queue. To do this we must first introduce the $M/M/1$ queue definition and some background information on why it is a useful model to study. This will all be accomplished throughout the section and we will be mostly using the notes from the course STAT 4508 taught by Professor Yiqiang Zhao at Carleton University for the review of the $M/M/1$ queue as well as other papers that go into a bit more detail that are references throughout.

The $M/M/1$ queue is a single server queue with poison arrivals at rate λ and exponential service time with rate μ, thus the expected service times will be $\mathbb{E}(s) = 1/\mu$ where s is the service time. Customers are served in the order of their arrival, the queue waiting length is infinite and customers depart immediately after service. The number of customers at any point in time will form a discrete Markov Chain, going up with rate $\lambda/(\lambda + \mu)$ (i.e. an arrival happens before the next service time is completed), and down with rate $\mu/(\lambda+\mu)$ (i.e. a service completion before the next arrival).These limiting probabilities were all shown in detail during the coursework mentioned earlier and are relatively simple, thus the exact proofs are left out for this project and we can take these limiting probabilities as known.

Recall the definition of a Markov chain being a stochastic model that describes a sequence of possible different outcomes. The probability of a certain outcome depends only on the current state of the process. This means that the probability of going up or down to any number of people in the queue depends only on the number of people currently in the queue. Given the queue starts in state i the number of customers can go to $i + 1$ if an arrival occurs before the completion of service, or to $i - 1$ if a customer is served and leaves before the next arrival. All other probabilities are 0 as the queue can not jump up or down by more than one person at a time. The Markov chain has stationary and independent increments which is what gives this memoryless property where the next state depends only on the current one. This is a necessary distinction and will be used later on when we model the random walk by a Brownian motion. The strong Markov property will also be introduced later on to show a queue can be modelled by a sticky Brownian motion.

Now we can see that the $M/M/1$ queue is a stochastic process where $Q(t)$ is the number of customers in the system and has values$\{0, 1, 2, 3, ...\}$ and thus the $M/M/1$ queue has state spaces $\{0, 1, 2, 3, ...\}$, this number includes the person currently in service. The arrivals come at rate λ and when an arrival happens if the process started in state $i = 0, 1, 2, ...$ before the arrival than after the process will move from state i to $i+1$ after the i^{th} arrival. Similarly the rate of service is μ and thus the mean service rate is $1/\mu$, when a customer is served and leaves the number of customers in the system goes down to $i - 1$ after the i^{th} is served, given no new customers arrived during the service time.This is how we get the

probabilities of the process going up or down a state mentioned earlier, $\lambda/\lambda+\mu$ and $\mu/\lambda+\mu$.

We can now introduce the transition rate matrix, also called the infinitesimal generator or the "Q-matrix" by:

$$Q = \begin{pmatrix} -\lambda & \lambda & & & & \\ \mu & -(\mu-\lambda) & \lambda & & & \\ & \mu & -(\mu-\lambda) & \lambda & & \\ & & \mu & -(\mu-\lambda) & \lambda & \\ & & & \ddots & \ddots & \ddots \end{pmatrix} \tag{1}$$

Where all the entries that aren't specified in the matrix are 0.

This matrix describes the instantaneous rates of change between the different states (i.e going from i to $i+1$ or down to $i-1$). These are the transition rates of the Markov Chain and are used in calculating the probabilities of going up or down in customers depending on the values of λ and μ. Note that we also have the stabilizing condition that $\lambda < \mu$, otherwise the number of customers would increase to infinity as the number arriving on average would be greater than the number being served. This matrix is what gives us the Kolmogorov backwards and forwards equations that denote the probability of the process being in a state later on. The Kolmogorov backwards equation, for $\{Q(t) : t \geq 0\}$ a continuous time Markov Chain with rates $q_{i,j}$ and transition probability function $P_{i,j}(t)$ in a state space S and for any $s, t \geq 0$ will be as follows:

$$P_{i,j}(t) = \sum_{k \in S} P_{i,k}(t) P_{k,j}(s)$$

the proof was again gone over in detail during the course taught by Professor Zhao and thus is left out of this project. These equations are useful when looking at the process as a discrete time Markov chain and will be used further in our illustration of the $M/M/1$ queue as a sticky Brownian motion.

The limiting probabilities of this Markov chain have been calculated by Harrison [9]. The probability that the process is in state i at any given time including the customer in service is defined by the variable π_i where π_i was found by Harrison to be:

$$\pi_i = (1-\rho)^i \rho$$

where ρ is still the traffic intensity, $\rho = \lambda/\mu$.

The average number of customers in the system can be found as it is clear from the limiting probability π_i that the number of customers is distributed geometrically with parameter $1-\rho$. Thus by the properties of the geometric distribution, which are recalled

in the paragraph below, the average number in the system is $\rho/(1-\rho)$ and the variance is $\rho/(1-\rho)^2$ (Guillemin [5]).

Recall the geometric distribution is the probability distribution of bernoulli trials needed to obtain a succeed. The expectation and variance of this distribution can be calculated easily if the parameters of the distribution are known as they are in our case.

This concludes out review of the $M/M/1$ queue and now we can move on to our studies of the Brownian motion. We will come back to this queue process later when we use it to illustrate the heavy traffic property and how to get the reflected and sticky Brownian motions to model the number of customers in a queue. We will do this for the $M/M/1$ queue.

3 Brownian Motion Review

The Brownian motion (also sometimes called a Weiner process) is a continuous time stochastic process. The Brownian motion is a sped up random walk with stationary and independent increments. This random walk is sped up in order to make the process a continuous time process. The speeding up is done by shortening the time intervals and increasing the number of steps within that time interval as much as necessary until we have a continuous process. We will see in more detail how this is done in the next section when we will be discussing the Brownian motion and it's properties.

The Brownian motion plays a crucial role in stochastic calculus, diffusion processes and more. In this project we are focused on its applications to queuing theory and how we can use the diffusion or heavy traffic approximations of queues to obtain a Brownian motion for the queue length process. Thus we need to first introduce the Brownian motion which gives us the general idea behind speeding up the random walk process until it becomes continuous. We will then see a reflected Brownian motion that occurs when we have a process that will not go below 0, and it's properties. The sticky Brownian motion is then introduced and will be used in later sections to model the process where the first customer to an idle system has a longer wait time than a customer arriving to an already busy system.

This section will cover all of these topics regarding the Brownian motion. By the end of it we should have a thorough enough knowledge of how the Brownian motion, reflected Brownian motion, and sticky Brownian motion work to show that the heavy traffic approximations will cause queue lengths to converge to one of them. Then later we will apply these concepts. with our illustration of the $M/M/1$ queue.

3.1 Standard Brownian Motion

We will start by constructing a more formal definition of a Brownian motion from a random walk. Then explain what the different parts of the definition mean and why they are important for our studies in this project. This includes the extension of the Brownian motion to the reflected and the sticky Brownian motion. Throughout this subsection we will show how to speed up the random walk Markov chain so that it can be modelled by a Brownian motion. To accomplish this we will once again be using a lot of the material covered by Professor Yiqiang Zhao in the course STAT 4508 at Carleton University. We will also be using other papers as references that further the understanding of speeding up the random walk that are referenced throughout.

We start with a process $X(t)$ that is a random walk Markov chain with parameters:

$$\mathbb{E}[X_i] = 0 \text{ and } \text{Var}[X_i] = \sigma^2$$

The process is sped up by shortening the time intervals (i.e: $t_n - t_{n-1}$ shortened where we know t_n is the $n-$th point in time) and increasing the number of steps within that interval by $\Delta x = \sigma\sqrt{\Delta t}$. This increase in the number of steps is chosen such that the difference in the number of steps (Δx) is also proportional to the difference in the time intervals Δt and the standard deviation of the process σ. Choosing it this way allows us to have a minimum increase in the number of steps that is still high enough such that the random walk behaves like a continuous process and thus from this we have the continuous process that is the Brownian motion

Then we obtain the new parameters for the sped up random walk will be :

$$\mathbb{E}[X_i] = 0 \quad Var[X_i] = (\Delta x)^2 \left\lfloor \frac{t}{\Delta t} \right\rfloor$$

Note that the floor function used in $\left\lfloor \frac{t}{\Delta t} \right\rfloor$ is the function used to describe the closest integer value of $t/\Delta t$ without surpassing the value of $t/\Delta t$.

The proof of how these parameters are obtained was done in detail over the course and thus was left out for this project. The expectation or the mean of the process does not change when it is sped up and will thus, without loss of generality, we will assume that it will remain as 0. The variance now depends on how much the process is sped up by decreasing time intervals Δt and increasing the number of steps Δx. Obtaining this value of the variance is another reason why the increase in steps $\Delta x = \sigma\sqrt{\Delta t}$ was chosen and will give us that the variance of the new sped up process is: $\text{Var}X(t) = \sigma^2 t$ where $X(t)$ is the Brownian motion.

Thus we have now seen that the speeding up of the random walk by shortening the time interval and increasing the number of steps within those intervals is what allows us to view the process as continuous. Since it is known that each step is independent and identically distributed (comes form the fact it is a random walk) the speeding up of the process will not eliminate those traits but will give us a more continuous looking process which can then be considered a continuous time stochastic process, which is necessary for the Brownian motion. We will now give the formal definition and afterwards explain in detail each part and how it is relevant to queuing models:

Definition : A sequence $\{X(t) : t \geq 0\}$ is a Brownian Motion if it satisfies the following properties:

1. The starting point, $X(0) = 0$
2. The process $\{X(t) : t \geq 0\}$ is continuous and has stationary and independent increments
3. The distribution $X(t) \sim N(0, \sigma^2\ t\ )$

The initial condition that $X(0) = 0$ is the start of our brownian motion, in relation to queuing this is the very beginning of the process at time 0 and thus it makes sense that there would be no one in the system at this time. The continuity of the Brownian motion comes from how we shortened the time interval and increased the number of steps within that time interval as we've discussed earlier.

Independent increments of the brownian motion means that the process will assign a random variable X_t to each point in time t for the process. Thus the increments will be the difference between each point in time $X_s - X_t$ for time points s and t where $s > t$. To call these increments independent implies that for time points $s, t, u, v : s > t$ and $u > v$ where $s \neq t \neq u \neq v$ the increments $X_s - X_t$ and $X_u - X_v$ are independent random variables as long as the time intervals do not overlap with each other.

Stationary increments of the brownian motion means that the probability distribution of the interval $X_s - X_t$ where s and t are defined in the same way depends only on the length of that interval $s - t$ and not on what happened previously in the process. If we have $s - t = u - v$ for u and v also defined in the same way then these increments will have the exact same distribution.

The properties of independent and stationary increments of the Brownian motion are important when discussing its uses for modelling the number of customers in different queue's and makes sense intuitively as the probability of another customer coming or going from the system does not depend on how many are already in the system. This is what leads

to the memoryless property discussed in our review of the $M/M/1$ queue which comes from the fact that $X(t)$ is a Markov process. Thus we have that the stationary and independent increments of the Brownian motion which are inherited from the Markov process are maintained as is necessary when discussing the process of the number of customers in the queue.

The central limit theorem as well as the choices of Δx and Δt so that the variance of the sped up process is the $\text{Var}[X_i] = \sigma^2\ t$ given before the definition is what gives us the distribution of the Brownian motion

$$X(t) \sim N(0, \sigma^2\ t)$$

This determines how the Brownian motion will behave and tells the probability of it going up or down about the mean of 0. This allows us to model the Brownian motion graphically so we can see the trends in the number of customers in the system at any given time.

The standard Brownian motion occurs when the mean is 0 and the variance is 1. We can convert any Brownian motion to the standard one by simply dividing the process by the variance. The standard brownian motion $B(t)$ and its relation to the Brownian motion $X(t)$ is as follows:

$$B(t) = \frac{X(t)}{\sigma_x}$$

and we have that $B(t) \sim N(0, 1)$ is the standard form Brownian motion. This is useful to note as many of the calculation tables come for the $N(0, 1)$ distribution and thus in order to calculate the sample paths of the Brownian motion is is easier to do it using the standard one.

3.2 Reflected Brownian Motion

The reflected Brownian motion behaves like a standard BM outside the boundary on $(0, \infty)$, however at the boundary point 0 the process reflects, as opposed to a Brownian motion that can go below 0. This process is used to model the number of customers, denoted by $Q(t)$, at any given time t in queuing processes. Once the system is empty and there are no customers the process will stay at 0 until a new customer arrives. Once a new customer arrives they will begin service immediately upon arrival and then the process will reflect up and away from 0, after which it will behave like a normal Brownian motion until it returns to 0 again. The process will never dip below 0 which makes sense intuitively as we can never have a negative amount of people in the queue.

We will be using the one-dimensional reflected Brownian motion in this project however it is worth noting that the same concepts can be used to extend it to a $d-$dimensional process and we would have a multidimensional reflected Brownian motion. This. can be used

to model a queue in which there are multiple servers or other types of multi-dimensional processes. The concept remains the same where it will never go below 0 and will reflect instead. These types of processes are currently being researched extensively to find the distributions of multi-dimensional queue's and also in finding diffusion approximations and that is why it is worth noting here even though we'll only be looking at the one-dimensional case.

We will start with a formal definition as given by Harrison [8] of the reflected Brownian motion in one-dimension. By the end of this subsection we should have enough knowledge about the reflected Brownian motion in order to use it in our heavy traffic approximations when we show that in heavy traffic, a reflected Brownian motion is what the one-dimensional queue length process will converge to. We will also need to have enough knowledge on the reflected Brownian motion to use it in our illustration by the $M/M/1$ queue.

The definition of the reflected Brownian motion is now given and the details of the definition will be explained in more detail afterwards.

Definition : Reflected Brownian Motion The process $\tilde{Z}(t)$ is a one-dimensional reflecting brownian motion if it can be written as

$$\tilde{Z}(t) = X(t) + M(t)$$

Where $X(t)$ is the Brownian motion that has previously been defined in section 3.1 and we have that $M(t)$ is defined to be:

$$M(t) = \sup_{s \in [0,t]} X(s)$$

One can show that the derivative of $\tilde{Z}(t)$ will be equal to that of $X(t)$ away from 0 and thus they behave the same way away from the boundary, ie when the process is not at 0 it will behave like the Brownian motion discussed earlier The addition of $M(t)$ is what causes the Brownian motion to reflect. $M(t)$ only changes when the Brownian motion is at 0 and this it will not have any affect on the Brownian motion away from the boundary Yeo[26].

In our discussion of the reflected Brownian motion and our illustration of it using the $M/M/1$ queue later on we will be using this definition of the reflected Brownian motion, however we can also note that there is an alternate definition of the boundary behaviour using the generator of the process instead. The generator of the process is an operator on a space of functions defined by the infinitesimal generator of the process. This will be discussed in more detail in the sticky Brownian motion section. For now we can note that

where the generator $(\mathcal{L}f)(x) = \frac{1}{2}f''(x)$ for bounded and smooth functions f we have that

$$f(X_t) - f(X_0) - \int_0^t \frac{1}{2}f''(X_s)ds(*)$$

Where $(*)$ is an expression with respect to the reflected Brownian motion. This will be a sequence of random variables for which, at some time, the conditional expectation of the next value will be equal to the present value (i.e its a martingale) Yeo[26]. If we then focus on what happens at 0 and find $f'(0)$ we can see that the function $(*)$ gives:

$$f(X_t) - f(X_0) - \int_0^t \frac{1}{2}f''(X_s)ds - f'(0)M(t)$$

is also a martingale where $M(t)$ is the local time process defined earlier Yeo [26]. Thus we have that we can also write it in terms of the function on the domain of the infinitesimal generator. This will be important in our extension from a reflected to a sticky Brownian motion and will be discussed in section 3.3 how we get this formula from Ito's lemma.

Going back to our original definition now that we will be using when discussing reflected Brownian motions:

$$\tilde{Z}(t) = X(t) + M(t)$$

we have that in the one-dimensional case we know the marginal distribution from the fact $X(t)$ is a Brownian motion as well as from the fact the function describing the boundary behaviour is a single value that will describe the fact that the process will go upwards at the beginning of a new busy period and will never go below 0. Note that in the d-dimensional case we will have a matrix and a vector to define this boundary behaviour, however, since we will only be focused on the one-dimensional case we will only use $M(t)$.

For some value of time s in $[0, t]$ a given time interval that we wish to look at, with t a fixed value. We have that from the definition if the reflected Brownian motion and what we know about the supremum given by $M(t)$, we have that $M(t)$ will be non-decreasing in t and will only have an affect on the Brownian motion at the time in which $\tilde{Z}(t) = 0$ because of the fact it is the supremum of the Brownian motion $X(s)$. Thus it can be used to accurately describes the boundary behaviour of the one dimensional process.

Now that we have an idea of the behaviour of the reflected Brownian motion and the functions used to model it, we can calculate the distributions for general values. The marginal distribution of the reflected Brownian motion was found by Harrison [6] to be:

$$\mathbb{P}(\tilde{Z}(t) \leq z) = \Phi\Big(\frac{z - \mu t}{\sigma t^{1/2}}\Big) - e^{2\mu z/\sigma^2}\Phi\Big(\frac{-z - \mu t}{\sigma t^{1/2}}\Big)$$

for all points in time $t \geq 0$ with Φ being the cumulative distribution function (CDF) of the normal distribution and $\mu < 0$ and z an arbitrary value of $\tilde{Z}(t)$. For this project we

have left out the exact calculations of the distribution as they are not necessary to know for our studies. The focus will be on what happens to the process in the limiting case as $t \to \infty$ which will be discussed next.

What is important to note for our project is the limiting distribution that was found by taking $t \to \infty$. This was found to be the exponential distribution with the marginal distribution in the limiting case to be :

$$\mathbb{P}(\tilde{Z} \leq z) = 1 - e^{2\mu z/\sigma^2}$$

Thus for the one dimensional reflected brownian motion we have that the process $\tilde{Z}(t)$ tends towards an exponential distribution with parameter $(-2\mu)/\sigma^2$. This is what gives us the heavy traffic property as well and why reflected Brownian motions can be modelled using the heavy traffic property. The heavy traffic property and it's uses in modelling queue's will be proven and discussed in more detail in section 4 of this project.

3.3 Sticky Brownian Motion

The sticky Brownian motion if a type of Brownian motion where once the process reaches 0 it will stay there for a longer period of time than it would stay at any other value of Brownian motion The "stickiness" is how long the process will stay at 0 for before leaving the origin, 0, and behaving like a regular Brownian motion from that point onwards until it reaches 0 again. Once the process reaches 0 it will repeat the process sticking for some amount of time before going up and away from the origin once again. In this subsection we will look at the sticky Brownian motion and how it comes from the Brownian motion as well as it's applications to queueing theory which we will then illustrate using the $M/M/1$ queue later on in the project.

The sticky Brownian motion, in this project, will be used to model queue's where the first customer to an idle system has a much longer service time than a customer arriving to an already busy system. The length of this longer service time varies from process to process and there are many different types that will give us a sticky Brownian motion. This extra amount of time can be a constant, an exponential random variable, or another type of random variable with a mean that is longer than that of the mean of the service time distribution if the customer arrives to a busy system. This can come up in many real life cases such as if in order to complete service a specific machine is needed that is turned off if the system is idle. Restarting the machine may take a long time and thus the service distribution for the fist customer to an empty system will be entirely different than to a busy one. For this project we will focus on the case where the first customer to an empty queue has a longer service time which is what will give us the sticky Brownian motion.

To introduce the sticky Brownian motion we must first introduce the strong Markov property as to have a sticky Brownian motion we need the process $X(t)$ to have the strong Markov property. The reasoning behind needing the process $X(t)$ to have the strong Markov property is that this is what ensures that the process has stationary and independent increments which comes from the memoryless property of a Markov chain. The strong Markov property states that each stop time is independent if the times before and after, even on the boundary. This is necessary for the sticky Brownian motion as the first customer will have a different service time, with the strong Markov property this will not affect the memoryless properties. Building on this is how we get the properties of the sticky Brownian motion

The definition of the strong Markov property is given is as follows:

Strong Markov Property: If we have $X = (X_t : t \geq 0)$ is a stochastic process on a given probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with the natural filtration (the filtration that is associated to a stochastic process which records the past behaviour at each time) $\{\mathscr{F}_t\}_{t\geq 0}$ Then for any $t \geq 0$ we define $\mathscr{F}_{t+}$ to be the intersection of $\mathscr{F}_s$ for all $s > t$ then for any stopping time τ on Ω we define:

$$\mathscr{F}_{\tau+} = \{A \in \mathscr{F} : \{\tau = t\} \cup A \in \mathscr{F}_{t+}, \forall t \geq 0\}$$

Where A is an (S, S) valued stochastic process and (S, S) is a measurable space. Then X is said to have the strong Markov Property if for each stopping time τ conditioned on the event $\{\tau < \infty\}$ we have for each $t \geq 0, X_{\tau+t}$ is independent of $\mathscr{F}_{t+}$ given X_τ, Yu [19] .

Recall that the reflected Brownian motion was not required to spend any time on the boundary, it could reflect instantaneously if an arrival happens exactly as the last customer is leaving. With the reflected Brownian motion the time spent at 0 was not fixed to be any longer than the time spent in any other state of the process. The sticky Brownian motion however requires the process to spend some time on the boundary and thus the distribution of the function will be different at 0 than it is at any other state in the process. This is why we need the addition of the definition of what the Brownian motion will do at the boundary as will be described in the definition. For the one-dimensional case the sticky boundary behaviours was first found by Feller [3] when he studied the problem of the domains of the infinitesimal generator associated with a strong Markov process, say $\tilde{X}$ in $[0, \infty)$. A formal definition was then developed by Feller[3] and we will give this definition and as well as a more detailed explanation of it in the paragraphs after.

Recall that in our reflected Brownian motion discussion we introduced the generator $(\mathcal{L}f)(x) = \frac{1}{2}f''(x)$ for bounded, smooth functions f to describe the boundary behaviour. For the reflected Brownian motion we introduced it as an aside whereas for the sticky

Brownian motion it will be in our main definition of how we describe the boundary behaviours.

The formal definition of the sticky Brownian motion is:

Definition: Sticky Brownian Motion If we have $\tilde{X}(t)$ a strong Markov process (i.e it has the Strong Markov Property described above) in $[0, \infty)$, then in the bounds of $(0, \infty)$ X will behave like a standard brownian motion. Then at 0 the possible boundary behaviour if found to be described by:

$$f^{'}(0+) = \frac{1}{2\mu} f^{''}(0+)$$

Where $\mu \in (0, \infty)$ is a given fixed constant and f are functions belonging to the domain of the infinitesimal generator of $\tilde{X}(t)$ that we found earlier as $(\mathcal{L}f)(x) = \frac{1}{2}f^{''}(x)$ Yeo[26].

To get a better understanding of the function f and what it means for the function we need to introduce Ito's lemma which is used to determine the derivative of a time dependant function of a stochastic process. When we take $W(t)$ to be a Brownian motion and $f(W(t))$ is then a function of the Brownian motion we cannot simply apply the chain rule to differentiate like we can in normal calculus. We instead need to use Ito calculus since the Brownian motion $W(t)$ is a time dependant process. Instead of the normal chain rule then when differentiating we have the following:

$$\frac{f(W(t))}{dt} = f^{'}(W(t))dW(t) + \frac{1}{2}f^{''}(W(t)dt$$

If we then take the integral of it all we obtain:

$$\int_0^T \frac{df(W(t))}{dt} = \int_0^T f^{'}(W(t))dW(t) + \frac{1}{2}f^{''}(W(t)dt$$

$$f(W(T) - f(W(0)) = \int_0^T f^{'}(W(t))dW(t) + \frac{1}{2}\int_0^T f^{''}(W(t))dt$$

Thus we have it written as what happens to the function between time 0 and time T. The proof and intuition behind why we get this equation for the derivative of the function is beyond the scope of this project so for now we will take it as given and will continue to use Ito's formula to obtain the distributions of the function for the Brownian motion at 0. We can see though how the integral just discussed above will give us the equation we need to describe the behaviour on the boundary of the Brownian motion that is $f^{'}(0+) = \frac{1}{2\mu}f^{''}(0+)$ where μ is the mean of the process. This formula was first found by Ito[24] however, we also used the information in a presentation by Zhang[25] to get a better understanding of the formula and how it is derived.

The infinitesimal generator of $\tilde{X}(t)$ for Markov process is a partial differential operator, meaning it is an operator defined as a function of differentiation. This is the matrix Q described in section 2 when discussing the $M/M/1$ queue and can also be modified to represent any type of queue. This generator includes most of the information about the process and it is what's used in developing Kolmogorov backward equations that describes the probability of the process changing over time and how it will change. Thus the function f belongs to the domain of this generator as we've stated earlier. For our studies that we will be doing on the $M/M/1$ queue we can then assume that the changes can be modelled in the form of a function that we define to be f. It has the same behaviours as a function of a Brownian motion on $(0, \infty)$. The differences here is what happens at the boundary (i.e when $X = 0$). This case, when the process is at the boundary is described by the second derivative $f^{''}(0+)$ which we know by using Ito's calculus on the process to find the behaviour at 0 which we went over briefly above.

Thus we have that the local time limit at 0 is described by the boundary condition for the sticky Brownian motion,

$$f^{'}(0+) = \frac{1}{2\mu} f^{''}(0+)$$

which we can also take from the aside we talked about for the reflected Brownian motion but with a longer time spent at the boundary. This is how we can connect the different boundary behaviours of each process and see how one can be gotten from the other.

We can also obtain the definition of the sticky Brownian motion by looking at the reflected Brownian motion as done in the paper by Dai[1] who used the work first done by Ito and McKean[11]. The one-dimensional sticky Brownian motion $\tilde{X}$ can be constructed from the one-dimensional reflected Brownian motion $\tilde{Z}$ by the time change of $t \to T(t)$ which is done so that the time is now a function of t. This allows us to see the ways that the Brownian motion will stick over a period of time at the boundary 0. We have that the definition of $T(t)$ and the inverse of it, $S(t)$ is described by:

$$T(s) = S^{-1}(s) \text{ where } S(s) = s + \frac{1}{u} M_s$$

and $T(t) = s$ is determined by the equation:

$$t = s + \frac{1}{u} M_s$$

Where we can recall from the definition of the reflected Brownian motion that

$$M(t) = \sup_{s \in [0,t]} X(s) = M_s$$

This function M_s is a time change process that describes the amount of time that a reflected Brownian motion would spend at 0. Thus we have our equation and definition

for the sticky Brownian motion as a reflected Brownian motion that spends time on the boundary according to the function M_s or when its equal to it's running maximum as we know from our definition of $M(t)$.

For this project we will be using the function f to describe the boundary behaviour rather than the above substitution, however they describe the same behaviours of the reflected Brownian motion that will spend a different amount of time in state 0 than it will at any other state in the process. Th e function f is in the domain of the infinitesimal generator of the process and we can find using Ito's lemma for differentiation of time dependant processes. Thus the Brownian motion with the strong Markov property $\tilde{X}(t)$ can also have it's boundary behaviour described as a time change process. From this and the definition we can conclude that the time change function $T(s)$ allows us to better model the effects that the stickiness at the boundary described by $f^{''}(0+)$ will have on the Brownian motion as a whole. Thus we have a definition of the sticky Brownian motion and in section 6 following we will further these concepts in our illustration of the sticky Brownian motion by the $M/M/1$ queue where the first customer has an exceptional service time.

4 Heavy Traffic Approximation

The heavy traffic or the diffusion approximation is when a queueing model is matched with a diffusion process. This is done for when the queue is busy (i.e the queue is experiencing heavy traffic). This property will give us the general behaviours of the Brownian motion that comes from the queue length process and we will see how the heavy traffic approximation can be applied later in our $M/M/1$ queue illustration. The diffusion process can also be looked at as the solution to the stochastic differential equations. The result will be a continuous time Markov process where the sample paths are almost surly continuous. We will see throughout the remained of this project that the reflected and sticky brownian motion exhibits the heavy traffic property under some limiting conditions, this will be illustrated by the $M/M/1$ queue in the next section, section 5. The heavy traffic condition was first written about by Kingman[14] in 1961 and we will be using his work as a basis for the remainder of this section and in our illustration using the $M/M/1$ queue as well as the work of people after who built upon his work to find heavy traffic approximations for different queue's.

If we let $Q(t)$ be the number of customers in the system at time t we can then scale the process by a factor of n and we get the resulting process:

$$\hat{Q}_n(t) = \frac{Q(nt) - \mathbb{E}(Q(nt))}{\sqrt{n}}$$

The scaled process $\hat{Q}_n(t)$ is then taken as $n \to \infty$ so that we can see the limiting behaviours

of the process and thus see is it exhibits the heavy traffic property easily.

There are three classes under which the heavy traffic or diffusion approximations are considered. Each case has a different fixed and limiting parameters resulting in a different type of process and are given as follows:

1. The number of servers is fixed and the traffic intensity ρ is increased to 1 from below (i.e $\rho \uparrow 1$) The queue length can then be approximated by a reflected Brownian motion (Kingman [14]). In this project we will be illustrating the property using the $M/M/1$ queue so we will be fixing the number of servers to 1.

2. If the number of servers and the arrival rate of customers increases to ∞ from below while the traffic intensity ρ is fixed then the queue length limit can be approximated by a normal distribution, this specified as a diffusion rather than a heavy traffic approximation (Iglehart [10])

3. If a quantity β a function of ρ and the number of servers s defined by $\beta = (1-\rho)\sqrt{s}$ The traffic intensity and number of servers are then increased to ∞ the limiting distribution is then a hybrid of the reflected brownian motion and limit covering to a normal distribution. (Whitt [17])

For this project since we are focusing on the case where the number of servers is fixed and we are only concerned with the first case when $\rho \uparrow 1$. This will be achiever by allowing the arrival rate λ to converge to the service rate μ, and will see that the queue length can then be approximated by a reflected Brownian motion. We can focus on the scaled process $\hat{Q}_n(t)$ representing the number of customers in the queue rather than simply $Q(t)$ as it will make it easier for us to see if the process can be modelled by a reflected Brownian motion by determining if it exhibits the heavy traffic property of not.

The heavy traffic property was first stated by Kingman [14] and is essentially that if the distribution of the function of the equilibrium number of customers is asymptotically negative exponential as traffic intensity $\rho \uparrow 1$ then it will have the heavy traffic property. A formal theorem is given by Kingman as well as classes of queue's with the heavy traffic property. A formal definition for the heavy traffic property that covers all cases is as follows:

Definition : Heavy Traffic: the heavy traffic approximation is the process of matching a queuing model with a diffusion process. To do this we introduce limiting conditions on the parameters of the process. These limiting conditions were described above when discussing the cases and imposing these is what gives us a heavy traffic approximation, or diffusion approximation depending in the condition.

for our project we're focused on the $M/M/1$ queue, fixing the number of servers to 1 and increasing $\rho \uparrow 1$. The queueing system has the heavy traffic property if it can be

accurately approximated by a reflected Brownian Motion and the steady state distribution of this reflected Brownian motion tends towards an exponential distribution. Thus we need to have that the scaled queue length process that we defined earlier as $\hat{Q}(t)$ can be written as:

$$\hat{Q}(t) = X(t) - \inf_{0 \leq s \leq t} \{X(s)\}$$

Where $X(t)$ is a regular Brownian motion and

$$M(t) = \sup_{0 \leq s \leq t} \{X(s)\} \equiv - \inf_{0 \leq s \leq t} \{X(s)\}$$

Since the process $M(t)$ simply reflects and will be a single value as the Brownian motion is one-dimensional we can use that

$$\sup_{0 \leq s \leq t} \{X(s)\} \equiv - \inf_{0 \leq s \leq t} \{X(s)\}$$

.

We also need to be able to show that the distribution of the Brownian motion $X(t)$ tends towards an exponential which is done by finding the distribution of the number of customers in the queue at any time t, $\hat{Q}(t)$. For the heavy traffic condition we will then take the limiting distribution and thus we will loo at the case when $n \to \infty$. Knowing that a queueing process can be modelled by a reflected Brownian motion makes finding the distribution and by extension, the number of customers in the queue, much easier.

In order to prove that this scaled process that we looked at earlier with regards to our definition of the heavy traffic property we will focus on the scaled version of the number of customers in the queue which we can recall is

$$\hat{Q}_n(t) = \frac{Q(nt) - \mathbb{E}(Q(nt))}{\sqrt{n}}$$

We will shot that $\hat{Q}_n(t)$ converges to a reflected Brownian motion with the properties described in the definition under the heavy traffic approximation condition that $\rho \uparrow 1$. Since in order for the system to be consistent we need $\lambda < \mu$ we will be taking that the value of λ approaches that of μ from below. We will use the work don Whitt[6] as well as other presentations that were compiled, mainly the presentation by Shaikhet[20]. To accomplish this we must use the strong law of large numbers. We also need to introduce the Central Limit theorem which we will do after. Thus we will give definitions of these concepts now and then we will continue with our proof that it converges to the reflected Brownian motion afterwards. For our theorems of the strong law of large numbers and the central limit theorem we refer to the presentation by Shaikhet [20] as his definitions relate the general theorems better to queuing models.

Strong Law of Large Numbers (SLLN) States that the sample average converges almost surely to the expected value, i.e. we have that $\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}(X)$, that is that we have where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}(X)$:

$$\mathbb{P}\big(\lim_{n\to\infty} \bar{X}_n = \mu\big) = 1$$

Central Limit Theorem (CLT) We also note that the central limit theorem states that for some process $X(t) \sim D(\mu, \sigma^2)$ where D is any distribution,

$$\frac{(X_n - \mu n)}{(\sigma^2 n)^{1/2}} \to N(0,1)$$

as $n \to \infty$. Thus as an extension of the central limit theorem we also have that the scaled process of X_{nt} has:

$$\frac{(X_{nt} - \mu nt)}{n^{1/2}} \to \sigma N(0,t)$$

Thus we have for our definition that the process $M(t)$ the average of it $\bar{M}(t) = \mu t$ and for the Brownian motion part in the definition $X(t)$ we have that $\bar{X}(t) \to mt$ where m is the mean of the Brownian Motion. Both of these are obtained by using the SLLN that we just discussed. We can then use what we know about the means and the convergence of $M(t)$ and $X(t)$ by the SLLN and add into it the definition of the functional CLT just discussed to show that $\hat{Q}(t)$ will converge to a reflected Brownian motion under the heavy traffic approximations.

Since $\rho = \frac{\lambda}{\mu} \uparrow 1$. Since we also need the system to be consistent we, again, take λ approaches μ from below. This means that we have we're approaching the case where $\lambda = \mu$. We then take the value of $\hat{Q}(t)$ to be:

$$\hat{Q}(t) = X(t) + M(t)$$

Since we have that $\hat{Q}(t)$ is the number of customers in the queue we can model this by a Brownian motion which we've discussed in the section on the Brownian motion by shortening the time interval and increasing the number of steps within that interval. Thus we can model it by a brownian motion, say $X(t)$ here, then we need something to describe the behaviour at 0 and for this we use the process $M(t)$ in order to ensure it will reflect after reaching 0 and will not go below. We then need to find the distributions of both $\hat{Q}(t)$ and $M(t)$ and ensure that this is an appropriate model for the number of customers. We do this with the following:

$$\hat{Q}(t) \equiv \frac{Q(nt) - mnt}{\sqrt{n}} \tag{2}$$

$$\to \sigma X(t) \tag{3}$$

Where we get from (2) to (3) using the functional central limit theorem and strong law of large numbers stated earlier. Similarly we have that:

$$\hat{M}(t) \equiv \frac{M(nt) - \mu nt}{\sqrt{n}} \tag{4}$$

$$\rightarrow -\mu Q(t)(\mu t) \tag{5}$$

Where we know that $X(t)$ is the Brownian motion as defined earlier in section 3.1 and $Q(t)$ also converges to a Brownian motion that we found earlier to be $\sigma X(t)$. Thus we have that is is a reflected Brownian motion comes from the scaled process for the number of customers in the queue that we defined to be:

$$\hat{Q}_n(t) = \frac{Q(nt) - \mathbb{E}(Q(nt))}{\sqrt{n}} \rightarrow X(t) + \hat{M}(t)$$

Which we have shown above converge to Brownian motion and thus we can conclude that under the condition that $n \rightarrow \infty$ the number of customers in a one-dimensional queue will converge to a Brownian motion. We would need to prove all of this if we wish to show that whatever specific queue we choose to look at exhibits the heavy traffic property. We also need to show that it converges to the exponential distribution which we can do my examining the specific probability generating functions of the resulting Brownian motions.

We have shown in this section that the heavy traffic conditions can be used to show that the number of customers can be modelled by a reflected Brownian motion. This will lead us to the heavy traffic property that will also help us to find the distribution of this Brownian motion. These concepts will be illustrated in the next section as we discuss the $M/M/1$ queue and its distributions first in the case of normal service times and then later on in the case where the first customer has an exceptional service time.

5 Heavy Traffic Approximations as Illustrated by the $M/M/1$ Queue

In this section we will be using definition of the heavy traffic approximation discussed in the previous section which looked at the process as the traffic intensity $\rho \uparrow 1$ and $n \rightarrow \infty$. We showed that under these conditions the number of customers in a queue will approach a reflected Brownian motion. We will now be illustrating this using the $M/M/1$ queue and show that under the heavy traffic approximation conditions it can also be modelled by a reflected Brownian motion. We will take the limiting distribution as $t \rightarrow \infty$ and see that the process tends towards an exponential distribution. Kingman [14] was the first to calculate the parameters for the expectation and variance of the reflected Brownian motion under the heavy traffic conditions and in this project we will be using his work as a basis

as well as taking references from other work done by Haflin and Whitt [6] who built upon Kingman's work to further solidify the properties for different one-dimensional or single server queue's. We use their work to find the distributions in the $M/M/1$ queue.

Kingman wrote about the single server queue in heavy traffic for the generalized $G/G/1$ queue, we will be taking his work and specifying it to the case of the $M/M/1$ queue. Recall that for an $M/M/1$ queue the arrivals are a Poison process with rate λ and the service time is distributed exponentially with rate μ. To see the heavy traffic approximation we must show that the process can be modelled by a reflected Brownian motion, and then that the distribution tends towards an exponential. In order to do this we will introduce or recall the following notations for the different parameters in the $M/M/1$ queue:

1. The Arrival process is defined by $A(t)$ which we have as a poison process with rate λ
2. The Service process is $S(t)$ which is exponential with rate μ
3. We let $Y(t) \equiv A(t) - S(t), t \geq 0$ be the input process when the queue is experiencing heavy traffic
4. Then as before we have the number of people in the system $Q(t) = Y(t) - \inf_{0 \leq s \leq t}\{Y(s)\}$
5. The system starts off empty with $Q(0) = 0$.

For the third assumption note that if it were the net process for the entire queue length we would need to add the case when the queue is empty, however, we have that under the heavy traffic conditions the number of customers is approaching infinity (i.e $n \to \infty$). Thus we can assume that the probability that the system is idle will be 0 when calculating $Y(t)$. We will look at the system when it is experiencing heavy traffic as we can use the heavy traffic assumptions on it and when it is at 0 separately. We will see that this is what gives us the total number of people in the system to be $Q(t) = Y(t) - \inf_{0 \leq s \leq t}\{Y(s)\}$, the process when it is busy and then adding the case when it is at the boundary by using the function $-\inf_{0 \leq s \leq t}\{Y(s)\}$ which will describe the behaviour of the queue length process when it is at 0. This is how we get the fourth assumption for how we write the process for the number of people in the queue $Q(t)$.

Note that $Q(t)$ is a reflection of the process $Y(t)$, this means that however $Y(t)$ behaves, $Q(t)$ will behave in a similar fashion but will reflect at the origin (i.e at point 0) since it also takes the negative infimum of the process. This comes from the definition of the reflected Brownian motion that we found in section 3.2. Thus here we need to show that $Y(t)$ is a Brownian motion and that

$$M(t) = -\inf_{0 \leq s \leq t}\{Y(s)\}$$

which is the process $M(t)$ does not have any effect on the Brownian motion except at 0 and the only thing it will do is ensure that the Brownian motion will reflect as was out definition in section 3.2 when we discussed the reflected Brownian motion.

We have that by the strong law of large numbers discussed in section 4 that the mean of the arrival process $A(t)$ and $S(t)$ will converge to λt and μt respectively as $n \to \infty$. That is we have that:

$$\bar{A}(t) = \frac{A(nt)}{n} \to \lambda t$$
$$\bar{S}(t) = \frac{S(nt)}{n} \to \mu t$$

From this we will then use the central limit theorem on each of the arrival and service processes as well to show that the both will converge to a Brownian motion with their respective values depending on λ and μ. For now we will look at the arrival and service processes separately with their respective rates λ and μ and need to show that they will each converge to their own Brownian motions. Then we will take $\rho \uparrow 1$ with λ converging to μ from below so that we can show that the combined process $Y(t)$ will also converge to a Brownian motion.

For the arrival process $A(t)$ we scale it by taking nt as the time and then we subtract the mean λnt and divide by $\sqrt{n}$ in the same way we had the scaled process $\hat{Q}(nt)$ in the previous section. Using what we obtained above about the mean of the process and the central limit theorem (i.e. the process subtract the mean over the square root of n as $n \to \infty$ will converge to a $N(0,1)$) we have the following:

$$\frac{A(nt) - \lambda nt}{\sqrt{n}} \to \sqrt{\lambda} B_a(t)$$

Again, this is by the central limit theorem and the strong law of large numbers. The process $B_a(t)$ is the Brownian motion corresponding to the arrival process which will have parameters based on the value of λ. Thus the arrival process converges to a Brownian motion as $n \to \infty$.

Obtaining the distribution of the service process is extremely similar to that of the arrival process. The service process also uses the strong law of large numbers and the central limit theorem in the exact same way, the only difference being that we have its parameter is μ. Thus we have that $S(t)$ as $n \to \infty$ will converge to:

$$\frac{S(nt) - \mu nt}{\sqrt{n}} \to \sqrt{\mu} B_s(t)$$

Where $B_s(t)$ is the Brownian motion corresponding to the service times and the central limit theorem was used in the exact same way as it was to find the limiting distribution of

the arrival time.

We will now find the distribution of the combined process $Y(t) \equiv A(t) - S(t), t \geq 0$. to do this we will impose the heavy traffic condition that $\rho \uparrow 1$ which means that we're approaching the case when $\lambda \to \mu$ that we talked about earlier in the section. To find the limiting distribution of $Y(t)$ we use the same logic as we did with the arrival process $A(t)$ and the service process $S(t)$. We know already that the means of the processes $A(t)$ and $S(t)$ converge to λt and μt respectively by the strong law of large numbers. Then since we have that $\lambda \to \mu$ we have the following by the strong law of large numbers:

$$\begin{aligned} \bar{A}(t) - \bar{S}(t) &= \frac{A(nt)}{n} - \frac{S(nt)}{n} \\ &\to \lambda t - \mu t \\ &\equiv 0 \end{aligned}$$

Thus we have that the mean of the process for $Y(t)$ will also converge to 0 by the strong law of large numbers. Then using this fact as well as the central limit theorem we have that

$$\frac{Y(nt)}{\sqrt{n}} \equiv \frac{A(nt) - S(nt)}{\sqrt{n}} \tag{6}$$

$$\to \sqrt{\lambda} B_a(t) - \sqrt{\mu} B_s(t) \tag{7}$$

$$\stackrel{d}{=} \sqrt{2\lambda} B(t) \tag{8}$$

Where to get from equation (6) to (7) we use what we found previously about the limiting distributions of the arrival and service times are Brownian motions, also since we're assuming $\lambda \to \mu$ we can replace μ with λ without loss of generality. Also since $\lambda \to \mu$ we have that the distributions $B_a(t) = B_s(t)$ since the only difference between the two before was the values of λ and μ. Thus we can replace and them both with $B(t)$ a standard Brownian motion and that is how we obtain equation (8).

Then, from how we previously defined the scaled number of customers in the queue as:

$$\hat{Q}_n(t) = \frac{Q(nt) - \mathbb{E}(Q(nt))}{\sqrt{n}}$$

And also how we defined the queue length process with respect to the input process when the queue is experiencing heavy traffic, $Y(t)$ which we know from the strong law of large numbers that we discussed earlier has mean 0, thus $\mathbb{E}(Q(nt)) = 0$. And when the process is at 0 which the behaviour is given by $M(t)$, thus we have:

$$\hat{Q}_n(t) = \frac{Q(nt)}{\sqrt{n}} \to Q(t) \equiv Y(t) - \inf_{0 \le s \le t}\{Y(s)\}$$

From the work that we've done above we know that $Y(t) \equiv \sqrt{2\lambda}B(t)$, a Brownian motion. Thus this part matches with our definition of the reflected Brownian motion done in section 3.2. Now we need to see what the negative infimum of the process $Y(t)$ which we denote:

$$M(s) = -\inf_{0 \le s \le t}\{Y(s)\}$$

Recall our definition of a reflected Brownian motion in section 3.2 where we had the reflected Brownian motion $\tilde{Z}(t) = X(t) + M(t)$ where $X(t)$ was a Brownian motion and $M(t) = \sup_{s\in[0,t]} X(s)$. The negative of the infimum will be the supremum, as well as our change of notation from X to Y representing the Brownian motion, and thus we have that, $M(t) = -\inf_{0\le s\le t}\{Y(s)\}$ matches with the $M(t)$ in our definition of the reflected Brownian motion and therefore in the $M/M/1$ queue the number of customers will converge to a reflected Brownian motion under the heavy traffic conditions as we expected.

To now prove that the process has the heavy traffic property we must also prove that the limiting distribution of the reflected Brownian motion that comes from implementing the heavy traffic conditions on the the number of customers in the $M/M/1$ queue will tend towards that of an exponential. To achieve this we have similar assumptions and notation on the process as we had before except with the addition of a drift coefficient in order to cover all possible cases in the $M/M/1$ queue and we will show that it will always converge towards an exponential. For this we will only be looking at when the queue is experiencing heavy traffic and thus we again assume that the probability the number of customers is at 0 will be 0. We have:

1. as $n \to \infty$ with λ_n the rate of arrivals as a function of n
2. if $(\lambda_n - \mu)\sqrt{n} \to c$ for some constant c. This means that $\rho_n \equiv 1 - (c/\sqrt{n})$ then we have that:
3. Recall that the arrival process $[A_n(nt) - \lambda_n nt]/\sqrt{n} \to \sqrt{\mu}B_a(t)$ where $B_a(t)$ is the Brownian motion associated with the arrival process
4. Recall that theservice process $[S_n(nt) - \lambda_n nt]/\sqrt{n} \to \sqrt{\mu}B_s(t)$ where $B_s(t)$ is the Brownian motion associated with the service process

Since we now have the addition of the fact that $(\lambda_n - \mu)\sqrt{n} \to c$ we will again find the limiting distribution of the combines process $Y(t)$ with this addition as $n \to \infty$ and $\rho \uparrow 1$. We will use the strong law of large numbers and the central limit theorem again in the same way we uses them earlier and we obtain the following:

$$\frac{Y_n(nt)}{\sqrt{n}} \equiv \frac{A_n(nt) - S_n(nt) - (\lambda - \mu)nt}{\sqrt{n}} \tag{9}$$

$$\to \sqrt{\lambda}B_a(t) - \sqrt{\mu}B_s(t) - ct \tag{10}$$

$$\stackrel{d}{=} \sqrt{2\mu}B(t) - ct \tag{11}$$

These calculations follow directly from the assumptions made above as well as the fact that we're taking $\rho \uparrow 1$ and thus $\lambda = \mu$ and the same result obtained before that this implies that the means of the arrival subtracted from the service process will be 0. Then we add in what happens to the process at 0 with the same reasoning for using $M(t)$ that we've described earlier. We have that the queue length process that we had before:

$$\hat{Q}_n(t) = \frac{Q(nt)}{\sqrt{n}} \to Q(t) \equiv Y(t) - \inf_{0 \le s \le t}\{Y(s)\}$$

We found just above that $Y(t) \equiv \sqrt{2\mu}B(t) - ct$. The process for $M(s) = -\inf_{0 \le s \le t}\{Y(s)\}$ is the exact same as before. Thus we have that $Q(t)$ is a reflected Brownian motion with drift ct. We must now show that this process will tend towards a exponential.

For the calculations of the distribution and showing its exponential we mostly look at the paper by Kingman[13] where he found the distribution of the single server queue in heavy traffic to be asymptotically (i.e. the limiting distribution) exponential. To obtain this we have that given n is the length of the queue upon arrival of a customer, we have that for some value of z:

$$\mathbb{P}(\hat{Q}(t) \ge z) = \mathbb{P}(Y(t) - \inf_{0 \le s \le t}\{Y(s)\} \ge z) \tag{12}$$

$$= \mathbb{P}(-\inf_{0 \le s \le t}\{Y(s)\} + Y(t) \ge z) \tag{13}$$

$$= \mathbb{P}(\sup_{0 \le s \le t} \hat{Y}(s) \ge z) \tag{14}$$

Where to get from line (13) to (14) we have that we define $\hat{Y}(t) = Y(0) - Y(0-t)$ and we observe that this is also a Brownian motion with drift ct. We make this substitution so that we can observe the exact distributions of $\hat{Q}(t)$ easily. This substitution is possible as it results in the same Brownian motion since we are only changing the value when the process is at 0 and simply adding it as 0, and thus is equal.

We now have two cases depending on the drift. The first is when $ct \ge 0$ and the second is when $ct < 0$. We will see the distributions that each case gives us as follows:

Case 1: We then have that from the definition and basic properties of the Brownian motion that when the drift $ct \geq 0$ we have that $\sup_{s\geq 0} Y(s) = \infty$ Therefore we have that :

$$\mathbb{P}(\sup_{0\leq s\leq t} \hat{Y}(s) \geq z) \to 1 \text{ as } t \to \infty$$

Thus the limiting distribution does not exist and we do not have to worry about this case anymore.

Case 2: Is when we have that the drift $ct < 0$ then we have that

$$\mathbb{P}(\sup_{0\leq s\leq t} \hat{Y}(s) \geq z) \to exp(-2ct/\text{var}(Q(t)))$$

Thus we have that in this case $\hat{Q}(t)$ will converge to an exponential distribution under the heavy traffic conditions. Thus we have that the number of customers in the $M/M/1$ queue under the heavy traffic conditions will converge to a reflected Brownian Motion with an exponential distribution as was our goal for this section. This illustration by the $M/M/1$ queue allows us to see how the heavy traffic property and Brownian motion approximations work in queueing.

In the following section we will build upon these concepts for the $M/M/1$ queue when the first customer to an empty queue has a longer service time. Recall that when discussing the reflected Brownian motion we also introduced the boundary behaviour using the derivative of a function on the domain the infinitesimal generator of the queue length process $Q(t)$. We will be using this definition in the following section when discussing the sticky Brownian motion and will see the similarities and differences between the two processes.

6 Sticky Brownian Motion as Illustrated by the $M/M/1$ Queue

If we have queue that behaves like an ordinary $M/M/1$ queue when the system is busy but after the system is idle the next arriving customer will have an exceptional service time, usually much longer. In section 3.3 we discussed the general ideas of the sticky Brownian motion and how if can be a special version of the reflected Brownian motion where the possible boundary behaviour is described by

$$f^{'}(0+) = \frac{1}{2\mu} f^{''}(0+)$$

where the function f is as defined in section 3.3 and the behaviour at 0 being the derivative, which we found in section 3.3 as well using Ito's lemma. The function f as we described before is from the domain of the infinitesimal generator and the exact derivation

is beyond the scope of this project. Some explanation is given in section 3.3 with regards to Ito's formula however for this project we will mostly be taking it as given in order to focus more on the overall behaviours of the function rather than the details. The theory behind why f is chosen is, in reality, much more complicated than we've explained here. The main use of the function for us will be to describe the boundary behaviour, as stated before. To do this we will only show that the behaviour of our queue at 0 will converge to the function described in section 3.3, $f^{'}(0+) = \frac{1}{2\mu} f^{''}(0+)$ to prove it will converge to a sticky Brownian motion.

In the previous section we showed the the number of customers in a queue has the heavy traffic property and can be modelled by the reflected Brownian motion. In this section we will build on that for the special case where the first customer to an empty system has an exceptional service time. We will show that the number of people in the queue in thus case can be modelled by sticky Brownian motion. Recall in our definition of the reflected Brownian motion we also introduced it as a function of f and thus we can show that the definition can be extended to include the case where it sticks.

To show this we will use the work done by Yeo[18] and to begin we will introduce some of the nutation that he used in the paper as well as recall some of the notation that were introduced previously in the project:

1. $\tau_1, \tau_n, ...$ are the arrival instances of each customer
2. $t_n = \tau_n - \tau_{n-1}$ the inter arrival times where $t_0 = 0$. They are independently and identically distributed (taken from the fact $M/M/1$ with rate λ and expectation $\mathbb{E}(t_n) = 1/\lambda$)
3. s_n the actual service time of the $n-th$ customer if they join an already busy queue, with expectation $\mathbb{E}(s_n) = 1/\mu$ the expected service time in a regular $M/M/1$ queue. The cumulative frequency is defined to be $\phi(\theta) = \int_0^\infty e^{i\theta x} dB(x)$ where $B(x)$ is the distribution of the service time if the system is busy. We know $B(x)$ to be exponential in out case.
4. r_n the service time of the $n-th$ customer if they join an empty queue, with expectation $\mathbb{E}(r_n) = d$. Note that this service time begins immediately however it will be longer than s_n. The cumulative frequency is defined to be $\zeta(\theta) = \int_0^\infty e^{i\theta x} dD(x)$ where $D(x)$ is the distribution of the service time if the customer is arriving to an empty queue. This distribution is unknown to us and needs to be obtained through observation of the queue. We will be using a general $D(x)$ for our calculations
5. $u_n = s_n - t_{n+1}$ the difference between the $n-th$ service time and next inter arrival if the queue is busy.

6. $c_n = r_n - t_{n+1}$ the difference between the $n-th$ service time if they came to an empty queue and the next inter arrival time after them.

7. $R_n(m) : n = 0, 1, 2, ...$ is the probability the m-th arrival finds n in the queue

8. $Q_n(m) : n = 0, 1, 2, ...$ is the probability that when the m-th arriving customer leaves they leave n in the queue.

9. Note that $Q_n = R_n$ where Q_n and R_n is the limiting distribution of how many the departing and arriving customers see respectively, thus Q_n is the probability a departing customer sees n in the system.

From this we can then easily see that the wait time for this special type of $M/M/1$ queue will be:

$$w_n = \begin{cases} w_n + u_n & \text{if } w_n + u_n > 0, w_n > 0 \\ c_n & \text{if } c_n > 0, w_n = 0 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Yeo[19] also found the distributions of the wait times. Since in this project we are focused on the number of customers in the queue and not the wait times we will take his work on this as given and simply use to to further our studies into the queue length process. We have that the probability that a customer will arrive to find the system empty was found, again in the paper by Yeo[19] and modified for our case of the $M/M/1$ queue, to be:

$$W(0) = \frac{1 - \lambda/\mu}{1 - \lambda/\mu + \lambda d}$$

where we recall that d is the mean of the wait time before a customer arrives to an empty queue. He also found the expectation of the wait time distribution to be:

$$\mathbb{E}(w) = \frac{\lambda\mathbb{E}(r^2) - \lambda^2\mathbb{E}(r^2) + \lambda^2 d\mathbb{E}(s^2)}{2(1 - \lambda/\mu)(1 - \lambda/\mu + \lambda d)}$$

Now that we have a general idea of how the queue will behave we can find the distribution of number of people that will be in the queue at any time. We note that the Brownian motion will go from 0 to 1 after the first customer that arrived to an empty queue has been served, so for the duration of time the first customer is being served the Brownian motion will stay at 0 even though a new customer has arrived. After leaving the origin the process can be modelled by a regular Brownian motion, this means that more customers arrived while the first was being served and the queue is now busy. Since the distributions for this were already fond in the previous section we will only focus on what happens when the queue is empty.

Denote the number of people in the queue by $Q(z)$ as we did in the previous sections and we will be looking at some small value $|z| < 1$, as we now wish to focus on when the queue is empty and then becomes busy, and we will assume also some small value of k such that $\lambda/\mu < k$. This will ensure that the probability of the queue being empty is high enough that we can focus on what happens at 0 and how the queue length process will behave. Because of this is also why we look at the process as being discrete rather than continuous when looking at the boundary behaviour, the exact opposite of the assumptions in heavy traffic.

The arrival process which we can recall is denotes $A(t)$ is distributed poission with rate λ, thus we have that we can write the distribution of $A(z)$ for our value of $|z| < 1$ and the chosen value of k such that $\lambda/\mu < k$ as:

$$A(z) = \sum_{n=0}^{z} \frac{\lambda^n}{n!} e^{-\lambda}$$

Then we choose $z = k - 1$ which we know will satisfy our conditions as k is some small positive value then we have that:

$$A(z) = \sum_{n=0}^{k-1} \frac{\lambda^n}{n!} e^{-\lambda}$$

We can try and write out the service distribution $S(t)$ in a similar way however we now have the addition of the fact that the first to an empty queue has an exceptional service time. Thus we will start by noting that the probability generating function of the number of customers in the system is the summation from $n = 0$ up to infinity of the probability there are n in the system multiplied by the time we're looking at to the power of n thus we have:

$$Q(z) = \sum_{n=0}^{\infty} Q_n z^n$$

Where we can recall from out notation that Q_n is the probability that the departing customer will leave behind n in the system. We are now concerned with finding the probability generating function of $Q(z)$. We are not yet sure what type and if this process can be modelled by a Brownian motion and thus we will look at the discrete variables and we will try to characterize them and see if we can model the process with a type of Brownian motion, in our case we will see that it will be modelled by a sticky Brownian motion. To do this we will look at the process as discrete, as we've mentioned before, to see the behaviours at 0 and then the rest we will use the heavy traffic property to show it will behave like a Brownian motion after leaving the origin.

We then let f_n be the function represents the integral of the distribution function of for the customers that arrive to a busy system. This is a function defined by the domain of the infinitesimal generator of our process as we looked at in section 2 when defining the $M/M/1$ queue and in section 3.3 when defining the sticky Brownian motion. It is the function that will model how the Markov chain of the number of customers will behave. Then we have that the probability generating function will be $\sum_{n=0}^{\infty} f_n z^n$ by the same logic we used when finding the probability generating function of the number of customers $Q(z)$. We already know the distribution of this from our work done in the previous section under the heavy traffic condition. However, since we are now concerned with small values for the number of customers, for now we will look at the probability generating function of it as a discrete random variable as it will help us to better understand later what happens at 0.

In order to accomplish this we define the function f_n which is the function in the domain of the infinitesimal generator of the process and we are now going to assume that we do not know that service is exponentially distributed as it will help us set up the same type of equation f_n^* that we will use for the service time when the system is empty. f_n^* is the probability that the number of customers in the queue will be at n at some given time in the future, then we take $f(z)$ as the sum of these probabilities multiplied by the fractional amount $|z| < 1$ to the power n to give the probability generating function $f(z)$ Yeo[19] :

$$f_n = \int_0^{\infty} e^{-\lambda x} \sum_{r=nk}^{(n+1)k-1} \frac{(\lambda x)^r}{r!} dB(x) \tag{16}$$

$$f(z) = \sum_{n=0}^{\infty} f_n z^n \tag{17}$$

As mentioned earlier we already know the distribution of the service time when the system is busy,$B(x)$, to be exponential with rate μ since we're looking at the $M/M/1$ queue and we know the limiting distribution of the probability generating function will lead us towards a reflected Brownian motion in heavy traffic. This definition however is what leads us to the very similar definition of what happens at the boundary (i.e. at 0) which we will give by f_n^*, which is the integral of the distribution when the customer arrives to an empty system. We have by the same logic as for calculating $f(z)$ except now $D(x)$ is the distribution of the service time for the first customer to an empty system. This probability generating function $f^*(z)$ in this case will be:

$$f_n^* = \int_0^\infty e^{-\lambda x} \sum_{r=nk}^{(n+1)k-1} \frac{(\lambda x)^r}{r!} dD(x) \tag{18}$$

$$f^*(z) = \sum_{n=0}^{\infty} f_n^* z^n \tag{19}$$

The values when the customer arrives to the busy system, so the values of f_n and $f(z)$ were calculated in the previous section when we found the distribution when the system is experiencing heavy traffic can be modelled by a reflected Brownian motion with exponential distribution. However, like we mentioned earlier we are not necessarily looking at heavy traffic now as we are mostly concerned with what happens to the process at 0, thus we look at it as discrete variables. The distribution function $B(x)$ is known to be the exponential distribute with rate μ. Thus the main difference we have now is f_n^* which is different from f_n only through the distribution of the extra wait time denoted by $D(x)$, the distributions when the system is empty.

We now wish to calculate the probability generating function for the number of customers in the queue. The way to do this is as follows:

$$Q_n = Q_{n+1} f_0 + Q_n f_1 + \cdots + Q_1 f_n + Q_0 f_n^* \tag{20}$$

$$Q_n z^n = z^n [Q_{n+1} f_0 + Q_n f_1 + \cdots + Q_1 f_n + Q_0 f_n^*] \tag{21}$$

$$\sum_{n=0}^{\infty} Q_n z^n = \sum_{n=0}^{\infty} z^n [Q_{n+1} f_0 + Q_n f_1 + \cdots + Q_1 f_n + Q_0 f_n^*] \tag{22}$$

$$Q(z) = Q_0 \Big(\frac{f(z) - z f^*(z)}{f(z) - z} \Big) \tag{23}$$

The steps are as follows:

1. to get from (20) to (21) we simply multiply both sides by z^n
2. to get from (21) to (22) we take the summation of each side from $n = 0$ to ∞
3. to get from (22) to (23) we take the limit as $n \to \infty$ of both sides.

Thus we have that the number of customers $Q(z) = Q_0\{f(z) - z f^*(z)\}\{f(z) - z\}^{-1}$ where the value of $f(z)$ is known. The value for the number of customers in the queue is found under part of the heavy traffic conditions, i.e. at $\rho = 1$, however we are not assuming that $n \to \infty$ here as what happens when n is small also matters.

By this assumption we have that the probability of there being 0 customers in the queue, since we have that $f^{'}(1) = \lambda/\mu$ and $f^{'*} = \lambda d$ will now be:

$$\begin{aligned} W(0) &= \frac{1 - \lambda/\mu}{1 - \lambda/\mu + \lambda d} \\ &\to 0 \end{aligned} \tag{24}$$

Through differentiation of $Q(z)$ Yeo[19] found the average number of customers in the queue to be:

$$Q^{'}(1) = \frac{Q_0\{2f^{*'}(1) + f^{*''}(1)\}}{2(1 - f^{'}(1))}$$

Thus from what we've just seen through the work done previously on the sticky Brownian motion and the work by Yeo [19] we can see that the number of customers in this $M/M/1$ queue behaves like a Brownian motion away from the origin using the heavy traffic property, and then we can use the probabilities of the number of customers in the queue as a discrete random variable to calculate what happens to this special case of the reflected Brownian motion at 0, even with the first customer having an exceptional service time. We also illustrated that it has a similar distribution to the reflected Brownian motion except that we have the addition of f^* which represents the time that the process will spend on the boundary. This value is what will determine $f^{'}(0+) = \frac{1}{2\mu} f^{''}(0+)$ in our original definition.

The number of customers in the queue at time t found above matched our definitions of a sticky Brownian motion and exhibits the heavy traffic property as we know that the distributions of f and $f^{'}$ will be from the exponential distribution as we have an $M/M/1$ queue and it was proven in section 5 that this results in the reflected Brownian motion. For this case we have the addition of f^* which is what represents the stickiness of the process at the boundary point 0. This is shown in the following as we need to further differentiate it to find the number of customers in a queue. We also take the value as k approaches 0 rather than at 1 to see more of the boundary behaviour. We will still have that the probability $Q_0 = 0$ however we now have the case where $\mu \to \lambda$ rather than $\lambda = \mu$ Thus we take the value of $Q^{'}(0+)$ and we see that:

$$Q^{'}(0+) = \frac{Q_0\{2f^{*'}(0+)\}}{2(1 - f^{'}(0+))} + \frac{f^{*''}(0+)\}}{2(1f^{'}(0+))} \tag{25}$$

$$= \frac{1}{2(1 - f^{'}(0+))}\Big[Q_0(2f^{*'}(0+)) + f^{*''}(0+)\Big] \tag{26}$$

$$= \frac{1}{2(1 - (1 - \mu))}\Big[Q_0(2f^{*'}(0+)) + f^{*''}(0+)\Big] \tag{27}$$

$$\to \frac{1}{2\mu}[f\,*''(0+)] \tag{28}$$

This is exactly our definition of the sticky Brownian motion with the "stick" caused by the second derivative of $f*''(0+)$ in this case. Thus we have shown that the $M/M/1$ queue where the first customer to an empty system has exceptional service times will converge to a sticky Brownian Motion. Thus we can use the sticky Brownian motion to model this type of queue process.

7 Conclusion

We have now illustrated the concepts of the reflected and sticky Brownian motion by using the $M/M/1$ queue. For the queue with a standard exponential service rate for all of the customers we found that the process $Q(t)$ converges to a reflected Brownian motion by using the heavy traffic conditions that $\rho \uparrow 1$ and $n \to \infty$. With this we found that:

$$\hat{Q}(t) \to Y(t) - \inf_{0 \leq s \leq t} \{Y(s)\}$$

which is our definition of the reflected Brownian motion that we looked at in section 3.2. We also found that the distribution of this will be exponential with:

$$\mathbb{P}(\sup_{0 \leq s \leq t} \hat{Y}(s) \geq z) \to \exp(-2ct/\text{var}Q(t))$$

Thus we showed that under the heavy traffic conditions that the process will converge to a reflected Brownian motion with an exponential distribution which is exactly what the definition of the heavy traffic property stated it would do.Through this illustration we solidified out understanding of the reflected Brownian motion and the heavy traffic property by allowing us to see how the heavy traffic condition works in the $M/M/1$ queue.

We also found in section 6 the distribution of the number of customers in a queue, $Q(t)$ when the first customer to an empty queue has a longer service time than a customer that arrives to an already busy queue. To do this we used the fact that away from the origin we can use the heavy traffic property and what we already found about the behaviours of this and seen that it behaves like a Brownian motion. We then focused on what happens at 0. In doing this we found that the distribution of the Brownian motion that results fro the queue length process is a Brownian motion outside the boundary and at 0 the behaviour can be described by:

$$Q^{'}(0+) \to \frac{1}{2\mu}[f^{*''}(0+)]$$

which is the same as in our definition of a sticky Brownian motion and thus proved that when the first customer to an empty system has exceptional service time we can model the number of customers in a queue by using the sticky Brownian motion. This allowed us to see the properties of the sticky Brownian motion that were discussed in section 3.3 as well as furthering the concepts of the heavy traffic property by using it to find the distribution

away from 0. This is what gives us the sticky Brownian motion in this case and again, allowed us to see an actual example of these principles using the $M/M/1$ queue, furthering our understanding of them.

Through these two examples we saw the similarities and differences in the reflected and sticky Brownian motion and how we can use these to

The concepts that we illustrated using the $M/M/1$ queue were also discussed extensively in this project. The reflected Brownian motion where we used a lot of the work done by Harrison [8] as well as Inglehart and Whitt [10] and Dai [2] as a reference. As we mentioned before currently most research is being done on $d-$dimensional reflected Brownian motions in their applications to queueing theory. This involves multi server queues and distributions that are not Poission arrivals and exponential service times. With the information covered in this project one should be able to now grasp the concepts in those papers that more work is currently being done as this project gave an overview of the base knowledge needed to understand them.

The sticky Brownian motion was also covered and this was done using references from Dai [1] and Ito [11] as well as the research done by Welch [17] and Yeo [19] . In this project we combined all of this and more information into an understandable format so that the reader can have a decent understanding of the sticky Brownian motion and its properties. As with the reflected Brownian motion this project will allow us to better understand sone of the more difficult research that is currently being done in the field such as "Stationary Distributions for Two-Dimensional Sticky Brownian Motions: Exact Tail Asymptotics and Extreme Value Distributions" by Dai and Zhao [1] which covers the tail asymptotics of a sticky Brownian motion.These multi-dimensional servers are what is currently being researched in the field and the hope of this project was to give enough understanding of the sticky Brownian motion in order to better understand more complex research such as that paper.

Finally, one of the most important topics covered in this project is the heavy traffic property. This is what enabled us to find the reflected and sticky Brownian motions from queue's. This concept was first discovered by Kingman [13] and [14]. In this project we built upon his work as well as the work done by Hafflin and Whitt[6]. The heavy traffic property tells us the behaviours of the number of customers in the queue as $n \to \infty$ and $\rho \uparrow 1$. These conditions as we saw in section 4 as well as in our illustration by the $M/M/1$ queue id what allows us to see the limiting behaviour of the number of customers in the queue. This property is extremely useful today in queueing theory and is what is used in finding the distributions of multi dimensional queues. It is also used in fluid and diffusion approximations as we looked at briefly in section 4 as well when discussing the different types of conditions we could impose to get different versions of the heavy traffic property. We used this property in out illustration with the $M/M/1$ queue as well. With this we

now have a solid understanding the the heavy traffic property and it's uses, with this we have enough of a foundation to understand some of the more complex research mentioned.

In conclusion, over the course of this project and the research done for it we now have a good understanding of the heavy traffic property and the role it plays when modelling different types of queueing systems. This was illustrated by modelling the number of customers in the queue in the $M/M/1$ queue both under standard conditions and when the first customer to an empty system has a longer service time. We have used the heavy traffic approximations first introduced by Kingman [14] to model these queue systems. We have seen that a reflected and sticky Brownian motion can be used for each type of queue respectively. We have calculated the limiting distributions of each using this heavy traffic property and seen that the Brownian motion is in fact an accurate representation. Thus the goal of the project has been accomplished.

8 References

[1] Dai, Hongshuai, and Yiqiang Zhao. 'Stationary Distributions for Two-Dimensional Sticky Brownian Motions: Exact Tail Asymptotics and Extreme Value Distributions' , 31 Mar. 2020.

[2] Dai, J. G., and J. M. Harrison. 'Reflected Brownian Motion in an Orthant: Numerical Methods for Steady-State Analysis.' The Annals of Applied Probability, vol. 2, no. 1, 1992, pp. 65 − 86., doi:10.1214/aoap/1177005771.

[3] Feller, William. 'Diffusion Processes in One Dimension.' Transactions of the American Mathematical Society, vol. 77, no. 1, 1954, pp. 1?1., doi:10.1090/s0002-9947-1954-0063607-6.

[4] Gautam, Natarajan. 'Analysis of Queues.' 2012, doi:10.1201/b11858.

[5] Guillemin, Fabrice, and Jacqueline Boyer. Queueing Systems, vol. 39, no. 4, 2001, pp. 377 − 397., doi:10.1023/a:1013913827667.

[6] Halfin, Shlomo, and Ward Whitt. 'Heavy-Traffic Limits for Queues with Many Exponential Servers.' Operations Research, vol. 29, no. 3, 1981, pp. 567?588., doi:10.1287/opre.29.3.567.

[7] Harrison, J. Michael, and Austin J. Lemoine. 'Sticky Brownian Motion as the Limit of Storage Processes.' Journal of Applied Probability, vol. 18, no. 1, 1981, pp. 216?226., doi:10.2307/3213181.

[8] Harrison, J. M., and R. J. Williams. 'Brownian Models of Feedforward Queueing Networks: Quasireversibility and Product Form Solutions.' The Annals of Applied Probability, vol. 2, no. 2, 1992, pp. 263 − 293., doi:10.1214/aoap/1177005704.

[9] Harrison, Peter G., et al. 'Negative Customers Model Queues with Breakdowns.' Performance Engineering of Computer and Telecommunications Systems, 1996, pp. 153 − 167.

[10] Iglehart, Donald L., and Ward Whitt. 'Multiple Channel Queues in Heavy Traffic. II: Sequences, Networks, and Batches.' Advances in Applied Probability, vol. 2, no. 02, 1970, pp. 355 −369., doi:10.1017/s0001867800037435.

[11] Ito, K., and H. P. Mckean. 'Brownian Motions on a Half Line.' Illinois Journal of Mathematics, vol. 7, no. 2, 1963, pp. 181 − 231., doi:10.1215/ijm/1255644633.

[12] Karlin, Samuel, and James Mcgregor. 'Many Server Queueing Processes with Poisson Input and Exponential Service Times.?'Pacific Journal of Mathematics, vol. 8, no. 1, 1958, pp. 87 − 118., doi:10.2140/pjm.1958.8.87.

[13] Kingman, J. F. C. 'On Queues in Heavy Traffic.' Journal of the Royal Statistical Society: Series B (Methodological), vol. 24, no. 2, 1962, pp. 383?392., doi:10.1111/j.2517-6161.1962.tb00465.x.

[14] Kingman, J. F. C. 'The Single Server Queue in Heavy Traffic.' Mathematical Proceedings of the Cambridge Philosophical Society, vol. 57, no. 4, 1961, pp. 902 − 904., doi:10.1017/s0305004100036094.

[15] Pang, Guodong, et al. 'Martingale Proofs of Many-Server Heavy-Traffic Limits for Markovian Queues.' Probability Surveys, vol. 4, 2007, pp. 193 − 267., doi:10.1214/06-ps091.

[16] Shaikhet, G. 'Fluid and Diffusion Approximation of Queues.' Presentation Overview

[17] Welch, Peter D. 'On a GeneralizedM/G/1 Queuing Process in Which the First Customer of Each Busy Period Receives Exceptional Service.' Operations Research, vol. 12, no. 5, 1964, pp. 736 − 752., doi:10.1287/opre.12.5.736.

[18] Whitt, Ward. 'Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments.' Management Science, vol. 50, no. 10, 2004, pp. 1449?1461., doi:10.1287/mnsc.1040.0279.

[19] Yeo, G. F. 'Single Server Queues with Modified Service Mechanisms.' Journal of the Australian Mathematical Society, vol. 2, no. 4, 1962, pp. 499 − 507., doi:10.1017/s144678870002749x.

[20] Yu, Shang, et al. 'Experimental Investigation of Spectra of Dynamical Maps and Their Relation to Non-Markovianity.' Physical Review Letters, vol. 120, no. 6, 2018, doi:10.1103/physrevlett.120.060406.

[21] ' Skorokhod Mapping Theorem. Reflected Brownian Motion.' MIT OpenCourseWare - 15.070J / 6.265J Advanced Stochastic Processes, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 2013

[22] H. Chen and D. Yao, Fundamentals of queueing networks: Performance, asymptotics and optimization, Springer-Verlag, 2001.

[23] Yeo, Dominic. “Sticky Brownian Motion.” Eventually Almost Everywhere, 24 Apr. 2014, eventuallyalmosteverywhere.wordpress.com/2014/04/24/sticky-brownian-motion/.

[24] Itô, Kiyosi. Stochastic integral. Proc. Imp. Acad. 20 (1944), no. 8, 519–524. doi:10.3792/pia/1195572786. https://projecteuclid.org/euclid.pja/1195572786

[25] Zhang, Wenyu. Introduction to Ito's Lemma. 6 May 2015, pi.math.cornell.edu/ web6720/Wendy −slides.pdf.

[26] Yeo, Dominic. “Reflected Brownian Motion.” Eventually Almost Everywhere, 21 Apr. 2014, eventuallyalmosteverywhere.wordpress.com/2014/04/21/reflected-brownian-motion/.