Investigating Dependence Between the Arrival Process and Service Time Distribution of Single-Server Queueing Models

Miriam Slattery

May 2, 2022

Thesis submitted for the degree of
Master of Philosophy
in
Applied Mathematics
at The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Mathematical Sciences

THE UNIVERSITY
of ADELAIDE

Contents

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution, and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through the web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: ..... Date: .....

Acknowledgements

First and foremost, I would like to whole-heartedly thank Professor Nigel Bean for his unwavering dedication to support and guide me through every step of this research. Even through uncertain times and unexpected challenges, he has never been too busy to offer me words of wisdom and share his vast knowledge. This thesis would not have been possible without his steadfast encouragement and genuine kindness.

I would also like to acknowledge Professor Gary Glonek. While he may have joined this research journey late, he did not hesitate to fully support me and bring his own expertise and unique perspective. I am grateful for his insightful questions and interesting ideas, which pushed my research and writing to a higher level.

I would like to thank my husband, Josh Bean, who was at first my academic rival and is now my dearest friend and most avid supporter. He has held my hand every step of the way and will never let me stumble and fall. His love and assistance were essential to completing this thesis and will continue to be for all my trials beyond.

Finally, I would like to thank all my other friends and family who have supported me through my academic pursuits. I thank Tobin who made my academic journey more fun and interesting at every step. I thank Diana and Anma for their keen interest, care and delicious food. I thank my parents, Penny and Michael, and my siblings, Ben, Shannon, Grace, Ruth, Declan, Brianna, Jessica and Ronan, for all of their kindness and support through difficult times and joy and fun through happy times. I thank my grandfather, Jim, who is so generous with his time and his home and is always so excited for my achievements.

Abstract

Varney *et al.* studied Australian hospital ICUs using queueing theory. The modelling of the ICU occupancy gave inaccurate predictions when using standard queueing models. This was due to a dependence between the arrival and service processes, discovered through the use of semi-experiments.

In this thesis, we aim to develop methods to detect, classify and quantify dependence between the arrival and service processes in single-server queues. This involves constructing several queueing models with such dependence and then applying a suite of semi-experiments to differentiate between dependence types. To improve efficiency and gain a deeper understanding of how semi-experiments detect dependence, we construct queueing models to represent semi-experiment queues. A majority of all the queueing models are approached using quasi-birth-and-death (QBD) processes and the associated machinery.

We compare how different semi-experiments behave when applied to queues with different types of dependence. Using this we can classify the types of dependence in queueing data in order to propose appropriate models.

Chapter 1

Introduction

Queues have a wide variety of applications to our modern lives from traffic flow, to hospital management, to computer systems. A basic queue consists of 'customers' or 'jobs' arriving, potentially waiting for some time, receiving service from a server, and finally departing from the system. Being able to accurately model these queues is important for improving the performance and predicting requirements in increased infrastructure to accommodate expected changes in demands. In 2017, Varney *et al.* [1] investigated intensive care units (ICUs) in Australian and New Zealand hospitals. These can be modelled as queues since patients arrive to the ICU, they are considered to be receiving 'service' while in the ICU, and are eventually discharged (there is no waiting in this case as all patients are admitted immediately in the ICU by our definition of the queueing system of interest). However, sometimes current ICU patients can be discharged due to demand from more acute arriving patients, which is called 'demand-driven-discharge'. This suggests a correlation structure between the arrival process and the length-of-stay/ service process of the queue. Such a correlation would break the assumption of independence between the arrival process and service process required for the standard modelling of ICUs as queues.

Varney *et al.* [1] encountered this issue when applying the standard method for stochastic ICU modelling based on queueing theory. Hence, they constructed a method of detecting whether such a dependence is present in queueing data like the ICU data. This was using 'semi-experiments', which can be applied to data to break the connection between the arrival and service times. If the queue behaviour (such as length, time customers spend waiting etc.) changes after this connection is broken, this indicates there must have been a dependence between the arrivals and services that affected the queue behaviour. Varney *et al.* [1]

showed that this type of dependence structure existed in many of the Australian ICUs. This was true even when accounting for other possible causes such as differentiating between elective and emergency patients, the weekday and time-of-day of a patient's arrival, the diagnosis type of the patient (for example cardiovascular, respiratory, gastrointestinal, neurological, trauma), and the number of patients in the ICU. Including these in the model reduced the effect of the semi-experiment on the queueing data, indicating that some of the dependence between the arrival process and service process can be explained by these factors, but these were not sufficient to explain all of the dependence. When this dependence in not accounted for in the modelling of a queue, the predictions can be wildly inaccurate, as was observed by Varney *et al.*

The research performed by Varney *et al.* suggests that systems such as ICUs which are commonly modelled by queues can have unexplained dependence structures between the arrival and service processes. This motivates research into queues with such dependence. In this thesis, we seek to understand the nature of different forms this dependence can take and use the methodology of semi-experiments to provide insight in order to detect, quantify and classify the types of dependence between arrival and service processes in queueing data. Thus this understanding and methods for identifying dependence can be applied to propose more appropriate and accurate queueing models for queueing data with dependence than the standard queueing models which assume independence between the arrival and service processes.

Before summarising the details of this thesis, we introduce some important terminology. Our queueing data consists of *inter-arrival times*, which are the times between successive arrivals of customers to the queue, and *service times* which are the times that each customer takes to complete their service once it has begun. The *arrival rates* and *service rates* are the rates at which arrivals and services occur. For example, if the arrival rate is λ, then we expect on average that there will be λ arrivals per unit time. The *stationary queue-length distribution* is the probabilities of the queue having each possible queue length in the long-term.

The thesis is structured as follows. Chapter 2 introduces fundamentals of queueing theory including manipulating data, performance measures and simulation methods. We also introduce basic Markov chains as useful stochastic processes for modelling queues. This leads to more complex Markov chains called quasi-birth-and-death processes (QBDs), which can be used to model a much larger range of complex queues. We introduce theory to be able to evaluate the stationary queue-length distribution, perform simulations and apply truncation and augmentation methods for QBDs. Then, we explain the concepts of semi-experiments

and how they can be applied to queues and the two mains types that will be used: s-perm and a-perm semi-experiments. We also introduce methods for comparing distributions which are needed to draw conclusions from semi-experiments. Finally, we include a review of the relevant literature on queues with dependence.

Chapter **??** provides some useful preliminary features of the queues that we explore. This includes defining a labelling system to easily refer to the large number of related queueing models that will be considered. We also note and explain an interesting feature: the stationary probability that the queue is empty is identical for a queue and the queues formed by semi-experiments applied to it, despite the disruption of the semi-experiments.

Chapter 3 constructs several queues which have simple forms of dependence between the arrival and service processes, including a system where the inter-arrival and service times are correlated pairs of random variables, dependence between the arrival and service rates, customers belonging to classes with different arrival and service rates, and auto-dependence in the arrival and service streams. We simulate each of these models and apply empirical a-perm and s-perm semi-experiments to investigate the effect of dependence on the queue-length distributions. Then we construct queueing models to describe the queues resulting from semi-experiments and to analytically calculate the stationary queue-length distributions to avoid lengthy simulations and permutations.

This approach of constructing queueing models with dependence, applying semi-experiments, and then using queueing models to calculate the stationary queue-length distributions for both the original and semi-experiment queues is extended for more complex models in Chapters 4, 5, 6, and 7. These models provide a better understanding of how the semi-experiments disrupt the dependence and hence a better understanding of the role the dependence plays in the original queue, particularly in the queue-length distribution.

Chapter 4 introduces the first model with queue-length dependence, where the service rate depends on the queue length at the time when service begins. We use QBDs to model the original queue and the s-perm semi-experiment applied to this queue, utilising analysis of the embedded Markov chain at epochs when service begins. Then we compare the stationary queue-length distributions to explore how the semi-experiment can detect and quantify dependence. Finally, we demonstrate that this modelling approach is significantly more efficient for comparing the queue-length distribution than a simulation method, especially for more complex models.

Chapter 5 similarly introduces a queue-length-dependent model, where the arrival rate depends on the queue-length at the start of the inter-arrival period. Again, we use a QBD model to find the stationary queue-length distribution for this queue. We then use the embedded Markov chain at arrival times to construct a QBD model for the s-perm semi-experiment. However, this approach does not successfully capture the behaviour of the empirical s-perm semi-experiments. This is due to a dependence structure within the arrival times, and not just the arrival rates, that is retained after the semi-experiment is applied. We devise a new modelling approach using an 'underlying' queue to recreate the dependence structure in the arrival times. This is incorporated into a more complex QBD which accurately models the s-perm semi-experiment. We then can perform comparisons between the original and s-perm semi-experiment stationary queue-length distributions to investigate the dependence and how semi-experiments detect it.

Chapter 6 considers the two queue-length-dependent models from Chapters 4 and 5 and applies the a-perm semi-experiments to them. We construct similar QBD models to the s-perm semi-experiment models above.

Chapter 7 combines the two models with both arrival and service rates dependent on the queue length. We perform a lengthy analysis of the embedded Markov chains at arrival and start-of-service epochs, which we use in combining the techniques of the previous chapters to construct QBD models for both the a-perm and s-perm semi-experiments.

Chapter 8 introduces restricted semi-experiments. These are s-perm and a-perm semi-experiments whose permutations are restricted to some externally-defined customer classes, the queue length at the starts of service periods, or the queue length at the starts of inter-arrival periods. We introduce algorithms to perform these restricted permutations and then apply them to all the queueing models with dependence we have already introduced, comparing the queue-length distributions of original, standard semi-experiment and restricted semi-experiment queues.

Finally, Chapter 9 summarises the results from the previous chapters by describing how various types of semi-experiments can be used to differentiate between the different types of dependence in the queueing models.

In this thesis, we construct various queueing models with dependence between the arrival and service process, provide models for a variety of semi-experiments applied to these queues, and compare the queueing behaviour. This provides insight into ways to detect and identify types of dependence in queues

and leads towards proposing appropriate models that accurately capture this dependence.

Chapter 2

Background

2.1 Queues

The following introduction to queueing theory closely follows the work of Shortle *et al.* [2].

Individuals in a queue will be referred to as 'customers' or 'jobs'. A customer's path through a basic queue consists of

- Arriving and joining the queue,
- Waiting in the queue until a server is available to provide service (possibly not waiting if a server is already idle),
- Receiving a service for an amount of time from a server,
- Departing from the queue.

Queues and waiting are a consequence of the natural variation inherent in service processes. Investments in infrastructure and staff can reduce waiting times and queue length, but at the cost of reduced average efficiency. This means that it is worth understanding in depth the relevant tradeoffs. To address these tradeoffs, queueing theory often tries to answer questions such as 'How long will a customer wait in the queue?' and 'How many people will be in the queue?' and 'How long are the servers idle?'. Therefore, three performance measures commonly of interest in the system are

- The waiting time of customers (either the time waiting before being served, or the time in the whole system),
- The number of customers at one time (either in the queue before service or in the system),
- The idle time of servers (either individual servers or the whole system).

Since most queues have stochastic elements these quantities are random variables and so this discussion is all based on probability theory.

2.1.1 Characterising Queues

Queues are defined using several informative characteristics, which are often abbreviated using Kendall's notation (see Section 2.1.2). These important characteristics include

- The arrival pattern of customers,
- The service pattern of customers,
- The number of servers,
- The system capacity,
- The queueing discipline.

Arrival Pattern The arrival pattern of customers is usually stochastic and describes a process by which customers arrive. We often assume it is a renewal process and so write about it in terms of the probability distribution of inter-arrival times; the time between successive customers' arrivals. A common arrival pattern is the Poisson process (Section 2.2.3). It is also important to know if jobs can arrive simultaneously in batches, and if so then we need to know the probability distribution for the size of each batch. It is often assumed that the arrival process is a stationary process, independent of time. It may also be possible to have impatient customers who can decide not to enter the queue if it is too long (balk), or to leave while waiting (renege), or to switch between parallel waiting lines (jockey).

Service Pattern The service times of customers are also usually stochastic, requiring a probability distribution to describe the sequence of service times. The system may allow a single job to be worked on by multiple servers at once (such as a computer with parallel processing). The service times may depend on the number of jobs in the system. They may also depend on time, though are often assumed to be stationary.

Number of Servers In the case of multiple servers, there may be cases where distinct queues feed to distinct servers. However, we generally assume a single queue and servers act independently of each other. If there are infinite servers, then every job will begin service upon entry to the system.

System Capacity Some systems may have a physical limitation on how many customers can wait, so that no more customers may enter until space is available but are instead lost to the system.

Queueing Discipline The queueing discipline describes the manner in which customers are selected for service. Most commonly this is first come, first served (FCFS), but there are other possibilities. The system could be last come, first served (LCFS) or random service selection (RSS) in which jobs in the queue are randomly selected for the next service. There may also be processor sharing (PS) in which all jobs are served simultaneously, but individual jobs are processed slower when there are more jobs. Such a discipline can be found in computer systems. The system may use a priority scheme in which there is some preference for particular customers, such as admitting more critical patients first in an emergency department. Such a scheme can be preemptive, in which a lower priority customer's service is interrupted to be resumed later, or nonpreemptive, in which the current service is completed before taking the next highest priority customer.

Note: In this thesis, all queues under consideration have a single server, an infinite system capacity and operate under the FCFS queueing discipline.

2.1.2 Kendall's Notation

A shorthand for describing queueing processes with the characteristics above is a notation developed mostly by Kendall [3]. The notation is as follows

$$A/B/X/Y/Z,$$

where

- A denotes the inter-arrival time distribution,
- B denotes the service time distribution,
- X is the number of parallel servers,
- Y is the system capacity, and
- Z is the queueing discipline.

Some standard symbols are used for each of these characteristics, some of which are given in Table 2.1.

Characteristic	Symbol	Description
A and B	M	Exponential
	D	Deterministic
	E_k	Erlang type k ($k = 1, 2, \dots$)
	H_k	Mixture of k exponentials
	PH	Phase type
	MAP	Markov Arrival Process
	$MMAP$	Marked Markov Arrival Process
	G	General
X	$1, 2, \dots, \infty$	
Y	$1, 2, \dots, \infty$	
Z	FCFS	First come, first served
	LCFS	Last come, first served
	RSS	Random selection for service
	PR	Priority
	GD	General discipline

Table 2.1: Common symbols used for the queueing notation $A/B/X/Y/Z$.

For example, $M/PH/2/\infty/FCFS$ indicates a queueing system with a Poisson arrival process (see Section 2.2.3), phase-type distributed service times, 2 parallel servers, no maximum imposed on the number of customers allowed in the system, and jobs in the system are served in the order in which they arrive. Commonly, the assumption is that there is no restriction on capacity and that the queueing discipline is FCFS, so these descriptions may be omitted so that $M/PH/2$ describes the same queueing system.

While most are self-explanatory, some of the symbols in Table 2.1 need further comment. The symbol M is used for the exponential distribution to avoid confusion with the Erlang distribution by using an E. The M stands for Markovian or memoryless, indicating the process has the memoryless property of the exponential distribution. The symbol G represents a general distribution upon which no distinct assumptions are made. Sometimes GI is used to indicate that the sequence of times from the general distribution are independent of each other.

2.1.3 Queueing Data Representation

In observing queues, we require a method of recording the important and relevant information. Depending on the system, this can be time-oriented recording in which the state of the system is observed at consistent time intervals, or event-oriented recording in which the state of the system is observed after every event, along with the time of the event. In this thesis, we will be using event-oriented data recording.

There are two main forms of representing queueing data that will be used. The first is the time-queue form, which has a vector of time epochs of when the queue length changes, $\boldsymbol{t}$, and a vector of the queue lengths immediately after those epochs, $\boldsymbol{q}$. This form is focussed on the length of the queue. Note that here, and in the rest of the thesis, the queue length is defined to be the number of jobs waiting in the queue *and* in service. The second is the arrival-service form, which has a vector of arrival times for each job in the queue, $\boldsymbol{a}$, and a vector of their corresponding length of service times, $\boldsymbol{s}$. This form is focussed on the individual customers. The time-queue form is necessary for finding the stationary queue-length distribution, while the arrival-service form is essential for performing semi-experiments (see Section 2.3.4). We assume that the service discipline is FCFS and that each job that enters the queue during the time horizon under observation completes its service and departs the queue during the same time horizon. That is, each job has an arrival and departure event in the data. Then, both forms contain the same

information and we can transform between them.

Assume that over the time horizon T there are N events, so the vectors $\boldsymbol{t}$ and $\boldsymbol{q}$ have length N, and assume that there are $M = N/2$ jobs (assuming every arrival also has a departure) so the vectors $\boldsymbol{a}$ and $\boldsymbol{s}$ have length M.

Note that we explicitly show this procedure for single-server and infinite-server queues. The procedure for other finite-server queues is an intuitive combination of the two.

From time-queue form to arrival-service form The following steps are used to obtain the vectors $\boldsymbol{a}$ and $\boldsymbol{s}$ from $\boldsymbol{t}$ and $\boldsymbol{q}$ for a single-server or infinite-server queue.

- If $q_1 > 1$, then the queue is initially longer than 1, and there are arrivals that have not been observed. Append these additional arrivals at the start of the queue, occurring at time 0. That is, $\boldsymbol{t}$ now becomes $[0, 0, \ldots, 0, \boldsymbol{t}]$ and $\boldsymbol{q}$ now becomes $[1, 2, \ldots, q_1 - 1, \boldsymbol{q}]$.
- Let e_i be the event type that occurred at epoch i, for $i = 1, 2, \ldots, N$. Let $e_i = 1$ represent an arrival, and $e_i = -1$ represent a departure. Then $e_1 = 1$ (since we assume the first event must be an arrival) and for $i = 2, \ldots, N$,

$$e_i = \begin{cases} 1 & , \quad \text{if } q_i > q_{i-1}, \\ -1 & , \quad \text{if } q_i < q_{i-1}. \end{cases}$$

  Let $M = \sum_{i=1}^{N} \mathcal{I}\{e_i = 1\}$ be the total number of arrivals to the queue, where $\mathcal{I}$ is an indicator function.

- Let f_i be the customer's identifier of the ith event. If $f_i > 0$, then the ith event is the arrival of the f_ith job. If $f_i < 0$, then the ith event is the departure of the $(-f_i)$th job. These differ for single-server and infinite-server cases.
  - **Single-Server:** For $i = 1, 2, \ldots, N$,

$$f_i = \begin{cases} \sum_{k=1}^{i} \mathcal{I}\{e_k = 1\}, & \text{if } e_i = 1, \\ -\sum_{k=1}^{i} \mathcal{I}\{e_k = -1\}, & \text{if } e_i = -1. \end{cases}$$

– **Infinite-Server:** If the identity of each job was recorded, then the labelling is obvious. If the identity of each job has not been provided, then we cannot get an exact recovery of the system and use the following approximation. For arrivals, $e_i = 1$, then $f_i = \sum_{k=1}^{i} \mathcal{I}\{e_k = 1\}$. For departures, we randomly allocate their identity from those available. If $e_i = -1$, then let X be a vector of f_j $(j \leq i)$ that have appeared exactly once so far (those that have arrived but not departed). Then randomly sample x from X and let $f_i = x$.

- For $j = 1, 2, \ldots, M$, the arrival time of the jth job is

$$a_j = t_i, \qquad \text{where } i \text{ is such that } f_i = j,$$

and the departure time of the jth job is

$$d_j = t_i, \qquad \text{where } i \text{ is such that } -f_i = j.$$

Now we wish to obtain the vector of length of service times, $\boldsymbol{s}$, for each job. In the case of a single-server queue,

- There are two types of service in a single-server queue; the arrival is to either an idle server or to a busy server. Let $b_j = 1$ if the server is busy when the jth arrival occurs, and $b_j = 0$ if the server is idle when the jth arrival occurs. Then, for $j = 1, 2, \ldots, M$,

$$b_j = \begin{cases} 1, & \text{if } q_i \geq 2, \text{where } t_i = a_j, \\ 0, & \text{if } q_i = 1, \text{where } t_i = a_j. \end{cases}$$

- Let s_j be the service time of the jth job. If the server is idle when a customer arrives, then they can begin service immediately, so their departure time is simply their arrival time plus their service time. If the server is busy when a customer arrives, then they must wait until the customer ahead of them has departed before beginning service. Hence, their departure time is the previous customer's departure time plus their service time. That is, for $j = 1, 2, \ldots, M$,

$$s_j = \begin{cases} d_j - a_j, & \text{if } b_j = 0, \\ d_j - d_{j-1}, & \text{if } b_j = 1. \end{cases}$$

In the case of an infinite-server queue, there is always an idle server available when a customer arrives, hence all service times are simply given by

$$s_j = d_j - a_j,$$

for $j = 1, 2, \ldots, M$.

From arrival-service form to time-queue form The following steps are used to obtain the vectors $\boldsymbol{t}$ and $\boldsymbol{q}$ from $\boldsymbol{a}$ and $\boldsymbol{s}$ for a single-server or infinite-server queue.

- Let d_j be the departure time of the jth job. This differs for single-server and infinite-server queues.
  - **Single-Server:** Let $b_j = 1$ if the server is busy when the j arrival occurs, and $b_j = 0$ if the server is idle when the jth arrival occurs. Let $b_1 = 0$ since we assume the first event is an arrival to an empty queue. For $j = 1, 2, \ldots, M$,

    $$d_j = \begin{cases} a_j + s_j, & \text{if } b_j = 0, \\ d_{j-1} + s_j, & \text{if } b_j = 1. \end{cases}$$

    and

    $$b_{j+1} = \begin{cases} 0, & \text{if } a_{j+1} \geq d_j, \\ 1, & \text{if } a_{j+1} < d_j \end{cases}$$

  - **Infinite-Server:** For $j = 1, 2, \ldots, M$,

    $$d_j = a_j + s_j.$$

- Let $\boldsymbol{t} = \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{d} \end{bmatrix}$ and let $\boldsymbol{e} = \begin{bmatrix} \mathbf{1} \\ -\mathbf{1} \end{bmatrix}$ where $\mathbf{1}$ is a vector of ones of length M. Sort the vector $\boldsymbol{t}$ from smallest to largest, and sort the vector $\boldsymbol{e}$ according to the sorting of $\boldsymbol{t}$. Then $\boldsymbol{t}$ is a chronological vector of the event times and $\boldsymbol{e}$ is a vector indicating each of these events as an arrival (1) or departure (-1).
- Let $N = 2M$. Then $q_1 = 1$ and for $i = 2, \ldots, N$,

  $$q_i = q_{i-1} + e_i,$$

  or

  $$q_i = \sum_{k=1}^{i} e_i, \quad \text{for } i = 1, 2, \ldots, N.$$

Example To illustrate the above processes, consider a single-server queue which begins empty at time 0, then has three arrivals (a_1, a_2, a_3) and three departures (d_1, d_2, d_3) over the time horizon T. This queueing data example is displayed in Figure 2.1.

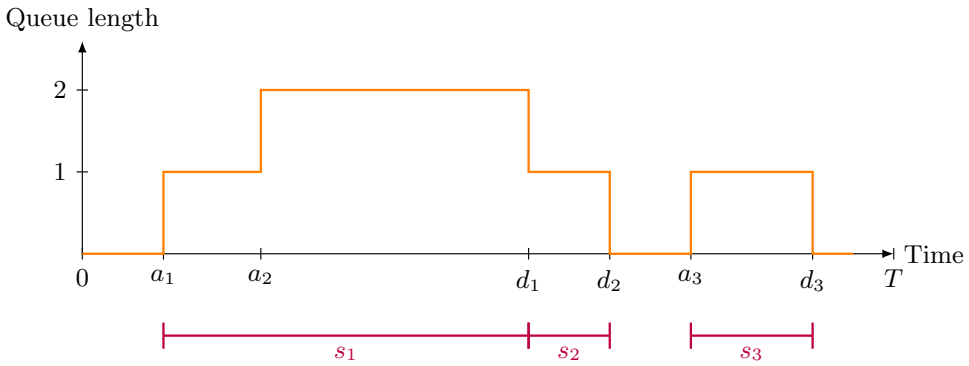


Figure 2.1: An example of queueing data, with 3 jobs arriving an departing over the time horizon, T.

First, suppose we know the vectors $\boldsymbol{t}$ and $\boldsymbol{q}$ for this queue,

$$\begin{aligned}\boldsymbol{t} =&[a_1, a_2, d_1, d_2, a_3, d_3]\\ \boldsymbol{q} =&[1, 2, 1, 0, 1, 0].\end{aligned}$$

So $N = 6$. Note that we can include the initial condition that the queue is empty at time 0 in these vectors, but it is not important for the purposes of this demonstration.

Then we can find the vectors $\boldsymbol{e}$ and $\boldsymbol{f}$,

$$\begin{aligned}\boldsymbol{e} =&[1, 1, -1, -1, 1, -1]\\ \boldsymbol{f} =&[1, 2, -1, -2, 3, -3].\end{aligned}$$

Then we can find the arrival time vector and deaprture time vector,

$$\begin{aligned}\boldsymbol{a} =&[t_1, t_2, t_5] = [a_1, a_2, a_3],\\ \boldsymbol{d} =&[t_3, t_4, t_6] = [d_1, d_2, d_3],\end{aligned}$$

as desired where $M = 3$.

Then we can construct the vector b,

$$\boldsymbol{b} = [0, 1, 0],$$

and the service time vector,

$$\boldsymbol{s} = [d_1 - a_1, d_2 - d_1, d_3 - a_3] = [s_1, s_2, s_3].$$

Now, assume that we know the vectors $\boldsymbol{a}$ and $\boldsymbol{s}$,

$$\begin{aligned}\boldsymbol{a} =&[a_1, a_2, a_3]\\ \boldsymbol{s} =&[s_1, s_2, s_3].\end{aligned}$$

where $M = 3$.

Then we can simultaneously find the vectors $\boldsymbol{b}$ and $\boldsymbol{d}$,

$$\begin{aligned}\boldsymbol{b} =&[0, 1, 0],\\ \boldsymbol{d} =&[a_1 + s_1, a_1 + s_1 + s_2, a_3 + s_3] = [d_1, d_2, d_3], \quad \text{from diagram.}\end{aligned}$$

Then, the sorted vectors $\boldsymbol{t}$ and $\boldsymbol{e}$ will be,

$$\begin{aligned}\boldsymbol{t} =&[a_1, a_2, d_1, d_2, a_3, d_3],\\ \boldsymbol{e} =&[1, 1, -1, -1, 1, -1].\end{aligned}$$

Then $N = 6$ and

$$\boldsymbol{q} = [1, 2, 1, 0, 1, 0].$$

2.1.4 Performance Measures of Queues

The aim in most investigations of queueing systems is to obtain the main performance measures of the system. These can include the number of the customers in the queue (including those in service), the utilisation of the server/s, the waiting time of customers in the queue, the idle time or busy time of a server.

The main focus for this investigation will be on the queue-length distribution. That is, the number of customers in the system at any time.

Let $Q(t)$ be the queue length at time $t \geq 0$. That is,

$$Q(t) = q_i, \quad \text{where } t_i \leq t < t_{i+1}.$$

Let $\{\pi_j : j \geq 0\}$ be the stationary queue length distribution. Then,

$$\pi_j = \lim_{t \to \infty} P(Q(t) = j),$$

for $j = 0, 1, 2, \ldots$.

For a single realisation of a queue over a finite time horizon, π_j is estimated by the proportion of time spent with a queue length of j.

Often, performance measures of queues are of interest when the queue is in a steady-state. That is, some equilibrium in which the average behaviour of the system can be observed. For this to occur we need the queue to be stable so that it does not continuously grow larger and never reach an equilibrium. Intuitively, this occurs when the average arrival rate is less than the average service rate, $\bar{\lambda} < \bar{\mu}$, as on average, customers are being served more quickly than they are arriving and so the queue length will not grow without limit. The more precise requirements are detailed in Sections 2.2.4, 2.2.5 and 2.3, in which queues are considered as Markov chains.

2.1.5 Simulating Queues

We can simulate a queueing system using discrete event simulation (DES). This method keeps track of events that change the state of the system and the times when these events occur. In the case of queues, the usual events that occur are arrivals, which increase the queue length by 1, and departures which decrease the queue length by 1.

The following is for a single-server queue.

Let E be a list of events yet to happen in the queue. The elements of E are ordered pairs (t, e) where t is the time the event occurs and e is the type of event (1 for arrival and -1 for departure). The queue begins with an arrival at time 0, so initialise $E = \{(0, 1)\}$.

For each step,

- Set the current time t and current event e to the first pair on the list E.
- If the current event is an arrival, $e = 1$,
  - Add 1 to the queue length

  - Generate a new arrival. Let $t' = t + a$ where a is sampled from the inter-arrival time distribution. Add a new arrival event $(t', 1)$ to E
  - If the current arrival was to an empty queue, generate a new departure. Let $t' = t + s$ where s is sampled from the service time distribution. Add a departure event $(t', -1)$ to E.

- If the current event is a departure, $e = -1$,
  - Subtract 1 from the queue length
  - If the queue is not empty after the current departure, generate a new departure. Let $t' = t + s$ where s is sampled from the service time distribution. Add a new departure event $(t', -1)$ to E.

- Remove the first pair (t, e) from the list E.

- Sort E into ascending order according to the first index of each pair so events are arranged chronologically.

These steps are repeated until some stopping criteria is met. This is usually when the current time reaches some maximum time limit, or when the number of jobs passed through the queue reaches some desired limit.

The simulation process is quite similar for an infinite-server queue. In this case, each arrival generates the next arrival and also a departure, since in this case, the service for each job begins immediately upon arrival. The departures do not generate any new events.

2.2 Stochastic Processes

This section will also follow the work of Shortle, Thompson, Gross and Harris [2].

This section is an overview of important concepts for stochastic processes that will be used. This includes the exponential distribution, the Poisson process and Markov chains (discrete-time and continuous-time). This is to lead towards developing an understanding of quasi-birth-and-death processes (QBDs) which are a class of continuous-time Markov chains (CTMC) and essential to this investigation of queues with dependence.

2.2.1 Exponential Distributions

The exponential distribution is often used in queueing to model the time until a particular event occurs, such as the time until the next arrival or the time until a service is completed. It is also closely related to the Poisson process which is also widely used in queues and is fundamental to the theory of continuous-time Markov chains.

T is defined to be an exponential random variable if it is a continuous random variable with probability density function (PDF),

$$f(t) = \lambda e^{-\lambda t}, \qquad t \geq 0,$$

where $\lambda > 0$ is a constant.

The cumulative density function (CDF) for an exponential random variable is given by

$$F(t) = P(T \leq t) = 1 - e^{-\lambda t}.$$

The expectation is $\mathbb{E}(T) = \frac{1}{\lambda}$ and the variance $Var(T) = \frac{1}{\lambda^2}$.

The parameter λ represents a rate of events per time. For example, if T is an exponential random variable for the time until the next arrival in a queue, then we expect there to be λ events per unit time, or on average the time until the next arrival is $\frac{1}{\lambda}$.

An important feature of the exponential distribution is the memoryless property,

$$P(T > t + s \mid T > s) = P(T > t), \qquad s, t \geq 0.$$

This can be proven as follows,

$$\begin{aligned} P(T > t + s \mid T > s) =& \frac{P(T > t + s, T > s)}{P(T > s)} \\ =& \frac{P(T > t + s)}{P(T > s)} \\ =& \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \\ =& e^{-\lambda t} \\ =& P(T > t). \end{aligned}$$

Note that the exponential distribution is the only continuous probability density function with the memoryless property.

Another useful feature of the exponential distribution in simulating queues is given below. Consider n independent exponential random variables, $T_1, T_2, \ldots, T_n$ with respective rates $\lambda_1, \lambda_2, \ldots, \lambda_n$. Let $T = \min\{T_1, \ldots, T_n\}$.

Then T is exponentially distributed with rate $\lambda_1 + \cdots + \lambda_n$. Also

$$P(T = T_i) = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_n},$$

where the event $\{T = T_i\}$ is independent of T.

2.2.2 Hyperexponential Distribution

The hyperexponential distribution is a continuous probability distribution and is a mixture of exponential distributions.

Let $Y_1, \ldots, Y_n$ be exponentially distributed random variables with respective rate parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$ and probability density functions $f_{Y_i}(x)$ for $i = 1, 2, \ldots, n$.

Let the random variable X have the following probability density function

$$f_X(x) = \sum_{i=1}^{n} f_{Y_i}(x) p_i,$$

where p_i is the probability that X will take on the form of the exponential distribution with rate λ_i. Then X has a hyperexponential distribution.

The mean of this distribution is

$$\mathbb{E}(X) = \sum_{i=1}^{n} \frac{p_i}{\lambda_i},$$

and the variance is

$$Var(X) = \left(\sum_{i=1}^{n} \frac{p_i}{\lambda_i} \right)^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \left(\frac{1}{\lambda_i} - \frac{1}{\lambda_j} \right)^2.$$

2.2.3 Poisson Distribution and Poisson Process

Poisson Distribution A discrete random variable X is said to have a Poisson distribution with parameter $\lambda > 0$ if the probability mass function of X is given by

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

for $k = 0, 1, 2, \ldots$.

The mean and variance are given by $\mathbb{E}(X) = Var(X) = \lambda$.

Poisson Process The Poisson Process is commonly used for modelling arrivals to queues. Intuitively, it describes events that occur randomly in time.

A counting process is a stochastic process where $N(t)$ takes non-negative integer values and is non-decreasing in time.

A Poisson process with rate $\lambda > 0$ is a counting process, $N(t)$, with the following properties:

- $N(0) = 0$
- The probability of 1 event occurring between times t and $t+\Delta t$ is $\lambda\Delta t+o(\Delta t)$
- The probability of 2 or more events occurring between times t and $t + \Delta t$ is $o(\Delta t)$
- The number of events in non-overlapping time intervals are statistically independent.

The notation $o(\Delta t)$ denotes a quantity that becomes negligible when compared to Δt as $\Delta t \to 0$. That is,

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0.$$

If $N(t)$ is a Poisson process with rate $\lambda > 0$, then the number of events occurring by time t is a Poisson random variable with mean λt. That is,

$$P(N(t) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

Another important feature of the Poisson process is that the times between successive events are independent and exponentially distributed with rate λ.

2.2.4 Discrete-Time Markov Chains

Consider a model in which the system transitions among a discrete set of states through time. Figure 2.2 shows an example of such a system with 4 states. If the system is in state 1 then it can transition to state 0 or state 3, for example. For queues, the states usually represent the number of customers in the system, so they are the non-negative integers.

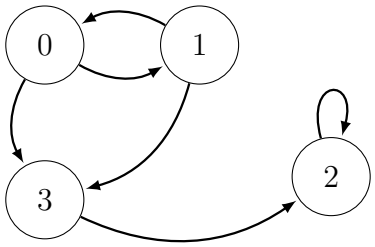


Figure 2.2: Simple example of a Markov Chain.

For a discrete-time Markov chain (DTMC), the transitions occur at discrete time points. Let X_n be the state of the system at time n where $n = 0, 1, 2, \ldots$. X_n takes values from the *state space* S, where S is countable. In Figure 2.2 the state space is given by $S = \{0, 1, 2, 3\}$.

The fundamental assumption of a Markov chain is the Markov property, for all $n \in \mathcal{N}$ and $i_0, \ldots, i_n, j \in S$,

$$P(X_{n+1} = j \mid X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = P(X_{n+1} = j \mid X_n = i_n).$$

That is, if the present state of the system is known, then the next state is independent of all the past states.

The probabilities $P(X_{n+1} = j \mid X_n = i)$ are the single-step transition probabilities. It is often assumed that these are independent of n, in which case the Markov chain is called time-homogeneous and the transition probabilities are written as

$$p_{ij} = P(X_{n+1} = j \mid X_n = i), \qquad i, j \in S.$$

These p_{ij} make up the elements of the transition matrix P. Clearly P is a non-negative matrix and its rows sum to 1 since the transition probabilities out of any state i must sum to 1. That is, for each i,

$$\sum_{j \in S} p_{ij} = 1.$$

Such a matrix is called a *stochastic matrix*.

Chapman-Kolmogorov Equations

The probability of being in state j exactly m steps after being in state i is the m-step transition probability,

$$p_{ij}^{(m)} = P(X_{n+m} = j \mid X_n = i), \qquad \forall i, j \in S, m \geq 0,$$

independent of n.

Let $P^{(m)}$ be the matrix of these probabilities. It can be shown that $P^{(m)} = P^m$. We can use this to find the probability of being in any state j at time m, which we define as $\pi_j^{(m)} = P(X_m = j)$. It follows from the law of total probability that

$$\pi_j^{(m)} = \sum_{i \in S} \pi_i^{(m-1)} p_{ij}, \quad \forall j \in S, m \geq 1,$$

which can be written in matrix form as

$$\boldsymbol{\pi}^{(m)} = \boldsymbol{\pi}^{(m-1)} P, \qquad \forall m \geq 1$$

Applying this rule recursively,

$$\boldsymbol{\pi}^{(m)} = \boldsymbol{\pi}^{(m-1)} P = \boldsymbol{\pi}^{(m-2)} P^2 = \cdots = \boldsymbol{\pi}^{(0)} P^m,$$

where $\boldsymbol{\pi}^{(0)}$ denotes the initial state distribution.

Long-Run Behaviour

We now define several properties

- State j is *accessible* from state i ($i \to j$) if there exists an $n \geq 0$ such that $p_{ij}^{(n)} > 0$. Note that a state j is always accessible from itself through a 0-step transition, $p_{jj}^{(0)} = 1$.

- Two states i and j *communicate* with each other ($i \rightleftarrows j$) if i is accessible from j and j is accessible from i.

- The states of a Markov chain can be partitioned into mutually exclusive subsets called *communicating class*. All states within a class communicate with each other and no states communicate outside the class.

- A Markov chain is *irreducible* if all the states form a single communicating class. Otherwise it is *reducible*.

- The *period* of state j is the greatest common divisor of positive integers m such that $p_{jj}^{(m)} > 0$. A state with period 1 is called *aperiodic*. Periodicity is a class property.

- Let $f_{jj}^{(n)}$ be the probability that the chain starting in state j returns to state j for the first time in n transitions. The probability of ever returning to state j, starting in state j is

$$f_{jj} = \sum_{n=1}^{\infty} f_{jj}^{(n)}.$$

  The state j is *recurrent* if $f_{jj} = 1$ and *transient* if $f_{jj} < 1$.

- If state j is recurrent then

$$m_{jj} = \sum_{n=1}^{\infty} n f_{jj}^{(n)}$$

  is the *mean recurrence time*. If $m_{jj} < \infty$, then state j is *positive recurrent*. If $m_{jj} = \infty$ then j is *null recurrent*.

- It can be shown that positive recurrence, null recurrence and transience are class properties and hold for every state within the same communicating class.

For each $j \in S$, consider

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n)}.$$

If this limit exists, is independent of i, and $\sum_{j\in S} \pi_j = 1$ then $\{\pi_j\}_{j\geq 0}$ is called a *limiting distribution*.

To find π_j we can solve a linear system of equations. If $|S| < \infty$, then for all $j \in S$,

$$\begin{aligned}\pi_j &= \lim_{m\to\infty} p_{ij}^{(m)} \\ &= \lim_{m\to\infty} \left(\sum_{k\in S} p_{ik}^{(m-1)} p_{kj} \right) \\ &= \left(\sum_{k\in S} \lim_{m\to\infty} p_{ik}^{(m-1)} \right) p_{kj} \\ &= \sum_{k\in S} \pi_k p_{kj}.\end{aligned}$$

This involves switching a sum and a limit and needs to be more carefully justified when there is an infinite number of states.

In matrix form this is written as $\boldsymbol{\pi} = \boldsymbol{\pi} P$, which leads to the stationary interpretation of $\boldsymbol{\pi}$. Any solution to this system of equations along with the normalising condition $\sum_{j\in S} \pi_j = 1$ is called a *stationary distribution*. So, if a Markov chain has a limiting distribution, then it also the stationary distribution.

An irreducible and positive recurrent discrete-time Markov chain has a unique solution to the stationary equations

$$\boldsymbol{\pi} = \boldsymbol{\pi} P, \qquad \sum_{j\in S} \pi_j = 1,$$

namely, $\pi_j = 1/m_{jj}$. Furthermore, if the chain is aperiodic, the limiting probability distribution exists and is equal to the stationary distribution.

Ergodicity

One other related concept is the idea of *ergodicity* which relates to whether the 'measures' of a stochastic process, $X(t)$, from a single infinitely long sample path, $X_0(t)$, can be determined or well approximated. We state that $X(t)$ is ergodic if time averages equal ensemble averages. Note that here the discussion is for a continuous time parameter t, but it is analogous in discrete time.

A *time average* is obtained from one sample path of the process. Over

an infinitely long time horizon, the time average is

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T f(X_0(t))dt,$$

for some measurable feature of the sample path, $f(\cdot)$.

An *ensemble average* is obtained from multiple realisations at a fixed point in time. With an infinite number of realisations, X_i, the ensemble average is

$$\mathbb{E}[f(X(t))] = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i(t)).$$

We assume that the ensemble averages converge as $t \to \infty$. Thus a process is ergodic (with respect to the feature f) if

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T f(X_0(t))dt = \lim_{t\to\infty} \mathbb{E}[f(X(t))] < \infty.$$

An irreducible Markov chain has a stationary distribution if and only if it is ergodic.

2.2.5 Continuous-Time Markov Chains

A (time-homogeneous) continuous-time Markov chain (CTMC) is a stochastic process $\{X(t), t \geq 0\}$ with a countable state space, S, such that:

- Each time the process enters state $i \in S$, it remains in that state for an exponentially distributed period of time with rate v_i (independent of the past).
- When the process departs state $i \in S$, it goes to state $j \neq i, j \in S$ with probability p_{ij} (independent of the past).

A CTMC is similar to a DTMC in that it transitions between states with different probabilities, but the time spent in each state is now an exponential

random variable in continuous time. The DTMC defined by the transition matrix of the p_{ij} probabilities is called the *embedded discrete-time Markov chain*.

In continuous time, for $s, t \geq 0$ and $i, j \in S$, the Markov property can be stated as

$$P(X(t+s) = j \mid X(t) = i, X(u), 0 \leq u < t) = P(X(t+s) = j \mid X(t) = i).$$

A CTMC has the Markov property because the remaining time spent in a particular state does not depend on how long the process has already been in that state due to the memoryless property of the exponential distribution, and the next state transition is independent of the past, given the present.

A CTMC can be parameterised by $\{v_i\}$ and $\{p_{ij}\}$. An alternate parameterisation is by $\{q_{ij}\}$ where $q_{ij} = v_i p_{ij}$ for $i \neq j$. The quantity v_i is the transition rate out of state i, so over a long time interval, v_i is approximately the number of transitions out of state i divided by the cumulative time spent in i. So, q_{ij} can be interpreted as the transition rate from state i to j. We can also determine v_i and p_{ij} from q_{ij},

$$v_i = \sum_{j \neq i} q_{ij}, \qquad p_{ij} = \frac{q_{ij}}{\sum_{j \neq i} q_{ij}}.$$

The matrix

$$Q = \begin{pmatrix} -v_0 & q_{01} & q_{02} & q_{03} & \cdots \\ q_{10} & -v_1 & q_{12} & q_{13} & \cdots \\ q_{20} & q_{21} & -v_2 & q_{23} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

is called the infinitesimal generator matrix. By construction, the rows sum to zero with the diagonal elements defined as $q_{ii} = -v_i = -\sum_{j \neq i} q_{ij}$.

Chapman-Kolmogorov Equations

For a DTMC, we can determine the m-step transition probabilities via the Chapman-Kolmogorov equations, $\boldsymbol{\pi}^{(m)} = \boldsymbol{\pi}^{(0)} P^m$. For a continuous-time process, we characterise the probability that the system is in a particular state at time t via a system of differential equations, known as the Kolmogorov differential equations.

Let $p_i(t)$ be the probability that the system is in state i at time t, let $\boldsymbol{p}(t)$ be the vector of these probabilities and let $\boldsymbol{p}'(t)$ be the vector of its derivatives.

Then

$$\boldsymbol{p}'(t) = \boldsymbol{p}(t)Q.$$

We can obtain a direct expression for $\boldsymbol{p}(t)$ by solving the system of differential equations,

$$\boldsymbol{p}(t) = \boldsymbol{p}(0)e^{Qt},$$

where the exponential of a matrix is defined using the Taylor expansion

$$e^{Qt} = \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!},$$

where Q^0 is defined to be the identity matrix.

Long-Run Behaviour

The same concepts for long-run behaviour can be applied to the continuous-time case. The limiting probabilities for a CTMC are defined as

$$\boldsymbol{\pi} = \lim_{t\to\infty} \boldsymbol{p}(t).$$

If the embedded DTMC is irreducible and positive recurrent, then there is a unique solution to the stationary equations,

$$\mathbf{0} = \boldsymbol{\pi}Q, \qquad \sum_{j\in S} \pi_j = 1.$$

Furthermore, if the mean holding times in all states are bounded ($v_i > 0$ for all i), the chain has a limiting probability distribution equal to the stationary distribution.

Embedded DTMCs

The most natural embedded DTMC for a CTMC is one that simply provides the probabilities for the next state the chain will jump to. However, there are other embedded DTMCs which can be relevant. For example, in a queue the system can be observed when a customer enters the queue so that X_1 would be the queue length when the first customer arrives and X_2 would be the queue length when the second customer arrives and so on. This is the DTMC embedded at epochs of arrivals.

Simulating CTMCs

In order to simulate a CTMC, we use a version of the so-called Gillespie algorithm [4].

For each iteration of the algorithm, two random variables need to be generated: the time of the next event, and the state of the process after the next event.

The time until transitioning out of state i is exponentially distributed with rate q_{ii}. Note that at any time the process is in state i the time until leaving the state is still exponentially distributed due to the memoryless property.

Given that the process is leaving state i, the probability that it transitions to state j is given by $\frac{q_{ij}}{q_{ii}}$, independent of the holding time in state i.

So, given the $m \times m$ rate matrix Q (for a finite state space), the initial state of the process, i, and the maximum time for the simulation to run, T, the process can be simulated by the Stochastic Simulation Algorithm 1. At its completion, the output should be the time of every transition and the state immediately after each transition.

Algorithm 1: The Stochastic Simulation Algorithm for simulating a basic CTMC.

Set current time $t := 0$ and initial state $i \in S$
while $t < T$ **do**
  Sample time until next event $t' \sim Exp(q_{ii})$
  Set new time $t := t + t'$
  Create PMF $\boldsymbol{P} = \left[\frac{q_{i1}}{q_{ii}}, \ldots, \frac{q_{i,i-1}}{q_{ii}}, 0, \frac{q_{i,i+1}}{q_{ii}}, \ldots, \frac{q_{im}}{q_{ii}}\right]$
  Sample $j \in \{1, 2, \ldots, i-1, i+1, \ldots, m\}$ from $\boldsymbol{P}$
  Set new state $i := j$
  Store t and i
end while

Birth-and-Death Processes

A birth-and-death process is a CTMC that is often used to describe queueing scenarios. It consists of a set of states $S = \{0, 1, 2, \ldots\}$ denoting the 'population' of the system. State transitions can only jump up or down one state. In state

$n \geq 0$ the time until the next arrival or 'birth' is exponentially distributed with rate λ_n and the state transition is $n \to n+1$. In state $n \geq 1$ the time until the next departure or 'death' is exponentially distributed with rate μ_n and the state transition is $n \to n-1$. Figure 2.3 shows such a rate-transition diagram.

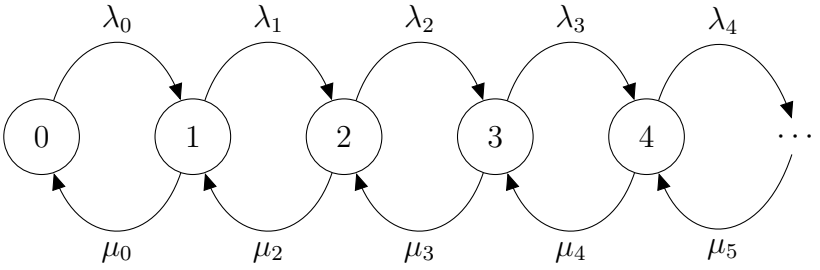


Figure 2.3: Rate-transition diagram for a birth-and-death process

The infinitesimal generator matrix of such a process is

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & \cdots & \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3+\mu_3) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Example: $M/M/1$ **Queue**

Consider a simple $M/M/1$ queue as an example of a CTMC. The M stands for 'Markovian' or 'memoryless'. The arrival process is a Poisson process with arrival rate λ, which means the inter-arrival times are exponentially distributed with rate λ. The service times are exponentially distributed with rate μ. Let $X(t)$ denote the number of customers in the queue at time $t \geq 0$, and the state space of $X(t)$ is $S = \{0, 1, 2, 3, \dots\}$. Then the state $X(t)$ can only increase by 1 with rate λ or decrease by 1 with rate μ. The infinitesimal generator matrix is given by

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ 0 & \mu & -(\lambda+\mu) & \lambda & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}.$$

Then to calculate the stationary queue-length distribution,

$$\boldsymbol{\pi} Q = \mathbf{0}$$

$$(\pi_0, \pi_1, \pi_2, \dots) \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ 0 & \mu & -(\lambda+\mu) & \lambda & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix} = (0, 0, 0, \dots).$$

So for the first column,

$$-\lambda\pi_0 + \mu\pi_1 = 0 \implies \pi_1 = \frac{\lambda}{\mu}\pi_0.$$

For the second column,

$$\lambda\pi_0 - (\lambda+\mu)\pi_1 + \mu\pi_2 = 0 \implies \pi_2 = \frac{\lambda}{\mu}\pi_1,$$

on substituting $\pi_0 = \mu/\lambda\pi_1$.

Continuing, the following pattern emerges for $i \geq 1$

$$\pi_i = \frac{\lambda}{\mu}\pi_{i-1}.$$

Applying this recursively,

$$\pi_i = \frac{\lambda}{\mu}\pi_{i-1} = \left(\frac{\lambda}{\mu}\right)^2 \pi_{i-2} = \cdots = \left(\frac{\lambda}{\mu}\right)^i \pi_0.$$

Now, we need to apply the normalisation,

$$\begin{aligned} \sum_{i=0} \pi_i &= 1 \\ \sum_{i=0} \left(\frac{\lambda}{\mu}\right)^i \pi_0 &= 1 \\ \pi_0 \sum_{i=0} \left(\frac{\lambda}{\mu}\right)^i &= 1 \\ \pi_0 \frac{1}{1-\lambda/\mu} &= 1, \quad \text{if } \lambda < \mu \\ \pi_0 &= 1 - \frac{\lambda}{\mu}. \end{aligned}$$

Therefore, the stationary queue-length distribution for the $M/M/1$ queue with $\lambda < \mu$ is given by

$$\pi_i = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^i, \quad \text{for } i \geq 0.$$

2.3 Quasi-Birth-and-Death Processes

Quasi birth-and-death processes (QBDs) will be key to introducing and exploring dependence between inter-arrival times and service times in a queueing model. They are an extension of the birth-and-death process in Section 2.2.5 which allows the modelling of more complex queues. The following is an introduction to homogeneous QBDs, and follows closely the work of Latouche and Ramaswami [5].

2.3.1 Level-Independent QBDs

Discrete-Time QBD

Consider a discrete-time Markov chain $\{X_t, t \in \mathbb{N}\}$ on the two-dimensional state space $S = \{(n, i) : n \geq 0, 1 \leq i \leq m\}$. The space can be partitioned as $S = \cup_{n\geq 0}\ell(n)$, where $\ell(n) = \{(n, 1), (n, 2), \ldots, (n, m)\}$ is known as *level* n. We refer to the index i as the *phase* within each level. The number of phases in each level, m, may be either finite or infinite.

Such a Markov chain is a Quasi-Birth-and-Death process (QBD) if the one-step transitions are restricted to the states in the same level, the level above, or the level below. That is, a transition from (n, i) to (n', i') is possible only if $n' = n, n + 1$ or $n - 1$; if $n = 0$ then it is possible only if $n' = n$ or $n + 1$.

For a homogeneous QBD, the transition probabilities are assumed to be *level-independent*. That is, for $n, n' \geq 1$, the probability of transition from (n, i) to (n', i') may depend on $i, i', n' - n$, but not on the values of n and n' individually. Therefore, the transition matrix is block-tridiagonal with the form

$$P = \begin{bmatrix} B_0 & A_+ & 0 & 0 & \cdots \\ A_- & A_0 & A_+ & 0 & \cdots \\ 0 & A_- & A_0 & A_+ & \cdots \\ 0 & 0 & A_- & A_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where A_-, A_0, A_+ and B_0 are non-negative square matrices of order m. Since this is the transition matrix of a discrete-time Markov chain, the row sums of P are equal to 1. That is,

$$\begin{aligned} (B_0 + A_+)\mathbf{1} &= \mathbf{1}, \\ (A_- + A_0 + A_+)\mathbf{1} &= \mathbf{1}, \end{aligned}$$

where $\mathbf{1}$ is the $m \times 1$ column vector of ones.

Continuous-Time QBD

A homogeneous continuous-time QBD is a continuous-time Markov chain $\{X(t), t \in \mathbb{R}^+\}$ on the same state space $S = \cup_{n \geq 0} \ell(n)$ with infinitesimal generator of the form

$$Q = \begin{bmatrix} B_0 & A_+ & 0 & 0 & \cdots \\ A_- & A_0 & A_+ & 0 & \cdots \\ 0 & A_- & A_0 & A_+ & \cdots \\ 0 & 0 & A_- & A_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where the matrices A_+ and A_- are non-negative, and the matrices A_0 and B_0 have non-negative off-diagonal elements and strictly negative diagonals. They are all square matrices of order m and the row sums of Q are 0. Hence,

$$\begin{aligned} (B_0 + A_+)\mathbf{1} &= \mathbf{0}, \\ (A_- + A_0 + A_+)\mathbf{1} &= \mathbf{0}, \end{aligned}$$

where $\mathbf{0}$ is the $m \times 1$ column vector of zeros.

In this research we shall focus mostly on continuous-time QBDs.

Important Quantities: R, U, G**.**

First, let us define the matrix N such that N_{ij}, for $1 \leq i, j \leq m$, is the expected total sojourn time in the state (n, j), starting from the state (n, i), before the first visit to any of the states in $\ell(n-1)$. Note that N is independent of $n \geq 1$ since the QBD is assumed to be level-independent.

Let τ be the first epoch of visit to $\ell(n-1)$. That is,

$$\tau = \inf\{t > 0 : X(t) \in \ell(n-1)\}.$$

Formally, we define the matrix N as follows

$$N_{ij} = \int_0^\infty {}_{\{\ell(n-1)\}}P_{ij}(u)du,$$

where ${}_{\{\ell(n-1)\}}P_{ij}(t)$ is the taboo transition function, with $i, j \in \ell(n)$,

$${}_{\{\ell(n-1)\}}P_{ij}(t) = P(X(t) = (n, j), \tau \geq t \mid X(0) = (n, i)).$$

That is, ${}_{\{\ell(n-1)\}}P_{ij}(t)$ is the probability that the process is in state (n, j) at time t, starting in (n, i) before visiting any of the states in $\ell(n-1)$.

Now, we can define

$$R = A_+ N, \tag{2.1}$$

which is such that, for any $n \geq 0$, R_{ij}, for $1 \leq i, j \leq m$, is the expected rate of earning sojourn time in $(n+1, j)$ per unit of local time in $\ell(n)$, given that the process started in (n, i). Here local time is the amount of time spent in a particular level.

Next, assume that the process begins in $X(0) \in \ell(n)$. Then, we define the matrix G such that G_{ij}, for $1 \leq i, j \leq m$, is the probability of first hitting level $n-1$ and doing so in state $(n-1, j)$ in finite time, given that the process started in state (n, i). Formally,

$$G_{ij} = P(\tau < \infty, X(\tau) = (n-1, j) \mid X(0) = (n, i)).$$

We also get

$$G = NA_-.$$

Now, consider the censored Markov process, starting in $X(0) \in \ell(n)$, restricted to $\ell(n)$ until the first visit to $\ell(n-1)$. That is, the censored process is the Markov process with the same rates, but is observed only for states in $\ell(n)$ and the time of the excursions above level n are excised. Then we define the matrix U as the infinitesimal generator of this censored process. We can write U as

$$U = A_0 + A_+ G,$$

and note that $N = (-U)^{-1}$.

The three matrices R, U, G are related by the set of equations

$$\begin{aligned} R &= A_+(-U)^{-1}, \\ G &= (-U)^{-1}A_-, \\ U &= A_0 + RA_- = A_0 + A_+G. \end{aligned}$$

A simple algebraic rearrangement of these equations gives that R, U, G must satisfy the following equations:

$$\begin{aligned} 0 &= A_+ + RA_0 + R^2A_-, \\ U &= A_0 + A_+(-U)^{-1}A_-, \\ 0 &= A_- + A_0G + A_+G^2. \end{aligned}$$

The Stationary Distribution

Assume that the QBD is positive recurrent. Therefore, the process returns to $\ell(0)$ in finite time with probability 1. Let $\boldsymbol{\pi}$ be the stationary probability vector, which we partition by levels of the QBD into subvectors $\boldsymbol{\pi}_n$ $(n \geq 0)$, where $\boldsymbol{\pi}_n = (\pi_{n,1}, \pi_{n,2}, \ldots, \pi_{n,m})$. Note that $\sum_{n\geq 0} \boldsymbol{\pi}_n \mathbf{1} = 1$. Then we can calculate these subvectors using the following equations:

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 R^n \quad \text{for } n \geq 0,$$

where R is the matrix defined in Equation (2.1).

To solve this system of equations, we need to satisfy the following boundary condition. We need to find $\boldsymbol{\pi}_0$, which is the unique solution to

$$\boldsymbol{\pi}_0(B_0 + A_+G) = \mathbf{0},$$

subject to normalisation,

$$\boldsymbol{\pi}_0(I - R)^{-1}\mathbf{1} = 1.$$

Note that the existence of the stationary distribution above is dependent on the QBD being positive recurrent. So we need to determine whether a QBD is positive recurrent.

Consider an irreducible continuous-time QBD, where the number of phases, m, is finite and the matrix $A = A_- + A_0 + A_+$ is irreducible. Then, the QBD is positive recurrent if and only if

$$\boldsymbol{\alpha}A_-\mathbf{1} > \boldsymbol{\alpha}A_+\mathbf{1},$$

where $\boldsymbol{\alpha}$ is the unique solution of the system $\boldsymbol{\alpha}A = \mathbf{0}, \boldsymbol{\alpha}\mathbf{1} = 1$. The QBD is null recurrent if $\boldsymbol{\alpha}A_-\mathbf{1} = \boldsymbol{\alpha}A_+\mathbf{1}$ and transient if $\boldsymbol{\alpha}A_-\mathbf{1} < \boldsymbol{\alpha}A_+\mathbf{1}$.

The vector $\boldsymbol{\alpha}$ is the stationary distribution of being in each phase as the level becomes infinitely larger, $n \to \infty$. That is, ignoring boundary effects at level 0. Then the condition for positive recurrence is that the probability of transitioning to lower levels is greater than transitioning to higher levels on average. That is, there is an average drift to level 0.

In the case that the number of phases is infinite, $m = \infty$, we can find conditions for the QBD to be positive recurrent, though it is more involved (see [5]). However, in this work we need to practically implement these QBDs in code and hence require finite phase spaces. Therefore, those with infinite phase spaces are truncated appropriately (see Section 2.3.3) and hence the QBDs for which the stationary distributions are calculated have finite state spaces.

The Stationary Distribution Algorithms

In order to calculate the stationary distribution for a positive recurrent, irreducible, continuous-time QBD, we need to calculate one of the matrices R, U or G, since any one of these can be used to find the other two. We generally choose to find G, since G is stochastic when the QBD is positive recurrent, so we know that we must have $G\mathbf{1} = \mathbf{1}$. We can then compute R as

$$R = A_+(-A_0 - A_+G)^{-1},$$

and hence compute $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 R^n$, for all n.

We also need to evaluate $\boldsymbol{\pi}_0$ using the boundary condition. This can be found numerically by solving the system $\boldsymbol{v}(B_0 + A_+G) = \mathbf{0}$, normalised by $\boldsymbol{v}\mathbf{1} = 1$ and then $\boldsymbol{\pi}_0 = [\boldsymbol{v}(I - R)^{-1}\mathbf{1}]^{-1}\boldsymbol{v}$.

In order to find G, consider the sequences $\{U(k), k \geq 1\}$ and $\{G(k), k \geq 1\}$ defined as

$$\begin{aligned} U(1) =& A_0, \\ G(k) =& (-U(k))^{-1}A_-, \\ U(k+1) =& A_0 + A_+G(k). \end{aligned}$$

These sequences are monotonically increasing and converge to the matrices U and G, respectively.

Note that if the QBD is recurrent, G is stochastic, and hence $(G - G_\bullet)\mathbf{1} = (\mathbf{1} - G_\bullet\mathbf{1})$. Since $G \geq G(k)$ for all k, a suitable exit condition for a small ϵ is $\|\mathbf{1} - G_\bullet\mathbf{1}\|_\infty \leq \epsilon$. This leads us to state this formally as in Algorithm 2.

Algorithm 2: The Linear Progression Algorithm [5]. The stopping criterion is chosen under the assumption that the QBD is recurrent.

$G_\bullet := (-A_0)^{-1}A_-$
while $\|\mathbf{1} - G_\bullet\mathbf{1}\|_\infty > \epsilon$ **do**
  $U_\bullet := A_0 + A_+G_\bullet$
  $G_\bullet := (-U_\bullet)^{-1}A_-$
end while

Another, more efficient approach is the Logarithmic Reduction Algorithm, shown in Algorithm 3.

Algorithm 3: The Logarithmic Reduction Algorithm [5]. A quadratically convergent algorithm for evaluating G, which is more efficient than the Linear Progresssion Algorithm.

$H^\bullet := (-A_0)^{-1}A_+$
$L^\bullet := (-A_0)^{-1}A_-$
$G_\bullet := L^\bullet$
$T := H^\bullet$
while $\|\mathbf{1} - G_\bullet \mathbf{1}\|_\infty > \epsilon$ **do**
  $U^\bullet := H^\bullet L^\bullet + L^\bullet H^\bullet$
  $H^\bullet := (-U^\bullet)^{-1}(H^\bullet)^2$
  $L^\bullet := (-U^\bullet)^{-1}(L^\bullet)^2$
  $G_\bullet := G_\bullet T L^\bullet$
  $T := TH^\bullet$
end while

Finally, we can numerically solve for the stationary distribution of a level-independent QBD using Algorithm 4.

Algorithm 4: An algorithm to numerically solve for the stationary distribution of a level-independent QBD.

Find G using Algorithm 3
Calculate $R = A_+(-A_0 - A_+G)^{-1}$
Solve the system: $\boldsymbol{v}(B_0 + A_+G) = \mathbf{0}, \quad \boldsymbol{v}\mathbf{1} = 1$
Calculate $\boldsymbol{\pi}_0 = [\boldsymbol{v}(I-R)^{-1}\mathbf{1}]^{-1}\boldsymbol{v}$
For $n \geq 1$, calculate $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 R^n$

Kronecker Product

It is often convenient to use a Kronecker product of matrices to define the model matrices for QBDs. Let A be an $m \times n$ matrix and B be a $p \times q$ matrix. Then the Kronecker product of A and B is given by

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

2.3.2 Level-Dependent QBDs

So far the discussion has concerned level-independent QBDs, whose transition rates do not depend on the current level n. A level-dependent continuous-time QBD (LDQBD), $X(t)$ with state space S, has a generator matrix of the form

$$Q = \begin{pmatrix} A_0^{(0)} & A_+^{(0)} & 0 & 0 & \cdots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & 0 & \cdots \\ 0 & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \cdots \\ 0 & 0 & A_-^{(3)} & A_0^{(3)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The state space is still two dimensional and partitioned into levels, and transitions are only allowed between adjacent levels, but transition rates from the state (n, j) may depend on n. Different levels may even have different numbers of phases, so while blocks on the diagonal are square matrices, the matrices on the secondary diagonals may be rectangular.

Stationary Distribution

When the LDQBD, $X(t)$, is irreducible and positive recurrent it has a stationary distribution. The limiting probability vector $\boldsymbol{\pi}$ satisfies

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1} R^{(n)} = \boldsymbol{\pi}_0 \prod_{m=1}^{n} R^{(m)}, \tag{2.2}$$

for $n \geq 1$, where the matrix $R^{(n)}$ records the expected time spent in the states in the level n between two visits to the level $n-1$, measured in units of local time at level $n-1$.

Note that in the level-independent case $R \equiv R^{(n)}$. There are also analogues to the matrices U and G in the level-dependent case. Define the sequences $\{G^{(n)} : n \geq 1\}$ and $\{U^{(n)} : n \geq 1\}$, where $G^{(n)}$ records the first passage probabilities from $\ell(n)$ to $\ell(n-1)$ and $U^{(n)}$ is the infinitesimal generator for the censored process, at $\ell(n)$, under the taboo of levels $\ell(0), \ldots, \ell(n-1)$.

As in the homogeneous case, these matrices are related by the following

equations

$$\begin{aligned} G^{(n)} =& (-U^{(n)})^{-1} A_-^{(n)}, \\ R^{(n)} =& A_+^{(n-1)} (-U^{(n)})^{-1}, \\ U^{(n)} =& A_0^{(n)} + A_+^{(n)} G^{(n+1)}, \\ U^{(n)} =& A_0^{(n)} + R^{(n+1)} A_-^{(n+1)}. \end{aligned}$$

By simple rearrangement of the equations above, the matrices also satisfy the equations

$$\begin{aligned} \mathbf{0} =& A_-^{(n)} + A_0^{(n)} G^{(n)} + A_+^{(n)} G^{(n+1)} G^{(n)}, \\ \mathbf{0} =& A_+^{(n-1)} + R^{(n)} A_0^{(n)} + R^{(n)} R^{(n+1)} A_-^{(n+1)}, \\ U^{(n)} =& A_0^{(n)} + A_+^{(n)} (-U^{(n+1)})^{-1} A_-^{(n+1)}, \end{aligned}$$

for $n \geq 1$.

For the stationary distribution to exist, $X(t)$ is required to be positive recurrent. This is true if and only if there exists a strictly positive solution of the system

$$\boldsymbol{\pi}_0 = \boldsymbol{\pi}_0 \left(A_0^{(0)} + A_+^{(0)} G^{(1)}) \right), \tag{2.3}$$

normalised by

$$\boldsymbol{\pi}_0 \sum_{n=0}^{\infty} \prod_{k=1}^{n} R^{(k)} \mathbf{1} = 1.$$

If the number of phases at each level is finite, then $X(t)$ is recurrent if and only if $G^{(1)}$ is stochastic.

Note that the system (2.3) is equivalent to

$$\boldsymbol{\pi}_0 = \boldsymbol{\pi}_0 \left(A_0^{(0)} + R^{(1)} A_-^{(1)}) \right).$$

Stationary Distribution Algorithm

Bright and Taylor [6] presented an algorithm for calculating the stationary distribution of LDQBDs, by creating an extension of the logarithmic reduction algorithm developed by Latouche and Ramaswami [5] for level-independent QBDs. They presented a method for evaluating the stationary distribution for a truncated process, where the level does not exceed some given upper limit, K. They also provided a method for selecting an appropriate value of K such that the equilibrium

probability of being in or above level K is approximately 0. Presented below is the basic equations behind the algorithm and the algorithm for when K is known is given in Algorithm 5.

Assume that the phase space of $X(t)$ is finite and that the state space is given by $\{(k,j) : k \geq 0, 1 \leq j \leq M_k\}$. That is the block matrices $A_+^{(k)}, A_0^{(k)}$ and $A_-^{(k)}$ are of order $M_k \times M_{k+1}, M_k \times M_k$ and $M_k \times M_{k-1}$, respectively.

Define $(\boldsymbol{\pi}_k^K)_j$, $0 \leq k \leq K$ to be the stationary probability that the LDQBD, $X(t)$, is in the state (k,j) conditional on being in the set $\{(i,j) : 0 \leq i \leq K, 1 \leq j \leq M_i\}$. From Equation (2.2), it is clear that

$$\boldsymbol{\pi}_k^K = \boldsymbol{\pi}_k^K R^{(k-1)} = \boldsymbol{\pi}_0^K \prod_{m=0}^{k-1} R^{(m)},$$

where $\boldsymbol{\pi}_k^K$ satisfies

$$\boldsymbol{\pi}_0^K = \boldsymbol{\pi}_0^K \left(A_0^{(0)} + R^{(1)} A_-^{(1)}) \right).$$

and

$$\boldsymbol{\pi}_0^K \sum_{k=0}^{K} \prod_{m=0}^{k-1} R^{(k)} \mathbf{1} = 1.$$

Note that $\boldsymbol{\pi}_k^K$is the invariant measure for level k normalised over states in and below K and that

$$\boldsymbol{\pi}_k \leq \boldsymbol{\pi}_k^K,$$
$$\boldsymbol{\pi}_k = \lim_{K \to \infty} \boldsymbol{\pi}_k^K.$$

Algorithm 5 calculates $\boldsymbol{\pi}_k^K, 0 \leq k \leq K$ for a given K.

Algorithm 5: The algorithm to calculate the stationary distribution $\{\boldsymbol{\pi}_k^K, 0 \leq k \leq K\}$ of a LDQBD.

Calculate R^{K-1} (See Algorithm 6).
for $k = K-2$ to $k = 0$ **do**
  Calculate $R^{(k)} = A_+^{(k)} \left(-A_0^{k+1} - R^{(k+1)} A_-(k+2) \right)^{-1}$
end for
Solve $\boldsymbol{\pi}_0^K \left(A_0^{(0)} + R^{(0)} A_-^{(1)} \right) = 0$ such that $\boldsymbol{\pi}_0^K \mathbf{1} = 1$
for $k = 1$ to $k = K$ **do**
  $\boldsymbol{\pi}_k^K = \boldsymbol{\pi}_{k-1}^K R^{(k-1)}$
  Normalise $\boldsymbol{\pi}_0^K, \boldsymbol{\pi}_1^K, \ldots, \boldsymbol{\pi}_k^K$ such that $\sum_{m=0}^{k} \boldsymbol{\pi}_m^K \mathbf{1} = 1$
end for

Note that in Algorithm 6, the UD-pair $UD(\ell, k)$ calculates the quantities U_k^ℓ and $D_{k+2^{\ell+1}}^\ell$.

Algorithm 6: The algorithm to calculate $R^{(K-1)}$.

$k = K - 1$
$\ell = 0$
Compute $UD(0, k)$ and store (See Algorithm 7)
$U = U_k^0$ and $D = D_{k+2}^0$
$\Pi = I$
$R^{(k)}(0) = U$
repeat
  $\Pi = D \times \Pi$
  for $i = 0$ to $i = \ell$ **do**
    for $j = k + (2^{\ell-i+1} - 1)2^i$ to $j = k + 2(2^{\ell-i+1} - 1)2^i$ in steps of size 2^i **do**
      Compute $UD(i, j)$ and store
    end for
  end for
  $\ell = \ell + 1$
  Compute $UD(\ell, k)$
  $U = U_k^\ell$ and $D = D_{k+2^{\ell+1}}^\ell$
  $R^{(k)}(\ell) = R^{(k)}(\ell - 1) + U \times \Pi$
  Remove all UD-pairs from storage except
  $UD(j, k + (2^{\ell-j} - 1)2^{j+1}), j = 0, 1, \ldots, \ell$
until $(R^{(K-1)}(\ell) - R^{(K-1)}(\ell - 1))_{max} < \varepsilon$
$L = \ell$
$R^{(k)} = R^{(k)}(L)$

Algorithm 7: The algorithm to calculate UD-pairs $UD(\ell, k)$. Note that this uses the UD-pairs stored up to the point when this algorithm is called in Algorithm 6.

if $\ell = 0$ **then**
  $U_k^0 = A_+^{(k)} \left(-A_0^{(k+1)}\right)^{-1}$
  $D_{k+2}^0 = A_-^{(k+2)} \left(-A_0^{(k+1)}\right)^{-1}$
else
  $U_k^\ell = U_k^{\ell-1} U_{k+2^{\ell-1}}^{\ell-1} \left(I - U_{k+2^\ell}^{\ell-1} D_{k+3\cdot 2^{\ell-1}}^{\ell-1} - D_{k+2^\ell}^{\ell-1} U_{k+2^{\ell-1}}^{\ell-1}\right)^{-1}$
  $D_{k+2^{\ell+1}}^\ell = D_{k+2^{\ell+1}} D_{k+3\cdot 2^{\ell-1}}^{\ell-1} \left(I - U_{k+2^\ell}^{\ell-1} D_{k+3\cdot 2^{\ell-1}}^{\ell-1} - D_{k+2^\ell}^{\ell-1} U_{k+2^{\ell-1}}^{\ell-1}\right)^{-1}$
end if

2.3.3 Truncating and Augmenting Infinite Blocks

Although the mathematical theory for calculating these stationary distributions for QBDs is proven, we require an algorithmic approach to actually solve for them. These computations require a finite phase space. If the block matrices B_0, A_-, A_0, A_+ have infinitely many rows and columns, then they need to be truncated. However, a simple truncation will result in the rows sums of Q not summing to 0, and hence the QBD will be not be positive recurrent and the stationary distribution will not exist. Therefore, the matrices must also be augmented. There is theory on how to augment the matrices in order to guarantee convergence of the stationary distribution in the limit, as described in [7].

In this thesis, we are more concerned with accurate approximations of the stationary distribution than limiting arguments. Hence, we choose physically meaningful truncations. These truncations are usually of two types. In one, we restrict the number of events between two other events to an upper bound. For example, the number of arrivals between two departures. Since there is a fixed time period for these events to occur in, a large number becomes more rare. The other type is to restrict the queue length of an underlying queue to an upper bound. Since the queues are defined to be positive recurrent, large queue lengths are rare. In both of these cases, at the upper bound the events that would violate this bound are suspended. In the first example, we simply do not allow any more arrivals until after the next departure. The negative sums of rates in the diagonal elements of Q are thus augmented so that the rows sum to 0.

2.3.4 Simulating QBDs

To simulate a QBD, we can use a version of the Stochastic Simulation in Algorithm 1, since it is simply a CTMC with a rate matrix Q.

2.4 Semi-experiments

The main aim of this thesis is to investigate dependence between the arrival process and service times in queues. This is motivated by Varney *et al.* [1] who found a dependence present between the arrival process and length-of-stay times (equivalent to service times in infinite-server queues) in intensive care units (ICUs) in hospitals. This dependence was shown to cause problems when modelling the

ICUs using standard queueing theory which assumes independence between these aspects. This dependence was detected and investigated through the use of *semi-experiments*.

Erramilli [8] first described *blockwise-shuffling* as a method to investigate long-range dependence in packet traffic data. A sequence of inter-arrival times is divided into blocks with m packets. First an 'external shuffle' is performed where the order of the blocks is shuffled and the sequence within each block is preserved. This essentially preserves the short-range dependence and eliminates the long-range dependence. Separately, an 'internal shuffle' is performed where the sequence within each block is randomised but the order of the blocks is unchanged. This removes the short-range dependence but preserves the long-range dependence. The delay times for these shuffled processes are compared with the original to learn about the importance and role of long-range dependence and short-range dependence separately on the process.

The term *semi-experiment* was coined in Hohn [9] as an extension to this block-wise shuffling idea. It describes a method of modifying a single dataset to virtually investigate 'what if' scenarios to determine the cause of statistical properties in the data. Ridoux *et al.* [10] described a typical semi-experiment as replacing a single specific aspect of the data with a neutral model substitute. Then some statistics before and after the manipulation are compared to draw conclusions about the role played by the removed structure.

Varney *et al.* [1] used this idea of semi-experiments to detect dependence between the arrival process and service times in queueing data of ICUs. The goal of these semi-experiments was to break the connection or dependence between the arrival stream and service stream. So, the arrival stream was kept exactly the same while the service times were randomly permuted. In this case, the aspect of the data being replaced is the specific connection between a job's arrival time and service time, replaced by a randomised permutation. The statistics compared are the occupancy (queue-length) distributions for these queues. Therefore, any difference between the occupancy distributions indicated that there must have been a dependence between the arrivals and services. Further, the semi-experiments were performed while taking careful consideration of classes of patients by restricting the permutations. For example, the service times of patients who arrived to the ICU for emergency reasons were permuted amongst like patients, and the same for those who arrived after elective surgeries. These restrictions allow the semi-experiment to detect only the dependence structure between the arrivals and services, not those within the arrival and service streams.

As with Varney *et al.* [1], our goal is to break the dependence between the arrival stream and service times using semi-experiments, though in our case the queues are single-server. In order to gain the most information about the dependence, we consider the types of semi-experiments possible for general queues. The quantities we could consider permuting are

- Arrival times/ inter-arrival (IA) times,
- Departure times/ inter-departure (ID) times,
- Service times,
- Length-of-stay (LOS) times,

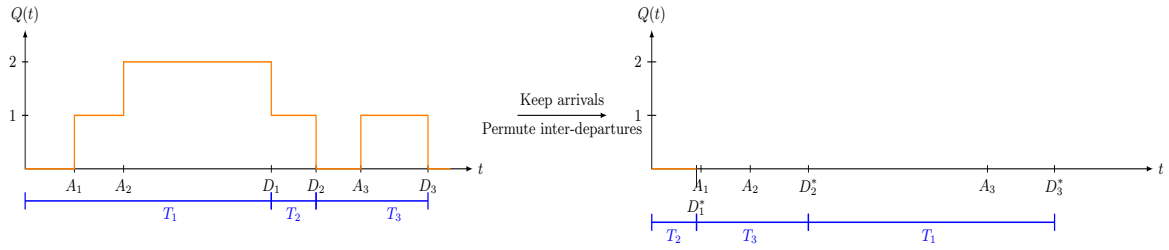
while keeping all other data the same.

Some of these are impossible since they violate features of a FCFS single-server queue. That is, only one customer is served at a time, all customers are served and depart in the order of their arrivals, and a customer's departure occurs after their arrival.

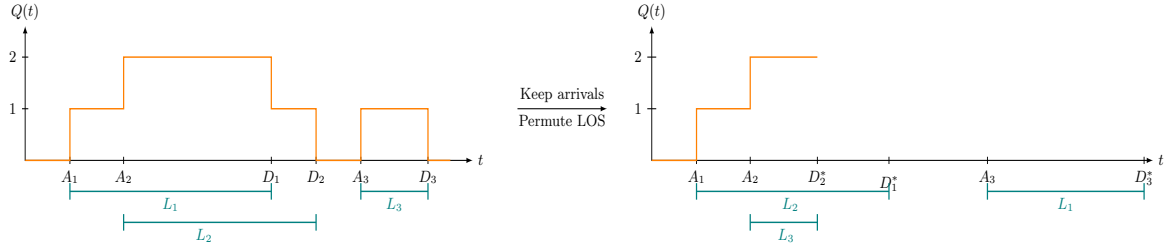
For example, if the arrival stream is maintained, then permuting inter-departure times can result in a customer's departure occurring before their arrival. Permuting length-of-stay times can result in departures occurring out of order. However, permuting service times works well. These examples are shown in Figure 2.4.

After exploration, there are two main styles of semi-experiments that can be used for single-server queues. The permuted services times semi-experiment (s-perm semi-experiment) is one in which the arrival stream in retained and the service times are randomly permuted. The permuted inter-arrival times semi-experiment (a-perm semi-experiment) is one in which the stream of service times is retained and the inter-arrival times are randomly permuted.

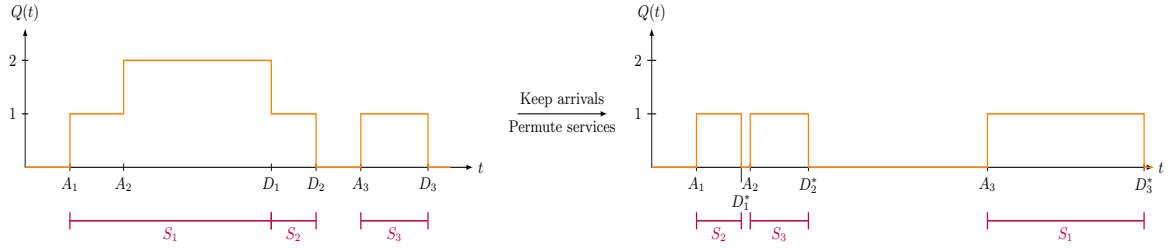
As done by Varney *et al.* [1], the statistics used to compare the data before and after permutation will be the queue-length distributions. When these semi-experiments are applied to queueing data, a significant change in the queue-length distribution indicates dependence between the arrival process and service times. The amount of difference is an indicator of the strength of dependence and also the type of dependence as the s-perm and a-perm semi-experiments will perform differently depending on the form of dependence, which is seen later.



(a) Keeping arrival times, permuting inter-departure times.



(b) Keeping arrival times, permuting length-of-stay times.



(c) Keeping arrival times, permuting service times.

Figure 2.4: Sample path of a queue with examples of different types of semi-experiments.

Restricted Semi-Experiments Varney *et al.* [1] was careful to permute length-of-stay times within intuitive patient classes that could reasonably be assumed to have the same distribution of service times. For example, whether a patient was in the ICU for emergency or elective reasons, the time of day and week a patient arrives, the types of ailments a patient has.

In this thesis, we use restricted semi-experiments to further investigate and diagnose forms of dependence in queues. In this case, we restrict the permutations to within externally defined customer classes (where appropriate), as was done by Varney *et al.* [1]. We also restrict the permutations according to the queue-length distribution at specific epochs. For example, we restrict the permutation of service times to within those with the same queue length at the start of service, or we restrict the permutation of inter-arrival times to those with the same queue length at arrival. See Chapter 8 for further details.

2.4.1 Comparing Distributions

In order to draw conclusions from the semi-experiments, we need to quantify the difference between the original queue-length distribution and the semi-experiment queue-length distribution, for example. It is also important to compare distributions in constructing models and testing whether they accurately model the simulated queues. While we can visually compare histograms, as well as means, standard deviations and other important statistics, one method to test whether two samples are from the same underlying distribution is the two-sample Kolmogorov-Smirnov test.

Two-Sample Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test (KS test) is a non-parametric test of the equality of one-dimensional probability distributions. It can be used to compare a sample with a reference distribution by quantifying the distance between the empirical distribution function of the sample and the cumulative distribution function (CDF) of the reference distribution [11].

The two-sample KS test compares two samples and tests whether they are from the same underlying probability distribution, by quantifying the distance between the empirical distribution functions.

Assume we have two samples of ordered observations $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$, of size n and m, respectively. Let F_1 and F_2 be the underlying CDFs of the first and second samples, and let $F_{1,n}$ and $F_{2,m}$ be the empirical distribution functions of the two samples, defined by

$$F_{1,n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_{[-\infty,x]}(X_i), \quad \text{and} \quad F_{2,m}(x) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{I}_{[-\infty,x]}(Y_i),$$

where $\mathcal{I}_{[-\infty,x]}(X_i)$ is the indicator function which is equal to 1 if $X_i \leq x$ and 0 otherwise.

The null hypothesis for this test is that the two underlying distributions are the same and the alternative hypothesis is that they differ. That is,

$$H_0 : F_1(x) = F_2(x), \quad \forall x \in \mathbb{R}.$$

The test statistic for the two-sample KS test is

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|.$$

To test the null hypothesis, we need to compare the test statistic $D_{n,m}$ against critical values, which are functions of the confidence level α and the sample sizes n and m. For large samples, the null hypothesis is rejected at confidence level α if

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}},$$

where

$$c(\alpha) = \sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)}.$$

One of the strengths of the two-sample KS test is that no assumptions are made about the forms of the underlying distributions F_1 and F_2, and the distribution of the test statistic is independent of these [11]. The test is less powerful than a test that directly compares the means if you just want to compare the means, however it has the advantage that it is sensitive to all departures from $F_1 = F_2$.

Note that this tests whether there is sufficient evidence to reject the hypothesis that the two samples come from same distribution. If the null hypothesis is accepted, it does not specify what the common distribution that the two samples come from is.

In this thesis, we will often be seeking analytic queue-length distributions which can only be evaluated numerically. We want to compare these to empirical queue-length distributions of queue simulations using the KS test. Note that the empirical queue-length is a sequence with autocorrelation and so we need to make sure we use the effective sample size in the calculations since the KS test assumes independent samples. When the KS test is implemented in this thesis, samples of equivalent effective sample size are taken from the queue-length distributions and the null hypothesis is accepted or rejected at the 5% confidence level.

2.5 Literature Review

In classical queueing systems, it is often assumed that the arrival process and service time distribution are independent of each other. However, there have been many queueing systems in which dependence between the arrivals and services of queues has been induced.

Correlated Inter-arrival and Service Times

Queues in which the inter-arrival time and service times are explicitly correlated obviously have such a dependence. Some early work in this area was by Conolly [12] and Conolly and Hadidi [13], in which they investigated a queue where the service time of a customer was proportional to the inter-arrival time between the customer and their predecessor. They found, first numerically and then analytically, that this correlation resulted in a drastic reduction in the mean and variance of the waiting time distribution, when compared to the standard $M/M/1$ queue. This model was then extended by Cidon *et al.* [14] in two ways. First, by adding an independent, generally distributed, non-negative random variable to the service time; and second by allowing the proportionality constant to be itself a random variable. This was in the context of a communication system, in which an example of correlation can arise from packet-switch networks where variable length packets can be sent from one node to another, and the finite speed of links can result in large inter-arrival times for large packets. The final model captures the ON-OFF behaviour of communication links in packet networks, and the Laplace-Stieltjes Transform (LST) and moments of the waiting time distribution were found. Hwang [15] extended the model from Cidon by allowing the proportionality constant to be equal to 1, and a more general ON-OFF process. Cidon [16] also used a similar model, but with subsequent inter-arrival times correlated with the

service times, rather than the prior inter-arrival times. Here, three models were explored. The first and simplest assumed a deterministic proportional dependency with additive inter-arrival random delay. The second considered a random proportional dependency and the third combined both by having a random proportional dependency, and an additive inter-arrival random delay. For each of these, the LST of the waiting time for a customer in the system is found, and numerical results are compared to an equivalent system ($GI/M/1$) without dependence, the latter exhibits much larger means and variances for waiting times.

The following work used various bivariate exponential distributions (BEDs), which are joint distribution of two random variables, with exponential marginal distributions. This requirement on the marginal distributions did not uniquely define the joint distribution, and it has been proven that infinitely many exist [17]. Mitchell and Paulson [18] studied an $M/M/1$ queue with a customer's service time and the inter-arrival time between the customer and their predecessor being correlated random variables with a bivariate exponential distribution. Through simulation, they found that a positive correlation reduces the mean and variance of the total waiting time, and that this reduction was statistically significant. This model was also explored by Conolly and Choo [19], who found a method to exactly calculate the statistics of the waiting time distribution without the use of cumbersome Taylor expansions or LST inversions. Extending this, Langaris [20] found for this model closed-form expressions for the LST of the joint probability density function (PDF) of the server busy period, and also the PDF of the number of customers served in the busy period. Kim and Kim [21] used an almost identical model and found that the waiting time, conditioned on the waiting time being positive, is exponentially distributed. Chao [22] used a similar bivariate exponential (Marshall-Olkin) distribution, but with the service time correlated with the subsequent inter-arrival time. It was found that the waiting time of this model is monotonically decreasing in dependency in increasing convex ordering sense. This result was generalised by Müller [23] for arbitrary joint distributions (as long as other customers are still independent) of service times and subsequent inter-arrival times. Langaris [24] also considered a model with infinitely many servers, with the inter-arrival time and following service time correlated. A bivariate distribution with exponential marginals is used. For this model, Langaris obtained the LST of the steady-state queue-length, as well as a closed-form expression for steady-state queue-length probabilities and moments.

State-Dependence

One way to induce dependence between the arrival process and service times is by having the arrival and/or service rates depend on the queue length process (which depends on both the arrival and service processes). This type of dependence is a large focus of this thesis.

One of the first works to explore such a queue was by Harris [25] in which the service rate was a stochastic process dependent on the queue length at the time when service began. The queueing behaviour was analysed using the embedded Markov chain at epochs of starts of services. Shanthikumar [26] considered an $M/G/1$ queue where the service rate can take two values depending on whether the queue behind is empty or not, and obtained the Laplace transform of the steady-state waiting time distribution. Courtis and Georges [27] considered a single-server queue where the arrival rates depend on the current queue length of the system and the service times depend on the queue length at the epochs when services begin. Gupta and Rao [28] also used this model, constructing recursive equations to calculate the stationary queue-length distribution at arbitrary times, as well as arrival and departure epochs. Abouee-Mehrizi and Baron [29] proposed a queueing system where the arrival rate depends on the current queue length, and the service rates can be modified at both arrival and departure epochs. This queue is analysed using embedded Markov chains and shows that the system is equivalent to a Markovian birth-and-death process.

Other Models

A different model used by Boxma and Perry [30] considers a correlated queue with dependence between service times and subsequent inter-arrival times as a model to explore a fluid production/inventory model with a two-state random environment. In this model, if the service time is less than a threshold, the following interarrival time is exponentially distributed, otherwise it is equal to the threshold. After establishing the link between the two models, Boxma and Perry presented an analysis of the workload and waiting time process of the queueing system.

Borst *et al.* [31] analysed a variant of the $M/G/1$ queue with dependence between service times and previous inter-arrival times. The paper proposed a batch arrival of customers which is considered a 'super-customer' with a longer service time. Hence the number of customers arriving and the number of service requests are positively correlated with the batch arrival times. For both individuals and

batches of customers, Borst *et al.* found the LST of the sojourn time, waiting time and queue length distributions, which are then compared to an $M/G/1$ queue without dependence.

Regterschot and De Smit [32] considered an $M/G/1$ queue with Markov modulated arrivals and service times, where both the arrival rates and service rates depend on an underlying Markov chain. Simiarly, Adan and Kulkani [33] considered a single-server queue in which the inter-arrival times and service times have a bivariate distribution that depends on an underlying Markov chain that jumps at arrival epochs. This creates a $MAP/PH/1$ model. Here, the LST of the waiting time and queue length distributions were derived, and used to obtain recursive equations to calculate the moments.

Iyer and Manjunath [34] explored a generalised $MAP/G/1$ queue where the joint density of inter-arrival and service times are described by a mixture of bivariate densities; in particular mixtures of two exponential densities, and a mixture of exponential and deterministic densities. The analysis for this model is focussed on finding the LST of the waiting time distribution.

Badila *et al.* [35] studied a generalisation of the general single-server queue $G/G/1$ by allowing dependence between service times and subsequent inter-arrival times. The distributions used for analysis are multivariate matrix exponential distributions, which by definition have rational LST. The steady-state waiting time distribution is obtained explicitly, and used to show that the classical relation between waiting time and workload distributions holds true when the independence assumption of the queue is relaxed. This model is directly furthered by Panda *et al.* [36] to include batch arrivals. By assuming rational LSTs, they derived closed-form expressions for the distributions of actual and approximate virtual waiting times for a single-arrival queue, idle periods and inter-departure times. They also derived an approximate queue-length distribution at arbitrary time epochs, and thus stationary distributions for actual and virtual waiting times for first and arbitrary customers in a batch.

Buchholz and Kriege [37] consider a model where the inter-arrival times are correlated, and also service times are correlated with inter-arrival times. It is shown that such a model can be interpreted as an $MMAP[K]/PH[K]/1$ queue. Unlike most other papers, instead of focusing on analysis of waiting times or queue length, the main focus here is to present algorithms to fit parameters relating to the arrival process, the service process, and the cross-correlation, given data containing the trace of corresponding inter-arrival and service times in a process. Two methods are presented; one is based on matching joint moments of the distribu-

tions, the other is an expectation maximisation (EM) algorithm.

Overall, the work done in the area of queues with dependence between the arrival process and service times is largely concerned with proposing a model and then analysing the model using various methods to find queue-length distributions, waiting time distributions, and other similar measures of queueing behaviour. In this thesis, the goal is to explore how different models of such dependence in queues effects the behaviour (specifically the queue-length distribution) and methods we could use to identify appropriate models of dependent queues for queueing data.

2.6 Some Specific Queueing Model Preliminaries

In the following chapters, we construct and explore several queueing models with dependence between the arrival process and service times. These examples allow the exploration of different types of dependence. The three broad types of dependence are pairwise dependence in which the inter-arrival time and service time of each customer are correlated, auto-dependence in which there is dependence within the arrival or service streams, but not between them, and queues in which the arrival and service rates are dependent on the queue length.

We can perform s-perm and a-perm semi-experiments on simulations of these models to show there is indeed dependence. To increase computational efficiency of calculating the queue-length distribution of these semi-experiment queues, we can model them as QBD queueing models. This also provides a greater insight into how the semi-experiments disrupt different types of dependence structures. By utilising these models for different types of semi-experiments on the constructed models, we can find a method to identify and diagnose different types of dependence in queues.

The following notes will be relevant to all of the remaining chapters.

2.6.1 Characterising The Queues In This Thesis

As shown in Section 2.1.1, there are several features generally needed to define a queue, namely

- The arrival pattern of customers,

- The service pattern of customers,
- The number of servers,
- The system capacity,
- The queueing discipline.

For the following sections, assume that the system has an infinite capacity so that no arriving customers are turned away from joining the queue. Also assume that the queueing discipline is a standard first in, first out (FIFO) system in which customers form a single queue and commence service in the order they arrived to the queue, upon completion of which they depart from the system. All queues are single-server queues.

The arrival pattern and service pattern is specified for each individual queue.

2.6.2 Labelling Queues In This Thesis

As there are many similar queues, we introduce a labelling system. Each queue will be labelled as (X, Y, Z) where

- X is the broad dependence type. Such as P for pairwise dependence queues, A for auto-dependence queues, and QL for queue-length-dependent queues.
- Y is the more specific dependence types. Such as A for dependence in arrival process, S for dependence in the service times, or AS for dependence in both. It is also used for types of pairwise dependence such as proportional service times, P, and bivariate exponentially distributed inter-arrival and service times, BED.
- Z is the queue. Such as O for the original queue, sSE for the s-perm semi-experiment queue, and aSE for the a-perm semi-experiment.

See Appendix A for a table of all labels used within this thesis.

2.6.3 Probability of an Empty Queue

It can be seen that for all the models, the queue-length distributions for the original queue, the s-perm semi-experiment, and the a-perm semi-experiment for all the models appear to have the same probability of the stationary queue length being 0. That is, π_0 is the same for a queue and its semi-experiments. This phenomenon is related to a well-known property in queueing theory that holds for $G/G/1$ queues. The following explanation follows the work of [2].

We introduce some notation that is summarised in Table 2.2. Let $\bar{\lambda}$ be the average arrival rate. Let S be the random length of service for a customer, and $\bar{\mu} = 1/\mathbb{E}[S]$ be the average service rate. We denote the traffic intensity of the queue as $\rho = \frac{\bar{\lambda}}{\bar{\mu}}$.

To denote the length of the queue, let N be the random number of customers in the (stationary) queue including those in service, and let N_q be the random number of customers in the (stationary) queue but not yet in service. Let $\pi_n = P(N = n)$, for $n = 0, 1, 2, \ldots$, be the stationary probability distribution of the number of customers in the system. Let W be the random waiting time of a customer before they leave the whole system, and let W_q be the waiting time of the customer in the queue before they begin service.

N, N_q	Number of customers in the system/ queue
W, W_q	Waiting time of a customer in the system/ queue
S	Service time of a customer
$\bar{\lambda}$	Average arrival rate
$\bar{\mu}$	Average service rate
$\rho = \overline{\lambda}/\overline{\mu}$	Traffic intensity
$\pi_n = P(N = n)$	The stationary probability distribution of N

Table 2.2: Summary of notation.

According to Little's Law,

$$\mathbb{E}(N) = \bar{\lambda}\mathbb{E}(W)$$

and

$$\mathbb{E}(N_q) = \bar{\lambda}\mathbb{E}(W_q).$$

It is also clear that the waiting time of a customer in the system is the waiting time of the customer in the queue plus the time they spend in service

before departing. Hence,

$$W = W_q + S.$$

Taking the expectation of this, we obtain

$$\mathbb{E}(W) = \mathbb{E}(W_q) + \frac{1}{\bar{\mu}}.$$

Using these equations,

$$\begin{aligned}\mathbb{E}(N) =& \bar{\lambda}\mathbb{E}(W) \\ =& \bar{\lambda}\left(\mathbb{E}(W_q) + \frac{1}{\bar{\mu}}\right) \\ =& \bar{\lambda}\mathbb{E}(W_q) + \frac{\bar{\lambda}}{\bar{\mu}} \\ =& \mathbb{E}(N_q) + \rho.\end{aligned}$$

We also note that,

$$\mathbb{E}(N) = \sum_{n=0}^{\infty} n\pi_n = \sum_{n=1}^{\infty} n\pi_n,$$

and for a single-server queue,

$$\mathbb{E}(N_q) = \sum_{n=2}^{\infty}(n-1)\pi_n = \sum_{n=1}^{\infty}(n-1)\pi_n.$$

Therefore, substituting these into 2.6.3 gives,

$$\begin{aligned}\rho =& \mathbb{E}(N) - \mathbb{E}(N_q) \\ =& \sum_{n=1}^{\infty} n\pi_n - \sum_{n=1}^{\infty}(n-1)\pi_n \\ =& \sum_{n=1}^{\infty} \pi_n \\ =& 1 - \pi_0.\end{aligned}$$

So we can conclude that for a single-server queue,

$$\pi_0 = 1 - \rho.$$

For each of the our queueing models, the value of π_0 is the same for the original, s-perm semi-experiment, and a-perm semi-experiment since the average arrival rate and average service rate are both retained through the permutation, giving the same value for ρ in each case.

Chapter 3

Simple Dependence Models

In this chapter, we present and investigate some simple forms of dependence between the arrival process and service times in queues. These models show how both s-perm and a-perm semi-experiments can affect queues with such dependence. These simpler models also provide a framework to introduce the methods for developing queueing models for the semi-experiment queues, namely constructing queueing models to analytically find and compare the stationary queue-length distributions of semi-experiment queues. This is expanded on for the queue-length-dependent queueing models in Chapters 4, 5, 6 and 7. In Chapter 9, we will use the simpler models presented here to compare to more complex queue-length-dependent models in order to develop methods to distinguish and classify various types of dependence.

3.1 Pairwise Inter-arrival and Service Time Dependence

In this section we will explore queueing models in which each customer's inter-arrival time and service time are a pair of correlated random variables. Note that the dependence here is between the *times*, and not the *rates*.

Let A_m and S_m be the arrival time and service time of the mth customer, respectively. Let $T_m = A_m - A_{m-1}$ be the inter-arrival period between the mth and $(m-1)$th customers arrivals.

3.1.1 Proportional Service Times

Consider a single server queue where the inter-arrival times are independent and exponentially distributed with $T_m \sim Exp(\lambda)$ and so the arrival process is a Poisson process with arrival rate λ. Then the service time for each customer is proportional to their inter-arrival time, $S_m = \nu T_m$, for some $\nu > 0$. This is the same model as presented by Conolly [12].

This queue is labelled as (P, P, O), where the first P represent a pairwise dependence, the second P represents the proportional service times, and the O represents the original (non-semi-experiment) queue.

Note that for the stationary queue-length distribution to exist for this model, we need the queue to be stable. Hence, the average arrival rate must be less than the average service rate, $\bar{\lambda} < \bar{\mu}$. In this case, the arrival rate is constant so $\bar{\lambda} = \lambda$.

Note that $\mathbb{E}[T_m] = 1/\bar{\lambda}$. Hence,

$$\begin{aligned}\mathbb{E}[S_m] =& \mathbb{E}[\nu T_m] \\ =& \nu \mathbb{E}[T_m] \\ =& \nu/\bar{\lambda}.\end{aligned}$$

Therefore, the average service rate is $\bar{\mu} = \lambda/\nu$.

So, this queue is stable when $\lambda < \lambda/\nu$ or $0 < \nu < 1$.

Empirical Semi-Experiments

Now that we have a queueing model with dependence between the arrival process and service times, we can apply s-perm and a-perm semi-experiments to demonstrate this dependence. To do this, the original queue is simulated by sampling exponentially distributed inter-arrival times, and then calculating the service times by multiplying by ν.

Figure 3.1 shows the queue length distribution for a single simulation of the original queue, a single s-perm semi-experiment and a single a-perm semi-experiment. Using the KS test at the 5% level as described in Section 2.4.1, we conclude that the original queue has a different queue-length distribution to the

semi-experiments. Note that this difference indicates that there is dependence between the arrival process and service times, as expected. Further, a KS test also indicates that the two semi-experiment queue-length distributions are from the same distribution.

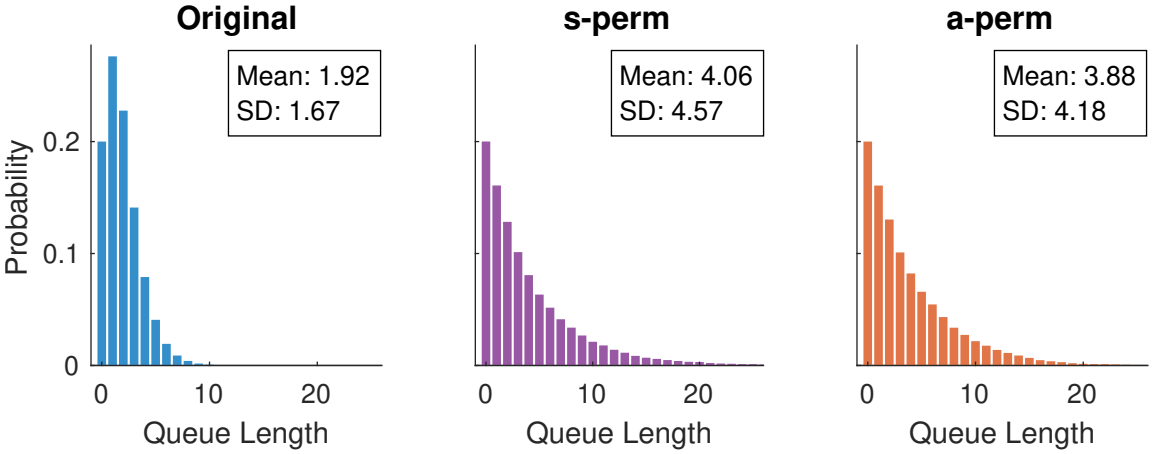


Figure 3.1: The empirical queue-length distributions for a single simulation of the (P, P, O) original process (blue), a single s-perm semi-experiment of the simulation (purple), and a single a-perm semi-experiment of the simulation (orange). Here the parameters are $\lambda = 2$ and $\nu = 0.8$ and the simulation is run for 80,000 customers. Note that all plots have the same y-axis.

Semi-Experiment Model

In order to perform semi-experiments we require data with service times and inter-arrival times for each customer, so that they can be permuted. However, to obtain accurate results for the original queue we need long simulations and for the semi-experiments we need many permutations for each simulation. This is time consuming. Therefore we seek to develop a queueing model to represent the semi-experiment process itself, from which we can analytically evaluate the stationary queue-length distribution of the semi-experiment without the need for simulated data. This modelling also allows us to better understand how the semi-experiments affect the queues.

First consider the s-perm semi-experiment in which the service times are permuted and the arrival process is retained. This permutation breaks the connection between the inter-arrival time and the service time of each customer. There is no additional dependence within the arrival process that is retained since the

inter-arrival times are simply a sequence of independent exponentially distributed times with rate λ. Hence, the arrival process for this semi-experiment is the same as the original; a Poisson process with arrival rate λ. The service times can be modelled as independent and exponentially distributed random variables with rate λ/ν, since they no longer depend on the inter-arrival times.

This proposed semi-experiment model is simply an $M/M/1$ queue with arrival rate λ and service rate λ/ν.

Now consider the a-perm semi-experiment in which the sequence of service times is retained and the inter-arrival times are permuted. This permutation again breaks the connection between the inter-arrival time and service time of each customer in a similar way to the s-perm semi-experiment. This leads to a semi-experiment queueing model with independent exponentially distributed inter-arrival times with rate λ and independent exponentially distributed service times with rate λ/ν. Hence, this is an identical model to the s-perm semi-experiment $M/M/1$ model. This is further supported by Figure 3.1 in which the two empirical semi-experiment queues have the same queue-length distribution.

This s-perm/a-perm semi-experiment queue is labelled as (P, P, SE).

The stationary queue-length distribution for this semi-experiment queue can be analytically evaluated. Let π_i be the stationary probability that the queue length is i in the semi-experiment queue. Then, for $i \geq 0$,

$$\begin{aligned}\pi_i =& \left(1 - \frac{\lambda}{\lambda/\nu}\right)\left(\frac{\lambda}{\lambda/\nu}\right)^i \\ =&(1-\nu)(\nu)^i,\end{aligned}$$

as shown in Section 2.2.5.

Figure 3.2 shows this stationary queue-length distribution compared to the empirical queue-length distribution of the semi-experiments found in Figure 3.1. The KS test concludes that there is no strong evidence of difference between the distributions, indicating this model of the semi-experiment accurately captures the behaviour of the empirical semi-experiments (in regards to the queue-length distribution). Note that we are very cautious in comparing the empirical semi-experiment and model semi-experiment here, because in more complex models later on, it is not so obvious.

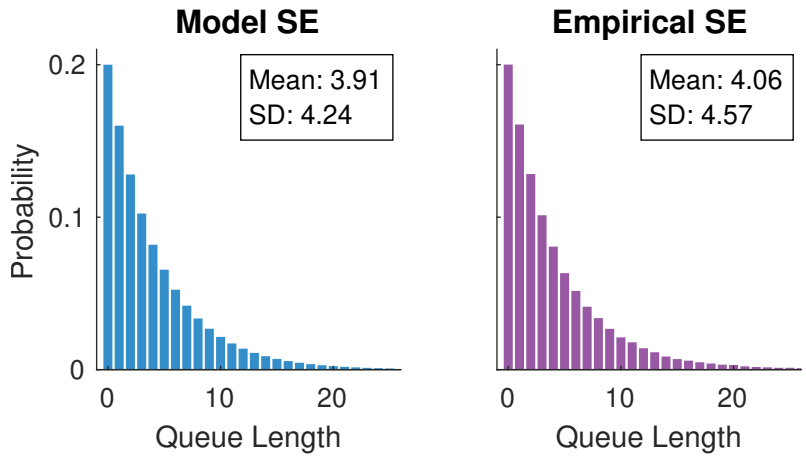


Figure 3.2: The stationary queue-length distribution for the (P, P, SE) model (blue) and the empirical queue-length distribution for an s-perm/a-perm semi-experiment on a single simulation of the original (P, P, O) queue with 80,000 customers. The parameters are $\lambda = 2$ and $\nu = 0.8$.

An important point to note is that the analytic semi-experiment model does not perfectly replicate the empirical semi-experiment. For example, all empirical semi-experiments applied to this model would have the property

$$\sum_{i \geq 1} S_i = \gamma \sum_{i \geq 1} T_i$$

with probability 1. However, since the $M/M/1$ model has independently generated arrival and service streams and would not have this property. There are various other differences that do not materially affect the queue-length distribution. The differences stem from the fact that empirical semi-experiments reuse the same data from a realisation of the original queue, while the model uses distributions to draw data from.

So, this model is only approximate. All the other semi-experiment models in this thesis will similarly be approximate. This (P, P, O) is the most extreme and direct form of dependence that we will consider, and we have shown that this does not have a significant effect on the queue-length distributions, which is the main feature of concern. As we will see, the same is true for all the other models.

3.1.2 BED Inter-arrival and Service Times

As described in Section 2.5, there are many models using bivariate exponential distributions (BEDs) to describe the relationship between inter-arrival times and service times of customers in queues. These models are a more sophisticated version of the (P, P, O) model above. They have correlated pairs of random variables with exponential marginals instead of directly proportional times.

For this model, assume that (T_m, S_m) are pairs of correlated random variables and that they have a Wicksell-Kibble bivariate exponential distribution as described by [38]. That is, the joint probability density function is given by

$$f(t, s) = (1 - \rho)\lambda\mu \exp(-\lambda t - \mu s)\mathcal{I}_0[2(\rho\lambda\mu ts)^{1/2}],$$

where

$$\mathcal{I}_0(x) = \sum_{m=0}^{\infty} \frac{(x/2)^{2m}}{m!m!}$$

is the zero-order modified Bessel function of the first kind and $0 \leq \rho < 1$ is the correlation coefficient between T_m and S_m.

The marginal distributions for T_m and S_m are exponential such that

$$T_m \sim Exp(\lambda(1 - \rho)), \quad \text{and} \quad S_m \sim Exp(\mu(1 - \rho)).$$

This queue is labelled as (P, BED, O).

Empirical Semi-Experiments

Now consider simulating this queue. To do this we need to sample pairs of correlated random variables, (T_m, S_m), from the probability density function $f(t, s)$. This is done using a simple Metropolis-Hastings algorithm, as described below in Algorithm 8.

Then the arrival stream for the simulation is given by

$$\begin{aligned} A_1 &= T_1 \\ A_m &= A_{m-1} + T_m, \quad \text{for } m = 2, 3, \ldots. \end{aligned}$$

We can perform an s-perm semi-experiment which permutes the service times and retains the arrival stream, and an a-perm semi-experiment which per-

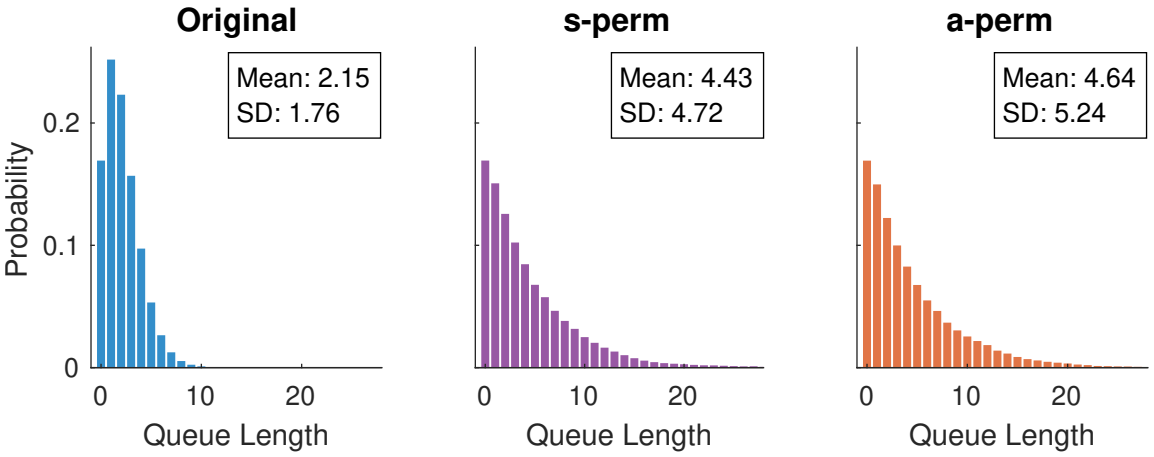


Figure 3.3: The empirical queue-length distributions of a single permutation of the original (P, BED, O) queue, a single s-perm semi-experiment of the simulation, and a single a-perm semi-experiment of the simulation. The parameters are $\lambda = 2.5, \mu = 3$ and $\rho = 0.9$.

mutes inter-arrival times and retains the service stream. The queue-length distributions for these are shown in Figure 3.3. Note that the semi-experiment distributions are different from the original queue-length distribution, demonstrating dependence exists between the arrival stream and service times. Also note that the two semi-experiment distributions appear very similar and we conclude from the KS test that they belong to the same distribution.

Metropolis-Hastings Algorithm

Let $\boldsymbol{x} = (t, s)$, where t is the inter-arrival time before a customer's arrival and s is their service time. The target distribution we wish to sample from is $f(\boldsymbol{x})$. We need a transition kernel, $Q(\boldsymbol{y} \mid \boldsymbol{x})$ to propose a new point $\boldsymbol{y}$ given the current point $\boldsymbol{x}$. For simplicity, we choose a bivariate normal distribution, so that $\boldsymbol{y} \mid \boldsymbol{x} \sim N(\boldsymbol{x}, \boldsymbol{\sigma})$, where $\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$ is chosen to give 'good' samples. Let N be the desired sample size.

Note that since the normal distribution is symmetric, the acceptance probability can be simplified to $A = \min\left(1, \frac{f(\boldsymbol{y})}{f(\boldsymbol{x})}\right)$ since $Q(\boldsymbol{x} \mid \boldsymbol{y}) = Q(\boldsymbol{y} \mid \boldsymbol{x})$. There are methods for choosing appropriate transition kernels and values of $\boldsymbol{\sigma}$. Often there is also a removal of 'burn-in' since it may take a number of iterations for the algorithm to find the accurate samples starting from the initial point ([39]).

Algorithm 8: Basic Metropolis-Hastings Algorithm for sampling from a target distribution.

Initialise $\boldsymbol{x}_1$
for $t = 1, 2, \ldots, N$ **do**
  Sample $\boldsymbol{y}$ from $Q(\boldsymbol{y} \mid \boldsymbol{x})$. $\boldsymbol{y}$ is the proposed value for $\boldsymbol{x}_{t+1}$.
  Compute $A = \min\left(1, \frac{f(\boldsymbol{y})Q(\boldsymbol{x}_t|\boldsymbol{y})}{f(\boldsymbol{x}_t)Q(\boldsymbol{y}|\boldsymbol{x}_t)}\right)$ This is the acceptance probability.
  With probability A accept the proposed value and set $\boldsymbol{x}_{t+1} = \boldsymbol{y}$. Otherwise, set $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t$.
end for

The Metropolis-Hastings algorithm also produces correlated samples, rather than independent bivariate samples. Hence, we use 'thinning', where we discard all but every kth sample to reduce the impact of the correlation.

Semi-Experiment Models

Now consider a model for the s-perm semi-experiment. The permutation of service times breaks the connection of the correlated pairs of random variables. There is no retained dependence in the order of the arrivals. Hence, the interarrival times are all simply independent exponentially distributed random variables with rate $(1 - \rho)\lambda$. Similarly, the permuted service times are independent exponentially distributed random variables with rate $(1 - \rho)\mu$.

Therefore, this semi-experiment queue is simply an $M/M/1$ queue where the Poisson arrival process has rate $(1 - \rho)\lambda$ and the exponential service times have rate $(1 - \rho)\mu$.

Now consider the a-perm semi-experiment queue. The permutation of the inter-arrival times breaks the connection of the correlated pairs of random variables in exactly the same way. Hence, this queue is identical to the s-perm semi-experiment.

This $M/M/1$ queue will be called the semi-experiment queue from here on and be labelled as (P, BED, SE).

Since this is just a simple $M/M/1$ queue, the stationary queue-length distribution for this can easily be evaluated. Let π_i be the stationary probability

that the queue length is i. Then, it can be shown that for $i \geq 0$,

$$\pi_i = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^i$$

3.1.3 Pairwise Arrival and Service Rate Dependence

Unlike the previous models which have correlated inter-arrival and service times, from this point on we will largely focus on dependence through the arrival and service rates. This leads to modelling using Markov chains (and QBDs). This model will demonstrate such a dependence.

Consider K classes of customer, where the class is determined by some external random factor. Then, if the mth customer is of class $k \in \{1, 2, \ldots, K\}$, then the inter-arrival time $T_m \sim Exp(\lambda_k)$ and service time $S_m \sim Exp(\mu_k)$. Assume that the probability of an arriving customer having class k is given by π_k^*, where $\sum_{k=1}^{K} \pi_k^* = 1$.

This queue is labelled as (P, C, O), where P represents the pairwise dependence between inter-arrival and service rates, C represents that the dependence is through customer classes, and O indicates this is the original queue.

Empirical Semi-Experiments

Now, as before, we can perform empirical s-perm and a-perm semi-experiments. To simulate this queue, we generate a sequence of M customers and their classes according to the distribution $\{\pi_k : 1 \leq k \leq K\}$. Then we generate the exponentially distributed inter-arrival times and service times of each customer, according to their class.

Figure 3.4 shows the empirical queue-length distributions for the original queues and the s-perm and a-perm semi-experiments. Using the KS test, we conclude again that the two semi-experiments have the same queue-length distribution, and it is different to the original queue. Note that the difference between the original and semi-experiment queue is less obvious than the previous models. This is because the dependence is through the rate (or the mean) of inter-arrival times and service times, rather than the exact realisations themselves.

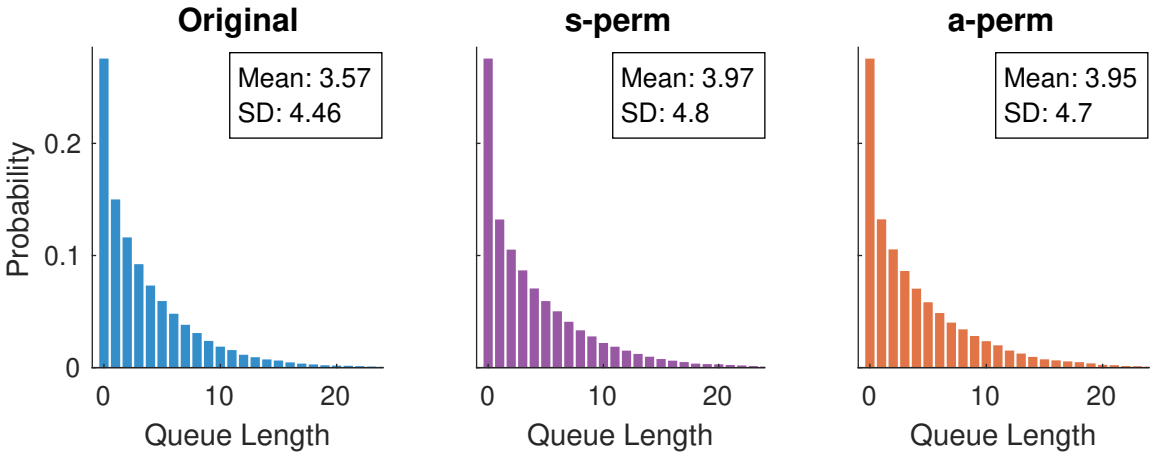


Figure 3.4: The empirical queue-length distributions for a single simulation of the original (P, C, O) queue (blue), a single s-perm semi-experiment of the simulation (purple), and a single a-perm semi-experiment of the simulation (orange). The parameters are $K = 4$, $\boldsymbol{\lambda} = [1, 4, 8, 13], \boldsymbol{\mu} = [1.4, 4.2, 8.1, 13.1]$ and $\boldsymbol{\pi}^* = [0.3, 0.3, 0.2, 0.2]$.

Semi-Experiment Model

As with the previous models, the s-perm and a-perm semi-experiments have the same model here. This is because once a permutation is performed, the arrival stream and service stream become independent of each other since the connection through the particular customer is broken. It only is significant that the connection between the customer's arrival time and service time is severed and not which stream is permuted. Hence, the inter-arrival times are simply a mixture of exponentially distributed times with rates $\lambda_k, 1 \leq k \leq K$, with the proportion according to $\boldsymbol{\pi}^*_{\boldsymbol{k}}$. Similarly, after the permutation, the service times are a mixture of exponentially distributed times with rates $\mu_k, 1 \leq k \leq K$, with the proportion according to $\boldsymbol{\pi}^*_{\boldsymbol{k}}$.

Hence, the inter-arrival times and service times each have a hyperexponential distribution with mixture $\{\pi^*_k : 1 \leq k \leq K\}$ and rates $\{\lambda_k : 1 \leq k \leq K\}$ and $\{\mu_k : 1 \leq k \leq K\}$, respectively.

This semi-experiment model is labelled as (P, C, SE).

This queue can be represented as a continuous-time Markov chain (CTMC), $\{X(t) : t \geq 0\}$. Let $X(t)$ be a triplet (n, i, j). n is the length of the queue at time

t, i is an indicator that the next customer will arrive with rate λ_i, and j is an indicator that the next customer to be served will have a service rate of μ_j. Then, the state space for this process is $S = \{(n, i, j) : n \geq 0, 1 \leq i, j \leq K\}$. The transition rates are given in Table 3.1. An arrival event occurs with rate λ_i and then n increases by 1 and the next arrival rate λ_ℓ is chosen with probability π^*_ℓ. A service is completed with rate μ_j (if the queue is not empty), and then n decreases by 1 and the next service rate is μ_m chosen with probability π^*_m.

Note that n only increases or decreases by 1 in a single transition. Hence, this can be represented as a QBD, where n is the level and (i, j) is the phase. The phase space here has size K^2.

From	To	Rate	For
(n, i, j)	$(n+1, \ell, j)$	$\lambda_i \pi^*_\ell$	$(n, i, j) \in S$
(n, i, j)	$(n-1, i, m)$	$\mu_j \pi^*_m$	$n \geq 1$

Table 3.1: Transition rates for the (P, C, SE) QBD model.

Now we can construct the $K^2 \times K^2$ model matrices A_+, A_0, A_-, B_0 for the QBD model. First we order the phases as follows $(1,1), (1,2), (1,3), \ldots,$ $(1,K), (2,1), \ldots (K,K)$.

A_+ contains the rates in which $n \to n+1$. So,

$$A_+ = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_K \end{bmatrix} \begin{bmatrix} \pi^*_1 & \pi^*_2 & \cdots & \pi^*_K \end{bmatrix} \otimes \mathcal{I}_K.$$

A_- contains the rates in which $n \to n-1$. So,

$$A_- = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{bmatrix} \begin{bmatrix} \pi^*_1 & \pi^*_2 & \cdots & \pi^*_K \end{bmatrix} \otimes \mathcal{I}_K.$$

Now consider the matrix A_0. It is a $K^2 \times K^2$ matrix with diagonal elements equal to the negative sum of all the rates out of each phase and all other

elements are 0. Consider phase (i, j). The sum of all the rates out of this phase is

$$\begin{aligned}&\sum_{k=1}^{K} \lambda_i \pi_k^* + \sum_{k=1}^{K} \mu_j \pi_k^* \\ =&\lambda_i \sum_{k=1}^{K} \pi_k^* + \mu_j \sum_{k=1}^{K} \pi_k^* \\ =&\lambda_i + \mu_j.\end{aligned}$$

Therefore, the diagonal element in A_0 corresponding to phase (i, j) is $-(\lambda_i + \mu_j)$.

Then,

$$A_0 = -\sum_{i=1}^{K} W_0^{(i)} \otimes \left(\lambda_i \mathcal{I}_K + diag(\mu_1, \mu_2, \ldots, \mu_K)\right),$$

where $W_0^{(i)}$ is a $K \times K$ matrix of zeros with a 1 in the (i, i)th position.

Finally, B_0 is similar to A_0 but for when $n = 0$. Hence there are no departures. So the diagonal elements on B_0 are $-\lambda_i$ when the first phase index is i.

Then,

$$B_0 = -\sum_{i=1}^{K} W_0^{(i)} \otimes \lambda_i \mathcal{I}_K.$$

Therefore, the transition matrix for this (P, C, SE) semi-experiment queue is given by

$$Q = \begin{bmatrix} B_0 & A_+ & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ A_- & A_0 & A_+ & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & A_- & A_0 & A_+ & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & A_- & A_0 & A_+ & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Now we have a QBD model for this semi-experiment queue, the stationary queue-length distribution can be easily evaluated using Algorithm 4.

3.2 Auto-Dependence Queues

The next few models here demonstrate an important feature of using semi-experiments to detect dependence between the arrival process and service times. Depending on the permutation type, the semi-experiment can detect dependence that is solely within the arrival stream or service stream of a queue. Such dependence shall be referred to as auto-dependence. This is complementary to the cross-dependence that is strictly between the arrival and service streams seen in the previous few models. It is possible to have both auto-dependence and cross-dependence in the arrival and service streams, as will be explored in Chapters 4, 5, 6 and 7. If it is suspected that the dependence in a queue is solely due to auto-dependence, using time series methods would be appropriate to analyse the queue. We consider them here because it is important to be aware that auto-dependence can be detected by semi-experiments when investigating forms of dependence in a particular queue.

3.2.1 Auto-Dependence in the Arrival Stream

Consider the following queueing model with auto-dependence in the arrival stream. For this example, the arrival rate of a customer depends on the previous customer's arrival rate through a DTMC, $Y_t, t \geq 0$.

Assume that there are K possible arrival rates, λ_k, $1 \leq k \leq K$. The probability of a customer having an arrival rate of λ_k, given the previous customer had an arrival rate of λ_i is given by p_{ik}. The service times are independent and exponentially distributed with constant rate μ. So, $p_{ik} = P(T_m \sim Exp(\lambda_k) \mid T_{m-1} \sim Exp(\lambda_i))$. Let P be the transition matrix containing the probabilities p_{ik}, so that $P = [p_{ik}]$.

This queue is labelled as (A, A, O) where the first A represents an auto-dependence model, and the second A shows that the dependence is within the arrival stream.

This model can be represented as a QBD. Let $X(t)$ be a Markov chain representing this queue. The state for this process is $X(t) = (n, i)$ with state space $S = \{(n, i) : n \geq 0, 1 \leq i \leq K\}$. Let the level, n, be the current length of the queue and let the phase, i, indicate that the next arrival will occur with rate λ_i.

Table 3.2 shows the transition rates for this process.

From	To	Rate	For
(n, i)	$(n+1, j)$	$\lambda_i p_{ij}$	$(n, i) \in S$
(n, i)	$(n-1, i)$	μ	$n \geq 1$

Table 3.2: Transition rates for the (A, A, O) process.

Then the model matrices for this QBD can be constructed as follows,

$$\begin{aligned} A_+ =& diag(\lambda_1, \lambda_2, \ldots, \lambda_K) \times P, \\ A_- =& \mu \mathcal{I}_K, \\ A_0 =& diag(q_1, q_2, \ldots, q_K), \\ B_0 =& - diag(\lambda_1, \lambda_2, \ldots, \lambda_K), \end{aligned}$$

where $q_i = -(\lambda_i + \mu)$.

Empirical Semi-Experiments

Now we can simulate this queue and perform empirical semi-experiments. This queue is simulated as described in Section 2.1.5. Figure 3.5 compares the queue-length distribution for these. Using the KS test, we conclude that the original and s-perm semi-experiment queues have the same queue-length distribution and that the a-perm semi-experiment queue-length distribution is different. This difference is explored below in the construction of the semi-experiment models.

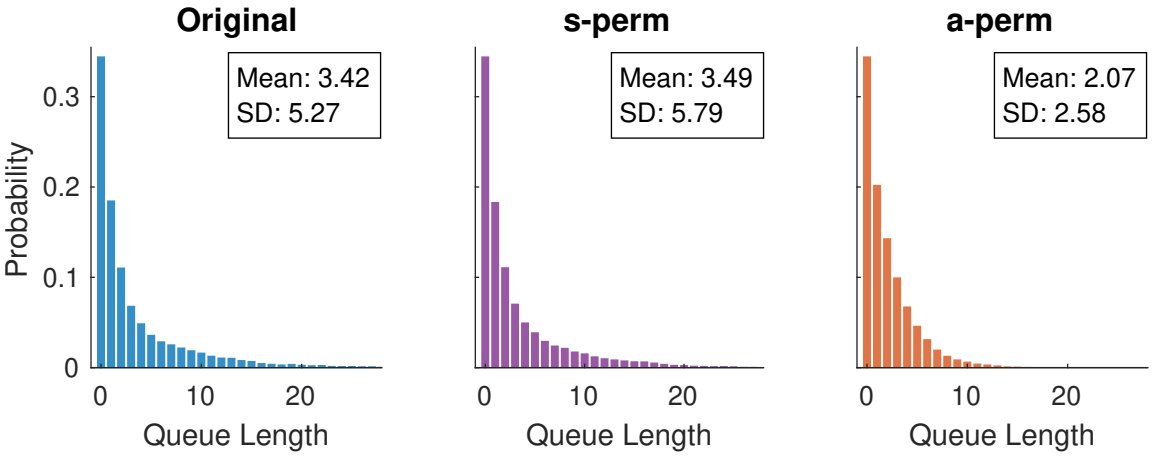


Figure 3.5: The empirical queue-length distributions of a single simulation of the original (A, A, O) queue (blue) run for 20,000 time steps, a single s-perm semi-experiment of the simulation (purple), and a single a-perm semi-experiment of the simulation (orange). The parameters are $K = 2, \mu = 2, \boldsymbol{\lambda} = [0.95, 1.95]$ and $P = [0.98, 0.02; 0.02, 0.98]$.

Semi-Experiment Models

As observed in Figure 3.5, the s-perm and original queues have the same queue-length distribution. This is logical since the auto-dependence in this model is within the arrival stream so permuting service times does not disrupt the dependence. The service times are independent and identically exponentially distributed random variables. When these are permuted they are still independent and exponentially distributed random variables. Hence, this permutation creates a queue which is simply another realisation of the original queue. So a model for the s-perm semi-experiment is simply the original model (A, A, O).

Now consider the a-perm semi-experiment. When the inter-arrival times are permuted the auto-dependence is disrupted. The inter-arrival times for this semi-experiment model will be a mixture of exponential distributions with rates λ_k with $1 \leq k \leq K$. The mixture of this distribution is equivalent to the proportion of each rate λ_k appearing in the original queue.

Let $\boldsymbol{\pi}^*$ be the stationary distribution of Y_t which can be evaluated as shown in Section 2.2.4. It gives the overall probability of observing the arrival rate indicator equal to each $k \in \{1, 2, \ldots, K\}$.

Therefore, the arrival process for this queue is a hyperexponential distribution with rates λ_k, $1 \leq k \leq K$ and mixture $\boldsymbol{\pi}^*$. The service times are independent and exponentially distributed with constant rate μ.

This queue is labelled as (A, A, aSE).

This queue can be represented as a QBD. Let $X(t)$ be the Markov chain for this queue. The state for this process is (n, i) with state space $S_X = \{(n, i) : n \geq 0, 1 \leq i \leq K\}$, where the level, n, is the current queue length and the phase i indicates that the next customer will arrive with rate λ_i.

Table 3.3 shows the transition rates for this process.

From	To	Rate	For
(n, i)	$(n+1, j)$	$\lambda_i \pi_j^*$	$(n, i) \in S_X$
(n, i)	$(n-1, i)$	μ	$n \geq 1$

Table 3.3: Transition rates for the (A, A, aSE) process.

The model matrices for this semi-experiment model are the same as those for the original model, except that

$$A_+ = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_K \end{bmatrix} \begin{bmatrix} \pi_1^* & \pi_2^* & \cdots & \pi_K^* \end{bmatrix},$$

3.2.2 Auto-dependence in the Service Stream

Now consider a model in which the auto-dependence is in the service stream. This model will be very similar to the previous model with auto-dependence in the arrival stream so the explanations here will be brief.

Consider a queue in which the arrival process is a Poisson process with constant rate λ. The service times are exponentially distributed with rates μ_k, $1 \leq k \leq K$. The probability that a service has a rate of μ_k given the previous service had a rate of μ_i is given by p_{ik}. That is, $p_{ik} = P(S_m \sim Exp(\lambda_k) \mid S_{m-1} \sim Exp(\lambda_i))$. Let P be the matrix containing these probabilities.

This queue is labelled as (A, S, O).

This queue can be modelled as a QBD. Let the state be (n, i) where n is the current queue length and i indicates that the next service will have a rate of μ_i. The state space is $S = \{(n, i) : n \geq 0, 1 \leq i \leq K\}$.

Table 3.4 shows the transition rates for this process.

From	To	Rate	For
(n, i)	$(n+1, i)$	λ	$(n, i) \in S$
(n, i)	$(n-1, j)$	$\mu_i p_{ij}$	$n \geq 1$

Table 3.4: Transition rates for the (A, S, O) process.

The model matrices for the process are

$$\begin{aligned} A_+ =& \lambda \mathcal{I}_K, \\ A_- =& diag(\mu_1, \mu_2, \ldots, \mu_K) \times P, \\ A_0 =& diag(q_1, q_2, \ldots, q_K), \\ B_0 =& -A_+, \end{aligned}$$

where $q_i = -(\lambda + \mu_i)$.

Empirical Semi-Experiments

Now we can simulate this queue and perform empirical semi-experiments. This queue is also simulated as described in Section 2.1.5. Figure 3.6 compares the queue-length distribution for these. Using the KS test, we conclude that the original and a-perm semi-experiment queues have the same queue-length distribution and that the s-perm semi-experiment queue-length distribution is different.

Semi-Experiment Models

As expected, Figure 3.6 shows that the a-perm semi-experiment produces the same queue-length distribution as the original queue. This is because permuting the inter-arrival times simply produces another sequence of independent and exponentially distributed times with rate λ and does not disrupt the dependence structure in the service stream. Hence the a-perm semi-experiment is another realisation of the original queue and can be modelled by the original queue.

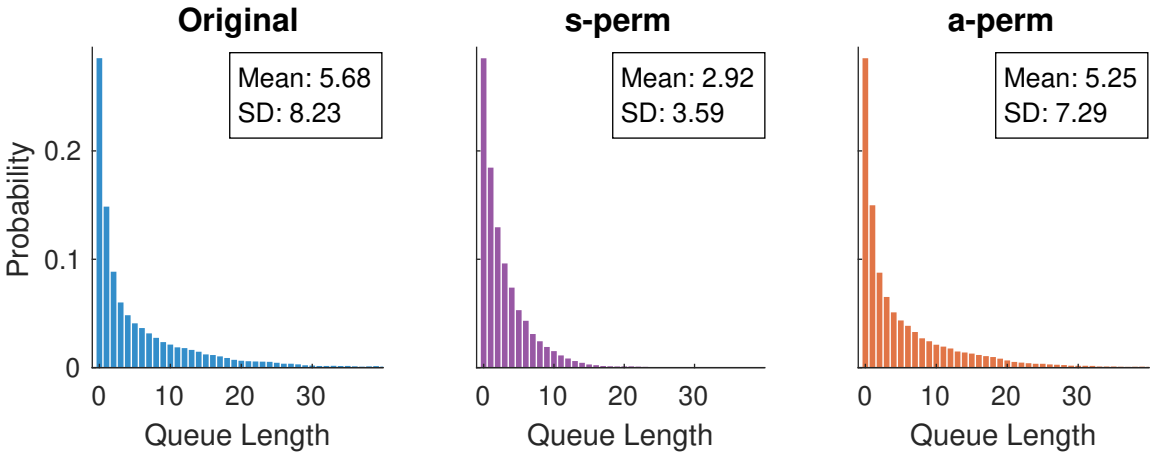


Figure 3.6: The empirical queue-length distributions of a single simulation of the original (A, S, O) queue (blue) run for 20,000 time steps, a single s-perm semi-experiment of the simulation (purple), and a single a-perm semi-experiment of the simulation(orange). The parameters are $K = 2, \lambda = 2, \boldsymbol{\mu} = [2, 5]$ and $P = [0.98, 0.02; 0.02, 0.98]$.

From	To	Rate	For
(n, i)	$(n+1, i)$	λ	$(n, i) \in S$
(n, i)	$(n-1, j)$	$\mu_i \pi_j^*$	$n \geq 1$

Table 3.5: Transition rates for th e$(A, S, 1, sSE)$ process.

Now consider a model for the s-perm semi-experiment. The arrival process for this is a Poisson process with constant arrival rate λ. The service times have a hyperexponential distributions with rates $\mu_k, 1 \leq k \leq K$ and mixture $\{\pi_k^*\}$. $\{\pi_k^*\}$ is the stationary distribution of the DTMC, Y_t, with state space $S = 1, 2, \ldots, K$ and transition probability matrix P.

This queue is labelled as (A, S, sSE) and it can be modelled by a QBD. The state for this QBD is (n, i) with state space $S = \{(n, i) : n \geq 0, 1 \leq i \leq K\}$, where n is the current queue length and i indicates that the next service will have a rate of μ_i.

Table 3.5 shows the transition rates for this process.

The model matrices for this process are the same as the original, except

that

$$A_{-} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{bmatrix} \begin{bmatrix} \pi_1^* & \pi_2^* & \cdots & \pi_K^* \end{bmatrix},$$

3.2.3 Auto-dependence in Arrival and Service Streams

Now we consider a model with auto-dependence in both the arrival and service streams, but with no cross-dependence between the streams. The model we will consider is a combination of the (A, A, O) and (A, S, O) models.

Consider the following queue. The inter-arrival times are exponentially distributed with rates λ_k, $1 \leq k \leq K_1$, chosen by the DTMC on the state space $\{1, 2, \ldots, K_1\}$ with transition probability matrix $P^A = [p_{ij}^A]$. The service times are exponentially distributed with rates $\mu_\ell, 1 \leq \ell \leq K_2$, chosen by the DTMC on state space $\{1, 2, \ldots, K_2\}$ with transition probability matrix $P^S = [p_{ij}^S]$. This queue is labelled as (A, AS, O).

This queue can be modelled as a QBD. Let the state of the process be (n, i, j) where n is the current queue length, i indicates that next arrival will occur with rate λ_i and j indicates that the next service will occur with rate μ_j. So the state space is $S = \{(n, i, j) : n \geq 0, 1 \leq i \leq K_1, 1 \leq j \leq K_2\}$. Here we think of n as the level and (i, j) as the phase, so the phase space has size $K_1 K_2$.

Table 3.6 shows the transition rates for this process.

From	To	Rate	For
(n, i, j)	$(n+1, \ell, j)$	$\lambda_i p_{i\ell}^A$	$(n, i, j) \in S$
(n, i, j)	$(n-1, i, m)$	$\mu_j p_{jm}^S$	$n \geq 1$

Table 3.6: Transition rates for the (A, AS, O) process.

In order to construct the model matrices the phases are ordered as follows $(1, 1), (1, 2), (1, 3), \ldots (1, K_2), (2, 1), \ldots, (K_1, K_2)$.

Let $U_+^{(i,\ell)} = \lambda_i p_{i\ell}^A \mathcal{I}_{K_2}$, for $1 \le i, \ell \le K_1$. Then,

$$A_+ = \begin{bmatrix} U_+^{(1,1)} & U_+^{(1,2)} & \cdots & U_+^{(1,K_1)} \\ U_+^{(2,1)} & U_+^{(2,2)} & \cdots & U_+^{(2,K_1)} \\ \vdots & \vdots & \ddots & \vdots \\ U_+^{(K_1,1)} & U_+^{(K_1,2)} & \cdots & U_+^{(K_1,K_1)} \end{bmatrix} = diag(\lambda_1, \lambda_2, \ldots, \lambda_{K_1}) P^A \otimes \mathcal{I}_{K_2}.$$

Let $U_- = diag(\mu_1, \mu_2, \ldots, \mu_{K_2}) \times P^S$, then

$$A_- = \begin{bmatrix} U_- & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & U_- & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & U_- \end{bmatrix} = U_- \otimes \mathcal{I}_{K_1}.$$

Let

$$U_0^{(i)} = -\left(\lambda_i \mathcal{I}_{K_2} + U_-\right), \quad \textbf{and} \quad V_0^{(i)} = -\lambda_i \mathcal{I}_{K_2}.$$

Then,

$$A_0 = \begin{bmatrix} U_0(1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & U_0(2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & U_0(K_1) \end{bmatrix} = \sum_{i=1}^{K_1} U_0^{(i)} \otimes W_0^{(i)},$$

$$B_0 = \begin{bmatrix} V_0(1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & V_0(2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & V_0(K_1) \end{bmatrix} = \sum_{i=1}^{K_1} V_0^{(i)} \otimes W_0^{(i)}.$$

Empirical Semi-Experiments

Just as before, we perform a simulation of the original (A, AS, O) queue and then apply s-perm and a-perm semi-experiments and compare their empirical queue-length distributions, as shown in Figure 3.7. The KS test indicates that all three distributions are different.

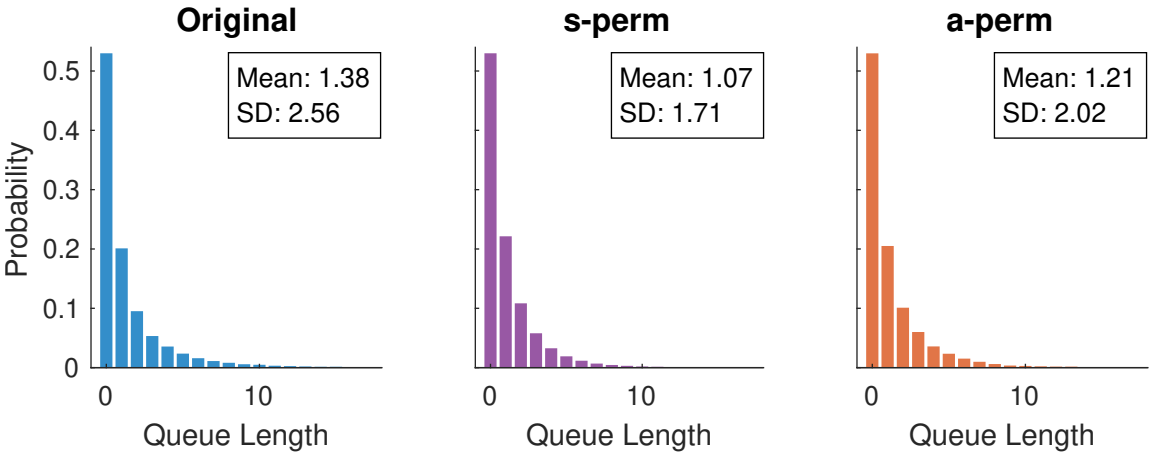


Figure 3.7: The empirical queue-length distributions of a single simulation of the original (A, AS, O) queue (blue) run for 20,000 time steps, a single s-perm semi-experiment of the simulation (purple), and a single a-perm semi-experiment of the simulation(orange). The parameters are $K_1 = 3, K_2 = 2, \boldsymbol{\lambda} = [1, 1.5, 2], \boldsymbol{\mu} = [2, 6]$ and $P^A = [0.95, 0.05, 0; 0, 0.95, 0.05; 0.05, 0, 0.95]$ and $P^S = [0.98, 0.02; 0.02, 0.98]$.

Semi-Experiment Models

The approach to modelling both the a-perm and s-perm semi-experiment queues is very similar to the previous two models.

First, consider the s-perm model. When the service times are permuted the auto-dependence within the service stream is disrupted, and the arrival stream remains untouched. So the model for this queue will have the exact same arrival process as the original model. The service times for this model have a hyperexponential distribution with rates μ_ℓ, $1 \leq \ell \leq K_2$ and mixture $\boldsymbol{\pi}^{*S}$ where $\boldsymbol{\pi}^{*S}$ is the stationary distribution of the DTMC with transition probability matrix P^S. This queue is labelled as (A, AS, sSE) and it can be modelled as a QBD. Let (n, i, j) be the state of this process where n,i,j are all defined the same as in the original model and have the same state space. Table 3.7 shows the transition rates for this model.

From	To	Rate	For
(n, i, j)	$(n+1, \ell, j)$	$\lambda_i p^A_{i\ell}$	$(n, i, j) \in S$
(n, i, j)	$(n-1, i, m)$	$\mu_j \pi^{*S}_m$	$n \geq 1$

Table 3.7: Transition rates for the (A, AS, O) process.

The model matrices are the same as the original, but with

$$U_- = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{K_2} \end{bmatrix} \begin{bmatrix} \pi_1^{*S} & \pi_2^{*S} & \cdots & \pi_{K_2}^{*S} \end{bmatrix}.$$

Now consider the a-perm semi-experiment. This model has the exact same service time distribution as the original queue. The inter-arrival times have a hyperexponential distribution with rates λ_k, $1 \leq k \leq K_2$ and mixture $\boldsymbol{\pi}^{*A}$ where $\boldsymbol{\pi}^{*A}$ is the stationary distribution of the DTMC with transition probability matrix P^A. This queue is labelled as (A, AS, aSE) and it can be modelled as a QBD. Again, let (n, i, j) be the state of this process where n,i,j are all defined the same as in the original model and have the same state space. Table 3.8 shows the transition rates for this model.

From	To	Rate	For
(n, i, j)	$(n+1, \ell, j)$	$\lambda_i \pi_\ell^{*A}$	$(n, i, j) \in S$
(n, i, j)	$(n-1, i, m)$	$\mu_j p_{jm}^S$	$n \geq 1$

Table 3.8: Transition rates for th e(A, AS, aSE) process.

The model matrices are the same as the original, but with the diagonal elements of $U_+^{(i,\ell)} = \lambda_i \pi_\ell^{*A} \mathcal{I}_{K_2}$, for $1 \leq i, \ell \leq K_1$, and

$$A_+ = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{K_1} \end{bmatrix} \begin{bmatrix} \pi_1^A & \pi_2^A & \cdots & \pi_{K_1}^A \end{bmatrix} \otimes \mathcal{I}_{K_2}.$$

3.3 Conclusion

In this chapter we have introduced three pairwise dependence models and three auto-dependence models as examples of simple dependence between the arrival and service processes of queues. We have constructed queueing models for the s-perm and a-perm semi-experiments applied to each of these dependent queues to find the stationary queue-length distributions and compared them to the original queue-length distributions. Through this we have found that the a-perm and s-perm semi-experiments have the same effect on the pairwise dependence queues. Further,

we have shown that the effect of dependence is stronger when the dependence is through the inter-arrival and service times, compared to dependence through arrival and service rates. In the auto-dependence queues, only the semi-experiment that permutes the stream with dependence has a significant effect on the queue-length distribution. This suggests that semi-experiments can be used to detect different types of dependence in queueing data. We return to this topic in Chapters 8 and 9 after investigating further forms of dependence in the intervening chapters.

Structure of Results Note that for each section in this chapter the following formula was followed:

- Defining a queueing model with some form of dependence between the arrival and service streams,
- Performing s-perm and a-perm semi-experiments and comparing the queue-length distributions to that of the original queue, and
- Constructing queueing models for the semi-experiment processes in order to both more fully understand the empirical results, and also analytically/numerically evaulate the queue-length distribution of the semi-experiments to remove the need for computationally inefficient simulations and permutations.

This structure allows the reader to first understand the original queue and then immediately see the results of empirical semi-experiments before more deeply exploring the semi-experiments through models. It also keeps a focus on the importance of empirical semi-experiments, which are necessary for queueing data with an unknown underlying model. However, in subsequent chapters we will simplfy this formula for efficiency by performing empirical semi-experiments after the model construction.

Chapter 4

Queue-Length-Dependent Service Rates

4.1 Original Model

Now we consider a queue-length-dependent queue, which induces dependence between the arrival process and the service times through a dependence of the service times on the queue length. It is a single-server queue with service rates dependent on the length of the queue *at the time when the service period begins*. We can formally specify the arrival and service patterns as follows, where the

- **Arrival Process** is a Poisson process with constant rate λ. That is, the inter-arrival times are independent and identically distributed from the exponential distribution with rate λ.

- **Service Times** are exponentially distributed with rate μ_i, where $i \geq 1$ is the length of the queue *at the time when the service begins.*

This process is labelled as the (QL, S, O) queue. That is, the original queue with queue-length-dependent service rates and a single server.

Note that this queue is not a general birth-and-death process, since the service rate depends on a historical state of the system, depending on when the service started. Hence, the following characterisation of the process is necessary.

This process can be represented as a continuous-time Markov chain (CTMC) $\{X(t), t \geq 0\}$ on the two-dimensional state space $S = \{(n, i) : n \geq 1, 1 \leq i \leq n\} \cup \{(0, 1)\}$. The first coordinate, or level, n represents the current length of the queue and the second coordinate, or phase, i represents the queue length at the time when the current service began. That is, the phase is used to inform the service rate. The level, n, takes values from the non-negative integers as it is counting people in the queue. The phase, i, is greater than or equal to 1, since there must be at least one person in the queue for a service to begin. Further, $i \leq n$ as during a service period the only events are arrivals, so the current queue length n must be greater than or equal to the queue length at the start of the service, i. Note that when $n = 0$, we have set $i = 1$. While this does not fit neatly into the interpretation since there is no ongoing service in an empty queue, it makes the notation simpler. The choice of $i = 1$ in this case is sensible since if the queue is empty, the next service will begin as soon as an arrival occurs, starting the service with a queue length of 1. We can also write $S = \bigcup_{n \geq 0} \ell(n)$ where $\ell(n) = \{(n, 1), \ldots (n, n)\}$ is the nth level for $n \geq 1$ and $\ell(0) = \{(0, 1)\}$.

The possible transitions for this process are depicted in the state transition diagram in Figure 4.1 and detailed in Table 4.1. For an arrival event, the current queue length increases by 1 with rate λ, and the phase does not change as an arrival does not affect the current service period. For a departure event leaving a non-empty queue, the queue length decreases by 1 from (n, i) with rate μ_i to $(n - 1, n - 1)$. The phase changes to $n - 1$ since when there is a departure a new service immediately begins with the new current queue length of $n - 1$.

From	To	Rate	For
(n, i)	$(n + 1, i)$	λ	$n \geq 1, 1 \leq i \leq n$
(n, i)	$(n - 1, n - 1)$	μ_i	$n \geq 2, 1 \leq i \leq n$
(0,1)	(1,1)	λ	
(1,1)	(0,1)	μ_1	

Table 4.1: Possible transitions for the (QL, S, O) process.

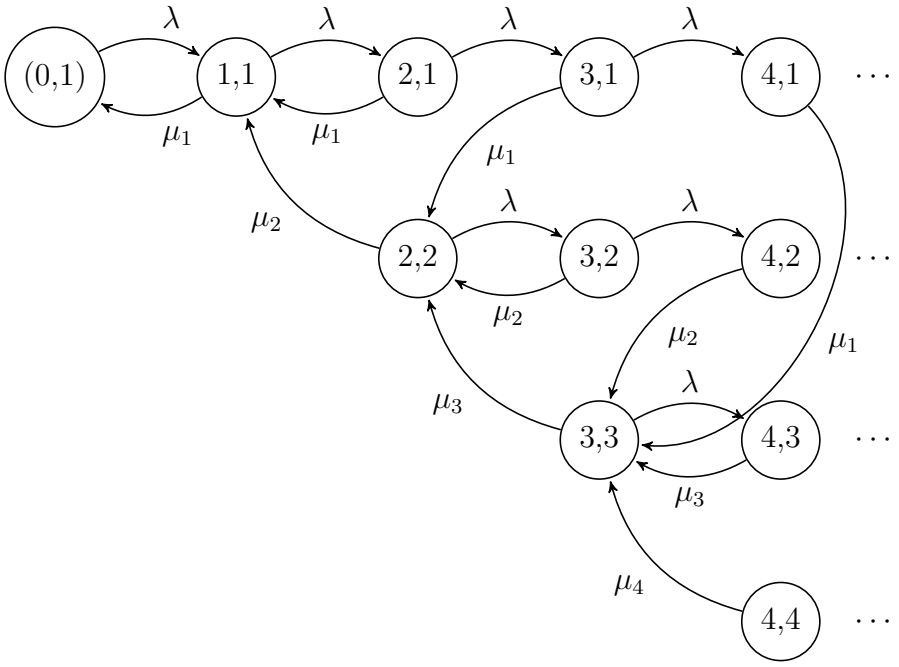


Figure 4.1: State transition diagram for the (QL, S, O) process. The arrows represent possible transitions and are labelled with the corresponding instantaneous rates.

The generator matrix for the process is then

$$Q = \begin{array}{c|cc|cc|ccc|cccc|c} & (0,1) & (1,1) & (2,1) & (2,2) & (3,1) & (3,2) & (3,3) & (4,1) & (4,2) & (4,3) & (4,4) & \cdots \\ \hline (0,1) & -\lambda & \lambda & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\ (1,1) & \mu_1 & q_1 & \lambda & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\ \hline (2,1) & \cdot & \mu_1 & q_1 & \cdot & \lambda & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\ (2,2) & \cdot & \mu_2 & \cdot & q_2 & \cdot & \lambda & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\ \hline (3,1) & \cdot & \cdot & \cdot & \mu_1 & q_1 & \cdot & \cdot & \lambda & \cdot & \cdot & \cdot & \cdots \\ (3,2) & \cdot & \cdot & \cdot & \mu_2 & \cdot & q_2 & \cdot & \cdot & \lambda & \cdot & \cdot & \cdots \\ (3,3) & \cdot & \cdot & \cdot & \mu_3 & \cdot & \cdot & q_3 & \cdot & \cdot & \lambda & \cdot & \cdots \\ \hline (4,1) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mu_1 & q_1 & \cdot & \cdot & \cdot & \cdots \\ (4,2) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mu_2 & \cdot & q_2 & \cdot & \cdot & \cdots \\ (4,3) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mu_3 & \cdot & \cdot & q_3 & \cdot & \cdots \\ (4,4) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mu_4 & \cdot & \cdot & \cdot & q_4 & \cdots \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

where the dots represent 0 and $q_i = -(\mu_i + \lambda)$, in order to have row sums of zero.

Each of the blocks in this matrix contain rates of transitions between the levels we defined earlier. Note that the level can only increase by 1 or decrease by 1, and this is shown in the structure in these blocks. Hence, we can represent this process as a level-dependent QBD as follows,

$$Q = \begin{array}{c} \\ \ell(0) \\ \ell(1) \\ \ell(2) \\ \ell(3) \\ \ell(4) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} \ell(0) & \ell(1) & \ell(2) & \ell(3) & \ell(4) & \cdots \end{array} \\ \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & \cdot & \cdot & \cdot & \cdots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \cdot & \cdot & \cdots \\ \cdot & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \cdot & \cdots \\ \cdot & \cdot & A_-^{(3)} & A_0^{(3)} & A_+^{(3)} & \cdots \\ \cdot & \cdot & \cdot & A_-^{(4)} & A_0^{(4)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Here,

$$A_-^{(n)} = \begin{bmatrix} 0 & \cdots & 0 & \mu_1 \\ 0 & \cdots & 0 & \mu_2 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \mu_n \end{bmatrix}, \qquad A_+^{(n)} = \begin{bmatrix} \lambda & 0 & \cdots & 0 & 0 \\ 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \lambda & 0 \end{bmatrix},$$

$$A_0^{(n)} = \begin{bmatrix} q_1 & 0 & \cdots & 0 \\ 0 & q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & q_n \end{bmatrix},$$

for $n \geq 1$, where $A_-^{(n)}$ is $n \times (n-1)$, $A_+^{(n)}$ is $n \times (n+1)$, and $A_0^{(n)}$ is $n \times n$. Also, $A_+^{(0)} = \lambda$ and $A_0^{(0)} = -\lambda$.

As this queue can be represented in this way, we can use the machinery of QBDs to numerically evaluate the stationary queue length distribution (See Section 2.3.2). Note that this level-dependent QBD is necessarily positive recurrent if there exists some N such that for all $i \geq N$, $\mu_i > \lambda$.

4.1.1 Delayed Queue-Length-Dependence

In this queue, the rates of service depend on the queue length at the start of a customer's service, that is a 'past' queue length. In the queue-length-dependent queues, the rates of arrival or service will also depend on a 'past' queue length. For example, in Chapter 5 the current arrival rate depends on the queue length immediately after the previous service. We will now justify this choice However, a dependence on the current queue length at any time would also induce dependence between the arrival process and service times. We will now justify the use of these 'delayed queue-length-dependence' models.

Consider a model where the service rate depends on the current queue length. The service time is then a sum of exponential times with changing rates as the queue length changes during this service period. This means the service time is no longer a property of the customer, but of the behaviour of the system. Since we seek to model the semi-experiment queues, which permutes inter-arrival or service times, we consider the data in a customer-centric approach. Hence, we justify using this delayed queue-length-dependence since each inter-arrival time or service time is an exponentially distributed random variable with a constant rate, and hence is easier to model in a customer-centric view.

4.2 Semi-Experiment Model

Note that we focus on the s-perm semi-experiment here as the a-perm semi-experiment model is more complex (see Chapter 6).

We apply an s-perm semi-experiment to this queue, which demonstrates there is indeed dependence since the queue-length distribution for the original and semi-experiment are significantly different (see Figure 4.2). This is achieved by simulating the (QL, S, O) queue, then retaining the exact arrival process while randomly permuting the service times for each customer.

As shown above, the stationary queue length distribution for the original queue can be calculated numerically using QBD techniques. So to avoid the need for lengthy simulations, we again seek to model the semi-experiment process as a queue in order to evaluate the stationary queue-length distribution.

Note that empirical s-perm semi-experiment keeps the same arrival stream, so the arrival process for the semi-experiment model is the same as the original queue; a Poisson process with rate λ.

In the empirical s-perm semi-experiment, the service times are randomly permuted, breaking the connection to the queue length. The service times are realisations from exponential distributions with rates μ_i, $i \geq 1$, hence the service times for the semi-experiment are a mixture of exponentials with these rates. That is, the service times in the semi-experiment are realisations from a hyperexponential distribution. The probability of getting a service time with rate μ_i is the probability that a service in the original queue begins with a queue length of i. Let this probability be denoted by $\widehat{\pi}_i$.

We therefore need to calculate $\{\widehat{\pi}_i : i \geq 1\}$, the stationary queue length distribution of the discrete-time Markov chain (DTMC) embedded at the epochs of service commencement. Let $\widehat{Q}_n$ be the queue length at the time when the nth service starts in the original (QL, S, O) queue at equilibrium. Then,

$$\widehat{\pi}_i = P(\widehat{Q}_n = i), \quad i \geq 1.$$

The following calculations for $\widehat{\boldsymbol{\pi}}$ are equivalent to those used by Harris [25].

Note that a service can begin in two ways: either a departure occurs leaving a non-empty queue in which the next customer can begin service, or an arrival occurs to an empty queue, commencing a service with a queue length of 1 immediately.

In order to find the queue lengths at the starts of service, we need to know how many (if any) arrivals occur during a service period. Let $k_{m|i}$ be the probability that m arrivals occur during a service period that began with a queue length of i, before a departure. So

$$k_{m|i} = P(\text{exactly } m \text{ arrivals during service period} \mid \text{service rate is } \mu_i),$$

for $i \geq 1$ and $m \geq 0$.

We can define $k_{m|i}$ more technically as follows. Let

$$\tau = \inf\{t \geq 0 : \exists k \in \mathbb{Z}_+ \text{ with } X(0) \in \ell(k-1), X(t) \in \ell(k)\}.$$

Then, for $i \geq 1$ and $m \geq 0$,

$$k_{m|i} = P(X(\tau) \in \ell(i+m-1) \mid X(0) = (i, i)).$$

To calculate these probabilities, assume T is the length of the service period. Then T is an exponentially distributed random variable with rate μ_i. Since the arrival pattern for this queue is a Poisson process, the number of arrivals during the time period T is Poisson distributed with parameter λT.

So, integrating over all possible values of T, we get for $i \geq 1$, $m \geq 0$,

$$\begin{aligned} k_{m|i} &= \int_0^\infty \frac{(\lambda t)^m}{m!} e^{-\lambda t} \mu_i e^{-\mu_i t} dt \\ &= \frac{\lambda^m \mu_i}{m!} \int_0^\infty t^m e^{-\mu_i t} dt \\ &= \frac{\lambda^m \mu_i}{(\lambda + \mu_i)^{m+1}}, \quad \text{integrating by parts.} \end{aligned}$$

Note that this value also has an intuitive interpretation. The probability of an arrival before a departure (during this service period) is given by $\frac{\lambda}{\lambda+\mu_i}$. So for m arrivals before departure, the probability is $\left(\frac{\lambda}{\lambda+\mu_i}\right)^m$. Finally the probability that there is a departure before the next arrival is given by $\frac{\mu_i}{\lambda+\mu_i}$. Hence,

$$k_{m|i} = \left(\frac{\lambda}{\lambda + \mu_i}\right)^m \frac{\mu_i}{\lambda + \mu_i}.$$

Note that this argument depends on the memoryless proiperty of the service time distribution.

Let $\widehat{p}_{ij}$ be the probability that a service begins with a queue length of j in the original queue, given that the previous service began with a queue length of i. That is,

$$\widehat{p}_{ij} = \lim_{n\to\infty} P(\widehat{Q}_n = j \mid \widehat{Q}_{n-1} = i).$$

These form the one-step transition probabilities for the DTMC embedded at epochs when services begin in the original (QL, S, O) queue.

If a service period begins with a queue length of i, then there may be any number of arrivals before the service period ends with a departure which allows a new service to begin. Hence, to start a service with a queue length of j, there must be $j - i + 1$ arrivals before the departure. Since the number of arrivals must be non-negative, $j \geq i - 1$. So, $\widehat{p}_{ij}$ is equivalent to the probability that there are $j - i + 1$ arrivals during a service period with rate μ_i.

So for i, j in the set $\{(i, j) : i \geq 3, j \geq i - 1\} \cup \{(i, j) : i = 1, 2, j \geq 2\}$,

$$\widehat{p}_{ij} = k_{j-i+1|i} = \frac{\lambda^{j-i+1}\mu_i}{(\lambda + \mu_i)^{j-i+2}},$$

These are the only non-zero values of $\widehat{p}_{ij}$, except for $\widehat{p}_{1,1}$ defined below.

Now consider the case when $i = j = 1$. There are two ways in which this can occur. First, a service may begin with a queue length of 1, then a single arrival occurs followed by a departure. The probability of this is simply $k_{1|1}$. Second, a service begins with a queue length of 1, then a departure occurs before any arrivals and the queue becomes empty. Once this occurs the only event that can occur is an arrival, so with probability 1 there will be an arrival which will immediately commence a service with a queue length of 1. Hence, the probability of this is $k_{0|1}$.

Therefore,

$$\widehat{p}_{11} = k_{1|1} + k_{0|1} = \frac{\lambda\mu_1}{(\lambda+\mu_1)^2} + \frac{\mu_1}{\lambda+\mu_1}.$$

So, the transition matrix for this embedded DTMC is given by

$$\widehat{P} = \begin{bmatrix} \widehat{p}_{11} & \widehat{p}_{12} & \widehat{p}_{13} & \widehat{p}_{14} & \cdots \\ \widehat{p}_{21} & \widehat{p}_{22} & \widehat{p}_{23} & \widehat{p}_{24} & \cdots \\ 0 & \widehat{p}_{32} & \widehat{p}_{33} & \widehat{p}_{34} & \cdots \\ 0 & 0 & \widehat{p}_{43} & \widehat{p}_{44} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Then $\{\widehat{\boldsymbol{\pi}} : j \geq 1\}$ is the stationary distribution for this DTMC. This can be calculated as shown in Section 2.2.4.

The s-perm semi-experiment queueing model is a single-server queue such that the:

- **Arrival Process** is given by a Poisson process with constant rate λ.
- **Service Times** have a hyperexponential distribution. This is a mixture of exponential distributions with rates μ_i for $i \geq 1$. The probability of getting rate μ_i is given by $\widehat{\pi}_i$, the (long-term) probability of the original queue beginning a service with queue length of i.

This process is labelled as the (QL, S, sSE) process. That is, the semi-experiment queue for the queue-length-dependent service rates process with a single server.

This process can also be represented as a CTMC $\{Y(t), t \geq 0\}$ on the two-dimensional state space $S = \{(n, i) : n \geq 0, i \geq 1\}$ where the level, n, represents

the current length of the queue and the phase, i, is an indicator for the current service rate. Since semi-experiments break the connection of service rates to queue length, the values that the phase i can take do not depend on n, in contrast to the original model.

The possible transitions for this semi-experiment process are given in Table 4.2. When there is an arrival, the queue length increases by 1 with rate λ, and the phase remains unchanged as the current service is not yet completed. For a departure, the queue length decreases by 1 with rate μ_i, where i is the current phase, and the next service rate is randomly chosen to be j with probability $\widehat{\pi}_j$. This is again a QBD since the level only increases or decreases by 1 for each transition.

From	To	Rate	
(n,i)	$(n+1,i)$	λ	
(n,i)	$(n-1,j)$	$\mu_i\widehat{\pi}_j$	$n \geq 1$

Table 4.2: Possible transitions for the (QL, S, sSE) process.

The generator matrix for the (QL, S, sSE) process is given by

$$\begin{bmatrix} B_0 & A_+ & \cdot & \cdot & \cdots \\ A_- & A_0 & A_+ & \cdot & \cdots \\ \cdot & A_- & A_0 & A_+ & \cdots \\ \cdot & \cdot & A_- & A_0 & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix},$$

where

$$A_+ = \begin{bmatrix} \lambda & \cdot & \cdot & \cdots \\ \cdot & \lambda & \cdot & \cdots \\ \cdot & \cdot & \lambda & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \qquad A_- = \begin{bmatrix} \mu_1\widehat{\pi}_1 & \mu_1\widehat{\pi}_2 & \mu_1\widehat{\pi}_3 & \cdots \\ \mu_2\widehat{\pi}_1 & \mu_2\widehat{\pi}_2 & \mu_2\widehat{\pi}_3 & \cdots \\ \mu_3\widehat{\pi}_1 & \mu_3\widehat{\pi}_2 & \mu_3\widehat{\pi}_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad A_0 = \begin{bmatrix} q_1 & \cdot & \cdot & \cdots \\ \cdot & q_2 & \cdot & \cdots \\ \cdot & \cdot & q_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$B_0 = \begin{bmatrix} -\lambda & \cdot & \cdot & \cdots \\ \cdot & -\lambda & \cdot & \cdots \\ \cdot & \cdot & -\lambda & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

and $q_i = -(\lambda + \mu_i)$.

Note that these matrices are not finite and need to be truncated in order to be practically implemented, as shown below.

This level-independent QBD representation of the (QL, S, sSE) process allows QBD machinery to analytically calculate the stationary queue-length distribution, $\boldsymbol{\pi}$, without the need for lengthy simulations and many permutations. Note that this queue is not necessarily positive recurrent even if the original queue is positive recurrent. Instead, it needs to satisfy the conditions for a positive recurrent queue as presented in Section 2.3.1.

4.2.1 Truncation and Augmentation

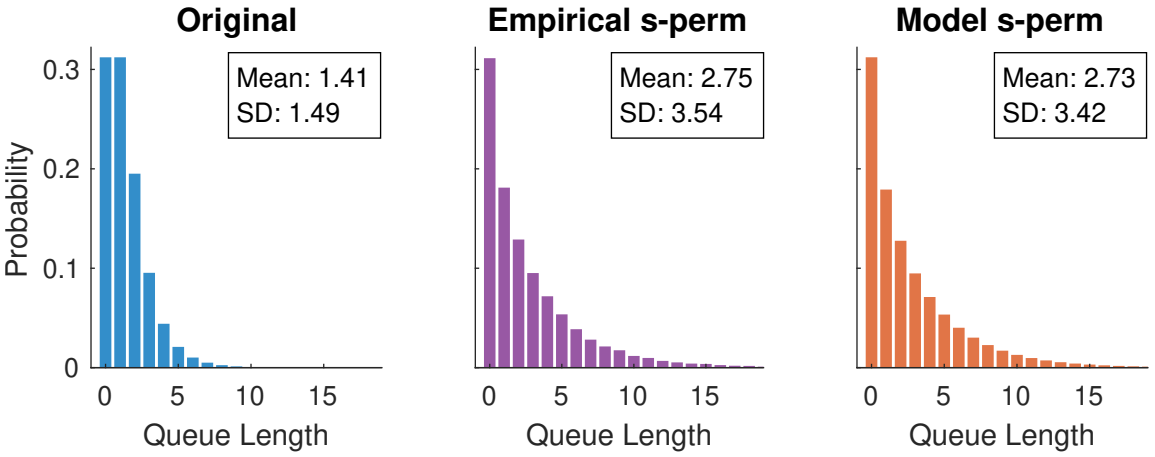
The phase-space for the semi-experiment QBD is infinite in size. In order to practically implement the model, the matrices A_+, A_-, A_0 and B_0 need to be truncated and augmented. It is very simple in this case. Let M be an artificial imposed upper bound of the phase space in the semi-experiment QBD; $S = \{(n, i) : n \geq 0, 1 \leq i \leq M\}$. Then, the matrices A_+, A_-, A_0 and B_0 are all truncated to $M \times M$ matrices. We also redefine the probability $\widehat{\pi}_M = 1 - \sum_{j=1}^{M-1} \widehat{\pi}_j$ so that the probabilities will still sum to 1. This probability $\widehat{\pi}_M$ can be thought of as the probability of a service beginning with a queue length of M or larger.

4.3 Results

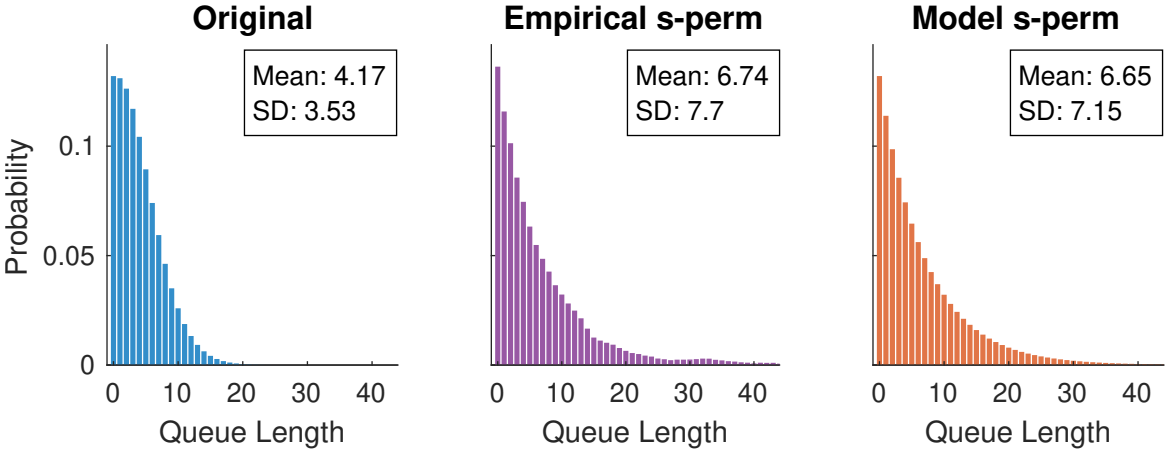
Now we can compare the empirical s-perm semi-experiments to the model constructed above. The stationary queue-length distributions for each, shown in Figure 4.2, pass the KS test, indicating that they are from the same distribution. Hence, the model accurately replicates the empirical process in term of the queue-length distribution. Note that the values for μ_i here are chosen arbitrarily to produce the reasonably different distributions to demonstrate that the method can be applied to a range of queues. We also compare the semi-experiment queue-length distribution to the original queue's stationary queue-length distribution in Figure 4.2. The differences between these, confirmed by the KS test, indicate a dependence between the arrival process and service times, as expected. The 'strength' of this difference can be measured by the KS test statistic and is an

indication of how significantly the dependence affected the queue length of the original queue.

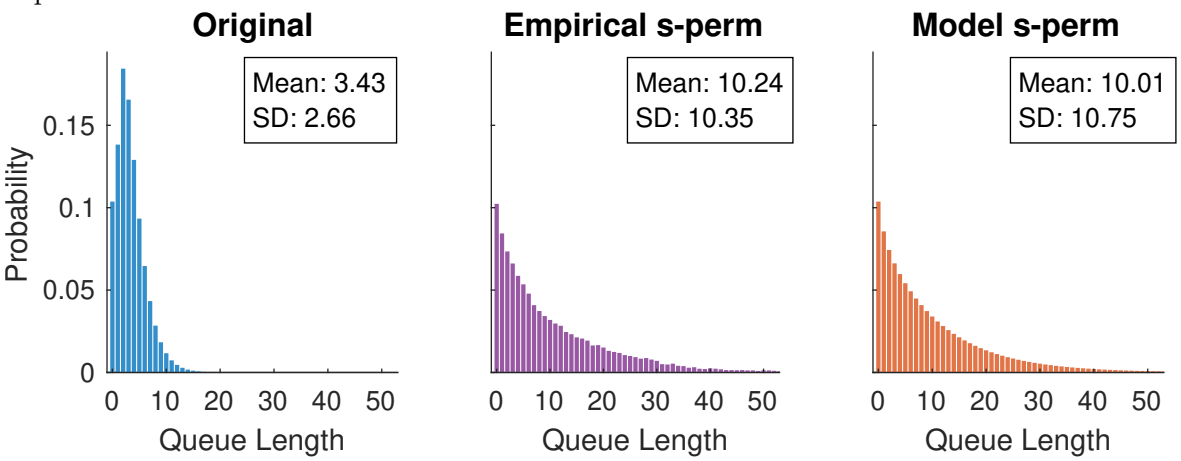
Note that the s-perm semi-experiment queue-length distributions appear very similar to an exponential shape, like that of a standard $M/G/1$ queue with no dependence. That is, the s-perm semi-experiment has disrupted the dependence completely so that the s-perm queue has no dependence between the arrivals and services. This is supported by the model for the s-perm queues, which is an $M/G/1$ queue with no dependence.



(a) Service rate of $\mu_i = 2i^2$. The KS statistic between the original and semi-experiment models is 0.202400.



(b) Service rate of $\mu_i = \log(i + 6.5)$. The KS statistic between the original and semi-experiment models is 0.169800.



(c) Service rate of $\mu_i = 1.5$, if $i \leq 2$ and $\mu_i = 3.5$ if $i \geq 3$. The KS statistic between the original and semi-experiment models is 0.384000.

Figure 4.2: The blue and orange plots show the stationary queue-length distribution of the original (QL, S, O) model and s-perm semi-experiment (QL, S, sSE) model, respectively. The purple plots show the empirical queue length distribution for a single empirical s-perm semi-experiment. That is, a single realisation of the original queue (run for 20,000 time steps) with a single random permutation of the service times. Each queue has the arrival rate $\lambda = 2$.

4.3.1 Efficiency Comparison

The benefits of using a QBD model instead of simulations is especially evident in the computational efficiency of calculating the queue-length distribution for both the original queue and the semi-experiment queue. The following is an example to illustrate this using $\lambda = 2$ and $\mu_i = 3i$.

Original Queue First consider the original queue. To have a meaningful comparison, the simulation must be run long enough to have an accurate queue length distribution. To determine an appropriate run time, we take a single long simulation of 10,000 time steps and make regular 'cuts' along its run time. Each cut contains 100 events. Taking the simulation from the beginning to each of these cuts, we evaluate the queue-length distribution and compare it to the distribution found using the QBD model via the KS test. The first time cut after which the KS test is always passed is recorded. Note that since we are using a single long-run simulation, once we have sufficient data to pass the KS test, adding more data (which we know is from the same distribution) means that the test will continue to be passed (except for possibly some small variation at this changeover point). This process is repeated for 30 simulations and the average of these times is taken to be the appropriate run time for a simulation to sufficiently capture the queue-length distribution, T.

Note that the QBD model's queue-length distribution is evaluated numerically and so subject to a prescribed level of accuracy. However, we only need to compare how long the simulation method takes to reach the same level of accuracy as the QBD method.

The time taken to run a single simulation for T time steps and also evaluate the queue-length distribution is then compared to the time taken to construct the QBD matrices and to evaluate the stationary queue-length distribution. This is repeated several times to establish average timings.

For the example with the parameters $\lambda = 2$ and $\mu_i = 3i$. The value for T was found to be 4670, averaging over 50 simulations with 30,000 time steps. The average time over 50 evaluations for each method is

- Simulations: 3.9461 seconds,
- QBD Model: 0.0054 seconds.

So the QBD model is over 700 times faster than the simulation method.

Semi-Experiment Queue Now consider the semi-experiment queue. Here we chose $M = 20$. The comparison here is between the time taken to run a single simulation for T time steps, permute the service times once and then evaluate the queue-length distribution; and the time taken to construct the QBD matrices and evaluate the stationary queue-length distribution.

Using the same parameters as above, the average time over 50 evaluations for each method is

- Simulations: 4.8466 seconds,
- QBD Model: 0.6860 seconds.

So the QBD method is 7 times faster. Note that for accurate results, a long simulation and many permutations of each simulation would need to be performed, making the computational time even larger.

4.4 Conclusion

In this chapter we introduced a queue where the service rates depend on the queue length at the times when service begins, and constructed a level-dependent QBD model for this. Then we considered an s-perm semi-experiment applied to this queue. We found the stationary distribution for the embedded DTMC at the epochs when service begins to find the distribution of queue-lengths at times when service begins. This gives the distribution for the service rates without being conditioned on the queue-length process, which is needed to describe the service process after the permutation of service times. This was used to construct a QBD model for the s-perm semi-experiment and allowed a comparison of the stationary queue-length distributions of the original and semi-experiment queue. Finally, we showed the significant increase in computational efficiency of using these models to find the queue-length distributions, compared to simulation methods.

Chapter 5

Queue-Length-Dependent Arrival Rates

5.1 Original Model

Now consider a single-server queue with arrival rates that are dependent on the length of the queue at the time when the inter-arrival period begins; that is, the queue length immediately after the last arrival. This also induces dependence between the arrival process and service time distribution through a dependence between the queue length and the inter-arrival times. This model is more formally defined below.

- **Arrival Process** has inter-arrival times which are exponentially distributed with rate λ_j, where $j \geq 1$ is the queue length immediately after the last arrival.
- **Service Times** are exponentially distributed with constant rate μ.

This process is labelled as the (QL, A, O) process. That is, the original queue with queue-length-dependent arrival rates and a single server.

As before, we cannot represent this process as a general birth-and-death process since the queue depends on the state in the past. However, it can be represented as a continuous-time Markov chain $\{X(t), t \geq 0\}$ on the two-dimensional state space $S = \{(n, i) : n \geq 0, i \geq 0, n + i \geq 1\}$. The level, n, represents the

current length of the queue and the phase, i, represents the number of departures since the last arrival. That is, $n+i$ is the queue length immediately after the last arrival, and informs the arrival rate. Hence, $n+i \geq 1$ since the queue must be at least length 1 after an arrival. The level, n, takes values from the non-negative integers as it is counting people in a queue. The phase, $i \geq 0$, since it is possible to have had any (non-negative) number of departures since the last arrival. Again, we can also write $S = \bigcup_{n \geq 0} \ell(n)$ where $\ell(n) = \{(n,0),(n,1),\dots\}$ is the nth level for $n \geq 1$ and $\ell(0) = \{(0,1),(0,2),\dots\}$.

However, for simplicity in notation, we also include the state $(0,0)$ in $\ell(0)$, which can never be entered and define $\lambda_0 = 0$.

The possible transitions for this process are depicted in the state transition diagram in Figures 5.1 and 5.2 and detailed in Table 5.1. For a departure event, the current queue length decreases by 1 with rate μ, and the phase, number of departures since the last arrival, increases by 1. For an arrival event, the queue length increases by 1 with rate λ_{n+i} and the phase changes to 0 as there are no departures since this latest arrival. Hence, this Markov chain is also a QBD since the level can only increase or decrease by 1, and is in fact level-dependent.

From	To	Rate	For
(n,i)	$(n-1,i+1)$	μ	$(n,i) \in S, n \geq 1$
(n,i)	$(n+1,0)$	λ_{n+i}	$(n,i) \in S$

Table 5.1: Possible transitions for the (QL, A, O) process.

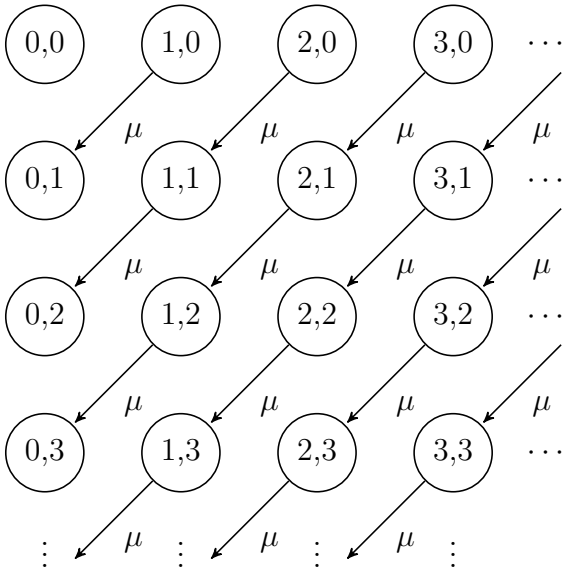


Figure 5.1: State transition diagram for the (QL, A, O) departure transitions.

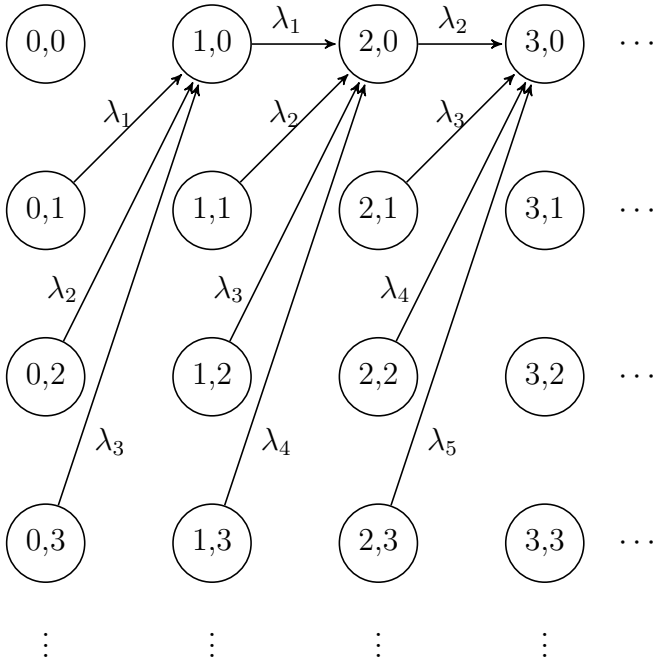


Figure 5.2: State transition diagram for the (QL, A, O) arrival transitions.

The generator matrix for the process is

$$Q = \begin{array}{c} \\ \ell(0) \\ \ell(1) \\ \ell(2) \\ \ell(3) \\ \ell(4) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} \ell(0) & \ell(1) & \ell(2) & \ell(3) & \ell(4) & \cdots \end{array} \\ \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & \cdot & \cdot & \cdot & \cdots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \cdot & \cdot & \cdots \\ \cdot & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \cdot & \cdots \\ \cdot & \cdot & A_-^{(3)} & A_0^{(3)} & A_+^{(3)} & \cdots \\ \cdot & \cdot & \cdot & A_-^{(4)} & A_0^{(4)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array},$$

where, for $n \geq 1$,

$$A_-^{(n)} = \begin{bmatrix} \cdot & \mu & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \mu & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \mu & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \mu & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \qquad A_0^{(n)} = \begin{bmatrix} q_n & \cdot & \cdot & \cdot & \cdots \\ \cdot & q_{n+1} & \cdot & \cdot & \cdots \\ \cdot & \cdot & q_{n+2} & \cdot & \cdots \\ \cdot & \cdot & \cdot & q_{n+3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

for $n \geq 0$,

$$A_+^{(n)} = \begin{bmatrix} \lambda_n & \cdot & \cdot & \cdots \\ \lambda_{n+1} & \cdot & \cdot & \cdots \\ \lambda_{n+2} & \cdot & \cdot & \cdots \\ \lambda_{n+3} & \cdot & \cdot & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

and

$$A_0^{(0)} = \begin{bmatrix} -\lambda_0 & \cdot & \cdot & \cdots \\ \cdot & -\lambda_1 & \cdot & \cdots \\ \cdot & \cdot & -\lambda_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

with $q_n = -(\mu + \lambda_n), n \geq 1$.

Again, this QBD representation allows for the efficient analytic calculation of the stationary queue-length distribution for the (QL, A, O) process. Note that this QBD is positive recurrent if there exists some N such that for all $i \geq N$, $\lambda_i < \mu$.

5.1.1 Truncation and Augmentation

The sub-matrices $A_0^{(n)}, A_+^{(n)}, A_-^{(n)}$ have infinite size and so need to be truncated and augmented in order to be used practically. Let M be an imposed artifical maximum on the phase space so that the state space becomes $= \{(n,i) : n \geq 0, 0 \leq i \leq M, n+i \geq 1\}$. That is, M is an imposed upper bound on the number of departures possible between two arrivals. So all of the model matrices are truncated to be $(M+1) \times (M+1)$ (since we can have 0 departures). The matrices $A_-^{(n)}$ is augmented so that the $(M+1)$th row is a row of zeros. We also adjust the $(M+1)$th row of A_0 to be $(0, 0, \ldots, -\lambda_{n+M+1})$. This means when the process is in phase M, it can have no departures and the next event will be an arrival. This will not affect the queue behaviour if M is sufficiently large that the probability of more than M departures occurring between two arrivals is small, since this can be considered as a Poisson process.

5.2 The First Semi-Experiment Model

Again we only consider the s-perm semi-experiment for this model. The a-perm semi-experiment model is considered in Chapter 6.

Once again, we seek to model the semi-experiment process to avoid the lengthy computation times needed for simulating queues and permuting service times.

The following section attempts to construct a queueing model for the semi-experiment process for the queue-length-dependent arrival queue, using a similar method to that used for the queue-length-dependent service queue. We shall see later that this model is not sufficient and the correct model is presented in Section 5.4. However, we present it here because it is a natural approach, particularly given the (QL, S, sSE) modelling approach, and it can be used to demonstrate the dependence in this model is more complex than it may seem at first glance.

Recall that the service times in the original (QL, A, O) queue were independent and identically distributed exponential random variables with rate μ. Hence after the permutation the service times are still independent and identically distributed exponential random variables with rate μ.

The inter-arrival times in the permuted queue are no longer dependent on

that queue length at the last arrival. However, since the arrival stream is preserved some dependence structure will also be preserved.

Consider a simple example with a linearly decreasing function for λ_i. If the original queue length immediately after an arrival is small, then the arrival rate would be relatively large and the time until the next arrival would be relatively short on average. Hence, there is likely to be relatively few departures during this period, resulting in a similar queue length immediately after the next arrival. If the original queue length immediately after an arrival is large, then the arrival rate would be relatively small and the time until the next arrival would be relatively long. Hence there is likely to be a relatively large number of departures, resulting in a smaller queue length immediately after the next arrival.

The original (QL, A, O) queue can be considered as an example of a $G/M/1$ queue. The analysis for this type of queue commonly involves the use of the embedded DTMC at points of arrival. We can more precisely explain the dependence using such a DTMC.

Let A_m be the time of the mth arrival and $\widetilde{Q}_m$ be the queue length immediately after the mth arrival in the original queue. Suppose that $\widetilde{Q}_{m-1} = i \geq 1$ and $\widetilde{Q}_m = j \geq 1$. During the inter-arrival period between the $(m-1)$th and mth arrivals, the only events that can happen are departures. In order to observe $\widetilde{Q}_m = j$, there must have been $i - j + 1$ departures. The number of possible departures is $0, 1, \ldots, i$, hence j must be such that $1 \leq j \leq i + 1$. Furthermore, the probability of observing $i - j + 1$ departures depends on the length of the inter-arrival period, $(A_m - A_{m-1})$, which is exponentially distributed with rate λ_i. Hence, the probability distribution of j depends on i and, through this, the $(m+1)$th arrival rate λ_j depends on the previous arrival rate λ_i.

Let $\widetilde{p}_{ij}$ be the probability of seeing an arrival rate λ_j after an arrival rate λ_i in the original queue; that is, the probability of having a queue length j immediately after an arrival, given the queue length was i immediately after the previous arrival. So,

$$\widetilde{p}_{ij} = \lim_{m \to \infty} P(\widetilde{Q}_m = j \mid \widetilde{Q}_{m-1} = i),$$

where i, j are in the set $\{(i, j) : 1 \leq j \leq i+1, i \geq 1\}$. So then $\widetilde{p}_{ij}$ is equivalent to the probability of observing $i - j + 1$ departures during an inter-arrival period starting with a queue length of i. If we assume the the length of the inter-arrival period is T, then T is an exponentially distributed random variable with rate λ_i. Since the service times are independent exponentially distributed random variables with rate μ, the number of departures during a time period T has a Poisson distribution

with rate μT. Hence, to calculate this probability we integrate over all possible values of T.

First consider the case $2 \leq j \leq i+1$,

$$\begin{aligned}\widetilde{p}_{ij} =& \int_0^\infty \frac{(\mu t)^{i-j+1}}{(i-j+1)!} e^{-\mu t} \lambda_i e^{-\lambda_i t} dt \\ =& \frac{\mu^{i-j+1}\lambda_i}{(\mu+\lambda_i)^{i-j+2}}, \qquad \text{after integrating by parts } i-j+1 \text{ times.}\end{aligned}$$

Now consider the case when $j = 1$. This implies that the queue became empty during the inter-arrival period, which ended with an arrival to this empty queue. When the queue is empty, there can only be arrival events and no departure events and hence once i departures have occurred, the probability that the next event is an arrival is 1. Therefore,

$$\begin{aligned}\widetilde{p}_{i1} =& P(\widetilde{Q}^{(m)} = 1 \mid \widetilde{Q}^{(m-1)} = i) \\ =& P(i \text{ departures before next arrival} \mid \widetilde{Q}^{(m-1)} = i) \\ =& \frac{\mu^i}{(\mu+\lambda_i)^i}.\end{aligned}$$

The state space for this DTMC is $\{\widetilde{Q}_m : m \geq 1\}$ and the transition probabilities are $\{\widetilde{p}_{ij} : i \geq 1, 1 \leq j \leq i+1\}$. That is, the transition matrix is

$$\widetilde{P} = \begin{bmatrix} \widetilde{p}_{11} & \widetilde{p}_{12} & 0 & 0 & \cdots \\ \widetilde{p}_{21} & \widetilde{p}_{22} & \widetilde{p}_{23} & 0 & \cdots \\ \widetilde{p}_{31} & \widetilde{p}_{32} & \widetilde{p}_{33} & \widetilde{p}_{34} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

The proposed s-perm semi-experiment queueing model is a single-server queue such that the:

- **Arrival Process** has exponentially distributed inter-arrival times with rates λ_j where j is determined by a sample path of the above DTMC with state space $\{\widetilde{Q}_m : m \geq 1\}$ and transition matrix $\widetilde{P}$.
- **Service Times** are independent and identically exponentially distributed with rate μ.

Now we can represent the proposed semi-experiment model as a level-independent QBD, $\{Y(t), t \geq 0\}$ on the two-dimensional state space $S = \{(n, i) : n \geq 0, i \geq 1\}$. Here, the level n represents the current queue length and the phase i is an indicator for the rate of the next arrival. Since there is no connection between the inter-arrival times and the semi-experiment queue length, the values the phase can take do not depend on n. The transitions for this process are given in Table 5.2.

From	To	Rate	
(n, i)	$(n+1, j)$	$\lambda_i \widetilde{p}_{ij}$	$n \geq 0, i, j \geq 1$
(n, i)	$(n-1, i)$	μ	$n \geq 1, i \geq 1$

Table 5.2: Transitions for the semi-experiment model of the (QL, A, O) original queue.

The transition matrix for this process is given by

$$\begin{bmatrix} B_0 & A_+ & \cdot & \cdot & \cdots \\ A_- & A_0 & A_+ & \cdot & \cdots \\ \cdot & A_- & A_0 & A_+ & \cdots \\ \cdot & \cdot & A_- & A_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where

$$A_+ = \begin{bmatrix} \lambda_1\widetilde{p}_{11} & \lambda_1\widetilde{p}_{12} & 0 & 0 & \cdots \\ \lambda_2\widetilde{p}_{21} & \lambda_2\widetilde{p}_{22} & \lambda_2\widetilde{p}_{23} & 0 & \cdots \\ \lambda_3\widetilde{p}_{31} & \lambda_3\widetilde{p}_{32} & \lambda_3\widetilde{p}_{33} & \lambda_3\widetilde{p}_{34} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \end{bmatrix}, \qquad A_- = \begin{bmatrix} \mu & 0 & 0 & \cdots \\ 0 & \mu & 0 & \cdots \\ 0 & 0 & \mu & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$B_0 = \begin{bmatrix} -\lambda_1 & 0 & 0 & \cdots \\ 0 & -\lambda_2 & 0 & \cdots \\ 0 & 0 & -\lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \qquad A_0 = \begin{bmatrix} q_1 & 0 & 0 & \cdots \\ 0 & q_2 & 0 & \cdots \\ 0 & 0 & q_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

with $q_i = -(\lambda_i + \mu), i \geq 1$.

5.2.1 Truncation and Augmentation

Note that these sub-matrices that define the QBD have infinitely many rows and columns. Hence, we need to truncate and augment them in order to implement this model. Let M be an enforced artificial maximum on the phase space such that the state space becomes $S = \{(n, i) : n \geq 0, 1 \leq i \leq M\}$. The matrices A_+, A_-, B_0, A_0 are truncated to be $M \times M$ and the probabilities are augmented so that $\widetilde{p}_{M,M}$ is replaced by $\widetilde{p}_{M,M} + \widetilde{p}_{M,M+1}$.

5.3 Inter-Arrival Time Dependence Problem

Figure 5.3 shows the queue-length distribution for a single realisation of the original queue, with the service times permuted (empirical semi-experiments) and the stationary queue-length distribution of the above QBD model. The two distributions are clearly different, which is also confirmed by the KS test. This indicates that the model presented above is not in fact an accurate representation of the semi-experiment process for the queue-length-dependent arrival queue.

Figure 5.3: The left plot shows the stationary queue-length distribution of the proposed QBD semi-experiment model as defined in Section 5.2. The right plot shows the empirical queue-length distribution of the empirical semi-experiment run for 20,000 time steps. Here the parameters are $\mu = 2$ and $\lambda_j = 6/j$.

We want to understand where the model is failing to reproduce the behaviour of the empirical semi-experiment.

Firstly, the marginal distributions of service times and inter-arrival times produced by the QBD model are the same as those produced by the empirical semi-experiments. So the issue must be due to some correlation structure in the arrival process that is not captured in the current model.

One concern is that the reuse of actual (but permuted) service times in the semi-experiment might carry some unexpected correlation. However, simulating the queue and then replacing the service times with an independent sample of exponentially distributed service times with rate μ gives the same results as a semi-experiment of a separate simulation.

The semi-experiment model assumes that the arrival rates only depend on the previous arrival rate (in the absence of knowledge of the original queue lengths at arrival epochs). However, there may be more correlation within the arrival process, so we consider the autocorrelation function. Let $\lambda^{(k)}$ be the kth arrival rate in a particular realisation of a queue. Let C be the total number of arrivals in this realisation. Let $\bar{\lambda}$ be the sample mean arrival rate and σ_λ^2 be the sample variance of the arrival rates. Then for lag $k = 0, 1, 2, \ldots$, the sample autocorrelation is defined to be

$$\rho_k = \frac{c_k}{\sigma_\lambda^2},$$

where

$$c_k = \frac{1}{C} \sum_{t=1}^{C-k} (\lambda^{(t)} - \bar{\lambda})(\lambda^{(t+k)} - \bar{\lambda}).$$

Note that there is always an autocorrelation of 1 at lag 0 since this is the correlation between the same observation. That is, $\rho_0 = \frac{\sigma_\lambda^2}{\sigma_\lambda^2} = 1$.

Simplistically, the sample autocorrelation at lag k describes similarity of observations as a function of the time lag k between them. The sample partial autocorrelation is similar, but it adjusts for the linear effects of $\lambda^{(t+1)}, \ldots, \lambda^{(t+k-1)}$. That is, the partial autocorrelation is the autocorrelation of a time series, regressed on the values of the time series at all shorter lags.

Figure 5.4 shows the sample partial autocorrelations of the arrival rates in both the model in Section 5.2 and the empirical semi-experiment. There are a few significant lags for each, with more in the model. The most striking difference is the partial autocorrelation at lag 1, which is significantly higher for the model simulation than the empirical semi-experiment. This one-step dependence is expected from a level-independent QBD. This is another indication that the model is not accurately capturing the correlation structure of the arrival rates.

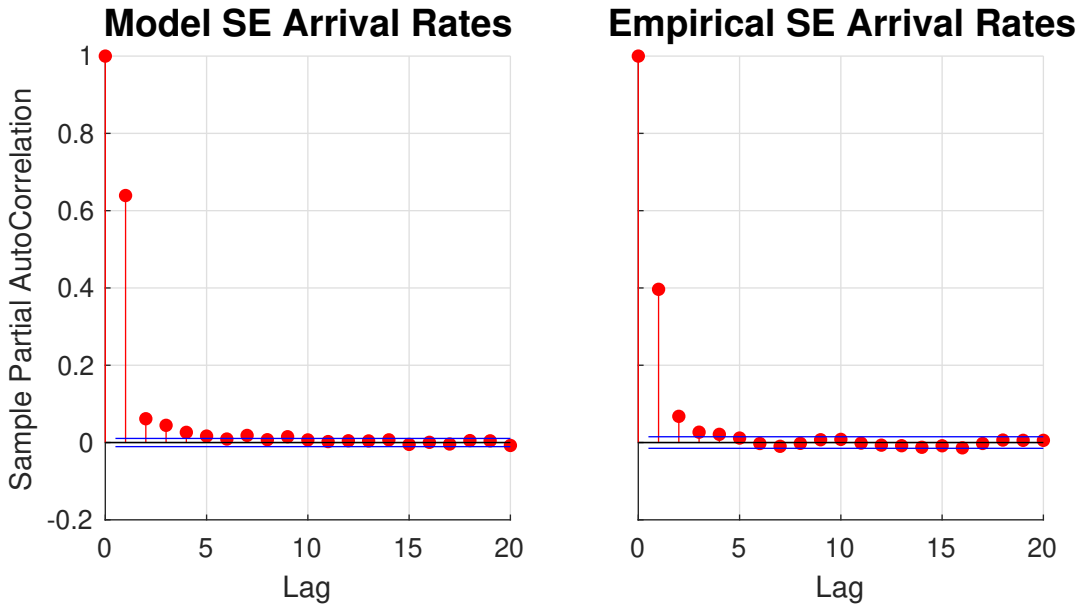


Figure 5.4: The partial autocorrelation functions for arrival rates of a simulation of the proposed semi-experiment model as in Section 5.2 (left) and the empirical semi-experiment (right), with $\mu = 2$ and $\lambda_j = 6/j$, each run for 30,000 time steps.

The issue with the semi-experiment model in Section 5.2 is that there is a dependence on the actual inter-arrival *times*, over and above the inter-arrival *rate*, which are retained while the service times are permuted. The length of an inter-arrival time influences how many departures will occur during that period in the original queue. After permuting service times, the length of each inter-arrival *time* is the same and hence, one can expect a similar number of departures in those periods as was observed in the original queue. In the original queue, the number of departures in each inter-arrival period exactly informs the queue length at the next arrival epoch and hence the next arrival rate. Hence, in the empirical semi-experiment the *length* of the inter-arrival period somewhat informs the next arrival rate. In the current semi-experiment model, this feature is not captured. The model assumes that the next arrival *rate* only depends on the value of the previous arrival rate. The true model for the semi-experiment needs to replicate this type of arrival process with independent service times.

5.4 The Second Semi-Experiment Model

The idea for the new semi-experiment model is as follows. We create a model of an underlying queue which is exactly equivalent to the original queue model. Then, the arrival process from this underlying queue is used as the arrival process for the semi-experiment queue, and the services for the semi-experiment queue are independent of the services in the underlying queue.

That is, this is a single-server queue such that the:

- **Arrival Process** is the same arrival process as the underlying queue which replicates the original queue (QL, A, O)
- **Service Times** are independent and identically exponentially distributed with rate μ.

This queue is labelled as (QL, A, sSE).

Let the state for this semi-experiment model be $(n, (k, i))$ where

- n is the queue length of the semi-experiment queue
- k is the queue length of the underlying original process
- i is the number of departures since the last arrival in the underlying original process.

The state space is $S = \{(n, (k, i)) : n \geq 0, k \geq 0, i \geq 0\}$. Let $m(n, k) = \{(n, (k, i)) : i \geq 0\}$ and $\ell(n) = \{(n, (k, i)) : k \geq 0, i \geq 0\} = \bigcup_{k \geq 0} m(n, k)$. Hence, $S = \bigcup_{n \geq 0} \ell(n) = \bigcup_{n \geq 0} \bigcup_{k \geq 0} m(n, k)$.

The possible transitions are given in Table 5.3. The rates here show that the semi-experiment queue and underlying queue both increase at the same time with the arrival rate of the underlying queue, the semi-experiment queue decreases independently with rate μ and the underlying queue decreases independently with rate μ. This process can be represented as a level-independent QBD, where n is the level and (k, i) is the phase.

From	To	Rate	For	Description
$(n,(k,i))$	$(n+1,(k+1,0))$	λ_{k+i}	$(n,(k,i)) \in S$	Underlying and SE arrival
$(n,(k,i))$	$(n-1,(k,i))$	μ	$n \geq 1$	SE departure
$(n,(k,i))$	$(n,(k-1,i+1))$	μ	$k \geq 1$	Underlying departure

Table 5.3: Possible transitions for the (QL, A, sSE) process.

Note that the phase (k,i) is defined to take doubly infinitely many values, and hence we again need to truncate and augment the block matrices for the QBD model.

First consider the QBD representation of the underlying queue with level k and phase i. The state space for this model is simply $S_U = \{(k,i) : k \geq 0, i \geq 0\}$. The rate sub-matrices for this process are identical to those for the original (QL, A, O) QBD model in Section 5.1. They are labelled here using U to represent the underlying process.

The transition matrix for this underlying process (for all $n \geq 0$) is given by

$$Q_U = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ m(n,3) \\ m(n,4) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} m(n,0) & m(n,1) & m(n,2) & m(n,3) & m(n,4) & \cdots \end{array} \\ \begin{bmatrix} U_0^{(0)} & U_+^{(0)} & \cdot & \cdot & \cdot & \cdots \\ U_-^{(1)} & U_0^{(1)} & U_+^{(1)} & \cdot & \cdot & \cdots \\ \cdot & U_-^{(2)} & U_0^{(2)} & U_+^{(2)} & \cdot & \cdots \\ \cdot & \cdot & U_-^{(3)} & U_0^{(3)} & U_+^{(3)} & \cdots \\ \cdot & \cdot & \cdot & U_-^{(4)} & U_0^{(4)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}$$

where, for $k \geq 1$,

$$U_-^{(k)} = U_- = \begin{bmatrix} \cdot & \mu & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \mu & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \mu & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \mu & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \qquad U_0^{(k)} = \begin{bmatrix} q_k & \cdot & \cdot & \cdot & \cdots \\ \cdot & q_{k+1} & \cdot & \cdot & \cdots \\ \cdot & \cdot & q_{k+2} & \cdot & \cdots \\ \cdot & \cdot & \cdot & q_{k+3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

and for $k \geq 0$,

$$U_+^{(k)} = \begin{bmatrix} \lambda_k & \cdot & \cdot & \cdots \\ \lambda_{k+1} & \cdot & \cdot & \cdots \\ \lambda_{k+2} & \cdot & \cdot & \cdots \\ \lambda_{k+3} & \cdot & \cdot & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

and

$$U_0^{(0)} = \begin{bmatrix} -\lambda_0 & \cdot & \cdot & \cdots \\ \cdot & -\lambda_1 & \cdot & \cdots \\ \cdot & \cdot & -\lambda_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

with $q_k = -(\mu + \lambda_k), k \geq 1$.

Now, consider the overall semi-experiment queue. First, consider the block matrix A_+ which contains all rates in which the semi-experiment queue length increases by 1 ($n \to n+1$). When this occurs, we also see $k \to k+1$ and $i \to 0$, as shown in Table 5.3. Therefore, the only non-zero blocks within A_+ are those in which $k \to k+1$, or the upper diagonal.

Thus, we can construct A_+ as

$$A_+ = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} m(n+1,0) & m(n+1,1) & (m(n+1,2) & m(n+1,3) & \cdots \end{array} \\ \begin{bmatrix} \mathbf{0} & U_+^{(0)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & U_+^{(1)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & U_+^{(2)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Similarly, we consider the matrix A_-, which contains transitions such that $n \to n-1, n \geq 1$. From Table 5.3, when this occurs, the phase (k, i) does not change and this transition occurs with rate μ. Hence, the only non-zero blocks within A_- is where $k \to k$ (the diagonal).

So, we can construct A_-,

$$A_- = \begin{bmatrix} \mu & \cdot & \cdot & \cdot & \cdots \\ \cdot & \mu & \cdot & \cdot & \cdots \\ \cdot & \cdot & \mu & \cdot & \cdots \\ \cdot & \cdot & \cdot & \mu & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Now consider the block matrix A_0, which contains transitions such that $n \to n$, $n \geq 1$. This is the transition such that $n \to n$, $k \to k-1$ and $i \to i+1$, occurring at rate μ (from Table 5.3). These rates are contained in the matrices $U_-^{(k)} = U_-$, and put into the lower diagonal of A_0. Also, the diagonal of A_0 must contain the negative sum of all transition rates of each phase. So we define the diagonal block matrices of A_0 as

$$V_0^{(k)} = U_0^{(k)} - \begin{bmatrix} \mu & \cdot & \cdot & \cdot & \cdots \\ \cdot & \mu & \cdot & \cdot & \cdots \\ \cdot & \cdot & \mu & \cdot & \cdots \\ \cdot & \cdot & \cdot & \mu & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

since $U_0^{(k)}$ is a diagonal matrix containing the negative sum of transition rates for the underlying process, and so we just need to subtract an additional μ for the departures of the semi-experiment process.

So, we can construct A_0,

$$A_0 = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} m(n,0) & m(n,1) & m(n,2) & \cdots \end{array} \\ \begin{bmatrix} V_0^{(0)} & \mathbf{0} & \mathbf{0} & \cdots \\ U_- & V_0^{(1)} & \mathbf{0} & \cdots \\ \mathbf{0} & U_- & V_0^{(2)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Finally, we need to construct the matrix B_0 which contains the rates in which $n \to n$ when $n = 0$. This is very similar to A_0 as the lower diagonal still contains the departure rates from the underlying queue contained in $U_-^{(k)}$. The only difference is that when $n = 0$, there are no departures from the overall semi-experiment queue. Hence, the diagonal block matrices of B_0 are simply $U_0^{(k)}$.

So,

$$B_0 = \begin{array}{c} \\ m(0,0) \\ m(0,1) \\ m(0,2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} m(0,0) & m(0,1) & m(0,2) & \cdots \end{array} \\ \begin{bmatrix} U_0^{(0)} & \mathbf{0} & \mathbf{0} & \cdots \\ U_- & U_0^{(1)} & \mathbf{0} & \cdots \\ \mathbf{0} & U_- & U_0^{(2)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

So the overall transition matrix for the (QL, A, sSE) process is

$$Q = \begin{array}{c} \\ \ell(0) \\ \ell(1) \\ \ell(2) \\ \ell(3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} \ell(0) & \ell(1) & \ell(2) & \ell(3) & \cdots \end{array} \\ \begin{bmatrix} B_0 & A_+ & \mathbf{0} & \mathbf{0} & \cdots \\ A_- & A_0 & A_+ & \mathbf{0} & \cdots \\ \mathbf{0} & A_- & A_0 & A_+ & \cdots \\ \mathbf{0} & \mathbf{0} & A_- & A_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix} \end{array}$$

5.4.1 Truncation and Augmentation

Note that the phase (k, i) is defined to take doubly infinitely many values. Hence, to practically implement this model, we need to appropriately truncate and augment the phase space to have finite block matrices B_0, A_+, A_0, A_-. Let M_1 and M_2 be some large artificial maximum values such that the state space for this process is truncated to be $S = \{(n, (k, i)) : n \geq 0, 0 \leq k \leq M_1, 0 \leq i \leq M_2\}$. This means that the underlying matrices $U_-, U_+^{(k)}, U_0^{(k)}$ are all of size $(M_2 + 1) \times (M_2 + 1)$ and the matrices B_0, A_+, A_0, A_- are all of size $(M_1 + 1)(M_2 + 1) \times (M_1 + 1)(M_2 + 1)$. We augment this truncated process by not allowing any arrivals in the underlying queue or the semi-experiment queue when the underlying queue length, k, is at its maximum M_1. We also do not allow departures in the underlying queue when the number of underlying departures between two arrivals, i, is at its maximum M_2.

Therefore, the truncated and augmented matrices for the underlying process are

For $1 \leq k \leq M_1$,

$$U_-^{(k)} = U_- = \begin{array}{c} \\ (k,0) \\ (k,1) \\ (k,2) \\ (k,3) \\ \vdots \\ (k,M_2-1) \\ (k,M_2) \end{array} \begin{array}{c} \begin{array}{cccccc} (k-1,0) & (k-1,1) & (k-1,2) & (k-1,3) & \cdots & (k-1,M_2) \end{array} \\ \begin{bmatrix} \cdot & \mu & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \mu & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \mu & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdot & \cdot & \cdots & \mu \\ \cdot & \cdot & \cdot & \cdot & \cdots & 0 \end{bmatrix} \end{array}.$$

For $0 \leq k \leq M_1 - 1$,

$$U_+^{(k)} = \begin{array}{c} \\ (k,0) \\ (k,1) \\ (k,2) \\ (k,3) \\ \vdots \\ (k,M_2) \end{array} \begin{array}{c} \begin{array}{ccccc} (k+1,0) & (k+1,1) & (k+1,2) & \cdots & (k+1,M_2) \end{array} \\ \begin{bmatrix} \lambda_k & \cdot & \cdot & \cdots & \cdot \\ \lambda_{k+1} & \cdot & \cdot & \cdots & \cdot \\ \lambda_{k+2} & \cdot & \cdot & \cdots & \cdot \\ \lambda_{k+3} & \cdot & \cdot & \cdots & \cdot \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{k+M_2} & \cdot & \cdot & \cdots & \cdot \end{bmatrix} \end{array},$$

and $U_+^{(M_1)}$ is an $M_2 \times M_2$ matrix of zeros.

For $1 \leq k \leq M_1 - 1$,

$$U_0^{(k)} = \begin{array}{c} \\ (k,0) \\ (k,1) \\ (k,2) \\ (k,3) \\ \vdots \\ (k,M_2-1) \\ (k,M_2) \end{array} \begin{array}{c} \begin{array}{ccccccc} (k,0) & (k,1) & (k,2) & (k,3) & \cdots & (k,M_2-1) & (k,M_2) \end{array} \\ \begin{bmatrix} q_k & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & q_{k+1} & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & q_{k+2} & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & q_{k+3} & \cdots & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \cdot & \cdot & \cdot & \cdot & \cdots & q_{k+M_2-1} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & -\lambda_{k+M_2} \end{bmatrix} \end{array},$$

and

$$U_0^{(0)} = \begin{array}{c} \\ (0,0) \\ (0,1) \\ (0,2) \\ (0,3) \\ \vdots \\ (0,M_2) \end{array} \begin{array}{c} \begin{array}{ccccccc} (0,0) & (0,1) & (0,2) & (0,3) & \cdots & (0,M_2) \end{array} \\ \begin{bmatrix} -\lambda_0 & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & -\lambda_1 & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & -\lambda_2 & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & -\lambda_3 & \cdots & \cdot \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdot & \cdot & \cdots & -\lambda_{M_2} \end{bmatrix} \end{array},$$

and

$$U_0^{(M_1)} = \begin{array}{c} \\ (M_1,0) \\ (M_1,1) \\ (M_1,2) \\ (M_1,3) \\ \vdots \\ (M_1,M_2-1) \\ (M_1,M_2) \end{array} \begin{array}{c} \begin{array}{ccccccc} (M_1,0) & (M_1,1) & (M_1,2) & (M_1,3) & \cdots & (M1,M_2-1) & (M_1,M_2) \end{array} \\ \begin{bmatrix} -\mu & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & -\mu & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & -\mu & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & -\mu & \cdots & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \cdot & \cdot & \cdot & \cdot & \cdots & -\mu & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & 0 \end{bmatrix} \end{array},$$

with $q_k = -(\mu + \lambda_k), k \geq 1$.

Then, the matrices A_+, A_-, A_0 and B_0 are given by

$$A_+ = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \\ m(n,M_1-1) \\ m(n,M_1) \end{array} \begin{array}{c} \begin{array}{cccccc} m(n+1,0) & m(n+1,1) & m(n+1,2) & m(n+1,3) & \cdots & m(n+1,M_1) \end{array} \\ \begin{bmatrix} \mathbf{0} & U_+^{(0)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & U_+^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & U_+^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & U_+^{(M_1-1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \end{array},$$

$$A_- = \begin{array}{c} \\ (n,0,0) \\ (n,0,1) \\ (n,0,2) \\ (n,0,3) \\ \vdots \\ (n,M_1,M_2) \end{array} \begin{array}{c} \begin{array}{cccccc} (n-1,0,0) & (n-1,0,1) & (n-1,0,2) & (n-1,0,3) & \cdots & (n-1,M_1,M_2) \end{array} \\ \begin{bmatrix} \mu & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \mu & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \mu & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \mu & \cdots & \cdot \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdot & \cdot & \cdots & \mu \end{bmatrix} \end{array},$$

$$A_0 = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \\ m(n,M_1) \end{array} \begin{array}{c} \begin{array}{cccccc} m(n,0) & m(n,1) & m(n,2) & \cdots & m(n,M_1-1) & m(n,M_1) \end{array} \\ \begin{bmatrix} V_0^{(0)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ U_- & V_0^{(1)} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & U_- & V_0^{(2)} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & U_- & V_0^{(M_1)} \end{bmatrix} \end{array},$$

and

$$B_0 = \begin{array}{c} \\ m(0,0) \\ m(0,1) \\ m(0,2) \\ \vdots \\ m(0,M_1) \end{array} \begin{array}{c} \begin{array}{cccccc} m(0,0) & m(0,1) & m(0,2) & \cdots & m(0,M_1-1) & m(n,M_1) \end{array} \\ \begin{bmatrix} U_0^{(0)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ U_- & U_0^{(1)} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & U_- & U_0^{(2)} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & U_- & U_0^{(M_1)} \end{bmatrix} \end{array}.$$

The semi-experiment transition matrix is still given by

$$Q = \begin{array}{c} \\ \ell(0) \\ \ell(1) \\ \ell(2) \\ \ell(3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} \ell(0) & \ell(1) & \ell(2) & \ell(3) & \cdots \end{array} \\ \begin{bmatrix} B_0 & A_+ & \mathbf{0} & \mathbf{0} & \cdots \\ A_- & A_0 & A_+ & \mathbf{0} & \cdots \\ \mathbf{0} & A_- & A_0 & A_+ & \cdots \\ \mathbf{0} & \mathbf{0} & A_- & A_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix} \end{array}.$$

5.4.2 Kronecker Product Notation

These matrices can be expressed more concisely by using Kronecker products.

For $0 \leq k \leq M_1 - 1$, let $W_+^{(k)}$ be $(M_1+1) \times (M_1+1)$ matrices of zeros with a 1 in the $(k+1, k+2)$th entry. Let $W_+^{(M_1)}$ be an $(M_1+1) \times (M_1+1)$ matrix of zeros.

Then we can write the truncated $(M_1+1)(M_2+1) \times (M_1+1)(M_2+1)$ matrix A_+ as

$$A_+ = \sum_{k=0}^{M_1} W_+^{(k)} \otimes U_+^{(k)}.$$

The truncated A_- is simply μ times the $(M_1+1)(M_2+1) \times (M_1+1)(M_2+1)$ identity matrix.

For $0 \leq k \leq M_1$, let $W_0^{(k)}$ be $(M_1+1) \times (M_1+1)$ matrices of zeros with a 1 in the $(i+1, i+1)$th entry. Let W_0^* be an $(M_1+1) \times (M_1+1)$ matrix of zeros with 1's on the lower diagonal.

Then we can write the truncated $(M_1+1)(M_2+1) \times (M_1+1)(M_2+1)$ matrix A_0 as

$$A_0 = W_0^* \otimes U_- + \sum_{k=0}^{M_1} W_0^{(k)} \otimes V_0^{(k)},$$

and the truncated $(M_1+1)(M_2+1) \times (M_1+1)(M_2+1)$ matrix B_0 can be written as

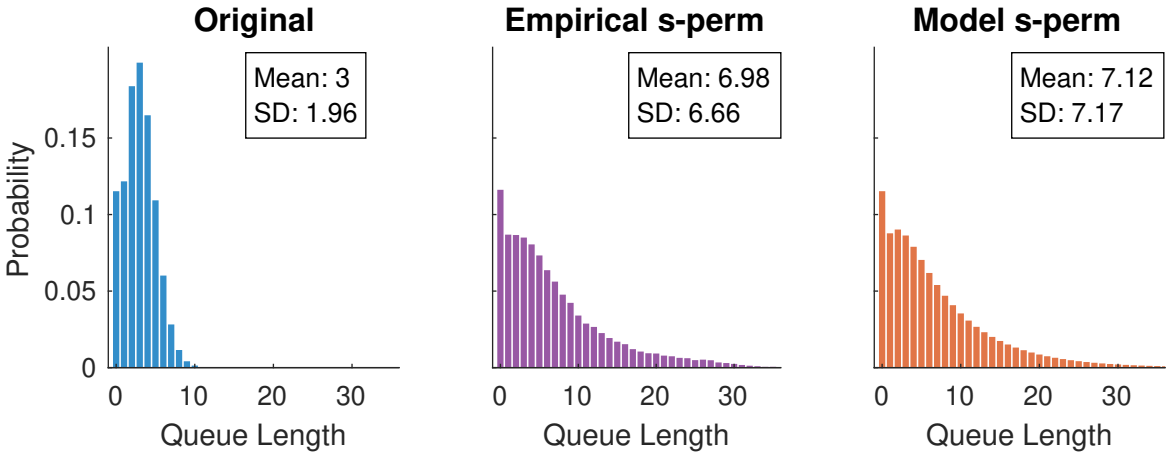
$$B_0 = W_0^* \otimes U_- + \sum_{k=0}^{M_1} W_0^{(k)} \otimes U_0^{(k)}.$$

5.5 Results

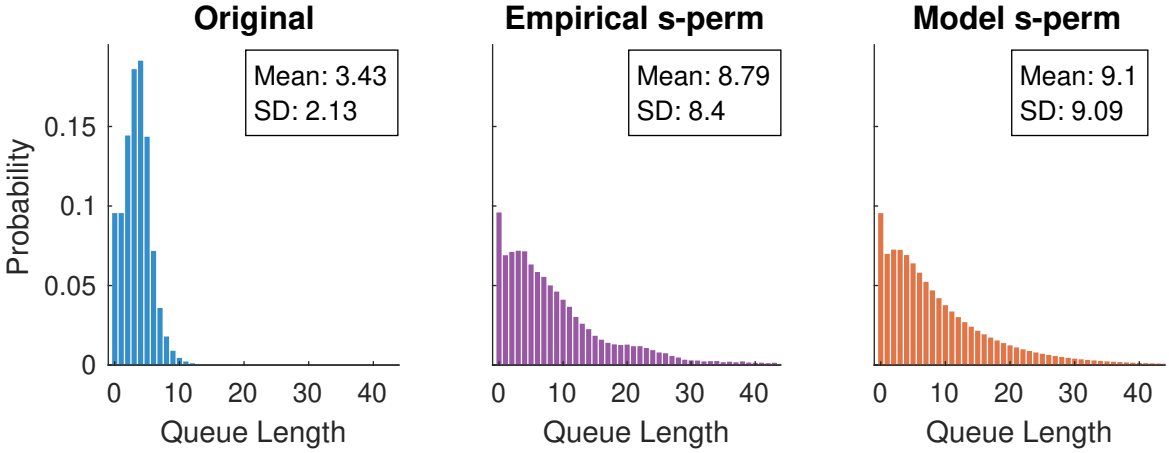
This new s-perm semi-experiment model accurately captures the features of the empirical semi-experiment (particularly in queue-length distribution). Figure 5.5 supports this by comparing the original, empirical s-perm and model s-perm queue-length distributions for different parameters.

Note that the effect of the s-perm semi-experiment on this queue is less extreme than that on the queue-length-dependent service rates queue in Chapter 4. Recall that in that chapter, the s-perm semi-experiment was an $M/G/1$ queue with no dependence between the arrival and service processes, giving the queue-length distribution an exponential shape. In this case, the s-perm semi-experiments do not have such a strong effect, as they still retain a 'hump' at lower queue-lengths and do not have an exponential shape. This is because the dependence is in the arrival process and so permuting service times does not disrupt the dependence as completely (since some dependence remained in the queueing data). In the s-perm model, a dependence structure is present in the arrival process, explaining some of this retained dependence in the s-perm semi-experiment queue.

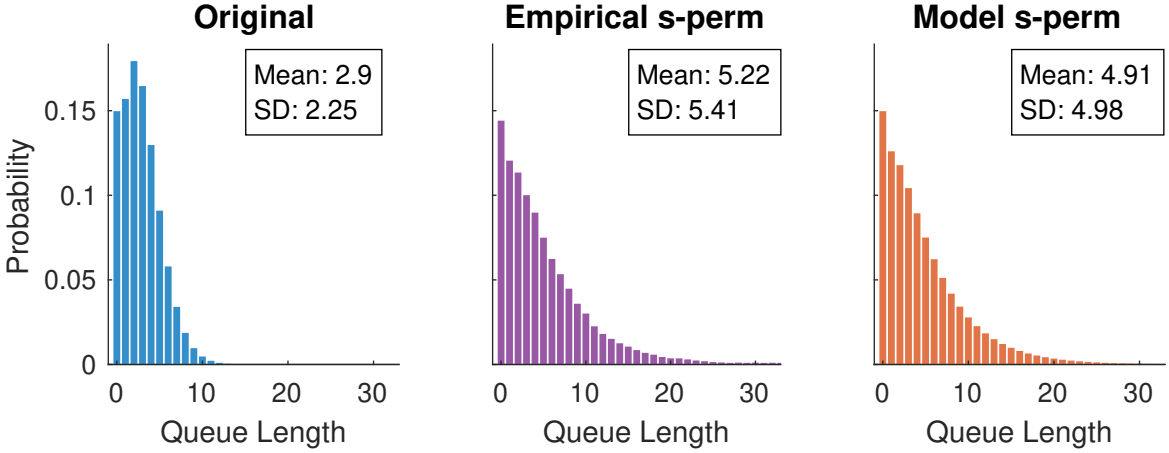
These differences in the effect of semi-experiments can be used to identify the type of dependence in the original queue. This is explored in Chapter 9.



(a) Arrival rate of $\lambda_i = 6/i$. The KS statistic between the original and semi-experiment models is 0.370250.



(b) Arrival rate of $\lambda_i = 6 - i$ for $i \leq 4$ and $\lambda_i = 1$ for $i \geq 5$. The KS statistic between the original and semi-experiment models is 0.429500.



(c) Service rate of $\lambda_i = 3/\sqrt{i}$. The KS statistic between the original and semi-experiment models is 0.204800.

Figure 5.5: The blue and orange plots show the stationary queue-length distribution of the original (QL, A, O) model and s-perm semi-experiment (QL, A, sSE) model, respectively. The purple plots show the empirical queue length distribution of an empirical s-perm semi-experiment. That is, a single realisation of the original queue (run for 20,000 time steps) with a single random permutation of the service times. Each queue has the service rate $\mu = 2$.

5.6 Conclusion

In this chapter we considered a queue-length-dependent queue where the arrival rates depend on the queue-length immediately after the previous arrival (that is, at the start of the inter-arrival period). We constructed a level-dependent QBD model to find the stationary queue-length distribution for this queue. Then we sought a QBD model for the s-perm semi-experiment applied to this queue by analysing the DTMC embedded at arrival epochs, using a similar method to that used in Chapter 4. This was found to not accurately model the semi-experiment queue due to a dependence on the actual inter-arrival times and not just the arrival rates. We then proposed a new model for the s-perm semi-experiment using an innovative method, an 'underlying' queue, to create an arrival process equivalent to the original queue and then adding independent service times. This new QBD model accurately captured the behaviour of the s-perm semi-experiment applied to this queue. We then compared the original and semi-experiment stationary queue-length distributions to observe the dependence behaviour.

Chapter 6

a-perm Semi-Experiments for Queue-Length-Dependent Rates

Now we consider an alternate semi-experiment in which the service times stream is preserved and the inter-arrival times are randomly permuted (a-perm semi-experiment). This semi-experiment will be applied to both the queue-length-dependent service rates queue, (QL, S, O), and the queue-length-dependent arrival rates queue, (QL, A, O).

6.1 Queue-Length-Dependent Arrival Rates

For this section, the original queue is the (QL, A, O) queue from Chapter 5. This is a single-server queue with constant service rates μ and queue-length-dependent arrival rates λ_j where $j \geq 1$ is the length of the queue immediately after the previous arrival.

6.1.1 a-perm Semi-Experiment Model

Now we can apply an a-perm semi-experiment to this model in which the service stream is retained and the inter-arrival times are randomly permuted.

The construction of the QBD model for the a-perm semi-experiment in this case is very similar to the (QL, S, sSE) QBD model. The a-perm semi-

experiment randomly permutes the inter-arrival times while keeping the same sequence of service times. Hence, the service times for this model are simply exponentially distributed with constant rate μ. The inter-arrival times no longer depend on the queue length at arrivals. So each inter-arrival time in the inter-arrival semi-experiment is exponentially distributed and has the rate λ_i with probability given by that of observing an arrival rate λ_i in the original queue. That is, the inter-arrival times have a hyperexponential distribution with rate parameters $\{\lambda_j : j \geq 1\}$ and mixture distribution $\{\widetilde{\pi}_j : j \geq 1\}$, where $\widetilde{\pi}_j$ is the long-term probability that the queue length immediately after an arrival is j in the original queue.

In Section 5.2, the DTMC embedded at epochs immediately after arrivals was defined to have state space $\{\widetilde{Q}^{(m)} : m \geq 1\}$ and probability transition matrix

$$\widetilde{P} = \begin{bmatrix} \widetilde{p}_{11} & \widetilde{p}_{12} & 0 & 0 & \cdots \\ \widetilde{p}_{21} & \widetilde{p}_{22} & \widetilde{p}_{23} & 0 & \cdots \\ \widetilde{p}_{31} & \widetilde{p}_{32} & \widetilde{p}_{33} & \widetilde{p}_{34} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where

$$\widetilde{p}_{ij} = \begin{cases} \dfrac{\mu^{i-j+1}\lambda_i}{(\mu+\lambda_i)^{i-j+2}}, & \text{for } 2 \leq j \leq i+1, \\ \dfrac{\mu^i}{(\mu+\lambda_i)^i}, & \text{for } j = 1, i \geq 1. \end{cases}$$

Now, $\{\widetilde{\boldsymbol{\pi}} : j \geq 1\}$ is the stationary distribution for this DTMC. This can be calculated as shown in Section 2.2.4.

Hence, the (QL, S, aSE) queue is defined as a single-server queue such that the

- **Arrival Process** has inter-arrival times with a hyperexponential distribution with rates λ_j for $j \geq 1$ and mixture distribution given by $\widetilde{\pi}_j$ for $j \geq 1$.
- **Service Times** are exponentially distributed with constant rate μ.

We summarise the transition rates for the a-perm semi-experiment model for queue-length-dependent arrival rates in Table 6.1.

From	To	Rate	For
(n,i)	$(n+1,j)$	$\lambda_i\widetilde{\pi}_j$	$(n,i)\in S, 1\leq j\leq i+1$
(n,i)	$(n-1,i)$	μ	$n\geq 1, i\geq 1$

Table 6.1: Transition rates for the (QL, A, aSE) process.

Now we can construct a QBD model for this process. Let $\{X(t) : t \geq 0\}$ be a Markov chain with state space $S = \{(n,i) : n \geq 0, i \geq 1\}$, where n represents the current queue length and i is an indicator for the rate of the next arrival. Then the block matrices for the transition matrix are given by

$$A_+ = \begin{bmatrix} \lambda_1\widetilde{\pi}_1 & \lambda_1\widetilde{\pi}_2 & \lambda_1\widetilde{\pi}_3 & \cdots \\ \lambda_2\widetilde{\pi}_1 & \lambda_2\widetilde{\pi}_2 & \lambda_2\widetilde{\pi}_3 & \cdots \\ \lambda_3\widetilde{\pi}_1 & \lambda_3\widetilde{\pi}_2 & \lambda_3\widetilde{\pi}_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \qquad A_- = \begin{bmatrix} \mu & 0 & 0 & \cdots \\ 0 & \mu & 0 & \cdots \\ 0 & 0 & \mu & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$A_0 = \begin{bmatrix} q_1 & 0 & 0 & \cdots \\ 0 & q_2 & 0 & \cdots \\ 0 & 0 & q_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \qquad B_0 = \begin{bmatrix} -\lambda_1 & 0 & 0 & \cdots \\ 0 & -\lambda_2 & 0 & \cdots \\ 0 & 0 & -\lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where $q_i = -(\lambda_i + \mu), i \geq 1$.

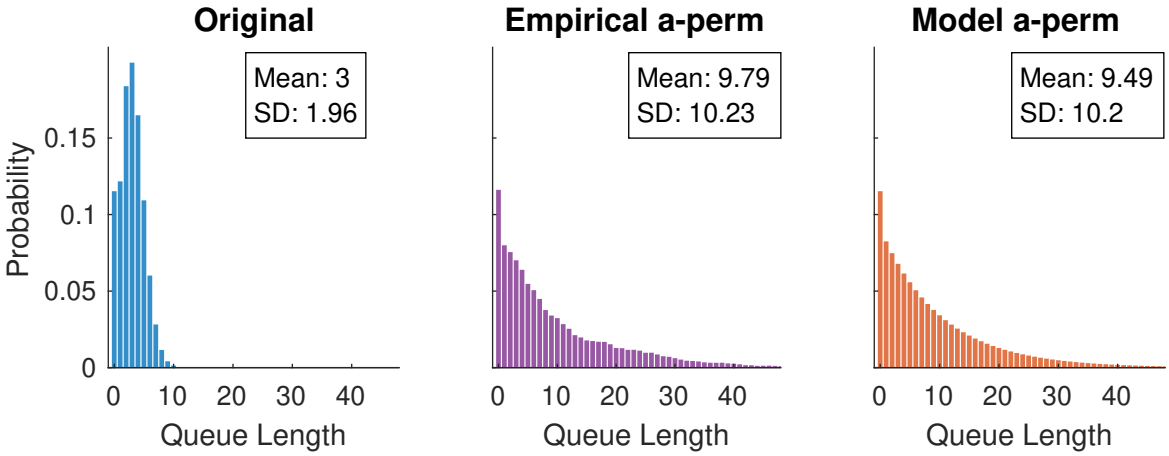
6.1.2 Truncation and Augmentation

Once again, the matrices A_+, A_-, A_0, B_0 are infinitely large in size and so need to be truncated and augmented appropriately. Let M be the imposed upper bound on the phase space (equivalently, the set of possible arrival rates) such that $S = \{(n,i) : n \geq 0, 1 \leq i \leq M\}$. Then each of the matrices are truncated to $M \times M$ and we augment $\{\widetilde{\pi}_j : j \geq 1\}$ so that $\widetilde{\pi}_M = 1 - \sum_{i=1}^{M-1} \widetilde{\pi}_i$.

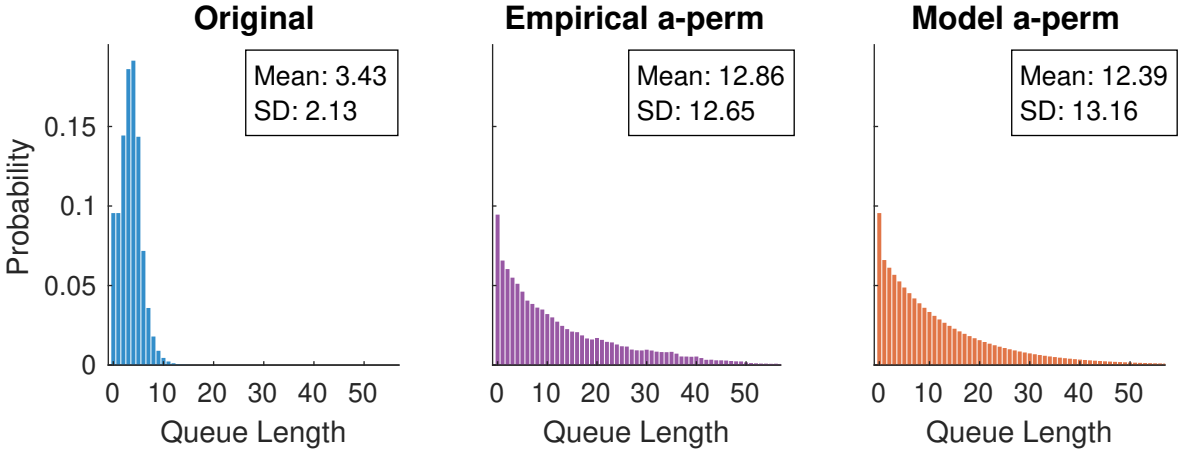
6.1.3 Results

Figure 6.1 shows the original, empirical a-perm and model a-perm queue-length distributions for different parameters. The a-perm semi-experiments completely

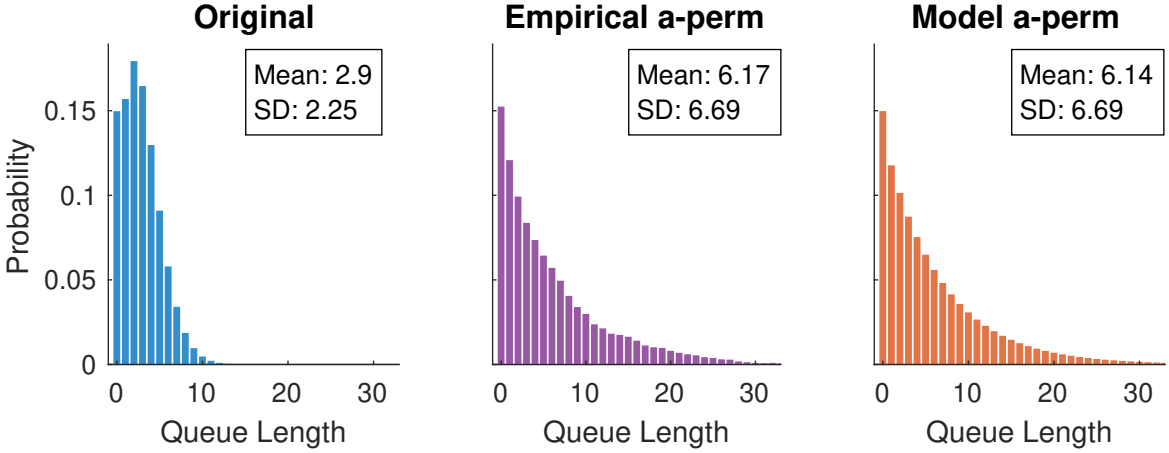
disrupt the dependence in the arrival process, compared to the s-perm semi-experiments in Chapter 5. Hence, the a-perm semi-experiment queue-length distributions have an exponential shape, indicating there is no dependence between the arrival and service processes.



(a) Arrival rate of $\lambda_i = 6/i$. The KS statistic between the original semi-experiment models is 0.448750.



(b) Arrival rate of $\lambda_i = 6 - i$ for $i \leq 4$ and $\lambda_i = 1$ for $i \geq 5$. The KS statistic between the original and semi-experiment models is 0.498050.



(c) Service rate of $\lambda_i = 3/\sqrt{i}$. The KS statistic between the original and semi-experiment models is 0.279850.

Figure 6.1: The blue and orange plots show the stationary queue-length distribution of the original (QL, A, O) model and the a-perm semi-experiment (QL, A, aSE) model, respectively. The purple plots show the empirical queue length distribution of an empirical a-perm semi-experiment. That is, a single realisation of the original queue (run for 20,000 time steps) with a single random permutation of the inter-arrival times. Each queue has the service rate $\mu = 2$.

6.2 Queue-Length-Dependent Service Rates

Now consider the original queue to be the (QL, S, O) queue. That is, a single-server queue with a constant arrival rate, λ, and queue-length-dependent service rates, μ_i, where i is the length of the queue at the time when service begins.

6.2.1 a-perm Semi-experiment Model

The construction of this semi-experiment queue and QBD model is very similar to the (QL, A, sSE) model.

When the inter-arrival times are randomly permuted, they are still independent exponentially distributed times with constant rate λ.

After the permutation the service rates no longer depend directly on the queue length at the start of services. However, since the order of service *times* is preserved some dependence structure is maintained. For example, if a particular service *time* was long (irrespective of the associated service rate), then there is likely to be relatively more arrivals during that time. Similarly, during a short service *time* there is likely to be less arrivals. This affects the queue length at the next service, which informs the next service *rate*. Therefore, the whole process depends on the sequence of actual service *times*, and not simply the sequence of service *rates*. Hence, to model the service times for the inter-arrival semi-experiment, we need to replicate the original process that produces them.

We propose a model in which we track two queues: the underlying queue that is almost identical to the original queue, and the a-perm semi-experiment queue. These two queues have independent arrival processes, which are Poisson processes with the constant rate λ. However, they share the same sequence of service times, generated by the underlying queue. That is, when a service occurs in the underlying queue at rate μ_i, the same service will also occur in the inter-arrival semi-experiment queue.

Empty Queue Issues

There are a few issues with this model that need to be addressed, and so this proposed model will be modified.

If the underlying queue becomes empty, but the semi-experiment queue was not empty, then it would continue needing to have services, but the underlying queue could not provide the appropriate sequence of service rates as there would be no departures. Hence, we require that the underlying queue is never empty. If a service is completed when the underlying queue has a length of 1, then the queue skips to the next busy period which occurs after the next arrival. Hence, the queue length remains at 1. This ensures that it is always possible for the underlying queue to generate service times for the semi-experiment queue, and skipping over the underlying queue's idle periods does not affect the sequence of service times.

If the semi-experiment queue becomes empty while the underlying queue is busy, then the underlying queue would continue to have services and departures while the semi-experiment queue would not. This means that when the semi-experiment queue becomes busy again, the sequence of service times would be inaccurate as some may have been missed. Therefore, we require that when the semi-experiment queue is empty, the underlying queue is 'paused'. That is, it remains in the same state and has no events occur during the semi-experiment queue's idle period. This means that the semi-experiment queue does not miss any services in the underlying queue, and this pausing does not change the sequence of service times.

The underlying queue is only important to the model for generating the required sequence of service times, and any other behaviour of the queue is irrelevant. Therefore, these adjustments do not disrupt its main purpose.

Note that similar adjustments were not needed in the construction of the (QL, A, sSE) model since the dependence was with the arrival rates which are not affected by the boundary condition on queue lengths in the way that the service times are.

So this queue is the a-perm semi-experiment on the queue-length-dependent service rates model. It is a single-server queue such that the:

- **Arrival Process** is a Poisson process with constant rate λ.
- **Service Times** are the sequence of service times generated by the modified underlying queue.

We label this queue as (QL, S, aSE).

Let the state for this model be $(n, (k, i))$, where

- n is the queue length of the a-perm semi-experiment process,
- k is the queue length of the modified underlying original process,
- i is the queue length at the start of the last service in the modified underlying original process.

Now, the state space for this model is $S = \{(n,(k,i)) : n \geq 0, k \geq 1, 1 \leq i \leq k\}$. Let $m(n,k) = \{(n,k,i) : k \geq 1, 1 \leq i \leq k\}$ and $\ell(n) = \{(n,k,i) : 1 \leq k \leq i\} = \bigcup_{k\geq 1} m(n,k)$ so that $S = \bigcup_{n\geq 0} \ell(n) = \bigcup_{n\geq 0}\bigcup_{k\geq 1} m(n,k)$. Table 6.2 shows the possible transitions for this process.

From	To	Rate	For	Description
$(n,(k,i))$	$(n+1,(k,i))$	λ	$(n,(k,i)) \in S$	SE arrival
$(n,(k,i))$	$(n,(k+1,i))$	λ	$(n,(k,i)) \in S$	Underlying arrival
$(n,(k,i))$	$(n-1,(k-1,k-1))$	μ_i	$n \geq 1, k \geq 2$	Underlying and SE departure
$(n,1,1)$	$(n-1,1,1)$	μ_1	$n \geq 1$	Underlying and SE departure

Table 6.2: Possible transitions for the (QL,S,aSE) process.

First consider the modified underlying process which is almost identical to the process (QL,S,O) except that there is no level 0. It has state space $S_U = \{(k,i) : k \geq 1, 1 \leq i \leq k\}$ and the transition matrix

$$Q_U = \begin{array}{c} \\ m(n,1) \\ m(n,2) \\ m(n,3) \\ m(n,4) \\ m(n,5) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} m(n,1) & m(n,2) & m(n,3) & m(n,4) & m(n,5) & \cdots \end{array} \\ \begin{bmatrix} U_0^{(1)} & U_+^{(1)} & \cdot & \cdot & \cdot & \cdots \\ U_-^{(2)} & U_0^{(2)} & U_+^{(2)} & \cdot & \cdot & \cdots \\ \cdot & U_-^{(3)} & U_0^{(3)} & U_+^{(3)} & \cdot & \cdots \\ \cdot & \cdot & U_-^{(4)} & U_0^{(4)} & U_+^{(4)} & \cdots \\ \cdot & \cdot & \cdot & U_-^{(5)} & U_0^{(5)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

For $k \geq 2$,

$$U_-^{(k)} = \begin{bmatrix} 0 & \cdots & 0 & \mu_1 \\ 0 & \cdots & 0 & \mu_2 \\ 0 & \cdots & 0 & \mu_3 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \mu_k \end{bmatrix},$$

and for $k \geq 1$,

$$U_0^{(k)} = \begin{bmatrix} q_1 & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & q_2 & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & q_3 & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & q_4 & \cdots & \cdot \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdot & \cdot & \cdots & q_k \end{bmatrix}, \qquad U_+^{(k)} = \begin{bmatrix} \lambda & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \lambda & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \lambda & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \ddots & \cdot & \cdot \\ \vdots & \vdots & \vdots & \cdots & \lambda & \cdot \end{bmatrix},$$

where $q_k = -(\lambda + \mu_k)$. The matrices $U_-^{(k)}, U_0^{(k)}, U_+^{(k)}$ have sizes $k \times (k-1), k \times k, k \times (k+1)$, respectively.

We also define $U_-^{(1)} = \mu_1$, which is not explicitly used in the non-modified underlying process, but will be needed to define the semi-experiment process.

Now consider the semi-experiment queue. First, the block matrix A_+ contains all the rates in which the semi-experiment queue increases its queue length by 1 ($n \to n+1$). When this occurs there is no change to the phase (k, i), as shown in Table 6.2. The transition rate is λ, independent of the phase (k, i). Hence,

$$A_+ = \begin{bmatrix} \lambda & \cdot & \cdot & \cdot & \cdots \\ \cdot & \lambda & \cdot & \cdot & \cdots \\ \cdot & \cdot & \lambda & \cdot & \cdots \\ \cdot & \cdot & \cdot & \lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

The block matrix A_- contains all the transition rates such that $n \to n-1$. When this occurs, the underlying departure also occurs, causing the phase change $(k, i) \to (k-1, k-1)$ (Table 6.2). The rates of this phase change $i \to k-1$ are contained in the matrices $U_-^{(k)}$. These matrices are placed in the lower diagonal of A_- to create the transition $k \to k-1$. Note that when $k = 1$ and there is a departure from the semi-experiment queue, $n \to n-1$, then the phase change does not change from $(1, 1)$. This transition happens with rate $U_-^{(1)} = \mu_1$. Hence,

$$A_- = \begin{array}{c} \\ m(n,1) \\ m(n,2) \\ m(n,3) \\ m(n,4) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} m(n-1,1) & m(n-1,2) & m(n-1,3) & \cdots \end{array} \\ \begin{bmatrix} U_-^{(1)} & \mathbf{0} & \mathbf{0} & \cdots \\ U_-^{(2)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & U_-^{(3)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & U_-^{(4)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

The block matrix A_0 contains all the transition rates such that $n \to n, n \geq 1$. This includes the underlying arrival phase change $(k, i) \to (k+1, i)$, the transition rates of which are contained within the matrices $U_+^{(k)}$. These are placed in the upper diagonal of A_0. The diagonal elements of A_0 contain the negative sum of all the transition rates out of each phase. Let $V_0^{(k)}$ for $k \geq 1$ be the $k \times k$ block matrices for the diagonal of A_0. Then,

$$V_0^{(k)} = U_0^{(k)} - \lambda \mathcal{I}_k,$$

since $U_0^{(k)}$ contains the negative sum of transition rates for the modified underlying process, and so we need to subtract an additional λ from each phase for the arrivals in the overall semi-experiment process.

So,

$$A_0 = \begin{array}{c} \\ m(n,1) \\ m(n,2) \\ m(n,3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} m(n,1) & m(n,2) & m(n,3) & m(n,4) & \cdots \end{array} \\ \begin{bmatrix} V_0^{(1)} & U_+^{(1)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & V_0^{(2)} & U_+^{(2)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & V_0^{(3)} & U_+^{(3)} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix} \end{array}.$$

Finally, consider the matrix B_0 which contains the transition rates such that $n \to n$ when $n = 0$. In this case the modified underlying process is 'paused' and so there are no transitions for the phase (k, i) and B_0 only contains the negative transition rates of the arrival transitions in the semi-experiment queue since there are also no departures in the semi-experiment queue when $n = 0$. Hence,

$$B_0 = -A_+.$$

Then, the overall transition matrix for the (QL, S, aSE) process is given by

$$Q = \begin{array}{c} \\ \ell(0) \\ \ell(1) \\ \ell(2) \\ \ell(3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} \ell(0) & \ell(1) & \ell(2) & \ell(3) & \cdots \end{array} \\ \begin{bmatrix} B_0 & A_+ & \mathbf{0} & \mathbf{0} & \cdots \\ A_- & A_0 & A_+ & \mathbf{0} & \cdots \\ \mathbf{0} & A_- & A_0 & A_+ & \cdots \\ \mathbf{0} & \mathbf{0} & A_- & A_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix} \end{array}.$$

6.2.2 Truncation and Augmentation

The phase space for this model takes infinitely many values and so the block matrices A_+, A_0, A_-, B_0 are not finite. As with previous models, in order to practically implement this QBD, these matrices need to be truncated and augmented. Let M be a large artificial maximum such that the process is truncated to the state space $S = \{(n,(k,i)) : n \geq 0, 1 \leq i \leq k \leq M\}$, representing a truncation of the underlying queue length to be at most M. Hence, each of the sizes of the matrices A_+, A_0, A_-, B_0 is $\sum_{i=1}^{M} i \times \sum_{i=1}^{M} i$ or $M(M+1)/2 \times M(M+1)/2$. To augment this truncated process, the underlying queue length k cannot exceed M. Hence, if $k = M$ there are no arrivals in the underlying process. Therefore, the truncated and augmented matrices are defined below.

The matrices $U_-^{(k)}$ for $1 \leq k \leq M$ are unchanged. For $1 \leq k \leq M$ the matrices $U_0^{(k)}$ and U_+ are also unchanged. Since there cannot be arrivals when $k = M$,

$$U_+^{(M)} = \mathbf{0}_{M\times M} \quad \text{and} \quad U_0^{(M)} = \begin{bmatrix} -\mu_1 & \cdot & \cdot & \cdots & \cdot \\ \cdot & -\mu_2 & \cdot & \cdots & \cdot \\ \cdot & \cdot & -\mu_3 & \cdots & \cdot \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdot & \cdots & -\mu_M \end{bmatrix}.$$

A_+ is λ times the identity matrix of size $M(M+1)/2 \times M(M+1)/2$ and $B_0 = -A_+$.

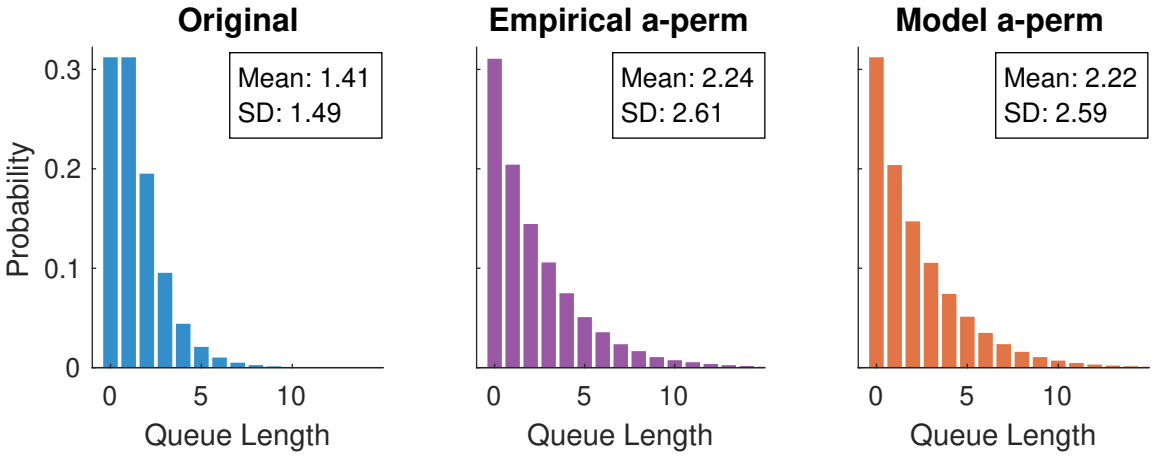
$$A_- = \begin{array}{c} \\ m(n,1) \\ m(n,2) \\ m(n,3) \\ m(n,4) \\ \vdots \\ m(n,M) \end{array} \begin{array}{c} \begin{array}{cccccc} m(n{-}1,1) & m(n{-}1,2) & m(n{-}1,3) & \cdots & m(n{-}1,M{-}1) & m(n{-}1,M) \end{array} \\ \begin{bmatrix} U_-^{(1)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ U_-^{(2)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & U_-^{(3)} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & U_-^{(4)} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & U_-^{(M)} & \mathbf{0} \end{bmatrix} \end{array},$$

and

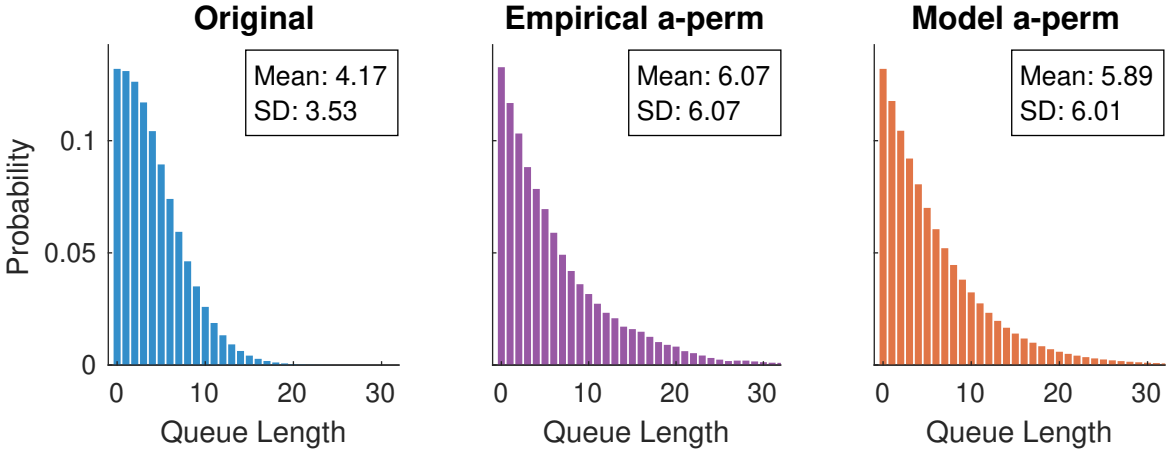
$$A_0 = \begin{array}{c} \\ m(n,1) \\ m(n,2) \\ m(n,3) \\ \vdots \\ m(n,M-1) \\ m(n,M) \end{array} \begin{array}{c} \begin{array}{ccccccc} m(n,1) & m(n,2) & m(n,3) & m(n,4) & \cdots & m(n,M-1) & m(n,M) \end{array} \\ \begin{bmatrix} V_0^{(1)} & U_+ & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & V_0^{(2)} & U_+ & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & V_0^{(3)} & U_+ & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & V_0^{(M-1)} & U_+ \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & V_0^{(M)} \end{bmatrix} \end{array}.$$

6.2.3 Results

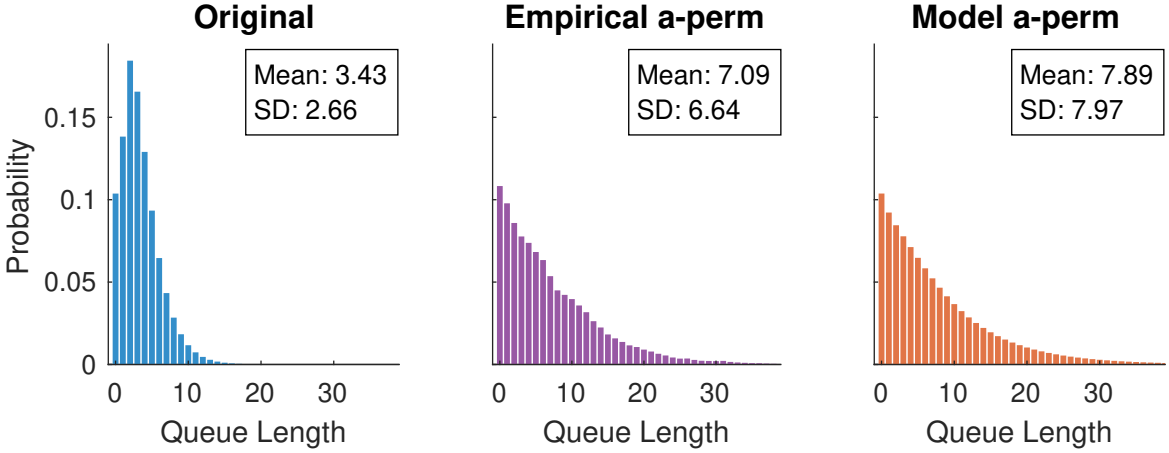
Figure 6.2 shows the original, empirical a-perm and model a-perm queue-length distributions for different parameters. The a-perm semi-experiment should not as completely disrupt the dependence as the s-perm semi-experiment since the dependence is in the service stream. However, it is clear that the a-perm semi-experiment queue-length distribution has a more exponential shape than the analogue of the (QL, A, sSE) queue. This is due to the difference in the nature of queue-length-dependent services and queue-length-dependent arrivals in a queue. In the (QL, A, O) queue, a customer's arrival rate depends on the queue-length at the previous arrival, so it is completely independent of that customer's service time. So permuting the service times in the (QL, A, sSE) does not have a strong effect on the dependence structure in the arrival process. In the (QL, S, O) queue, a customer's service rate depends on the the queue-length when service begins. However, this queue-length at service is related to the queue-length at that customer's arrival due to the nature of the queue-length process. Hence, the arrival process is related to the length of the queue at the start of each service, and hence the service rates. So permuting the inter-arrival times in the (QL, S, aSE) queue more effectively disrupts the dependence structure in the service process. This difference between dependence in arrivals and dependence in services is more thoroughly explored in Section 8.5.1.



(a) Service rate of $\mu_i = 2i^2$. The KS statistic between the original and semi-experiment models is 0.151800.



(b) Service rate of $\mu_i = \log(i + 6.5)$. The KS statistic between the original and semi-experiment models is 0.124300.



(c) Service rate of $\mu_i = 1.5$, if $i \leq 2$ and $\mu_i = 3.5$ if $i \geq 3$. The KS statistic between the original and semi-experiment models is 0.325450.

Figure 6.2: The blue and orange plots show the stationary queue-length distribution of the original (QL, S, O) model and the a-perm semi-experiment (QL, S, aSE) model, respectively. The purple plots show the empirical queue length distribution of an empirical a-perm semi-experiment. That is, a single realisation of the original queue (run for 20,000 time steps) with a single random permutation of the inter-arrival times. Each queue has the arrival rate $\lambda = 2$.

6.2.4 Conclusion

In this chapter we took the queue-length-dependent service rates model (QL, S, O) from Chapter 4 and the queue-length-dependent arrival rates model (QL, A, O) from Chapter 5 and applied the a-perm semi-experiments to them. Since the dependence structure for the (QL, A, O) model is only within the arrival process, applying the a-perm semi-experiment is very similar to applying the s-perm semi-experiment to the (QL, S, O) which has a dependence structure only in the service process. Hence, we constructed a QBD model for the a-perm semi-experiment applied to the queue-length-dependent arrival rates model by finding the stationary distribution of the embedded DTMC at arrival epochs to get the unconditional probability distribution for arrival rates after the inter-arrival times are permuted. The stationary queue-length distribution for this model was compared to the original and shown to have a stronger and more complete disruption of the dependence compared to the s-perm semi-experiment in Chapter 5. Then we applied the a-perm semi-experiment to the (QL, S, O). This is modelled very similarly to the (QL, A, sSE) model by using an underlying queue to replicate the service process of the original queue in a QBD model and adding an independent arrival process. There were some extra considerations for this underlying queue. We needed a continuous stream of available service times when the semi-experiment queue is non-empty. Hence, we do not allow the underlying queue to be empty, so the service process is not paused, and when the semi-experiment queue is empty, we pause the service process in the underlying queue. We compared the stationary queue-length distributions of the original and a-perm semi-experiment queues, showing a different effect than observed in the s-perm semi-experiment.

Chapter 7

Queue-Length-Dependent Arrival and Service Rates

7.1 Original Model

We now consider a model with both queue-length-dependent arrival rates and service rates. This is a natural combination of the previous two queue-length-dependent queueing models. We consider a single-server queue with arrival rates dependent upon the queue length immediately after the previous arrival, and service rates dependent upon the queue length at the start of the service period. More formally,

- **Arrival Process** has inter-arrival times which are exponentially distributed with rate λ_{i^*}, where $i^* \geq 1$ is the queue length immediately after the previous arrival (that is, at the start of the current inter-arrival period).
- **Service Times** are exponentially distributed with constant rate μ_j, where $j \geq 1$ is the queue length at the start of the current service period.

This process is labelled as the (QL, AS, O) queue.

This process can be modelled as a CTMC. Let the state for this process be $(n, (i, j))$, where

- n is the current queue length

- i is the number of departures since the last arrival. Note that we can use this to determine arrival rates using $i^* = n + i$

- j is the queue length at the start of the current service, when the queue is busy ($n \geq 1$), or it is 1 when the queue is idle ($n = 0$). This choice has been made because when the queue is idle, the queue length at the start of the next service will always be 1.

The state space is given by $S = \{(n, i, j) : n \geq 0, 1 \leq j \leq n, i \geq 0, n+i \geq 1\}$. If n is considered the level and (i, j) is considered the phase, then this process is a queue-length-dependent QBD. For $n \geq 1$, let $m(n, i) = \{(n, i, j) : 1 \leq j \leq n\}$ and $m(0, i) = \{(0, i, 1)\}$. For $n \geq 1$, let $\ell(n) = \bigcup_{i=0}^{\infty} m(n, i)$ and let $\ell(0) = \bigcup_{i=1}^{\infty} m(0, i)$. The transitions for this process are given in Table 7.1.

From	To	Rate	For
(n, i, j)	$(n + 1, 0, j)$	λ_{n+i}	$n \geq 1, i \geq 0, 1 \leq j \leq n$
$(0, i, 1)$	$(1, 0, 1)$	λ_i	$i \geq 1$
(n, i, j)	$(n - 1, i + 1, n - 1)$	μ_j	$n \geq 2, i \geq 0, 1 \leq j \leq n$
$(1, i, 1)$	$(0, i + 1, 1)$	μ_1	$i \geq 0$

Table 7.1: Possible transitions for the (QL, AS, O) process.

Below, we detail the construction of the model matrices, $A_+^{(n)}, A_0^{(n)}, A_-^{(n)}$, for this level-dependent QBD.

For $n \geq 1, i \geq 0$, let $U_+^{(n,i)}$ be $n \times (n + 1)$ matrices which contain the transition rates in which $n \to n + 1$ and $i \to 0$. These are sub-matrices of the matrices $A_+^{(n)}$.

$$U_+^{(n,i)} = \begin{bmatrix} \lambda_{n+i} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_{n+i} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \lambda_{n+i} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{n+i} & 0 \end{bmatrix}.$$

So for $n \geq 1$,

$$A_+^{(n)} = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} m(n+1,0) & m(n+1,1) & m(n+1,2) & \cdots \end{array} \\ \begin{bmatrix} U_+^{(n,0)} & \mathbf{0} & \mathbf{0} & \cdots \\ U_+^{(n,1)} & \mathbf{0} & \mathbf{0} & \cdots \\ U_+^{(n,2)} & \mathbf{0} & \mathbf{0} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array} \quad \text{and}$$

$$A_+^{(0)} = \begin{array}{c} \\ (0,1,1)) \\ (0,2,1) \\ (0,3,1) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} (1,0,1) & (1,1,1) & (1,2,1) & \cdots \end{array} \\ \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots \\ \lambda_2 & 0 & 0 & \cdots \\ \lambda_3 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Note that the $n = 0$ case is different because the the state space is slightly different since we cannot have both $n = 0$ and $i = 0$.

Now, for $n \geq 2$ and $i \geq 0$, let $U_-^{(n,i)}$ be $n \times (n-1)$ matrices containing the rates in which $n \to n-1$ and $i \to i+1$. These form sub-matrices for the matrices $A_-^{(n)}$.

$$U_-^{(n,i)} = \begin{bmatrix} 0 & \cdots & 0 & \mu_1 \\ 0 & \cdots & 0 & \mu_2 \\ 0 & \cdots & 0 & \mu_3 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \mu_n \end{bmatrix}.$$

So for $n \geq 2$,

$$A_-^{(n)} = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} m(n-1,0) & m(n-1,1) & m(n-1,2) & m(n-1,3) & \cdots \end{array} \\ \begin{bmatrix} \mathbf{0} & U_-^{(n,0)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & U_-^{(n,1)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & U_-^{(n,2)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array} \quad \text{and}$$

$$A_-^{(1)} = \begin{array}{c} \\ (1,0,1) \\ (1,1,1) \\ (1,2,1) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} (0,1,1) & (0,2,1) & (0,3,1) & \cdots \end{array} \\ \begin{bmatrix} \mu_1 & 0 & 0 & \cdots \\ 0 & \mu_1 & 0 & \cdots \\ 0 & 0 & \mu_1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array} = \mu_1 \mathcal{I}.$$

Now for $n \geq 1, i \geq 0$ let $U_0^{(n,i)}$ be $n \times n$ matrices which contain the negative sums of the transition rates for each row. These sub-matrices form the matrices $A_0^{(n)}$.

$$U_0^{(n,i)} = \begin{bmatrix} -(\lambda_{n+i} + \mu_1) & 0 & \cdots & 0 \\ 0 & -(\lambda_{n+i} + \mu_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -(\lambda_{n+i} + \mu_n) \end{bmatrix} = -\lambda_{n+i}\mathcal{I} - diag(\mu_1, \mu_2, \dots).$$

So for $n \geq 1$,

$$A_0^{(n)} = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} m(n,0) & m(n,1) & m(n,2) & \cdots \end{array} \\ \begin{bmatrix} U_0^{(n,0)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & U_0^{(n,1)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & U_0^{(n,2)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array} \quad \text{and}$$

$$A_0^{(0)} = \begin{array}{c} \\ (1,0,1) \\ (1,1,1) \\ (1,2,1) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} (0,1,1) & (0,2,1) & (0,3,1) & \cdots \end{array} \\ \begin{bmatrix} -\lambda_1 & 0 & 0 & \cdots \\ 0 & -\lambda_2 & 0 & \cdots \\ 0 & 0 & -\lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array} = -diag(\lambda_1, \lambda_2, \dots).$$

So, the generator matrix for this level-dependent QBD is given by

$$Q = \begin{array}{c} \\ \ell(0) \\ \ell(1) \\ \ell(2) \\ \ell(3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} \ell(0) & \ell(1) & \ell(2) & \ell(3) & \cdots \end{array} \\ \begin{bmatrix} A_0^{(0)} & A_+^{(0)} & \mathbf{0} & \mathbf{0} & \cdots \\ A_-^{(1)} & A_0^{(1)} & A_+^{(1)} & \mathbf{0} & \cdots \\ \mathbf{0} & A_-^{(2)} & A_0^{(2)} & A_+^{(2)} & \cdots \\ \mathbf{0} & \mathbf{0} & A_-^{(3)} & A_0^{(3)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

7.1.1 Truncation and Augmentation

Note that the phase space for this process is not finite since there is no upper bound on the values taken by i. To practically apply this model, the state space

and model matrices need to be truncated. Let M be an artificial maximum number of departures between two arrivals. Then the truncated state space becomes $S = \{(n,i,j) : n \geq 1, 1 \leq j \leq n, 0 \leq i \leq M, n+i \geq 1\} \cup \{(0,i,1) : 1 \leq i \leq M\}$.

The matrices $A_+^{(n)}$ are simply truncated as follows. For $n \geq 1$,

$$A_+^{(n)} = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \\ m(n,M) \end{array} \begin{array}{c} \begin{array}{cccc} m(n+1,0) & m(n+1,1) & \cdots & m(n+1,M) \end{array} \\ \begin{bmatrix} U_+^{(n,0)} & \mathbf{0} & \cdots & \mathbf{0} \\ U_+^{(n,1)} & \mathbf{0} & \cdots & \mathbf{0} \\ U_+^{(n,2)} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ U_+^{(n,M)} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \end{array} \quad \text{and}$$

$$A_+^{(0)} = \begin{array}{c} \\ (0,1,1)) \\ (0,2,1) \\ (0,3,1) \\ \vdots \\ (0,M,1) \end{array} \begin{array}{c} \begin{array}{cccc} (1,0,1) & (1,1,1) & \cdots & (1,M,1) \end{array} \\ \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ \lambda_2 & 0 & \cdots & 0 \\ \lambda_3 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_M & 0 & \cdots & 0 \end{bmatrix} \end{array},$$

where $A_+^{(n)}$ is an $n(M+1) \times (n+1)(M+1)$ matrix for $n \geq 1$ and $A_+^{(0)}$ is an $M \times (M+1)$ matrix.

The matrices $U_-^{(n,i)}$ are defined only for $0 \leq i \leq M-1$ and when $i = M$ there are no departures allowed since the maximum number of departures has already been attained. This means that the rows of $A_-^{(n)}$ where $i = M$ are rows of zeros. So the truncated and augmented matrices $A_-^{(n)}$ are given as follows. For

$n \geq 2$,

$$A_{-}^{(n)} = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \\ m(n,M-1) \\ m(n,M) \end{array} \begin{array}{c} \begin{array}{cccccc} m(n-1,0) & m(n-1,1) & m(n-1,2) & m(n-1,3) & \cdots & m(n-1,M) \end{array} \\ \begin{bmatrix} \mathbf{0} & U_{-}^{(n,0)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & U_{-}^{(n,1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & U_{-}^{(n,2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & U_{-}^{(n,M-1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \end{array} \quad \text{and}$$

$$A_{-}^{(1)} = \begin{array}{c} \\ (1,0,1) \\ (1,1,1) \\ (1,2,1) \\ \vdots \\ (1,M-1,1) \\ (1,M,1) \end{array} \begin{array}{c} \begin{array}{ccccc} (0,1,1) & (0,2,1) & (0,3,1) & \cdots & (0,M,1) \end{array} \\ \begin{bmatrix} \mu_1 & 0 & 0 & \cdots & 0 \\ 0 & \mu_1 & 0 & \cdots & 0 \\ 0 & 0 & \mu_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mu_1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \end{array},$$

where $A_{-}^{(n)}$ is an $n(M+1) \times (n-1)(M+1)$ matrix for $n \geq 2$ and $A_{-}^{(1)}$ is an $(M+1) \times M$ matrix.

Finally, since there are no departures when $i = M$, $U_0^{(n,M)}$ is an $n \times n$ matrix defined by

$$U_0^{(n,M)} = \begin{bmatrix} -\lambda_{n+i} & 0 & 0 & \cdots & 0 \\ 0 & -\lambda_{n+i} & 0 & \cdots & 0 \\ 0 & 0 & -\lambda_{n+i} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_{n+i} \end{bmatrix}.$$

Then the truncation for the matrices $A_0^{(n)}$ is given below. For $n \geq 1$,

$$A_0^{(n)} = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \\ m(n,M) \end{array} \begin{array}{c} \begin{array}{ccccc} m(n,0) & m(n,1) & m(n,2) & \cdots & m(n,M) \end{array} \\ \begin{bmatrix} U_0^{(n,0)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & U_0^{(n,1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & U_0^{(n,2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & U_0^{(n,M)} \end{bmatrix} \end{array} \quad \text{and}$$

$$A_0^{(0)} = \begin{array}{c} \\ (0,1,1) \\ (0,1,1) \\ (0,2,1) \\ \vdots \\ (0,M,1) \end{array} \begin{array}{c} \begin{array}{ccccc} (0,1,1) & (0,2,1) & (0,3,1) & \cdots & (0,M,1) \end{array} \\ \begin{bmatrix} -\lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & -\lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_M \end{bmatrix} \end{array},$$

where $A_0^{(n)}$ is an $n(M+1) \times n(M+1)$ matrix for $n \geq 1$ and $A_0^{(0)}$ is an $M \times M$ matrix.

7.1.2 Truncating Queue Length

Now we divert slightly to consider this original model, but where there is an upper bound on the queue length. This is not required to model the original queue, but becomes necessary later for the models of the semi-experiment queues. In both the s-perm and a-perm models, this original queue features as an 'underlying' queue in the phase space (similar to the semi-experiment models in Chapters 5 and 6) and is required to be truncated. For ease of referring back to the original model, we perform that truncation here.

Let N be the maximum queue length that this queue can attain. Therefore, we have $0 \leq n \leq N$. Further, we know that the queue length at arrival epochs cannot exceed this maximum, so we also have $n + i \leq N$. Combining this with the already established state space, we get that $0 \leq i \leq N - n$ for $n \geq 1$ and $1 \leq i \leq N$ for $n = 0$. So the state space for this truncated process is $S = \{(n, i, j) : 1 \leq n \leq N, 1 \leq j \leq n, 0 \leq i \leq N - n\} \cup \{(0, i, 1) : 1 \leq i \leq N\}$. Note this is different to the truncation where we simply imposed an upper limit on the number of departures between two arrivals, i, since we now impose an upper limit on the actual queue length.

Then, the $A_+^{(n)}$ matrices are truncated for $1 \leq n \leq N-1$,

$$A_+^{(n)} = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \\ m(n,N-n) \end{array} \begin{array}{c} \begin{array}{cccc} m(n+1,0) & m(n+1,1) & \cdots & m(n+1,N-(n+1)) \end{array} \\ \begin{bmatrix} U_+^{(n,0)} & \mathbf{0} & \cdots & \mathbf{0} \\ U_+^{(n,1)} & \mathbf{0} & \cdots & \mathbf{0} \\ U_+^{(n,2)} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ U_+^{(n,N-n)} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \end{array} \quad \text{and}$$

$$A_+^{(0)} = \begin{array}{c} \\ (0,1,1)) \\ (0,2,1) \\ (0,3,1) \\ \vdots \\ (0,N,1) \end{array} \begin{array}{c} \begin{array}{cccc} (1,0,1) & (1,1,1) & \cdots & (1,N-1,1) \end{array} \\ \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ \lambda_2 & 0 & \cdots & 0 \\ \lambda_3 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_N & 0 & \cdots & 0 \end{bmatrix} \end{array},$$

where $A_+^{(n)}$ is an $n(N-n+1) \times (n+1)(N-n)$ matrix for $1 \leq n \leq N-1$ and $A_+^{(0)}$ is an $N \times N$ matrix.

$A_+^{(N)} = \mathbf{0}_N$, since there can be no arrivals when the queue length is N.

The matrices $A_-^{(n)}$ are truncated for $2 \leq n \leq N$,

$$A_-^{(n)} = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \\ m(n,N-n) \end{array} \begin{array}{c} \begin{array}{cccccc} m(n-1,0) & m(n-1,1) & m(n-1,2) & m(n-1,3) & \cdots & m(n-1,N-(n-1)) \end{array} \\ \begin{bmatrix} \mathbf{0} & U_-^{(n,0)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & U_-^{(n,1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & U_-^{(n,2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & U_-^{(n,N-n)} \end{bmatrix} \end{array} \quad \text{and}$$

$$A_-^{(1)} = \begin{array}{c} \\ (1,0,1) \\ (1,1,1) \\ (1,2,1) \\ \vdots \\ (1,N-1,1) \end{array} \begin{array}{c} \begin{array}{ccccc} (0,1,1) & (0,2,1) & (0,3,1) & \cdots & (0,N,1) \end{array} \\ \begin{bmatrix} \mu_1 & 0 & 0 & \cdots & 0 \\ 0 & \mu_1 & 0 & \cdots & 0 \\ 0 & 0 & \mu_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mu_1 \end{bmatrix} \end{array} = \mu_1 \mathcal{I}_N,$$

where $A_-^{(n)}$ is an $n(N-n+1) \times (n-1)(N-n+2)$ matrix for $2 \leq n \leq N$ and $A_-^{(1)}$ is an $N \times N$ matrix.

Finally, there can be no more departures when $i = N$, which can only occur when $n = 0$. As there can be no departures when $n = 0$ in any queue, we do not need to worry about this condition (unlike the previous truncation).

So, the truncated matrices $A_0^{(n)}$ are given below for $1 \leq n \leq N-1$,

$$A_0^{(n)} = \begin{array}{c} \\ m(n,0) \\ m(n,1) \\ m(n,2) \\ \vdots \\ m(n,N-n) \end{array} \begin{array}{c} \begin{array}{ccccc} m(n,0) & m(n,1) & m(n,2) & \cdots & m(n,N-n) \end{array} \\ \begin{bmatrix} U_0^{(n,0)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & U_0^{(n,1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & U_0^{(n,2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & U_0^{(n,N-n)} \end{bmatrix} \end{array} \quad \text{and}$$

$$A_0^{(0)} = \begin{array}{c} \\ (0,1,1) \\ (0,1,1) \\ (0,2,1) \\ \vdots \\ (0,N,1) \end{array} \begin{array}{c} \begin{array}{ccccc} (0,1,1) & (0,2,1) & (0,3,1) & \cdots & (0,N,1) \end{array} \\ \begin{bmatrix} -\lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & -\lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_N \end{bmatrix} \end{array}, \quad \text{and}$$

$$A_0^{(N)} = \begin{array}{c} \\ (N,1,1) \\ (N,1,1) \\ (N,2,1) \\ \vdots \\ (N,N,1) \end{array} \begin{array}{c} \begin{array}{ccccc} (N,1,1) & (N,2,1) & (N,3,1) & \cdots & (N,N,1) \end{array} \\ \begin{bmatrix} -\mu_1 & 0 & 0 & \cdots & 0 \\ 0 & -\mu_2 & 0 & \cdots & 0 \\ 0 & 0 & -\mu_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\mu_N \end{bmatrix} \end{array},$$

where $A_0^{(n)}$ is an $n(N-n+1) \times n(N-n+1)$ matrix for $n \geq 1$ and $A_0^{(0)}$ and $A_0^{(N)}$ are $N \times N$ matrices.

7.2 Stationary Distributions of Embedded DTMCs

As before, we wish to construct models for the s-perm and a-perm semi-experiments applied to this original model. The construction of these is similar to those constructed for the previous queue-length-dependent models. Hence, we need to calculate the stationary probability distribution for the queue-length at the start of service, and the stationary probability distribution for the queue-length immediately after an arrival. That is, we need to evaluate the stationary distributions

for the DTMC embedded at the starts of services, and for the DTMC embedded immediately after arrivals.

Note that in all of the following, if $j > k$, we define

$$\prod_{i=j}^{k} f(i) = 1$$

for any function $f(i)$.

Introducing Notation

First, we introduce notation we will be using to find these stationary distributions.

Let $\widehat{X}_n$ be the queue length when the nth service begins (immediately after a departure in a non-empty queue or immediately after an arrival to an empty queue) for $n \geq 1$. Let $\widehat{T}_n$ be the time when the nth service begins.

Let $\widetilde{X}_n$ be the queue length immediately after the nth arrival for $n \geq 1$, and let $\widetilde{T}_n$ be the time of the nth arrival.

The stationary probability distributions we seek are $\widehat{\boldsymbol{\pi}}$ and $\widetilde{\boldsymbol{\pi}}$, where for $j \geq 1$,

$$\widehat{\pi}_j = \lim_{n\to\infty} P(\widehat{X}_n = j),$$
$$\widetilde{\pi}_j = \lim_{n\to\infty} P(\widetilde{X}_n = j).$$

That is $\widehat{\boldsymbol{\pi}}$ is the stationary distribution of the DTMC embedded at the epochs when service begins and $\widetilde{\boldsymbol{\pi}}$ is the stationary distribution of the DTMC embedded at the epochs immediately after arrivals.

Also, let $\widetilde{Y}_n$ be the queue length immediately after the most recent arrival before the nth service starts, and let $\widehat{Y}_n$ be the queue length at the start of the most recent service before the nth arrival occurs. That is,

$$\widetilde{Y}_n \equiv \widetilde{X}_m : \widetilde{T}_m \leq \widehat{T}_n < \widetilde{T}_{m+1}$$
$$\widehat{Y}_n \equiv \widehat{X}_m : \widehat{T}_m \leq \widetilde{T}_n < \widehat{T}_{m+1}.$$

Note that the definitions of $\widehat{Y}_n$ and $\widetilde{Y}_n$, are concerned with the arrival rate and service rate in progress at the time. Hence, when there is an arrival to

an empty queue in which the arrival and service begin simultaneously,

$$\widetilde{Y}_n \equiv \widetilde{X}_m : \widetilde{T}_m = \widehat{T}_n$$
$$\widehat{Y}_n \equiv \widehat{X}_m : \widehat{T}_m = \widetilde{T}_n.$$

Note that implicit in the definition of $\widetilde{Y}_n$ is that a service must occur before the next arrival.

Therefore, when we consider $P(\widetilde{Y}_n = \ell)$, this is the probability that an arrival occurs with a queue length of ℓ, *and* that the nth service occurs before the next arrival. That is,

$$P(\widetilde{Y}_n = \ell) = P(\widetilde{X}_m = \ell, \text{ and } \exists n \text{ s.t. } \widetilde{T}_m \leq \widehat{T}_n < \widetilde{T}_{m+1}).$$

Also, the probability $P(\widehat{X}_n = i \mid \widetilde{Y}_n = \ell)$ is the probability that a service begins with a queue length of i, given that the most recent arrival has a queue length of ℓ, *and* that at least one service begins before the next arrival.

There is an analogous argument for $\widehat{Y}_n$.

Calculating $\widehat{\pi}_j$

To calculate the value of $\widehat{\pi}_j, j \geq 1$, we introduce the conditional probabilities $\widehat{p}_{ij}$, defined as the probability of having a service begin with a queue length of j given that the previous service began with a queue length of i. That is,

$$\widehat{p}_{ij} = P(\widehat{X}_{n+1} = j \mid \widehat{X}_n = i).$$

Note that we assume the queue is time-homogeneous.

Note that if $i \geq 2$, then the nth service starts due to the $(n-1)$th departure. Now there may be $0, 1, 2, \ldots$ arrivals during this service period. Therefore, the queue length at the start of the nth service period may be $j = i-1, i, i+1, i+2, \ldots$. If $i = 1$, then we cannot have $j = i - 1$, since the queue length at the start of service must be at least 1 ($j \geq 1$), so $j = 1, 2, 3, \ldots$. Hence the set of non-zero values for $\widehat{p}_{ij}$ is $\{(i,j) : i \geq 2, j \geq i-1\} \cup \{(1,j) : j \geq 1\}$.

Therefore, the global balance equations are

$$\widehat{\pi}_j = \sum_{i=1}^{j+1} \widehat{\pi}_i \widehat{p}_{ij}, \qquad \text{for } j \geq 1,$$

subject to

$$\sum_{j=1}^{\infty} \widehat{\pi}_j = 1.$$

Calculating $\widetilde{\pi}_j$

To calculate $\widetilde{\pi}_j$ for $j \geq 1$, we define $\widetilde{p}_{ij}$ to be the conditional probability that the queue length immediately after an arrival is j given that the queue length immediately after the previous arrival was i. That is,

$$\widetilde{p}_{ij} = P(\widetilde{X}_{n+1} = j \mid \widetilde{X}_n = i).$$

Note that there may be $0, 1, 2, \ldots, i$ departures before the next arrival, so $j = 1, 2, \ldots i{+}1$. So the set of non-zero values for $\widetilde{p}_{ij}$ is $\{(i, j) : 1 \leq j \leq i{+}1, i \geq 1\}$.

Therefore, the global balance equations are

$$\widetilde{\pi}_j = \sum_{i=j-1}^{\infty} \widetilde{\pi}_i \widetilde{p}_{ij}, \qquad \text{for } j \geq 2,$$

and

$$\widetilde{\pi}_1 = \sum_{i=1}^{\infty} \widetilde{\pi}_i \widetilde{p}_{i1},$$

subject to

$$\sum_{j=1}^{\infty} \widetilde{\pi}_j = 1.$$

Calculating $\widehat{p}_{ij}$

Now we need to calculate the probabilities $\widehat{p}_{ij}$ for $\{(i, j) : i \geq 2, j \geq i{-}1\} \cup \{(1, j) : j \geq 1\}$, where

$$\widehat{p}_{ij} = P(\widehat{X}_{n+1} = j \mid \widehat{X}_n = i).$$

If $j \geq 2$, then $\widehat{p}_{ij}$ is equivalent to the probability that there are exactly $j - i + 1$ arrivals during a service period with rate μ_i that ends with a departure. We will consider the particular case when $j = 1$ later.

In order to construct the probabilities $\widehat{p}_{ij}$, we need to consider the initial event during the nth service period that begins with a queue length of i. This can either be a departure with rate μ_i or an arrival. Assume that λ_ℓ is the arrival rate at the time when the nth service period begins. Then the probability that the first event in the service period is an arrival is given by

$$\frac{\lambda_\ell}{\lambda_\ell + \mu_i},$$

and the probability that the first event is a departure is given by

$$\frac{\mu_i}{\lambda_\ell + \mu_i}.$$

Now, we require the probability that the initial arrival rate is λ_ℓ. Let

$$\widetilde{q}_{\ell|i} = P(\widetilde{Y}_n = \ell \mid \widehat{X}_n = i).$$

That is, $\widetilde{q}_{\ell|i}$ is the probability, given that a service begins with a queue length of i, that the queue length immediately after the preceding arrival was ℓ. These probabilities will be calculated later.

Now we consider the values that ℓ can take. If $i \geq 2$, the service period must have begun due to a departure. Hence, $\ell \geq i+1$ since ℓ is the queue length immediately after the most recent arrival before the departure that left the queue with length i. If $i = 1$, then the service may have begun due to an arrival to an empty queue, and in this case the initial arrival rate is λ_1. Hence when $i = 1$, $\ell \geq 1$. So if we condition on the initial arrival rate being λ_ℓ and apply the law of total probability, then the probability that the first event in the nth service period is an arrival is

$$F_i^A = \begin{cases} \displaystyle\sum_{\ell=i+1}^{\infty} \frac{\lambda_\ell}{\lambda_\ell + \mu_i} \widetilde{q}_{\ell|i}, & \text{for } i \geq 2 \\ \displaystyle\sum_{\ell=1}^{\infty} \frac{\lambda_\ell}{\lambda_\ell + \mu_1} \widetilde{q}_{\ell|1}, & \text{for } i = 1, \end{cases}$$

and the probability that the first event in the nth service period is a departure is

$$F_i^D = \begin{cases} \displaystyle\sum_{\ell=i+1}^{\infty} \frac{\mu_i}{\lambda_\ell + \mu_i} \widetilde{q}_{\ell|i}, & \text{for } i \geq 2 \\ \displaystyle\sum_{\ell=1}^{\infty} \frac{\mu_i}{\lambda_\ell + \mu_1} \widetilde{q}_{\ell|1}, & \text{for } i = 1. \end{cases}$$

Now consider the following cases to construct the probabilities $\widehat{p}_{ij}$.

Case: $2 \leq i \leq j$ or $i = 1, j \geq 2$ In this case, there must be $j - i + 1 \geq 1$ arrivals during the nth service period. The probability that the first event is an arrival is given by F_i^A.

Now the rates of the subsequent $j - i$ arrivals during this service period are known to be $\lambda_{i+1}, \lambda_{i+2}, \ldots, \lambda_j$ because the queue length after the first arrival must be $i + 1$. So the probability that there are $j - i$ subsequent arrivals before the departure is given by

$$\prod_{x=i+1}^{j} \frac{\lambda_x}{\lambda_x + \mu_i}.$$

Since we require *exactly* $j - i + 1$ arrivals, the service period must end (a departure must occur) after these arrivals. The probability that a departure is the next event is given by

$$\frac{\mu_i}{\lambda_{j+1} + \mu_i}.$$

So for $2 \leq i \leq j$ or $i = 1, j \geq 2$,

$$\widehat{p}_{ij} = F_i^A \times \prod_{x=i+1}^{j} \frac{\lambda_x}{\lambda_x + \mu_i} \times \frac{\mu_i}{\lambda_{j+1} + \mu_i}.$$

Case: $i \geq 2, j = i - 1$ In this case there are 0 arrivals during the nth service period. So the first and only event must be a departure. Therefore,

$$\widehat{p}_{i,i-1} = F_i^D, \quad i \geq 2.$$

Case: $i = j = 1$ There are two ways we can have $j = 1$ when the service period starts with a queue length of $i = 1$. Either there is one arrival during the service period and then the next service begins after a departure, leaving a queue length of 1; or there is a departure immediately and then an arrival to an empty queue which begins a new service with a queue length of 1. For the first case, the probability is given by

$$F_1^A \times \frac{\mu_1}{\lambda_2 + \mu_1}.$$

Note that for the second case the probability of an arrival when the queue is empty is 1 and so the probability for the second case is

$$F_1^D.$$

Therefore, using the fact that the first event has to be either an arrival or a service,

$$\begin{aligned}\widehat{p}_{1,1} =& F_1^A \times \frac{\mu_1}{\lambda_2+\mu_1} + F_1^D \\ =& F_1^A \times \frac{\mu_1}{\lambda_2+\mu_1} + (1 - F_1^A) \\ =& 1 - \left(1 - \frac{\mu_1}{\lambda_2+\mu_1}\right) F_1^A \\ =& 1 - \frac{\lambda_2}{\lambda_2+\mu_1} F_1^A \\ =& 1 - \frac{\lambda_2}{\lambda_2+\mu_1} \sum_{\ell=1}^{\infty} \frac{\lambda_\ell}{\lambda_\ell+\mu_1} \widetilde{q}_{\ell|1}.\end{aligned}$$

So to summarise,

$$\widehat{p}_{ij} = \begin{cases} 1 - \dfrac{\lambda_2}{\lambda_2+\mu_1} \displaystyle\sum_{\ell=1}^{\infty} \frac{\lambda_\ell}{\lambda_\ell+\mu_1} \widetilde{q}_{\ell|1}, & \text{for } i=j=1, \\ \displaystyle\sum_{\ell=1}^{\infty} \frac{\lambda_\ell}{\lambda_\ell+\mu_1} \widetilde{q}_{\ell|1} \times \prod_{x=2}^{j} \frac{\lambda_x}{\lambda_x+\mu_1} \times \frac{\mu_1}{\lambda_{j+1}+\mu_1}, & \text{for } i=1, j \geq 2, \\ \displaystyle\sum_{\ell=i+1}^{\infty} \frac{\mu_i}{\lambda_\ell+\mu_i} \widetilde{q}_{\ell|i}, & \text{for } j=i-1, i \geq 2, \\ \displaystyle\sum_{\ell=i+1}^{\infty} \frac{\lambda_\ell}{\lambda_\ell+\mu_i} \widetilde{q}_{\ell|i} \times \prod_{x=i+1}^{j} \frac{\lambda_x}{\lambda_x+\mu_i} \times \frac{\mu_i}{\lambda_{j+1}+\mu_i}, & \text{for } 2 \leq i \leq j. \end{cases}$$

Also, the following equations must hold,

$$\begin{aligned}\sum_{j=i-1}^{\infty} \widehat{p}_{ij} =& 1, \qquad \text{for } i \geq 2, \\ \sum_{j=1}^{\infty} \widehat{p}_{1j} =& 1,\end{aligned}$$

since the next service must begin with some positive queue length.

Calculating $\widetilde{p}_{ij}$

Now we calculate the probabilities $\widetilde{p}_{ij}$ for $\{(i,j) : 1 \leq j \leq i+1, i \geq 1\}$, where

$$\widetilde{p}_{ij} = P(\widetilde{X}_{n+1} = j \mid \widetilde{X}_n = i).$$

The probability $\widetilde{p}_{ij}$ is equivalent to the probability of having $i - j + 1$ departures during an inter-arrival period with arrival rate λ_i.

Similar to the calculation of the $\widehat{p}$'s, we need to consider the first event in the nth inter-arrival period which begins with a queue length of i. This can either be an arrival with rate λ_i or a departure. Assume that μ_ℓ is the service rate at the time when the nth inter-arrival period begins. Then the probability that the first event in the inter-arrival period is a departure is given by

$$\frac{\mu_\ell}{\mu_\ell + \lambda_i}$$

and the probability that the first event is an arrival is given by

$$\frac{\lambda_i}{\mu_\ell + \lambda_i}.$$

Now, we require the probability that the initial service rate is μ_ℓ. Let

$$\widehat{q}_{\ell|i} = P(\widehat{Y}_n = \ell \mid \widetilde{X}_n = i).$$

That is, $\widehat{q}_{\ell|i}$ is the probability, given that an arrival occurs with a queue length of i, that the queue length at the start of the preceding service was ℓ. These probabilities will be calculated later.

Now we consider the values that ℓ can take. If $i \geq 2$, then the nth arrival was to a non-empty queue with length i and hence the current service began before this arrival occurred, with a queue length of $1 \leq \ell \leq i - 1$. If $i = 1$, then the nth arrival was to an empty queue and a service began with a queue length of $\ell = 1$ immediately upon arrival. Therefore, using the law of total probability, the probability that the first event in the nth inter-arrival period is a departure is given by

$$G_i^D = \begin{cases} \displaystyle\sum_{\ell=1}^{i-1} \frac{\mu_\ell}{\mu_\ell + \lambda_i} \widehat{q}_{\ell|i}, & \text{for } i \geq 2, \\ \dfrac{\mu_1}{\mu_1 + \lambda_1}, & \text{for } i = 1, \end{cases}$$

and the probability that the first event in the nth inter-arrival period is an arrival is given by

$$G_i^A = \begin{cases} \displaystyle\sum_{\ell=1}^{i-1} \frac{\lambda_i}{\mu_\ell + \lambda_i} \widehat{q}_{\ell|i}, & \text{for } i \geq 2, \\ \dfrac{\lambda_1}{\mu_1 + \lambda_1}, & \text{for } i = 1. \end{cases}$$

Now consider the following cases to construct the probabilities $\widetilde{p}_{ij}$.

Case: $2 \leq j \leq i$ In this case, there are $i - j + 1 \geq 1$ departures during the nth inter-arrival period. The probability that the first event is an arrival is given by

$$G_i^D.$$

The service rates for the subsequent $i-j$ departures during the nth inter-arrival period are $\mu_{i-1}, \mu_{i-2}, \ldots, \mu_j$ since the queue length after the initial arrival is $i - 1$. So the probability that these departures occur is given by

$$\prod_{x=j}^{i-1} \frac{\mu_x}{\mu_x + \lambda_i},$$

noting that the product notation reverses the conventional order of writing the probability of the sequence of events.

Finally, we want *exactly* $i - j + 1$ departures during this inter-arrival period, so we require that the inter-arrival period ends before the next departure occurs. This probability is given by

$$\frac{\lambda_i}{\mu_{j-1} + \lambda_i}.$$

So, for $2 \leq j \leq i$,

$$\widetilde{p}_{ij} = G_i^D \times \prod_{x=j}^{i-1} \frac{\mu_x}{\mu_x + \lambda_i} \times \frac{\lambda_i}{\mu_{j-1} + \lambda_i}.$$

Case: $i \geq 1, j = i + 1$ In this case there are 0 departures during the nth inter-arrival period. This occurs when the first and only event is an arrival. Therefore,

$$\widetilde{p}_{i,i+1} = G_i^A, \quad i \geq 1.$$

Case: $i \geq 1, j = 1$ When $j = 1$, the nth arrival occurs into an empty queue. This means that there are i departures during the inter-arrival period and the queue length reaches 0 before the nth arrival. Once the queue is empty, the only event that can occur is an arrival. Therefore, once there have been i departures, the probability that the next event is an arrival, ending the inter-arrival period, is 1. So, for $i \geq 1$,

$$\widetilde{p}_{i,1} = G_i^D \times \prod_{x=1}^{i-1} \frac{\mu_x}{\mu_x + \lambda_i},$$

and

$$\widetilde{p}_{1,1} = G_1^D.$$

So, to summarise,

$$\widetilde{p}_{ij} = \begin{cases} \dfrac{\mu_1}{\mu_1+\lambda_1}, & \text{for } i=j=1 \\ \dfrac{\lambda_1}{\mu_1+\lambda_1}, & \text{for } i=1, j=2 \\ \displaystyle\sum_{\ell=1}^{i-1} \frac{\mu_\ell}{\mu_\ell+\lambda_i}\widehat{q}_{\ell|i} \times \prod_{x=1}^{i-1}\frac{\mu_x}{\mu_x+\lambda_i}, & \text{for } i\geq 2, j=1 \\ \displaystyle\sum_{\ell=1}^{i-1} \frac{\lambda_i}{\mu_\ell+\lambda_i}\widehat{q}_{\ell|i}, & \text{for } i\geq 2, j=i+1 \\ \displaystyle\sum_{\ell=1}^{i-1} \frac{\mu_\ell}{\mu_\ell+\lambda_i}\widehat{q}_{\ell|i} \times \prod_{x=j}^{i-1}\frac{\mu_x}{\mu_x+\lambda_i} \times \frac{\lambda_i}{\mu_{j-1}+\lambda_i}, & \text{for } 2\leq j\leq i. \end{cases}$$

The following equation also must hold,

$$\sum_{j=1}^{i+1} \widetilde{p}_{ij} = 1, \qquad \text{for } i \geq 1,$$

since the next arrival must occur with some queue length.

Calculating $\widetilde{q}_{\ell|i}$

In this section, we calculate the probabilities $\widetilde{q}_{\ell|i}$, where

$$\begin{aligned} \widetilde{q}_{\ell|i} =& P(\widetilde{Y}_n = \ell \mid \widehat{X}_n = i) \\ =& P(\widetilde{X}_m = \ell : \widetilde{T}_m \leq \widehat{T}_n < \widetilde{T}_{m+1} \mid \widehat{X}_n = i). \end{aligned}$$

That is, $\widetilde{q}_{\ell|i}$ is the probability, given a service begins with a queue length of i, that the queue length immediately after the preceding arrival was ℓ.

If we assume that the nth service begins due to a departure from a non-empty queue, then there is at least one departure since the most recent arrival. Hence, we can have $\ell = i+1, i+2, \ldots$. If the nth service begins due to an arrival into an empty queue, then $\ell = i = 1$. Hence, the non-zero values for $\widetilde{q}_{\ell|i}$ are given by $\ell \geq i+1, i \geq 1$ and $i = \ell = 1$.

To solve for these probabilities, we apply Bayes' Rule

$$\begin{aligned}\widetilde{q}_{\ell|i} =& \frac{P(\widehat{X}_n = i \mid \widetilde{Y}_n = \ell)P(\widetilde{Y}_n = \ell)}{P(\widehat{X}_n = i)} \\ =& \frac{P(\widehat{X}_n = i \mid \widetilde{Y}_n = \ell)P(\widetilde{Y}_n = \ell)}{\widehat{\pi}_i},\end{aligned}$$

by assuming stationarity.

We consider each of these terms separately. For ease, in part (a) we treat $P(\widehat{X}_n = i \mid \widetilde{Y}_n = \ell)$ and in part (b) we treat $P(\widetilde{Y}_n = \ell)$. In part (c) we combine them.

Part (a) Consider $P(\widehat{X}_n = i \mid \widetilde{Y}_n = \ell)$. This is the probability that a service begins with a queue length of i, given that the most recent arrival had a queue length of ℓ and that at least one service begins before the next arrival (this is implicit in the definition of $\widetilde{Y}_n$). This means that the first service occurs with probability 1, and there needs to be at least $\ell - i - 1 \geq 0$ additional services beginning before the next arrival.

Consider $1 \leq i \leq \ell - 1$. Since $\ell \geq 2$, all services begin due to a departure from a non-empty queue. So,

$$P(\widehat{X}_n = i \mid \widetilde{Y}_n = \ell) = \prod_{x=i+1}^{\ell-1} \frac{\mu_x}{\mu_x + \lambda_\ell},$$

and

$$P(\widehat{X}_n = 1 \mid \widetilde{Y}_n = 1) = 1,$$

since when $\ell = i = 1$ there was an arrival to an empty queue which immediately begins a service with a queue length of 1.

Part (b) Now, $P(\widetilde{Y}_n = \ell)$ is the probability that a arrival occurs with a queue length of ℓ and that a service begins before the next arrival. So,

$$\begin{aligned}P(\widetilde{Y}_n = \ell) =& P(\widetilde{X}_m = \ell, \text{and } \exists n \text{ s.t. } \widetilde{T}_m \leq \widehat{T}_n < \widetilde{T}_{m+1}) \\ =& P(\exists n \text{ s.t. } \widetilde{T}_m \leq \widehat{T}_n < \widetilde{T}_{m+1} \mid \widetilde{X}_m = \ell)P(\widetilde{X}_m = \ell) \\ =& P(\exists n \text{ s.t. } \widetilde{T}_m \leq \widehat{T}_n < \widetilde{T}_{m+1} \mid \widetilde{X}_m = \ell)\widetilde{\pi}_\ell,\end{aligned}$$

assuming stationarity.

Now $P(\exists n \text{ s.t. } \widetilde{T}_m \leq \widehat{T}_n < \widetilde{T}_{m+1} \mid \widetilde{X}_m = \ell)$ is the probability that at least one service begins between the mth and $(m+1)$th arrivals given that the queue length at the mth arrival was ℓ.

Note that if $\ell = 1$ then a service began when the arrival occurred into an empty queue, so $P(\exists n \text{ s.t. } \widetilde{T}_m \leq \widehat{T}_n < \widetilde{T}_{m+1} \mid \widetilde{X}_m = 1) = 1$. So,

$$P(\widetilde{Y}_n = 1) = \widetilde{\pi}_1.$$

For $\ell \geq 2$, we need to condition on the initial service rate at the time of the mth arrival. So,

$$\begin{aligned}
&P(\widetilde{Y}_n = \ell) \\
=&P(\exists n \text{ s.t. } \widetilde{T}_m \leq \widehat{T}_n < \widetilde{T}_{m+1} \mid \widetilde{X}_m = \ell)\widetilde{\pi}_\ell \\
=&\sum_{k=1}^{\ell-1} P(\exists n \text{ s.t. } \widetilde{T}_m \leq \widehat{T}_n < \widetilde{T}_{m+1} \mid \widetilde{X}_m = \ell, \widehat{Y}_m = k) P(\widehat{Y}_m = k \mid \widetilde{X}_m = \ell)\widetilde{\pi}_\ell \\
=&\sum_{k=1}^{\ell-1} \frac{\mu_k}{\lambda_\ell + \mu_k} \widehat{q}_{k|\ell} \widetilde{\pi}_\ell.
\end{aligned}$$

Part (c) Now we can put these parts together.

$$\widetilde{q}_{\ell|i} = \begin{cases} \dfrac{\widetilde{\pi}_1}{\widehat{\pi}_1}, & \text{for } i = \ell = 1, \\ \displaystyle\prod_{x=i+1}^{\ell-1} \frac{\mu_x}{\mu_x + \lambda_\ell} \times \sum_{k=1}^{\ell-1} \frac{\mu_k}{\lambda_\ell + \mu_k} \widehat{q}_{k|\ell} \times \frac{\widetilde{\pi}_\ell}{\widehat{\pi}_i}, & \text{for } 1 \leq i \leq \ell - 1. \end{cases}$$

Note that the following equations must also hold,

$$\begin{aligned}
\sum_{\ell=i+1}^{\infty} \widetilde{q}_{\ell|i} =& 1, \qquad \text{for } i \geq 2 \\
\sum_{\ell=1}^{\infty} \widetilde{q}_{\ell|1} =& 1,
\end{aligned}$$

since there must be some arrival rate when a service begins.

Calculating $\widehat{q}_{\ell|i}$

Now we calculate the probabilities $\widehat{q}_{\ell|i}$, where

$$\begin{aligned}\widehat{q}_{\ell|i} =& P(\widehat{Y}_n = \ell \mid \widetilde{X}_n = i)\\ =& P(\widehat{X}_m = \ell, \widehat{T}_m \le \widetilde{T}_n < \widehat{T}_{m+1} \mid \widetilde{X}_n = i).\end{aligned}$$

That is, $\widehat{q}_{\ell|i}$ is the probability, given an arrival occurs with a queue length of i, that the preceding service starts with a queue length of ℓ.

Since there must be at least one arrival (the nth arrival) after the mth service begins with a queue length of $\ell \ge 2$, then the queue length at the nth arrival must be $i \ge \ell + 1$. If $\ell = 1$, then the nth arrival was to an empty queue, which means that a service immediately begins with a queue length of 1, so $\ell = i = 1$. Hence, the non-zero values for $\widehat{q}_{\ell|i}$ are given by $1 \le \ell \le i - 1$ and $\ell = i = 1$.

First, note that if $i = 1$, then the arrival occurs into an empty queue, which starts a service with a queue length of 1. So, the only possibility is that $\ell = 1$ and $\widehat{q}_{1|1} = 1$.

When $i \ge 2$ and $1 \le \ell \le i - 1$, we apply Bayes' Rule,

$$\begin{aligned}\widehat{q}_{\ell|i} =& \frac{P(\widetilde{X}_n = i \mid \widehat{Y}_n = \ell)P(\widehat{Y}_n = \ell)}{P(\widetilde{X}_n = i)}\\ & \frac{P(\widetilde{X}_n = i \mid \widehat{Y}_n = \ell)P(\widehat{Y}_n = \ell)}{\widetilde{\pi}_i},\end{aligned}$$

assuming stationarity.

As before, we consider each term separately. Part (a) concerns $P(\widetilde{X}_n = i \mid \widehat{Y}_n = \ell)$, part (b) concerns $P(\widehat{Y}_n = \ell)$, and in part (c) these are combined.

Part (a) Consider $P(\widetilde{X}_n = i \mid \widehat{Y}_n = \ell)$. This is the probability that there is an arrival with queue length i given the most recent service began with a queue length of ℓ and there is at least 1 arrival before the next service. So the probability that the first arrival occurs is 1. Therefore we need at least $\ell - i - 1 \ge 0$ additional arrivals before the next service begins, with service rate μ_ℓ.

So for $1 \leq \ell \leq i-1$,

$$P(\widetilde{X}_n = i \mid \widehat{Y}_n = \ell) = \prod_{x=\ell+1}^{i-1} \frac{\lambda_x}{\lambda_x + \mu_\ell}.$$

Part (b) Now, $P(\widehat{Y}_n = \ell)$ is the probability that a service begins with a queue length of ℓ and that there is an arrival before the next service. So, for $2 \leq \ell \leq i-1$,

$$\begin{aligned}
&P(\widehat{Y}_n = \ell) \\
=&P(\widehat{X}_m = \ell, \text{and } \exists \text{s.t. } \widehat{T}_m \leq \widetilde{T}_n < \widehat{T}_{m+1}) \\
=&P(\exists n \text{ s.t. } \widehat{T}_m \leq \widetilde{T}_n < \widehat{T}_{m+1} \mid \widehat{X}_m = \ell) P(\widehat{X}_m = \ell) \\
=&P(\exists n \text{ s.t. } \widehat{T}_m \leq \widetilde{T}_n < \widehat{T}_{m+1} \mid \widehat{X}_m = \ell) \widehat{\pi}_\ell, \qquad \text{(assuming stationarity)}, \\
=&\sum_{k=\ell+1}^{\infty} P(\exists n \text{ s.t. } \widehat{T}_m \leq \widetilde{T}_n < \widehat{T}_{m+1} \mid \widehat{X}_m = \ell, \widetilde{Y}_n = k) P(\widetilde{Y}_n = k \mid \widehat{X}_m = \ell) \widehat{\pi}_\ell \\
=&\sum_{k=\ell+1}^{\infty} \frac{\lambda_k}{\lambda_k + \mu_\ell} \widetilde{q}_{k|\ell} \widehat{\pi}_\ell.
\end{aligned}$$

Note that if $\ell = 1$ then the initial arrival rate could also be λ_1, so for $\ell = 1, i \geq 2$,

$$P(\widehat{Y}_n = \ell) = \sum_{k=1}^{\infty} \frac{\lambda_k}{\lambda_k + \mu_1} \widetilde{q}_{k|1} \widehat{\pi}_1.$$

Part (c) So we can put these parts together,

$$\widehat{q}_{\ell|i} = \begin{cases} 1, & \text{for } \ell = i = 1, \\ \displaystyle\prod_{x=2}^{i-1} \frac{\lambda_x}{\lambda_x + \mu_1} \times \sum_{k=1}^{\infty} \frac{\lambda_k}{\lambda_k + \mu_1} \widetilde{q}_{k|1} \times \frac{\widehat{\pi}_1}{\widetilde{\pi}_i}, & \text{for } \ell = 1, i \geq 2, \\ \displaystyle\prod_{x=\ell+1}^{i-1} \frac{\lambda_x}{\lambda_x + \mu_\ell} \times \sum_{k=\ell+1}^{\infty} \frac{\lambda_k}{\lambda_k + \mu_\ell} \widetilde{q}_{k|\ell} \times \frac{\widehat{\pi}_\ell}{\widetilde{\pi}_i}, & \text{for } 2 \leq \ell \leq i-1. \end{cases}$$

Note the following equations must also hold,

$$\sum_{\ell=1}^{i-1} \widehat{q}_{\ell|i} = 1, \qquad \text{for } i \geq 2,$$

since there must be some service rate when an arrival occurs.

7.2.1 Full System of Equations

The following is the full system of equations to be solved in order to evaluate $\{\widehat{\pi}_j : j \geq 1\}$ and $\{\widetilde{\pi}_j : j \geq 1\}$.

$$\widehat{q}_{\ell|i} = \begin{cases} 1, & \text{for } \ell = i = 1, \\ \displaystyle\prod_{x=2}^{i-1} \frac{\lambda_x}{\lambda_x + \mu_1} \times \sum_{k=1}^{\infty} \frac{\lambda_k}{\lambda_k + \mu_1} \widetilde{q}_{k|1} \times \frac{\widehat{\pi}_1}{\widetilde{\pi}_i}, & \text{for } \ell = 1, i \geq 2, \\ \displaystyle\prod_{x=\ell+1}^{i-1} \frac{\lambda_x}{\lambda_x + \mu_\ell} \times \sum_{k=\ell+1}^{\infty} \frac{\lambda_k}{\lambda_k + \mu_\ell} \widetilde{q}_{k|\ell} \times \frac{\widehat{\pi}_\ell}{\widetilde{\pi}_i}, & \text{for } 2 \leq \ell \leq i-1. \end{cases}$$

$$\widetilde{q}_{\ell|i} = \begin{cases} \dfrac{\widetilde{\pi}_1}{\widehat{\pi}_1}, & \text{for } i = \ell = 1, \\ \displaystyle\prod_{x=i+1}^{\ell-1} \frac{\mu_x}{\mu_x + \lambda_\ell} \times \sum_{k=1}^{\ell-1} \frac{\mu_k}{\lambda_\ell + \mu_k} \widehat{q}_{k|\ell} \times \frac{\widetilde{\pi}_\ell}{\widehat{\pi}_i}, & \text{for } 1 \leq i \leq \ell - 1. \end{cases}$$

$$\widetilde{p}_{ij} = \begin{cases} \dfrac{\mu_1}{\mu_1 + \lambda_1}, & \text{for } i = j = 1 \\ \dfrac{\lambda_1}{\mu_1 + \lambda_1}, & \text{for } i = 1, j = 2 \\ \displaystyle\sum_{\ell=1}^{i-1} \frac{\mu_\ell}{\mu_\ell + \lambda_i} \widehat{q}_{\ell|i} \times \prod_{x=1}^{i-1} \frac{\mu_x}{\mu_x + \lambda_i}, & \text{for } i \geq 2, j = 1 \\ \displaystyle\sum_{\ell=1}^{i-1} \frac{\lambda_i}{\mu_\ell + \lambda_i} \widehat{q}_{\ell|i}, & \text{for } i \geq 2, j = i + 1 \\ \displaystyle\sum_{\ell=1}^{i-1} \frac{\mu_\ell}{\mu_\ell + \lambda_i} \widehat{q}_{\ell|i} \times \prod_{x=j}^{i-1} \frac{\mu_x}{\mu_x + \lambda_i} \times \frac{\lambda_i}{\mu_{j-1} + \lambda_i}, & \text{for } 2 \leq j \leq i. \end{cases}$$

$$\widehat{p}_{ij} = \begin{cases} 1 - \dfrac{\lambda_2}{\lambda_2 + \mu_1} \displaystyle\sum_{\ell=1}^{\infty} \dfrac{\lambda_\ell}{\lambda_\ell + \mu_1} \widetilde{q}_{\ell|1}, & \text{for } i = j = 1, \\ \displaystyle\sum_{\ell=1}^{\infty} \dfrac{\lambda_\ell}{\lambda_\ell + \mu_1} \widetilde{q}_{\ell|1} \times \prod_{x=2}^{j} \dfrac{\lambda_x}{\lambda_x + \mu_1} \times \dfrac{\mu_1}{\lambda_{j+1} + \mu_1}, & \text{for } i = 1, j \geq 2, \\ \displaystyle\sum_{\ell=i+1}^{\infty} \dfrac{\mu_i}{\lambda_\ell + \mu_i} \widetilde{q}_{\ell|i}, & \text{for } j = i - 1, i \geq 2, \\ \displaystyle\sum_{\ell=i+1}^{\infty} \dfrac{\lambda_\ell}{\lambda_\ell + \mu_i} \widetilde{q}_{\ell|i} \times \prod_{x=i+1}^{j} \dfrac{\lambda_x}{\lambda_x + \mu_i} \times \dfrac{\mu_i}{\lambda_{j+1} + \mu_i}, & \text{for } 2 \leq i \leq j. \end{cases}$$

$$\widetilde{\pi}_j = \begin{cases} \displaystyle\sum_{i=j-1}^{\infty} \widetilde{\pi}_i \widetilde{p}_{ij}, & \text{for } j \geq 2 \\ \displaystyle\sum_{i=1}^{\infty} \widetilde{\pi}_i \widetilde{p}_{i1}, & \text{for } j = 1 \end{cases}$$

$$\widehat{\pi}_j = \sum_{i=1}^{j+1} \widehat{\pi}_i \widehat{p}_{ij}, \qquad \text{for } j \geq 1.$$

Subject to the normalisations:

$$\sum_{j=1}^{\infty} \widetilde{\pi}_j = 1,$$
$$\sum_{j=1}^{\infty} \widehat{\pi}_j = 1.$$

7.2.2 Truncated Stationary Distributions of Embedded DTMCs

To actually implement the system of equations above, we need to practically truncate and augment the system appropriately. Let M be the maximum possible queue length. The main changes to note are that if the queue length is ever M then there can be no arrivals and the probability of a departure is 1. Also, the queue length at the start of service must be at most $M - 1$ since we cannot have a

departure from a queue-length higher than M to start a service at a queue length of M.

Note that under this truncation, when the queue length after an arrival is M, there must be a departure as the next event. Hence, the inter-arrival period that *should* begin immediately after this arrival with rate λ_M does not actually begin until this departure has occurred. Since the process has exponential holding times which are memoryless, this does not affect the dynamics of the system. Note that this is very similar to how a standard queue deals with the boundary condition at 0. That is, when the queue is empty, there can be no departures until an arrival occurs.

Calculating $\widehat{\pi}_j$

We augment the value of $\widehat{\pi}_{M-1}$ since we are removing some possible values by truncating. Hence, we calculate $\widehat{\pi}_{M-1}$ using the normalisation condition instead.

Therefore,

$$\widehat{\pi}_j = \begin{cases} \displaystyle\sum_{i=1}^{j+1} \widehat{\pi}_i \widehat{p}_{ij}, & \text{for } 1 \leq j \leq M-2, \\ \displaystyle 1 - \sum_{j=1}^{M-2} \widehat{\pi}_j, & \text{for } j = M-1. \end{cases}$$

Calculating $\widetilde{\pi}_j$

Similarly, we augment $\widetilde{\pi}_M$ by using the normalisation condition.

$$\widetilde{\pi}_j = \begin{cases} \displaystyle\sum_{i=1}^{\infty} \widetilde{\pi}_i \widetilde{p}_{i1}, & \text{for } j = 1, \\ \displaystyle\sum_{i=j-1}^{\infty} \widetilde{\pi}_i \widetilde{p}_{ij}, & \text{for } 2 \leq j \leq M-1, \\ \displaystyle 1 - \sum_{j=1}^{M-1} \widehat{\pi}_j & \text{for } j = M. \end{cases}$$

Calculating $\widehat{p}_{ij}$

The set of i and j such that $\widehat{p}_{ij}$ is non-zero is $\{(i,j) : 2 \leq i \leq M-1, i-1 \leq j \leq M-1\} \cup \{(1,j) : 1 \leq j \leq M-1\}$.

The first change is that all the sums over ℓ are truncated to a maximum of M, the highest possible arrival rate. Further, when $j = M-1$, this means the queue length must have been M and so the departure that starts the service with a queue length of $j = M-1$ occurs with probability 1. So,

$$\widehat{p}_{ij} = \begin{cases} 1 - \dfrac{\lambda_2}{\lambda_2+\mu_1} \displaystyle\sum_{\ell=1}^{M} \frac{\lambda_\ell}{\lambda_\ell+\mu_1} \widetilde{q}_{\ell|1}, & \text{for } i=j=1, \\ \displaystyle\sum_{\ell=1}^{M} \frac{\lambda_\ell}{\lambda_\ell+\mu_1} \widetilde{q}_{\ell|1} \times \prod_{x=2}^{j} \frac{\lambda_x}{\lambda_x+\mu_1} \times \frac{\mu_1}{\lambda_{j+1}+\mu_1} & \text{for } i=1, 2 \leq j \leq M-2, \\ \displaystyle\sum_{\ell=1}^{M} \frac{\lambda_\ell}{\lambda_\ell+\mu_1} \widetilde{q}_{\ell|1} \times \prod_{x=2}^{M-1} \frac{\lambda_x}{\lambda_x+\mu_1} & \text{for } i=1, j=M-1, \\ \displaystyle\sum_{\ell=i+1}^{M} \frac{\mu_i}{\lambda_\ell+\mu_i} \widetilde{q}_{\ell|i} & \text{for } j=i-1, 2 \leq i \leq M-1, \\ \displaystyle\sum_{\ell=i+1}^{M} \frac{\lambda_\ell}{\lambda_\ell+\mu_i} \widetilde{q}_{\ell|i} \times \prod_{x=i+1}^{j} \frac{\lambda_x}{\lambda_x+\mu_i} \times \frac{\mu_i}{\lambda_{j+1}+\mu_i} & \text{for } 2 \leq i \leq j \leq M-2, \\ \displaystyle\sum_{\ell=i+1}^{M} \frac{\lambda_\ell}{\lambda_\ell+\mu_i} \widetilde{q}_{\ell|i} \times \prod_{x=i+1}^{M-1} \frac{\lambda_x}{\lambda_x+\mu_i} & \text{for } 2 \leq i \leq M-1, j=M-1. \end{cases}$$

Calculating $\widetilde{p}_{ij}$

The set of i and j such that $\widehat{p}_{ij}$ is non-zero is $\{(i,j) : 1 \leq i \leq M-1, 1 \leq j \leq i+1\} \cup \{(M,j) : 1 \leq j \leq M\}$.

The main change to carefully consider is that when the queue length is M there can be no arrivals, only departures. So when the queue reaches a queue length of M, a departure will occur with probability 1. Also, when this departure occurs it immediately begins a new service period with a service rate of μ_{M-1} and the inter-arrival period is restarted with an arrival rate of λ_M. Hence, whenever the arrival rate is λ_M, then the service rate at the start of the inter-arrival period must be μ_{M-1}.

$$\widetilde{p}_{ij} = \begin{cases} \dfrac{\mu_1}{\mu_1+\lambda_1}, & \text{for } i=j=1, \\ \dfrac{\lambda_1}{\mu_1+\lambda_1}, & \text{for } i=1, j=2, \\ \displaystyle\sum_{\ell=1}^{i-1} \frac{\mu_\ell}{\mu_\ell+\lambda_i}\widehat{q}_{\ell|i} \times \prod_{x=1}^{i-1} \frac{\mu_x}{\mu_x+\lambda_i}, & \text{for } 2 \le i \le M-1, j=1, \\ \displaystyle\prod_{x=1}^{i-1} \frac{\mu_x}{\mu_x+\lambda_i}, & \text{for } i=M, j=1, \\ \displaystyle\sum_{\ell=1}^{i-1} \frac{\lambda_i}{\mu_\ell+\lambda_i}\widehat{q}_{\ell|i}, & \text{for } 2 \le i \le M-1, j=i+1, \\ \displaystyle\sum_{\ell=1}^{i-1} \frac{\mu_\ell}{\mu_\ell+\lambda_i} \times \prod_{x=j}^{i-1} \frac{\mu_x}{\mu_x+\lambda_i} \times \frac{\lambda_i}{\mu_{j-1}+\lambda_i}, & \text{for } 2 \le j \le i \le M-1, \\ \displaystyle\prod_{x=j}^{M-1} \frac{\mu_x}{\mu_x+\lambda_M} \times \frac{\lambda_M}{\mu_{j-1}+\lambda_M}, & \text{for } 2 \le j \le M, i=M. \end{cases}$$

Calculating $\widetilde{q}_{\ell|i}$

The set of ℓ and i such that $\widetilde{q}_{\ell|i}$ is non-zero is given by $\{(\ell,i) : 1 \le i \le M-1, i+1 \le \ell \le M\} \cup \{(1,1)\}$.

The only change here is that when $\ell = M$, there must always be a departure from the queue length M. Hence, $P(\widetilde{Y}_n = M) = 1 \times \widetilde{\pi}_M$. So,

$$\widetilde{q}_{\ell|i} = \begin{cases} \dfrac{\widetilde{\pi}_1}{\widehat{\pi}_1}, & \text{for } i=\ell=1, \\ \displaystyle\prod_{x=i+1}^{\ell-1} \frac{\mu_x}{\mu_x+\lambda_\ell} \times \sum_{k=1}^{\ell-1} \frac{\mu_k}{\lambda_\ell+\mu_k}\widehat{q}_{k|\ell} \times \frac{\widetilde{\pi}_\ell}{\widehat{\pi}_i}, & \text{for } 1 \le i \le \ell-1 \le M-2, \\ \displaystyle\prod_{x=i+1}^{M-1} \frac{\mu_x}{\mu_x+\lambda_M} \times \frac{\widetilde{\pi}_M}{\widehat{\pi}_i}, & \text{for } 1 \le i \le M-1, \ell=M. \end{cases}$$

Calculating $\widehat{q}_{\ell|i}$

The set of ℓ and i such that $\widetilde{q}_{\ell|i}$ is non-zeros is given by $\{(\ell, i) : 1 \leq i \leq M, 1 \leq \ell \leq i-1\} \cup \{(1,1)\}$.

There are a few changes to be made here. Firstly, the sums over k are truncated at M, the maximum possible index for the arrival rate.

Recall that when the queue length after an arrival is M, there must be a departure as the next event. Hence, the inter-arrival period that *should* begin immediately after this arrival with rate λ_M does not actually begin until this departure has occurred.

Hence, when $i = M$, the only non-zero probability for $\widehat{q}_{\ell|M}$ is when $\ell = M-1$, since if the arrival rate is λ_M, then this inter-arrival period begins when there is a departure leaving the queue with a length of $M-1$. This also immediately begins a service with a queue length of $M-1$.

$$\widehat{q}_{\ell|i} = \begin{cases} 1, & \text{for } \ell = i = 1, \\ \displaystyle\prod_{x=2}^{i-1} \frac{\lambda_x}{\lambda_x + \mu_1} \times \sum_{k=1}^{\infty} \frac{\lambda_k}{\lambda_k + \mu_1} \widetilde{q}_{k|1} \times \frac{\widehat{\pi}_1}{\widetilde{\pi}_i}, & \text{for } \ell = 1, 2 \leq i \leq M-1, \\ \displaystyle\prod_{x=\ell+1}^{i-1} \frac{\lambda_x}{\lambda_x + \mu_\ell} \times \sum_{k=\ell+1}^{\infty} \frac{\lambda_k}{\lambda_k + \mu_\ell} \widetilde{q}_{k|\ell} \times \frac{\widehat{\pi}_\ell}{\widetilde{\pi}_i}, & \text{for } 2 \leq \ell \leq i-1 \leq M-2, \\ 1, & \text{for } i = M, \ell = M-1. \end{cases}$$

7.2.3 Verifying the System of Equations

We need to ensure that this system of equations is solvable and produces the desired distributions $\{\widehat{\pi}_j : 1 \leq j \leq M-1\}$ and $\{\widetilde{\pi}_j : 1 \leq j \leq M\}$.

First, we apply the MATLAB `solve` function. This proves capable of solving the system, but as M gets larger, the size of the system increases rapidly and the solver becomes computationally inefficient. The solver works in reasonable time for $M \leq 18$. Hence, we construct a simple iterative approach to solve the system for larger values of M more quickly. We begin with a uniform initial distribution for the unknowns.

Specifically, let

$$\begin{aligned}\widehat{\pi}_j =& \frac{1}{M-1}, \qquad \text{for } 1 \le j \le M-1,\\ \widetilde{\pi}_j =& \frac{1}{M}, \qquad \text{for } 1 \le j \le M.\end{aligned}$$

For each i, set the non-zero values of $\widehat{q}_{\ell|i}$ and $\widetilde{q}_{\ell|i}$ to have uniform distributions. That is,

$$\begin{aligned}\widehat{q}_{\ell|i} =& \frac{1}{i-1}, \quad \text{for } 2 \le i \le M, 1 \le \ell \le i-1,\\ \widehat{q}_{1|1} =& 1,\\ \widetilde{q}_{\ell|i} =& \frac{1}{M-i}, \quad \text{for } 2 \le i \le M-1, i+1 \le \ell \le M,\\ \widehat{q}_{\ell|1} =& \frac{1}{M}, \quad \text{for } 1 \le \ell \le M.\end{aligned}$$

Then for each step and using the system of equations, we calculate new values for $\widehat{q}_{\ell|i}$ and $\widetilde{q}_{\ell|i}$, then evaluate new values for $\widehat{p}_{ij}$ and $\widetilde{p}_{ij}$, then evaluate new values for $\widehat{\pi}_j$ and $\widetilde{\pi}_j$. We iterate this process until some convergence criterion is met. In this case, we simply checked whether the new values were within an acceptable tolerance of the previous values.

This method converges reasonably quickly to the same solution found by the MATLAB solver, as shown in Figure 7.1.

To further confirm that both the iterative approach and the MATLAB solver converge to the correct solution we simulate the original queue, truncated so that the queue length will never exceed M.

This truncated simulation is similar to the simulation process described in Section 2.1.5, with some minor changes. When the queue length after an arrival event is M, then a new arrival is not generated. Instead, if the queue length after a departure event is $M-1$, then a new arrival event is generated. From this truncated simulation, the empirical distributions of $\{\widehat{\pi}_j : 1 \le j \le M-1\}$ and $\{\widetilde{\pi}_j : 1 \le j \le M\}$ can be calculated. These converge to the same solution found by the iterative method to solving the system of equations, as shown in Figure 7.2.

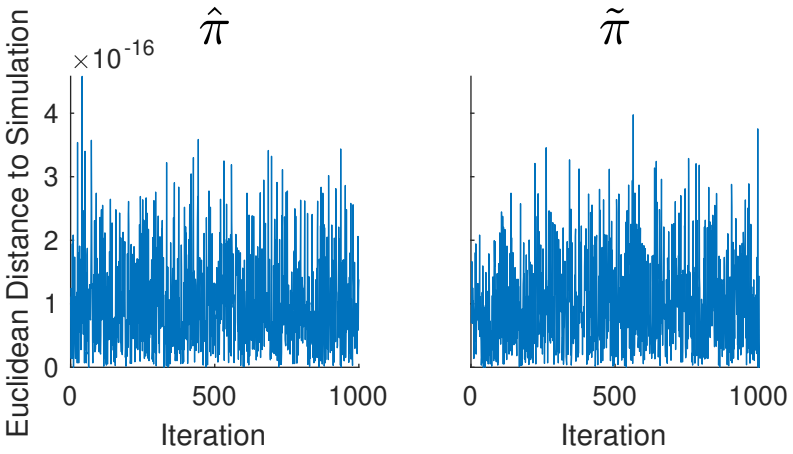
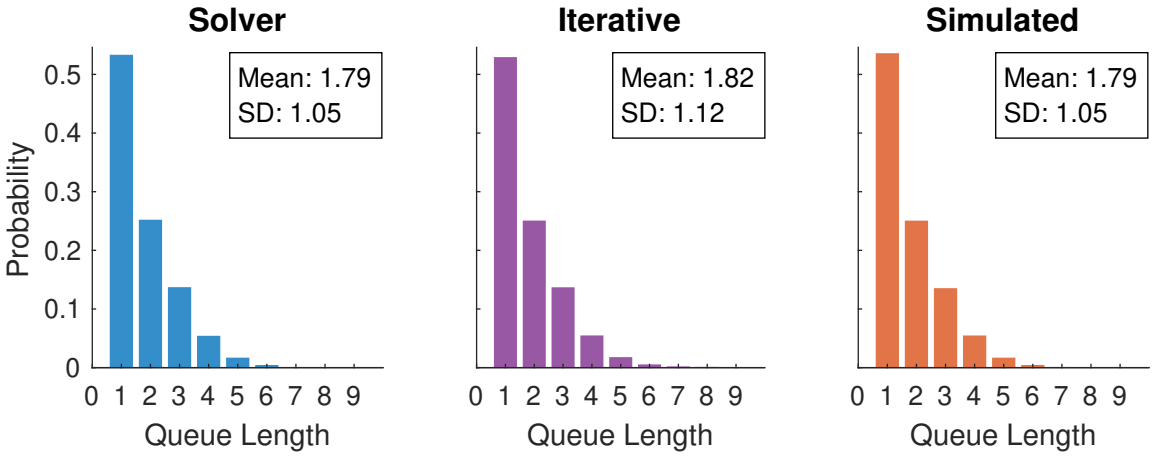
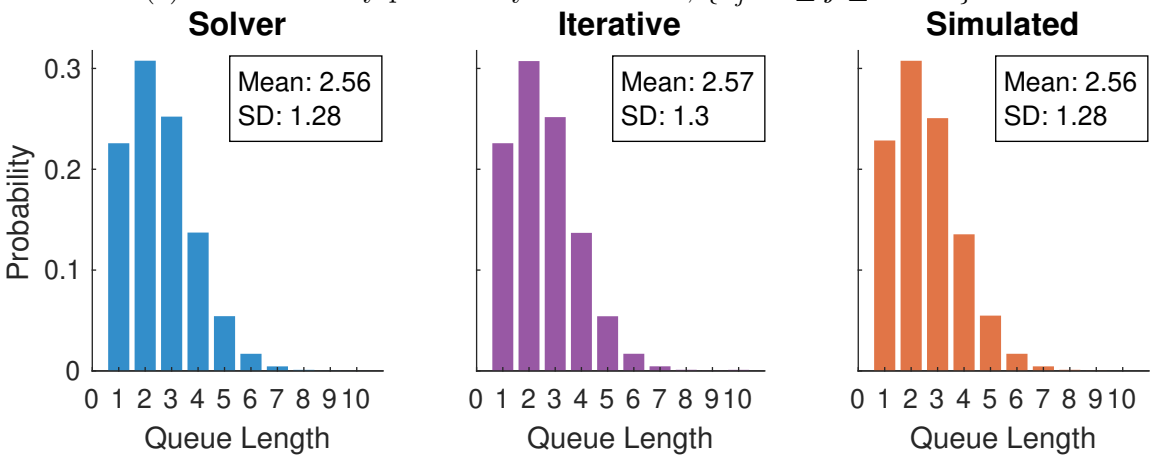


Figure 7.1: The Euclidean distance between the each step of the iterative process and the MATLAB solver solution. The parameters are $\lambda_i = 5/i$, $\mu_j = 1 + j$ and $M = 10$.



(a) The stationary probability distribution, $\{\widehat{\pi}_j : 1 \leq j \leq M-1\}$.



(b) The stationary probability distribution, $\{\widetilde{\pi}_j : 1 \leq j \leq M\}$.

Figure 7.2: The embedded DTMC stationary probability distributions, calculated using the MATLAB solver, truncated simulations (run for 20,000 time steps), and the iterative method. The parameters are $\lambda_i = 5/i$, $\mu_j = 1 + j$ and $M = 10$.

7.3 s-perm Semi-Experiment Model

Now we can construct a QBD model for the s-perm semi-experiment applied to this queue with queue-length-dependent arrival and service rates. This is a combination of the (QL, S, sSE) and (QL, A, sSE) QBD models.

As with the (QL, S, sSE), the service times will have a hyperexponential distribution with rates $\{\mu_j : j \geq 1\}$ and mixture distribution $\{\widehat{\pi}_j : j \geq 1\}$.

The arrival process will be similar to that in the (QL, A, sSE) model. We include an underlying queue which replicates the original queue. This semi-experiment queue will share the exact arrival process of the underlying queue and the underlying queue's service process is independent of the rest of the semi-experiment queue.

So this s-perm semi-experiment queue is a single-server queue such that the:

- **Arrival Process** is the same arrival stream as generated by the underlying queue which replicates the original queue.
- **Service Times** have a hyperexponential distribution with rates $\{\mu_j : j \geq 1\}$ and mixture distribution $\{\widehat{\pi}_j : j \geq 1\}$.

Now we consider the state space of the QBD model for this queue. There are a number of pieces of information to track. Let the state for this model be (n, m, k, i, j), where

- n is the queue length of the s-perm semi-experiment queue,
- m indicates that the current service rate for the s-perm semi-experiment queue is μ_m,
- k is the queue length of the underlying queue,
- i is the number of departures in the underlying queue since the last arrival,
- j is the queue length at the start of the current service in the underlying queue when the queue is busy $(k \geq 1)$ and 1 if the queue is idle $(k = 0)$.

The state space for this process is given by $S = \{(n, m, k, i, j) : n \geq 0, m \geq 1, k \geq 0, i \geq 0, k + i \geq 1, 1 \leq j \leq k\}$.

The full set of possible transitions for this QBD are given in Table 7.2.

From	To	Rate	For	Description
(n, m, k, i, j)	$(n+1, m, k+1, 0, j)$	λ_{k+i}	$k \geq 1, i \geq 0$	Underlying and sSE arrival
$(n, m, 0, i, 1)$	$(n+1, m, 1, 0, 1)$	λ_i	$i \geq 1$	Underlying and sSE arrival
(n, m, k, i, j)	$(n, m, k-1, i+1, k-1)$	μ_j	$k \geq 2$	Underlying departure
$(n, m, 1, i, 1)$	$(n, m, 0, i+1, 1)$	μ_1		Underlying departure
(n, m, k, i, j)	$(n-1, r, k, i, j)$	$\mu_m \widehat{\pi}_r$	$n \geq 1$	sSE departure

Table 7.2: Possible transitions for the $(QL, AS, 1, sSE)$ process (full state space).

Note that the sub-phases k, i, j all relate to the underlying queue. The underlying queue only affects the s-perm queue when an arrival occurs and otherwise operates independently. Further, the underlying queue replicates the original queue. Let $U_+^{(k)}, U_-^{(k)}, U_0^{(k)}$ be the same matrices as $A_+^{(k)}, A_-^{(k)}, A_0^{(k)}$ of the original queue in Section 7.1. Then $U_+^{(k)}, U_-^{(k)}, U_0^{(k)}$ are the model matrices for the underlying process with k as the level and (i, j) as the phase.

Let $\ell(k)$ be the set of all sub-phases i, j when the underlying queue length is k, such that $(n, m, k, i, j) \in (n, m, \ell(k))$. That is $\ell(k) = \{(k, i, j) : 1 \leq j \leq k, i \geq 0, k + i \geq 1\}$. Note that in this case $\ell(k)$ refers to the states in the underlying queue which has a queue length of k and not the overall semi-experiment queue, as has been used previously. Then the transitions can be simplified as shown in Table 7.3. The matrices $U_0^{(k)}$ contain the negative sums of rates and are not shown in the table, but are used in the A_0 matrix to achieve the correct negative sums of rates of the overall system.

From	To	Rate	For	Description
$(n, m, \ell(k))$	$(n+1, m, \ell(k+1))$	$U_+^{(k)}$		Underlying and sSE arrival
$(n, m, \ell(k))$	$(n, m, \ell(k-1))$	$U_-^{(k)}$	$k \geq 1$	Underlying departure
$(n, m, \ell(k))$	$(n-1, r, \ell(k))$	$\mu_m \widehat{\pi}_r$	$n \geq 1$	sSE departure

Table 7.3: Possible transitions for the $(QL, AS, 1, sSE)$ process (compact state space).

Now we can construct the model matrices, A_+, A_-, A_0, B_0 for the s-perm semi-experiment QBD model.

First, let V_+ be a matrix containing the transition rates in which $n \to n+1$ for a fixed $m \to m$. When this occurs $\ell(k) \to \ell(k+1)$ and these rates are contained

in the matrices $U_+^{(k)}$. Then,

$$V_+ = \begin{array}{c} \\ \ell(0) \\ \ell(1) \\ \ell(2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} \ell(0) & \ell(1) & \ell(2) & \ell(3) & \cdots \end{array} \\ \begin{bmatrix} \mathbf{0} & U_+^{(0)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & U_+^{(1)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & U_+^{(2)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Let $L(n,m) = \{(n,m,\ell(k)) : k \geq 0\}$ be the set of phases when the s-perm queue length is n and the current s-perm service rate is m.

Then, for $n \geq 0$,

$$A_+ = \begin{array}{c} \\ L(n,1) \\ L(n,2) \\ L(n,3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} L(n+1,1) & L(n+1,2) & L(n+1,3) & \cdots \end{array} \\ \begin{bmatrix} V_+ & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & V_+ & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & V_+ & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Let $V_-^{(m,r)}$ contain the transition rates in which $n \to n-1$ and $m \to r$ and let $Z_-^{(m,r)} = \mu_m \widehat{\pi}_r \mathcal{I}$. Then,

$$V_-^{(m,r)} = \begin{array}{c} \\ \ell(0) \\ \ell(1) \\ \ell(2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} \ell(0) & \ell(1) & \ell(2) & \cdots \end{array} \\ \begin{bmatrix} Z_-^{(m,r)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & Z_-^{(m,r)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & Z_-^{(m,r)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

That is, $V_-^{(m,r)}$ is a diagonal matrix where every diagonal element is $\mu_m \widehat{\pi}_r$.

Then, for $n \geq 1$,

$$A_- = \begin{array}{c} \\ L(n,1) \\ L(n,2) \\ L(n,3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} L(n-1,1) & L(n-1,2) & L(n-1,3) & \cdots \end{array} \\ \begin{bmatrix} V_-^{(1,1)} & V_-^{(1,2)} & V_-^{(1,3)} & \cdots \\ V_-^{(2,1)} & V_-^{(2,2)} & V_-^{(2,3)} & \cdots \\ V_-^{(3,1)} & V_-^{(3,2)} & V_-^{(3,3)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Let $V_0^{(m)}$ contain the transition rates in which $n \to n$ and $m \to m$ for a fixed m. We also need to ensure that the diagonal of A_0 contains the negative sum of all the rates out of each state. Since $V_0^{(m)}$ will lie on the diagonal of A_0, we must have the negative sum of rates in the diagonals of the matrices $V_0^{(m)}$. Let $Z_0^{(m,k)}$ be a diagonal matrix containing the negative sum of the transition rates out of the states $(n, m, \ell(k))$. Then, $Z_0^{(m,k)} = U_0^{(k)} - \mu_m \mathcal{I}$, since $U_0^{(k)}$ is already a diagonal matrix containing the negative sum of transition rates for the underlying process and the only other rate to consider is the s-perm queue service rates, μ_m. Hence,

$$V_0^{(m)} = \begin{array}{c} \\ \ell(0) \\ \ell(1) \\ \ell(2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} \ell(0) & \ell(1) & \ell(2) & \cdots \end{array} \\ \begin{bmatrix} Z_0^{(m,0)} & \mathbf{0} & \mathbf{0} & \cdots \\ U_-^{(1)} & Z_0^{(m,1)} & \mathbf{0} & \cdots \\ \mathbf{0} & U_-^{(2)} & Z_0^{(m,2)} & \cdots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix} \end{array}.$$

Then, for $n \geq 1$,

$$A_0 = \begin{array}{c} \\ L(n,1) \\ L(n,2) \\ L(n,3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} L(n,1) & L(n,2) & L(n,3) & \cdots \end{array} \\ \begin{bmatrix} V_0^{(1)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & V_0^{(2)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & V_0^{(3)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

The structure of B_0 is very similar except that since $n = 0$, we cannot have s-perm queue departures, so the negative sum of rates is simply $U_0^{(k)}$. So, let

$$V_B = \begin{array}{c} \\ \ell(0) \\ \ell(1) \\ \ell(2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} \ell(0) & \ell(1) & \ell(2) & \cdots \end{array} \\ \begin{bmatrix} U_0^{(0)} & \mathbf{0} & \mathbf{0} & \cdots \\ U_-^{(1)} & U_0^{(1)} & \mathbf{0} & \cdots \\ \mathbf{0} & U_-^{(2)} & U_0^{(2)} & \cdots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix} \end{array},$$

$$B_0 = \begin{array}{c} \\ L(0,1) \\ L(0,2) \\ L(0,3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} L(0,1) & L(0,2) & L(0,3) & \cdots \end{array} \\ \begin{bmatrix} V_B & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & V_B & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & V_B & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

7.3.1 Truncation and Augmentation

As before, we need to truncate the state space for this model to make it finite and augment the transition rate matrices appropriately. Let M_1 be the artificial maximum queue length for the original queue so that $1 \leq m \leq M_1$. Let M_2 be the artificial maximum queue length for the underlying queue, so that $0 \leq k \leq M_2$ and $0 \leq i \leq M_2$. It would generally be suitable to have $M_1 = M_2$ since the underlying queue replicates the original queue, and so an appropriate upper bound for one would be appropriate for the other. The truncated state space is then $S = \{(n, m, k, i, j) : n \geq 0, 1 \leq m \leq M_1, 0 \leq k \leq M_2, 1 \leq j \leq k, 0 \leq i \leq M_2, 1 \leq k + i \leq M_2\}$.

From Section 7.1.2, and letting $N = M_2$, note that

$U_+^{(k)}$ is $k(M_2 - k + 1) \times (k + 1)(M_2 - k)$, for $k \geq 1$
$U_+^{(0)}$ is $M_2 \times M_2$,
$U_-^{(k)}$ is $k(M_2 - k + 1) \times (k - 1)(M_2 - k + 2)$, for $k \geq 2$
$U_-^{(1)}$ is $M_2 \times M_2$,
$U_0^{(k)}$ is $k(M_2 - k + 1) \times k(M_2 - k + 1)$, for $k \geq 1$
$U_0^{(0)}$ is $M_2 \times M_2$.

These matrices are as shown in Section 7.1.2.

So, V_+ is now a square matrix with dimension $M_2 + \sum_{k=1}^{M_2} k(M_2 - k + 1) = M_2 + \frac{1}{6}M(M + 1)(M + 2)$. In the states $\ell(M_2)$, there cannot be further arrivals since M_2 is the maximum queue length for the underlying queue. So,

$$V_+ = \begin{array}{c} \\ \ell(0) \\ \ell(1) \\ \ell(2) \\ \vdots \\ \ell(M_2-1) \\ \ell(M_2) \end{array} \begin{array}{c} \begin{array}{cccccc} \ell(0) & \ell(1) & \ell(2) & \ell(3) & \cdots & \ell(M_2) \end{array} \\ \begin{bmatrix} \mathbf{0} & U_+^{(0)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & U_+^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & U_+^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & U_+^{(M_2-1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \end{array}.$$

The structure of A_+ remains the same, with it being truncated to be a square matrix of size $M_1(M_2 + \frac{1}{6}M(M + 1)(M + 2))$.

$V_-^{(m,r)} = \mu_m \widehat{\pi}_r \mathcal{I}$ of size $M_2 + \frac{1}{6}M(M+1)(M+2)$. Since $\widehat{\pi}_M = 0$, $V_-^{(m,M)} = \mathbf{0}$ of the same size.

Then, A_- has the same structure and is truncated to be a square matrix of size $M_1(M_2 + \frac{1}{6}M(M+1)(M+2))$.

$V_0^{(m)}$ is a square matrix of size $M_2 + \frac{1}{6}M(M+1)(M+2)$ and V_B is also a square matrix of size $M_2 + \frac{1}{6}M(M+1)(M+2)$.

A_0 and B_0 are truncated to be square matrices of size $M_1(M_2 + \frac{1}{6}M(M+1)(M+2))$.

7.3.2 Results

Figure 7.3 shows the queue-length distributions for the original, empirical s-perm semi-experiment and model s-perm semi-experiment queues. This, along with a KS test, indicates that the model accurately captures the behaviour of the empirical semi-experiment. As with previous models, the semi-experiment queue-length has a more exponential shape than the original, indicating that some of the dependence has been accounted for. However, since there is still dependence in the arrival process that is not disrupted by an s-perm semi-experiment, which causes the departures away from the exponential shape.

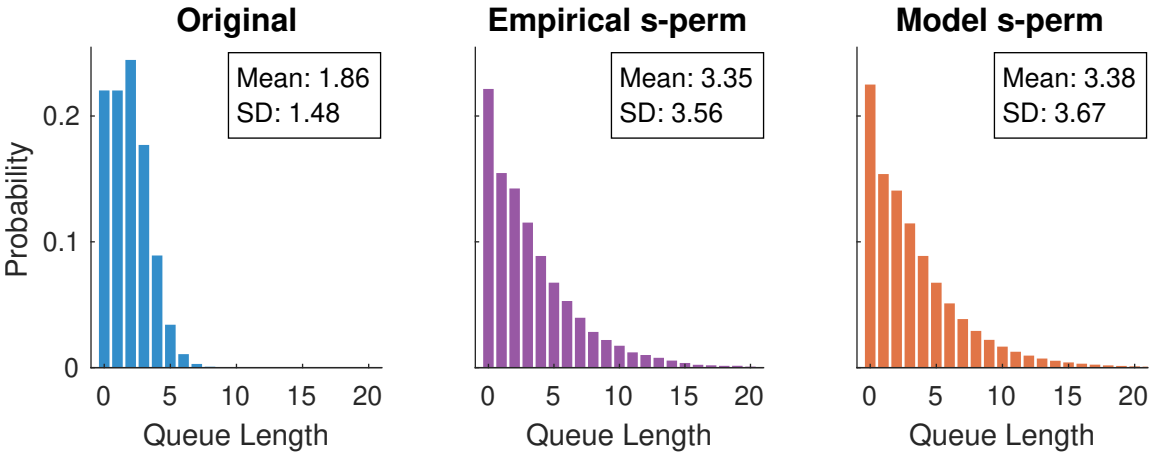


Figure 7.3: The blue and orange plots show the stationary queue-length distribution of the original (QL, AS, O) model and s-perm semi-experiment (QL, AS, sSE) model, respectively. The purple plots show the empirical queue length distribution of an empirical s-perm semi-experiment. That is, a single realisation of the original queue (run for 20,000 time steps) with a single random permutation of the service times. The arrival rate is $\lambda_i = 5/i$ and the service rate is $\mu_j = j + 1$. The KS statistic between the original and semi-experiment models is 0.227600.

7.4 a-perm Semi-Experiment Model

Now we can construct a QBD model for the a-perm semi-experiment applied to this model with both arrival rates and service rates dependent on the queue length. It is basically a combination of the (QL, S, aSE) and (QL, A, aSE) models and is in some sense the reverse of the s-perm model above.

The inter-arrival times are permuted, so they are exponentially distributed with the rates λ_i, according to the proportions $\widetilde{\pi}_i$ in which they occur in the original queue. That is, they have a hyperexponential distribution.

We create an underlying queue which replicates the original queue in order to generate an appropriate sequence of service times. This underlying queue is modified in exactly the same way as for the (QL, S, aSE) in Section 6.2.1. That is, the underlying queue will never become empty (it skips its idle period) and the underlying queue is 'paused' when the a-perm semi-experiment queue is empty. This ensures a correct sequence of service times is generated.

This a-perm semi-experiment model (QL, AS, aSE) is a single-server queue such that the:

- **Arrival Process** has inter-arrival times with a hyperexponential distribution with rates $\{\lambda_i : i \geq 1\}$ and mixture distribution $\{\widetilde{\pi}_i : i \geq 1\}$.
- **Service Times** are the sequence of service times generated by the modified underlying queue.

Now we can consider the state space of the QBD model of this queue. Let the state be (n, m, k, i, j), where

- n is the queue length of the a-perm semi-experiment queue,
- m indicates that the current arrival rate for the a-perm semi-experiment queue is λ_m,
- k is the queue length of the underlying queue,
- i is the number of departures in the underlying queue since the last arrival,
- j is the queue length at the start of the current service in the underlying queue.

The state space is given by $S = \{(n, m, k, i, j) : n \geq 0, m \geq 1, k \geq 1, 1 \leq j \leq k, i \geq 0\}$. As in the previous section, this state space can be compacted by letting $\ell(k) = \{(k, i, j) : 1 \leq j \leq k, i \geq 0\}$.

Let $U_+^{(k)}, U_-^{(k)}$ and $U_0^{(k)}$ be the matrices containing the transition rates for the modified underlying process. They are almost identical to the matrices $A_+^{(n)}, A_-^{(n)}, A_0^{(n)}$ for the original queue in Section 7.1, except that they do not allow $k = 0$. That is, we do not have $U_+^{(0)}$ and $U_0^{(0)}$. Further, $U_-^{(1)}$ is not $A_-^{(1)}$ from Section 7.1 since we cannot transition to $k = 0$. Instead, this transition represents a departure in the a-perm queue, and a departure in the underlying queue, where the idle period is immediately skipped back to the next arrival when $k = 1$. Hence, we should have the transition $i \to 0$ since there has been an arrival. So, define

$$U_-^* = \begin{array}{c} \\ (1,0,1) \\ (1,1,1) \\ (1,2,1) \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} (1,0,1) & (1,1,1) & (1,2,1) & (1,3,1) & \cdots \end{array} \\ \begin{bmatrix} \mu_1 & 0 & 0 & 0 & \cdots \\ \mu_1 & 0 & 0 & 0 & \cdots \\ \mu_1 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Then the transitions for this process are given in Table 7.4.

From	To	Rate	For	Description
$(n, m, \ell(k))$	$(n+1, r, \ell(k))$	$\lambda_m \widetilde{\pi}_r$		aSE arrival
$(n, m, \ell(k))$	$(n, m, \ell(k+1))$	$U_+^{(k)}$	$n \geq 1$	Underlying arrival
$(n, m, \ell(k))$	$(n-1, m, \ell(k-1))$	$U_-^{(k)}$	$n \geq 1, k \geq 2$	Underlying and aSE departure
$(n, m, \ell(1))$	$(n-1, m, \ell(1))$	U_-^*	$n \geq 1$	Underlying and aSE departure

Table 7.4: Possible transitions for the (QL, AS, aSE) process (compact state space).

Now we can construct the model matrices for this QBD process.

Let $Z_+^{(m,r)} = \lambda_m \widetilde{\pi}_r \mathcal{I}$. Let $V_+^{(m,r)}$ be matrices containing the tranistion rates such that $n \to n+1$ and $m \to r$. Then,

$$V_+^{(m,r)} = \begin{array}{c} \\ \ell(1) \\ \ell(2) \\ \ell(3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} \ell(1) & \ell(2) & \ell(3) & \cdots \end{array} \\ \begin{bmatrix} Z_+^{(m,r)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & Z_+^{(m,r)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & Z_+^{(m,r)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

That is, $V_+^{(m,r)}$ is a diagonal matrix where each diagonal element is $\lambda_m \widetilde{\pi}_r$.

Let $L(n, m) = \{(n, m, \ell(k)) : k \geq 1\}$ be the set of phases when the a-perm queue length is n and the current a-perm arrival rate is m.

Then, for $n \geq 0$,

$$A_+ = \begin{array}{c} \\ L(n,1) \\ L(n,2) \\ L(n,3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} L(n+1,1) & L(n+1,2) & L(n+1,3) & \cdots \end{array} \\ \begin{bmatrix} V_+^{(1,1)} & V_+^{(1,2)} & V_+^{(1,3)} & \cdots \\ V_+^{(2,1)} & V_+^{(2,2)} & V_+^{(2,3)} & \cdots \\ V_+^{(3,1)} & V_+^{(3,2)} & V_+^{(3,3)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Let V_- be a matrix containing all the transition rates such that $n \to n-1$

and $m \to m$ for a fixed m. Then,

$$V_- = \begin{array}{c} \\ \ell(1) \\ \ell(2) \\ \ell(3) \\ \ell(4) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} \ell(1) & \ell(2) & \ell(3) & \cdots \end{array} \\ \begin{bmatrix} U_-^{(1)} & \mathbf{0} & \mathbf{0} & \cdots \\ U_-^{(2)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & U_-^{(3)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & U_-^{(4)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Note that the rates $U_-^{(1)}$ between $\ell(1)$ and $\ell(1)$ allow there to be a departure in the a-perm semi-experiment queue, while not allowing the underlying queue to become empty. The transition is back to $\ell(1)$ because after an empty period, the next service rate will begin when a customer enters the queue with a queue length of $k = 1$.

Then, for $n \geq 1$,

$$A_- = \begin{array}{c} \\ L(n,1) \\ L(n,2) \\ L(n,3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} L(n+1,1) & L(n+1,2) & L(n+1,3) & \cdots \end{array} \\ \begin{bmatrix} V_- & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & V_- & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & V_- & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Let $Z_0^{(m,k)}$ be a diagonal matrix containing the negative sum of transition rates out of the states $(n, m, \ell(k))$. Then, $Z_0^{(m,k)} = U_0^{(k)} - \lambda_m \mathcal{I}$. Let $V_0^{(m)}$ contain the transition rates in which $n \to n$ and $m \to m$. Then,

$$V_0^{(m)} = \begin{array}{c} \\ \ell(1) \\ \ell(2) \\ \ell(3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} \ell(1) & \ell(2) & \ell(3) & \cdots \end{array} \\ \begin{bmatrix} Z_0^{(m,1)} & U_+^{(1)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & Z_0^{(m,2)} & U_+^{(2)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & Z_0^{(m,3)} & U_+^{(3)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

Then, for $n \geq 1$,

$$A_0 = \begin{array}{c} \\ L(n,1) \\ L(n,2) \\ L(n,3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} L(n+1,1) & L(n+1,2) & L(n+1,3) & \cdots \end{array} \\ \begin{bmatrix} V_0^{(1)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & V_0^{(2)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & V_0^{(3)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}.$$

When $n = 0$, the underlying queue is 'paused' so there are no underlying arrivals or departures. The only events that can occur are the a-perm arrivals. So $B_0 = -diag(A_+\boldsymbol{e})$.

7.4.1 Truncation and Augmentation

The truncation for this process is very similar to that for the s-perm semi-experiment queue. Let M_1 be an artificial maximum for the queue length of the original queue, so that $1 \leq m \leq M_1$. Let M_2 be the artificial maximum queue length for the underlying queue. As before, it would be sensible to have $M_1 = M_2$. Now the truncated state space is $S = \{(n, m, k, i, j) : n \geq 0, 1 \leq m \leq M_1, 1 \leq k \leq M_2, 0 \leq i \leq M_2, 1 \leq j \leq k\}$.

From Section 7.1, by setting $N = M_2$, and adjusting for $k \neq 0$, we have that

$$\begin{aligned} U_+^{(k)} &\text{ is } k(M_2 - k + 1) \times (k+1)(M_2 - k), \text{ for } k \geq 1 \\ U_-^{(k)} &\text{ is } k(M_2 - k + 1) \times (k-1)(M_2 - k + 2), \text{ for } k \geq 2 \\ U_-^{*} &\text{ is } M_2 \times M_2, \\ U_0^{(k)} &\text{ is } k(M_2 - k + 1) \times k(M_2 - k + 1), \text{ for } k \geq 1 \end{aligned}$$

These matrices are as shown in Section 7.1.2.

Now, $V_+^{(m,r)} = \lambda_m \widetilde{\pi}_r \mathcal{I}$ with dimension $\frac{1}{6}M_2(M_2^2 + 3M_2 + 2)$. Also, V_- and $V_0^{(m)}$ are square matrices of dimension $\frac{1}{6}M_2(M_2^2 + 3M_2 + 2)$.

Then, A_+, A_-, A_0 and B_0 are square matrices with dimension $\frac{1}{6}M_1M_2(M_2^2 + 3M_2 + 2)$.

7.4.2 Results

Figure 7.4 shows the queue-length distributions for the original, empirical a-perm semi-experiment and model a-perm semi-experiment queues. This, along with a KS test, indicates that the model accurately captures the behaviour of the empirical semi-experiment. We can see a similar shape to the semi-experiment queue-length distribution as in the (QL, AS, sSE) model where the dependence is broken in the arrival stream, but the dependence in the service stream is still present, resulting in a shape that is not quite exponential.

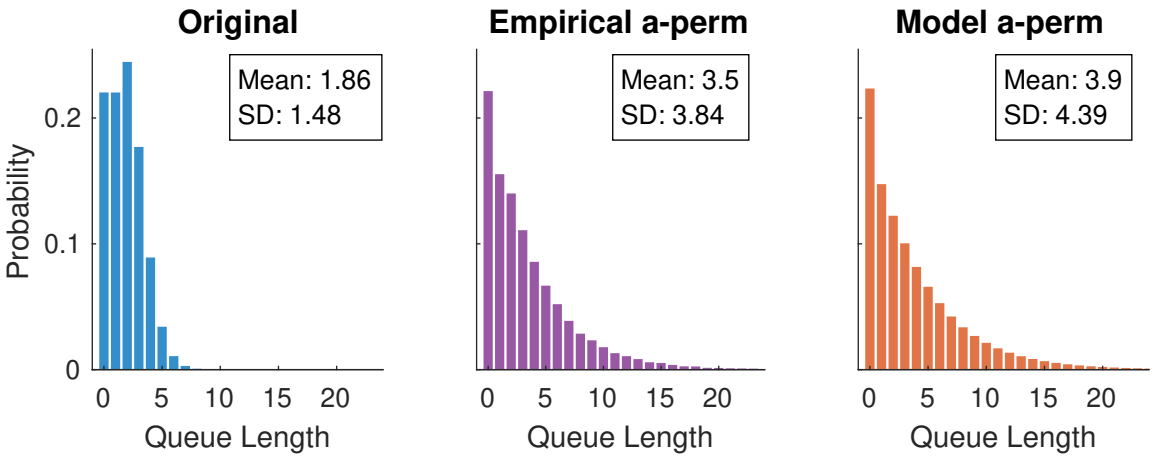


Figure 7.4: The blue and orange plots show the stationary queue-length distribution of the original (QL, AS, O) model and s-perm semi-experiment (QL, AS, aSE) model, respectively. The purple plots show the empirical queue length distribution of an empirical a-perm semi-experiment. That is, a single realisation of the original queue (run for 20,000 time steps) with a single random permutation of the inter-arrival times. The arrival rate is $\lambda_i = 5/i$ and the service rate is $\mu_j = j + 1$. The KS statistic between the original and semi-experiment models is 0.269500.

7.5 Conclusion

In this chapter we constructed a queue with both queue-length-dependent arrival and service rates and used a queue-length-dependent QBD model to find the stationary queue-length distribution. We then considered the embedded DTMCs at the arrival epochs and the start of service epochs. In order to construct QBD models for the s-perm and a-perm semi-experiments, we required the stationary probability distribution of these embedded DTMCs. So we constructed a system of equations and an iterative method of solution to evaluate these. Then we constructed the s-perm and a-perm semi-experiment model, which were largely combinations of the previous semi-experiment models in Chapters 4, 5 and 6. Finally, we compared the stationary queue-length distributions of the semi-experiment queues to the original and found that the s-perm and a-perm semi-experiments have different effects.

Chapter 8

Restricted Semi-Experiments

8.1 Introduction

Now we explore restricted semi-experiments. So far, we have seen two types of semi-experiments, the s-perm, which permutes service times, and the a-perm, which permutes inter-arrival times. We have already seen that these give different results depending on the form of dependence in the original queue. Restricted semi-experiments can provide even more information about the nature of the dependence in the original queueing data.

First we will briefly consider restricting by customer classes. Then we will focus on queue-length-restricted semi-experiments: restricting a-perm semi-experiments by the queue-length at the start of inter-arrival periods, and restricting s-perm semi-experiments by the queue-length at the start of service periods. These are only two examples of queue-length-restricted semi-experiments. It would be possible to restrict at any epochs of queue-lengths for either a-perm or s-perm semi-experiments, but the two we use are the most logical to apply to the queue-length-dependent queue models we have developed.

Note that the queue-length-restricted semi-experiments actually use sampling with replacement rather than permutation (sampling without replacement), which is explained below. However, we will still use the terms a-perm and s-perm to refer to these semi-experiments for consistency and to highlight the similarity to previous unrestricted semi-experiments.

Note that this sampling with replacement does not cause the restricted

semi-experiments to have the same probability of an empty queue as the original queue, in contrast to the discussion in Section 2.6.3.

8.2 Class-Restricted SEs

The first and most intuitive way to restrict semi-experiments is to separate customers into sensible classes. This was done by Varney *et al.* [1] where patients entering the ICU were classified by whether they were there for emergency reasons or elective surgery, the time of day they were admitted, the type of ailment they suffered from and more. The reason to classify customers like this is because it may be reasonable to assume that their behaviour is different as they may require different lengths of service or may arrive with different rates. Then the restricted semi-experiment can either permute the service times or inter-arrival times of each customer within those of the same class. If the class is known with certainty, then this will return a realisation of the original queue for models like the one with pairwise rates such as those in Section 3.1 and auto-dependence such as those in Section 3.2.

So if there is no significant difference between the original and restricted semi-experiment, then the dependence between the arrivals and services can be fully explained by the customer classes. Even if there is additional dependence in the queue, by restricting permutations within the classes, this reduces the dependence that may then require further exploration.

8.3 Performing Queue-Length-Restricted SEs

The goal is to perform queue-length-restricted s-perm and a-perm semi-experiments. This restriction should be logical from a modelling point of view, such as relating the queue length to each service time being permuted. It should also be possible to perform. This restriction should also be able to inform about the nature of dependence in the models that we have. In particular, we use queue-length-restricted semi-experiments to identify the dependence present in the (QL, S, O) and (QL, A, O) queues.

One way we might consider performing these queue-length-restricted semi-experiments is as follows. We label each customer in the original queue with the class of the queue length at the start of their service. Then we can permute all

of the service times within customer classes. However, once this permutation is performed, the queue length process is changed. Hence, those service times associated with a particular queue-length at the start of service no longer are attributed to customers of that class. So the pre-labelling of customer classes before permutation is a problem and we should perform the semi-experiment process and dynamically select appropriate service times for each customer once we know their queue length at the start of service in the semi-experiment queue. However, there may be a different number of customers in each class than in the original queue, so we cannot perform a permutation (sampling without replacement) as we may not have sufficient service times.

Therefore a different method is proposed. Consider an s-perm semi-experiment which is restricted according to the queue-length at the start of service. The basic steps of the restricted semi-experiment is as follows.

- Take the arrival stream and service stream from the observed queue
- Evaluate the queue-length at the start of service for each customer
- Keep the arrival stream and pool the service times according to their queue-length at the start of service
- 'Run' the semi-experiment queue by having each customer arrive according to the arrival stream. Then calculate what the queue-length will be at the start of their service.
- Sample with replacement a service time for that customer from the pool with the appropriate queue-length at the start of service.

Note that in this case, we sample with replacement instead of permuting because the number of required service times from each pool may be greater than the number in that pool. With 'enough' data, the change to sampling with replacement instead of permuting should not be a significant weakness. Also note that the semi-experiment queue may reach queue-lengths at the start of service that were not attained in the original queue and hence no service times will be available to sample from. In this case, the service times are sampled from the nearest queue-length pool. This only occurs when the required queue-length is larger than the maximum queue-length at the start of service observed in the original queue, so the nearest pool is that of the maximum. Note that when these queue-lengths get large, the amount of data available becomes very small. This can result in strange and insignificant tail behaviour of distributions. So, let c be

the cut-off for the acceptable number of data points in each pool, and let P_{K^*} be the last pool with at least c data points. All pools after P_{K^*} are merged into P_{K^*}.

This method is formally described in Algorithm 9. Let A^*_m, S^*_m and Q^*_m be the arrival time, length of service time and queue-length at the start of service of the mth customer in the original queue, respectively. Let $A_m, S_m, D_m, \widehat{t}_m$ and $\widehat{Q}_m$ be the arrival time, length of service time, departure time, start of service time, and queue-length at the start of service of the mth customer in the restricted semi-experiment queue, respectively. Let M be the number of customers that visit the original and semi-experiment queue during the simulation. It is assumed that the queues start empty so the first service will start when the first arrival occurs with a queue length of 1.

A similar method is used for the queue-length-restricted a-perm semi-experiment which is restricted by the queue-length immediately after the previous arrival. The general method is

- Take the arrival stream and service stream from a simulated queue
- Evaluate the queue-length immediately after each arrival
- Keep the service stream and pool the inter-arrival times according to queue-length at the start of the inter-arrival period (the queue length immediately after the preceding arrival)
- One customer at a time, calculate the queue-length immediately after their arrival in the semi-experiment queue
- Sample with replacement an inter-arrival time for the next customer from the pool with the appropriate queue-length at the start of the inter-arrival period.

This method is formally described in Algorithm 10. Let Let A^*_m, S^*_m and Q^*_m be the arrival time, length of service time and queue-length at the start of service of the mth customer in the original queue, respectively. Let A_m, S_m, D_m and $\widetilde{Q}_m$ be the arrival time, length of service time, departure time, and queue-length immediately after arrival of the mth customer in the restricted semi-experiment queue, respectively. Let $T^*_m = A^*_m - A^*_{m-1}$ and $T_m = A_m - A_{m-1}$ be the inter-arrival times of the mth customer in the original queue and the semi-experiment queue, respectively and $T^*_1 = A^*_1$ and $T_1 = A_1$. Let M be the total number of customers. It is assumed that the queue starts empty and that the first inter-arrival time has

an associated queue-length of 1 by definition. Note that we once again merge pools of size less than c.

Algorithm 9: The algorithm for a queue-length-restricted s-perm semi-experiment which is restricted at the queue-length at the start of service.

Input: $A_m^*, S_m^*, \widehat{Q}_m^*$ for $m = 1, \ldots, M$, c
Set $A_m = A_m^*$ for $m = 1, \ldots, M$
Set $K = \max\{\widehat{Q}_m^* : m = 1, \ldots, M\}$
Set $P_k = \{S_m^* : \widehat{Q}_m^* = k, m = 1, \ldots, M\}$ for $k = 1, \ldots, K$
Set $K^* = \inf\{k = 1, 2, \ldots, K \mid |P_k| \leq c\}$
Set $P_{K^*} = \bigcup_{k=K^*}^{K} P_k$
Set $\widehat{Q}_1 = 1$ and $\widehat{t}_1 = A_1$
Sample with replacement S_1 from P_1
Set $D_1 = A_1 + S_1$
for $m = 2, \ldots, M$ **do**
  if $A_m > D_{m-1}$ **then**
    Set $\widehat{t}_m = A_m$
    Set $\widehat{Q}_m = 1$
    Sample with replacement S_m from P_1
    Set $D_m = A_m + S_m$
  else
    Set $\widehat{t}_m = D_{m-1}$
    Calculate $n = \sum_{i=1}^{M} \mathcal{I}\{\widehat{t}_{m-1} \leq A_i < \widehat{t}_m\}$
    Set $\widehat{Q}_m = \widehat{Q}_{m-1} + n - 1$
    if $\widehat{Q}_m \leq K^*$ **then**
      Sample with replacement S_m from $P_{\widehat{Q}_m}$
    else
      Sample with replacement S_m from P_{K^*}
    end if
    Set $D_m = D_{m-1} + S_m$
  end if
end for

Algorithm 10: The algorithm for a queue-length-restricted a-perm semi-experiment which is restricted at the queue-length at the start of inter-arrival periods.

Input: $A^*_m, S^*_m, \widetilde{Q}^*_m$ for $m = 1, \ldots, M$, c
Set $S_m = S^*_m$ for $m = 1, \ldots, M$
Set $K = \max\{\widetilde{Q}^*_m : m = 1, \ldots, M\}$
Set $P_k = \{T^*_m : \widetilde{Q}^*_{m-1} = k, m = 1, \ldots, M\}$ for $k = 1, \ldots, K$
Set $K^* = \inf\{k = 1, 2, \ldots, K \mid |P_k| \leq c\}$
Set $P_{K^*} = \bigcup_{k=K^*}^{K} P_k$
Set $\widetilde{Q}_1 = 1$
Sample with replacement T_1 from P_1
Set $A_1 = T_1$
Set $D_1 = A_1 + S_1$
for $m = 2, \ldots, M$ **do**
  if $\widetilde{Q}_{m-1} \leq K^*$ **then**
    Sample with replacement T_m from $P_{\widetilde{Q}_{m-1}}$
  else
    Sample with replacement T_m from P_{K^*}
  end if
  Set $A_m = A_{m-1} + T_m$
  if $A_m > D_{m-1}$ **then**
    Set $D_m = A_m + S_m$
    Set $\widetilde{Q}_m = 1$
  else
    Set $D_m = D_{m-1} + S_m$
    Calculate $n = \sum_{i=1}^{M} \mathcal{I}\{A_{m-1} \leq D_i < A_m\}$
    Set $\widetilde{Q}_m = \widetilde{Q}_{m-1} - n + 1$
  end if
end for

8.4 Queue-Length-Dependent Models

In this section, we apply these queue-length-restricted semi-experiments to the queue-length-dependent models.

8.4.1 Queue-Length-Dependent Service Rates

First, consider the original (QL, S, O) model, with queue-length-dependent service rates. We can apply the restricted s-perm and a-perm semi-experiments to simulations of this model. These queues will be labelled as $(QL, S, RsSE)$ and $(QL, S, RaSE)$, respectively. Figure 8.1 shows the stationary queue-length distributions for the original model, the standard s-perm and a-perm semi-experiment models, and the empirical queue-length distribution for the simulated restricted s-perm and a-perm semi-experiments.

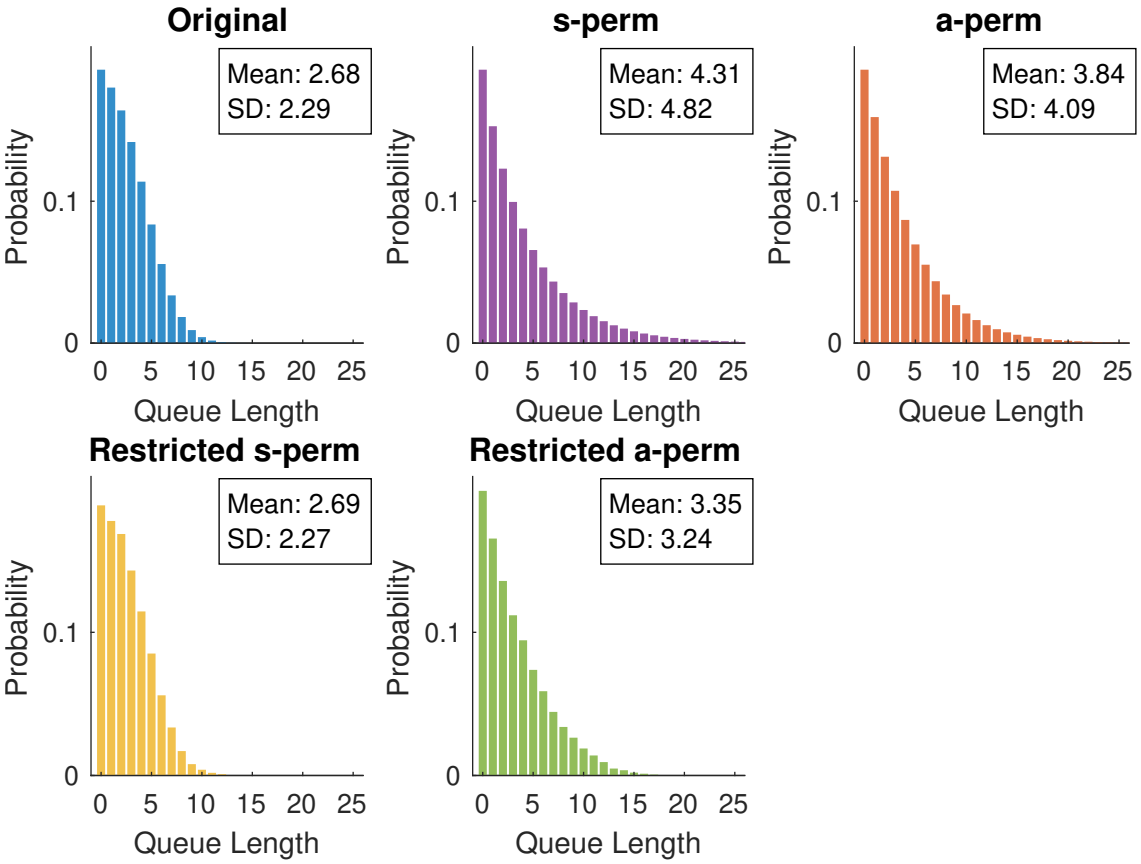


Figure 8.1: The blue plot shows the stationary queue-length distribution for the (QL, S, O) queue with $\lambda = 2$ and $\mu_i = 2.1 + 0.04i^2, i \geq 1$. The purple and orange plots show the stationary queue-length distribution for the (QL, S, sSE) and (QL, S, aSE). The yellow and green plots show the queue-length distribution for restricted s-perm and a-perm semi-experiments performed on a simulation of the original queue with 20,000 time steps. Here we chose $c = 5$.

Using the KS test, the queue-length distributions for the original queue and the restricted s-perm semi-experiment queue are statistically indistinguishable. This is as expected since the original queue's dependence was entirely through the queue-length at the start of service. Hence, restricting the permutations to have the same queue-length at those points produced another simulation of the original queue.

As explored in Chapter 6, the standard s-perm semi-experiment has a larger variance than the a-perm semi-experiment since permuting the service times more completely disrupted the dependence.

The restricted a-perm semi-experiment is very similar to the standard a-perm semi-experiment. This would be expected as the inter-arrival times are independent and identically distributed, so permuting with the restriction should have no significant difference to the standard permutation. However, there are differences in the tails of the distributions. This can be explained by the limited data available for larger queue-lengths, and the merging of smaller pools of data, which change the distribution at these extreme values.

We demonstrate that this is responsible for the differences in the queue-length distributions of the a-perm and restricted a-perm semi-experiments, as follows. Since the original model is known, when the restricted semi-experiment has a queue length greater than K^* at arrival, the inter-arrival times are sampled independently from an exponential distribution with rate λ. This method produces simulations which pass the KS test and show that the a-perm and restricted a-perm semi-experiments are from the same distribution, as expected.

Note that this issue does not have such a large effect on the restricted s-perm semi-experiment. Since the dependence is retained, the queue-length is kept low and to a similar range as the original queue. Hence, there is a suitable amount of data for this semi-experiment.

8.4.2 Queue-Length-Dependent Arrival Rates

Now consider the original model (QL, A, O) with queue-length-dependent arrival rates. We can apply the restricted s-perm and a-perm semi-experiments to this queue. These will be labelled $(QL, A, RsSE)$ and $(QL, A, RaSE)$, respectively. These are shown in Figure 8.2.

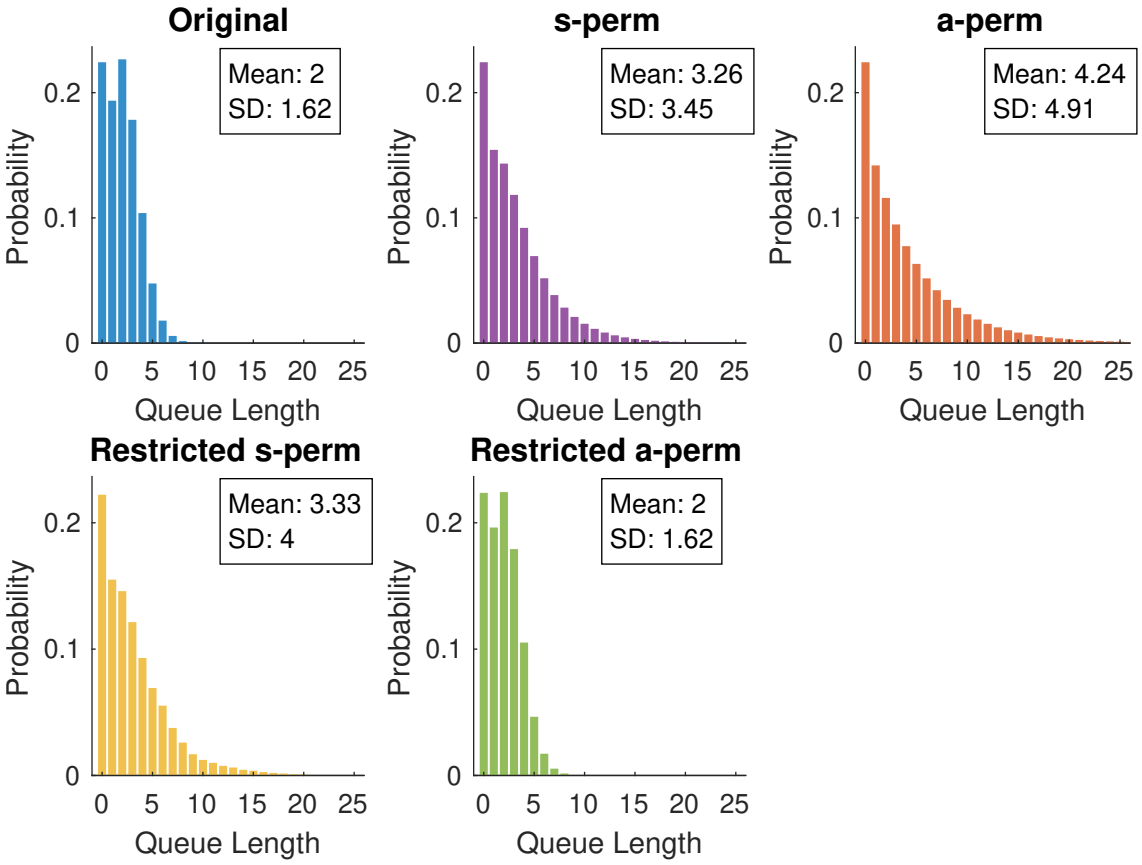


Figure 8.2: The blue, purple and orange plots are the stationary queue-length distributions for the (QL, A, O) queue, the s-perm semi-experiment (QL, A, sSE) and the a-perm semi-experiment (QL, A, aSE), respectively, with $\lambda_j = 4/j$ and $\mu = 2$. The yellow and green plots are the empirical queue-length distributions for the restricted s-perm and a-perm semi-experiments applied to a simulation of the original queue with 20,000 time steps. Here we chose $c = 5$.

The restricted a-perm semi-experiment on this model only samples inter-arrival times which had the same queue length immediately after the previous arrival. This exactly describes the dependence of this model and hence the restricted a-perm semi-experiment is simply another realisation of the original model. This is observed in Figure 8.2 and the KS test for these distributions indicate that they are not statistically significantly different.

Similar to before, the restricted s-perm semi-experiment appears to be quite similar to the s-perm semi-experiment, but the variance is quite a bit smaller and the distributions have differing behaviour in the tails. This is due to similar

issues as above where there is less data at more extreme queue lengths and hence we merge these pools. This can be shown by sampling those service times from queue-lengths with no data from the true distribution, $Exp(\mu)$. This creates a queue that is not significantly different to the standard s-perm semi-experiment, according to the KS test.

8.4.3 Queue-Length-Dependent Arrival and Service Rates

Now consider the (QL, AS, O) original model where both the arrival rates and service rates depend on the queue length at the previous arrival and start of service, respectively. We can apply the queue-length-restricted semi-experiments to this model, as shown in Figure 8.3. This shows that the queue-length-restricted semi-experiments are both more similar to the original queue than the standard semi-experiments. This is expected as the queue-length-restricted semi-experiments retain some of the dependence structure in the original queue. However, neither is completely a realisation of the original queue since some of the dependence is still disrupted. For example, the restricted s-perm semi-experiment accounts for the dependence structure within the service stream, reducing the variance of the queue-length distribution. However, the sampling with replacement of the service times still disrupts the overall dependence in the queue, since it depends on the queue length which depends on both the arrival and service times.

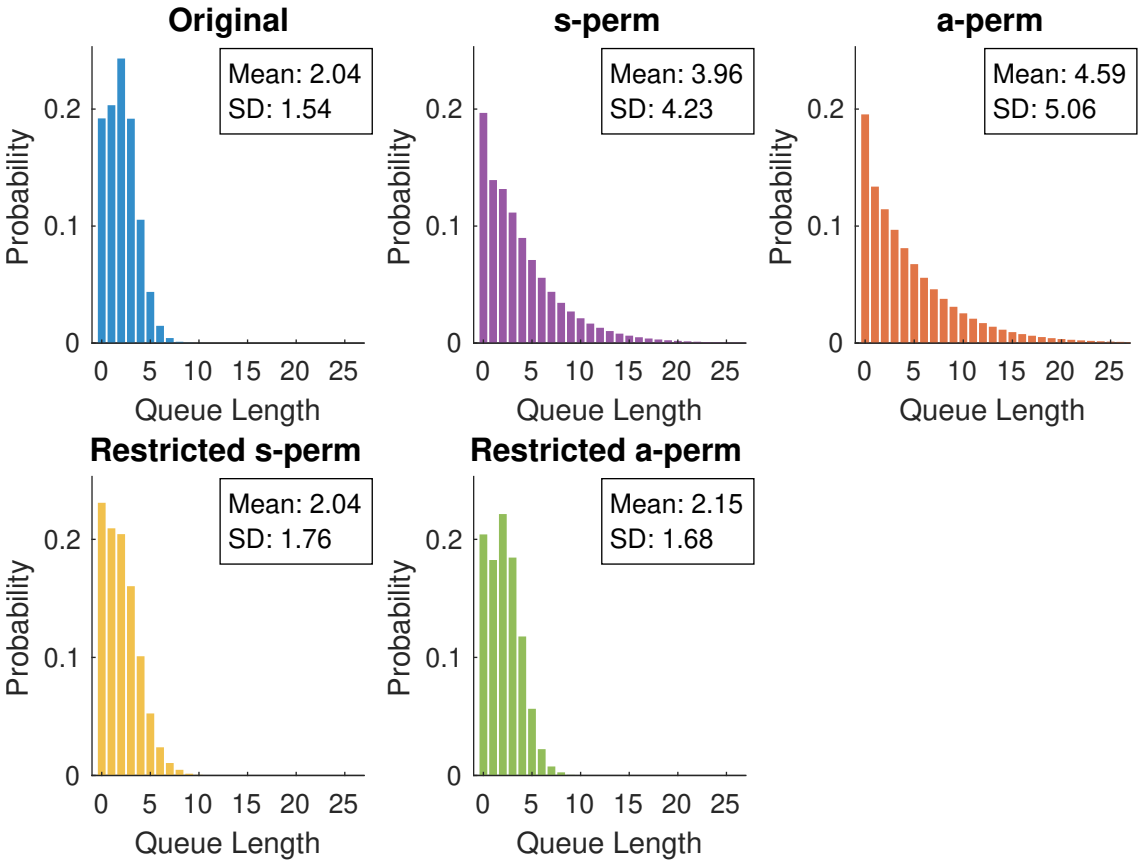


Figure 8.3: The blue, purple and orange plots are the stationary queue-length distributions for the original (QL, AS, O) queue, the s-perm semi-experiment (QL, AS, sSE) and the a-perm semi-experiment (QL, AS, aSE), respectively, with $\lambda_i = 5/i$ and $\mu_j = 1 + 0.8j$. The yellow and green plots are the empirical queue-length distributions for the restricted s-perm and a-perm semi-experiments applied to a simulation of the original queue with 20,000 time steps. Here we chose $c = 5$.

8.5 Pairwise Models

8.5.1 Proportional Service Times

Now consider the original model to be (P, P, O) where the service time of each customer is proportional to their inter-arrival time. While the queue-length-restricted semi-experiments do not specifically target the dependence present in this model, it is interesting to explore the effect. More importantly, we wish to know whether

the effect is different to the previous models to work towards identifying the different types dependence present in queueing data. Figure 8.4 shows the queue-length distributions for these restricted semi-experiments.

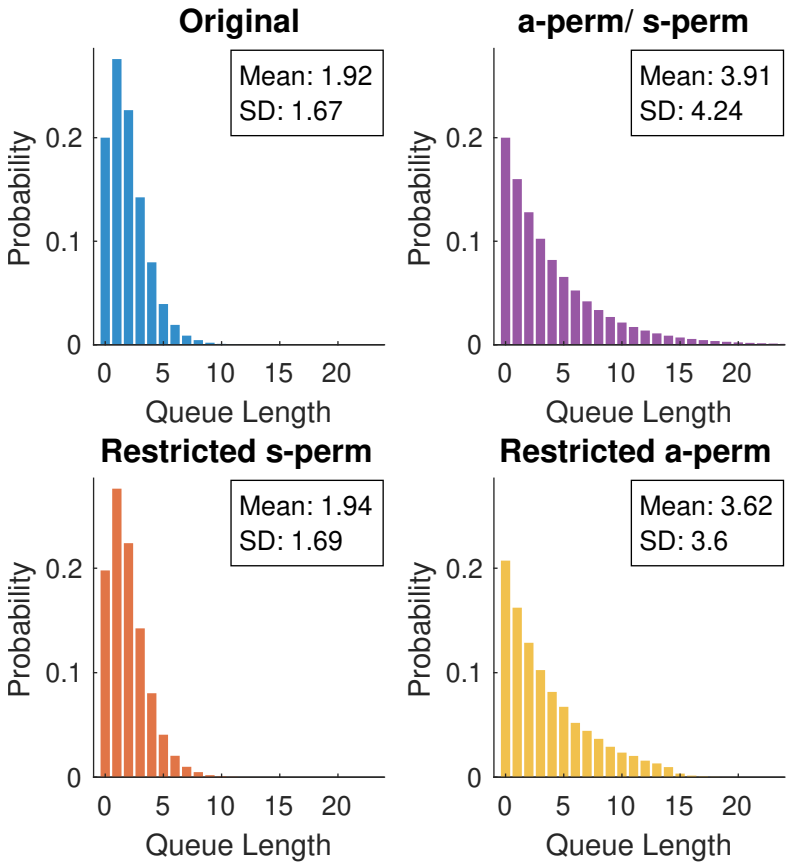


Figure 8.4: The blue plot shows the stationary queue-length distribution for the original (P, P, O) queue, where $\lambda = 2$ and $\nu = 0.8$. The purple plot shows the stationary queue-length distribution for the s-perm and a-perm semi-experiments, (P, P, SE), which are the same. The orange plot and yellow plot show the empirical stationary queue-length distribution for the restricted s-perm and a-perm semi-experiments performed on a simulation of the original queue run for 60,000 customers. Here we chose $c = 5$.

Recall that for this model, the s-perm and a-perm semi-experiments were the same $M/M/1$ queue since they both broke the direct dependence between each customer's inter-arrival time and service time. Since the dependence in this model is not directly through queue-length, it would seem logical to assume that the restricted a-perm and s-perm semi-experiments would be no different to the

unrestricted versions since they still break the connection between inter-arrival and service times. The restricted a-perm semi-experiment appears to be quite similar to the standard semi-experiments. The queue-length distribution has some differing behaviour in the tail, but this is explained by the lack of data at more extreme queue-lengths, as seen in the previous models.

However, the restricted s-perm semi-experiment is clearly quite different to the other semi-experiments and consistently passes the KS test to have the same queue-length distribution as the original queue. So, consider the differences between the restricted s-perm and a-perm semi-experiments. For the restricted a-perm, the inter-arrival times are chosen according to the queue-length immediately after the previous arrival. This queue-length is determined before the inter-arrival time (and service time) of the customer is even selected since it depends on all the inter-arrival times and service times prior to the previous arrival epoch. Hence, there is no dependence between the queue-length and the following inter-arrival time in the original queue, which is sampled independently from $Exp(\lambda)$.

The restricted s-perm chooses service times according to the queue-length at the start of service. In the original queue there is a dependence between this queue-length and the service time. Consider the following:

A particularly long inter-arrival time before mth arrival

$\downarrow$

More departures (on average) before mth arrival

$\downarrow$

Short queue length at mth arrival

$\downarrow$

mth customer's waiting time until they begin service is short

$\downarrow$

Fewer arrivals (on average) during this waiting time

$\downarrow$

Short queue length at the start of the mth service.

So a particularly long inter-arrival time can lead to a short queue length at the start of service, and vice versa. Also, if the mth inter-arrival time is long, then the service time is directly proportional and also particularly long. So a short queue length at service is associated with long service times, and vice versa.

Hence, when the restricted s-perm semi-experiment is performed, the service times selected from the pools will be quite similar to each other. This leads to a strong similarity to the original queue since the sequence of service times is similar.

These ideas are demonstrated in Figure 8.5 and Figure 8.6. These figures compare the the inter-arrival and service times grouped by the relevant queue-lengths to independent exponential samples in the groups of the same size. This comparison in Figure 8.5 shows that the inter-arrival times are independent of the queue-length at the prior arrival since the distribution of the inter-arrival times is not significantly different to the random exponential sample. Figure 8.6 shows that the service times are dependent on the queue length at the start of service where smaller queue lengths have larger service times and vice versa, as this plot is very different to the random exponential samples.

There does seem to be a stronger and more formal relationship where the queue-length restricted s-perm semi-experiment preserves some queue-length dependence for a broad range of models. However, we leave this as potential further work.

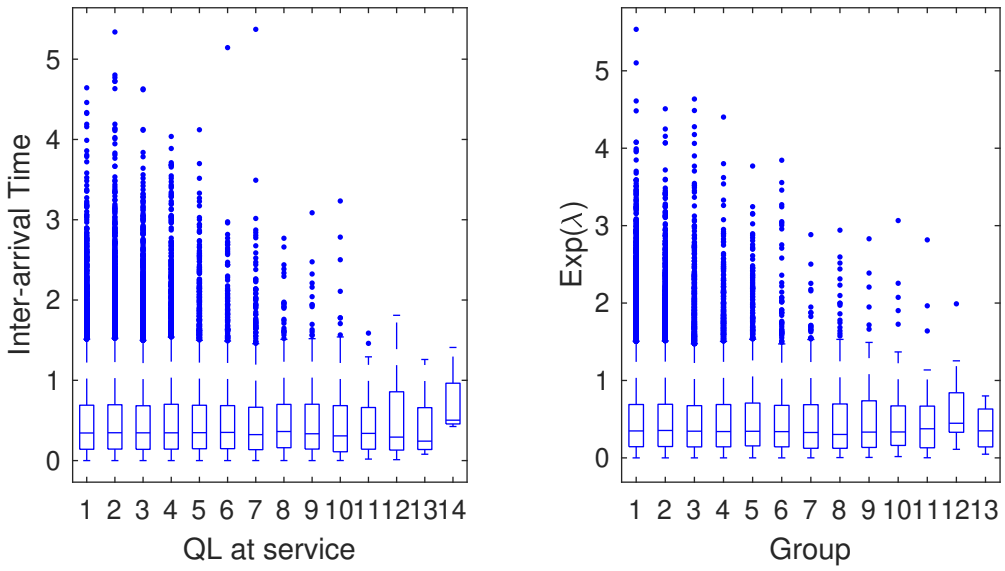


Figure 8.5: The left plot shows box-plots of the inter-arrival times in a simulation of the original queue with $\lambda = 2$ and $\nu = 0.8$ and 60,000 customers, grouped by the queue length at the start of the inter-arrival period. The right plot shows box-plots of random independent samples from an exponential distribution with rate λ, in groups of the same size as the right plot.

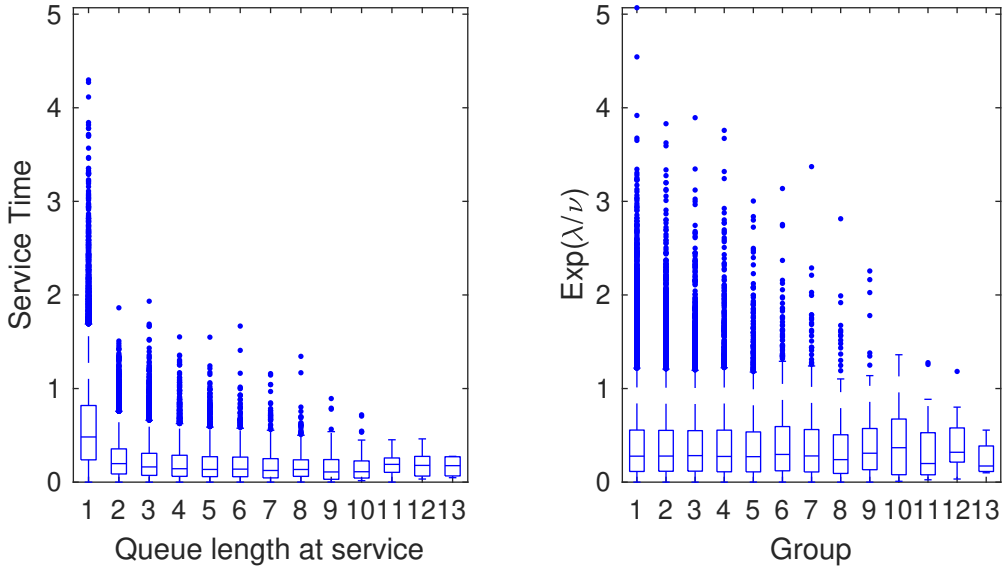


Figure 8.6: The left plot shows box-plots of the service times in a simulation of the original queue with $\lambda = 2$ and $\nu = 0.8$ and 60,000 customers, grouped by the queue length at the start of service. The right plot shows box-plots of random independent samples from an exponential distribution with rate λ/ν, in groups of the same size as the right plot.

8.5.2 BED Inter-arrival and Service Times

Now consider the original model to be (P, BED, O), where the inter-arrival times and service times have a bivariate exponential distribution. Again, we apply the queue-length-restricted semi-experiments to this model, shown in Figure 8.7. The results are very similar to the previous model: the restricted a-perm performed similarly to the standard semi-experiments with the exception of differing tail behaviour due to lack of data; and the restricted s-perm appeared as a realisation of the original queue due to the dependence between the queue length at start of service and the service times.

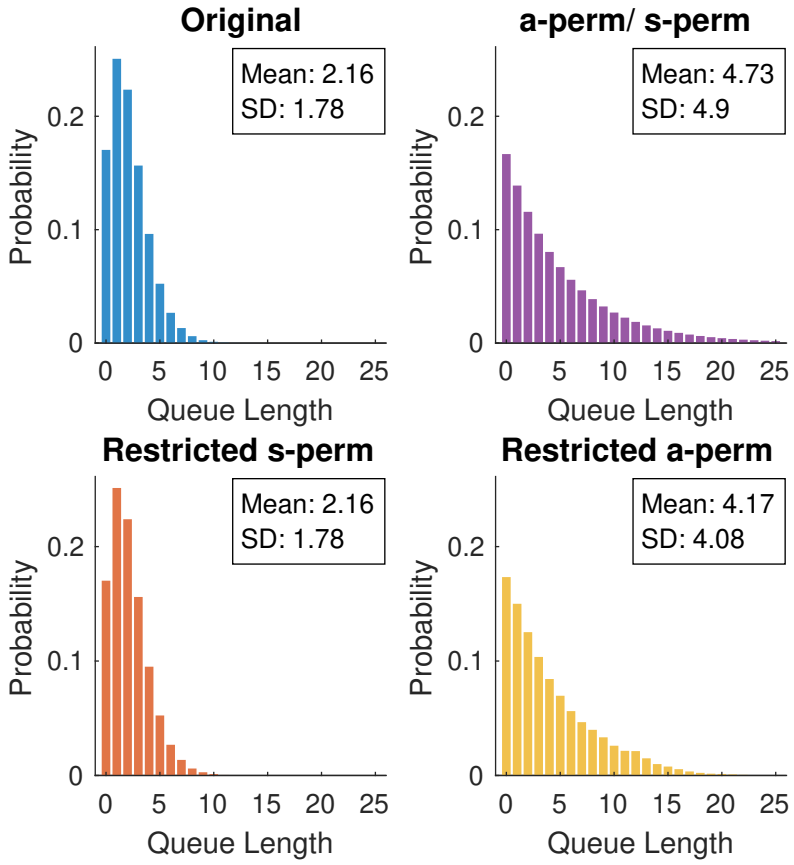


Figure 8.7: The blue plot shows the stationary queue-length distribution for the original (P, BED, O) queue, where $\lambda = 2.5, \mu = 3$ and $\rho = 0.9$. The purple plot shows the stationary queue-length distribution for the s-perm and a-perm semi-experiments, (P, BED, SE), which are the same. The orange plot and yellow plot show the empirical stationary queue-length distribution for the restricted s-perm and a-perm semi-experiments performed on a simulation of the original queue run for 60,000 customers. Here we chose $c = 5$.

8.5.3 Pairwise Arrival and Service Rates Dependence

The original model here is (P, C, O) where the the arrival rate and service rate of each customer depends on an external class of the customer. Figure 8.8 shows the application if the queue-length-restricted semi-experiments are applied to this original model. The results again are very similar to the previous pairwise models, however the effect can be seen as slightly weaker since the dependence in this model is through the rates, rather than the actual times. A high rate does not

guarantee a small time, though it is more likely. Despite this, the restricted a-perm semi-experiment appears close to the standard semi-experiments, with the exception in the tail of the distribution due to a lack of data at higher queue lengths. The restricted s-perm semi-experiment appears to be a realisation of the original queue. This is confirmed by KS tests.

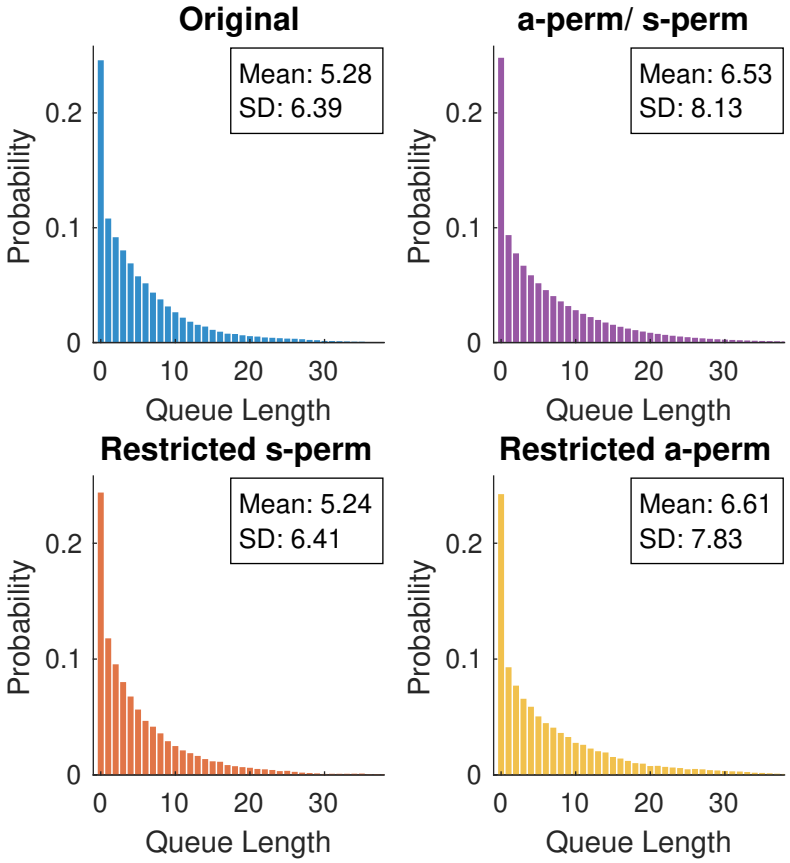


Figure 8.8: The blue plot shows the stationary queue-length distribution for the original (P, C, O) queue, where $\boldsymbol{\lambda} = [1, 4, 8, 12], \boldsymbol{\mu} = [1.4, 4.4, 8.1, 13]$ and $\boldsymbol{\pi}^* = [0.4, 0.2, 0.1, 0.3]$. The purple plot shows the stationary queue-length distribution for the s-perm and a-perm semi-experiments, (P, C, SE), which are the same. The orange plot and yellow plot show the empirical stationary queue-length distribution for the restricted s-perm and a-perm semi-experiments performed on a simulation of the original queue run for 60,000 customers. Here we chose $c = 5$.

8.6 Auto-Dependence Models

8.6.1 Auto-dependent Arrival Rates

Now we apply the queue-length-restricted semi-experiments to the (A, A, O) original queue with auto-dependent arrival rates. These semi-experiments are shown in Figure 8.9. The KS test confirms that the queue-length distributions for the original queue and the restricted s-perm queue are from the same distribution. This is expected since the dependence is only within the arrival stream and the service times are all independent and exponentially distributed with rate μ. So the s-perm and restricted s-perm are simply permuting identical and independent samples. Hence, there is no disruption of the dependence structure and they appear as realisations of the original queue.

The restricted a-perm semi-experiment is different. The example in Figure 8.9 shows that the standard a-perm semi-experiment has a much smaller variance in queue length than the original queue. This is because in this example the original queue has long periods with a larger arrival rate and long periods with a smaller arrival rate, since $\boldsymbol{\lambda} = [1.2, 1.95]$ and $P^* = \begin{bmatrix} 0.98 & 0.02 \\ 0.02 & 0.98 \end{bmatrix}$. This leads to longer queue lengths as seen in the plot of the distribution. The variance for the restricted a-perm semi-experiment is smaller than the original, but larger than the standard a-perm. This is due to some dependence between the inter-arrival times and the queue-length immediately after the previous arrival. Suppose the queue length at an arrival is particularly large. Then it is likely that the arrival rate has been large for some time preceding this. In our example, this means that it is likely to keep the large arrival rate, due to the transition probabilities in P^*. More generally, the queue-length at the previous arrival can provide some information about the the previous few arrival rates, and P^* informs the most likely rate for the new inter-arrival period. This can be shown analytically as follows.

Let $\widetilde{Q}_m$ be the queue length immediately after the mth arrival and let $\lambda^{(m)}$ be the rate of the inter-arrival time between the $(m-1)$th and mth arrivals. If $\lambda^{(m+1)}$ and $\widetilde{Q}_m$ were independent, then $P(\lambda^{(m+1)} = \lambda_k \mid \widetilde{Q}_m = i) = P(\lambda^{(m+1)} = \lambda_k) = \pi^*_k$. However, it can be shown this is not the case.

$$
\begin{aligned}
&P(\lambda^{(m+1)} = \lambda_k \mid \widetilde{Q}_m = i) \\
&\quad = \sum_{\ell=1}^{K} P(\lambda^{(m+1)} = \lambda_k \mid \widetilde{Q}_m = i, \lambda^{(m)} = \lambda_\ell) P(\lambda^{(m)} = \lambda_\ell \mid \widetilde{Q}_m = i) \\
&\quad = \sum_{\ell=1}^{K} P(\lambda^{(m+1)} = \lambda_k \mid \lambda^{(m)} = \lambda_\ell) P(\lambda^{(m)} = \lambda_\ell \mid \widetilde{Q}_m = i) \\
&\quad = \sum_{\ell=1}^{K} p^*_{\ell,k} P(\lambda^{(m)} = \lambda_\ell \mid \widetilde{Q}_m = i).
\end{aligned}
$$

Now let $q^*_{i,\ell} = P(\lambda^{(m)} = \lambda_\ell \mid \widetilde{Q}_m = i)$. For $1 \le \ell \le K$ and $i \ge 2$,

$$
\begin{aligned}
&q^*_{i,\ell} \\
&= \sum_{n=1}^{K} P(\lambda^{(m)} = \lambda_\ell \mid \widetilde{Q}_m = i, \lambda^{(m-1)} = \lambda_n) P(\lambda^{(m-1)} = \lambda_n \mid \widetilde{Q}_m = i) \\
&= \sum_{n=1}^{K} p^*_{n,\ell} P(\lambda^{(m-1)} = \lambda_n \mid \widetilde{Q}_m = i) \\
&= \sum_{n=1}^{K} p^*_{n,\ell} \sum_{j=i-1}^{\infty} P(\lambda^{(m-1)} = \lambda_n \mid \widetilde{Q}_m = i, \widetilde{Q}_{m-1} = j) P(\widetilde{Q}_{m-1} = j \mid \widetilde{Q}_m = i) \\
&= \sum_{n=1}^{K} p^*_{n,\ell} \sum_{j=i-1}^{\infty} \frac{P(\widetilde{Q}_m = i \mid \lambda^{(m-1)} = \lambda_n, \widetilde{Q}_{m-1} = j) P(\lambda^{(m-1)} = \lambda_n \mid \widetilde{Q}_{m-1} = j)}{P(\widetilde{Q}_m = i \mid \widetilde{Q}_{m-1} = j)} \\
&\qquad\qquad \times \frac{P(\widetilde{Q}_m = i \mid \widetilde{Q}_{m-1} = j) P(\widetilde{Q}_{m-1} = j)}{P(\widetilde{Q}_m = i)} \\
&= \sum_{n=1}^{K} p^*_{n,\ell} \sum_{j=i-1}^{\infty} \frac{\widetilde{\pi}_j}{\widetilde{\pi}_i} P(\widetilde{Q}_m = i \mid \lambda^{(m-1)} = \lambda_n, \widetilde{Q}_{m-1} = j) P(\lambda^{(m-1)} = \lambda_n \mid \widetilde{Q}_{m-1} = j) \\
&= \sum_{n=1}^{K} p^*_{n,\ell} \sum_{j=i-1}^{\infty} \frac{\widetilde{\pi}_j}{\widetilde{\pi}_i} P(\widetilde{Q}_m = i \mid \lambda^{(m-1)} = \lambda_n, \widetilde{Q}_{m-1} = j) q^*_{j,n},
\end{aligned}
$$

where $\widetilde{\pi}_i = \lim_{m \to \infty} P(\widetilde{Q}_m = i)$ and is calculated below. Note that $P(\widetilde{Q}_m = i \mid \lambda^{(m-1)} = \lambda_n, \widetilde{Q}_{m-1} = j)$ is equivalent to the probability that there are exactly $j - i + 1$ departures during an inter-arrival period with arrival rate λ_n. This is given by $\left(\frac{\mu}{\mu+\lambda_n}\right)^{j-i+1} \frac{\lambda_n}{\mu+\lambda_n}$.

Now note that when $i = 1$, the possible queue lengths at the previous arrival is $j = 1, 2, \ldots$. Also note that we need to calculate the probability of j departures during an inter-arrival period that begins with a queue length of j. This means the queue will be empty and hence the probability that an arrival will occur next is 1. Hence this probability is given by $\left(\frac{\mu}{\mu+\lambda_n}\right)^{j-i+1}$.

So,

$$q^*_{i,\ell} = \begin{cases} \sum_{n=1}^{K} p^*_{n,\ell} \sum_{j=i-1}^{\infty} \frac{\widetilde{\pi}_j}{\widetilde{\pi}_i} \left(\frac{\mu}{\mu+\lambda_n}\right)^{j-i+1} \frac{\lambda_n}{\mu+\lambda_n} q^*_{j,n}, & \text{for } 1 \le \ell \le K, i \ge 2 \\ \sum_{n=1}^{K} p^*_{n,\ell} \sum_{j=1}^{\infty} \frac{\widetilde{\pi}_j}{\widetilde{\pi}_1} \left(\frac{\mu}{\mu+\lambda_n}\right)^{j} q^*_{j,n}, & \text{for } 1 \le \ell \le K, i = 1. \end{cases}$$

Now, we need to calculate $\widetilde{\pi}_i$ for $i \ge 1$. For $i \ge 2$,

$$\begin{aligned} \widetilde{\pi}_i &= \lim_{m\to\infty} P(\widetilde{Q}_m = i) \\ &= \lim_{m\to\infty} \sum_{j=i-1}^{\infty} P(\widetilde{Q}_m = i \mid \widetilde{Q}_{m-1} = j) P(\widetilde{Q}_{m-1} = j) \\ &= \sum_{j=i-1}^{\infty} \widetilde{\pi}_j \lim_{m\to\infty} P(\widetilde{Q}_m = i \mid \widetilde{Q}_{m-1} = j) \\ &= \sum_{j=i-1}^{\infty} \widetilde{\pi}_j P(\widetilde{Q}_m = i \mid \widetilde{Q}_{m-1} = j), \quad \text{since independent of the limit} \\ &= \sum_{j=i-1}^{\infty} \widetilde{\pi}_j \sum_{k=1}^{K} \lim_{m\to\infty} P(\widetilde{Q}_m = i \mid \widetilde{Q}_{m-1} = j, \lambda^{(m-1)} = \lambda_k) P(\lambda^{(m-1)} = \lambda_k \mid \widetilde{Q}_{m-1} = j) \\ &= \sum_{j=i-1}^{\infty} \widetilde{\pi}_j \sum_{k=1}^{K} \left(\frac{\mu}{\mu+\lambda_k}\right)^{j-i+1} \frac{\lambda_k}{\mu+\lambda_k} q^*_{j,k}. \end{aligned}$$

Note that we can justify swapping the sum and limit since all the values are non-negative and the sum is clearly finite (≤ 1).

Using arguments as above, for $i = 1$,

$$\widetilde{\pi}_1 = \sum_{j=1}^{\infty} \widetilde{\pi}_j \sum_{k=1}^{K} \left(\frac{\mu}{\mu+\lambda_k}\right)^{j} q^*_{j,k}.$$

So,

$$\widetilde{\pi}_i = \begin{cases} \sum_{j=i-1}^{\infty} \widetilde{\pi}_j \sum_{k=1}^{K} \left(\frac{\mu}{\mu+\lambda_k}\right)^{j-i+1} \frac{\lambda_k}{\mu+\lambda_k} q^*_{j,k} & \text{for } i \geq 2 \\ \sum_{j=1}^{\infty} \widetilde{\pi}_j \sum_{k=1}^{K} \left(\frac{\mu}{\mu+\lambda_k}\right)^{j} q^*_{j,k} & \text{for } i = 1, \end{cases}$$

subject to

$$\sum_{i=1}^{\infty} \widetilde{\pi}_i = 1.$$

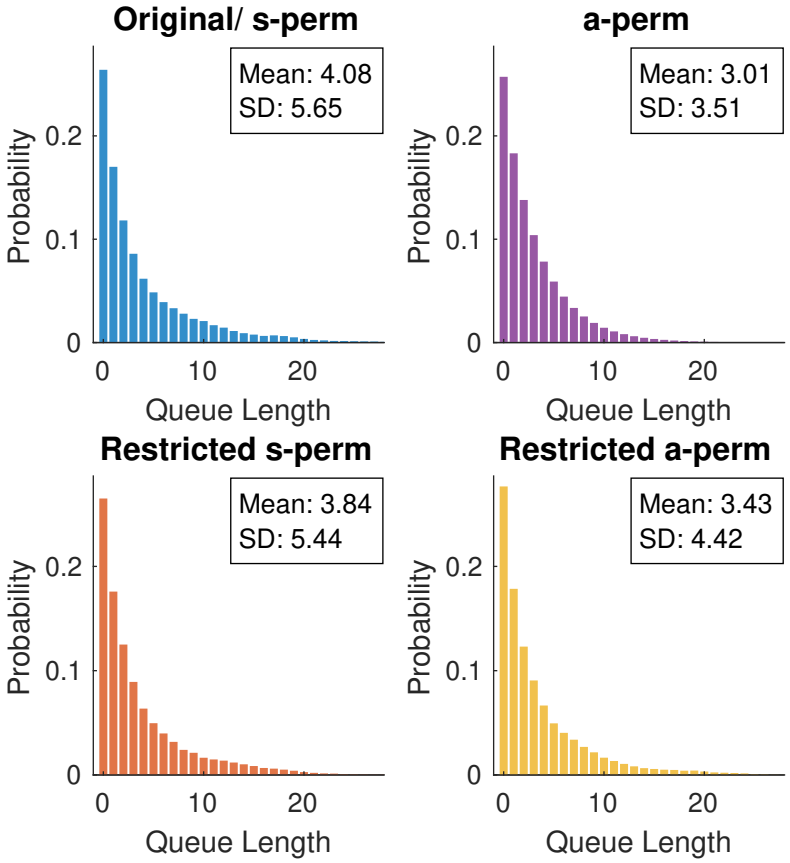


Figure 8.9: The blue plot shows the stationary queue-length distribution for the original (A, A, O) queue and the a-perm semi-experiment (A, A, aSE) queue, where $\boldsymbol{\lambda} = [1.2, 1.95]$, $\mu = 2$ and $P^* = [0.98, 0.02; 0.02, 0.98]$. The purple plot shows the stationary queue-length distribution for the a-perm semi-experiments. The orange plot and yellow plot show the empirical stationary queue-length distribution for the restricted s-perm and a-perm semi-experiments performed on a simulation of the original queue run for 20,000 time steps. Here we chose $c = 5$.

Figure 8.10 demonstrates the relationship between the queue length immediately after an arrival and the probability of being in each customer class. In this example, when the queue length at arrival is low, it is more likely that the next arriving customer will be from class 1, which has a smaller arrival rate, and when the queue length is high, it is more likely that the customer has class 2 with a larger arrival rate. This is because, with these parameters, there is a high probability that customers who arrive at similar times have the same arrival rate class. So the queue length at arrival can inform us which arrival rate class is most likely.

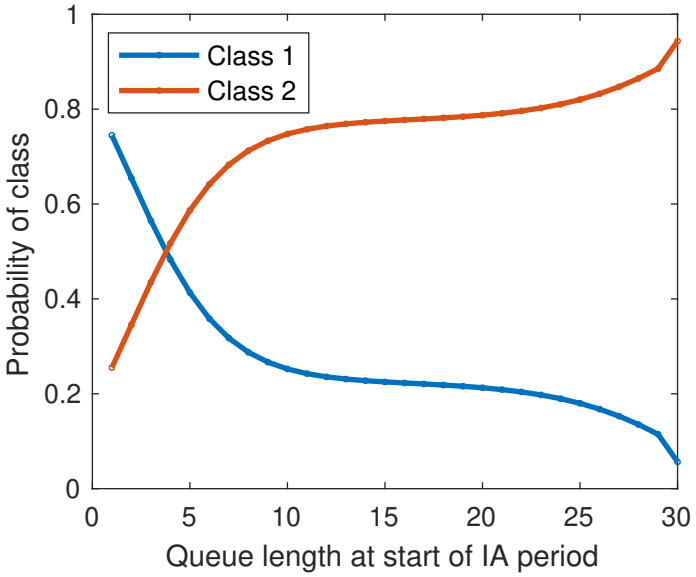


Figure 8.10: This plot shows the probability of being in arrival class 1 (with $\lambda_1 = 1.2$) in the blue line and arrival class 2 (with $\lambda_2 = 1.95$) in the orange line as the queue length at the start of inter-arrival periods increases. This is for the same parameters in Figure 8.9.

8.6.2 Auto-dependent Service Rates

Now we apply the queue-length-restricted semi-experiments to the (A, S, O) original queue with auto-dependent service rates. These are shown in Figure 8.11. The KS test shows that the original, a-perm, and restricted a-perm queues have the same queue-length distribution. Since the dependence is only within the service stream, permuting the inter-arrival times does not disrupt the dependence. Hence the restricted a-perm is the same as the standard a-perm. The restricted s-perm has a greater variance than the standard s-perm and is more similar to the original

queue. This is because there is a dependence between the service rates and the queue length at the start of service. If the queue length is large, then it is more likely that recent service rates were small and hence this can inform the next likely service rate. This is an almost identical feature to the one noted for the restricted a-perm semi-experiments in the auto-dependent arrival rates model.

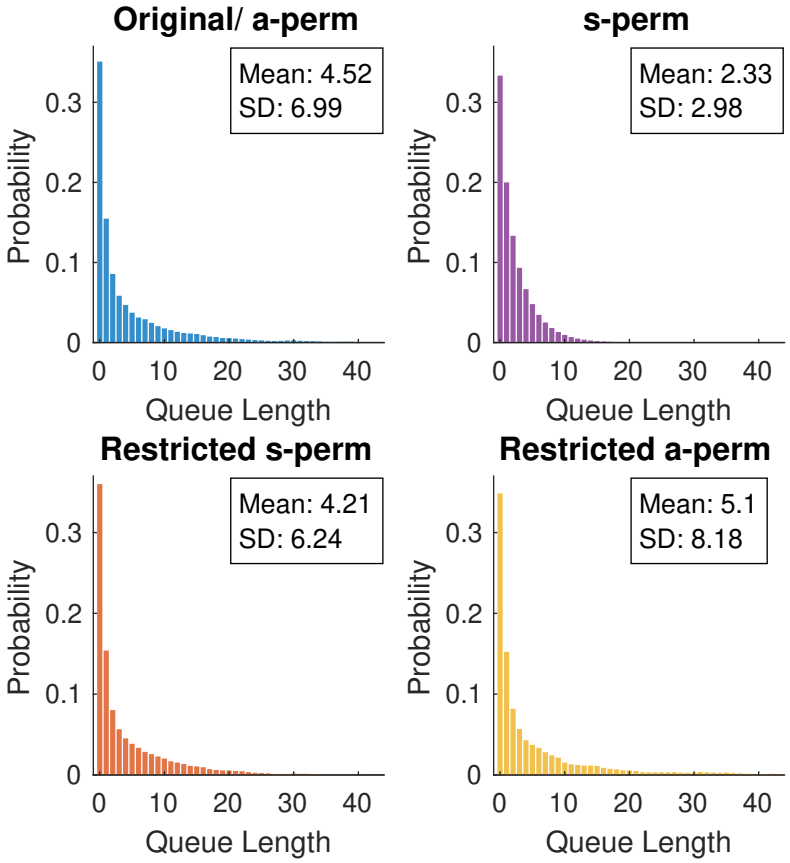


Figure 8.11: The blue plot shows the stationary queue-length distribution for the original (A, S, O) queue and the s-perm semi-experiment (A, S, sSE) queue, where $\boldsymbol{\mu} = [2, 6]$, $\lambda = 2$ and $P^* = [0.98, 0.02; 0.02, 0.98]$. The purple plot shows the stationary queue-length distribution for the a-perm semi-experiments. The orange plot and yellow plot show the empirical stationary queue-length distribution for the restricted s-perm and a-perm semi-experiments performed on a simulation of the original queue run for 20,000 time steps. Here we chose $c = 5$.

8.6.3 Auto-dependent Arrival and Service Rates

Now consider the (A, AS, O) original model with auto-dependent arrival rates and auto-dependent service rates with no cross-dependence. The queue-length-restricted semi-experiments for this model are shown in Figure 8.12. The KS test applied to all of these queue-length distributions indicates that none of them are from the same distribution. As discussed in Chapter 3, the original, s-perm and a-perm queues are all different in general. Now consider the restricted s-perm and a-perm semi-experiments. The restricted s-perm semi-experiment has a larger variance (in this example) than the standard s-perm and hence looks more like the original queue. This is due to dependence between the service times and queue-lengths at the start of services as seen in the restricted s-perm semi-experiment for the model with auto-dependent service rates. The restricted a-perm semi-experiment has a greater variance than the a-perm semi-experiment here due to the analogous dependence between the inter-arrival times and the queue length at the previous arrival, as seen before. Note that the tail ends of these queue-length distributions are 'wobbly' due to the lack of data at higher queue-lengths, also as seen before.

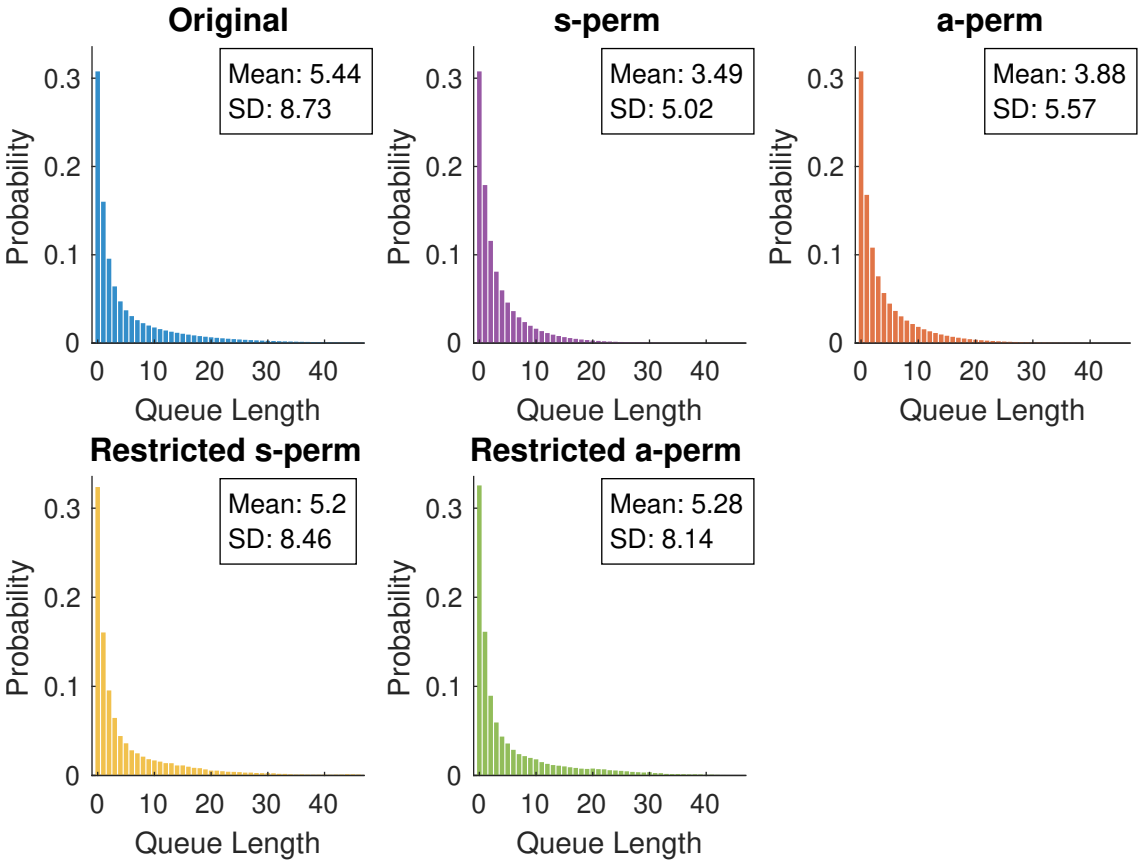


Figure 8.12: The blue plot shows the stationary queue-length distribution for the original (A, AS, O) queue, where $\boldsymbol{\lambda} = [1, 1.5, 2]$, $\boldsymbol{\mu} = [1.5, 3]$, and $P^{*A} = [0.97, 0.03, 0; 0.97, 0.03; 0.03, 0, 0.97]$ and $P^{*S} = [0.98, 0.02; 0.02, 0.98]$. The purple plot and orange plot show the stationary queue-length distributions for the a-perm (A, AS, aSE) and s-perm (A, AS, sSE) semi-experiment models. The yellow and green plot show the empirical stationary queue-length distribution for the restricted s-perm and a-perm semi-experiments performed on a simulation of the original queue run for 20,000 time steps. Here we chose $c = 5$.

8.7 Conclusion

In this chapter we considered different ways to restrict semi-experiments, including restricting permutations to customer classes, permuting service times within customers with the same queue-length at the start of service periods, and permuting inter-arrival times within customers with the same queue-length at the

start of inter-arrival periods. We then provided algorithms for the latter two restricted semi-experiments. These queue-length-restricted semi-experiments were applied to all the dependence queue models and the queue-length distributions were compared to the original and standard semi-experiment queues.

We found that applying the appropriate queue-length-restricted semi-experiment to the queue-length-dependent queues gives a queue-length distribution like the original queue, as expected since the restricted permutation does not disrupt the dependence structure. The other queue-length-restricted semi-experiment was the same as the standard semi-experiment except for some tail behaviour due to small sample sizes in extreme queue-length values.

When applying the restricted semi-experiments to the pairwise dependence models, we found that the restricted s-perm semi-experiment reproduced the original queue, while the restricted a-perm semi-experiment was equivalent to the standard s-perm/a-perm semi-experiment. This was due to dependence between a customer's inter-arrival time and the queue-length when they begin service.

In the auto-dependence models, the restricted s-perm applied to the (A, A, O) queue is equivalent to the standard s-perm which is equivalent to the original queue since this does not disrupt the dependence. The restricted a-perm semi-experiment is different to the standard a-perm since there is a remaining dependence between the queue-length at the previous arrival and the current arrival rate, which we explained analytically. Very similar results are found for the (A, S, O) queue. The queue-length-restricted semi-experiments applied to the (A, AS, O) queue have the same additional dependence between the queue-length at arrivals and services to arrival and service rates as the previous two models.

Chapter 9

Diagnosing Dependence Types

To summarise the previous chapters, we have constructed models for three broad types of dependence between the arrival process and service times of queues. These are:

- **Pairwise Dependence**, including
  - the service times proportional to inter-arrival times (P, P, O),
  - the inter-arrival times and service times have a bivariate exponential distribution (P, BED, O), and
  - the customers having classes with related inter-arrival and service rates (P, C, O).
- **Auto-dependence**, where the inter-arrival and service rates depend on a customer class within the arrival or service stream, including
  - auto-dependence within the arrival stream (A, A, O),
  - auto-dependence within the service stream (A, S, O), and
  - auto-dependence within both the arrival and service stream without and cross-dependence (A, AS, O).
- **Queue-Length-Dependence**, where the inter-arrival and service rates depend on the queue length, including
  - the service rates depend on the queue length at the start of the service period (QL, S, O),

  - the arrival rate depends on the queue length immediately after the previous arrival (QL, A, O), and
  - a combination of the two previous queue-length-dependent models (QL, AS, O).

The aim is to distinguish between these types of dependence in queueing data in order to suggest an appropriate model. This is achieved by analysing semi-experiments of these queues. To summarise the types of semi-experiments we have considered are

- **s-perm** where the arrival stream is maintained and the service times are randomly permuted (sSE),
- **a-perm** where the service stream is maintained and the inter-arrival times are randomly permuted (aSE),
- **Class restricted s-perm/ a-perm** where the service times/ inter-arrival times are permuted within some known external classes,
- **Queue-length-restricted s-perm** where the arrival stream is maintained and the service times are permuted within those customers who had the same queue length at the start of service ($RsSE$), and
- **Queue-length-restricted a-perm** where the service stream is maintained and the inter-arrival times are permuted within those customers who had the same queue length at the start of the inter-arrival period ($RaSE$).

Assume we have queueing data with suspected dependence between the arrival process and service times. First, if there are known classes of customers which are likely to have different arrival and/or service rates, then all semi-experiments should be performed restricted to these classes. If this restriction significantly reduces the difference between the semi-experiments and the original queue, then these classes account for some of the percieved dependence in the naive semi-experiments. Hence, they should be included in the models. Then, the flow chart in Figure 9.1 shows an approach to distinguish between the dependence models we have explored. The flow chart summarises the information already discussed in previous chapters. The symbols $O, sSE, aSE, RsSE, RaSE$ represent the original, s-perm, a-perm, queue-length-restricted s-perm and queue-length-restricted a-perm queues. To interpret the questions posed in the blue boxes, consider the example '$sSE = O$?'. This questions asks whether the s-perm queue and the original queue have significantly different queue-length distributions (perhaps according to the KS test, or some other measures). If they are significantly different,

the answer is 'No', otherwise the answer is 'Yes'. The yellow diamonds show the decisions that can be made from the flow chart, mostly the type of dependence in the queue.

Potential Pairwise Dependence Note that the 'Potential Pairwise Dep' decision represents all queues with a pairwise dependence between the inter-arrival times and service times, as well as some other possible examples. We have not explored these other cases here and leave them for further work, but such an example could be when a service time depends on the previous two inter-arrival times. If we assume pairwise dependence, a little more work is required to further distinguish between the three models considered in Chapter 3 using the correlation coefficient. First, the simplest model (P, P, O) can easily be identified since all the service times are equal to some constant multiple of the inter-arrival times for each customer. This means that the correlation coefficient between inter-arrival times and service times is always 1. For the (P, BED, O) model, the correlation is specified by the parameter $0 < \rho < 1$. This means the model has the capacity for a very high correlation coefficient The (P, C, O) model has the dependence through the rates/means of the inter-arrival times and service times, rather than the actual times themselves. This means that the correlation coefficient will generally be smaller than what is attainable in the (P, BED, O) model, and increases as the number of classes increases. Therefore, if the correlation coefficient is large, then the (P, BED, O) model is likely more suitable. If the correlation coefficient is small, then either model may be suitable. Since a queue with smaller correlation has a weaker dependence and effects from semi-experiments, the difference between these models is negligible when seeking a useful model.

Other The 'Other' decision represents other queues which we may not have considered here. This also includes the (QL, AS, O) queue. To help identify this queue, we note that the RaSE and RsSE queues both become much more similar to the original than the standard s-perm and a-perm semi-experiments.

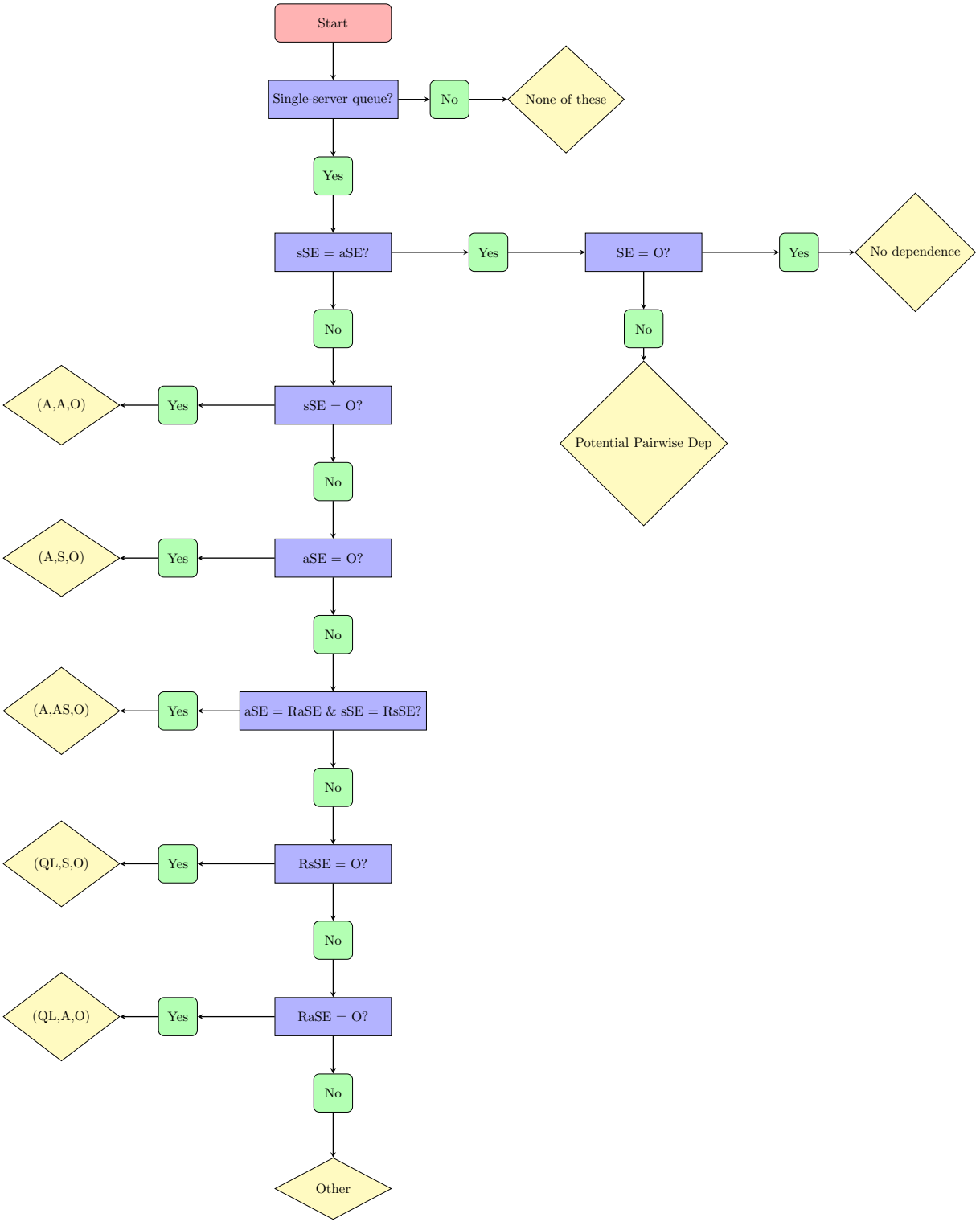


Figure 9.1: Flow chart for diagnosing dependence types.

Chapter 10

Conclusion

Due to the many applications of queues in our modern society, it is important to be able to accurately model various and complex queues. We have seen examples of queues with unexplained dependence between the arrival and service processes in the ICUs studied by Varney *et al.* [1]. Hence, in this thesis we have proposed and explored several different queueing models with dependence between the arrival and service processes, and developed methods for detecting, classifying and gaining insight into the various forms that this dependence can take. We used a variety of semi-experiments to disrupt the dependence structure in these queues and compare them to the original queues to identify the dependence structure and the role it plays in the queue's behaviour. In particular, we chose a modelling approach to applying semi-experiments in order to gain an understanding of how they interact with dependence structures in queues as well as provide more computationally efficient methods for calculating the stationary queue-length distributions, compared to simulation methods. This investigation culminated in a proposed method for exploring and differentiating between different forms of dependence between arrivals and services in queues, and suggested appropriate models.

In Chapter 3, we proposed some models with simple forms of dependence and demonstrated how semi-experiments can be modelled and used to detect and identify this dependence. This was extended in Chapters 4, 5, 6, and 7, where we proposed more complex queue-length-dependent queues and constructed models for both the a-perm and s-perm semi-experiments. In Chapter 8, we considered restricted semi-experiments to further classify the types of dependence. Finally, Chapter 9 summarised the results of the previous chapters and proposed a flow chart to use various semi-experiments to differentiate between queueing models with dependence between the arrival and service processes. So, if a real-world

queue is found to have dependence between the arrival process and service time distribution, one could use a range of empirical semi-experiments and compare the results to this flow chart to narrow down a set of appropriate models to implement to accurately describe the data.

10.1 Future Work

There are several directions for future work. Firstly, we have restricted all the queues to be single-server queues, with infinite capacity and a first-come, first-served service protocol. There are many other types of queueing systems which we could consider and which can be useful for modelling real-world scenarios. We could also consider queues with multiple or infinitely many servers or queues with a finite capacity in which arrivals are lost when the queue length is at its limit. We could have impatient customers who choose not to enter the queue if it is too long (balk) or choose to leave while waiting (renege). We could also consider various queueing disciplines, such as last-come, first-served or random service selection, or arriving customers are served according to their particular priority. This last discipline in particular could induce a dependence between the arrival and service process if the priority scheme is not observed. We could also consider systems with parallel queues that customers can choose between, or a network of queues.

Secondly, there are other forms of dependence between the arrival and service processes of queues. To have a more robust method, we need to consider how we can detect and classify as many forms of dependence as possible. Some particular models are explored in the literature review in Section 2.5.

These other forms of dependence and various types of queues would lead to the need for a larger suite of semi-experiments. For example, in infinite-server queues, it is possible to permute the service times and retain the departure points, since the customers do not need to be served sequentially.

Another topic of further investigation was alluded to in Chapter 9. This is that equivalent queue-length distributions can occur even when the semi-experiment model is distinct from the original model, as we saw when the restricted s-perm semi-experiment was applied to models with pairwise dependence. While we provided some explanation for why this may occur, it is not a well-understood phenomenon there is evidence that there is a stronger and more general relationship that should be investigated further in order to more fully understand how various semi-experiments behave and how to interpret them in terms of the dependence

within the queue.

Finally, this method of detection and classification is just one step in the process of accurately model queueing data. We would ultimately like to develop a tool that can be used to detect, classify and quantify dependence between the arrival and service processes in queueing data, including a variety of queueing systems with different protocols. Then we would be able to propose an appropriate model for the data and propose methods for fitting the parameters to the data.

Appendix A

Labels of Queues in This Thesis

In this thesis queues are labelled using (X, Y, Z), where X is the broad dependence type, Y is the more spcific dependence type, and Z specifies which type of queue. Table A.1 displays all the possible combinations of X and Y used in the thesis. Table A.2 displays all the possible values of Z used in this thesis.

X	Y	Description
P	P	Pairwise dependence where service times are proportional to arrival times
P	BED	Pairwise dependence where arrival and service rates have a bivariate exponential distribution
P	C	Pairwise dependence where arrival and service rates depend on customer class
A	A	Auto-dependence where arrival rates depend on previous arrival rate
A	S	Auto-dependence where service rates depend on previous service rate
A	AS	Auto-dependence where arrival rates depend on previous arrival rate, and service rates depend on previous service rate
QL	S	Queue-length dependence where service rates depend on queue-length at the start of service
QL	A	Queue-length dependence where arrival rates depend on queue-length at start of inter-arrival period
QL	AS	Queue-length dependence where arrival rates depend on queue-length at start of inter-arrival period, and service rates depend on queue-length at the start of service

Table A.1: All combinations of the X and Y components in the (X, Y, Z) labels of queues used in this thesis.

Z	Description
O	Original queue
SE	Semi-experiment queue when s-perm and a-perm are indentical
sSE	s-perm semi-experiment queue
aSE	a-perm semi-experiment queue
$RsSE$	s-perm semi-experiment queue, restricted to permuting within customer classes of the same queue-length at the start of service
$RaSE$	a-perm semi-experiment queue, restricted to permuting within customer classes of the same queue-length at the start of inter-arrival period

Table A.2: All possibilities of the Z component in the (X, Y, Z) labels of queues used in this thesis.

Bibliography

[1] J. Varney, N. Bean, and M. Mackay, "The self-regulating nature of occupancy in ICUs: stochastic homoeostasis," *Health Care Management Science*, vol. 22, pp. 615–634, 2018.

[2] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris, *Fundamentals of Queueing Theory*. John Wiley and Sons, Inc., fifth ed., 2018.

[3] D. G. Kendall, "Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain," *The Annals of Mathematical Statistics*, vol. 24, pp. 338–354, 1953.

[4] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of Computational Physics*, vol. 22, pp. 403–434, 1976.

[5] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial and Applied Mathematics, 1999.

[6] L. Bright and P. G. Taylor, "Equilibirum distributions for level-dependent quasi-birth-and-death processes," *Matrix-analytic Methods in Stochastic Models*, vol. 183, pp. 359–375, 1997.

[7] N. Bean and G. Latouche, "Approximations to quasi-birth-and-death processes with infinite blocks," *Advances in Applied Probability*, vol. 42, pp. 1102–1125, 2010.

[8] A. Erramilli, O. Narayan, and W. Willinger, "Experimental queueing analysis with long-range dependent packet traffic," *IEEE/SCM Transactions on Networking*, vol. 4, pp. 209–223, 1996.

[9] N. Hohn, D. Veitch, and P. Abry, "Does fractal scaling at the IP level depend on TCP flow arrival processes?," 2002.

[10] J. Ridoux, A. Nucci, and D. Veitch, "Seeing the difference in IP traffic: Wireless versus wireline," 2006.

[11] I. T. Young, "Proof without predjudice : Use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources," *The Journal of Histochemistry and Cytochemistry*, vol. 25, pp. 935–941, 1977.

[12] B. W. Conolly, "The waiting time process for a certain correlated queue," *Operations Research*, vol. 16, pp. 1006 – 1015, 1968.

[13] B. W. Conolly and N. Hadidi, "A correlated queue," *Journal of Applied Probability*, vol. 6, pp. 122–136, 1969.

[14] I. Cidon, R. Gucrin, A. Khamisy, and M. Sidi, "Analysis of a correlated queue in a communication system," *IEEE Transactions on Information Theory*, vol. 39, pp. 456–465, 1993.

[15] G. U. Hwang and K. Sohraby, "Performance of correlated queues: The impact of correlated service and inter-arrival times," *Performance Evaluation*, vol. 55, pp. 129–145, 2004.

[16] I. Cidon, R. Guréin, A. Khamisy, and M. Sidi, "On queues with interarrival times proportional to service times," *Probability in the Engineering and Informational Sciences*, vol. 10, pp. 87–107, 1996.

[17] E. T. Gumbel, "Bivariate exponential distributions," *Journal of the American Statistical Association*, vol. 55, pp. 698–707, 1960.

[18] C. R. Mitchell and A. S. Paulson, "M/M/1 queues with interdependent arrival and service processes," *Naval Research Logistics Quarterly*, vol. 26, pp. 47–56, 1979.

[19] B. W. Conolly and Q. H. Choo, "The waiting time process for a generalized correlated queue with exponential demand and service," *SIAM Journal on Applied Mathematics*, vol. 37, pp. 263–275, 1979.

[20] C. Langaris, "Busy-period analysis of a correlated queue with exponential demand and service," *Applied Probability Trust*, vol. 24, pp. 476–485, 1987.

[21] B. Kim and J. Kim, "The waiting time distribution for a correlated queue with exponential interarrival and service times," *Operations Research Letters*, vol. 46, pp. 268–271, 2018.

[22] X. Chao, "Monotone effect of dependency between interarrival and service times in a simple queueing system," *Operations Research Letters*, vol. 17, pp. 47–51, 1995.

[23] A. Muller, "On the waiting times in queues with dependency between interarrival and service times," *Operations Research Letters*, vol. 26, pp. 43–47, 2000.

[24] C. Langaris, "A correlated queue with infinitely many servers," *Applied Probability Trust*, vol. 23, pp. 155–165, 1986.

[25] C. Harris, "Queues with stochastic service rates," *Naval Research Logistics Quarterly*, vol. 14, pp. 219–230, 1967.

[26] J. Shanthikumar, "On a single-server queue with state-dependent service," *Naval Research Logistics Quarterly*, vol. 26, pp. 305–309, 1979.

[27] P. J. Courtois and J. Georges, "On a single-server finite queuing model with state-dependent arrival and service processes," *Operations Research*, vol. 19, pp. 424–435, 1971.

[28] U. C. Gupta and T. S. S. S. Rao, "On the analysis of single server finite queue with state dependent arrival and service processes: M(n)/G(n)/1/K," *OR Spektrum*, vol. 20, pp. 83–89, 1998.

[29] H. Abouee-Mehrizi and O. Baron, "State-dependent M/G/1 queueing systems," *Queueing Systems*, vol. 82, pp. 121–148, 2 2016.

[30] O. J. Boxma and D. Perry, "Queueing model with dependence between service and interarrival times," *European Journal of Operational Research*, vol. 128, pp. 611–624, 2001.

[31] S. Borst, O. Boxama, and M. Combe, "An M/G/1 queue with dependence between interarrival and service times," *Stochastic Models*, vol. 9, pp. 341–371, 1993.

[32] G. J. K. Regterschot and J. H. A. D. Smit, "The queue M/G/1 with Markov modulated arrivals and services," *Mathematics of Operations Research*, vol. 11, pp. 465–483, 1986.

[33] I. Kulkarni and V. Adan, "Single-server queue with Markov-dependent interarrival and service times," *Queueing Systems*, pp. 113–134, 2003.

[34] S. K. Iyer and D. Manjunath, "Queues with dependency between interarrival and service times using mixtures of bivariates," *Stochastic Models*, vol. 22, pp. 3–20, 2006.

[35] E. S. Badila, O. J. Boxma, and J. A. Resing, "Queues and risk processes with dependencies," *Stochastic Models*, vol. 30, pp. 390–419, 2014.

[36] G. Panda, A. D. Banik, and M. L. Chaudhry, "Stationary distributions of the R[X]/R/1 cross-correlated queue," *Communications in Statistics - Theory and Methods*, vol. 46, pp. 8666–8689, 2017.

[37] P. Buchholz and J. Kriege, "Fitting correlated arrival and service times and related queueing performance," *Queueing Systems*, vol. 85, pp. 337–359, 4 2017.

[38] N. Hadidi, "Queues with partial correlation," *SIAM Journal on Applied Mathematics*, vol. 40, pp. 467–475, 1981.

[39] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, pp. 97–109, 1970.