*Author:*
**Dean, Justin**

*Title:*
**A Queueing Theoretic Model of a Gene Regulatory Network**

# A Queueing Theoretic Model of a Gene Regulatory Network

Justin Dean

Submitted to the University of Bristol in accordance with the requirements of the degree of PhD in Mathematics in the Faculty of Science.

School of Mathematics

September 2018

Word Count: 39494

# Abstract

We consider an infinite server queue into which customers arrive according to a Cox process and have independent service times with a general distribution. The model is motivated by a linear feed-forward gene regulatory network, in which the rate of protein synthesis is modulated by the number of RNA molecules present in a cell. The system can be modelled as a nonstandard tandem of infinite server queues, in which the number of customers present in a queue modulates the arrival rate into the next queue in the tandem. We first study second order statistics of the equilibrium queue length by making a simplifying assumption on the service time distribution. We then establish a Large Deviations Principle for this queueing system in the asymptotic regime in which the arrival process is sped up, while the service process is not scaled.

# Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .......................

DATE: ...........................

# Acknowledgements

I would like to thank my supervisor, Ayalvadi Ganesh, for all the support he gave me. I am extremely grateful for his unending patience, academic guidance, constant encouragement and positive attitude, all of which were invaluable to me. I also owe him a debt of gratitude for first inspiring me through his excellent lectures on applied probability and introducing me to the topics in this thesis.

Thanks also must go to Edward Crane who generously gave up a lot of his own time to discuss the work in this thesis and whose ideas contributed substantially to the work in chapter 5. My thanks also go to Angus Davidson and Sean Ledger for many useful discussions and a great deal of patience, and to the rest of the Probability group at Bristol for many fruitful conversations.

Financial support from the Engineering and Physical Sciences Research Council is gratefully acknowledged.

Finally I must thank my family and friends for their unwavering support.

# Contents

# List of Figures

# Notation

## List of Symbols

| | |
|---|---|
| $\overset{d}{=}$ | equality in distribution |
| $\approx$ | approximately equal to |
| $\varnothing$ | the empty set |
| $\mathbb{C}$ | the set of complex numbers |
| $\mathbb{N}$ | the set of natural numbers $\{1,2,...\}$ |
| $\mathbb{N}_0$ | $\mathbb{N} \bigcup \{0\} = \{0,1,2,...\}$ |
| $\mathbb{R}$ | the set of real numbers |
| $\mathbb{R}_+$ | the set of non-negative real numbers |
| $\mathbb{R}^*$ | the extended real numbers, $\mathbb{R} \bigcup \{+\infty\}$ |
| $\mathbb{R}_+^*$ | the non-negative extended real numbers, $\mathbb{R}_+ \bigcup \{+\infty\}$ |
| $\mathbb{R}^d$ | $d$-dimensional Euclidean space |
| $\mathbb{Z}$ | the set of integers $\{...,-2,-1,0,1,2,...\}$ |
| $\lfloor x \rfloor$ | the largest integer smaller than or equal to $x \in \mathbb{R}$ |
| $\lceil x \rceil$ | the smallest integer larger than or equal to $x \in \mathbb{R}$ |
| $x!$ | $x \cdot (x-1) \cdot ... \cdot 2 \cdot 1$, the factorial of $x \in \mathbb{N}$ |
| $x^+$ | $\max\{x,0\}$, the positive part of $x$ |
| $x \wedge y$ | $\min\{x,y\}$, the minimum of $x$ and $y$ |
| $\langle x,y \rangle$ | $\sum_{i=1}^d x_i y_i$, the inner product of $x, y \in \mathbb{R}^d$ |
| $\mathbf{1}$ | the all one vector |
| $\mathbf{v}^T$ | the transpose of a vector $\mathbf{v}$ |
| $B^\circ$ | the interior of a set $B$ |
| $\overline{B}$ | the closure of a set $B$ |
| $B^c$ | the complement of a set $B$ |
| $\|\cdot\|_\infty$ | the uniform norm |
| $\mathbb{1}_B(\cdot)$ | the indicator function of a set $B$ |
| $f^{-1}(A)$ | the inverse image of a set $A$ under a function $f$ |
| $(f \circ g)(x)$ | $f(g(x))$, composition of functions $f$ and $g$ |
| $f^*$ | convex conjugate or Fenchel-Legendre transform of $f$ |
| $\mathcal{D}_I$ | $\{x : I(x) < \infty\}$, the effective domain of a function $I$ |
| $Lip(\alpha)$ | the set of Lipschitz functions with Lipschitz constant $\alpha$ |
| $C_b(X)$ | the set of bounded continuous functions on $X$ |
| $D$ | the space of càdlàg functions |

| | |
|---|---|
| $\mathbb{E}(X)$ | expectation of a random element $X$ |
| $Var(X)$ | variance of a random element $X$ |
| $Cov(X,Y)$ | covariance of random elements $X$ and $Y$ |
| $C_X^2$ | the squared coefficient of variation, $Var(X)/\mathbb{E}(X)^2$ |
| $\mathbb{P}(A)$ | probability of an event $A$ |
| $F_X(t)$ | $\mathbb{P}(X \le t)$, the cumulative distribution function of $X$ evaluated at $t$ |
| $\overline{F}_X(t)$ | $1 - F_X(t) = \mathbb{P}(X > t)$, the c.c.d.f. of $X$ evaluated at $t$ |
| $P_n \Rightarrow P$ | weak convergence of a sequence of probability measures $P_n$ to $P$ |
| $\mathcal{N}(\mu, \sigma^2)$ | the Normal distribution with mean $\mu$ and variance $\sigma^2$ |
| $H_k$ | Hyperexponential distribution with $k$ phases |
| $E_k$ | Erlang distribution with $k$ phases |
| $genE_k$ | Generalised Erlang or Hypoexponential distribution with $k$ phases |
| $PH_k$ | Phase-type distribution with $k$ phases |
| $(E,d)$ | a Polish space $E$ equipped with a metric $d$ |
| $\mathcal{B}_E$ | the Borel $\sigma$-algebra on $E$ |
| $\mathcal{M}_1(\mathcal{X})$ | the space of probability measures on a set $\mathcal{X}$ |
| $\mathcal{M}_{\le 1}(\mathcal{X})$ | the space of subprobability measures on a set $\mathcal{X}$ |
| $\mathcal{M}_+^f(\mathcal{X})$ | the space of finite non-negative Borel measures on a set $\mathcal{X}$ |
| $\delta_x$ | the Dirac measure at $x$ |
| $H(\nu, \mu)$ | the relative entropy of $\nu$ with respect to $\mu$ |
| $\Phi$ | a point process |
| $\Pi_\lambda$ | the distribution of a Poisson point process with intensity measure $\lambda$ |

## Abbreviations

| | |
|---|---|
| a.s. | almost surely |
| i.i.d. | independent and identically distributed |
| c.c.d.f. | complementary cumulative distribution function |
| càdlàg | right continuous with left limits |
| LLN | Law of Large Numbers |
| FLLN | Functional Law of Large Numbers |
| CLT | Central Limit Theorem |
| FCLT | Functional Central Limit Theorem |
| LDP | Large Deviations Principle |
| FLDP | Functional Large Deviations Principle |
| WLDP | Weak Large Deviations Principle |
| MDP | Moderate Deviations Principle |

# 1 Introduction

## 1.1 Biological Motivation

The central dogma of molecular biology is that DNA makes RNA, which makes protein [28]. Each RNA molecule synthesises proteins, so the more RNA molecules there are, the faster the total rate of protein production. Similarly, the rate of RNA transcription depends upon the number of active genes. Networks of genes, RNA molecules and proteins are often called gene regulatory networks. In practice these networks include feedback mechanisms in addition to the obvious feed-forward network topology suggested by the central dogma. For instance, if there are enough molecules of a certain species of protein in the cell, a signal is sent to repress the gene responsible for its production.

The life of a cell is inherently stochastic. Chemical reactions in cells happen according to the random collisions of molecules. This means that cellular complexes are formed at random times. Additionally, organic molecules have both finite and random lifespans. The external environment is also constantly changing. Many molecular species in a cell are present in very low numbers (sometimes with as few as one copy present); this means that stochastic fluctuations in molecular abundances can have profound effects on cellular concentration levels - the signal to noise ratio can be very low. This cellular noise is typically harmful, reducing efficiency and constraining functionality. In principle, if one knew the positions and momenta of all particles, one could model the system deterministically with Newton's equations of motion. But it is clear that measuring all this information is not possible in any practical sense and such an approach does not scale. So modelling the system stochastically provides a simpler interpretation and acts as a good approximation.

The goal of a cell, loosely speaking, is to carry out a multitude of different tasks. This is principally achieved by carefully balancing molecular concentration levels within the cell. It tries to reach an optimal state for the desired functionality and then maintain this state by keeping the chemical composition roughly constant. However keeping concentration levels within narrow bands is not straightforward due to the destabilising forces caused by a changing environment and a whole host of sources of stochastic fluctuations. So it is important that cells are robust in the face of noise. Efficiency is also essential. It is not enough to produce sufficient numbers of molecules of the desired types, the cell must also be careful to reduce wastage and not form redundant molecules as this uses up valuable shared resources. Of course in reality the cell is not an autonomous agent making decisions. The behaviour we observe is a consequence of a large number of feedback loops.

There is a great deal of interest amongst biologists in the statistical fluctuation properties of gene regulatory networks. They would like to have a better mechanistic understanding of cellular noise. In particular, they would like to understand how the correlation structure in molecular count data that have been observed experimentally actually arises. Indeed, there are sometimes marked correlations between ostensibly unrelated molecular species [27]. But since there is so much happening within a cell, with millions of species and reactions, all linked in a hugely complicated web of interdependencies, not to mention the low signal to noise ratio, it becomes nigh on impossible to extract causal relations from statistical data analysis. The classical scientific approach to inferring causality from data is to perform a controlled trial. However in this instance it is hopeless to think one could control the chemical compositions and positions and momenta of all molecules across multiple cells. Even if one could in principle do this, the experimental design problem would be combinatorially astronomical, as a priori it is not clear which species directly or indirectly affect each other. A bottom-up mathematical model can circumvent some of these problems. With a model one can perform pseudo-experiments in which one artificially varies just one model parameter in isolation. Additionally, finding outputs of the model explicitly in terms of the input parameters can shed light on the mechanistic or causal structure. Of course the model predictions could be tested against real data. This description motivates the use of a stochastic model of a gene regulatory network. Our goal is to analyse this model to shed light on the causal structure. The stochastic nature of the model is particularly important in capturing behaviour of those species which typically have low copy counts. Mere averages do not suffice in such cases. For the sake of mathematical tractability we shall omit the effects of feedback.

## 1.2 A Motivating Markov Chain Model

We consider a simple model of a gene regulatory network, which is a simplification of that in [83] in that we do not keep track of the number of active genes present. We let $N_1$ and $N_2$ represent the number of RNA molecules and proteins in a cell respectively. An RNA molecule is transcribed by genes after an Exponential time of rate $\lambda_1$ (we are essentially assuming that the combined action of genes is given by a constant rate Poisson process). An individual RNA transcript synthesises protein molecules after an Exponential time of rate $\lambda_2$. The time until the next protein synthesis is therefore the minimum of $N_1$ independent $Exponential(\lambda_2)$ times, that is an $Exponential(\lambda_2 N_1)$ time. But since $N_1$ is a random number which changes over time, proteins are formed at the increments of a Poisson process with a stochastic intensity - a Cox process (see section 2.1 for a formal definition). So the rate of protein production is modulated by the number of transcripts. Additionally, each RNA and protein molecule degrades after an independent Exponential time of rate $\mu_1$ and $\mu_2$ respectively. So the number of RNA and protein molecules decreases by one upon the independent degradation of any one of the molecules present and so is again linear in the molecular count and so happens after an Exponential time of rate $N_1\mu_1$ and $N_2\mu_2$ respectively. We summarise this information into a set of four elementary reactions:

$$\begin{aligned} N_1 &\xrightarrow{\lambda_1} N_1 + 1, \\ N_1 &\xrightarrow{N_1\mu_1} N_1 - 1, \\ N_2 &\xrightarrow{\lambda_2 N_1} N_2 + 1, \\ N_2 &\xrightarrow{N_2\mu_2} N_2 - 1. \end{aligned} \tag{1}$$

These reactions can be interpreted as describing the transition rates of a Markov chain in state $(N_1, N_2)$. We can use the reaction rates from (1) to write down the Chapman-Kolmogorov forward equations for this Markov chain:

$$\begin{aligned} \frac{dP_t(N_1,N_2)}{dt} &= \lambda_1 P_t(N_1-1,N_2) - \lambda_1 P_t(N_1,N_2) \\ &+ (N_1+1)\mu_1 P_t(N_1+1,N_2) - N_1\mu_1 P_t(N_1,N_2) \\ &+ \lambda_2 N_1 P_t(N_1,N_2-1) - \lambda_2 N_1 P_t(N_1,N_2) \\ &+ (N_2+1)\mu_2 P_t(N_1,N_2+1) - N_2\mu_2 P_t(N_1,N_2), \end{aligned}$$

where $P_t(N_1, N_2)$ is the probability of being in state $(N_1, N_2)$ at time $t$. Because of the linearity inherent in the reaction rates for this system, one can explicitly find exact analytical expressions for the time-dependent moments of the molecular count processes using generating functions [83]. Note it is very easy to generalise this set-up to a feed-forward network of arbitrary length. Before continuing our study of this model in section 1.5, we make a brief foray into an area of mathematics known as queueing theory.

## 1.3 A Brief Introduction to Queueing Theory

Queueing theory is the formal mathematical study of waiting lines. The original motivation came from the pioneering work of the Danish engineer Erlang [41], who wished to model the amount of traffic experienced by a telephone exchange. From an engineering point of view, there is a trade-off to be struck between having enough wires and machinery so that callers have a good service (in the sense that they do not have to wait a long time to make a call) whilst not having a large amount of expensive and redundant equipment sitting around. In order to reach a sensible compromise, it is important that one has an understanding of the levels of traffic at the exchange. The time evolution of the number of people wanting to make a telephone call in Denmark is evidently stochastic, so mere averages do not adequately capture the behaviour of the system. What is preferable is a probabilistic analysis that can capture the inherent fluctuations.

The stochastic model Erlang used is called a queueing model (or just a queue for short). At a high level this typically describes a system into which customers (or jobs) arrive, queue up in a buffer, then receive service before exiting the system. Mathematically speaking, a queue is described by an arrival process (a rule governing how work arrives to the system), a distribution of service times (specifying the amount of work each job comes bearing), some number of servers working on the jobs, and a policy that determines how service is allocated. This information is succinctly conveyed by Kendall's $X/Y/Z-W$ notation [60]. The $X$ specifies the stochastic process of arrivals, $Y$ the service time distribution, $Z$ the number of servers, and $W$ the service discipline which dictates how servers divide their attention amongst customers. For instance, for the $M/G/\infty$ queue, the $M$ tells us that the arrival process is Markovian - specifically that customers arrive at the increments of a Poisson process (equivalently that the interarrival times are i.i.d. Exponential random variables), the $G$ that the service times follow some general (arbitrary) distribution, and there is an infinite number of servers.

Due to the infinite number of servers, whenever a customer arrives to the system there is always a free server that instantaneously begins servicing it until its completion, whereupon it exits the system. Hence, its sojourn time in the system is simply its own service requirement. Calling this system a queue is something of a misnomer as no customer ever waits in line (by queue length we really just mean the number of customers in service at a given time). For this reason it was not necessary to specify a service discipline. Customers do not feel the effects of each other, so there is a great deal of independence which makes this model particularly tractable. Of course if the service time distribution changes over time (for instance according to some background process), then customer sojourn times can be correlated. The assumption of an infinite number of servers may seem unrealistic, but in some applications (see for instance the model studied in chapter 4) this makes physical sense. Alternatively, sometimes infinite server queueing models are used as approximations to systems with a large number of servers, which arriving customers typically find to be underloaded.

Since the early work of Erlang much has been discovered about a wide range of queueing models. These models have been successfully employed in a diverse range of applications including (but not limited to) retail, telecommunications, computing, industrial engineering and many more. It is a testament to its versatility that the same theory can be used to describe customers standing in line in a shop, traffic on a road, or internet traffic in a fibre optic cable.

## 1.4 Some Preliminaries on Infinite Server Queues

The simplest infinite server queue is the $M/M/\infty$ queueing model. For this model customers arrive according to the increments of a Poisson process of fixed rate $\lambda$ and their service times are i.i.d. $Exponential(\mu)$ distributed random variables. The stochastic process $(X_t, t \geq 0)$ tracking the time evolution of the number of customers in the system is clearly a continuous time irreducible birth-death Markov chain, and as such it admits a unique invariant distribution. By considering the infinitesimal generator of the chain, it is a trivial exercise to solve the detailed balance equations and normalisation condition to reveal that the equilibrium distribution for the number of customers in an $M/M/\infty$ queue follows a Poisson distribution with mean $\rho := \frac{\lambda}{\mu}$ (see section 5.5.2 of [75], for example, for further details).

The $M/G/\infty$ queue is a generalisation of this where job sizes are i.i.d. and generally distributed according to some distribution $F$ supported on $\mathbb{R}_+$ with mean $\frac{1}{\mu}$. In particular we are no longer assuming that the service time distribution is Exponential. This means that the queue length process is no longer a Markov chain. Nevertheless, it turns out that the $M/G/\infty$ queue exhibits what is known as the insensitivity property, meaning that the equilibrium queue length distribution still follows a $Poisson(\rho)$ distribution. It is insensitive in the sense that the stationary distribution for the number of customers in the queue only depends upon the job size distribution through its mean. Note, the insensitivity property is not robust to changes in the arrival process (for an example see [77]). One could reasonably additionally keep track of all of the residual service times and obtain a Markov chain representation on a much larger state space. But the unwieldy size of the state space means that the chain is no longer especially amenable to analysis by traditional Markov chain methods. Instead, it is convenient and fruitful to analyse this queueing model by appealing to the theory of point processes. For a formal description of point processes, see section 2.1.

## 1.5 An Equivalent Tandem Queueing Model

A natural analogy can be drawn between the number of customers in a queue and the number of molecules of a certain species in a cell. Each customer represents an individual molecule, while its service requirement can be thought of as its lifetime. So the arrival or departure of a customer portrays the formation or degradation respectively of a molecule. It is assumed that all molecules have independent lifespans, in this case it makes sense to have an infinite server queue. This means that all customers are served in parallel from the instant they arrive - equivalently, no molecule need wait after its formation to start ageing.

By virtue of this analogy we can translate the Markov chain model into a nonstandard tandem queueing model whose arrival and service rates are given by the reaction rates in (1) of section 1.2 and which satisfies exactly the same Chapman-Kolmogorov equations. It is nonstandard in the sense that there is no routing of customers between queues as in the normal tandem. The number of RNA transcripts can be thought of as the occupancy of an $M/M/\infty$ queue and the number of proteins present is given by the occupancy of a $Cox/M/\infty$ queue. The Cox process allows one to capture the burstiness experimentally observed in protein production where there are sporadic periods of high level activity interspersed between long stretches of low level translation (see section 2.1 for a formal definition). Figure 1 shows a schematic of the tandem queueing model.

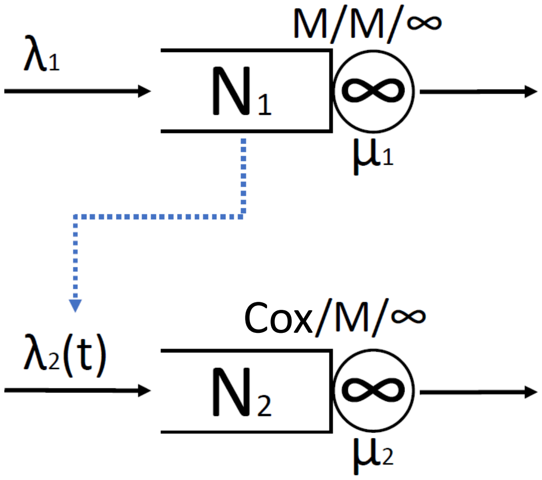

**Figure 1** – Schematic of the tandem queueing model.

The tandem queueing model is mathematically equivalent to the simple Markov model described earlier. However the queueing model permits some generalisations. In particular, we would like to relax the assumption of Exponential lifetimes of molecules. So we take the service time distributions for both queues to be i.i.d. from arbitrary distributions. This means that the queueing model now consists of an $M/G/\infty$ and

$Cox/G/\infty$ queue. The biological motivation for this more general model of decay is that the degradation pathway of a molecule is typically a multistage and incremental process; this means that the assumption of a memoryless Exponential lifetime may not be a justified one. Now that service times are not Exponential, the queue length process is no longer Markovian since knowledge of the past of the process carries information about the elapsed service times and hence the residual service requirements of customers and is thus informative for the future of the queue length process.

Rather than viewing the $Cox/G/\infty$ queue as a continuous time Markov process on a very large state space, we instead interpret it using a point process on the upper half plane. This is depicted for a $\bullet/G/\infty$ queue in Figure 2. The horizontal axis represents time and the vertical half axis corresponds to service requirement. A point (a hollow circle in the figure) $(t, x) \in \mathbb{R} \times \mathbb{R}_+$ represents a customer that arrived at time $t$, bearing a service requirement of size $x$. The 45 degree line between a point and the horizontal axis shows the residual service requirement of a customer over time. Hence the number of points falling in the orange wedge is exactly the number of customers in the queue at time 0. The points in the intersection of the green and orange wedges represents those customers that are in the queue at both times 0 and $\tau$. For a sample path of the $Cox/G/\infty$ queue, the locations of the points in the upper half plane are given by a realisation of a Cox point process on $\mathbb{R} \times \mathbb{R}_+$. We shall study properties of this spatial point process and the corresponding queue whose arrivals are given by it.

Note, just as we could extend the simple Markov model to a feed-forward network of arbitrary length, we can do the same with the queueing model. We can simply add $Cox/G/\infty$ facilities, where the arrival rate into each system is modulated by the occupancy of the previous facility. Note that this is not the usual tandem queueing model as there is no routing of customers between queues. The effect of the previous facility on each queue is only felt through its arrival rate. The behaviour of this nonstandard queueing network model is the basis of study of this thesis. We are yet to think of a more appropriate name than a nonstandard tandem. Note, rather than working with point processes, an alternative approach may have been to write down the joint Laplace transform of the two queues and analyse this.

## 1.6 Outline

The $M/G/\infty$ queue is a well understood model. The $Cox/G/\infty$ queue conversely has received very little attention. It is an instance of an infinite server queue in a random environment. This simply means that the arrival and or service process are modu-

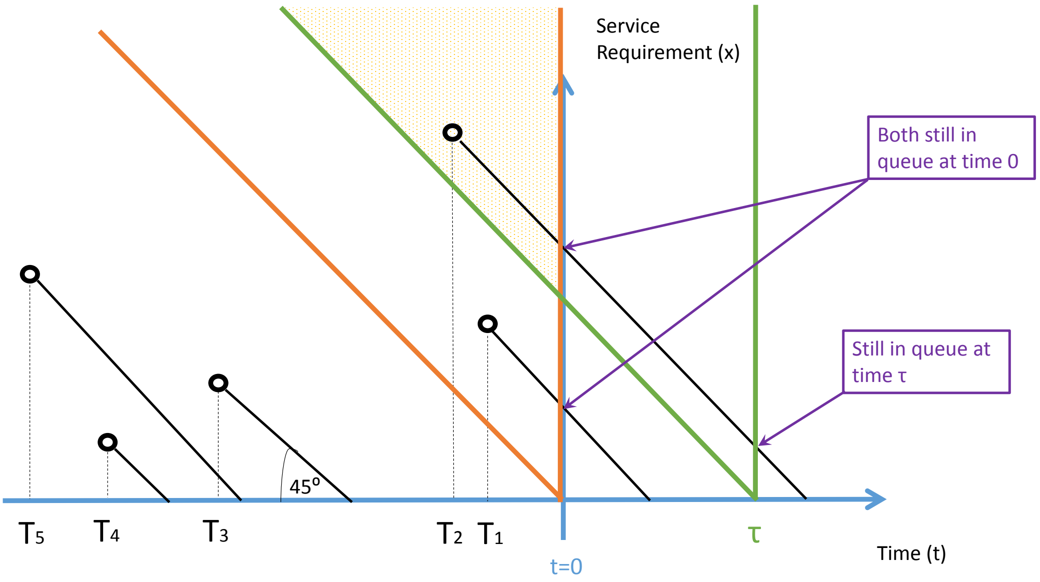

**Figure 2** – Diagram representing a spatial Poisson process model of a $\bullet/G/\infty$ queue. Hollow circles represent the arrivals of customers, which happen at random times and with random job sizes. The black 45 degree lines show the residual service requirement of a customer. The orange and green cones give the regions where customers would have to arrive (that is, their arrival time and service requirements) so that they are still present in the queue at times 0 and $\tau$ respectively. The dotted conical region is the overlap between the first two conical regions, and so represents those customers that are in the queue at both times 0 and $\tau$.

lated by some background process. In the motivating example above, only the arrival process is modulated and the modulation mechanism is simply that the arrival rate is proportional to the number of jobs in the previous facility. In principle the Coxian arrival process need not be modulated in this way, one could imagine many other driving processes.

The outline of this thesis is as follows. First we introduce a number of mathematical preliminaries formally that are needed for the rest of the thesis. We provide a brief formal description of point processes and their relevant properties in section 2.1. Then we introduce Phase-type distributions in section 2.2. We give some background on large deviations in section 2.3 that is used in chapter 5.

We proceed to give an account of the relevant literature in chapter 3. This starts by reviewing some classes of deterministic models of biochemical reaction networks in section 3.1 and discusses their strengths and weaknesses. Then we turn our attention to stochastic models, first touching on some Markov chain models in section 3.2 and then models of a queueing theoretic nature in section 3.3. In section 3.4 we report what is known about infinite server queues in random environment. The discussion then changes focus in section 3.5 to large deviations for random point measures induced by spatial Poisson and Cox point processes. The empirical process approach of section 3.6 provides an alternative viewpoint of empirical measures. Instead of viewing them as random measures, one thinks of them as random functionals indexed by a class of functions which one integrates against the measure. Literature on limit theorems and large deviations for infinite server queues of a similar nature to our own is the subject of section 3.7. We lastly review, in section 3.8, some asymptotic theory in a field of geometric probability called stochastic geometry.

Chapter 4 is concerned with second order statistics of the queueing model when viewed as a stochastic process on the space of càdlàg functions. Specifically, we derive closed form expressions for the autocovariance of the queue length process at equilibrium. To obtain these results in closed form we have to make the simplifying assumption that the service time law has a so called Phase-type distribution. These distributions are a generalisation of mixtures of Exponential and Gamma distributions - see section 2.2 for a formal definition and further details. There is little loss in making this assumption since the Phase-type distributions are dense in the class of probability distributions supported on $\mathbb{R}_+$. So in principle one could approximate any service time distribution arbitrarily well within this framework.

In section 4.1 we provide some biological motivation and interpretation for the calculations performed in the rest of chapter 4. Essentially we are interested in quantifying the speed of noise suppression and dissipation of fluctuations in RNA and protein counts. In section 4.2 we calculate the average number of proteins at equilibrium when molecular lifetimes are generally distributed. We then assume that all molecular lifetimes follow a Phase-type distribution which enables us to explicitly calculate the autocovariance function for the $M/PH_k/\infty$ queue at stationarity in section 4.3. This gives a quantification of the noise suppression ability of RNA molecules. Then in section 4.4 we do the same for proteins by studying the $Cox/PH_k/\infty$ queue. The result in both cases is essentially that fluctuations dissipate exponentially fast, where the rate of decay is related to the spectral properties of a parameter of the Phase-type distribution. We then perform some simulations, calculating the stationary autocorrelation function for queues with two particularly simple Phase-type distributions in section 4.5. The results obtained match what the theory predicts very well. The accompanying code for the simulations can be found in chapter 7. In section 4.6 we calculate the power spectral densities associated to the autocovariance functions mentioned above. These tell us about the behaviour of fluctuations at different frequencies. Finally, we discuss the robustness of the Phase-type approximation scheme in section 4.7.

In the next chapter we consider the rare event behaviour of the system. A particular species of protein may typically only have one or two copies present in the cell at a given time. Due to the randomness inherent in the gene regulatory network, one may often observe small fluctuations around this average copy count. But very rarely one may see a substantial excursion away from the typical number and observe as many as ten proteins, say. Since many resources needed in multiple reactions are shared and scarce, such a large deviation can be potentially damaging to the cell. For this reason we wish to estimate the probability of seeing such a large deviation and ask how quickly this decays with the size of the excursion away from the mean. This is exactly the subject of chapter 5. It uses the theory of large deviations to make these statements about rare events precise. We refer the reader to section 2.3 for the definition of the Large Deviations Principle which formalises the description of rare events. We prove a Large Deviations Principle for the $Cox/G/\infty$ queue length process at stationarity. This essentially shows, in a very precise sense, that the probability of observing a large excursion away from the mean decays exponentially fast in the size of the deviation, however the rate of this exponential is only given implicitly as the solution of a particular optimisation problem.

Section 5.1 provides the motivation which inspires the work in this chapter and outlines the main results, whereas section 2.3 introduces all the relevant definitions and theorems from the theory of large deviations relevant to our work. The proofs are split into a few parts. First we prove an LDP for the empirical measure of a Cox process on a Polish space in the projective limit topology in section 5.2. This is then strengthened to the weak topology by showing exponential tightness. Finally, an LDP is proven for the occupancy measure of a $Cox/G/\infty$ queue in section 5.3.

# 2 Mathematical Preliminaries

In this chapter we collect a number of definitions and theorems used throughout the rest of this thesis for the convenience of the reader.

## 2.1 Point Processes

We briefly introduce a number of definitions and results related to point processes that will be used throughout the rest of this thesis. Intuitively speaking, a point process is just a mathematical description of a collection of points scattered according to some random rule on some space. We now provide a more formal description. The material in this section relies heavily on the treatment of point processes in [66]. For a much more detailed account the reader is referred to [29].

Let $(\mathbb{X}, \mathcal{X})$ be a measurable space. Let the space of all measures $\mu$ on $\mathbb{X}$ with the property that $\mu(A) \in \mathbb{N}_0$ for any $A \in \mathcal{X}$, be denoted by $\mathbf{N}_{<\infty}(\mathbb{X})$, or just $\mathbf{N}_{<\infty}$ for short. Now write $\mathbf{N}(\mathbb{X})$ (or just $\mathbf{N}$ for short) for the space of all measures that can be written as a countable sum of elements of $\mathbf{N}_{<\infty}$. We define the $\sigma$-algebra $\mathcal{N}$ on $\mathbf{N}$ by

$$\mathcal{N} := \sigma\left(\{\mu \in \mathbf{N} : \mu(A) = k\} : A \in \mathcal{X}, k \in \mathbb{N}_0\right).$$

**Definition 2.1.** *([66] Definition 2.1)* ***Point Process***
*Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a point process on $\mathbb{X}$ is a measurable mapping $\eta : \Omega \to \boldsymbol{N}$.*

So a point process is just a random element of $(\mathbf{N}, \mathcal{N})$, that is a random counting measure. It counts the number of points that fall in any measurable set. One might ask what the average number of points is for a given measurable set. The answer is given by the intensity measure.

**Definition 2.2.** *([66] Definition 2.5)* ***Intensity Measure***
*The intensity measure (or first moment measure) of a point process $\eta$ on $\mathbb{X}$, is the measure $\lambda$ defined by $\lambda(A) := \mathbb{E}[\eta(A)]$, where $A \in \mathcal{X}$.*

Note that $\lambda(A)$ is well-defined, though possibly infinite because $\eta(A)$ is a non-negative random variable.

Before we introduce the Poisson point process, we first define what it means for a measure to be $s$-finite.

**Definition 2.3.** *([66] Page 10)* ***s-finite***
*We say that a measure $\nu$ on $\mathbb{X}$ is s-finite if it is a countable sum of finite measures.*

In this thesis, we work with finite measures, which are trivially $s$-finite. We can now define the Poisson point process which is a model of complete spatial randomness of a collection of points in the sense that they do not interact (there are no repulsive or attractive forces between points).

**Definition 2.4.** *([66] Definition 3.1)* ***Poisson Process***
*Let $\lambda$ be an s-finite measure on $\mathbb{X}$. A Poisson process with intensity measure $\lambda$ is a point process $\eta$ on $\mathbb{X}$ with the following two properties:*

1. *For every $B \in \mathcal{X}$, the distribution of $\eta(B)$ is Poisson with parameter $\lambda(B)$.*
2. *For every $m \in \mathbb{N}$ and all pairwise disjoint sets $B_1, ..., B_m \in \mathcal{X}$ the random variables $\eta(B_1), ..., \eta(B_m)$ are mutually independent.*

One of the main objects of study of this thesis is the Cox point process. The rest of this subsection is devoted to introducing them. A Cox process is a generalisation of a Poisson point process and is the result of a two-stage construction: first one generates a random intensity measure, then constructs a Poisson point process with that measure as its intensity measure. As there are two layers of randomness this is sometimes called a doubly stochastic process. We first introduce the notion of random measures and then formally define the Cox process.

As before, let $(\mathbb{X}, \mathcal{X})$ be a measurable space. Denote by $\mathbf{M}(\mathbb{X}) = \mathbf{M}$, the set of $s$-finite measures on $\mathbb{X}$. We denote by $\mathcal{M}(\mathbb{X}) = \mathcal{M}$ the $\sigma$-algebra generated by $\{\mu \in \mathbf{M} : \mu(B) \leq t\}, B \in \mathcal{X}, t \in \mathbb{R}_+$. This $\sigma$-algebra is the smallest possible to make the mappings $\mu \mapsto \mu(B)$ measurable for all $B$ in $\mathcal{X}$.

In what follows there is a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random elements are defined.

**Definition 2.5.** *([66] Definition 13.1)* ***Random Measure***
*A random measure on $\mathbb{X}$ is a random element $\xi$ of the space $(\boldsymbol{M}, \mathcal{M})$, that is, a measurable mapping $\xi : \Omega \to \boldsymbol{M}$.*

We use the shorthand $\xi(B)$ to denote the random variable $\omega \mapsto \xi(\omega, B) := \xi(\omega)(B)$. In fact every point process is a random measure, so this is a more general notion.

Now let $\lambda \in \mathbf{M}(\mathbb{X})$ and write $\Pi_\lambda$ to denote the distribution of a Poisson point process with intensity measure $\lambda$. Theorem 3.6 of [66] guarantees that such a process exists.

**Definition 2.6.** *([66] Definition 13.5)* ***Cox Process***
*Let $\xi$ be a random measure on $\mathbb{X}$. A point process $\eta$ on $\mathbb{X}$ is called a Cox process directed by $\xi$ if*

$$\mathbb{P}(\eta \in A|\xi) = \Pi_{\xi}(A), \quad \mathbb{P}\text{-a.s.}, \; A \in \mathcal{N}.$$

*Then $\xi$ is called a directing random measure of $\eta$.*

For a given probability measure $\mathbb{Q}$ on $(\mathbf{M}, \mathcal{M})$, we always have the that the Cox process directed by a random measure with distribution $\mathbb{Q}$ exists (see [66] section 13.2, page 130).

## 2.2 Phase-Type Distributions

Phase-type distributions are a class of probability distributions of non-negative random variables. Loosely speaking, they are a generalisation of mixtures of Exponential and Gamma distributions. They have two main advantages: the first is that they are dense in the class of probability distributions supported on $\mathbb{R}_+$ (see [3] Chapter III.6, Theorem 6.2). This means they can in principle be used to approximate any such probability distribution arbitrarily well. Secondly they retain a lot of the tractability of the Exponential distribution (which in some sense is the trivial Phase-type distribution). This owes to the fact that they are made up of Exponential distributions.

Before we define the Phase-type distribution in general we introduce two specific examples that we shall meet again in section 4.5. These are the Hyperexponential distribution and the (generalised) Erlang distribution. These were motivated by applications in which there was a need to model non-negative random variables that exhibited significantly narrower or wider dispersion than an Exponential distribution. A quantification of this notion is given by the following definition.

**Definition 2.7.** ***Squared Coefficient of Variation***
*The squared coefficient of variation of a random variable $X$ is defined to be*

$$C_X^2 = \frac{Var(X)}{\mathbb{E}(X)^2}.$$

It is a normalised version of the variance and can be viewed as a noise to signal ratio. If $X$ follows an Exponential distribution, for example, then $C_X^2 = 1$, while a deterministic random variable (almost surely taking some constant value) has squared coefficient of variation equal to 0. In general having a very low value for $C^2$ (near 0) implies that the random variable is nearly deterministic and a high value (much greater than 1) implies that it is highly variable.

**Definition 2.8.** ***Hyperexponential Distribution***
*We say that the random variable $T$ has the Hyperexponential distribution with $k$ phases (and write $T \sim H_k$) if its probability density function is given by*

$$f_T(t) = \sum_{i=1}^{k} p_i f_{X_i}(t),$$

*where $X_i \sim Exp(\mu_i)$ and $p_i \geq 0$ for all $i \in \{1, 2, ..., k\}$ and $\sum_{i=1}^{k} p_i = 1$.*

So the $H_k$ distribution is just $k$ Exponential distributions arranged in parallel with distinct parameters $\mu_1, ..., \mu_k$ and routing probabilities $p_1, ..., p_k$ (see Figure 3). The squared coefficient of variation for this distribution is greater than 1, so it is an appropriate model for distributions which are more variable than the Exponential distribution.

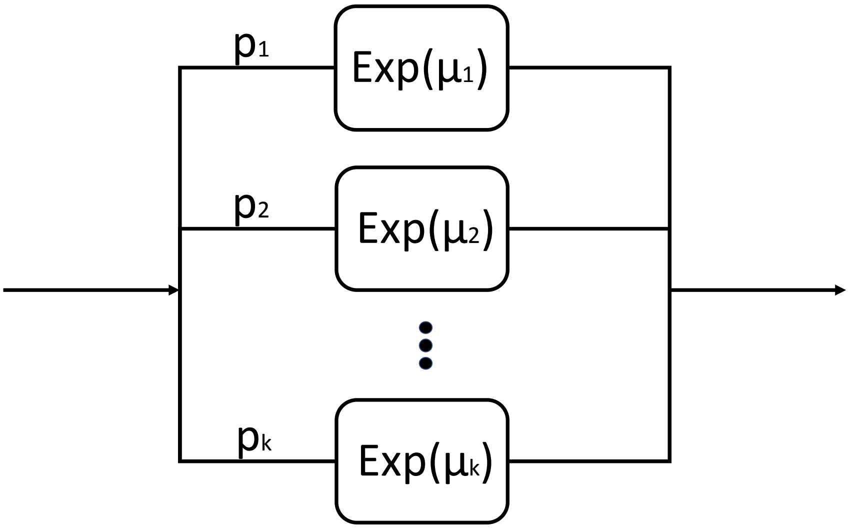

**Figure 3** – Hyperexponential distribution with $k$ phases.

**Definition 2.9.** ***Erlang and Generalised Erlang Distribution***
*We say that the random variable $T$ has the Erlang distribution with $k$ phases (and write $T \sim E_k$) if $T = \sum_{i=1}^{k} X_i$, where $X_1, ..., X_k$ are i.i.d. $Exp(\mu)$ distributed. If $X_1, ..., X_k$ are independent Exponential random variables with different rates $\mu_1, ..., \mu_k$ respectively, then we say that $T$ follows a generalised Erlang (or Hypoexponential) distribution (and write $T \sim genE_k$).*

So the $genE_k$ distribution is simply $k$ Exponential distributions arranged in series (see Figure 4). The squared coefficient of variation for this distribution is less than 1, so it is appropriate for modelling distributions that are less variable than the Exponential

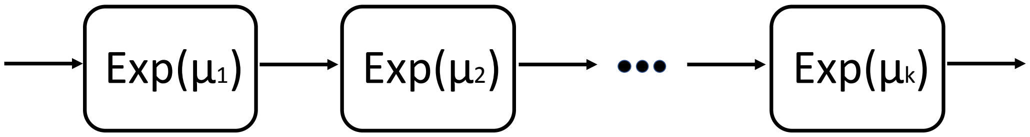

**Figure 4** – Generalised Erlang distribution with $k$ phases.

distribution.

By taking a general mixture of Exponential distributions in series and parallel we obtain a Phase-type distribution. The general definition given below is adapted from [3] Chapter III.6, page 74.

Consider a continuous time Markov chain on a state space with $k+1$ states (where $k \geq 1$). States $1, 2, ..., k$ are transient and state 0 is absorbing. Define a $(k+1)$-dimensional row vector $\widehat{\alpha} := (\alpha_0, \boldsymbol{\alpha^T})$, where $\boldsymbol{\alpha^T} = (\alpha_1, \alpha_2, ..., \alpha_k)$, whose entries give the probabilities of starting in each possible state of the chain. Let $Q$ be the $(k+1)\times(k+1)$ infinitesimal generator matrix of the Markov chain given by

$$Q = \begin{pmatrix} 0 & \mathbf{0} \\ S^0 & S \end{pmatrix},$$

where $\mathbf{0}$ is the $k$-dimensional row vector of zeroes, $S$ is the $k \times k$ subgenerator matrix whose entries give the jump rates between states $1, 2, ..., k$ and $S^0 = -S\mathbf{1}$, where $\mathbf{1}$ is the $k$-dimensional column vector of all ones. With this notation we can now define the Phase-type distribution.

**Definition 2.10.** ***Phase-type Distribution***
*The Phase-type distribution with k phases is the distribution of the time to absorption of the $(k+1)$-state continuous time Markov chain with initial distribution $\widehat{\alpha}$ and generator $Q$. It is parameterised by the initial probability vector $\boldsymbol{\alpha}$ and the subgenerator matrix $S$. If $X$ is a random variable with this distribution we write $X \sim PH_k(\boldsymbol{\alpha}, S)$.*

The distribution function of $X \sim PH_k(\boldsymbol{\alpha}, S)$ can be explicitly calculated from the definition and is given by $F_X(t) = 1 - \boldsymbol{\alpha}^T e^{St}\mathbf{1}$ (where $e^{St}$ is understood to be the usual matrix exponential).

The following result makes precise what we mean by the denseness of Phase-type distributions in the non-negative probability distributions.

**Theorem 2.11.** *([3] Chapter III.6, Theorem 6.2)* ***Denseness of PH Distributions*** *The class $\mathcal{PH}$ of Phase-type distributions is dense in the set $\mathcal{P}$ of probability distributions on $(0,\infty)$. More generally, to any $F \in \mathcal{P}$ with $\mu(F;p) = \int_0^\infty x^p dF(x) < \infty$ for some $p \geq 0$, there are $F_k \in \mathcal{PH}$ with $F_k$ converging in distribution to $F$, and $\mu(F_k;q) \to \mu(F;q)$ for $q \leq p$.*

The main idea of the proof is to approximate the degenerate distribution at a point on $\mathbb{R}_+$ with a sequence of Erlang distributions. Then taking a linear combination of point masses, we can construct discrete distributions on $\mathbb{R}_+$. These in turn are dense in the set of probability distributions supported on $[0,A]$ for $A \in \mathbb{R}_+$. A simple truncation argument completes the proof. For a more detailed discussion of Phase-type distributions see [3] (part A, chapter III, section 6).

## 2.3 Large Deviations Theory

The purpose of this section is to formally introduce what we mean by large deviations. For background on large deviations see the very well known book [38] and references therein. Indeed, the definitions and theorems of this section are taken almost verbatim from this reference and the accompanying discussion is heavily influenced by it. Note that in this section, for the convenience of the reader, we pick out those concepts and results from [38] that are relevant to this thesis. For large deviations theory applied to queueing systems see [44] and its bibliography. Now let us begin with a motivating example from section 1.1 of [38].

Consider a sequence $X_1, X_2, \dots$ of independent Standard Normal random variables. Let $S_n := \sum_{i=1}^n X_i$ be their $n^{th}$ partial sum. Then

$$\frac{S_n}{n} \overset{d}{=} \mathcal{N}\left(0, \frac{1}{n}\right) \quad \text{and} \quad \frac{S_n}{\sqrt{n}} \overset{d}{=} \mathcal{N}(0,1),$$

where $\overset{d}{=}$ denotes equality in distribution. Now fix a $\delta > 0$. By the LLN we know

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| \geq \delta\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Now fixing any interval $A \subset \mathbb{R}$, the classical CLT ensures

$$\mathbb{P}\left(\frac{S_n}{\sqrt{n}} \in A\right) \to \frac{1}{\sqrt{2\pi}} \int_A e^{-\frac{x^2}{2}} dx \quad \text{as} \quad n \to \infty.$$

Using the exact distribution of $\frac{S_n}{\sqrt{n}}$ we have

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| \geq \delta\right) = 1 - \mathbb{P}\left(\left|\frac{S_n}{n}\right| < \delta\right) = 1 - \mathbb{P}\left(\left|\frac{S_n}{\sqrt{n}}\right| < \delta\sqrt{n}\right) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} e^{-\frac{x^2}{2}} dx.$$

Hence,

$$\frac{1}{n} \log \mathbb{P}\left(\left|\frac{S_n}{n}\right| \geq \delta\right) \to -\frac{\delta^2}{2} \quad \text{as} \quad n \to \infty.$$

Thus, $\frac{S_n}{n}$ concentrates around 0, and the probability of being far from 0 decays exponentially in $n$. Such a scaling law is much more widely prevalent, e.g., for sums of random variables with other distributions and which are weakly dependent. Large deviations theory deals with sequences that satisfy such limiting behaviour. In this thesis we often concern ourselves with measure-valued random variables, so we need to describe large deviations in a more abstract setting. The definitions and results given in [38] are presented in a great deal of generality. All of the random variables we work with in this thesis are defined on Polish spaces, so we do not always require the full generality provided.

We are now ready to put this all on a more formal grounding. We say that a topological space is Hausdorff if for any two distinct points, one can find a neighbourhood of each point such that the neighbourhoods are disjoint. So let $\mathcal{X}$ be a Hausdorff topological space with Borel $\sigma$-algebra $\mathcal{B}$, and let $\{\mu_\varepsilon\}$ be a family of probability measures on $(\mathcal{X}, \mathcal{B})$. We use the notation $\mathbb{R}^* = \mathbb{R} \bigcup \{+\infty\}$, and for a set $B$ we denote its interior and closure by $B^\circ$ and $\overline{B}$ respectively.

**Definition 2.12.** *([38] Definition, Page 4)* ***Lower Semicontinuity***
*A function $f : \mathcal{X} \to \mathbb{R}^*$ is lower semicontinuous if $f^{-1}((-\infty\ , \alpha]) = \{x : f(x) \leq \alpha\}$ is closed for all $\alpha \in \mathbb{R}$.*

**Definition 2.13.** *([38] Definition, Page 4)* ***Rate Function***
*A function $I : \mathcal{X} \to \mathbb{R}^*$ is called a rate function if:*

1. *$I(x) \geq 0$ for all $x \in \mathcal{X}$,*

2. *$I(\cdot)$ is lower semicontinuous.*

*$I(\cdot)$ is called a good rate function if its level sets are compact, i.e. $\Psi_I(\alpha) := \{x : I(x) \leq \alpha\}$ is a compact set for all $\alpha \in \mathbb{R}$. We denote the effective domain of $I$ by $\mathcal{D}_I := \{x : I(x) < \infty\}$.*

**Definition 2.14.** *([38] Definition, Page 5)* ***Large Deviations Principle***
*We say that the family of probability measures $\{\mu_\varepsilon\}$ satisfies the Large Deviations Principle (LDP) with rate function $I$, if for all $B \in \mathcal{B}$,*

$$-\inf_{x \in B^\circ} I(x) \le \liminf_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(B) \le \limsup_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(B) \le -\inf_{x \in \overline{B}} I(x).$$

We adopt the convention that the infimum of a mapping over an empty set is $-\infty$. Note the topology enters into this statement by dictating what the interior and closure are as it specifies the open and closed sets. The rightmost and leftmost inequalities are termed the large deviation upper and lower bound respectively. The large deviation upper and lower bounds can be stated equivalently as follows:

([38] Page 6) For any closed set $F \subseteq \mathcal{X}$,

$$\limsup_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(F) \le -\inf_{x \in F} I(x). \tag{2}$$

For any open set $G \subseteq \mathcal{X}$,

$$\liminf_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(G) \ge -\inf_{x \in G} I(x). \tag{3}$$

If the upper bound only holds for compact sets, we say the family $\{\mu_\varepsilon\}$ satisfies a weak LDP (WLDP).

If a family of probability measures concentrates all but an exponentially small amount of probability mass on a compact set we say that the family is exponentially tight. We now define this formally.

**Definition 2.15.** *([38] Definition, Page 8)* ***Exponential Tightness***
*A family of probability measures $\{\mu_\varepsilon\}$ on $\mathcal{X}$ is exponentially tight if for every $\alpha < \infty$, there exists a compact set $K_\alpha \subset \mathcal{X}$ such that*

$$\limsup_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(K_\alpha^c) < -\alpha.$$

This concept plays an important role in the theory of large deviations because it provides a way to upgrade a weak LDP to a full LDP. This is intuitive because if the large deviations upper bound holds for all compact sets and essentially all of the probability mass is concentrated on these sets, then there is no obstacle when extending to all closed sets. This is summarised by the following result.

**Lemma 2.16.** *([38] Lemma 1.2.18)*
*Let $\{\mu_\varepsilon\}$ be an exponentially tight family.*

- *If the upper bound (2) holds for all compact sets, then it also holds for all closed sets.*

- *If the lower bound* (3) *holds for all open sets, then $I(\cdot)$ is a good rate function.*

In other words, a WLDP plus exponential tightness gives a full LDP with a good rate function. Exponential tightness also facilitates the strengthening of an LDP from a coarser to a finer topology.

**Lemma 2.17.** *([38] Corollary 4.2.6)*
*Let $\{\mu_\varepsilon\}$ be an exponentially tight family of probability measures on $\mathcal{X}$ equipped with the topology $\tau_1$. If $\{\mu_\varepsilon\}$ satisfies an LDP with respect to a Hausdorff topology $\tau_2$ on $\mathcal{X}$ that is coarser than $\tau_1$, then the same LDP holds with respect to the topology $\tau_1$.*

One incredibly useful but easily proved result is the Contraction Principle. This states that LDPs are preserved under continuous mappings and so too is the goodness of the rate function.

**Theorem 2.18.** *([38] Theorem 4.2.1)* ***Contraction Principle***
*Let $\mathcal{X}$ and $\mathcal{Y}$ be Hausdorff topological spaces and $f : \mathcal{X} \to \mathcal{Y}$ a continuous function. Consider a good rate function $I : \mathcal{X} \to \mathbb{R}_+^*$.*

- *For each $y \in \mathcal{Y}$, define*

  $$I'(y) := \inf \{I(x) : x \in \mathcal{X}, \quad y = f(x)\}.$$

  *Then $I'$ is a good rate function on $\mathcal{Y}$.*

- *If $I$ controls the LDP associated with a family of probability measures $\{\mu_\varepsilon\}$ on $\mathcal{X}$, then $I'$ controls the LDP associated with the family of probability measures $\{\mu_\varepsilon \circ f^{-1}\}$ on $\mathcal{Y}$.*

Another useful notion is that of exponential equivalence.

**Definition 2.19.** *([38] Definition 4.2.10)* ***Exponential Equivalence***
*Let $(\mathcal{Y}, d)$ be a metric space. The probability measures $\{\mu_\varepsilon\}$ and $\{\widetilde{\mu}_\varepsilon\}$ on $\mathcal{Y}$ are called exponentially equivalent if there exist probability spaces $\{(\Omega, \mathcal{B}_\varepsilon, P_\varepsilon)\}$ and two families of $\mathcal{Y}$-valued random variables $\{Z_\varepsilon\}$ and $\left\{\widetilde{Z}_\varepsilon\right\}$ with joint laws $\{P_\varepsilon\}$ and marginals $\{\mu_\varepsilon\}$ and $\{\widetilde{\mu}_\varepsilon\}$, respectively, such that the following condition is satisfied: For each $\delta > 0$, the set $\left\{\omega : \left(\widetilde{Z}_\varepsilon, Z_\varepsilon\right) \in \Gamma_\delta\right\}$ is $\mathcal{B}_\varepsilon$ measurable, and*

$$\limsup_{\varepsilon \to 0} \varepsilon \log P_\varepsilon(\Gamma_\delta) = -\infty,$$

*where $\Gamma_\delta := \{(\widetilde{y}, y) : d(\widetilde{y}, y) > \delta\} \subset \mathcal{Y} \times \mathcal{Y}$.*

Note the measurability condition is satisfied whenever $\mathcal{Y}$ is a separable space ([38] Remark 4.2.10(b)), and in particular, whenever $\mathcal{Y}$ is a Polish space, as in this thesis. Exponential equivalence essentially means that two families of probability measures are superexponentially close. As a consequence they are indistinguishable on the exponential scale. So if an LDP holds for one family, then the same LDP holds for the other.

**Theorem 2.20.** *([38] Theorem 4.2.13)*
*If an LDP with a good rate function $I(\cdot)$ holds for the probability measures $\{\mu_\varepsilon\}$, which are exponentially equivalent to $\{\widetilde{\mu}_\varepsilon\}$, then the same LDP holds for $\{\widetilde{\mu}_\varepsilon\}$.*

One of the most fundamental results in the theory of large deviations is that of Cramér's Theorem. This describes the LDP for sums of i.i.d. random vectors (or random variables in the one dimensional case). So let $X_1, X_2, \ldots$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$, each with law $\mu \in \mathcal{M}_1(\mathbb{R}^d)$, and denote their empirical mean by $\widehat{S}_n := \frac{1}{n}\sum_{i=1}^n X_i$ whose law we denote by $\mu_n$. For vectors $\lambda, x \in \mathbb{R}^d$ let $\langle \lambda, x\rangle := \sum_{i=1}^d \lambda_i x_i$ be the usual $\mathbb{R}^d$ inner product. Denote the logarithmic moment generating function (or cumulant generating function) of $\mu$ by $\Lambda : \mathbb{R}^d \to \mathbb{R}^*$,

$$\Lambda(\lambda) := M_{X_1}(\lambda) = \mathbb{E}\left[e^{\langle \lambda, X_1\rangle}\right].$$

It is not hard to see that $\Lambda$ is convex, lower semicontinuous and that $\Lambda(0) = 0$. Further it is differentiable in the interior of its effective domain ([38] Lemma 2.2.5 and Lemma 2.2.31). The rate function in Cramér's Theorem turns out to be a special type of transform of the cumulant generating function which we now introduce.

**Definition 2.21.** *([44] Definition 2.5)* ***Convex Conjugate***
*Let $f : \mathbb{R}^d \to \mathbb{R}^*$ be a function which is not identically infinite. The convex conjugate (or Fenchel-Legendre transform) of $f$ is another function $f^* : \mathbb{R}^d \to \mathbb{R}^*$, defined by*

$$f^*(\theta) := \sup_{x \in \mathbb{R}^d} \left(\langle \theta, x\rangle - f(x)\right).$$

We now have all the necessary ingredients to state Cramér's Theorem in $\mathbb{R}^d$.

**Theorem 2.22.** *([38] Theorem 2.2.30)* ***Cramér's Theorem***
*Assume $\mathcal{D}_\Lambda = \mathbb{R}^d$. Then $\{\mu_n\}$ satisfies the LDP on $\mathbb{R}^d$ with the good convex rate function $\Lambda^*(\cdot)$.*

In fact the condition on the effective domain can be relaxed to $0 \in \mathcal{D}_\Lambda^\circ$ instead ([38] Corollary 6.1.6). It is natural to ask for finiteness in a neighbourhood of the origin as the moments are found by differentiating the moment generating function and evaluating at zero. So knowing that limits exist at zero tells you almost everything about the random variable. Upon meeting Cramér's Theorem a natural question to ask is whether the assumption that the random vectors are i.i.d. is necessary in order to obtain an LDP. This turns out not to be the case and is the content of a more general result called the Gärtner-Ellis Theorem. We now introduce the relevant definitions to state a version (by no means the most general) of this result. So let $Z_1, Z_2, \ldots$ be a sequence of random vectors in $\mathbb{R}^d$ and denote the law of $Z_n$ by $\mu_n$ and its cumulant generating function by $\Lambda_n(\lambda) := \log \mathbb{E}\left[e^{\langle \lambda, Z_n \rangle}\right]$. Define the limiting cumulant generating function by

$$\Lambda(\lambda) := \lim_{n\to\infty} \frac{1}{n} \Lambda_n(n\lambda).$$

The rate function will be given by $\Lambda^*(\cdot)$, the convex conjugate of $\Lambda(\cdot)$. But before we state the theorem we need one more definition.

**Definition 2.23.** *([44] Definition 2.6)* ***Essentially Smooth***
*Let $f : \mathbb{R}^d \to \mathbb{R}^*$ be a function. We say that $f$ is essentially smooth if $\mathcal{D}_f^\circ \neq \varnothing$, $f$ is differentiable in the interior of its effective domain, and $f$ is steep - namely, for any sequence $\theta_n$ which converges to a boundary point of the effective domain of $f$, $\lim_{n\to\infty} |\nabla f(\theta_n)| = +\infty$.*

**Theorem 2.24.** *([44] Theorem 2.11)* ***Gärtner-Ellis Theorem***
*Suppose that $\Lambda(\lambda)$, the limiting cumulant generating function of $Z_n$, exists as an extended real number for each $\lambda \in \mathbb{R}^d$ and that it is finite in a neighbourhood of the origin, essentially smooth and lower semicontinuous. Then the sequence $\{\mu_n\}$ satisfies the LDP in $\mathbb{R}^d$ with good convex rate function $\Lambda^*$.*

For examples of what types of mild dependence can be tolerated, see the exercises of section 2.3 of [38]. Another consequence of Cramér's Theorem is Sanov's Theorem. This describes an LDP for the empirical distribution of i.i.d. random variables.

**Theorem 2.25.** *([44] Theorem 4.13)* ***Sanov's Theorem***
*Let $(X_i, i \in \mathbb{N})$ be a sequence of i.i.d. random variables taking values in a Polish space $\mathcal{X}$, with distribution $\mu$. The sequence of empirical measures*

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[X_i \in A] \quad \text{for } A \subset \mathcal{X}$$

*satisfies an LDP in $\mathcal{M}_1(\mathcal{X})$ with good convex rate function $H(\cdot|\mu)$ given by:*

$$H(\nu|\mu) = \begin{cases} \int_{\mathcal{X}} \log(d\nu/d\mu) d\nu & \text{if } \nu << \mu \text{ and } \int |\log(d\nu/d\mu)| d\nu < \infty \\ \infty & \text{otherwise.} \end{cases}$$

*The function $H(\nu|\mu)$ is called the relative entropy or Kullback-Leibler divergence of $\nu$ with respect to $\mu$.*

The next theorem formalises the idea that if we have an LDP on each of an increasing sequence of spaces, then an LDP holds on the infinite limiting space. The following definitions, needed to make this rigorous, are taken (almost verbatim) from [38] section 4.6, page 162. Let $(J, \leq)$ be a partially ordered, right-filtering set. The latter notion means that for any $i, j$ in $J$, there exists $k \in J$ such that both $i \leq k$ and $j \leq k$. A projective system $(\mathcal{Y}_j, p_{ij})_{i\leq j\in J}$ consists of Hausdorff topological spaces $\{\mathcal{Y}_j\}_{j\in J}$

and continuous maps $p_{ij} : \mathcal{Y}_j \to \mathcal{Y}_i$ such that $p_{ik} = p_{ij} \circ p_{jk}$ whenever $i \leq j \leq k$. The projective limit of this system, denoted by $\mathcal{X} := \lim_j \mathcal{Y}_j$, is the subset of the topological product space $\mathcal{Y} = \prod \mathcal{Y}_j$, consisting of all the elements $\mathbf{x} = (y_j)_{j\in J}$ for which $y_i = p_{ij}(y_j)$ whenever $i \leq j$, equipped with the topology induced by $\mathcal{Y}$. Projective limits of closed subsets $F_j \subseteq \mathcal{Y}_j$ are defined analogously and denoted $F = \lim_j F_j$. The canonical projections of $\mathcal{X}$, which are the restrictions $p_j : \mathcal{X} \to \mathcal{Y}_j$ of the coordinate maps from $\mathcal{Y}$ to $\mathcal{Y}_j$, are continuous.

**Theorem 2.26.** *([38] Theorem 4.6.1)* ***Dawson-Gärtner Theorem for Projective Limits***
*Let $\{\mu_\varepsilon\}$ be a family of probability measures on $\mathcal{X}$, such that for any $j \in J$ the Borel probability measures $\mu_\varepsilon \circ p_j^{-1}$ on $\mathcal{Y}_j$ satisfy the LDP with the good rate function $I_j(\cdot)$. Then $\mu_\varepsilon$ satisfies the LDP with the good rate function*

$$I(\boldsymbol{x}) = \sup_{j\in J} \{I_j(p_j(\boldsymbol{x}))\}, \quad \boldsymbol{x} \in \mathcal{X}.$$

# 3 Literature Review

## 3.1 Deterministic Models of Biochemical Systems

There are a wide variety of mathematical modelling techniques employed to shed light on biomolecular networks, each possessing its own strengths and weaknesses. A common approach is to describe the time evolution of the network using systems of differential equations (tracking, for example, the evolution of gene expression levels or molecular counts - see [37, 96] for a list of such works). The rich theory for finding analytical or approximate solutions to systems of ODEs is the principal advantage of this approach. There are however two main drawbacks. The first is that this treats an inherently stochastic system as though it were purely deterministic. This simplification is particularly pronounced in cases where there are low molecular copy counts. This means that stochastic fluctuations are of comparable size to mean values and so ignoring them may lead to quite different behaviour. If fluctuations are however small compared to mean copy counts, then deterministic models capturing the average behaviour of the system will usually do well. The second issue is that differential equations are continuous, whereas real molecular abundances are discrete. A protein, for instance, is either present or not. When molecular counts are very high this is not really an issue, but if they are low then continuous models can be misleading. Furthermore, numerically solving such systems can be computationally demanding, so there is a de facto limit on the size or complexity that can be handled.

A different approach is taken in [58] which uses Boolean networks to describe the dynamics of gene regulation. A Boolean network is a directed graph, in which each node has an associated Boolean function and the state of the network is updated sequentially at each discrete time step. In this biological example nodes represent either the presence or absence of a molecule, or the expression or inactivation of a gene. The topology of the edge set describes the regulatory relationships. The principal drawback of this approach is that every node is either on or off - the relative abundances of molecules and the range of gene expression levels are overlooked. This can be a very crude oversimplification, but the flip side is that the resulting computational complexity is extremely low. This allows for the simulation of huge networks - many orders of magnitude larger than those that can be tackled with other approaches.

## 3.2 Stochastic Models of Biochemical Systems

A substantial review of Markov chain models of gene regulatory networks is conducted in [83]. Most of the references therein fit into (at least part of) the central dogma, with genes transcribing RNA molecules which translate proteins. The dynamics are governed by a small number of stochastic elementary reactions describing the feed-forward network, genes flipping between their active and inactive states and the degradation pathway of each species. A continuous time Markov chain tracks the stochastic time evolution of molecular abundances. Time-dependent moments of copy numbers are then calculated exactly using generating functions, leveraging the linearity of reaction rates and the different sources of noise in the network are characterised in terms of the model parameters.

A generalisation of the results in [83] for the three stage model of a gene regulatory network are given in [43]. The authors calculate the stationary mean and variance of the number of proteins. However, they relax the assumption of Exponentially distributed molecular lifetimes. Lifetimes of molecules are assumed to be i.i.d. from a general distribution supported on $\mathbb{R}_+$. Some of the results obtained are similar to those of chapter 4, as is the approach to find them. They also rely on the analysis of a marked Poisson point process on the upper half plane. They do not calculate the autocovariance function or prove anything about queues, as we do. Their calculation of the variance proceeds by differentiating the generating function, conditional on the state of the gene, to obtain the first two factorial moments, and then averaging over this state.

A three stage gene regulatory network model that incorporates feedback is studied in [39]. Proteins are assumed to have an autoregulatory function, with the gene deactivation rate proportional to the number of proteins. This negative feedback complicates the analysis, so the stationary variance of the number of proteins is not calculated explicitly. Instead the authors study a limiting regime in which genes flip state and RNAs are produced on a timescale much faster than protein evolution. Asymptotically the number of proteins behaves as a birth-death process whose birth rate is proportional to $1/k$ when there are $k$ individuals. The model with and without feedback are compared to examine to what extent the autoregulation mechanism reduces protein count variability.

A far more general biochemical reaction network is analysed in [70] using Markov chain methods. Rather than considering genes, RNA and protein molecules as in [83], the model allows for arbitrary numbers, $n$ and $m$, of molecular species and elementary reactions respectively. A vector $\mathbf{x}(t) \in \mathbb{N}_0^n$ is a continuous time Markov process tracking the molecular abundance of each species over time. The rate of the elementary reaction $i \in \{1, ..., m\}$ is given by $W_i(\mathbf{x}(t)) \geq 0$ and is naturally a function of the chemical composition of the system at the time of occurrence. The vector $r_i \in \mathbb{Z}^n$ describes the change in molecular composition of the system that results from reaction $i$. This could be a combination of reactants being consumed and new products being formed. The set of elementary reactions, therefore, can be concisely expressed as

$$\mathbf{x}(t) \xrightarrow{W_i(\mathbf{x}(t))} \mathbf{x}(t) + \mathbf{r}_i, \quad i = 1, 2, ..., m.$$

Hence, if we denote by $P(\mathbf{k}, t/\mathbf{k}_0, t_0)$, the probability that the chain is in some state $\mathbf{x}(t) = \mathbf{k}$ given that it started in state $\mathbf{x}(t_0) = \mathbf{k}_0$, the Chapman-Kolmogorov forward equation for the system is given by

$$\frac{dP(\mathbf{k}, t/\mathbf{k}_0, t_0)}{dt} = \sum_i W_i(\mathbf{k} - \mathbf{r}_i)P(\mathbf{k} - \mathbf{r}_i, t/\mathbf{k}_0, t_0) - \sum_i W_i(k)P(\mathbf{k}, t/\mathbf{k}_0, t_0).$$

From this one can write down differential equations for transient and stationary first and second moments for $\mathbf{x}(t)$. In general it is not clear how to solve them, but in the special case where reaction rates are linear (or affine), explicit solutions are readily available. Once again, the assumption of Exponential reactions is essential to the analysis.

Since Markov chain models can be computationally expensive to simulate, particularly if they have a large state space, sometimes a diffusion approximation can offer a less cumbersome alternative [68]. Here a Markov chain model of a chemical reaction network is approximated by a reflected diffusion. Since concentration levels cannot go negative, it makes sense to reflect inside the positive orthant. The diffusion is obtained as the weak limit of a sequence of rescaled jump Markov processes which have been centered by subtracting their means and then rescaled appropriately.

## 3.3 Queueing Theoretic Models

Compared with the number of Markov chain models of gene regulatory networks, queueing network models are relatively scarce in the literature and the connection has largely been made only fairly recently. One instance of this can be found in [2], which leverages theory on G-networks to model a well known gene regulatory network present in e.coli known as the lac operon. The lac operon controls the metabolism of lactose in the absence of the more favourable energy source glucose. To explain what a G-network is we first describe so called Jackson networks.

A Jackson network is a network of $\bullet/M/1 - FCFS$ ('first come first served' service discipline) queues in which exogenous traffic arrives to the network according to independent Poisson streams and routing of customers between queueing facilities is Markovian. The network is called open if every customers visits just finitely many queues before exiting the network with probability one. These assumptions ensure that the vector of queue lengths is a continuous time Markov process. Under some additional assumptions one can show that the invariant distribution of the Markov chain is of product form. This means that the equilibrium joint queue length distribution is a product of the marginal equilibrium queue length distributions. For more details on Jackson networks and product form results see the original paper [55] and [25]. Our queueing system of interest falls outside the scope of the first reference as we do not have a single server and outside the second because exogenous arrivals to the system are not Poissonian (except in the first facility).

A G-network (the G stands for generalised or Gelenbe - their creator) is a generalisation of a Jackson network that allows for work deletion and reallocation, yet retains the favourable property of a product form stationary distribution for the network (see [46] for more details). The feature of removing or rerouting traffic provides a way of modelling feedback, such as the inhibition of the expression of a gene. Each queueing facility in the model of the lac operon of [2] represents the molecular count or expression level of some biological species. The main contribution of the paper is to compute the product form stationary distribution for this model. Another instance of a G-network model of a gene regulatory network is provided by [47]. This time queueing facilities represent concentration levels of various genes and the main result is again to compute the product form stationary distribution of the concentration levels. The restriction of having single server queueing facilities makes this an inappropriate model for our applications, because under this model ageing would not happen in parallel.

Certain molecular species in cells show marked correlations in their copy counts. Often this synchronised behaviour is due to a direct interaction between them. Direct coupling mechanisms include coordinated transcription or protein-protein interactions. In practice biologists sometimes observe correlations which are seemingly stronger than could be accounted for solely by direct synchronising mechanisms. Indeed, even seemingly unrelated species can exhibit striking correlations. This has led scientists to investigate potential indirect coupling mechanisms. Multiple molecular species in cells often share scarce resources, for instance some substrates are involved in chemical reactions catalysed by the same enzymes. It is believed that an indirect coupling between species is induced by sharing this common cellular machinery. This is sometimes referred to as cross talk or correlation resonance. Constructing mathematical models that exhibit this behaviour is the subject of a series of papers [27, 72, 73, 74]. These all rely on the theory of multiclass queueing networks. A multiclass queue is simply one in which customers belong to one of a number of distinguishable classes, each (possibly) with their own class-dependent arrival rate, service time distribution, routing rules and so on. It is interesting to note that under certain assumptions the stationary distribution of the network is of product form. See chapter 3 of [59] for more details.

The aim of [74] is to model a phenomenon called translational cross talk. This involves RNA molecules migrating to organelles called ribosomes, whereupon they bind and catalyse protein production. The process of translating proteins does not consume the RNA molecules and ribosomes, so they are free to rebind and catalyse multiple reactions. One hypothesis is that correlations in certain protein concentrations are the result of different RNA molecules competing for a scarce pool of ribosomes. The authors analyse a number of variants of stochastic models based on this description with multiclass queueing methods. The time evolution of a multiclass queue tracks the abundances of different molecular species. They compute steady state moments and covariances of two distinct protein species (translated by two distinct RNA species). Their conclusion is that when mutual translational processing resources are scarce, strong anticorrelations in protein counts can be induced if RNA molecules of one type keep rebinding to the same ribosomes.

A different indirect coupling mechanism is the focus of [73]. They consider two species of proteins flipping between their phosphorylated and unphosphorylated states due to the action of a common pool of enzymes. Phosphorylation refers to the joining of a phosphate group to the protein. New unphosphorylated proteins of both types are continually being translated, while existing proteins (regardless of their state) are degrading over time. The authors model this with the following multiclass queueing network. There are two queueing facilities keeping track of the number of phosphorylated and unphosphorylated proteins respectively. Each queue has two customer classes representing the two distinct protein species. Both have a fixed number of servers representing the number of phosphorylating and dephosphorylating enzymes present. New unphosphorylated proteins of each species arrive exogenously according to independent Poisson streams. Service corresponds to the phosphorylation or dephosphorylation process, so customers are simply routed around in a cycle. All proteins degrade after some time. This is captured by customers reneging from both queues and thereby exiting the system. By reneging we mean that customers depart the queue while waiting for or receiving service. By studying this multiclass queueing network the authors find that when the queues are critically loaded there is a spike in correlations between the two protein species. This corresponds to competition for the shared enzymatic processing resources.

## 3.4 Infinite Server Queues in Random Environment

A queue in a random environment is one in which the arrival and or service rate is modulated by some background process (often called the environment). For instance the arrival process may randomly flip between a low activity regime and intermittently exhibit short periods of high level activity. So this model provides us with a mechanism to capture bursty behaviour and means that the arrival and service rates are themselves stochastic processes. An early study of infinite server queues in random environment is conducted in [77]. The authors study the $M/M/\infty$ queue in a Markovian environment. By a Markovian environment, we mean that the arrival and service rates are given by the state of a finite state, irreducible, continuous time Markov chain. Sometimes this is called a Markov-modulated infinite server queue as the environment process is a Markov chain. Their contribution is to show that a necessary and sufficient condition for the queue to have a unique stationary distribution is that at least one of the possible service rates is strictly positive. They further derive the factorial moments of the queue length at stationarity and systems of partial differential equations for the transient moments.

Markov-modulated infinite server systems have been extensively studied recently (see [20] for a collation of the results summarised concisely into a diagram). For a queue in a Markovian environment there are two different ways in which the service rates can be modulated by the background Markov chain. Either the service rate of a job changes as the environment process jumps [16, 18, 19, 20], or alternatively a customer has a service rate (which stays fixed throughout its sojourn in the system) given by the state of the modulating Markov chain at the instant of its arrival [14, 15, 17, 19, 20]. These mechanisms are referred to in [20] as model 1 and model 2 respectively. The trivial case in which there is only one possible service rate (where models 1 and 2 coincide) is called model 0 [1]. Early work on these systems focussed on the steady state behaviour and transient dynamics of the queue length and the moments in both cases are derived [17, 77]. An extension to the $M/M/\infty$ queue in semi-Markovian random environment is made in [31], in which the authors compute all factorial moments for the steady state queue length. A semi-Markov process is one whose holding times in each state are not Exponentially distributed, but which is nevertheless Markov at its jump times.

More recently the focus has shifted to the asymptotic behaviour of these models. Large deviations results are proven [14, 15, 18, 51, 56], as are Central Limit Theorem type results for the convergence of the centered, normalised queue length at a fixed time $t$ to a Gaussian random variable [16, 17, 19]. Furthermore, Functional Central

Limit Theorem results are shown for the convergence of the queue length process to an Ornstein-Uhlenbeck process [1, 20, 51]. This matches what was known for the standard $M/M/\infty$ queue, where Ornstein-Uhlenbeck processes also arise as heavy traffic diffusion approximations for the queue length process. In the asymptotic regime in which arrival rates are sped up by a factor of $N$ while the Markov transition rates are scaled by a factor of $N^\alpha$ with $\alpha \in \mathbb{R}_+$, a dichotomy of behaviour is observed resulting in different scaling regimes. When the background process switches rapidly ($\alpha > 1$), the queue behaves like a homogeneous $M/M/\infty$ queue which sees only the average state of the background process, whereas for $\alpha < 1$ the behaviour depends upon the deviation matrix of the Markov chain [26]. When $\alpha = 1$ one gets an effective superposition of these two behaviours. In the fast changing environment regime the steady state number of jobs follows a Poisson distribution, just like the ordinary $M/G/\infty$ queue [17, 52]. A notable absentee from this plethora of results is the establishment of sample path large deviations for this model.

We elucidate in more detail the contributions of these papers to limit theorems and large deviations for Markov-modulated infinite server queues. In [18] logarithmic asymptotics for tail probabilities of both the transient and stationary queue length are derived in the setting of model 1 in both the slow and fast background regimes. Meanwhile, [14, 15] derive analogous results but in the setting of model 2. The latter deals with the fast environment case and the former with the slow switching regime. The Markovian assumption on the environment is relaxed in [56], where the background process for an $M/M/\infty$ queue is just a general càdlàg stochastic process which modulates both the arrival and service processes. A full LDP is then proven for the transient queue length for a fixed time in a setting slightly more general than both models 1 and 2.

The Markov-modulated model 2 $M/G/\infty$ queue is considered in [17] and explicit expressions for the transient mean and variance of the number of jobs in the system at a fixed time $t$ are derived under the condition that the queue started empty. In the special case of Exponential service times, a differential equation for the moment generating function of the number of customers in the system is found, which allows for the computation of moments at an Exponentially distributed time and at steady state. Finally, a CLT for the finite-dimensional marginals of the transient queue length process is shown in the fast background regime. These results are complemented by [16] which studies the Markov-modulated model 1 $M/M/\infty$ queue. CLTs for the transient and stationary scaled queue lengths are proved in both the fast and slow regimes. This is achieved by first finding differential equations for the probability generating functions of the stationary and transient queue length, then establishing the appropriate centering by showing weak laws of large numbers for each quantity and finally scaling appropriately. Interestingly, in the slow switching regime the CLT scaling is nonstandard, exhibiting smaller typical fluctuations than usual. The Central Limit Theorem picture in the case of Markovian service is completed by [19] which establishes a unified approach to tackle models 1 and 2 in the fast and slow regimes as opposed to the ad hoc approaches of [16] and [17].

A Functional Central Limit Theorem is established for the centered, normalised queue length process of a Markov-modulated model 0 $M/M/\infty$ queue in [1] in both the fast and slow regimes. In the fast regime one gets the usual square root CLT scaling, but in the slow regime the normalising polynomial is of higher degree. In the slow case, the behaviour of the limiting Ornstein-Uhlenbeck process depends upon the deviation matrix of the background Markov chain. These results are extended in [20] to models 1 and 2. The proofs in both papers rely heavily on martingale methods. In [51] the background process is a special type of stationary Cox process. A functional CLT for the scaled queue length is proved and logarithmic asymptotics of tail probabilities for the queue length distribution are investigated. A trichotomy in behaviour is observed depending on how rapidly the environment changes relative to the arrival rate. If the arrival rate changes faster, then one essentially recovers the familiar $M/M/\infty$ queue. If, however, it is slower, then the system exhibits overdispersion - fluctuating more wildly than Poissonian behaviour. More specifically the variance to mean ratio is greater than unity (which one obtains in the Poisson case), so this is essentially a measure of the noise to signal ratio. When the same scaling is applied to both, one obtains an effective superposition of the previous behaviours. There are effectively only two different causes for the number in the system to become unusually high. Either the background process takes on an unexpectedly high value and we see a typical number of arrivals given this state of the environment, or alternatively, although the background process is not itself especially large, we nevertheless see a rarely high number of arrivals into the queue. The scaling parameter of the background process dictates which of these paths to overflow dominates. The slow regime favours the former and the fast regime the latter behaviour.

The model we study is also an infinite server queue in a random environment. However, unlike the work described in this section that views the queue length as a càdlàg stochastic process, in chapter 5 we view it as living on a space of measures. Additionally, our model does not take the service process to be modulated.

## 3.5 Large Deviations for Random Point Measures

A functional LDP is proved in [69] for rescaled Poisson random measures using the Cramér's Theorem approach and subadditivity arguments. Essentially the same result was proven contemporarily in [42], but by a different method. The main goal of [42] is the following statistical application. Suppose one observes a spatial Poisson point process over a compact time interval, then just from the realisation of the Poisson random measure, one would like to obtain a parametric estimate of its intensity measure. The main contribution is to prove an LDP for the maximum likelihood estimator and explicitly identify the corresponding rate function. On the way to achieving this goal the authors prove an LDP for rescaled Poisson random measures. In more detail, let $E$ be a Polish space and let $N$ be a Poisson random measure on $\mathbb{R}_+ \times E$. The intensity measure of $N$ is denoted $q = \mathbb{E}(N)$ and is given by $q(dt, dx) = \nu(dx)dt$, where $\nu$ is a $\sigma$-finite positive measure on $E$ with its Borel $\sigma$-field $\mathcal{B}_E$. Fix a $0 < T < \infty$ and define the random measure $n_T$ on $E$ by

$$n_T(\Gamma) = \frac{1}{T} N([0,T] \times \Gamma) \quad \text{for any } \Gamma \in \mathcal{B}_E.$$

The authors then prove an LDP on $(E, \mathcal{B}_E)$ for $n_T$ as $T \to \infty$. The proof is based upon the Laplace-Varadhan Principle (see [101]) and makes use of the Dawson-Gärtner projective limit approach (for details see [33]). Some key steps are to first find the limiting cumulant generating function of the empirical measure $n_T$. Then find its Fenchel-Legendre transform, in other words the Gärtner-Ellis rate function. This is obtained as the solution to a certain variational problem. They then solve the optimisation problem to get the rate function explicitly. Finally, they prove the LDP by using the Dawson-Gärtner Projective Limit Theorem and the abstract Gärtner-Ellis approach and show that the rate function is both good and convex. The setting of these papers is similar to our own, but we work with the empirical measure of a spatial Cox point process.

The contribution of [91] is to prove a functional LDP for the empirical measure generated by a Cox process on a Polish space in the vague topology. The authors make the natural assumption that the stochastic intensity measure itself obeys an LDP with a good rate function. It is also assumed to be locally finite and satisfy a further technical condition - that the sequence of intensity measures dominate a fixed measure with full support on the Polish space. Our work in section 5.2 is similar in nature to this, but replaces the latter two assumptions with the condition that the intensity measure is a finite measure. Our proof also requires that the underlying Polish space is $\sigma$-compact. Additionally, we show the LDP with respect to a finer topology (this extension turns out to be nontrivial) and with a different method of proof. Just as is the case for us, the rate function is only given implicitly as the solution of an optimisation problem.

However in the special case of a mixed Poisson process an explicit solution to the variational problem is given. The proof uses Sanov's Theorem, the Contraction Principle and a lot of technical estimates.

## 3.6 Empirical Processes

For a thorough account of this subject see the books [95] and [99] and the many references therein. We present just a brief overview of relevant parts of [99]. Much of this section is therefore inspired by their treatment of empirical processes (see in particular chapter 2 and section 3.5).

Suppose $X_1, ..., X_n$ are a collection of random elements of a measurable space $(\mathcal{X}, \mathcal{A})$ with some common law $P$, then their empirical measure is given by

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}.$$

Imagine, for instance, that the $X_i$ give the location of points of a point process on Euclidean space. The empirical measure is a random measure (as the $X_i$ are random), that when given a measurable set, returns the proportion of points falling into that set. The empirical process perspective essentially boils down to viewing such random measures instead as random functionals over some class of functions. By random functional we mean integrate a function from the set of test functions against the measure - we use the notation $Pf := \int f dP$ for a measure $P$ and measurable function $f$. Suppose $\mathcal{F}$ is such a set of measurable test functions $f : \mathcal{X} \to \mathbb{R}$, then

$$\mathbb{P}_n f = \int f d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i),$$

a real-valued random variable. Testing against every function in the class yields an $\mathcal{F}$-indexed process $\{\mathbb{P}_n f : f \in \mathcal{F}\}$. The book [99] is principally interested in establishing the average behaviour and fluctuations about the average of the empirical process. To capture central limit type behaviour one naturally considers the centered rescaled version of the empirical measure to obtain the random signed measure

$$\sqrt{n}(\mathbb{P}_n - P) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\delta_{X_i} - P),$$

for which the corresponding process is denoted $\{\mathbb{G}_n f : f \in \mathcal{F}\}$, so that

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[f(X_i) - Pf].$$

Fixing a given $f$, the strong Law of Large Numbers guarantees $\mathbb{P}_n f \to Pf$ almost surely, provided we have the usual first moment condition that $Pf < \infty$. If we also have the second moment condition $Pf^2 < \infty$, then the Central Limit Theorem dictates that $\mathbb{G}_n f \Rightarrow Z \sim \mathcal{N}(0, P(f - Pf)^2)$.

For a fixed $f$, $\mathbb{P}_n f$ and $\mathbb{G}_n f$ are of course real valued random variables. But if one views them as processes indexed by a large class of functions $\mathcal{F}$, one might ask if the processes satisfy a corresponding LLN and CLT respectively, uniformly over $\mathcal{F}$. For conciseness let $\|Q\|_{\mathcal{F}} := \sup\{|Qf| : f \in \mathcal{F}\}$, where $Q$ is a signed measure. Then the uniform LLN, known as the Glivenko-Cantelli Theorem, states $\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$ outer almost surely (for a discussion of the mode of convergence see [99] sections 1.2 and 1.9). We call a class $\mathcal{F}$, for which this convergence holds, a Glivenko-Cantelli class or $P$-Glivenko-Cantelli - where we have emphasized the dependence on the underlying law $P$. For example, for the classical empirical process $\mathcal{X} = \mathbb{R}$ and $\mathcal{F}$ is the set of indicator functions of left half lines in $\mathbb{R}$.

A uniform version of the CLT requires that for every $x$

$$\sup_{f \in \mathcal{F}} |f(x) - Pf| < \infty.$$

This allows one to regard the empirical process $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ as a map into $\ell^\infty(\mathcal{F})$, the space of uniformly bounded functions from $\mathcal{F}$ to $\mathbb{R}$. So one might then ask when it is true that $\mathbb{G}_n \Rightarrow \mathbb{G}$, where the limit process $\mathbb{G}$ is a Borel measurable element of $\ell^\infty(\mathcal{F})$. Note we have been deliberately vague about the mode of convergence, but for details we direct the reader to section 1.3 of [99]. We call a class $\mathcal{F}$, for which this convergence holds, a Donsker class or $P$-Donsker. By considering the marginals and the associated covariance structure, the limit process can be determined to be the $P$-Brownian Bridge (see Appendix A of [99] for details). Unsurprisingly it turns out that every Donsker class of functions is necessarily Glivenko-Cantelli almost surely.

A natural question to pose at this point is what determines whether a given collection of functions is a Glivenko-Cantelli or Donsker class. Loosely speaking it is the 'size' of the class that matters. A notion of size of a class of functions is given by so called entropy numbers. The $\varepsilon$-entropy is the logarithm of the minimal number of balls (in the function space) of radius $\varepsilon$ that are needed to cover $\mathcal{F}$. So it is intuitively clear that as the radii get smaller, the entropy will increase. It is the rate of entropy increase in the small ball limit that provides sufficient conditions for a class to be Glivenko-Cantelli or Donsker (see chapters 2.4 and 2.5 of [99] for details). Once one has established certain classes of functions are P-Glivenko-Cantelli or P-Donsker, then one can construct other classes from them with operations on these sets that preserve these properties. These closure properties are known as permanence properties.

So far we have only considered empirical measures of a sample of fixed size $n$, but sometimes we are interested in the case of a random sample size. The key results of section 3.5 of [99] show that CLT type results carry over to the case of random sums under quite general circumstances. Specifically, if we have a sequence of random sample sizes $N_n$, such that $N_n/c_n$ converges in probability to an almost surely positive limit $\nu$, where the deterministic sequence $c_n \to \infty$, then $\mathbb{G}_{N_n} \Rightarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$. In the special case that $N_n$ are Poisson random variables, we call the corresponding empirical process the Poissonized empirical process, and if $N_n \sim Poisson(n)$ then we call it the Kac empirical point process. In this case $\mathbb{N}_n := \sum_{i=1}^{N_n} \delta_{X_i}$ is a Poisson point process with intensity measure $nP$ and provided $\|P\|_{\mathcal{F}} < \infty$ the Kac process $\mathbb{Z}_n := n^{-\frac{1}{2}}(\mathbb{N}_n - nP)$ converges weakly in $\ell^\infty(\mathcal{F})$ to a Brownian motion if and only if $\mathcal{F}$ is $P$-Donsker (see Theorem 3.5.5 of [99]).

Necessary and sufficient conditions for the empirical process to satisfy an LDP and MDP in the Banach space $\ell^\infty(\mathcal{F})$ are given in [105]. These results are extended to the setting of Poissonized empirical processes in [107].

## 3.7 Central Limit Theorems and Large Deviations for Infinite Server Queues

Providing an extensive review of the work in this area is beyond the scope of this thesis due to the vast number of papers. Instead, we must content ourselves with a description of the work most similar in spirit to our own. In section 3.4 we already discussed (amongst other things) limit theorems and large deviations results for infinite server queues in random environment. These references are of course relevant to this section too, but we do not repeat the discussion here.

There is a vast literature on limit theorems for infinite server queueing models, starting with the pioneering work of [53] on an FCLT for the $M/M/\infty$ queue. Section 10.3 of [104] and its many references provide a survey of key contributions in this area. If one considers a queue length jump process of an infinite server queue and rescales space and time appropriately, then under quite general circumstances the queue length process converges to an Ornstein Uhlenbeck process reflected at zero (as the queue length cannot go negative). The name limit theorem in this context comes from considering a sequence of queueing facilities, indexed by $n$, say, where the arrival rate into each grows with $n$ and the service time distribution remains fixed and then rescaling appropriately.

A common approach to obtaining a Functional Central Limit Theorem is the continuous mapping approach. We learn in a first course in analysis that if a sequence of numbers converges to a limit, then this limit is preserved under a continuous mapping. The Continuous Mapping Theorem asserts that the same is true of random sequences with an appropriate stochastic mode of convergence. There are a number of generalisations that basically allow one to get away with a mapping that is 'almost' continuous. The idea underlining the continuous mapping approach essentially entails combining Donsker's FCLT (see, for instance Theorem 4.3.2 of [104]) with the Continuous Mapping Theorem to establish new FCLTs - viewing queue length processes as random elements of $D$, the space of càdlàg functions (see section 3.4 of [104]). The usual approach is to write the queue length process as a reflection of a net input process to the queue and then show that the reflection mapping is continuous in the appropriate topology.

An alternative approach to establishing stochastic process limits for queueing (and other) models is the compactness approach (see [12] for a full account). This approach usually consists of showing convergence of the relevant finite dimensional distributions and that the sequence of probability measures is tight. Prokhorov's Theorem (Theorem 11.6.1 of [104]) is key, showing that there is no loss in focussing on tightness rather than relative compactness directly. Analogues of the Arzela Ascoli Theorem give a characterisation of compact sets of $D$, which in turn yield tightness criteria - see section 11.6 of [104] for a very abridged version of [12]. A further procedure that has been used to prove limit theorems for queueing models is the martingale method. A thorough review and survey is given by [78].

The model in [64] is somewhat similar to our own. The authors consider an open network of $\bullet/G/\infty$ queues whose exogenous arrival processes are a special type of Cox process called a shot noise process. There is some baseline arrival rate, but at random times (at the increments of a homogeneous Poisson process, independent of everything else) there are sudden spikes in the arrival rate - called shots. Following a shot, the arrival intensity gradually reverts to its original baseline. In this paper they assume the shot noise decays exponentially fast to the baseline. This makes the shot noise process particularly tractable as it is then a Markov process. Like the Markov modulated infinite server queue, this model captures burstiness in arrivals and is controlled by a background exogenous process. Such a shot noise process is used as a model for the arrival of insurance claims [30], where an individual catastrophe (like a natural disaster) may generate a cluster of incoming claims. They assume that service times are i.i.d. and generally distributed and that the network has no cycles (all traffic feeds forward, but not necessarily in a traditional linear tandem). They allow for some dependence structure between shot noise processes at different facilities. For instance in the case of a major incident in a city, one might expect the police and fire departments to receive a large number of emergency calls at roughly the same time. Their main contributions include calculating the joint Laplace transform of the queue lengths and shot noise processes and proving a Functional Central Limit Theorem for the rescaled number of jobs in an individual queue in the asymptotic regime in which the shot intensity is sped up linearly. The limiting process turns out to be a special type of Ornstein Uhlenbeck process and the proof combines the martingale FCLT with the continuous mapping approach. A major difference between this work and ours is that they work in the space $D$, whereas we work with the empirical measure.

Another work in a similar vein is [45]. The authors consider a $\bullet/G/\infty$ queue driven by a Hawkes arrival process. A Hawkes process is a type of self-exciting process. There is some baseline rate at which increments occur, but each arrival event triggers a spike in the arrival rate whose effect gradually dissipates. So this is much like the shot noise process, except the shots now are induced by arrivals themselves rather than being the result of an exogenous process. The bursty and self-exciting behaviour makes Hawkes processes a realistic model for (amongst other things) earthquakes - where shocks can trigger aftershocks, or gang violence - where one crime is met by an exchange of retaliatory crimes. Their key contributions are to prove an FCLT for stationary Hawkes processes in the large baseline intensity limit and then use this to prove an FCLT for an infinite server queue fed by accelerated Hawkes traffic. The convergence is of a sequence of scaled random elements of $D$ with respect to Skorohod's $J_1$ topology. The proof follows the compactness approach. For results of a different nature on infinite server queues with Hawkes traffic see [32] and [65].

Heavy traffic limits for the $G/GI/\infty$ queue are established in [90]. By heavy traffic, we mean that the arrival rate is sent to infinity while the service time distribution is not scaled. Consequently the number of busy servers tends to infinity. The processes and their limits live on the space of tempered distributions - the dual of the Schwartz space (see [57] for further details). Specifically two processes are considered, the age process and the residual service time process. The former records the age of each job in the queue, while the latter records their remaining service times and the time elapsed since former customers departed. The authors prove fluid (FLLN) and diffusion (FCLT) limits for rescaled versions of these processes. The proofs use a version of the martingale FCLT for tempered distribution valued processes and then use the continuous mapping approach upon showing the continuity of a certain regulator map. The diffusion limits are tempered distribution valued Ornstein Uhlenbeck processes. The work in this paper builds upon results in [36], which analyses the residual service time process in the same setting of tempered distributions of the $M/GI/\infty$ queue. Again a FLLN and

FCLT are proved for the rescaled residual service time process, using a combination of compactness and continuous mapping arguments. The age process is not dealt with.

There are a series of papers on various infinite server queueing models that also prove fluid and diffusion limits for appropriately rescaled age and residual service time processes living on a two-parameter function space - see [79, 80, 81, 82] and references therein. This is always done in the asymptotic regime in which arrivals are sped up, but service times remain unscaled. There is a progression of very similar results for the $G_t/GI/\infty$ queue (with a general time varying arrival process) [79], the $G_t/G_t/\infty$ queue (with general time varying service times that depend upon the arrival times, but are conditionally independent given these arrival times) [81], and the $G_t/G^D/\infty$ queue (where the service times are weakly dependent) [80, 82].

The literature on large deviations for queueing models is also sizeable, though not nearly as extensive. The book [44] and the many references therein gives a fairly broad survey. An early work in applying large deviations theory to infinite server queueing systems is [48]. This establishes approximations of tail probabilities for the queue length of a $GI/G/\infty$ queue at a fixed time on the large deviations scale. The author considers a sequence of systems in which arrivals are sped up (either by having a single fast source or a large number of sources operating at moderate speed) but the service times are not scaled.

Sample path large and moderate deviations behaviour of a tandem of $M/G/\infty$ queues with nonhomogeneous arrivals is the subject of [107]. The main results are full Large Deviations and Moderate Deviations Principles for the joint distribution of all of the queue length processes and departure process for the network viewed as random càdlàg functions. These results allow for the study of the most likely path to extreme queue lengths and abnormally large numbers of departures. Key to the main results is expressing the queue length and departure processes as Poisson random measures, which in turn are viewed as Poissonized empirical measures. The author proves a result concerning LDPs and MDPs for Poissonized empirical processes and this is applied to obtain the main results for the queueing model. Said result is actually more general than what is needed for the queueing applications but there is no extra difficulty in proving it in the more general setting.

In some ways the complementary large deviations counterpart to the diffusion limit results of [79] and [90] is [13]. It is not a direct large deviations analogue but there are many similarities. The authors consider sample path large deviations for a two-parameter process which is akin to the residual service time process. Specifically they let $Q_\lambda(t,y)$ represent the number of jobs in the system present at time $t$ with residual service requirement at least $y$ in an infinite server queue whose arrival rate is $\lambda$. The arrival process is general, assumed to be of nonlattice renewal type and service times are general following some continuous distribution. Their main contribution is to prove an LDP with a good rate function for $\frac{Q_\lambda(\cdot,\cdot)}{\lambda}$ with respect to the uniform topology on $[0,T] \times \mathbb{R}_+$ in the asymptotic regime in which arrivals are sped up but service times remain unscaled. As an application they find the most likely path to ruin for a portfolio of life insurance policies. As the two-parameter process is discrete valued, they consider a polygonalised version of the process (shown to be exponentially equivalent) which is more amenable to analysis. They first prove the LDP assuming the service time distribution has bounded support, but then show via a truncation argument that this result can be extended to the case of unbounded service times. The process that one obtains by discounting customers with service times larger than some threshold is shown to be an exponentially good approximation thanks to a simple Chernoff bound. The LDP for the truncated family is first proven with respect to the topology of pointwise convergence owing in part to an application of the Gärtner Ellis Theorem. Then it is lifted to the finer uniform topology by showing exponential tightness.

## 3.8 Central Limit Theorems and Large Deviations in Stochastic Geometry

Stochastic Geometry is a branch of probability that deals with random spatial patterns. An extensive account is given, for instance, by [98]. This allows one to study random phenomena that have a geometric structure. The most common and simplest models concern random point configurations in some (often Euclidean) space - see [29] for a thorough exposition. There are a multitude of applications but some examples include astronomy [71], where point patterns represent celestial bodies in space, ecology [97], where a point configuration might describe the spatial distribution of organisms of some species in some habitat, or wireless networks [49], where points in space correspond to the locations of transmitters and receivers. In our case the application in mind is queueing theory, where we have a point pattern representing the arrival times and service requirements of customers at a queueing facility.

Complicated stochastic geometric systems may not be amenable to exact analysis, however one may nevertheless be able to study their asymptotic behaviour. These asymptotic properties are made precise by limit theorems - essentially spatial analogues of those found in classical probability theory. A thorough survey of such results and relevant concepts and techniques is provided by [94]. The rest of this section is essentially a very abridged and incomplete version of this review. These proof techniques typically use tools from classical probability while leveraging the geometric structure in the problem.

A common setting involves a geometric functional on a locally finite point pattern on some space. To be more precise, let $\chi$ be a locally finite point configuration on $\mathbb{R}^d$ and $x \in \chi$ a point. Then let $\xi(\cdot,\cdot)$ be a real-valued geometric functional defined on all pairs $(x,\chi)$. For $x \notin \chi$ we slightly abuse notation and write $\xi(x,\chi) = \xi(x, \chi \bigcup \{x\})$. It is commonly assumed that $\xi$ is translation invariant, meaning that for $v \in \mathbb{R}^d$ we have that $\xi(x+v, \chi+v) = \xi(x,\chi)$. We define the empirical measure associated to $\xi$ and $\chi$ by

$$\mu^\xi(\chi) := \sum_{x \in \chi} \xi(x,\chi)\delta_x.$$

The point pattern $\chi$ is typically a realisation of some point process $\mathcal{P}$. If the underlying point process is clear then notation is often shortened from $\mu^\xi(\mathcal{P})$ to $\mu^\xi$. Denote the cube centered at the origin of volume $\lambda > 0$ by

$$Q_\lambda = \left[\frac{-\lambda^{\frac{1}{d}}}{2}, \frac{\lambda^{\frac{1}{d}}}{2}\right]^d.$$

Given an underlying point process $\mathcal{P}$, the rescaled empirical measure on the unit cube is given by

$$\mu^\xi_\lambda := \sum_{x \in \mathcal{P} \bigcap Q_\lambda} \xi(x, \mathcal{P} \bigcap Q_\lambda)\delta_{\lambda^{-1/d}x}.$$

The object of interest is usually the centered scaled empirical measure $\overline{\mu}^\xi_\lambda := \mu^\xi_\lambda - \mathbb{E}\left[\mu^\xi_\lambda\right]$ in the large $\lambda$ asymptotic regime.

One important concept in the study of limit theorems in this framework is that of stabilisation. Loosely speaking we say that a geometric functional is stabilising for some input point process if its value at a point is determined locally. In other words there is some finite (and possibly random) radius, outside of which the environment does not affect the value of the functional at that location. This is known as a radius of stabilisation. For instance, suppose we sample a point process on $\mathbb{R}^2$, and construct the Voronoi tesselation on this point set. Then the Voronoi cell that contains the origin is unaffected by changes to the point process that are sufficiently far away (there is a random radius depending on the original point process, outside of which no effect is felt). There are a number of different notions of stabilisation of differing strengths - for precise statements and a detailed comparison see [94]. The concept was first formulated in somewhat different language in [61] and [67] and first considered in its present form in [84], [85] and [86]. In the simplest case that the underlying point process is a spatial Poisson point process, a great deal is known about the behaviour of stabilising geometric functionals.

Provided a stabilising geometric functional satisfies sensible moment bounds, one obtains Law of Large Numbers results which describe the typical behaviour of the associated empirical measure (see [8, 84, 85, 86]). Once one derives variance asymptotics of the random measures [8, 89], it is possible to obtain Central Limit Theorem type results that make precise the convergence of the centered normalised empirical meaure to a Gaussian random field. See for instance [8, 84, 85, 88] and [89] and references therein. There are also Berry-Esseen type results which give the rate at which Gaussian approximation kicks in by bounding the maximal error of approximation [7]. Proofs of the CLT tend to use martingale techniques [87], comparison of cumulants [8] or Stein's method [4, 89]. As well as CLT scale fluctuations of the empirical measures, there are also results on the moderate [9] and large [92] deviations scales. Further there are so called de-Poissonisation techniques [84] that allow one to transfer many of these results to the case of non Poissonian input. This is important for the asymptotic theory of applications in stochastic geometry such as: Voronoi and Delaunay tessellations [4], germ grain models [76], random sequential packing [93], spatial birth and growth models [98], nearest neighbour graphs [102] and sphere of influence graphs [84].

# 4 Second Order Statistics and a Phase-Type Approximation

## 4.1 Motivation

The number of RNA and protein molecules is constantly fluctuating around their typical copy counts. Biological cells have evolved mechanisms to suppress this noise and to try to maintain cellular concentration levels within narrow bands. Most noise is detrimental; the underproduction of relevant proteins can constrain the functionality of the cell, while overproduction can use up scarce resources common to other reactions and costs unnecessary energy. In this chapter our aim is to understand how fast such noise is suppressed. If, for example, there is a fluctuation in the number of proteins away from the mean, how long does it take for the number to return to typical levels? One potential application of this is in synthetic biology, where one tries to engineer artificial biological structures. Understanding their noise suppression capabilities is an important part of making them efficient. It is hoped that insights into noise suppression in naturally occurring systems will help shed light on this and provide a benchmark for performance.

We shall derive closed form expressions for typical molecular counts in section 4.2. Then we study the fluctuations around them in sections 4.3 and 4.4. The speed at which these fluctuations dissipate will be quantified by calculating the autocovariance function of the stationary queue length for both RNA molecules and proteins. This in turn allows us to calculate the associated power spectral densities in section 4.6 which tell us about the level of fluctuations at different frequencies. To calculate these quantities explicitly in closed form requires us to relax the assumption of arbitrary service time distributions (or equivalently molecular lifetime distributions). Instead, we will assume the service time distribution to be of Phase-type (defined in section 2.2). These are analytically very tractable but can provide arbitrarily good approximations of any non-negative probability distribution. In [34], these were calculated in the simpler setting of an Exponential service time distribution (essentially the trivial Phase-type distribution). The work in this chapter is a generalisation of these results. We compare the theoretical answers of sections 4.3 and 4.4 with simulated sample paths of the model in section 4.5. Finally, we discuss the robustness of the Phase-type approximation scheme in section 4.7.

## 4.2 Statistical Properties of the Queueing Model

First we briefly recall the point process description of the queueing model. At a high level, a point process is just a mathematical description of a collection of points scattered (possibly) randomly on some underlying space. In general, a realisation of a point process can be thought of as either a random point set $\{x_1, x_2, \ldots, x_k\}$ (assuming the point process is simple - otherwise we may need a multiset), or as a random counting measure $\sum_{i=1}^{k} \delta_{x_i}$, where $\delta_x$ denotes a Dirac measure at $x$. We call the latter the empirical measure corresponding to the realisation of the point set. The arrival process into an $M/G/\infty$ queue with arrival rate $\lambda$ and job size distribution $F$, can be represented as an inhomogeneous Poisson process on $\mathbb{R} \times \mathbb{R}_+$ with intensity measure $\lambda \otimes F$. If a realisation of this Poisson point process has a point at $(t, x)$, it denotes that a customer arrives at time $t$ bringing a service requirement of $x$. The queue length at time $t$ is simply the total number of points of the Poisson process in the set

$$A_t := \{(s, x) : s \leq t, x > t - s\},$$

as a customer arriving at time $s$ will still be in the system at time $t$ if and only if its service requirement is greater than $t - s$. We follow the convention of defining the queue length process to be right continuous. Likewise, the queue length process during a time interval $[s, t]$ can be described in terms of the empirical measure of the above Poisson process on the wedge-shaped set

$$A_{[s,t]} = \bigcup_{u \in [s,t]} A_u;$$

see Figure 5 for a visualisation.

Since we let time run over the interval $(-\infty, \infty)$, then an infinite amount of time has already elapsed by the time we reach time zero (or indeed any other finite time), hence the system has settled down to stationarity, and therefore we can deduce that the equilibrium distribution for the number of customers in the queue is a Poisson random variable with mean given by:

$$\begin{aligned} \int_{t=0}^{\infty} \int_{x=t}^{\infty} \lambda dF(x) dt &= \int_{x=0}^{\infty} \left[ \int_{t=0}^{x} \lambda dt \right] dF(x) \qquad \text{(by Tonelli's Theorem)} \\ &= \int_{x=0}^{\infty} \lambda x dF(x) = \lambda E(X) = \frac{\lambda}{\mu} = \rho. \end{aligned}$$

So this matches the claimed invariant distribution from section 1.4.

So for a single $M/G/\infty$ queue, with customers arriving according to a homogeneous Poisson process of rate $\lambda_1$, and mean service time given by $1/\mu_1$, the stationary queue

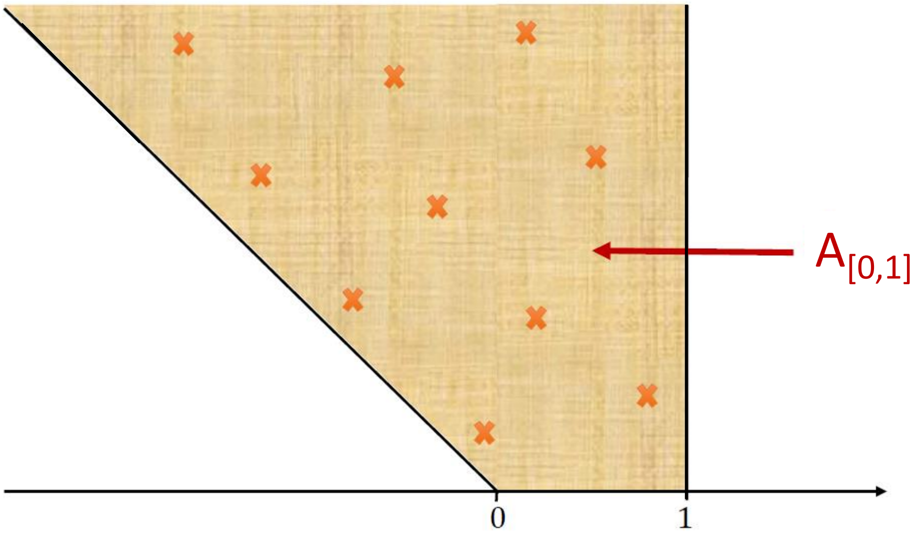

**Figure 5** – The set $A_{[0,1]}$ represents the region in time-service requirement space, which affects the queue length on the interval $[0,1]$.

length follows a $Poisson(\rho_1)$ distribution, where $\rho_1 = \lambda_1/\mu_1$. Therefore, if $(N_1(t), t \in \mathbb{R})$ represents the number of customers in the queue at time $t$, we have $\mathbb{E}[N_1(0)] = \rho_1$. Consider the tandem queueing model of section 1.5, and append some arbitrary number of $Cox/G/\infty$ facilities in the same feed-forward structure. Due to the linearity of expectation it is simple to find an exact analytical expression for the stationary mean number of customers in any queue in the series.

**Proposition 4.1.**
*Let $n \in \mathbb{N}$. Consider a tandem of $n$ $\bullet/G/\infty$ queues, whose queue length stochastic processes are denoted by $N_k(t)$ for $k \in \{1,...,n\}$ respectively. Let the first queue have Markovian arrivals at rate $\lambda_1 \in \mathbb{R}_+$ and every subsequent queue have a Cox arrival process with stochastic intensity $\Lambda_k(t) = \lambda_k N_{k-1}(t)$ for the $k^{th}$ facility, where $k \in \{2,...,n\}$ and $\lambda_k \in \mathbb{R}_+$. Let the $k^{th}$ facility have i.i.d. service times from an arbitrary distribution $F_k$ with mean $\frac{1}{\mu_k}$, where $\mu_k \in \mathbb{R}_+$ for $k \in \{1,...,n\}$. Then the steady-state expected queue length of the $\ell^{th}$ facility is given by $\mathbb{E}(N_\ell) = \frac{\lambda_\ell ... \lambda_1}{\mu_\ell ... \mu_1}$ for all $\ell \in \{1,...,n\}$.*

*Proof.* We show this by induction: For the base case $\ell = 1$, we have an $M/G/\infty$ queue whose invariant distribution is $Poisson(\rho_1)$, where $\rho_1 = \frac{\lambda_1}{\mu_1}$; and so $\mathbb{E}[N_1] = \frac{\lambda_1}{\mu_1}$. Suppose the claim is true for some $\ell = k$, i.e. assume $\mathbb{E}[N_k] = \frac{\lambda_k ... \lambda_1}{\mu_k ... \mu_1}$. Then consider the expected queue length of the $(k+1)^{th}$ queue:

$$\begin{aligned}
\mathbb{E}[N_{k+1}] &= \mathbb{E}[N_{k+1}(0)] \\
&= \mathbb{E}[\mathbb{E}(N_{k+1}(0)|N_k(t), t \in \mathbb{R})] \\
&= \mathbb{E}\left[\int_{t=0}^{\infty}\int_{x=t}^{\infty} \lambda_{k+1} N_k(t) dF_{k+1}(x) dt\right] \\
&= \lambda_{k+1}\mathbb{E}[N_k(t)] \int_{x=0}^{\infty}\int_{t=0}^{x} dt dF_{k+1}(x) && \text{(by Tonelli's Theorem)} \\
&= \lambda_{k+1}\mathbb{E}[N_k] \int_{x=0}^{\infty} x dF_{k+1}(x) \\
&= \lambda_{k+1} \cdot \frac{\lambda_k \dots \lambda_1}{\mu_k \dots \mu_1} \cdot \frac{1}{\mu_{k+1}} && \text{(by the inductive assumption)} \\
&= \frac{\lambda_{k+1} \dots \lambda_1}{\mu_{k+1} \dots \mu_1},
\end{aligned}$$

note, we have used the fact that all queues are in stationarity at time $t = 0$ as the system has been evolving since $t = -\infty$. So we have shown by induction that the claim is true. □

It was later pointed out that the result follows immediately from Little's Law (with a one line proof), without the need to consider point processes at all. We leave the original proof here as it illustrates how one can use the point process interpretation to say something about the queueing system. For a statement of Little's Law see Theorem 6.1 of [50].

One question of biological interest is to understand how quickly noise caused by fluctuations in molecular counts dissipates. To this end let us now find an expression for the stationary autocovariance function of the $M/G/\infty$ queue length process, $Cov[N(s), N(t)]$ (we drop the subscripts for convenience). To do so we define the complementary cumulative distribution function (henceforth referred to as the c.c.d.f. for short) as

$$\overline{F}(x) := 1 - F(x) = \int_{y=x}^{\infty} dF(y).$$

We can think about the autocovariance function pictorially; so consider two vertical half-lines, one at $t = 0$ and the other at $t = u$ where $u > 0$. Then drawing on the 45 degree diagonal half-lines from the base of each vertical line creates two conical regions. The intersection of these regions creates a third conical region, representing those customers that are present in the queue at both times 0 and $u$ (see Figure 6). Points in this region are the only ones which contribute to the covariance (by the independence of disjoint regions of the spatial Poisson point process). It is clear that the expectation of the random number of points in this region only depends upon the distance between 0 and $u$ and not on their position on the line. So when finding $Cov[N(s), N(t)]$ it suffices to consider $Cov[N(0), N(u)]$ and set $u = |t - s|$.

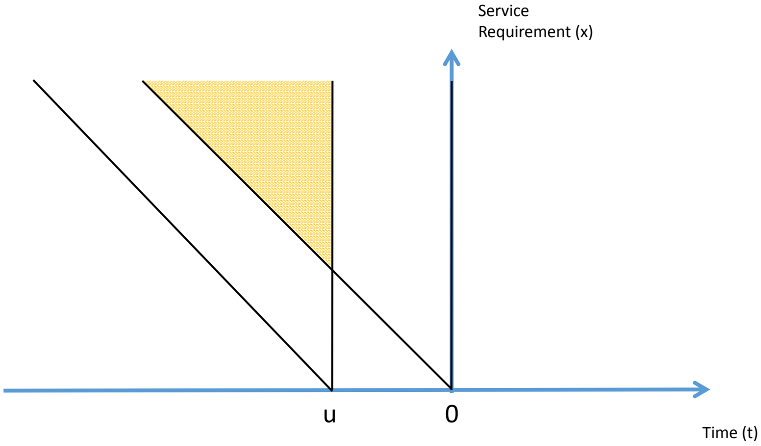

**Figure 6** – Common region contributing to autocovariance, representing customers that are present at times 0 and $u$.

Let $Z$ be the Poisson random variable for the number of points in the common region. Then:

$$\begin{aligned} Cov[N(0), N(u)] &= Var(Z) \\ &= \mathbb{E}(Z) \qquad\qquad \text{(as } Z \text{ is a Poisson random variable)} \\ &= \int_{t=u}^{\infty}\int_{x=t}^{\infty} \lambda dF(x)dt \\ &= \lambda \int_{t=u}^{\infty} \overline{F}(t)dt. \qquad (4) \end{aligned}$$

Any further simplification requires us to restrict ourselves to a specific service time distribution.

## 4.3 Second Order Statistics of the $M/PH_k/\infty$ Queue

In section 4.2 we began calculating the autocovariance function of the stationary $M/G/\infty$ queue length. We got as far as equation (4) saying

$$Cov[N(0), N(u)] = \lambda \int_{t=u}^{\infty} \overline{F}(t)dt. \qquad (5)$$

It is hard to proceed any further in complete generality, so let us now instead suppose that the service time distribution is of Phase-type (the reader is referred back to Definition 2.10 for details). This will allow us to write down the complementary cumulative distribution function (c.c.d.f.) and hence compute the integral. So we now consider an $M/PH_k/\infty$ queue (where $k \in \mathbb{N}$ is the number of phases). Note, if we had just one phase (the case $k = 1$), we would recover Exponential service times (and therefore have an $M/M/\infty$ queue). To proceed we need a functional form for the c.c.d.f. of a $PH_k$ distribution; so we first prove a lemma to express this in a convenient form.

We assume that the subgenerator matrix parameterising the Phase-type distribution is diagonalisable. A sufficient (but not necessary) condition for this is that all of its eigenvalues are unique. Biologically speaking this means that no two reaction rates in the degradation process are identical. A necessary and sufficient condition for an $n \times n$ matrix to be diagonalisable is that the sum of dimensions of the eigenspaces is $n$. Repeated eigenvalues are one possible obstruction to diagonalisability when this results in a lack of eigenvectors. Consider, for instance, the matrix

$$\begin{pmatrix} -2 & 1 \\ 0 & -2 \end{pmatrix},$$

which is a perfectly acceptable subgenerator matrix. This has $-2$ as a repeated eigenvalue with algebraic multiplicity two. Its corresponding eigenvector is $(1,0)^T$ and its geometric multiplicity is one. So it is not diagonalisable. So our assumption rules out genuine cases that can arise.

It is easy to see that the real part of the eigenvalues of the subgenerator matrix are negative. This is a consequence of the structure of the subgenerator matrix (diagonal entries negative, all other entries positive and row sums at most zero - indeed at least one row sum strictly negative) and Gershgorin's Theorem. Let $R_i$ be the sum of off diagonal elements in row $i$ of the subgenerator matrix. Then all eigenvalues lie in one of the closed discs of radius $R_i$ centered at the corresponding diagonal element.

**Lemma 4.2.**
*Let $k \in \mathbb{N}$. Let $X \sim PH_k(\boldsymbol{\alpha}, S)$, where $S$ is assumed to be diagonalisable and has eigenvalues $-\eta_i \in \mathbb{C}$, whose real parts are negative, and which occur in conjugate pairs, for $i = 1, 2, ..., k$. Then, there exist constants $c_i \in \mathbb{R}$, $i = 1, 2, ..., k$ (that can be calculated from the eigenvectors of the subgenerator matrix) such that*

$$\overline{F}_X(t) = \sum_{i=1}^{k} c_i e^{-\eta_i t}.$$

*Proof.* Let $\mathbf{1}$ denote the all one vector. It is known that for a Phase-type distribution:

$$\overline{F}_X(t) = \boldsymbol{\alpha}^T e^{St}\mathbf{1} = \boldsymbol{\alpha}^T \sum_{\ell=0}^{\infty} \frac{S^\ell t^\ell}{\ell!}\mathbf{1}.$$

Now let us diagonalise $S$ to obtain:

$$S = P^{-1}\Lambda P \quad \text{where} \quad \Lambda = \begin{pmatrix} -\eta_1 & & \\ & \ddots & \\ & & -\eta_k \end{pmatrix},$$

i.e. $\Lambda$ is the diagonal matrix of eigenvalues of $S$ and $P$ is the matrix whose columns are the corresponding eigenvectors. So we have that

$$\begin{aligned} \overline{F}_X(t) &= \boldsymbol{\alpha}^T \sum_{\ell=0}^{\infty} \frac{\left(P^{-1}\Lambda P\right)^{\ell} t^{\ell}}{\ell!} \mathbf{1} \\ &= \boldsymbol{\alpha}^T P^{-1} \sum_{\ell=0}^{\infty} \frac{\Lambda^{\ell} t^{\ell}}{\ell!} P\mathbf{1} \\ &= \boldsymbol{\alpha}^T P^{-1} \sum_{\ell=0}^{\infty} \begin{pmatrix} \frac{(-\eta_1 t)^{\ell}}{\ell!} & & \\ & \ddots & \\ & & \frac{(-\eta_k t)^{\ell}}{\ell!} \end{pmatrix} P\mathbf{1} \\ &= \boldsymbol{\alpha}^T P^{-1} \begin{pmatrix} e^{-\eta_1 t} & & \\ & \ddots & \\ & & e^{-\eta_k t} \end{pmatrix} P\mathbf{1}. \end{aligned}$$

But note that $\boldsymbol{\alpha}^T P^{-1}$ is just a row vector and $P\mathbf{1}$ is just a column vector. So we will label these vectors as follows with $\mathbf{a}^T := \boldsymbol{\alpha}^T P^{-1}$ and $\mathbf{b} := P\mathbf{1}$. But this yields

$$\begin{aligned} \overline{F}_X(t) &= \mathbf{a}^T \begin{pmatrix} e^{-\eta_1 t} & & \\ & \ddots & \\ & & e^{-\eta_k t} \end{pmatrix} \mathbf{b} \\ &= \sum_{i=1}^{k} a_i b_i e^{-\eta_i t} \\ &= \sum_{i=1}^{k} c_i e^{-\eta_i t}, \end{aligned}$$

letting $c_i = a_i b_i$, as required. Note any complex eigenvalues come in conjugate pairs, so the c.c.d.f. is indeed real valued. □

This allows us to find an analytically exact expression for the autocovariance function of the number of customers in an $M/PH_k/\infty$ queue in its stationary regime.

**Theorem 4.3.**
*Let $u > 0$ and $k \in \mathbb{N}$. Let $N(t)$ denote the number of customers in an $M/PH_k/\infty$ queue at time $t$, whose Phase-type service time distribution is assumed to have a diagonalisable subgenerator matrix $S$ with eigenvalues $-\eta_i \in \mathbb{C}$, whose real parts are negative, and which*

*occur in conjugate pairs, for $i = 1, ..., k$. Assume in addition that the c.c.d.f. of the Phase-type service time distribution is given by*

$$\overline{F}(t) = \sum_{i=1}^{k} c_i e^{-\eta_i t},$$

*where $c_i \in \mathbb{R}$ for $i = 1, ..., k$. Then the autocovariance function for the number of customers in the queue at stationarity is of the form*

$$Cov\left[N(0), N(u)\right] = \sum_{i=1}^{k} a_i e^{-\eta_i u},$$

*where $a_i \in \mathbb{R}$ for $i \in \{1, 2, ..., k\}$ are constants that can be calculated from the parameters of the arrival and service distributions.*

*Proof.* By equation (4) we have that

$$\begin{aligned} Cov\left[N(0), N(u)\right] &= \lambda \int_{t=u}^{\infty} \overline{F}(t) dt \\ &= \lambda \sum_{i=1}^{k} c_i \int_{t=u}^{\infty} e^{-\eta_i t} dt \\ &= \lambda \sum_{i=1}^{k} \frac{c_i}{\eta_i} e^{-\eta_i u} \\ &= \sum_{i=1}^{k} a_i e^{-\eta_i u}. && \text{(by setting } a_i := \tfrac{\lambda c_i}{\eta_i}\text{)} \end{aligned}$$

□

So now it follows immediately that:

$$Cov[N(t), N(s)] = \sum_{i=1}^{k} a_i e^{-\eta_i |s-t|}.$$

So in summary what we have found is that statistical fluctuations in molecular counts of RNA molecules dissipate exponentially fast. This happens according to a mixture of decaying exponentials whose rates are given by the eigenvalues of the subgenerator matrix parameterising the Phase-type distribution, while the coefficients $a_1, ..., a_k$ come from the corresponding eigenvectors. Eyeballing this expression it is clear that the smallest eigenvalue will predominantly determine the behaviour as the other terms are exponentially smaller.

## 4.4 Second Order Statistics of the $Cox/PH_k/\infty$ Queue

Now let us consider the statistical fluctuation properties of the number of proteins. To this end we find, in the following theorem, the autocovariance function for the number of customers in the $Cox/PH_k/\infty$ queue at stationarity. We shall assume that the stochastic intensity of the Cox process has autocovariance function of a specific form. The reason for this form is that it corresponds to the input process being given by the length of the RNA queue.

**Theorem 4.4.**
*Let $u > 0$ and $k, \ell \in \mathbb{N}$. Let $\kappa_1, ..., \kappa_\ell \in \mathbb{C}$, whose real parts are positive, and which occur in conjugate pairs, and $a_1, ..., a_\ell \in \mathbb{R}$. Consider a stationary Cox process, whose stochastic intensity $\Lambda(t)$ is a stationary and ergodic stochastic process with finite mean $\mathbb{E}[\Lambda(0)]$ and satisfies*

$$Cov\left[\Lambda(0), \Lambda(u)\right] = \sum_{i=1}^{\ell} a_i e^{-\kappa_i u}.$$

*Let $N(t)$ denote the number of customers in a $Cox/PH_k/\infty$ queue at time $t$, whose Cox arrival process has stochastic intensity $\Lambda(t)$ and whose Phase-type service time distribution is assumed to have a diagonalisable subgenerator matrix $S$ with eigenvalues $-\eta_i \in \mathbb{C}$, whose real parts are negative, and which occur in conjugate pairs, for $i = 1, ..., k$. Assume in addition that the c.c.d.f. of the Phase-type service time distribution is given by*

$$\overline{F}(t) = \sum_{i=1}^{k} c_i e^{-\eta_i t},$$

*where $c_i \in \mathbb{R}$ for $i = 1, ..., k$. Then the autocovariance function for the number of customers in the queue at stationarity is of the form*

$$Cov\left[N(0), N(u)\right] = \sum_{i=1}^{r} \gamma_i e^{-\theta_i u},$$

*where $r = k + \ell$, $\gamma_i \in \mathbb{R}$ and $\theta_i \in \mathbb{C}$, whose real parts are positive, and which occur in conjugate pairs, for $i \in \{1, 2, ..., r\}$ are constants that can be calculated from the parameters of the arrival and service distributions.*

*Proof.* First lighten notation by defining $\mathbf{\Lambda} = \{\Lambda(t), t \in \mathbb{R}\}$. We shall consider $N(\cdot)|\mathbf{\Lambda}$, the random variable $N(\cdot)$ representing the number of customers in the queue conditional on the entire sample path of the $\Lambda(\cdot)$ process on the interval $(-\infty, \infty)$. By the law of total covariance we have

$$Cov\left[N(0), N(u)\right] = \mathbb{E}\left\{Cov\left[N(0), N(u)|\mathbf{\Lambda}\right]\right\} + Cov\left\{\mathbb{E}[N(0)|\mathbf{\Lambda}], \mathbb{E}[N(u)|\mathbf{\Lambda}]\right\}.$$

For the first term we can consider the random number of points, say $Z$, in an appropriate overlapping region which accounts for all the conditional covariance (that is, one representing all those customers that are in the queue at both times 0 and $u$):

$$\begin{aligned}
\mathbb{E}\left\{Cov\left[N(0), N(u)|\mathbf{\Lambda}\right]\right\} &= \mathbb{E}[Var(Z|\mathbf{\Lambda})] \\
&= \mathbb{E}[\mathbb{E}(Z|\mathbf{\Lambda})] && \text{(as } Z \text{ is conditionally Poisson)} \\
&= \mathbb{E}\left[\int\limits_{t=u}^{\infty}\int\limits_{x=t}^{\infty} \Lambda(t) dF(x) dt\right] \\
&= \int\limits_{t=u}^{\infty} \mathbb{E}[\Lambda(t)]\overline{F}(t)dt && \text{(by Tonelli's Theorem)} \\
&= \mathbb{E}[\Lambda(0)]\int\limits_{t=u}^{\infty} \overline{F}(t)dt && (6) \\
&= \mathbb{E}[\Lambda(0)]\sum_{i=1}^{k} c_i \int\limits_{t=u}^{\infty} e^{-\eta_i t}dt && \text{(by Lemma 4.2)} \\
&= \mathbb{E}[\Lambda(0)]\sum_{i=1}^{k} \frac{c_i}{\eta_i}e^{-\eta_i u} \\
&= \sum_{i=1}^{k} d_i e^{-\eta_i u}. && \text{(where } d_i = \mathbb{E}[\Lambda(0)]\cdot \tfrac{c_i}{\eta_i}\text{)}
\end{aligned}$$

Consider now

$$\mathbb{E}[N(0)|\mathbf{\Lambda}] = \int\limits_{t=0}^{\infty}\int\limits_{x=t}^{\infty} \Lambda(t)dF(x)dt = \int\limits_{t=0}^{\infty} \Lambda(t)\overline{F}(t)dt, \tag{7}$$

and similarly

$$\mathbb{E}[N(u)|\mathbf{\Lambda}] = \int\limits_{t=u}^{\infty}\int\limits_{x=t-u}^{\infty} \Lambda(t)dF(x)dt = \int\limits_{t=u}^{\infty} \Lambda(t)\overline{F}(t-u)dt. \tag{8}$$

Then it follows that

$$\begin{aligned}
&Cov\left\{\mathbb{E}[N(0)|\mathbf{\Lambda}], \mathbb{E}[N(u)|\mathbf{\Lambda}]\right\} \\
&= \int\limits_{t=0}^{\infty}\int\limits_{s=u}^{\infty} Cov[\Lambda(t), \Lambda(s)]\overline{F}(t)\overline{F}(s-u)dsdt \qquad (9) \\
&= \int\limits_{t=0}^{\infty}\int\limits_{s=u}^{\infty} \sum_{i=1}^{\ell} a_i e^{-\kappa_i|s-t|}\sum_{j=1}^{k} c_j e^{-\eta_j t}\sum_{m=1}^{k} c_m e^{-\eta_m(s-u)}dsdt \\
&= \sum_{i=1}^{\ell}\sum_{j=1}^{k}\sum_{m=1}^{k} a_i c_j c_m e^{\eta_m u}\int\limits_{t=0}^{\infty}\int\limits_{s=u}^{\infty} e^{-\kappa_i|s-t|}e^{-\eta_j t}e^{-\eta_m s}dsdt.
\end{aligned}$$

Now considering the double integral separately we obtain:

$$\begin{aligned}
&\int_{t=0}^{\infty}\int_{s=u}^{\infty} e^{-\kappa_i|s-t|}e^{-\eta_j t}e^{-\eta_m s}dsdt \qquad\qquad \text{(now split to remove the absolute value)}\\
&= \int_{t=0}^{u}\int_{s=u}^{\infty} e^{-\kappa_i(s-t)}e^{-\eta_j t}e^{-\eta_m s}dsdt\\
&+ \int_{t=u}^{\infty}\left(\int_{s=u}^{t} e^{-\kappa_i(t-s)}e^{-\eta_j t}e^{-\eta_m s}ds + \int_{s=t}^{\infty} e^{-\kappa_i(s-t)}e^{-\eta_j t}e^{-\eta_m s}ds\right)dt\\
&= \int_{t=0}^{u} e^{t(\kappa_i-\eta_j)}\int_{s=u}^{\infty} e^{-s(\kappa_i+\eta_m)}dsdt\\
&+ \int_{t=u}^{\infty}\left(e^{-t(\kappa_i+\eta_j)}\int_{s=u}^{t} e^{s(\kappa_i-\eta_m)}ds + e^{t(\kappa_i-\eta_j)}\int_{s=t}^{\infty} e^{-s(\kappa_i+\eta_m)}ds\right)dt\\
&= e^{-u(\eta_j+\eta_m)}\left[\frac{1}{(\kappa_i+\eta_m)(\kappa_i-\eta_j)} + \frac{1}{(\kappa_i-\eta_m)(\eta_j+\eta_m)} - \frac{1}{(\kappa_i-\eta_m)(\kappa_i+\eta_j)}\right.\\
&\qquad\qquad \left. + \frac{1}{(\kappa_i+\eta_m)(\eta_j+\eta_m)}\right] - \frac{e^{-u(\kappa_i+\eta_m)}}{(\kappa_i+\eta_m)(\kappa_i-\eta_j)}\\
&= \beta_1 e^{-u(\eta_j+\eta_m)} + \beta_2 e^{-u(\kappa_i+\eta_m)},
\end{aligned}$$

where we have used $\beta_1$ and $\beta_2$ to simplify the presentation of the unwieldy coefficients in the penultimate line. Thus, we have that

$$\begin{aligned}
Cov\left\{\mathbb{E}[N(0)|\mathbf{\Lambda}], \mathbb{E}[N(u)|\mathbf{\Lambda}]\right\} &= \sum_{i=1}^{\ell}\sum_{j=1}^{k}\sum_{m=1}^{k} a_i c_j c_m e^{\eta_m u}\left[\beta_1 e^{-u(\eta_j+\eta_m)} + \beta_2 e^{-u(\kappa_i+\eta_m)}\right]\\
&= \sum_{i=1}^{\ell}\sum_{j=1}^{k}\sum_{m=1}^{k} a_i c_j c_m \left[\beta_1 e^{-u\eta_j} + \beta_2 e^{-u\kappa_i}\right].
\end{aligned}$$

And therefore

$$\begin{aligned}
Cov\left[N(0), N(u)\right] &= \sum_{i=1}^{k} d_i e^{-\eta_i u} + \sum_{i=1}^{\ell}\sum_{j=1}^{k}\sum_{m=1}^{k} a_i c_j c_m \left[\beta_1 e^{-u\eta_j} + \beta_2 e^{-u\kappa_i}\right]\\
&= \sum_{i=1}^{r} \gamma_i e^{-u\theta_i},
\end{aligned}$$

□

where $\gamma_i$ and $\theta_i$ are placeholders to make the expression less messy.

So what this result says is that if the Cox process has an autocovariance function which is a mixture of decaying exponentials, then so too is the autocovariance function of the stationary queue length. So essentially we have a closure property: if the input is a mixture of decaying exponentials then so too is the output. This also shows that the result propagates through any number of queues in this nonstandard tandem. So in biological terms, since RNA molecules suppress noise exponentially fast as a mixture of decaying exponentials, protein fluctuations too dissipate as a mixture of decaying exponentials. Moreover, since $r = k + \ell$, the expressions obtained for tandems only grow linearly in the number of stages. The coefficients are unwieldy, but very tractable numerically.

Instead of analysing the point process we anticipate one could perform such computations as in the previous proof by studying the joint Laplace transform and in particular differentiating it to find the moments. The computations would essentially be equivalent, though possibly slightly less transparent. The big advantage of this approach would be that in principle all moments could now be found.

## 4.5 Simulations

We now present some simulations that act as a sanity check to the above theory. We simulate the feed-forward tandem queueing model in two slightly different settings. In both cases we have an $M/PH_k/\infty$ queue whose occupancy modulates the arrival rate into a $Cox/PH_k/\infty$ facility. In one instance we give the queues non-identical Hyperexponential service time distributions (see Definition 2.8), and in the other, non-identical generalised Erlang distributions (see Definition 2.9).

We use the two simplest nontrivial Phase-type distributions which fit into the above framework of a diagonalisable subgenerator matrix parameterising the Phase-type distribution - namely $H_2$ and $genE_2$ service time distributions. It is vital that all of the parameters of the component Exponential distributions are distinct, otherwise diagonalisability is lost. Using exactly the same methods as in Theorem 4.4, one can explicitly calculate the autocovariance function of the equilibrium queue length of a $Cox/H_2/\infty$ and $Cox/genE_2/\infty$ queue. It is no more complicated to add more phases in each of these scenarios, the calculations simply become more arduous. In fact even in the non-diagonalisable case, such as an $E_2$ service time distribution, explicit calculations can still be performed. These lead to closed form expressions for the autocovariance function which are not simply linear combinations of decaying exponentials of the lag, but have polynomials of the lag as prefactors too.

We simulate a sample path of the stationary $Cox/H_2/\infty$ queue length process whose arrival rate is modulated by a stationary $M/H_2/\infty$ queue length, plot the autocorrelation function of the simulated process and overlay the theoretical autocorrelation function predicted by Theorem 4.4. We then do the same with $genE_2$ service time distributions. The simulations were performed using the statistical software R. The results are shown in Figures 7 and 8 and the code is contained in the Appendix.

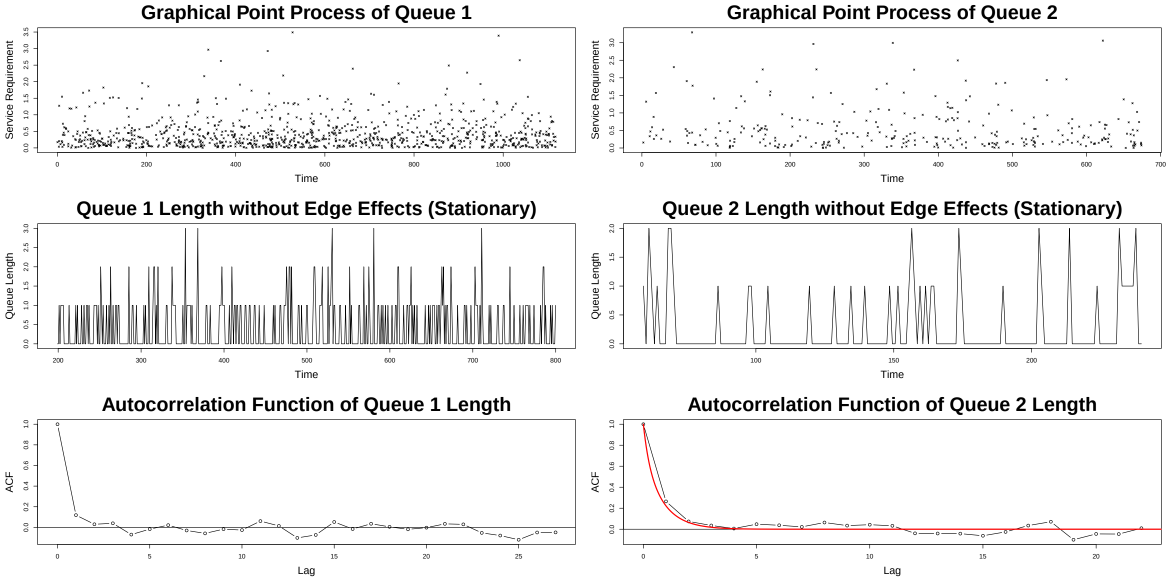

**Figure 7** – $M/H_2/\infty$ queue modulating $Cox/H_2/\infty$ queue. For both facilities we plot the point process representing customer arrival times and corresponding service requirements, the equilibrium queue length process, and the autocorrelation function of the queue length process. We overlay the theoretical autocorrelation function in red.

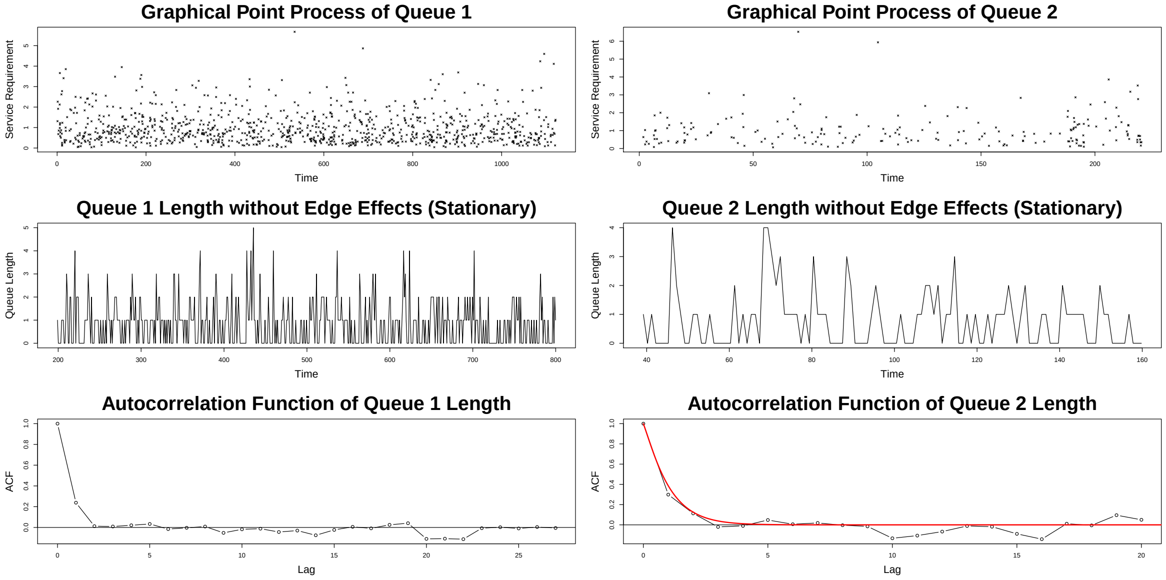

**Figure 8** – $M/genE_2/\infty$ queue modulating $Cox/genE_2/\infty$ queue. For both facilities we plot the point process representing customer arrival times and corresponding service requirements, the equilibrium queue length process, and the autocorrelation function of the queue length process. We overlay the theoretical autocorrelation function in red.

The simulations match the theory well. Since the total number of customers that ever enter the system is finite in the simulations, it is necessary to remove edge effects. We want to consider the queues in equilibrium, so we remove the initial and final part of the path. There is a burn-in period at the start as the queue starts empty, and a burn-out period at the end due to the number of arrivals being finite - meaning the queue eventually empties out. The correlograms start with a correlation of one at lag zero, as they should. The decay in autocorrelation as the lag increases matches what the theory tells us for the first few lags. The rest of the values are small and move around a bit either side of zero. This is because the queue lengths fluctuate around their averages with fairly small excursions away. This means the queue length processes only take on a few different values, so we should expect some small correlations due to chance. Increasing arrival rates and or decreasing service rates increases the number of customers in the system, but one still sees essentially the same picture with autocorrelations decaying exponentially fast. Again, only the first few lags are non-neglible and the later ones are slightly smaller than in the shorter queues as the outcomes are less random.

## 4.6 Power Spectral Densities for the $Cox/PH_k/\infty$ Queue

In time series analysis it is common to view a signal in the frequency domain as well as the time domain. The frequency domain representation offers an alternative (and often more easily interpretable) view of the signal. The Fourier transform is the tool that takes us between the time and frequency perspectives. We can further our analysis of the statistical properties of the queue length process by finding its spectral density. This will tell us about the behaviour of fluctuations at different frequencies.

**Corollary 4.5.**
*Let $N(t)$ be the queue length process in stationarity of a $Cox/PH_k/\infty$ queue satisfying the assumptions of Theorem 4.4. Then the power spectral density of $N(t)$ is given by*

$$f(\omega) = \frac{2}{\pi}\sum_{j=1}^{r}\frac{\theta_j\gamma_j}{\theta_j^2+\omega^2},$$

*where $\theta_i$ and $\gamma_i$, i=1,...,r are as in the statement of Theorem 4.4.*

*Proof.* A straightforward calculation yields

$$\begin{aligned}
f(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} Cov\left[N(0), N(\tau)\right] e^{-i\omega\tau} d\tau \\
&= \frac{1}{\pi}\sum_{j=1}^{r}\gamma_j \int_{-\infty}^{\infty} e^{-\theta_j|\tau|} e^{-i\omega\tau} d\tau \\
&= \frac{1}{\pi}\sum_{j=1}^{r}\gamma_j \left[\int_{-\infty}^{0} e^{\theta_j\tau} e^{-i\omega\tau} d\tau + \int_{0}^{\infty} e^{-\theta_j\tau} e^{-i\omega\tau} d\tau\right] \\
&= \frac{1}{\pi}\sum_{j=1}^{r}\gamma_j \left[\frac{1}{\theta_j - i\omega} + \frac{1}{\theta_j + i\omega}\right] \\
&= \frac{2}{\pi}\sum_{j=1}^{r}\frac{\theta_j\gamma_j}{\theta_j^2 + \omega^2}.
\end{aligned}$$

□

## 4.7 Robustness of the Phase-type Approximation

The use of Phase-type distributions for service times not only gives us tractable results, but also a hypothetical route to approximate arbitrary service time distributions. The reason is that Phase-type distributions are dense in the space of probability distributions on $\mathbb{R}_+$ equipped with the weak topology (see Theorem 2.11). However, this property is only useful if approximating the input to a queue yields approximations to the quantities of interest, such as means and covariances of queue lengths. In other words, we need these quantities to be continuous functions of the service time distribution. This is a question of robustness and is discussed further in chapter VIII.5 of [3]. We show now that this robustness does indeed hold in this case.

**Theorem 4.6.**

*Let $k \in \mathbb{N}$ and $A, L, M \in \mathbb{R}_+$. Consider a sequence of stationary Cox processes, indexed by $n$, whose stochastic intensities $\Lambda_n(t), n \in \mathbb{N}$ are stationary and ergodic stochastic processes with finite means $\mathbb{E}[\Lambda_n(0)]$ respectively. Let $N_n(t), n \in \mathbb{N}$ denote the number of customers in a sequence of stationary $Cox/PH_k/\infty$ queues at time $t$, whose Cox arrival processes have stochastic intensities $\Lambda_n(t)$ and whose Phase-type service time distributions $S_n$ have means $\mathbb{E}(S_n) < L$ which are uniformly bounded over $n \in \mathbb{N}_+$ and distribution functions $F_n$. Assume additionally that there exists an $\varepsilon > 0$ and $A < \infty$ such that $\mathbb{E}\left[S_n^{1+\varepsilon}\right] \leq A$ for all $n \in \mathbb{N}$. Let $N(t)$ denote the number of customers in a stationary $Cox/G/\infty$ queue at time $t$, whose Cox arrival process $\Lambda(t)$ is a stationary and ergodic stochastic process with finite mean $\mathbb{E}[\Lambda(0)]$, and whose service time* *distribution $S$ has finite mean $\mathbb{E}(S)$ and distribution function $F$. Assume*

$$\begin{aligned}
&\mathbb{E}[\Lambda_n(0)] \to \mathbb{E}[\Lambda(0)],\\
&\mathbb{E}[\Lambda_n(0)] < M \quad \textit{and} \quad Var[\Lambda_n(0)] < M \quad \textit{for all } n \in \mathbb{N}_+,\\
&Cov[\Lambda_n(0), \Lambda_n(t)] \to Cov[\Lambda(0), \Lambda(t)] \quad \textit{pointwise } \forall t \in \mathbb{R}.
\end{aligned}$$

*Then*

$$\begin{aligned}
&\mathbb{E}[N_n(0)] \to \mathbb{E}[N(0)],\\
&\mathbb{E}[N_n(0)] \quad \textit{and} \quad Var[N_n(0)] \quad \textit{are bounded uniformly over } n \in \mathbb{N}_+,\\
&Cov[N_n(0), N_n(t)] \to Cov[N(0), N(t)] \quad \textit{pointwise } \forall t \in \mathbb{R}.
\end{aligned}$$

*Proof.* By the denseness of the Phase-type distributions in the non-negative probability distributions (Theorem 2.11), we can find Phase-type service time distributions $S_n$ with distribution functions $F_n$, such that $F_n$ converges to $F$, both in distribution and expectation. We then use Tonelli's Theorem (which we can do due to the integrand being non-negative) to conclude

$$\mathbb{E}[N_n(0)] = \mathbb{E}\left[\int_{t=0}^{\infty} \Lambda_n(t)\overline{F}_n(t)dt\right] = \int_{t=0}^{\infty} \mathbb{E}[\Lambda_n(t)]\overline{F}_n(t)dt = \mathbb{E}[\Lambda_n(0)]\mathbb{E}[S_n].$$

Hence

$$\mathbb{E}[N_n(0)] = \mathbb{E}[\Lambda_n(0)]\mathbb{E}[S_n] \to \mathbb{E}[\Lambda(0)]\mathbb{E}[S] = \mathbb{E}[N(0)].$$

We have just seen that

$$\mathbb{E}[N_n(0)] = \mathbb{E}[\Lambda_n(0)]\mathbb{E}[S_n] < ML$$

and so is clearly bounded uniformly. Now we will show $Var[N_n(0)]$ is uniformly bounded too. By equation (6)

$$\mathbb{E}[Var(N_n(0)|\Lambda_n(s), s \in \mathbb{R})] = \mathbb{E}[\Lambda_n(0)] \int_0^{\infty} \overline{F}_n(t)dt = \mathbb{E}[\Lambda_n(0)]\mathbb{E}[S_n] < ML.$$

And by equation (9)

$$\begin{aligned}
Var[\mathbb{E}(N_n(0)|\Lambda_n(s), s \in \mathbb{R})] &= \int_0^{\infty} \overline{F}_n(t) \int_0^{\infty} Cov(\Lambda_n(t), \Lambda_n(s))\overline{F}_n(s)dsdt\\
&\leq \int_0^{\infty} \overline{F}_n(t) \int_0^{\infty} Var[\Lambda_n(0)]\overline{F}_n(s)dsdt < ML^2.
\end{aligned}$$

The uniform boundedness of $Var[N_n(0)]$ is now just a consequence of the law of total variance.

Now we compute the stationary autocovariance function of a single $Cox/G/\infty$ queue in the series using the law of total covariance. First, by equation (6)

$$\begin{aligned}
&\mathbb{E}[Cov(N_n(0), N_n(u)|\Lambda_n(s), s \in \mathbb{R})] \\
&= \mathbb{E}[\Lambda_n(0)] \int_{t=u}^{\infty} \overline{F}_n(t)dt \\
&= \mathbb{E}[\Lambda_n(0)] \left[ \int_{t=0}^{\infty} \overline{F}_n(t)dt - \int_{t=0}^{u} \overline{F}_n(t)dt \right] \\
&= \mathbb{E}[\Lambda_n(0)] \left[ \mathbb{E}[S_n] - \int_{t=0}^{u} \int_{x=t}^{\infty} dF_n(x)dt \right] \\
&= \mathbb{E}[\Lambda_n(0)] \left[ \mathbb{E}[S_n] - \int_{x=0}^{\infty} \int_{t=0}^{u \wedge x} dt dF_n(x) \right] && \text{(by Tonelli's Theorem)} \\
&= \mathbb{E}[\Lambda_n(0)] \left[ \mathbb{E}[S_n] - \int_{x=0}^{\infty} (u \wedge x) dF_n(x) \right] \\
&= \mathbb{E}[\Lambda_n(0)] \left[ \mathbb{E}[S_n] - \mathbb{E}[S_n \wedge u] \right] \\
&= \mathbb{E}[\Lambda_n(0)] \mathbb{E}[(S_n - u)^+],
\end{aligned}$$

The notation $x^+ := \max\{0, x\}$ means the positive part of $x$ and $x \wedge y := \min\{x, y\}$. Therefore,

$$\mathbb{E}[Cov(N_n(0), N_n(u)|\Lambda_n(s), s \in \mathbb{R})] = \ \mathbb{E}[\Lambda_n(0)]\mathbb{E}[(S_n - u)^+] \to \mathbb{E}[\Lambda(0)]\mathbb{E}[(S - u)^+]$$

because

$$\mathbb{E}[S_n] = \mathbb{E}[(S_n - u)^+] + \mathbb{E}[S_n \wedge u].$$

But $\mathbb{E}[S_n]$ converges to $\mathbb{E}[S]$; and since $(S_n \wedge u)$ is a bounded and continuous function, we have that by weak convergence of $F_n$ to $F$, that $\mathbb{E}[S_n \wedge u] \to \mathbb{E}[S \wedge u]$. This implies that we must also have $\mathbb{E}[(S_n - u)^+] \to \mathbb{E}[(S - u)^+]$. Finally, use the fact that the limit of a product of sequences is the product of the limits.

By equations (7) and (8), we have

$$\mathbb{E}(N_n(0)|\Lambda_n(s), s \in \mathbb{R}) = \int_{t=0}^{\infty} \Lambda_n(t)\overline{F}_n(t)dt,$$

and

$$\mathbb{E}(N_n(u)|\Lambda_n(s), s \in \mathbb{R}) = \int_{t=u}^{\infty} \Lambda_n(t)\overline{F}_n(t - u)dt.$$

So it follows by equation (9) that

$$\begin{aligned} g_n(u) &:= Cov[\mathbb{E}(N_n(0)|\Lambda_n(s), s \in \mathbb{R}), \mathbb{E}(N_n(u)|\Lambda_n(s), s \in \mathbb{R})] \\ &= \int\limits_{t=0}^{\infty}\int\limits_{s=u}^{\infty} Cov[\Lambda_n(t), \Lambda_n(s)]\overline{F}_n(t)\overline{F}_n(s-u)dsdt \\ &= \int\limits_{t=0}^{\infty}\int\limits_{v=0}^{\infty} Cov[\Lambda_n(t), \Lambda_n(v+u)]\overline{F}_n(t)\overline{F}_n(v)dvdt \quad \text{(via the substitution } v = s-u) \\ &= \int\limits_{t=0}^{\infty} \overline{F}_n(t) \int\limits_{v=0}^{\infty} Cov[\Lambda_n(t), \Lambda_n(v+u)]\overline{F}_n(v)dvdt. \end{aligned}$$

Observe that by Markov's inequality

$$\overline{F}_n(t) = \mathbb{P}\left(S_n \geq t\right) = \mathbb{P}\left(S_n^{1+\varepsilon} \geq t^{1+\varepsilon}\right) \leq \frac{\mathbb{E}\left[S_n^{1+\varepsilon}\right]}{t^{1+\varepsilon}} \leq \frac{A}{t^{1+\varepsilon}},$$

and further that

$$\overline{F}_n(t) \leq h(t) := \min\left\{1, \frac{A}{t^{1+\varepsilon}}\right\},$$

while it is easy to see that

$$\int\limits_{0}^{\infty} h(t)dt < \infty.$$

So now note that for all $t$

$$\begin{aligned} &\overline{F}_n(t)\left[\int\limits_{v=0}^{\infty}\left|Cov[\Lambda_n(t), \Lambda_n(v+u)]\right|\overline{F}_n(v)dv\right] \leq h(t)\int\limits_{v=0}^{\infty} Var[\Lambda_n(t)]\overline{F}_n(v)dv \\ &\leq h(t)M\int\limits_{v=0}^{\infty}\overline{F}_n(v)dv \leq h(t)ML. \end{aligned}$$

But,

$$\int\limits_{t=0}^{\infty} h(t)MLdt < \infty.$$

So by the Dominated Convergence Theorem

$$\begin{aligned} \lim_{n\to\infty} g_n(u) &= \int\limits_{t=0}^{\infty} \lim_{n\to\infty}\overline{F}_n(t)\left[\int\limits_{v=0}^{\infty} Cov[\Lambda_n(t), \Lambda_n(v+u)]\overline{F}_n(v)dv\right]dt \\ &= \int\limits_{t=0}^{\infty}\overline{F}(t)\left[\lim_{n\to\infty}\int\limits_{v=0}^{\infty} Cov[\Lambda_n(t), \Lambda_n(v+u)]\overline{F}_n(v)dv\right]dt, \end{aligned}$$

as the limit of a product is the product of the limits. But also

$$Cov[\Lambda_n(t), \Lambda_n(v+u)]\overline{F}_n(v) \leq Var[\Lambda_n(t)]h(v) \leq Mh(v),$$

and

$$\int_{v=0}^{\infty} Mh(v)dv < \infty.$$

So by the Dominated Convergence Theorem

$$\begin{aligned}\lim_{n\to\infty} g_n(u) &= \int_{t=0}^{\infty} \overline{F}(t)\left[\int_{v=0}^{\infty} \lim_{n\to\infty} Cov[\Lambda_n(t), \Lambda_n(v+u)]\overline{F}_n(v)dv\right]dt \\ &= \int_{t=0}^{\infty} \overline{F}(t)\left[\int_{v=0}^{\infty} Cov[\Lambda(t), \Lambda(v+u)]\overline{F}(v)dv\right]dt,\end{aligned}$$

using again the fact that the limit of a product is the product of the limits. So in summary

$$Cov[N_n(0), N_n(t)] \to Cov[N(0), N(t)].$$

□

Note that we have not shown convergence in distribution of $N_n$ to $N$, that is we have not shown that the equilibrium queue length distribution converges to the equilibrium queue length distribution of the limiting system. We have made the weaker assumption that the mean and covariance of the input process converge, and this is enough to get analogous results for the output. We anticipate that something similar can be shown for higher moments too, the calculations will however become more arduous. This is enough to propagate the result through the tandem queueing system, where the arrival process into a queue is given by a constant multiplied by the occupancy of the previous queue. It remains an open question exactly what assumptions are needed for distributional convergence. One way to approach this may be to compute the characteristic function of the stationary number of customers in the queue and then use Lévy's Continuity Theorem. To get the desired convergence in distribution then reduces to showing pointwise convergence of the corresponding sequence of characteristic functions.

# 5 LDP for Cox Processes and $Cox/G/\infty$ queues

Bar a few small (mostly typographical) changes, the work in this chapter has been submitted as an article that is available as a preprint [35] and is joint work with my supervisor Ayalvadi Ganesh and Edward Crane. A less general version of the weakly compact set of Proposition 5.9 used to show exponential tightness was originally suggested by Edward Crane.

## 5.1 Motivation and Outline

The work in this chapter is motivated by the problem of modelling fluctuations in the number of protein molecules in a cell. The synthesis of proteins is catalysed by RNA molecules, which in turn are transcribed from DNA molecules. Both RNA and protein molecules degrade spontaneously after some random time. It is important for proper functioning of the cell that protein numbers are maintained within certain limits, and biologists are interested in understanding the regulatory mechanisms involved in controlling their fluctuations. Consequently, the problem of modelling stochastic fluctuations has attracted interest, and there has been considerable work on Markovian models of such systems; see, e.g., [70, 83]. These models assume that each copy of a gene creates RNA molecules according to a Poisson process (while active), that each RNA molecule generates protein molecules according to a Poisson process, and that the lifetimes of RNA and protein molecules are Exponentially distributed. The assumption of Exponential lifetimes is biologically unrealistic; for example, inhomogeneities in the cellular environment could result in lifetimes that are mixtures of Exponential distributions, or the denaturing of molecules could be a multistage process.

Our approach relies on modelling the chemical kinetics using $\bullet/G/\infty$ queues rather than Markov processes, which correspond to $\bullet/M/\infty$ queues. Customer arrivals into the queue correspond to the synthesis of molecules of a specified type; after independent lifetimes with a general distribution, the molecules decay which equates to service (and departure) of the corresponding customers. For the problem described above, we have two such queues in series, one for RNA molecules and one for proteins. However, unlike in a tandem queueing network, where departures from one queue enter the next queue in series, here departures just leave the system; the way influence propagates is that the arrival rate into the protein queue is modulated by the occupancy of the preceding queue (here, RNA) in the series. We consider a very simple form of modulation, in which the arrival rate into a queue is proportional to the occupancy of the preceding queue, and the arrival process is conditionally Poisson given the occupancy. Thus, this results in a Cox process model for the arrivals into a queue, and the system is modelled as a series of $Cox/G/\infty$ queues interacting as described.

We briefly recall the description of the queue length process in an $M/G/\infty$ queue with arrival rate $\lambda$ and service distribution $F$. The arrival process into this queue can be represented as an inhomogeneous Poisson process on $\mathbb{R} \times \mathbb{R}_+$ with intensity measure $\lambda \otimes F$. If a realisation of this point process has a point at $(t, y)$, it denotes that a customer arrives at time $t$ bringing a service requirement of $y$. The queue length at time $t$ is simply the total number of points of the Poisson process in the set

$$A_t = \{(s, y) : s \leq t, y > t - s\},$$

as a customer arriving at time $s$ will still be in the system at time $t$ if and only if its service requirement is greater than $t - s$. (We follow the convention of defining the queue length process to be right continuous.) Likewise, the queue length process during a time interval $[s, t]$ can be described in terms of the empirical measure of the above Poisson process on the wedge-shaped set

$$A_{[s,t]} = \bigcup_{u \in [s,t]} A_u.$$

In the problem we want to study, the intensity of the arrival process is modulated by the number of customers present in the previous queue. Hence, we need to model it as a Cox process and study the corresponding $Cox/G/\infty$ queue. As described above, this requires us to study the empirical measure of a Cox process on a subset of $\mathbb{R}^2$. We shall in fact study them in a more general setting of $\sigma$-compact Polish spaces, namely Polish spaces that can be covered by countably many compact subsets. Our goal is to obtain functional large deviation principles (FLDPs) for the corresponding queue length processes; we shall obtain these by contraction from LDPs for the empirical measure of the Cox process. We have not been able to drop the technical assumption of $\sigma$-compactness from the proof, but do not know if it is essential for the stated results.

We now set out our Cox process model. Let $(E, d)$ be a $\sigma$-compact Polish space, and let $\Lambda$ be a random finite Borel measure on $E$; in other words, $\Lambda$ is a random variable taking values in $\mathcal{M}_+^f(E)$, the space of finite non-negative Borel measures on $E$. A Cox process $\Phi$ with stochastic intensity $\Lambda$ is a point process which is conditionally Poisson, with intensity measure $\lambda$ on the event that $\Lambda = \lambda$. Note that the point process $\Phi$ is almost surely finite. A realisation of $\Phi$ can be thought of as either a point set $\{x_1, x_2, \ldots, x_k\}$, or as a counting measure $\sum_{i=1}^k \delta_{x_i}$, where $k \in \mathbb{N}$ is the total number of points of the process (in general this is random, but once we have sampled a realisation of the process it is fixed). We call the latter the empirical measure corresponding to the realisation of the point set, and note that it is also an element of $\mathcal{M}_+^f(E)$. There are two topologies on $\mathcal{M}_+^f(E)$ which will be of interest to us. We say that a sequence of measures $\mu_n \in \mathcal{M}_+^f(E)$ converges to $\mu \in \mathcal{M}_+^f(E)$ in the weak topology if $\int_E f d\mu_n$ converges to $\int_E f d\mu$ for all bounded continuous functions $f : E \to \mathbb{R}$; we say the measures converge in the vague topology if the integrals converge only for continuous functions with compact support (which are necessarily bounded).

We now consider a sequence of Cox point processes $\Phi_n$, with corresponding stochastic intensities $\Lambda_n$. Our first major result in this chapter is a large deviations principle (see Definition 2.14) for their scaled empirical measures:

**Theorem 5.1.**
*Suppose that $(\Lambda_n, n \in \mathbb{N})$ is a sequence of random finite Borel measures on a $\sigma$-compact Polish space $(E, d)$, and that the sequence $\Lambda_n/n$ satisfies an LDP in $\mathcal{M}_+^f(E)$ equipped with the weak topology, with good rate function $\mathfrak{I}_1(\cdot)$. Let $\Phi_n$ be a Cox process with stochastic intensity $\Lambda_n$, i.e., a random counting measure on E equipped with its Borel $\sigma$-algebra. Then the sequence of measures $\Phi_n/n$ satisfies an LDP in $\mathcal{M}_+^f(E)$ equipped with the weak topology, with good rate function*

$$\mathfrak{I}_2(\mu) = \begin{cases} \inf_\lambda \left\{\mathfrak{I}_1(\lambda) + \lambda(E)\right\}, & \text{if } \mu \equiv 0, \\ \inf_\lambda \left\{\mathfrak{I}_1(\lambda) + I_{Poi}(\mu(E), \lambda(E)) + \mu(E) H\left(\frac{\mu}{\mu(E)} \,\middle|\, \frac{\lambda}{\lambda(E)}\right)\right\}, & \text{if } \mu \not\equiv 0, \end{cases}$$

*where $H$ is defined in the statement of Theorem 5.5 and $I_{Poi}$ in the statement of Lemma 5.6.*

A slightly different version of this theorem, with only local finiteness of the measures $\Lambda_n$ assumed, has been established by Schreiber [91], albeit in the vague rather than the weak topology; his result also requires a technical assumption about the measures $\Lambda_n/n$ dominating a fixed measure with full support on $E$, which we do not need. However, his result does not require that the space be $\sigma$-compact. The extension of the result to the weak topology is nontrivial, and relies on the finiteness assumption on the intensity measures. In addition, our proof techniques are very different. A functional LDP for rescaled Poisson random measures is proved in [42] using projective limits, and in [69] using Cramér's Theorem and subadditivity arguments.

The claim of Theorem 5.1 appears intuitive from the assumed LDP for the intensity measures $\Lambda_n/n$, the LDP for a Poisson random variable, and Sanov's Theorem for the empirical distribution. However, a number of technical conditions need to be checked. Moreover, while these imply an LDP, goodness of the rate function is not immediate. We show this indirectly by establishing exponential tightness (see Definition 2.15); this is the step where finiteness of the measures is crucial.

Next, we consider a sequence of stationary $Cox/G/\infty$ queues where the arrival processes are sped up by the index $n \in \mathbb{N}$, while the service process remains unchanged. More precisely, the service times are i.i.d. with some fixed distribution $F$ that does not depend on $n$, while the arrival process into the $n^{\text{th}}$ queue is a Cox process with stochastic intensity (directing measure) $\Lambda_n$ on $\mathbb{R}$. We make the following assumptions.

### Assumptions

A1 $(\Lambda_n, n \in \mathbb{N})$ is a sequence of random $\sigma$-finite measures on $\mathbb{R}$, whose laws are translation invariant, such that $\mathbb{E}[\Lambda_n([a,b])] = n\lambda(b-a)$, for some fixed $\lambda > 0$, and any compact interval $[a,b] \subset \mathbb{R}$.

A2 For any interval $[a,b]$, the sequence $(\Lambda_n/n)|_{[a,b]}$ obeys an LDP on $\mathcal{M}_+^f([a,b])$ equipped with the weak topology, with good rate function $I_{[a,b]}$.

A3 Define

$$\psi_n(\theta) = \log \mathbb{E}\left[e^{\frac{\theta\Lambda_n([0,1])}{n}}\right].$$

There is a neighbourhood of 0 on which $\psi_n(n\theta)/n$ is bounded, uniformly in $n$.

A4 The mean service time, given by $\int_0^\infty x dF(x) = \int_0^\infty \overline{F}(x)dx$, is finite; here $\overline{F} = 1-F$ denotes the complementary cumulative distribution function of the service time.

Let $Q_n(t)$ denote the number of customers at time $t$ in the infinite-server queue with Cox process arrivals with intensity $\Lambda_n$ and i.i.d. service times with distribution $F$. Let $L_n$ denote the measure on $\mathbb{R}$ which is absolutely continuous with respect to Lebesgue measure, with density $Q_n(\cdot)$. Our second main result in this chapter is the following:

**Theorem 5.2.**
*Consider a sequence of $Cox/G/\infty$ queues indexed by $n \in \mathbb{N}$, where the arrival process into the $n^{\text{th}}$ queue is a Cox process with directing measure $\Lambda_n$, and service times are i.i.d. with common distribution $F$. Suppose the arrival and service processes satisfy Assumptions [A1]-[A4]. Let $Q_n(t)$ denote the number of customers in the $n^{\text{th}}$ queue at time $t$, and let $L_n$ denote the random measure on $\mathbb{R}$ which is absolutely continuous with respect to Lebesgue measure and has density $Q_n(\cdot)$. Then the sequence of measures $L_n$ satisfies Assumptions [A1]-[A3]. In particular, for any compact interval $[a,b] \subset \mathbb{R}$, the measures $(L_n/n)|_{[a,b]}$ satisfy an LDP on $\mathcal{M}_+^f([a,b])$ equipped with the weak topology, with good rate function $J_{[a,b]}$.*

A fuller description of the rate function $J_{[a,b]}$ is provided in the proof of this theorem, in section 5.3. The theorem shows that the sequence of queue occupancy measures $L_n$ also satisfies the above assumptions and, in particular, that they satisfy an LDP. This implies that our analysis extends easily to an arbitrary number of $Cox/G/\infty$ queues in tandem, where the arrivals into each queue constitute a Cox process with directing measure given by the number in the previous queue. This is the set-up that motivated this work. The theorem yields an LDP for the occupancy measure of each of these queues.

The $Cox/G/\infty$ model studied is an instance of a queue in a random environment. The first study of infinite-server queues in random environment was in [77]: factorial moments in stationarity are derived for the $M/M/\infty$ queue in a Markovian environment, namely one in which the arrival and service rates are modulated by a finite state, irreducible, continuous time Markov chain. There has recently been extensive further study of this model, including moments for steady state and transient distributions, and large deviation and central limit asymptotics for the marginal distribution of the queue length; see [20] for a collation of the results.

The Markovian assumption on the environment is relaxed in [56], where the background process modulating arrivals and services in an $M/M/\infty$ queue is just a general càdlàg stochastic process. An LDP is proved for the queue length at an arbitrary fixed time, $t$, whereas we establish a process level LDP, without assuming (conditionally) Exponential service times. A special type of Cox background process is considered in [51], which proves a functional CLT for the scaled queue length process. In all of these cases the queue length is viewed as a random càdlàg function, whereas we view it as living on a space of measures.

## 5.2 Proof of Empirical Measure LDP

Our proof of Theorem 5.1 relies on a theorem of Chaganty [24], which essentially states that a sequence of probability measures on a product space satisfies an LDP if the corresponding sequences of marginal and conditional probability distributions do so, and certain additional technical conditions are satisfied. For completeness, we include below a statement of this theorem, together with an extension of Sanov's Theorem by Baxter and Jain [11] which is needed to check its conditions, and relevant definitions.

**Definition 5.3.** *([24] Page 2)*
*Let $(\Omega_1, \mathcal{B}_1)$ and $(\Omega_2, \mathcal{B}_2)$ be two Polish spaces with their associated Borel $\sigma$–fields.*

*Let $\{\nu_n(\cdot,\cdot)\}$ be a sequence of transition functions on $\Omega_1 \times \mathcal{B}_2$, i.e., $\nu_n(x_1,\cdot)$ is a probability measure on $(\Omega_2, \mathcal{B}_2)$ for each $x_1 \in \Omega_1$ and $\nu_n(\cdot, B_2)$ is a measurable function on $\Omega_1$ for each $B_2 \in \mathcal{B}_2$. We say that the sequence of probability transition functions $\{\nu_n(x_1,\cdot), x_1 \in \Omega_1\}$ satisfies the LDP continuously in $x_1$ with rate function $J(x_1,x_2)$, or simply the LDP continuity condition holds, if:*

1. *For each $x_1 \in \Omega_1$, $J(x_1,\cdot)$ is a good rate function on $\Omega_2$, i.e., it is non-negative, lower semicontinuous, and has compact level sets.*

2. *For any sequence $\{x_{1n}\}$ in $\Omega_1$ such that $x_{1n} \to x_1$, the sequence of measures $\{\nu_n(x_{1n},\cdot)\}$ on $\Omega_2$ obeys the LDP with rate function $J(x_1,\cdot)$.*

3. *$J(x_1,x_2)$ is lower semicontinuous as a function of $(x_1,x_2)$.*

**Theorem 5.4.** *([24] Theorem 2.3)*
*Let $(\Omega_1,\mathcal{B}_1)$, $(\Omega_2,\mathcal{B}_2)$ be two Polish spaces with their associated Borel $\sigma-$fields. Let $\{\mu_{1n}\}$ be a sequence of probability measures on $(\Omega_1,\mathcal{B}_1)$. Let $\{\nu_n(x_1,B_2)\}$ be a sequence of probability transition functions defined on $\Omega_1 \times \mathcal{B}_2$. We define the joint distribution $\mu_n$ on the product space $\Omega_1 \times \Omega_2$, and the marginal distribution $\mu_{2n}$ on $\Omega_2$ by*

$$\mu_n(B_1 \times B_2) = \int\limits_{B_1} \nu_n(x_1, B_2) d\mu_{1n}(x_1), \quad \mu_{2n}(B_2) = \mu_n(\Omega_1 \times B_2).$$

*Suppose that the following two conditions are satisfied:*

1. *$\{\mu_{1n}\}$ satisfies an LDP with good rate function $I_1(x_1)$.*

2. *$\{\nu_n(\cdot,\cdot)\}$ satisfies the LDP continuity condition with a rate function $J(x_1,x_2)$.*

*Then the sequence of joint distributions $\{\mu_n\}$ satisfies a weak LDP on the product space $\Omega_1 \times \Omega_2$, with rate function*

$$I(x_1,x_2) = I_1(x_1) + J(x_1,x_2).$$

*The sequence of marginal distributions $\mu_{2n}$ satisfies an LDP with rate function*

$$I_2(x_2) = \inf_{x_1 \in \Omega_1} \left[ I_1(x_1) + J(x_1,x_2) \right].$$

*Finally, $\{\mu_n\}$ satisfies the LDP if $I(x_1,x_2)$ is a good rate function.*

**Remark.**
Recall that a sequence of probability measures (or random variables) is said to satisfy a weak LDP if the large deviations upper bound holds for all compact sets, and to satisfy a (full) LDP if it holds for all closed sets. For both, the large deviations lower bound holds for all open sets (see Definition 2.14).

**Theorem 5.5.** *([11] Theorem 5)*
*Let $(S,d)$ be a Polish space. Let $\{\alpha_n\}$ be a sequence of probability measures on $(S,d)$ converging weakly to a probability measure $\alpha$. For each $n$, let $X_i^n$, $i \in \mathbb{N}$ be i.i.d. $S-$valued random variables with common distribution $\alpha_n$. Let $\mathcal{M}_1(S)$ denote the space of probability measures on $S$ and let $\overline{\mu}_n \in \mathcal{M}_1(S)$ denote the empirical distribution, $\left(\delta_{X_1^n} + \ldots + \delta_{X_n^n}\right)/n$. Then $\{\overline{\mu}_n\}$ satisfies the LDP with good rate function $H(\cdot|\alpha)$ given by:*

$$H(\beta|\alpha) = \begin{cases} \int \log(d\beta/d\alpha)d\beta & \text{if } \beta << \alpha \text{ and } \int |\log(d\beta/d\alpha)|d\beta < \infty \\ \infty & \text{otherwise.} \end{cases}$$

*The function $H(\beta|\alpha)$ is called the relative entropy or Kullback-Leibler divergence of $\beta$ with respect to $\alpha$.*

The proof of Theorem 5.1 proceeds through a sequence of lemmas. We begin with an elementary LDP for a sequence of Poisson random variables.

**Lemma 5.6.**
*Let $N_n, n \in \mathbb{N}$ be a sequence of Poisson random variables with parameter $n\alpha_n$, and suppose that $\alpha_n$ tends to $\alpha \geq 0$. Then the sequence $N_n/n$ obeys an LDP in $\mathbb{R}_+$ with good rate function $I_{Poi}(\cdot,\alpha)$ given by*

$$I_{Poi}(x,\alpha) = \begin{cases} x\log\frac{x}{\alpha} - x + \alpha, & \text{if } \alpha > 0, \\ 0, & \text{if } \alpha = 0, x = 0, \\ +\infty, & \text{if } \alpha = 0, x > 0. \end{cases}$$

*Proof.* We apply the Gärtner-Ellis Theorem [38, Theorem 2.3.6] to the sequence $N_n/n$. By direct calculation,

$$\frac{1}{n}\log\mathbb{E}\left[e^{n\theta\frac{N_n}{n}}\right] = \alpha_n\left(e^\theta - 1\right).$$

This sequence of scaled log-moment generating functions converges pointwise to the limit $\alpha(e^\theta - 1)$, which is finite and differentiable everywhere (hence also continuous, so in particular lower semicontinuous) and essentially smooth. Hence, by the Gärtner-Ellis Theorem, the sequence of random variables $N_n/n$ obeys an LDP with a rate function which is the convex conjugate of $\alpha(e^\theta - 1)$. A straightforward calculation confirms that this is the function $I_{Poi}(\cdot,\lambda)$ in the statement of the lemma, and that it is lower semicontinuous with compact level sets for each $\alpha$. □

The next two lemmas establish conditional LDPs for the scaled empirical measures of Poisson processes whose scaled intensities converge to a limit.

**Lemma 5.7.**

*Let $\Phi_n, n \in \mathbb{N}$ be a sequence of Poisson point processes with intensity measures $n\lambda_n$, and suppose that $\lambda_n$ converge weakly in $\mathcal{M}_+^f(E)$ to the zero measure. Then, $\Phi_n/n, n \in \mathbb{N}$ satisfy the LDP in $\mathcal{M}_+^f(E)$ equipped with the weak topology, with good rate function*

$$\mathcal{I}_0(\mu) = \begin{cases} 0, & \textit{if } \mu \equiv 0, \\ +\infty, & \textit{otherwise.} \end{cases}$$

*Proof.* As the map $\mu \mapsto \mu(E)$ is weakly continuous (the indicator of $E$ is one everywhere, in particular it is a bounded, continuous function), it is limit preserving and so it follows that $\lambda_n(E)$ tends to $\lambda(E) = 0$. Let $N_n = \Phi_n(E)$ denote the total number of points in the Poisson process $\Phi_n$. Then, $N_n$ is a Poisson random variable with parameter $n\lambda_n(E)$, and it follows from Lemma 5.6 that $(N_n/n, n \in \mathbb{N})$ obey an LDP with good rate function

$$I_{Poi}(x,0) = \begin{cases} 0, & \text{if } x = 0, \\ +\infty, & \text{if } x > 0. \end{cases}$$

Let $F \subset \mathcal{M}_+^f(E)$ be closed in the weak topology (so it contains its weak limits), and suppose that it does not contain the zero measure. Define

$$x_F = \inf\{\mu(E) : \mu \in F\}.$$

We claim that $x_F > 0$. Indeed, if $x_F = 0$, then we can find a sequence of measures $\mu_n \in F$ such that $\mu_n(E)$ tends to zero, i.e., $\int_E 1 d\mu_n$ tends to zero. It follows that $\int_E f d\mu_n$ tends to zero for all bounded functions $f$, and so in particular for all bounded, continuous functions. Hence, the sequence $\mu_n$ converges weakly to the zero measure, contradicting the assumption that $0 \notin F$ and $F$ is closed.

Using the inequality

$$N_n = \Phi_n(E) \geq \inf_{\mu \in F} \{\mu(E)\} = x_F$$

we now have the large deviations upper bound for $F$:

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{\Phi_n}{n} \in F\Big) &\leq \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{\Phi_n(E)}{n} \geq x_F\Big) \\ &= \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{N_n}{n} \geq x_F\Big) = -\infty, \end{aligned}$$

where we have used the LDP for $N_n/n$ with rate function $I_{Poi}(\cdot, 0)$ and the fact that $x_F > 0$ to obtain the last equality.

The large deviations lower bound is trivial for open sets $G$ not containing the zero measure, as the infimum of the rate function is infinite on such sets. Now, for $G$

containing the zero measure, we have

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big(\frac{\Phi_n}{n}\in G\Big) &\geq \liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big(\frac{\Phi_n}{n}\equiv 0\Big) = \liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}(N_n=0)\\ &= \liminf_{n\to\infty}(-\lambda_n(E)) = -\lambda(E)=0,\end{aligned}$$

as $N_n \sim Poi(n\lambda_n(E))$. This completes the proof of the lemma. □

**Lemma 5.8.**
*Let $\Phi_n, n\in\mathbb{N}$ be a sequence of Poisson point processes with intensity measures $n\lambda_n$, and suppose that the sequence $\lambda_n$ converges in the weak topology on $\mathcal{M}_+^f(E)$ to $\lambda\not\equiv 0$. Then, $\Phi_n/n, n\in\mathbb{N}$ satisfy the LDP in $\mathcal{M}_+^f(E)$ equipped with the weak topology, with good rate function*

$$\mathcal{I}_1(\mu)=\begin{cases} I_{Poi}(\mu(E),\lambda(E))+\mu(E)H\Big(\frac{\mu}{\mu(E)}\,\Big|\,\frac{\lambda}{\lambda(E)}\Big), & \textit{if } \mu\not\equiv 0,\\ I_{Poi}(0,\lambda(E)), & \textit{if } \mu\equiv 0.\end{cases}$$

*Here, $I_{Poi}(\cdot,\cdot)$ and $H(\cdot|\cdot)$ are as defined in Lemma 5.6 and Theorem 5.5 respectively.*

*Proof.* We will prove the lemma by first establishing an LDP for the sequence $N_n/n$, then verifying that conditional on this, $\Phi_n/n$ satisfies the LDP continuously, and invoking Theorem 5.4.

The LDP for $N_n/n$, with rate function $I_{Poi}(\cdot,\lambda(E))$, is immediate from Lemma 5.6 since $\lambda_n(E)$ tends to $\lambda(E)$. We now prove an LDP for $\Phi_n/n$, conditional on $N_n/n$. Fix a sequence $N_n$ such that $N_n/n\to x\geq 0$. If $x=0$, then the proof of the LDP follows that of Lemma 5.7, and yields $\mathcal{I}_0$ as the rate function.

It remains to consider $x>0$. We can write

$$\Phi_n=\delta_{X_1^n}+\delta_{X_2^n}+\ldots+\delta_{X_{N_n}^n},$$

where the $X_i^n$ are i.i.d., with law $\frac{\lambda_n}{\lambda_n(E)}$. Note that the probability law of $X_i^n$ is well-defined for all $n$ sufficiently large, as $\lambda_n(E)$ tends to $\lambda(E)>0$. Define

$$\hat{\Phi}_n=\delta_{X_1^n}+\delta_{X_2^n}+\ldots+\delta_{X_{\lfloor nx\rfloor}^n},$$

where the dependence of $\hat{\Phi}_n$ on $x$ has been suppressed in the notation. We claim that the sequences $\Phi_n/n$ and $\hat{\Phi}_n/n$ are exponentially equivalent (see Definition 2.19). To see this, we use the fact that the weak topology on $\mathcal{M}_+^f(E)$ can be metrised, for instance by the Kantorovich-Rubinstein metric,

$$d_{\mathrm{KR}}(\mu,\nu)=\sup_{f\in\mathrm{Lip}(1),\|f\|_\infty\leq 1}\int_E f d\mu-\int_E f d\nu.$$

It is easy to see that

$$d_{KR}\left(\frac{\Phi_n}{n},\frac{\hat{\Phi}_n}{n}\right)\leq\frac{1}{n}\sup_{f\in\text{Lip}(1),\|f\|_\infty\leq1}\Big|\int_E fd\Phi_n-\int_E fd\hat{\Phi}_n\Big|\leq\frac{\|f\|_\infty}{n}\Big|N_n-\lfloor nx\rfloor\Big|,$$

and so, $d_{\text{KR}}(\Phi_n/n,\hat{\Phi}_n/n)$ tends to zero deterministically, because $N_n/n$ tends to $x$ deterministically. This establishes the exponential equivalence of the two sequences.

Now, we have from Theorem 5.5 and the observation that $\lambda_n(\cdot)/\lambda_n(E)$ converges weakly to $\lambda(\cdot)/\lambda(E)$ (by continuous mapping), that $(\hat{\Phi}_n/\lfloor nx\rfloor,\lfloor nx\rfloor\in\mathbb{N})$ obey an LDP in $\mathcal{M}_1(E)$ with good rate function $H\big(\cdot\ \big|\ \frac{\lambda}{\lambda(E)}\big)$, and hence also in $\mathcal{M}_+^f(E)$ with rate function which is the same on $\mathcal{M}_1(E)$, and infinite outside it. It follows that $(\hat{\Phi}_n/n, n\in\mathbb{N})$ obey an LDP in $\mathcal{M}_+^f(E)$ with rate function

$$H_x(\mu)=\begin{cases}xH\big(\frac{\mu}{x}\ \big|\ \frac{\lambda}{\lambda(E)}\big), & \text{if } \frac{\mu}{x}\in\mathcal{M}_1(E),\\ +\infty, & \text{otherwise.}\end{cases}\tag{10}$$

Finally, by Theorem 2.20, $(\Phi_n/n, n\in\mathbb{N})$ obey an LDP in $\mathcal{M}_+^f(E)$ with the same rate function $H_x$, as they are exponentially equivalent to $\hat{\Phi}_n/n$.

Having established conditional LDPs for $\Phi_n/n$, conditional on $N_n/n$ tending to $x$, we now need to check the LDP continuity conditions in Definition 5.3 with $\Omega_1=\mathbb{R}_+$ and $\Omega_2=\mathcal{M}_+^f(E)$, and transition function $\nu_n(x,\cdot)$ defined as the law of $\Phi_n$ conditional on $N_n=\lfloor nx\rfloor$. We defne the function

$$J(x,\mu)=\begin{cases}\mathcal{I}_0(\mu), & \text{if } x=0,\\ H_x(\mu) & \text{if } x>0,\end{cases}$$

where $\mathcal{I}_0$ is defined in Lemma 5.7 and $H_x$ in (10). Note that $J$ is non-negative as $\mathcal{I}_0$ and $\{H_x, x\geq0\}$ are all non-negative.

The first condition in Definition 5.3 holds trivially if $x=0$, as all level sets are singletons comprised of the zero measure; if $x>0$, the condition follows from the goodness of the relative entropy function, which is well known from Sanov's Theorem (see, e.g., [38, Theorem 6.2.10]). In a bit more detail, given $\alpha>0$, the level set

$$L_\alpha=\left\{\mu\in\mathcal{M}_1(E):H\left(\mu\,\middle|\,\frac{\lambda}{\lambda(E)}\right)\leq\frac{\alpha}{x}\right\}$$

is compact in $\mathcal{M}_1(E)$ equipped with the weak topology; hence, so is its image under the continuous map $\mu\mapsto x\mu$ from $\mathcal{M}_1(E)$ to $\mathcal{M}_+^f(E)$.

The second condition in Definition 5.3 is precisely the content of the conditional LDPs that we just obtained. That leaves us to check the third condition, which is that $J(x,\mu)$ is lower semicontinuous in $(x,\mu)$. As $\mathbb{R}_+ \times \mathcal{M}_+^f(E)$ is a metric space, we can check this along sequences. Consider a sequence $(x_n,\mu_n)$ converging to $(x,\mu)$. If $(x,\mu) = (0,0)$, then $J(x,\mu) = 0$, which is no bigger than $\liminf J(x_n,\mu_n)$. If $x = 0$ and $\mu \not\equiv 0$, then $\mu(E) > 0$ and so, for all $n$ sufficiently large, $x_n < \mu_n(E)$; consequently, $\mu_n/x_n$ is not a probability measure, and $J(x_n,\mu_n) = +\infty$. The same reasoning applies if $x > 0$ and $\mu/x \notin \mathcal{M}_1(E)$. Finally, suppose $x > 0$ and $\mu/x \in \mathcal{M}_1(E)$, so that $\mu_n/x_n$ converges weakly to $\mu/x$ in $\mathcal{M}_+^f(E)$. We may restrict attention to the subsequence of $\mathbb{N}$ for which $\mu_n/x_n$ are probability measures, as $J(x_n,\mu_n) = +\infty$ otherwise. Along this subsequence, the desired inequality $\liminf H_{x_n}(\mu_n) \geq H_x(\mu)$ follows from the lower semicontinuity of $H$, the relative entropy function.

We are now in a position to invoke Theorem 5.4, with $\Omega_1 = \mathbb{R}_+$ and $\Omega_2 = \mathcal{M}_+^f(E)$. The second condition in the theorem is a conditional LDP for $\Phi_n/n$ given that $N_n/n$ tends to $x$, which we have just verified. The first condition is an LDP for $N_n/n$, which was proved in Lemma 5.6. Hence, the conclusion of Theorem 5.4 holds, i.e., we have an LDP for $\Phi_n/n$ with rate function

$$I_2(\mu) = \inf_{x\in\mathbb{R}_+} \left\{ I_{Poi}(x,\lambda(E)) + J(x,\mu) \right\}.$$

As $J(x,\mu) = +\infty$ unless $x = \mu(E)$, it is clear that the infimum is attained at $x = \mu(E)$, and we have

$$I_2(\mu) = I_{Poi}(\mu(E),\lambda(E)) + J(\mu(E),\mu).$$

This coincides with the rate function in the statement of the lemma, and concludes its proof. □

We now have all the ingredients required to complete the proof of Theorem 5.1.

**Proof of Theorem 5.1.**

We invoke Theorem 5.4 with $\Omega_1$ and $\Omega_2$ both being the space of finite non-negative measures on $E$, equipped with the weak topology and the corresponding Borel $\sigma$-algebra. The sequence $\mu_{1n}$ will denote the laws of the directing (intensity) measures $\Lambda_n$, and the probability transition functions $\nu_n(\lambda,\cdot)$ will denote the law of the scaled Poisson random measures $\Phi_n/n$, where $\Phi_n$ has intensity $n\lambda$. We now check the assumptions of the theorem.

The first condition in Theorem 5.4 is an LDP for $(\Lambda_n/n, n \in \mathbb{N})$ with a good rate function, which holds by assumption. To check the second condition in Theorem 5.4, define

$$J(\lambda,\mu) = \begin{cases} \mathcal{I}_0(\mu), & \text{if } \lambda \equiv 0, \\ \mathcal{I}_1(\mu), & \text{otherwise,} \end{cases}$$

where $\mathcal{I}_0$ and $\mathcal{I}_1$ are as defined in Lemmas 5.7 and 5.8. We need to check that the conditions in Definition 5.3 are satisfed. The first condition is satisfied as $\mathcal{I}_0$ and $\mathcal{I}_1$ are both good rate functions, as shown in Lemmas 5.7 and 5.8. The second condition is the content of the conditional LDPs established in these lemmas. That leaves us to check the third condition, that $J(\cdot,\cdot)$ is lower semicontinuous. As the weak topology on $\mathcal{M}_+^f(E)$ is metrisable, so is the product topology on $\mathcal{M}_+^f(E) \times \mathcal{M}_+^f(E)$, and we can check lower semicontinuity along sequences. Consider a sequence $(\lambda_n, \mu_n)$ converging to $(\lambda, \mu)$, i.e., $\lambda_n$ converges weakly to $\lambda$, and $\mu_n$ to $\mu$. We distinguish four cases:

1. If $\lambda \equiv 0$ and $\mu \equiv 0$, then $J(\lambda, \mu) = \mathcal{I}_0(\mu) = 0$, which is no bigger than the limit infimum of a non-negative sequence.

2. If $\lambda \equiv 0$ and $\mu \not\equiv 0$, then $J(\lambda, \mu) = \mathcal{I}_0(\mu) = +\infty$. But note that $\lambda_n(E) \to \lambda(E) = 0$ and $\mu_n(E) \to \mu(E) > 0$, and so $I_{Poi}(\mu_n(E), \lambda_n(E)) \to +\infty$. As
$$J(\lambda_n, \mu_n) = \mathcal{I}_1(\mu_n) \geq I_{Poi}(\mu_n(E), \lambda_n(E)),$$
we see that $J(\lambda_n, \mu_n)$ also tends to infinity.

3. If $\lambda \not\equiv 0$ and $\mu \equiv 0$, then $J(\lambda, \mu) = \mathcal{I}_1(\mu) = I_{Poi}(0, \lambda(E))$. On the other hand, $J(\lambda_n, \mu_n) \geq I_{Poi}(\mu_n(E), \lambda_n(E))$, which tends to $I_{Poi}(0, \lambda(E))$ as $n$ tends to infinity (as $I_{Poi}$ is continuous and hence limit preserving).

4. Finally, suppose that $\lambda \not\equiv 0$ and $\mu \not\equiv 0$. In this case, for all $n$ sufficiently large, both $\lambda_n$ and $\mu_n$ are non-zero measures, and we have $J(\lambda_n, \mu_n) = \mathcal{I}_1(\mu_n)$. As $\lambda_n(E)$ and $\mu_n(E)$ converge to $\lambda(E)$ and $\mu(E)$ respectively, it is easy to see that $I_{Poi}(\mu_n(E), \lambda_n(E))$ tends to $I_{Poi}(\mu(E), \lambda(E))$. Hence, to verify lower semicontinuity, it suffices to show that $H(\beta|\alpha)$ is jointly lower semicontinuous in its arguments. Recall the Donsker-Varadhan variational formula for the relative entropy (see, e.g., [40, Sec. C.2]):
$$H(\beta|\alpha) = \sup_{g \in C_b(E)} \left\{ \int_E g d\beta - \log \int_E e^g d\alpha \right\},$$
where $C_b(E)$ denotes the set of bounded continuous functions on $E$. But if $g \in C_b(E)$, so is $e^g$, and the map
$$(\alpha, \beta) \longmapsto \int_E g d\beta - \log \int_E e^g d\alpha$$
is continuous. Consequently, $H(\beta|\alpha)$, being the supremum of continuous functions of $(\alpha, \beta)$, is lower semicontinuous.

Thus, we have checked all the conditions of Theorem 5.4. Hence, the conclusion of the theorem holds, and yields that $(\Phi_n/n, n \in \mathbb{N})$ obey an LDP on $\mathcal{M}_+^f(E)$, with rate function

$$\mathfrak{I}_2(\mu) = \inf_{\lambda \in \mathcal{M}_+^f(E)} \{\mathfrak{I}_1(\lambda) + J(\lambda,\mu)\},$$

where $J(\lambda,\mu)$ equals $\mathcal{I}_0(\mu)$ if $\lambda \equiv 0$ and $\mathcal{I}_1(\mu)$ otherwise, and $\mathcal{I}_0$ and $\mathcal{I}_1$ are defined in Lemmas 5.7 and 5.8 respectively. Using those definitions, we can write the rate function more explicitly as follows:

$$\mathfrak{I}_2(\mu) = \begin{cases} \inf_\lambda \{\mathfrak{I}_1(\lambda) + \lambda(E)\}, & \text{if } \mu \equiv 0, \\ \inf_\lambda \left\{\mathfrak{I}_1(\lambda) + I_{Poi}(\mu(E),\lambda(E)) + \mu(E) H\left(\frac{\mu}{\mu(E)} \mid \frac{\lambda}{\lambda(E)}\right)\right\}, & \text{if } \mu \not\equiv 0, \end{cases}$$

where the infimum is taken over all finite Borel measures $\lambda$ on $E$. The expression above coincides with that in the statement of the theorem.

It remains only to check that the rate function $\mathfrak{I}_2$ is good. This is a consequence of Lemma 5.10 below, which establishes the exponential tightness of the scaled empirical measures $\Phi_n/n$, and Lemma 2.16. This completes the proof of Theorem 5.1. □

We first state a proposition which provides an explicit construction of compact subsets of $\mathcal{M}_+^f(E)$, and which we will need for the proof of Lemma 5.10. This is where $\sigma$-compactness is key. The proof of the proposition is deferred until after the lemma.

**Proposition 5.9.**
*Let $K_1 \subseteq K_2 \subseteq \dots$ be a nested sequence of compact subsets of $E$, whose union is equal to $E$; such a sequence exists by the assumption that $E$ is $\sigma$-compact. Let $\varepsilon_0 \geq \varepsilon_1 \geq \dots$ be a sequence of real numbers decreasing to zero. Define $K_0$ to be the empty set. Then, the set*

$$L_{(K_n,\varepsilon_n)} = \left\{\mu \in \mathcal{M}_+^f(E) : \mu(K_n^c) \leq \varepsilon_n \ \forall\ n \in \mathbb{N}\right\},$$

*is compact in the weak topology on $\mathcal{M}_+^f(E)$. Moreover, if $\mathcal{K}$ is any compact subset of $\mathcal{M}_+^f(E)$, and $\varepsilon_n, n \in \mathbb{N}_+$ any sequence decreasing to 0, then there exist $\varepsilon_0 > 0$ and compact $K_1 \subseteq K_2 \subseteq \dots \subseteq E$ such that $\mathcal{K} \subseteq L_{(K_n,\varepsilon_n)}$.*

**Lemma 5.10.**
*Suppose that $(\Lambda_n, n \in \mathbb{N})$ is a sequence of random finite Borel measures on a Polish space $(E,d)$, which satisfy the assumptions of Theorem 5.1. Let $(\Phi_n, n \in \mathbb{N})$ be a sequence of Cox point processes on $E$, with stochastic intensities $\Lambda_n$. Then, the sequence of random measures $\Phi_n/n$ is exponentially tight in $\mathcal{M}_+^f(E)$ equipped with the weak topology.*

*Proof.* We have to show that for every $\alpha < \infty$, there is a compact $\mathcal{K}_\alpha \subseteq \mathcal{M}_+^f(E)$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \log\ \mathbb{P}\left(\frac{\Phi_n}{n} \in \mathcal{K}_\alpha^c\right) < -\alpha. \tag{11}$$

By the assumptions of Theorem 5.1, the sequence $\Lambda_n/n$ satisfies an LDP in $\mathcal{M}_+^f(E)$, with *good* rate function $\mathfrak{I}_1$, therefore the sequence is exponentially tight. Hence, there is a compact set $\hat{\mathcal{K}}_\alpha \subseteq \mathcal{M}_+^f(E)$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{\Lambda_n}{n} \notin \hat{\mathcal{K}}_\alpha\Big) < -\alpha. \tag{12}$$

By Proposition 5.9, $\hat{\mathcal{K}}_\alpha$ is contained in a compact set of the form $L_{(K_n,\varepsilon_n)}$, where $\varepsilon_n, n \geq 1$ can be chosen to decrease to zero arbitrarily. We will show that, for a suitably chosen sequence $\delta_n \downarrow 0$, the set $L_{(K_n,\delta_n)}$ satisfies the upper bound in (11).

Observe that

$$\begin{aligned}\mathbb{P}\Big(\frac{\Phi_n}{n} \notin L_{(K_i,\delta_i)}\Big) &\leq \mathbb{P}\Big(\frac{\Phi_n}{n} \notin L_{(K_i,\delta_i)} \Big| \frac{\Lambda_n}{n} \in L_{(K_i,\varepsilon_i)}\Big) + \mathbb{P}\Big(\frac{\Lambda_n}{n} \notin L_{(K_i,\varepsilon_i)}\Big) \\ &\leq \mathbb{P}\Big(\frac{\Phi_n}{n} \notin L_{(K_i,\delta_i)} \Big| \frac{\Lambda_n}{n} \in L_{(K_i,\varepsilon_i)}\Big) + \mathbb{P}\Big(\frac{\Lambda_n}{n} \notin \hat{\mathcal{K}}_\alpha\Big). \end{aligned} \tag{13}$$

Now, conditional on $\Lambda_n$, $\Phi_n$ is a Poisson point process, and $\Phi_n(K_i^c)$ is a Poisson random variable with mean $\Lambda_n(K_i^c)$. Thus, conditional on $\Lambda_n/n \in L_{(K_i,\varepsilon_i)}$, the random variable $\Phi_n(K_i^c)$ is stochastically dominated by a Poisson random variable with mean $n\varepsilon_i$, for each $i \in \mathbb{N}$. Also, the event $\{\Phi_n/n \notin L_{(K_i,\delta_i)}\}$ is the union of the events $\{\Phi_n(K_i^c) > n\delta_i\}$ over $i \in \mathbb{N}$. Define $m_n = \sup\{i : n\delta_i > 1\}$. Since $\Phi_n$ is a counting measure, the event $\{\Phi_n(K_i^c) > n\delta_i\}$ coincides with $\{\Phi_n(K_i^c) \geq 1\}$ for $i > m_n$ (as then $0 < n\delta_i \leq 1$). Hence, we obtain using the union bound that

$$\begin{aligned}\mathbb{P}\Big(\frac{\Phi_n}{n} \notin L_{(K_i,\delta_i)} \Big| \frac{\Lambda_n}{n} \in L_{(K_i,\varepsilon_i)}\Big) &\leq \sum_{i=0}^{\infty} \mathbb{P}\big(Poi(n\varepsilon_i) > n\delta_i\big) \\ = \sum_{i=0}^{m_n} \mathbb{P}\big(Poi(n\varepsilon_i) > n\delta_i\big) &+ \sum_{i=m_n+1}^{\infty} \mathbb{P}\big(Poi(n\varepsilon_i) \geq 1\big). \end{aligned} \tag{14}$$

Without loss of generality, we can take $\varepsilon_0 \geq 1$. Take $\varepsilon_i = e^{-i}$ and $\delta_i = \kappa/i$ for $i \geq 1$, for a constant $\kappa$ to be determined, depending on $\alpha$. Take $\delta_0 = \kappa\varepsilon_0$. Then $m_n = \lfloor \kappa n \rfloor$, and we obtain using Markov's inequality that

$$\sum_{i=m_n+1}^{\infty} \mathbb{P}\big(Poi(n\varepsilon_i) \geq 1\big) \leq \sum_{i=\lceil \kappa n \rceil}^{\infty} ne^{-i} \leq \frac{ne^{-\kappa n}}{1-e^{-1}}. \tag{15}$$

We also have the large deviations (Chernoff) bound for a Poisson random variable that, for $\mu > \lambda$,

$$\mathbb{P}\big(Poi(\lambda) > \mu\big) \leq \exp\Big(-\mu \log\frac{\mu}{\lambda} + \mu - \lambda\Big),$$

from which it follows that

$$\mathbb{P}\big(Poi(n\varepsilon_i) > n\delta_i\big) \leq \begin{cases} \exp(-n\varepsilon_0(\kappa\log\kappa - \kappa + 1)), & i = 0, \\ \exp\big(-n\kappa\frac{\log\kappa + i - 1 - \log i}{i}\big), & i \geq 1. \end{cases}$$

Now, $\varepsilon_0 \geq 1$ by assumption and, if $\kappa$ is chosen sufficiently large, then it is easy to verify that $(\log\kappa + i - 1 - \log i)/i$ is bigger than $1/2$ for all $i \geq 1$. Hence, we obtain that

$$\sum_{i=0}^{m_n} \mathbb{P}\big(Poi(n\varepsilon_i) > n\delta_i\big) \leq e^{-n(\kappa\log\kappa - \kappa + 1)} + \kappa n e^{-\kappa n/2}, \tag{16}$$

as $m_n = \lfloor \kappa n \rfloor$. Substituting (15) and (16) in (14), we get

$$\mathbb{P}\Big(\frac{\Phi_n}{n} \notin L_{(K_i,\delta_i)} \Bigm| \frac{\Lambda_n}{n} \in L_{(K_i,\varepsilon_i)}\Big) \le \frac{ne^{-\kappa n}}{1-e^{-1}} + e^{-n(\kappa \log \kappa - \kappa + 1)} + \kappa n e^{-\kappa n/2}.$$

It is clear from this that we can choose $\kappa$ sufficiently large to ensure that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{\Phi_n}{n} \notin L_{(K_i,\delta_i)} \Bigm| \frac{\Lambda_n}{n} \in L_{(K_i,\varepsilon_i)}\Big) \le -\alpha. \tag{17}$$

Finally, combining (12), (13) and (17), we conclude that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{\Phi_n}{n} \notin L_{(K_i,\delta_i)}\Big) \le -\alpha.$$

This concludes the proof of the lemma. □

**Proof of Proposition 5.9.**

The weak topology on the space of finite measures on a Polish space is metrisable [100]), and so it suffices to check sequential compactness. Let $(\mu_n, n \in \mathbb{N})$ be a sequence of finite measures on $E$ satisfying the assumptions of the proposition with respect to a nested sequence of compact sets $K_n$ whose union is equal to $E$, and a sequence $\varepsilon_n$ decreasing to zero. In particular, the measures are bounded; $\mu_n(E) \le \varepsilon_0$ for all $n \in \mathbb{N}$. We want to show that $(\mu_n, n \in \mathbb{N})$ contains a convergent subsequence.

We show that the space of subprobability measures on a compact set $K$ is compact in the weak topology; recall the Banach-Alaoglu Theorem which states that if $X$ is a Banach space, then the closed unit ball of the dual space $X^*$ is compact with respect to the weak star topology. Of course on a compact set, the weak and weak star topologies coincide. Consider the Banach space $C(K)$ of continuous functions on $K$ equipped with the supremum norm. Then by the Riesz Representation Theorem its dual is the space of finite signed measures on $K$. The closed unit ball in this space gives all finite signed measures of mass at most unity. The non-negative elements of this set (the subprobability measures) are a closed subset of this weakly compact set - so they too form a weakly compact set. Hence, by Tychonoff's Theorem applied to the product space $\mathcal{M}_{\le 1}(K) \times [0, \varepsilon_0]$, the space of finite measures on $K$ bounded by an arbitrary constant $\varepsilon_0$ is also compact.

Thus, the measures $\mu_n$ restricted to $K_1$ all lie within a compact set; hence, there is a subsequence $\mu_{11}, \mu_{12}, \ldots$, whose restriction to $K_1$ converges weakly to some $\tilde{\mu}_1 \in \mathcal{M}_+^f(K_1)$. Similarly, the restriction of this subsequence to $K_2$ all lie within a compact set, and contain a convergent subsubsequence $\mu_{21}, \mu_{22}, \ldots$. We can extend this reasoning to $K_3$, $K_4$ and so on.

Formally, denote by $p_n$ the projection from $\mathcal{M}_+^f(E)$ to $\mathcal{M}_+^f(K_n)$ and by $p_{mn}$ the projection from $\mathcal{M}_+^f(K_m)$ to $\mathcal{M}_+^f(K_n)$ for $m \geq n$. Then, we can rewrite the above as:

$$p_1\mu_{1n} \to \tilde{\mu}_1 \in \mathcal{M}_+^f(K_1), \quad p_2\mu_{2n} \to \tilde{\mu}_2 \in \mathcal{M}_+^f(K_2), \quad , \dots ,$$

where the convergence is with respect to the weak topology on the corresponding spaces. Now consider the diagonal sequence $\mu_{kk}$. It is clear from the above that

$$p_n\mu_{kk} \overset{k\to\infty}{\to} \tilde{\mu}_n \in \mathcal{M}_+^f(K_n),$$

for each $n$. A natural question to ask is whether there is a measure $\tilde{\mu} \in \mathcal{M}_+^f(E)$ such that $\tilde{\mu}_n = p_n\tilde{\mu}$ for all $n$. The answer follows from a generalisation of Kolmogorov's Extension Theorem (to consistent finite measures) by Yamasaki [106, Proposition 2.1]; it is affirmative if the measures $\tilde{\mu}_n$ satisfy the consistency conditions $p_{mn}\tilde{\mu}_m = \tilde{\mu}_n$ for all $m > n$. It is straightforward to verify these.

We now show that the diagonal subsequence $\mu_{kk}$ converges weakly to the measure $\tilde{\mu}$ (whose existence we have just shown) in the weak topology on $\mathcal{M}_+^f(E)$, and moreover that the limit $\tilde{\mu}$ is in $L_{(K_n,\varepsilon_n)}$. We start with the latter. As $\tilde{\mu}$ is a finite measure on the Polish space $E$, it is regular (any measurable set can be approximated from within by compacts); therefore, as $K_n$ are compact sets increasing to $E$, $\tilde{\mu}(K_n)$ increases to $\tilde{\mu}(E)$. Hence, for any $m \in \mathbb{N}$,

$$\tilde{\mu}(K_m^c) = \lim_{n\to\infty} \tilde{\mu}(K_n) - \tilde{\mu}(K_m).$$

Now, for any fixed $i > n > m$, $\tilde{\mu}_i$ is the restriction (or projection) of $\tilde{\mu}$ to the set $K_i$, and so

$$\tilde{\mu}(K_n) - \tilde{\mu}(K_m) = \tilde{\mu}_i(K_n) - \tilde{\mu}_i(K_m) \leq \tilde{\mu}_i(K_m^c) \leq \varepsilon_m.$$

The last inequality holds because $\tilde{\mu}_i$ is the weak limit of measures whose mass on $K_m^c$ is bounded by $\varepsilon_m$, and $K_m^c$ is an open set. As this holds for each $n$, we conclude on taking limits that $\tilde{\mu}(K_m^c) \leq \varepsilon_m$. But $m$ was arbitrary, so $\tilde{\mu} \in L_{(K_n,\varepsilon_n)}$.

Next, given $\delta > 0$ and a bounded continuous function $g : E \to \mathbb{R}$, choose $\ell$ large enough that $\varepsilon_\ell \|g\|_\infty < \delta$. Next, pick $m \geq \ell$ large enough that

$$\Big| \int_{K_\ell} g d\mu_{\ell n} - \int_{K_\ell} g d\tilde{\mu}_\ell \Big| \leq \delta \quad \forall\, n \geq m,$$

which is possible since $\mu_{\ell n}$ converges weakly to $\tilde{\mu}_\ell$ as $n$ tends to infinity. Now, $\mu_{n\cdot}$ is a subsequence of $\mu_{\ell\cdot}$ for $n \geq \ell$, so the above inequality also holds for $\int_{K_\ell} g(d\mu_{nn} - d\tilde{\mu}_\ell)$ for all $n \geq m$. Thus, we can write

$$\Big| \int_E g d\mu_{nn} - \int_E g d\tilde{\mu} \Big| \leq \Big| \int_{K_\ell} g(d\mu_{nn} - d\tilde{\mu}_\ell) \Big| + \Big| \int_{K_\ell} g(d\tilde{\mu}_\ell - d\tilde{\mu}) \Big| + 2\|g\|_\infty \varepsilon_\ell,$$

as $\mu_{nn}(K_\ell^c)$ and $\tilde{\mu}(K_\ell^c)$ are both bounded above by $\varepsilon_\ell$. We have just shown that the first integral above is smaller than $\delta$ in absolute value, for all $n \geq m$. The second integral is zero as $\tilde{\mu}_\ell$ is the restriction or projection of $\tilde{\mu}$ to $K_\ell$. The last term is bounded by $2\delta$ by the choice of $\ell$. Thus, we have shown that we can choose $m$ in such a way that

$$\left| \int_E g d\mu_{nn} - \int_E g d\tilde{\mu} \right| \leq 3\delta$$

for all $n \geq m$. As $g$ was an arbitrary bounded continuous function, this proves that $\mu_{nn}$ converges to $\tilde{\mu}$. This completes the proof that $L_{(K_n, \varepsilon_n)}$ is compact.

For the converse, let $\mathcal{K}$ be compact in $\mathcal{M}_+^f(E)$ equipped with the weak topology. As the map $\mu \mapsto \mu(E)$ is continuous (the indicator of $E$ is a bounded continuous function $E \to \mathbb{R}$), its supremum over $\mathcal{K}$ is attained. Denote the supremum by $\varepsilon_0$. Then $\mu(E) = \mu(K_0^c) \leq \varepsilon_0$ for all $\mu \in \mathcal{K}$. Next, we invoke a generalisation of Prokhorov's Theorem by Bogachev [22, Theorem 8.6.2]), which states that the measures in a compact set are uniformly tight. In other words, given $\varepsilon_1 > 0$, we can find a compact subset $K_1$ of $E$ such that $\mu(K_1^c) \leq \varepsilon_1$ for all $\mu \in \mathcal{K}$. Similarly, we can find compact $K_2$ such that $\mu(K_2^c) \leq \varepsilon_2$ for all $\mu \in \mathcal{K}$. Without loss of generality, we can assume that $K_1 \subseteq K_2$; otherwise, re-define $K_2$ as their union. Continuing in the same vein, we obtain a sequence $K_n$ of nested compact sets such that $\mu(K_n^c) \leq \varepsilon_n$ for all $n \in \mathbb{N}$, for all $\mu \in \mathcal{K}$. If their union is not equal to $E$, it can be extended countably to have this property, by the assumption that $E$ is $\sigma$-compact. Now, $\mathcal{K} \subseteq L_{(K_n, \varepsilon_n)}$. □

## 5.3 Proof of LDP for the Queue Occupancy Measure

The proof of Theorem 5.2 is presented in this section. We begin by recalling how the queue occupancy measure is related to the input to the queue. First, we represent the input to the $n^{\text{th}}$ queue as a Cox process on $\mathbb{R} \times \mathbb{R}_+$ by marking each arrival with its service time; the resulting marked point process is a Cox process on $\mathbb{R} \times \mathbb{R}_+$ with stochastic intensity $\Lambda_n \otimes F$. Now, $Q_n(t)$ is equal to the number of points of this Cox process lying in the triangle

$$A_t = \{(s, x) \in \mathbb{R} \times \mathbb{R}_+ : s \leq t, x \geq t - s\}.$$

Furthermore, the queue length process $\{Q_n(t), t \in [a, b]\}$, is determined by the restriction of the above Cox process to the wedge

$$A_{[a,b]} := \bigcup_{t \in [a,b]} A_t,$$

as illustrated in Figure 9. Next, for $u \leq s \leq t$, we will also need to define the truncated sets

$$A_t^u = \{(s,x) \in \mathbb{R} \times \mathbb{R}_+ : u \leq s \leq t, x \geq t - s\}, \quad A_{[s,t]}^u := \bigcup_{x \in [s,t]} A_x^u.$$

Finally, recall that we are interested in the occupancy measure $L_n$, which is defined as the random measure that is absolutely continuous with respect to Lebesgue measure, and has density $Q_n(\cdot)$.

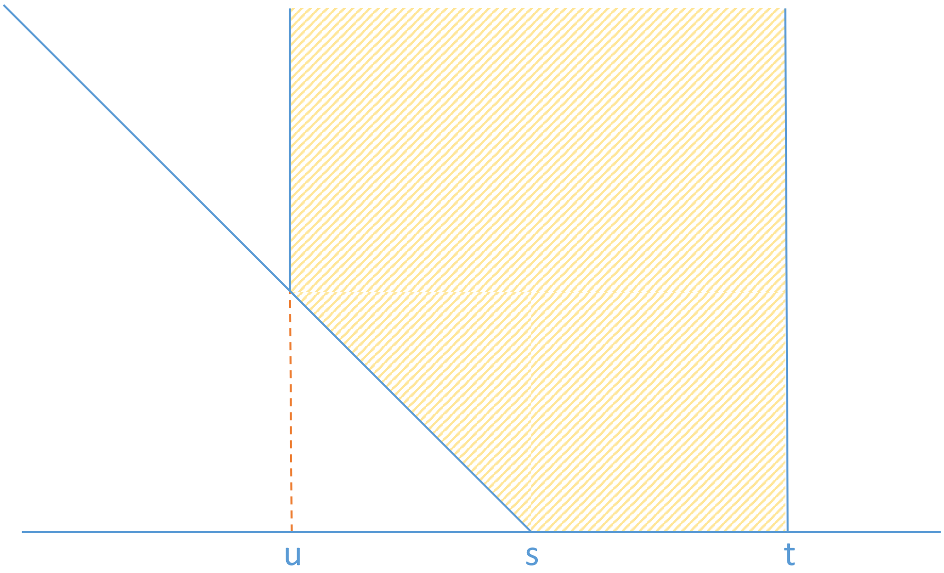

**Figure 9** – The wedge $A_{[s,t]}$ and the truncated wedge $A_{[s,t]}^u$.

Our goal is to prove an LDP for $L_n$, restricted to an arbitrary interval $[a,b]$. We start by establishing an LDP for the scaled directing measures $\frac{\Lambda_n}{n} \otimes F$, restricted to a truncated wedge $A_{[a,b]}^u$, for arbitrary $u < a$; this LDP is in the topology of weak convergence of measures restricted to the truncated wedge. Then, using the projective limit approach described below, we extend this family of LDPs to an LDP on the full wedge $A_{[a,b]}$, in the projective limit topology. However, the queueing map is not continuous in this topology, so we need to strengthen the LDP to the weak topology on the full wedge. We do this by establishing exponential tightness of the measures $\frac{\Lambda_n}{n} \otimes F$ in the weak topology on $A_{[a,b]}$. Next, we invoke Theorem 5.1 to deduce an LDP for the Cox process on $A_{[a,b]}$ with this intensity. Finally, we use weak continuity of the queueing map, and the Contraction Principle (Theorem 2.18), to obtain the LDP for $L_n$. Checking that $L_n$ also satisfies Assumptions [A1]-[A3] is fairly straightforward. The details of all these steps are presented below.

**Lemma 5.11.**

*Fix $u \leq a < b \in \mathbb{R}$ and consider the truncated wedge $A^u_{[a,b]}$. The sequence of random measures $\frac{\Lambda_n}{n} \otimes F\big|_{A^u_{[a,b]}}$, $n \in \mathbb{N}$, satisfy an LDP on $\mathcal{M}^f_+(A^u_{[a,b]})$ equipped with the weak topology, with good rate function*

$$I^u_{[a,b]}(\mu) = \inf\left\{ I_{[a,b]}(\lambda) : \lambda \in \mathcal{M}^f_+([a,b]),\ \mu = (\lambda \otimes F)\big|_{A^u_{[a,b]}} \right\}.$$

*Proof.* Define the map

$$T : \mathcal{M}^f_+([u,b]) \to \mathcal{M}^f_+([u,b] \times \mathbb{R}_+)$$

by $T(\mu) = \mu \otimes F$. We first show that this map is continuous in the weak topology. We will then use the Contraction Principle and the LDP for $\frac{\Lambda_n}{n}\big|_{[u,b]}$ to get the LDP on a rectangle of infinite height, and the Contraction Principle applied to the restriction map to get the LDP on the wedge. As the weak topology is metrisable, we can check continuity along sequences. To this end, consider a sequence of finite measures $\mu_n$ on $[u,b]$ converging weakly to a finite measure $\mu$, and let $g : [u,b] \times \mathbb{R}_+ \to \mathbb{R}$ be bounded and continuous. Define $h : [u,b] \mapsto \mathbb{R}$ by $h(x) = \int_0^\infty g(x,y)dF(y)$. We have

$$\int_{[u,b]\times\mathbb{R}_+} gd(T(\mu_n)) = \int_u^b \Big(\int_0^\infty g(x,y)dF(y)\Big)d\mu_n(x) = \int_u^b h(x)d\mu_n(x),$$

where the first equality follows from Fubini's Theorem, which we can use because

$$\int_{[u,b]\times\mathbb{R}_+} |g|\, d\,(\mu_n \otimes F) \leq \|g\|_\infty\, \mu_n([u,b])F(\mathbb{R}_+) < \infty.$$

If we can show that $h$ is continuous, then it will follow that $\int gd(T(\mu_n))$ converges to $\int gd(T(\mu))$, and, as $g$ was an arbitrary bounded continuous function, that $T(\mu_n)$ converges weakly to $T(\mu)$, thus proving that $T$ is continuous.

Now, to show that $h$ is continuous, fix $\varepsilon > 0$ and $x_0 \in \mathbb{R}$ such that $1 - F(x_0) \leq \varepsilon$. Now $g$ is uniformly continuous on the compact set $[u,b] \times [0,x_0]$ (as $g$ is continuous on a compact), so we can find $\delta > 0$ such that $|g(x,z) - g(y,z)| < \varepsilon$ provided $|x-y| < \delta$. It follows that

$$\begin{aligned}&|h(x) - h(y)|\\ &\leq \int_0^{x_0} |g(x,z) - g(y,z)|dF(z) + \int_{x_0}^\infty |g(x,z)|dF(z) + \int_{x_0}^\infty |g(y,z)|dF(z)\\ &\leq (1 + 2\|g\|_\infty)\varepsilon.\end{aligned}$$

This proves the continuity of $h$, and consequently of $T$.

Next, the map $S$ that restricts finite measures on $[u,b] \times \mathbb{R}_+$ to the wedge $A^u_{[a,b]}$ is trivially continuous, and hence so is the composition $S \circ T$. The claim of the lemma now follows from the assumed LDP for $\frac{\Lambda_n}{n}\big|_{[u,b]}$ and the Contraction Principle (Theorem 2.18). □

The family of LDPs on the truncated wedges $\{A^u_{[a,b]}, u < a\}$ can be extended to an LDP on the full wedge $A_{[a,b]}$ using the Dawson-Gärtner Theorem for projective limits (Theorem 2.26). This yields an LDP in the projective limit topology, which is generated by bounded continuous functions supported on the truncated wedges, $A^u_{[a,b]}$. In order to strengthen this LDP to the weak topology on $A_{[a,b]}$, we need to show exponential tightness of the measures $\frac{\Lambda_n}{n} \otimes F$ in the weak topology. The following lemma is a key ingredient in establishing this. This result is probably already known, but we could not find a reference for it, so we include a proof.

**Lemma 5.12.**
*Suppose $X, X_1, X_2, \ldots$ are identically distributed random variables with arbitrary joint distribution, and suppose $\alpha_i$, $i \in \mathbb{N}$ are non-negative coefficients whose sum is finite, and which we denote by $\alpha$. Then,*

$$W := \sum_{i=1}^{\infty} \alpha_i X_i \leq_{\text{cx}} \alpha X,$$

*where we write $Y \leq_{\text{cx}} Z$ to denote that $Y$ is dominated by $Z$ in the convex stochastic order, i.e., $\mathbb{E}[\phi(Y)] \leq \mathbb{E}[\phi(Z)]$ for all convex functions $\phi$ for which the expectations are defined, possibly infinite.*

*Proof.* By scaling the random variables, we assume $\alpha = 1$ without loss of generality. By Jensen's inequality, the inequality

$$\phi(W(\omega)) = \phi\Big(\sum_{i=1}^{\infty} \alpha_i X_i(\omega)\Big) \leq \sum_{i=1}^{\infty} \alpha_i \phi(X_i(\omega)) = \phi(X_i(\omega)),$$

holds pointwise on the probability space $\Omega$. Taking expectations on both sides yields the result if we can interchange expectation and summation on the right. We can certainly do so (by Tonelli's Theorem) if the functions $\phi$ are non-negative, obtaining

$$\mathbb{E}\left[\phi(W)\right] \leq \mathbb{E}\left[\phi(X_i)\right] = \mathbb{E}\left[\phi(X)\right]$$

as $X$ and $X_i$ are identically distributed. Hence the same is true if the functions $\phi$ are bounded below. To see this, suppose $\phi \geq -k$, where $k \in \mathbb{R}_+$. Then $u(x) := \phi(x) + k$ is non-negative and convex, so

$$\mathbb{E}\left[\phi\left(W\right)\right] + k = \mathbb{E}\left[u\left(W\right)\right] \leq \mathbb{E}\left[u(X)\right] = \mathbb{E}\left[\phi\left(X\right)\right] + k.$$

Now, for any $c \in \mathbb{R}$, the function $\phi_c$ defined by $\phi_c(x) = \max\{c, \phi(x)\}$ is convex and bounded below, so we get

$$\mathbb{E}\Big[\phi_c\Big(W\Big)\Big] \leq \sum_{i=1}^{\infty} \alpha_i \mathbb{E}\Big[\phi_c(X_i)\Big] = \Big(\sum_{i=1}^{\infty} \alpha_i\Big)\mathbb{E}[\phi_c(X)],$$

as the $X_i$ are identically distributed with the same law as $X$. Since $\phi \le \phi_c$, it follows that

$$\mathbb{E}\Big[\phi\Big(\sum_{i=1}^{\infty}\alpha_i X_i\Big)\Big] \le \Big(\sum_{i=1}^{\infty}\alpha_i\Big)\mathbb{E}[\phi_c(X)],$$

for all $c \in \mathbb{R}$. Letting $c$ decrease to $-\infty$ on the right now yields the claim of the lemma. This can be justified by splitting $\phi$ into its positive and negative parts and using the Monotone Convergence Theorem. □

We are now ready to show that the directing measures restricted to a wedge are exponentially tight in the weak topology.

**Proposition 5.13.**
*The sequence of random measures*

$$\left(\left(\frac{\Lambda_n}{n} \otimes F\right)\Big|_{A_{[a,b]}}\right)_{n\in\mathbb{N}}$$

*is exponentially tight in the weak topology.*

*Proof.* We have to show that for every $0 < \alpha < \infty$, there is a compact set $\mathcal{K}_\alpha \subseteq \mathcal{M}_+^f(A_{[a,b]})$ such that

$$\limsup_{n\to\infty}\frac{1}{n}\log\ \mathbb{P}\left(\left(\frac{\Lambda_n}{n} \otimes F\right)\Big|_{A_{[a,b]}} \in \mathcal{K}_\alpha^c\right) < -\alpha. \tag{18}$$

We will use the explicit construction of a weakly compact set of measures given in Proposition 5.9. We seek a nested sequence of compact sets $K_1 \subseteq K_2 \subseteq \ldots \subseteq A_{[a,b]}$, whose union is the wedge $A_{[a,b]}$, and a sequence of positive constants $\varepsilon_0 \ge \varepsilon_1 \ge \ldots$ decreasing to zero, such that

$$\mathbb{P}\left(\left(\frac{\Lambda_n}{n} \otimes F\right)\left(K_i^c\right) > \varepsilon_i\right) \le e^{-n(i+1)\alpha} \quad \forall\, i \ge 0, \tag{19}$$

where we define $K_0$ to be the empty set. If we can find such $K_i$ and $\varepsilon_i$, then the weakly compact set of measures

$$\mathcal{K}_\alpha = \left\{\mu \in \mathcal{M}_+^f(A_{[a,b]}) : \mu(K_i^c) \le \varepsilon_i \;\forall\, i \in \mathbb{N}\right\},$$

satisfies the inequality in (18), thus proving the proposition.

Each of the compact sets $K_i$, $i \ge 1$, will be specified by two real numbers $u_i$ and $h_i$ as shown in Figure 10:

$$K_i = \{[u_i, b] \times [0, h_i]\} \bigcap A_{[a,b]}.$$

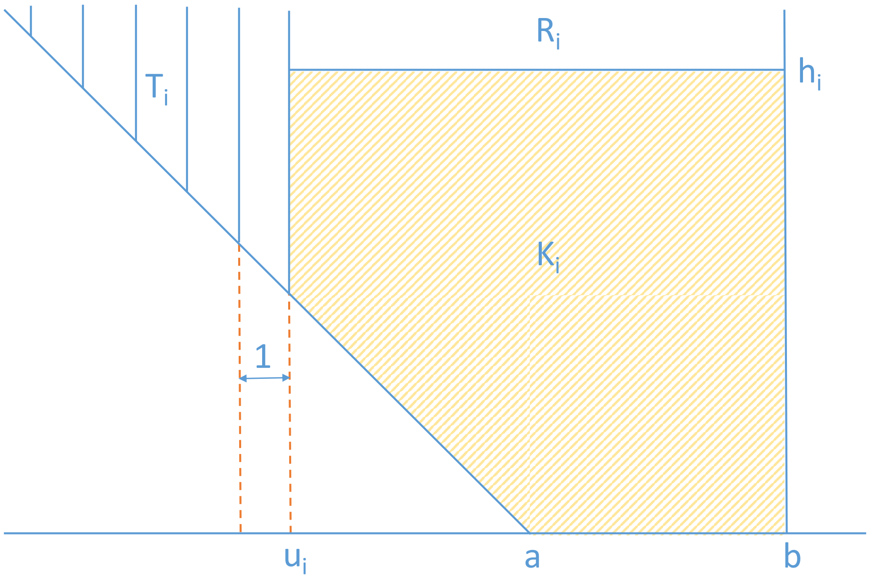

**Figure 10** – The wedge $A_{[a,b]}$ split into a compact set $K_i$, infinite rectangle $R_i$ and infinite triangle $T_i$. The triangle is split into strips of unit width.

We shall write $K_i^c$ to denote the complement of $K_i$ in $A_{[a,b]}$, and we decompose this set into a triangle

$$T_i = \{(s,x) \in \mathbb{R} \times \mathbb{R}_+ : s \le u_i, x \ge a - s\},$$

and a rectangle

$$R_i = \{(s,x) \in \mathbb{R} \times \mathbb{R}_+ : u_i \le s \le b, x \ge h_i\};$$

see Figure 10. Thus, we have

$$\frac{1}{n}(\Lambda_n \otimes F)(K_i^c) = \frac{1}{n}(\Lambda_n \otimes F)(T_i) + \frac{1}{n}(\Lambda_n \otimes F)(R_i). \tag{20}$$

Now, by the translation invariance of $\Lambda_n$ changing the horizontal coordinate makes no difference (though is slightly more convenient to work with and puts us in the setting of Lemma 5.14), so we have

$$(\Lambda_n \otimes F)(T_i) \overset{\mathrm{d}}{=} (\Lambda_n \otimes F)(T^{a-u_i}) \text{ and } (\Lambda_n \otimes F)(R_i) \overset{\mathrm{d}}{=} \big(\Lambda_n \otimes F\big)\big(R^{h_i}_{b-u_i}\big),$$

where $\overset{\mathrm{d}}{=}$ denotes equality in distribution, and the sets $T^\ell$ and $R^h_z$ are defined as

$$\begin{aligned} T^\ell &= \{(t,x) \in \mathbb{R} \times \mathbb{R}_+ : t \le 0, t + x \ge \ell\} \\ R^h_z &= \{(t,x) \in \mathbb{R} \times \mathbb{R}_+ : t \in [0,z], x \ge h\}. \end{aligned} \tag{21}$$

Thus, we obtain from (20) that

$$\begin{aligned} \mathbb{P}\left(\left(\frac{\Lambda_n}{n} \otimes F\right)\left(K_i^c\right) > \varepsilon_i\right) \ \le\ & \mathbb{P}\left(\left(\Lambda_n \otimes F\right)\left(T^{a-u_i}\right) > \frac{n\varepsilon_i}{2}\right) \\ & + \mathbb{P}\left(\left(\Lambda_n \otimes F\right)\left(R^{h_i}_{b-u_i}\right) > \frac{n\varepsilon_i}{2}\right). \end{aligned} \tag{22}$$

We show in Lemma 5.14 that, given $i \in \mathbb{N}$, $\varepsilon_i > 0$ and $\alpha > 0$, we can choose $u_i$ to make $a - u_i$ sufficiently large that

$$\mathbb{P}\Big(\big(\Lambda_n \otimes F\big)\big(T^{a-u_i}\big) > \frac{n\varepsilon_i}{2}\Big) \le e^{-n(i+1)\alpha}, \quad \forall n \in \mathbb{N};$$

to see this, take $\varepsilon = \varepsilon_i/2$ and $\beta = (i+1)\alpha$ in the statement of the lemma. Next, by the same lemma, given $u_i$, and hence $b - u_i$, we can choose $h_i$ sufficiently large to ensure that

$$\mathbb{P}\Big(\big(\Lambda_n \otimes F\big)\big(R^{h_i}_{b-u_i}\big) > \frac{n\varepsilon_i}{2}\Big) \le e^{-n(i+1)\alpha}, \quad \forall n \in \mathbb{N}.$$

Combining these two inequalities with (22), we conclude that for all $i \ge 1$,

$$\mathbb{P}\Big(\big(\Lambda_n \otimes F\big)\big(K_i^c\big) > n\varepsilon_i\Big) \le 2e^{-n(i+1)\alpha}, \quad \forall n \in \mathbb{N}, \tag{23}$$

which is essentially the same as (19). That leaves the case $i = 0$.

The same argument does not work for $K_0$ as we cannot choose this set (that is, we cannot choose a triangle and rectangle arbitrarily high); $K_0$ is the empty set and $K_0^c = A_{[a,b]}$. Instead, we need to show that we can choose $\varepsilon_0$ sufficiently large that

$$\mathbb{P}\Big(\big(\Lambda_n \otimes F\big)\big(A_{[a,b]}\big) > n\varepsilon_0\Big) \le e^{-n\alpha}, \quad \forall n \in \mathbb{N}. \tag{24}$$

We first note that $A_{[a,b]} \subset T_0 \cup \{[a-\ell, b] \times \mathbb{R}_+\}$, where $0 < \ell < a$ and

$$T_0 = \{(t,x) \in \mathbb{R} \times \mathbb{R}_+ : t \le a - \ell, t + x \ge a\}.$$

Hence

$$(\Lambda_n \otimes F)(A_{[a,b]}) \le (\Lambda_n \otimes F)(T_0) + \Lambda_n([a-\ell, b]).$$

Moreover, by translation invariance of $\Lambda_n$, we have

$$(\Lambda_n \otimes F)(T_0) \overset{\mathrm{d}}{=} (\Lambda_n \otimes F)(T^\ell),$$

where $T^\ell$ is defined in (21). Using Lemma 5.14 below, we conclude that we can choose $\ell$ sufficiently large that

$$\mathbb{P}\Big(\big(\Lambda_n \otimes F\big)\big(T_0\big) > n\Big) \le e^{-n\alpha}, \quad \forall n \in \mathbb{N}. \tag{25}$$

We also see from the proof of Lemma 5.14 that $\Lambda_n([a-\ell, b])$ is dominated, in the increasing convex order, by $\lceil \ell + b - a \rceil \Lambda_n([0,1])$; in particular,

$$\mathbb{E}\left[e^{\theta\Lambda_n([a-\ell,b])}\right] \le \mathbb{E}\left[e^{\theta(\ell+1+b-a)\Lambda_n([0,1])}\right] = \exp\Big(\psi_n\big(n\theta(\ell+1+b-a)\big)\Big),$$

where $\psi_n$ is defined in Assumption [A3]. By [A3], for given $a, b, \ell$, $\psi_n(n\theta(\ell+1+b-a))/n$ is bounded, for $\theta$ in a neighbourhood of the origin, uniformly in $n$, i.e, there exist constants $\theta, \delta > 0$ such that $\psi_n(n\theta) \le n\delta$ for all $n \in \mathbb{N}$. Consequently, by Markov's inequality,

$$\mathbb{P}\left(\Lambda_n([a-\ell, b]) \ge n(\varepsilon_0 - 1)\right) \le e^{-n\theta(\varepsilon_0-1)+n\delta}, \quad \forall n \in \mathbb{N}.$$

Clearly, we can choose $\varepsilon_0$ large enough to ensure that

$$\mathbb{P}\left(\Lambda_n([a-\ell, b]) \ge n(\varepsilon_0 - 1)\right) \le e^{-n\alpha}, \quad \forall n \in \mathbb{N}.$$

Combining the above equation with (25), we see that the inequality in (24) holds, up to a factor of two. This completes the proof that the inequality in (19) holds for all $i \ge 0$, up to a factor of two on the RHS. Now, using the union bound over $i$, we get

$$\mathbb{P}\Big(\exists\, i \ge 0 : \Big(\frac{\Lambda_n}{n} \otimes F\Big)\Big(K_i^c\Big) > \varepsilon_i\Big) \le \sum_{i=0}^{\infty} e^{-n(i+1)\alpha} = \frac{e^{-n\alpha}}{1-e^{-n\alpha}} \le 2e^{-n\alpha},$$

from which (18) is immediate, given the definition of $\mathcal{K}_\alpha$. This completes the proof of the proposition. □

**Lemma 5.14.**
*Let $\beta > 0$ be a given constant. For $\ell, h, z > 0$, let the triangle $T_\ell$ and the rectangle $R_z^h$ be defined as in (21). Then, we have the following:*

1. *Given $\varepsilon > 0$, we can choose $\ell$ sufficiently large that*

$$\mathbb{P}\Big(\big(\Lambda_n \otimes F\big)\big(T^\ell\big) > n\varepsilon\Big) \le e^{-n\beta}, \quad \forall n \in \mathbb{N}.$$

2. *Given $z > 0$ and $\varepsilon > 0$, we can choose $h$ sufficiently large that*

$$\mathbb{P}\Big(\big(\Lambda_n \otimes F\big)\big(R_z^h\big) > n\varepsilon\Big) \le e^{-n\beta}, \quad \forall n \in \mathbb{N}.$$

*Proof.* Fix an $\ell \in \mathbb{R}$. By splitting the triangle $T^\ell$ into vertical strips of unit width, we see that

$$\big(\Lambda_n \otimes F\big)\big(T^\ell\big) \le \sum_{k=0}^{\infty} \Lambda_n\big([-k-1, -k]\big)\overline{F}(\ell + k).$$

Now, by translation invariance of $\Lambda_n$, the random variables $\Lambda_n\big([-k-1,-k]\big)$ are identically distributed for all $k$. Moreover, the sum of the coefficients $\overline{F}(\ell+k)$ can be bounded as follows:

$$\sum_{k=0}^{\infty} \overline{F}(\ell + k) \le c_\ell := \int_{\ell-1}^{\infty} \overline{F}(x)dx = \mathbb{E}[S\mathbb{1}(S \ge \ell - 1)],$$

where $S$ denotes a random variable with the distribution $F$ of the service time, and $\mathbb{1}(E)$ denotes the indicator of the event $E$. This last expectation is finite by the assumption that the service time has finite mean. Hence, invoking Lemma 5.12, we obtain that

$$\big(\Lambda_n \otimes F\big)\big(T^\ell\big) \le_{icx} c_\ell \Lambda_n\big([0,1]\big),$$

where, for random variables $X$ and $Y$, we say that $X$ is dominated by $Y$ in the increasing convex order, written $X \leq_{icx} Y$, if $\mathbb{E}[\phi(X)] \leq \mathbb{E}[\phi(Y)]$ for all increasing convex functions $\phi$. Applying this bound to the increasing convex function $\phi(x) = e^{\theta x}$ for arbitrary $\theta > 0$, and using Markov's inequality, we get, for any $\varepsilon > 0$,

$$\mathbb{P}\Big(\big(\Lambda_n \otimes F\big)\big(T^\ell\big) \geq \frac{n\varepsilon}{2}\Big) \leq e^{-n\theta\varepsilon/2}\mathbb{E}\left[e^{\theta c_\ell \Lambda_n([0,1])}\right] = \exp\Big(-\frac{n\theta\varepsilon}{2} + \psi_n\big(n\theta c_\ell\big)\Big),$$

where the function $\psi_n$ was defined in Assumption [A3]. As $\theta > 0$ is arbitrary, it is convenient to rewrite the above inequality (replacing $\theta$ by $\theta/c_\ell$) as

$$\log\mathbb{P}\Big(\big(\Lambda_n \otimes F\big)\big(T^\ell\big) \geq \frac{n\varepsilon}{2}\Big) \leq -\frac{n\theta\varepsilon}{2c_\ell} + \psi_n(n\theta), \text{ where } c_\ell = \mathbb{E}[S\mathbb{1}(S \geq \ell - 1)]. \tag{26}$$

Now, by Assumption [A3], there exist positive constants $\delta$ and $\theta$ such that $\psi_n(n\theta) \leq n\delta$, uniformly in $n$. Morever, as $\mathbb{E}[S]$ is finite by Assumption [A4], it follows that $c_\ell$ tends to zero as $\ell$ tends to infinity. Hence, we see from (26) that, given $i \in \mathbb{N}$ and $\beta, \varepsilon > 0$, we can choose $\ell$ sufficiently large, and consequently $c_\ell$ sufficiently small, to ensure that

$$\mathbb{P}\left(\big(\Lambda_n \otimes F\big)\big(T^\ell\big) \geq n\varepsilon\right) \leq e^{-n\beta} \quad \forall\ n \in \mathbb{N}. \tag{27}$$

This completes the proof of the first claim of the lemma.

The proof of the second claim is very similar. We show that

$$\big(\Lambda_n \otimes F\big)\big(R^h_{b-a}\big) \leq_{icx} \lceil b - a\rceil \overline{F}(h)\Lambda_n([0,1]),$$

and apply Markov's inequality to the exponential of the random variable on the RHS. The details are omitted. □

Note there was nothing particularly special about triangles and rectangles (though they are relevant to the queueing applications) and the results would certainly extend to other similar sets. We now have all the ingredients required to establish an LDP for the scaled intensity measures $(\Lambda_n \otimes F)/n$, on the wedge $A_{[a,b]}$.

**Proposition 5.15.**
*Suppose that $\Lambda_n, n \in \mathbb{N}$ is a sequence of random measures satisfying Assumptions [A1]-[A3] and $F$ satisfies [A4]. Fix an interval $[a,b] \subset \mathbb{R}$. The sequence of random measures $\big(\frac{\Lambda_n}{n} \otimes F\big)\big|_{A_{[a,b]}}$, $n \in \mathbb{N}$, satisfy an LDP on $\mathcal{M}_+^f(A_{[a,b]})$ equipped with the weak topology, with good rate function*

$$I_{[a,b]}(\nu) = \sup_{u \leq a} I^u_{[a,b]}\big(\nu\,\big|_{A^u_{[a,b]}}\big), \quad \nu \in \mathcal{M}_+^f([a,b]).$$

*Proof.* We will use the Dawson-Gärtner Theorem for projective limits (Theorem 2.26). Letting

$$J := \left\{ A^u_{[a,b]} : u \in (-\infty, a) \right\},$$

it is clear that the collection $(J, \subseteq)$ of truncated wedges $A^u_{[a,b]}$ equipped with set inclusion is totally ordered, and hence also right-filtering. The set is indexed by $u$, and we will use $u$ to denote the element $A^u_{[a,b]}$, to simplify notation. Denote by $\mathcal{Y}_u$ the space $\mathcal{M}^f_+(A^u_{[a,b]})$ of finite measures on $A^u_{[a,b]}$, equipped with the weak topology.

If $t \le u$, i.e., $A^u_{[a,b]} \subseteq A^t_{[a,b]}$ (note that the order in the projective system reverses inequalities from the order on the real line), define the projection $p_{ut} : \mathcal{Y}_t \to \mathcal{Y}_u$ by the restriction of a measure on $A^t_{[a,b]}$ to the subset $A^u_{[a,b]}$. It is clear that this map is continuous, and also that the projections satisfy the consistency condition $p_{us} = p_{ut} \circ p_{ts}$ for $s \le t \le u$. Thus, $(\mathcal{Y}_u, p_{ut})_{t \le u}$ constitute a projective system. We can identify $\mathcal{M}^f_+(A_{[a,b]})$ with the projective limit, with canonical projections

$$p_u : \mathcal{M}^f_+(A_{[a,b]}) \to \mathcal{M}^f_+(A^u_{[a,b]})$$

defined as the restriction of a measure from the full wedge $A_{[a,b]}$ to its truncation $A^u_{[a,b]}$. These are clearly continuous in the weak topology.

Now, by Lemma 5.11, the projections

$$\left( \frac{\Lambda_n}{n} \otimes F \right) \Big|_{A^u_{[a,b]}} = p_u \left( \left( \frac{\Lambda_n}{n} \otimes F \right) \Big|_{A_{[a,b]}} \right), \ n \in \mathbb{N},$$

satisfy an LDP for each $u \in (\infty, a)$, with rate function $I^u_{[a,b]}$. Hence, by the Dawson-Gärtner Theorem, the sequence of measures $\left( \frac{\Lambda_n}{n} \otimes F \right)\big|_{A_{[a,b]}}$, $n \in \mathbb{N}$, satisfies an LDP in the projective limit topology, with good rate function

$$I_{[a,b]}(\nu) = \sup_{u \le a} I^u_{[a,b]}\big(\nu \,|_{A^u_{[a,b]}}\big), \quad \nu \in \mathcal{M}^f_+([a,b]).$$

Moreover, by Proposition 5.13, the measures $\left( \frac{\Lambda_n}{n} \otimes F \right)\big|_{A_{[a,b]}}$ are exponentially tight in the weak topology on $\mathcal{M}^f_+(A_{[a,b]})$. Hence, by Lemma 2.17, we obtain that the LDP holds in the weak topology. Exponential tightness also implies goodness of the rate function in the weak topology by Lemma 2.16. □

Next, we show the weak continuity of the queueing map, which is the prelude to obtaining the LDP for the queue occupancy measure. For a measure $\nu \in \mathcal{M}^f_+(A_{[a,b]})$, and $t \in [a,b]$, we define $Q^\nu(t) = \nu(A_t)$, where we recall that $A_t = A_{[t,t]}$ is the set

$$\{(s,x) \in \mathbb{R} \times \mathbb{R}_+ : s \le t, s + x \ge t\}.$$

The interpretation is that, if $\nu$ is a counting measure representing the marked arrival process into an infinite-server queue, where each arrival is marked with its service time, then $Q^\nu(t)$ denotes the number of customers in the queue at time $t$. Let $L(\nu)$ denote the measure on $[a, b]$ which is absolutely continuous with respect to Lebesgue measure, and has density $Q^\nu(\cdot)$; let $L$ denote the map from $\mathcal{M}_+^f(A_{[a,b]})$ to $\mathcal{M}_+^f([a, b])$ which takes $\nu$ to $L(\nu)$.

We want an explicit characterisation of the map $L$. We will describe $L(\nu)$ through its action on the dual space $C_b([a, b])$ of bounded, continuous functions on $[a, b]$, i.e., by specifying $\int_a^b g(t)dL(\nu)(t)$ for all $g \in C_b([a, b])$. By the Riesz Representation Theorem, $L(\nu)$ is uniquely determined by these integrals. From the description above, we have

$$\begin{aligned}\int_a^b g(t)dL(\nu)(t) &= \int_a^b g(t)Q^\nu(t)dt = \int_{t=a}^b g(t)\nu(A_t)dt \\ &= \int_{A_{[a,b]}} \Big( \int_{\max\{a,s\}}^{\min\{s+x,b\}} g(t)dt \Big) \nu(ds \times dx). \qquad (28)\end{aligned}$$

The last equality is obtained by interchanging the order of integration (which can be done by Fubini's Theorem), noting that an area element at $ds \times dx$ contributes to $\nu(A_t)$ for each $t$ between $\max\{a, s\}$ and $\min\{s + x, b\}$.

**Lemma 5.16.**
*The map $L : \mathcal{M}_+^f(A_{[a,b]}) \to \mathcal{M}_+^f([a, b])$, defined by (28) via the Riesz Representation Theorem, is continuous with respect to the weak topology on each of these sets.*

*Proof.* The weak topology on the space of finite measures on a Polish space is metrisable [100], so we can check continuity of $L$ along sequences. Suppose $\nu_n, n \in \mathbb{N}$ converge to $\nu$ in the weak topology on $\mathcal{M}_+^f(A_{[a,b]})$. Let $g : [a, b] \to \mathbb{R}$ be a bounded, continuous function. We have by (28) that

$$\begin{aligned}\int_a^b g(t)dL(\nu_n)(t) &= \int_{A_{[a,b]}} h(s, x)\nu_n(ds \times dx), \\ \text{where} \qquad h(s, x) &= \int_{\max\{a,s\}}^{\min\{s+x,b\}} g(t)dt. \qquad (29)\end{aligned}$$

It is clear that the the function $h : A_{[a,b]} \to \mathbb{R}$ is bounded and continuous, and so the RHS above converges to

$$\int_{A_{[a,b]}} h(s, x)\nu(ds \times dx).$$

This completes the proof of the lemma. □

We are now ready to prove the main result.

**Proof of Theorem 5.2.**
Let $\Phi_n$ denote the Cox process of arrivals into the $n^{\text{th}}$ queue, marked with their service times. Fix $[a,b] \subset \mathbb{R}$. By Proposition 5.15, the sequence of measures $\left(\frac{\Lambda_n}{n} \otimes F\right)\Big|_{A_{[a,b]}}$, satisfy an LDP on $\mathcal{M}_+^f(A_{[a,b]})$ equipped with the weak topology, with good rate function $I_{[a,b]}$ given therein. Hence, by Theorem 5.1, the sequence of Cox point measures $\frac{\Phi_n}{n}\Big|_{A_{[a,b]}}$ also satisfies an LDP on $\mathcal{M}_+^f(A_{[a,b]})$ equipped with the weak topology, with good rate function $\mathcal{I}_{[a,b]}$ given by

$$\mathcal{I}_{[a,b]}(\mathbf{0}) = \inf_{\lambda} \left\{ I_{[a,b]}(\lambda) + \lambda\left(A_{[a,b]}\right) \right\}, \tag{30}$$

where $\mathbf{0}$ denotes the zero measure, whereas, for $\mu \not\equiv \mathbf{0}$,

$$\begin{aligned} \mathcal{I}_{[a,b]}(\mu) &= \inf_{\lambda} \Big\{ I_{[a,b]}(\lambda) + I_{Poi}\left(\mu(A_{[a,b]}), \lambda(A_{[a,b]})\right) \\ &\qquad + \mu(A_{[a,b]}) H\Big( \frac{\mu}{\mu(A_{[a,b]})} \,\Big|\, \frac{\lambda}{\lambda(A_{[a,b]})} \Big) \Big\}, \end{aligned} \tag{31}$$

where $H$ and $I_{Poi}$ are defined in the statements of Theorem 5.5 and Lemma 5.6 respectively.

Now, the queue occupancy measures $L_n$ are given by $L_n/n = L(\Phi_n/n)$, where the map $L$ is defined by (28), and is linear and weakly continuous by Lemma 5.16. Hence, by the Contraction Principle (Theorem 2.18), the sequence of measures $L_n/n$ satisfies an LDP on $\mathcal{M}_+^f([a,b])$ equipped with the weak topology, with good rate function

$$J_{[a,b]}(\nu) = \inf \left\{ \mathcal{I}_{[a,b]}(\mu) : L(\mu) = \nu \right\}, \tag{32}$$

where the infimum of an empty set is defined to be $+\infty$. Thus, the sequence $L_n$ satisfies Assumption [A2]. The measures $L_n$ inherit translation invariance from $\Lambda_n$ via $\Lambda_n \otimes F$ and $\Phi_n$, while finiteness of the mean follows easily from that of $\lambda$ (the mean arrival intensity) and of the service time distribution. Thus, [A1] is verified. It remains to check [A3].

Observe that, analogous to (29), we have

$$\begin{aligned} L_n([0,1]) &= (L(\Phi_n))([0,1]) \\ &= \int_{(s,x)\in A_{[0,1]}} \big(\min\{s+x,1\} - \max\{s,0\}\big)\Phi_n(ds \times dx) \\ &\le \Phi_n(A_{[0,1]}). \end{aligned}$$

But, conditional on $\Lambda_n \equiv \boldsymbol{\lambda}$, $\Phi_n([0,1])$ is a Poisson random variable with mean $(\boldsymbol{\lambda} \otimes F)(A_{[0,1]})$. Hence, we have for $\theta \ge 0$ that

$$\mathbb{E}\left[e^{\theta L_n([0,1])}\right] \le \mathbb{E}\Big[\exp\Big(\left(e^{\theta} - 1\right)\left(\Lambda_n \otimes F\right)\left(A_{[0,1]}\right)\Big)\Big].$$

Moreover, it can be shown by splitting $A_{[0,1]}$ into vertical strips of unit width and invoking Lemma 5.12, as in the proof of Lemma 5.14, that

$$(\Lambda_n \otimes F)(A_{[0,1]}) \leq_{icx} (1 + \mathbb{E}[S])\Lambda_n([0,1]),$$

where $\mathbb{E}[S]$ denotes the mean service time, and is finite by Assumption [A4]. Hence, we obtain for $\theta \geq 0$ that

$$\mathbb{E}\left[e^{\theta L_n([0,1])}\right] \leq \mathbb{E}\Big[\exp\Big(\big(e^{\theta} - 1\big)\big(1 + \mathbb{E}[S]\big)\big(\Lambda_n([0,1])\big)\Big)\Big].$$

By Assumption [A3], there is a neighbourhood of 0 on which

$$\frac{\psi_n(n\eta)}{n} = \frac{1}{n}\log \mathbb{E}\left[e^{\eta\Lambda_n(0,1)}\right]$$

is bounded, uniformly in $n$. Setting $\eta = (e^{\theta} - 1)(1 + \mathbb{E}[S])$, we obtain uniform boundedness of

$$\frac{1}{n}\log \mathbb{E}\left[e^{\theta L_n([0,1])}\right]$$

for $\theta \geq 0$ sufficiently small, uniformly in $n$. Boundedness is automatic for $\theta < 0$ as the random variables $L_n([0,1])$ are non-negative. Thus, the sequence of measures $L_n$ satisfy [A3] as well. This completes the proof of the theorem. □

# 6 Concluding Remarks

In this final section we make some brief remarks on the models and results of this thesis and discuss some possible extensions. There are a number of obvious extensions that one could make to the gene regulatory network model that would make it more realistic. Firstly, rather than assuming that RNA molecules are transcribed at a constant rate, one could model the number of active genes explicitly (as is done in [83]). Rather than having Markovian arrivals, the RNA queue could have a Coxian arrival process too. But this still falls within the scope of chapter 5, so we have implicitly already handled this case.

Throughout this thesis we have ignored the role of feedback in the system. In practice proteins can be autocatalytic or autoregulatory, reacting to the concentration levels in the cell. If there are too few proteins of a certain type a signal is sent to increase the expression levels of the relevant genes upstream in the gene regulatory network. Conversely if there are too many or sufficiently many, then feedback in the system curtails the action of the responsible genes. In practice there are time lag effects in both cases. Introducing feedback into the model makes the mathematical analysis substantially more complicated. Incorporating feedback would involve working with networks that are more complicated than the feedforward tandem considered in this thesis in which influence only propagates forwards. For an example of a three stage model of a gene regulatory network that incorporates negative feedback, see [39]. Understanding the role of feedback is of great biological interest and to the best of our knowledge there are very few mathematical results in this direction. So this remains a big open problem.

One could study different biochemical reaction networks which might admit other network topologies. For instance if there were two types of molecules and the first underwent a reaction in which it was consumed and transformed into a copy of the second, then one could use a traditional tandem. Upon completing service at the first queue a customer could be routed to the second. Other biological motifs would require different network structures. Another possible extension is keep track of multiple types of proteins by using a queue with multiple distinguishable customer classes. The multiclass queueing framework has been used before as a model in biological settings, see for instance [73, 74]. The different job size distributions could be used to capture the distinct degradation pathways.

The main results of chapter 4 relied upon the assumption that the subgenerator matrix parameterising the Phase-type service time distribution be diagonalisable. This may seem like an inconvenience. But in any practical sense this is not a restrictive assumption. Matrices with repeated eigenvalues are atypical (both in a topological sense and in terms of having zero Lebesgue measure). In the application considered this would correspond to certain reaction or degradation rates being identical - which of course they are not. Although we do not have any general results relating to the non-diagonalisable case, for a given Phase-type distribution with such a subgenerator matrix, the same calculations can still be performed. Instead of diagonalising, it may help to decompose the matrix into its Jordan normal form. The resulting autocovariance functions are not simply mixtures of decaying exponentials, but also have polynomial prefactors of the lag and model parameters. The dominant behaviour, however, is still that the stationary autocovariance decays exponentially fast in the lag.

An alternative way to view an infinite server queue with a Phase-type service time distribution is to consider it as a network rather than a single facility. Each queue in the network would correspond to a phase of the Phase-type distribution - so that the $i^{th}$ subsystem counts the number of customers in phase $i$ of service for $i \in \{1, 2, ..., \ell\}$. The total number of customers in the system is then just the sum of the number of customers in each subsystem. Each subsystem is simpler than the original system, in that service times are Exponential. Analysing this network is still not entirely straightforward due to the Coxian arrival process - in particular assumptions on exogenous arrivals means that this falls outside the framework of BCMP networks [10].

One of the advantages of Phase-type distributions is that they are dense in the set of non-negative probability distributions. This means that an arbitrarily exotic service time distribution can in principle be arbitrarily well approximated by some Phase-type distribution. Finding a close Phase-type distribution in practice may not be straightforward and may require an impractically large number of phases. A common approach is to match the first few moments while trying to keep the number of phases low. This may not work well in practice, for instance if one wants to fit a Pareto distribution, where most moments do not exist. In particular, it may be hard to capture the tail behaviour. In that context it may make more sense to (roughly) minimise a metric on probability measures, or something similar like the relative entropy. One would try to optimise (over a class of approximating distributions with some constraint on the number of phases) the distance to a fixed target distribution. For more about various methods related to the fitting problem see [63] and references therein.

The proof of the main results of chapter 5 rely crucially on the assumption that the underlying Polish space on which the point process is defined is $\sigma$-compact. It is not clear to us whether the theorems still hold without this requirement, but it seems that one would require a substantially different proof. Of course for the queueing application, where the point process lives on a subset of $\mathbb{R}^2$, this condition is trivially satisfied. An LDP for the empirical measure of a Cox process on a Polish space is proved in [91], but with respect to a coarser topology and under different assumptions. In particular, it is neither assumed that the space is $\sigma$-compact, nor that the intensity measure is finite. So perhaps one can dispense with these assumptions in exchange for others.

The LDP for the biological application provides an approximation of the probability of rare events. But it is an asymptotic result, so will only begin to be an accurate approximation when $n$ becomes sufficiently large. If copy counts are very small, say just a handful of proteins of some type, then this approximation will be poor. If the number of copies is very large, say thousands or millions, then one might as well model the system deterministically by an ODE or PDE as the Law of Large Numbers will kick in. So the approximation is most useful on a mesoscopic scale. In practice, solving the variational problem to explicitly compute the rate function is hard. In particular this would involve optimising over spaces of measures, which are very large. If one could show that the rate function is convex, then numerical methods stand a better chance of approximately solving the optimisation problem. In some sense the results of this chapter are not that practical in terms of the biological applications, but we considered them worth pursuing because the mathematics is interesting in its own right.

It is worth noting that the large deviations asymptotics for the empirical measure of a Cox point process may be of independent interest. There are for instance many models in stochastic geometry that are based on an underlying Poisson point process, whose fluctuations and rare event behaviour have been studied (see section 3.8 for details). Many applications are given in [5] and references therein, and [6] touches on many examples in the field of wireless networks. It may be interesting to study similar models based on an underlying Cox point process.

In chapter 5 we investigated rare event behaviour of gene regulatory networks. This tells us about the probability of seeing highly atypical molecular count data. One could also ask about the much more commonly observed finer fluctuations about the typical behaviour of the system. The inherent stochasticity of biochemical reaction networks means that copy numbers will be constantly fluctuating around their mean values. But this begs two pertinent questions. How big are typically observed fluctuations (relative to the average copy counts)? And what distribution do these fluctuations follow? We suspect the fluctuations turn out to be Gaussian, with the usual $\sqrt{n}$ magnitude (where the number of proteins, say, is of order $n$).

Like in the large deviations case, we would begin by considering the fluctuations of a spatial Cox point process, before proceeding to describe the fluctuations of the associated $Cox/G/\infty$ queue. More precisely we would consider the empirical measure of the Cox process, appropriately centered and scaled. The centering isolates the fluctuations and the normalisation makes sure that we work on the appropriate scale to observe them.

The centered CLT scaled empirical measure lives on the space of finite signed measures. We would like to show that this sequence of measures converges weakly to a Gaussian random field (in a sense we hope to make precise in future work) given that the sequence of directing measures do so too. There are a couple of technical obstructions that make this difficult. Firstly, the space of finite signed measures equipped with the total variation norm is not a separable Banach space. Futher, the weak topology on the space of signed measures on a topological space is not metrisable. Rather than proving a bona fide Central Limit Theorem for spatial Cox point processes on a Polish space, it may be easier to show that the fluctuations are Gaussian in nature. Specifically, that the empirical measure evaluated on a broad class of test functions converges weakly to a family of real valued Gaussian random variables (a different one for each function), given that the intensity measure does so. Formally we would need to assume the following.

**Assumption 6.1.**
*Let $E$ be a Polish space, and let $\Lambda_n \in \mathcal{M}_+^f(E)$, $n \in \mathbb{N}$ be a sequence of random finite Borel measures. Suppose that there exists $\overline{\lambda} \in \mathcal{M}_+^f(E)$ such that, for all $f \in C_b(E)$, we have*

$$U_n(f) := \frac{\Lambda_n(f) - n\overline{\lambda}(f)}{\sqrt{n}} \Rightarrow U_f \sim \mathcal{N}(0, \sigma_f^2) \tag{33}$$

*where $\sigma_f^2 \in \mathbb{R}_+$.*

Let $\Phi_n$ denote the Cox point process on $E$ directed by $\Lambda_n$ and denote the centered, CLT scaled empirical measure integrated against $f \in C_b(E)$ by

$$Y_n(f) := \frac{\Phi_n(f) - n\overline{\lambda}(f)}{\sqrt{n}}. \tag{34}$$

We would like to show that $Y_n(f), n \in \mathbb{N}$ converges weakly to a real valued Gaussian distribution (a different one for each $f$).

Further work will involve trying to find the right language (space, topology, etc.) to state and prove a genuine functional CLT. For instance one could view the CLT-scaled empirical measure as a random element of the space of tempered distributions (the dual of the Schwartz space) as in [57], or take the empirical process perspective (described in section 3.6) and try to find a Donsker class of test functions. An earlier version of this thesis contained a chapter with partial progress towards an FCLT in each of these settings. This removed chapter was the main reason for including a discussion of empirical processes and random tempered distributions in the literature review.

# 7 Appendix: Code for Simulations

This is the code used to generate the plots in section 4.5.

```
##queue 1 - M/H2/Infty

n1 <- 1e3 #number of customers in 1st queue
m <- 1e3 #number of time points
lam1 <- 0.9 #arrival rate into 1st queue
mu11 <- 3 #service rate of first phase in parallel of 1st queue
mu12 <- 1.5 #service rate of second phase in parallel of 1st queue
alpha11 <- 0.7 #routing probability
alpha12 <- 1-alpha11 #routing probability

par(mfcol=c(3,2)) #grid of plots

arr.t1 <- rep(NA,n1)
for (j in 1:n1){
        if(j==1){
                arr.t1[j] <- rexp(1,rate=lam1)
                next
        }
        arr.t1[j] <- arr.t1[j-1] + rexp(1,rate=lam1)
} #arrival times for queue 1

U <- runif(n1,min=0,max=1) #flip to decide which parallel Exponential branch is service time
rate1s <- U<alpha11 #all flips less than alpha11
rate2s <- !rate1s #all other flips

srv.t1 <- rep(NA,n1) #service times H2 for queue 1
srv.t1[rate1s] <- rexp(sum(rate1s),rate=mu11)
srv.t1[rate2s] <- rexp(sum(rate2s),rate=mu12)

dep.t1  <- srv.t1 + arr.t1 #departure times for queue 1

plot(arr.t1,srv.t1,pch=4,cex=0.5,xlab="",ylab="")
title("Graphical␣Point␣Process␣of␣Queue␣1",xlab="Time",ylab="Service␣Requirement",cex.main=3,cex.lab=1.7)

T <- 1000 #end time
latt1 <- seq(from=0,to=T,length.out=m) #discretised times

cum.arr1 <- sapply(1:m, function(j) sum(arr.t1 <= latt1[j])) #cumulative arrivals up to t in queue 1
cum.dep1 <- sapply(1:m, function(j) sum(dep.t1 <= latt1[j])) #cumulative departures up to t in queue 1
```

```

q.length1 <- sapply(1:m, function(j) sum(arr.t1 <= latt1[j]) - sum(dep.t1 <= latt1[j])) #length of queue 1 at time t
q.length.eq1 <- q.length1[floor(m/5) : (m-floor(m/5))] #length of queue 1 with burn in and burn out removed (equilibrium)

latt1chop <- latt1[floor(m/5) : (m-floor(m/5))] #lattice without first and last fifth

plot(latt1chop,q.length.eq1,typ="l",xlab="",ylab="")
title("Queue 1 Length without Edge Effects (Stationary)",xlab="Time",ylab="Queue Length",cex.main=3,cex.lab=1.7)
acf(q.length.eq1,ci=0,main="",xlab="",ylab="",type="correlation",typ="b")
title("Autocorrelation Function of Queue 1 Length",xlab="Lag",ylab="ACF",cex.main=3,cex.lab=1.7)

##queue 2 - Cox/H2/Infty

n2 <- 300 #number of customers ever through queue 2 (significantly less than that of queue 1 as otherwise arrival rate drops to 0 eventually)
lam2 <- 0.9 #arrival rate constant into queue 2
mu21 <- 4 #service rate of first phase in parallel of 2nd queue
mu22 <- 4/3 #service rate of second phase in parallel of 2nd queue
alpha21 <- 0.4 #routing probability
alpha22 <- 1-alpha21 #routing probability

arr.t2 <- rep(NA,n2) #arrival times into queue 2
non_zero_q <- which(!(q.length1==0))
for (j in 1:n2){
        if(j==1){
                arr.t2[j] <- rexp(1,rate=lam2)
                next
        }
        above <- which(latt1 >= arr.t2[j-1])
        next.ind <- intersect(above,non_zero_q)[1]
        arr.t2[j] <- arr.t2[j-1] + latt1[next.ind]-latt1[above[1]] + rexp(1,rate=lam2*q.length1[next.ind])
}
#1st arrival time doesn't matter as we'll exclude burn in
#look at times when 1st queue is non-empty (otherwise arrival rate 0)
#find next time queue 1 is empty, then sample Exp(lam2*N1(t)) for next arrival plus time spent waiting for arrival rate to return above 0

U <- runif(n2,min=0,max=1) #decide H2 service times again
rate1s <- U<alpha21
rate2s <- !rate1s

```

```r
srv.t2 <- rep(NA,n2) #service times H2 for queue 2
srv.t2[rate1s] <- rexp(sum(rate1s),rate=mu21)
srv.t2[rate2s] <- rexp(sum(rate2s),rate=mu22)

dep.t2  <- srv.t2 + arr.t2 #departure times for queue 2

plot(arr.t2,srv.t2,pch=4,cex=0.5,xlab="",ylab="")
title("Graphical Point Process of Queue 2",xlab="Time",ylab="Service Requirement",cex.main=3,cex.lab=1.7)

m2 <- floor(m*(n2/n1))
T2 <- n2

latt2 <- seq(from=0,to=T2,length.out=m2) #discretised times

cum.arr2 <- sapply(1:m2, function(j) sum(arr.t2 <= latt2[j])) #cumulative arrivals up to t in queue 2
cum.dep2 <- sapply(1:m2, function(j) sum(dep.t2 <= latt2[j])) #cumulative departures up to t in queue 2

q.length2 <- sapply(1:m2, function(j) sum(arr.t2 <= latt2[j]) - sum(dep.t2 <= latt2[j])) #length of queue 2 at time t
q.length.eq2 <- q.length2[floor(m2/5) : (m2-floor(m2/5))] #length of queue 2 with burn in and burn out removed (equilibrium)

latt2chop <- latt2[floor(m2/5) : (m2-floor(m2/5))] #remove edge effects

plot(latt2chop,q.length.eq2,typ="l",xlab="",ylab="")
title("Queue 2 Length without Edge Effects (Stationary)",xlab="Time",ylab="Queue Length",cex.main=3,cex.lab=1.7)

## ACF plots
a <- mu11
b <- mu12
c <- mu21
d <- mu22
e <- lam1
f <- lam2
g <- alpha11
h <- alpha12
i <- alpha21
j <- alpha22

#theoretical acvf
theory_acvf <- function(x)
{
f^2*e*(g/a*(i^2*(c*exp(-a*x)-a*exp(-c*x))/c/(c+a)/(c-a)+j^2*(d*exp(-a*x)-
a*exp(-d*x))/d/(d+a)/(d-a)+
i*j*((2*a+c+d)*(exp(-c*x)+exp(-d*x))/(c+a)/(d+a)/(c+d)+(exp(-a*x)+exp(-c*x))/(d+a)/(c-a)+(exp(-a*x)+
exp(-d*x))/(c+a)/(d-a)))+h/b*(i^2*(c*exp(-b*x)-b*exp(-c*x))/c/(c+b)/(c-b)+j^2*(d*exp(-b*x)-
b*exp(-d*x))/d/(d+b)/(d-b)+i*j*((2*b+c+d)*(exp(-c*x)+exp(-d*x))/(c+b)/(d+b)/(c+d)+(exp(-b*x)+
exp(-c*x))/(d+b)/(c-b)+(exp(-b*x)+exp(-d*x))/(c+b)/(d-b))))+f*e*(g/a+h/b)*(i/c*exp(-c*x)+j/d*exp(-d*x))
}

theory_acf <- function(x)
{
theory_acvf(x)/theory_acvf(0)
} #theoretical acf

seq1 <- seq(from=0,to=40,length.out=1e3)
theo_acf <- theory_acf(seq1)
acf(q.length.eq2, ci=0,type="correlation",typ="b",main="",xlab="",ylab="")
title("Autocorrelation Function of Queue 2 Length",xlab="Lag",ylab="ACF",cex.main=3,cex.lab=1.7)
lines(seq1, theo_acf,add=T,col="red",lwd=2) #overlay theoretical line
```

```

##queue 1 - M/genE2/Infty

n1 <- 1e3 #number of customers in 1st queue
m <- 1e3 #number of time points
lam1 <- 0.9 #arrival rate into 1st queue
mu11 <- 3 #service rate of first phase in series of 1st queue
mu12 <- 1.5 #service rate of second phase in series of 1st queue

par(mfcol=c(3,2)) #grid of plots

arr.t1 <- rep(NA,n1)
for (j in 1:n1){
        if(j==1){
                arr.t1[j] <- rexp(1,rate=lam1)
                next
        }
        arr.t1[j] <- arr.t1[j-1] + rexp(1,rate=lam1)
} #arrival times for queue 1

srv.t1 <- rexp(n1,rate=mu11)+rexp(n1,rate=mu12) #service times genE2 for queue 1

dep.t1  <- srv.t1 + arr.t1 #departure times for queue 1
```

```
plot(arr.t1,srv.t1,pch=4,cex=0.5,xlab="",ylab="")
title("Graphical Point Process of Queue 1",xlab="Time",ylab="Service Requirement",cex.main=3,cex.lab=1.7)

T <- 1000 #end time

latt1 <- seq(from=0,to=T,length.out=m) #discretised times

cum.arr1 <- sapply(1:m, function(j) sum(arr.t1 <= latt1[j])) #cumulative arrivals up to t in queue 1
cum.dep1 <- sapply(1:m, function(j) sum(dep.t1 <= latt1[j])) #cumulative departures up to t in queue 1

q.length1 <- sapply(1:m, function(j) sum(arr.t1 <= latt1[j]) - sum(dep.t1 <= latt1[j])) #length of queue 1 at time t
q.length.eq1 <- q.length1[floor(m/5) : (m-floor(m/5))] #length of queue 1 with burn in and burn out removed (equilibrium)

latt1chop <- latt1[floor(m/5) : (m-floor(m/5))] #lattice without first and last fifth

plot(latt1chop,q.length.eq1,typ="l",xlab="",ylab="")
title("Queue 1 Length without Edge Effects (Stationary)",xlab="Time",ylab="Queue Length",cex.main=3,cex.lab=1.7)
acf(q.length.eq1,ci=0,main="",xlab="",ylab="",type="correlation",typ="b")
title("Autocorrelation Function of Queue 1 Length",xlab="Lag",ylab="ACF",cex.main=3,cex.lab=1.7)

##queue 2 - Cox/genE2/Infty

n2 <- 200 #number of customers ever through queue 2 (significantly less than that of queue 1 as otherwise arrival rate drops to 0 eventually)
lam2 <- 0.9 #arrival rate constant into queue 2
mu21 <- 4 #service rate of first phase in series of 2nd queue
mu22 <- 4/3 #service rate of second phase in series of 2nd queue

arr.t2 <- rep(NA,n2)
non_zero_q <- which(!(q.length1==0))
for (j in 1:n2){
        if(j==1){
                arr.t2[j] <- rexp(1,rate=lam2)
                next
        }
        above <- which(latt1 >= arr.t2[j-1])
        next.ind <- intersect(above,non_zero_q)[1]
        arr.t2[j] <- arr.t2[j-1] + latt1[next.ind]-latt1[above[1]] + rexp(1,rate=lam2*q.length1[next.ind])
} #arrival times into queue 2
#1st arrival time doesn't matter as we'll exclude burn in
#look at times when 1st queue is non-empty (otherwise arrival rate 0)
#find next time queue 1 is empty, then sample Exp(lam2*N1(t)) for next arrival plus time spent waiting for arrival rate to return above 0

srv.t2 <- rexp(n2,rate=mu21)+rexp(n2,rate=mu22) #service times genE2 of queue 2

dep.t2  <- srv.t2 + arr.t2 #departure times of queue 2

plot(arr.t2,srv.t2,pch=4,cex=0.5,xlab="",ylab="")
title("Graphical␣Point␣Process␣of␣Queue␣2",xlab="Time",ylab="Service␣Requirement",cex.main=3,cex.lab=1.7)

m2 <- floor(m*(n2/n1))
T2 <- n2

latt2 <- seq(from=0,to=T2,length.out=m2) #discretised times

cum.arr2 <- sapply(1:m2, function(j) sum(arr.t2 <= latt2[j])) #cumulative arrivals up to t in queue 2
cum.dep2 <- sapply(1:m2, function(j) sum(dep.t2 <= latt2[j])) #cumulative departures up to t in queue 2

q.length2 <- sapply(1:m2, function(j) sum(arr.t2 <= latt2[j]) - sum(dep.t2 <= latt2[j])) #length of queue 2 at time t
q.length.eq2 <- q.length2[floor(m2/5) : (m2-floor(m2/5))] #length of queue 2 with burn in and burn out removed (equilibrium)

latt2chop <- latt2[floor(m2/5) : (m2-floor(m2/5))] #remove edge effects

plot(latt2chop,q.length.eq2,typ="l",xlab="",ylab="")
title("Queue␣2␣Length␣without␣Edge␣Effects␣(Stationary)",xlab="Time",ylab="Queue␣Length",cex.main=3,cex.lab=1.7)

## ACF plots
a <- mu11
b <- mu12
c <- mu21
d <- mu22
e <- lam1
f <- lam2

#theoretical acvf
```

```r
theory_acvf <- function(x)
{
e*f^2/(a-b)/(c-d)^2*(a/b*(c^2*(d*exp(-b*x)-b*exp(-d*x))/d/(d+b)/(d-b)+d^2*(c*exp(-b*x)-b*exp(-c*x))/c/(c+b)/(c-b)-
c*d*((exp(-b*x)-exp(-d*x))/(d+b)/(d-b)+exp(-d*x)*(2*b+c+d)/(d+b)/(c+b)/(c+d)+(exp(-b*x)-exp(-c*x))/(c+b)/(c-b)+
exp(-c*x)*(2*b+c+d)/(d+b)/(c+b)/(c+d))-b/a*(c^2*(d*exp(-a*x)-a*exp(-d*x))/d/(d+a)/(d-a)+d^2*(c*exp(-a*x)-
a*exp(-c*x))/c/(c+a)/(c-a)-c*d*((exp(-a*x)-exp(-d*x))/(d+a)/(d-a)+exp(-d*x)*(2*a+c+d)/(d+a)/(c+a)/(c+d)+(exp(-a*x)-
exp(-c*x))/(c+a)/(c-a)+exp(-c*x)*(2*a+c+d)/(d+a)/(c+a)/(c+d))))+f*e*(1/a+1/b)/(c-d)*(c/d*exp(-d*x)-d/c*exp(-c*x))
}

theory_acf <- function(x)
{
theory_acvf(x)/theory_acvf(0)
} #theoretical acf

seq1 <- seq(from=0,to=40,length.out=1e3)
theo_acf <- theory_acf(seq1)
acf(q.length.eq2,ci=0,type="correlation",typ="b",main="",xlab="",ylab="")
title("Autocorrelation Function of Queue 2 Length",xlab="Lag",ylab="ACF",cex.main=3,cex.lab=1.7)
lines(seq1,theo_acf,add=T,col="red",lwd=2) #overlay theoretical line
```

# References

[1] Anderson,D., Blom,J., Mandjes,M., Thorsdottir,H., De Turck,K. (2016). A functional central limit theorem for a Markov-modulated infinite-server queue. *Methodology and Computing in Applied Probability*, 18(1):153-168.

[2] Arazi,A., Ben-Jacob,E., Yechiali,U. (2004). Bridging genetic networks and queueing theory. *Physica A*, 332:585-616.

[3] Asmussen,S. (1987). *Applied Probability and Queues (1st Edition).* Wiley, New York.

[4] Avram,F., Bertsimas,D. (1993). On Central Limit Theorems in Geometrical Probability. *Annals of Applied Probability*, 3:1033-1046.

[5] Baccelli,F., Blaszczyszyn, B. (2009). *Stochastic Geometry and Wireless Networks: Volume I: Theory.* NoW Publishers Inc., Boston.

[6] Baccelli,F., Blaszczyszyn, B. (2009). *Stochastic Geometry and Wireless Networks: Volume II: Applications.* NoW Publishers Inc., Boston.

[7] Barbour,A., Xia,A. (2006). Normal Approximation for Random Sums. *Advances in Applied Probability*, 38:693-728.

[8] Baryshnikov,Y., Yukich,J. (2005). Gaussian Limits for Random Measures in Geometric Probability. *Annals of Applied Probability*, 15:213-253.

[9] Baryshnikov,Y., Eichelsbacher,P., Schreiber,T., Yukich,J. (2008). Moderate Deviations for some Point Measures in Geometric Probability. *Annales de l'Institut Henri Poincare*, 44:422-446.

[10] Baskett,F., Chandy,K., Muntz,M., Palacios,R. (1975). Open, Closed and Mixed Networks of Queues with Different Classes of Customers. *Journal of the ACM*, 22(2):248-260.

[11] Baxter,J., Jain,N. (1988). A Comparison Principle for Large Deviations. *Proceedings of the American Mathematical Society*, 103:1235-1240.

[12] Billingsley,P. (1999). *Convergence of Probability Measures, 2nd Edition.* Wiley, New York.

[13] Blanchet,J., Chen,X., Lam,H. (2014). Two-Parameter Sample Path Large Deviations for Infinite-Server Queues. *Stochastic Systems*, 4(1):206-249.

[14] Blom,J., Mandjes,M. (2013). A large-deviations analysis of Markov-modulated infinite-server queues. *Operations Research Letters*, 41:220-225.

[15] Blom,J., De Turck,K., Mandjes,M. (2013). Rare event analysis of Markov-modulated infinite-server queues: a Poisson limit. *Stochastic Models*, 29:463-474.

[16] Blom,J., De Turck,K., Mandjes,M. (2013). A central limit theorem for Markov-modulated infinite-server queues. *In: Dudin,A., De Turck,K. (eds.): Proceedings ASMTA 2013, Ghent, Belgium. Lecture Notes in Computer Science (LNCS) Series*, 7984:81-95.

[17] Blom,J., Kella,O., Mandjes,M., Thorsdottir,H. (2014). Markov-modulated infinite server queues with general service times. *Queueing Systems*, 76:403-424.

[18] Blom,J., Kella,O., Mandjes,M., De Turck,K. (2014). Tail asymptotics of a Markov-modulated infinite-server queue. *Queueing Systems*, 78:337-357.

[19] Blom,J., De Turck,K., Mandjes,M. (2015). Analysis of Markov-modulated infinite-server queues in the central-limit regime. *Probability in the Engineering and Informational Sciences*, 29:433-459.

[20] Blom,J., De Turck,K., Mandjes,M. (2016). Functional central limit theorems for Markov-modulated infinite-server systems. *Mathematical Methods of Operations Research*, 83(3):351-372.

[21] Bochner,S. (1955). *Harmonic analysis and the theory of probability*. University of California.

[22] Bogachev,V. (2007). *Measure Theory (Volume 2)*. Springer.

[23] Chaganty,N. R. (1996). Some Properties of the Kullback-Leibler Number. *Sankhyā*, Series A 58:69-80.

[24] Chaganty,N. R. (1997). Large deviations for joint distributions and statistical applications. *Sankhyā*, Series A 59:147-166.

[25] Chao,X., Miyazawa,M., Pinedo,M. (1999). *Queueing Networks: Customers, Signals and Product Form Solutions*. Wiley, New York.

[26] Coolen-Schrijner,P., van Doorn,E. (2002). The deviation matrix of a continuous-time Markov chain. *Probability in the Engineering and Informational Sciences*, Series A 16:351-366.

[27] Cookson,N., Mather,W., Danino,T., Mondragón-Palomino,O., Williams,R., Tsimring,L., Hasty,J. (2011). Queueing up for enzymatic processing: correlated signaling through coupled degradation. *Molecular Systems Biology*. 7, Article number: 561.

[28] Crick,F. (1970). Central Dogma of Molecular Biology. *Nature*, 227:221-237.

[29] Daley,D., Vere-Jones,D. (1988). *An Introduction to the Theory of Point Processes*. Springer, New York.

[30] Dassios,A., Jang,J. (2003). Pricing of catastrophe reinsurance and derivatives using the Cox process with shot noise intensity. *Journal of Finance and Stochastics*, 7:73-95.

[31] D'Auria,B. (2008). $M/M/\infty$ queues in semi-Markovian random environment. *Queueing Systems*, 58:561-563.

[32] Daw,A., Pender,J. (2017). Queues driven by Hawkes processes. *Stochastic Systems*, 8(3):192-229.

[33] Dawson,D., Gärtner,J. (1987). Large deviations from the McKean-Vlasov limit for weakly interacting diffusions. *Stochastics*, 20:247-308.

[34] Dean,J. (2014). Stochastic Fluctuations in Gene Regulatory Networks. *Masters Thesis, University of Bristol.*

[35] Dean,J., Ganesh,A., Crane,E. (2018). Functional Large Deviations for Cox Processes and $Cox/G/\infty$ Queues, with a Biological Application. *https://arxiv.org/abs/1808.04347.*

[36] Decreusefond,L., Moyal,P. (2008). A functional central limit theorem for the $M/GI/\infty$ queue. *Annals of Applied Probability*, 18(6):2156-2178.

[37] De Jong,H. (2002). Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *Journal of Computational Biology*, 9(1):67-103.

[38] Dembo,A., Zeitouni,O. (1998). *Large Deviations Techniques and Applications (2nd edition).* Springer, New York.

[39] Dessalles,R., Fromion,V., Robert,P. (2017). A stochastic analysis of autoregulation of gene expression. *Journal of Mathematical Biology*, 75:1253-1283.

[40] Dupuis,P., Ellis,R. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley Series in Probability and Statistics. John Wiley and Sons, New York.

[41] Erlang,A. (1909). The Theory of Probabilities and Telephone Conversations. *Nyt Tidsskrift for Matematik*, 20:33-39.

[42] Florens,D., Pham,H. (1998). Large deviation probabilities in estimation of Poisson random measures. *Stochastic Processes and their Applications*, 76:117-139.

[43] Fromion,V., Leoncini,E., Robert,P. (2013). Stochastic Gene Expression in Cells: A Point Process Approach. *SIAM Journal on Applied Mathematics*, 73(1):195-211.

[44] Ganesh,A., O'Connell,N., Wischik,D. (2004). *Big Queues.* Lecture Notes in Mathematics, Springer.

[45] Gao,X., Zhu,L. (2018). Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. *Queueing Systems* 90(1-2):161-206.

[46] Gelenbe,E. (1994). G-networks: a unifying model for neural and queueing networks. *Annals of Operations Research*, 48:433-461.

[47] Gelenbe,E. (2007). Steady-state solution of probabilistic gene regulatory networks. *Physical Review E*, 76:031903.

[48] Glynn,P. (1995). Large Deviations for the Infinite Server Queue in Heavy Traffic. *Stochastic Networks, Vol. 71 of Mathematics and its Applications eds. F. Kelly and R. Williams, Springer, Berlin*, 387-394.

[49] Haenggi,M. (2012). *Stochastic Geometry for Wireless Networks.* Cambridge University Press.

[50] Harchol-Balter,M. (2013). *Performance Modeling and Design of Computer Systems: Queueing Theory in Action.* Cambridge University Press.

[51] Heemskerk,M., van Leeuwaarden,J., Mandjes,M. (2017). Scaling limits for infinite-server systems in a random environment. *Stochastic Systems*, 7(1):1-31.

[52] Hellings,T., Mandjes,M., Blom,J. (2012). Semi-Markov-modulated infinite-server queues: approximations by time-scaling. *Stochastic Models*, 28:452-477.

[53] Iglehart,D. (1965). Limiting Diffusion Approximations for the Many Server Queue and the Repairman Problem. *Journal of Applied Probability*, 2(2):429-441.

[54] Itô,K. (1983). Distribution-valued processes arising from independent Brownian motions. *Mathematische Zeitschrift*, 182(1):17-33.

[55] Jackson,J. (1957). Networks of Waiting Lines. *Operations Research*, 5(4):518-521.

[56] Jansen,H., Mandjes,M., De Turck,K., Wittevrongel,S. (2016). A large deviations principle for infinite-server queues in a random environment. *Queueing Systems*, 82:199-235.

[57] Kallianpur,G., Xiong,J. (1995). *Stochastic differential equations in infinite-dimensional spaces.* Institute of Mathematical Statistics Lecture Notes - Monograph Series, 26. Institute of Mathematical Statistics, Hayward, CA..

[58] Kauffman,S.A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22:437-467.

[59] Kelly,F. (1979). *Reversibility and Stochastic Networks.* John Wiley and Sons, New York.

[60] Kendall,D.G. (1953). Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. *Annals of Mathematical Statistics*, 24(3):338.

[61] Kesten,H., Lee,S. (1996). The Central Limit Theorem for Weighted Minimal Spanning Trees on Random Points. *Annals of Applied Probability*, 6:495-527.

[62] Kolmogorov,A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer.

[63] Komárková,Z. (2012). Phase-type Approximation Techniques. *Bachelor Thesis, Masaryk University, Faculty of Informatics.*

[64] Koops,D., Boxma,O., Mandjes,M. (2017). Networks of $\bullet/G/\infty$ queues with shot-noise-driven arrival intensities. *Queueing Systems*, 86(3-4):301-325.

[65] Koops,D., Saxena,M., Boxma,O., Mandjes,M. (2017). Infinite-server queues with Hawkes input. *https://arxiv.org/abs/1707.02196.*

[66] Last,G., Penrose,M. (2017). *Lectures on the Poisson Process.* Cambridge University Press.

[67] Lee,S. (1997). The Central Limit Theorem for Euclidean Minimum Spanning Trees. *Annals of Probability*, 7:996-1020.

[68] Leite,S., Williams,R. (2017). A Constrained Langevin Approximation for Chemical Reaction Networks. *http://www.math.ucsd.edu/ williams/biochem/biochem.html.*

[69] Léonard,C. (2000). Large deviations for Poisson random measures and processes with independent increments. *Stochastic Processes and their Applications*, 85:93-121.

[70] Lestas,I., Paulsson,J., Ross,N., Vinnicombe,G. (2008). Noise in Gene Regulatory Networks. *IEEE Transactions on Automatic Control*, 53:189-200.

[71] Martinez,V., Saar,E. (2002). *Statistics of the Galaxy Distribution.* Chapman and Hall / CRC: Boca Raton.

[72] Mather,W., Cookson,N., Hasty,J., Tsimring,L., Williams,R. (2010). Correlation resonance generated by coupled enzymatic processing. *Biophysical Journal*, 99:3172-3181.

[73] Mather,W., Hasty,J., Tsimring,L., Williams,R. (2011). Factorized time-dependent distributions for certain multiclass queueing networks and an application to enzymatic processing networks. *Queueing Systems*, 69:313-328.

[74] Mather,W., Hasty,J., Tsimring,L., Williams,R. (2013). Translational cross talk in gene networks. *Biophysical Journal*, 104:2564-2572.

[75] Mitrani,I. (1998). *Probabilistic Modelling.* Cambridge University Press.

[76] Molchanov,I., Stoyan,D. (1994). Asymptotic Properties of Estimators for Parameters of the Boolean Model. *Advances in Applied Probability*, 26:301-323.

[77] O'Cinneide,C., Purdue,P. (1986). The $M/M/\infty$ queue in a Random Environment. *Journal of Applied Probability*, 23:175-184.

[78] Pang,G., Talreja,R., Whitt,W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian Queues. *Probability Surveys*, 4:193-267.

[79] Pang,G., Whitt,W. (2010). Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems*, 65:325-364.

[80] Pang,G., Whitt,W. (2013). Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems*, 73:119-146.

[81] Pang,G., Zhou,Y. (2017). Two-parameter process limits for an infinite-server queue with arrival dependent service times. *Stochastic Processes and their Applications*, 127:1375-1416.

[82] Pang,G., Zhou,Y. (2018). Two-parameter process limits for infinite-server queues with dependent service times via chaining bounds. *Queueing Systems*, 88:1-25.

[83] Paulsson,J. (2005). Models of stochastic gene expression. *Physics of Life Reviews*, 2:157-175.

[84] Penrose,M., Yukich,J. (2001). Central Limit Theorems for some Graphs in Computational Geometry. *Annals of Applied Probability*, 11:1005-1041.

[85] Penrose,M., Yukich,J. (2002). Limit Theory for Random Sequential Packing and Deposition. *Annals of Applied Probability*, 12:272-301.

[86] Penrose,M., Yukich,J. (2003). Weak Laws of Large Numbers in Geometric Probability. *Annals of Applied Probability*, 13:277-303.

[87] Penrose,M. (2003). *Random Geometric Graphs.* Oxford University Press, Oxford, New York.

[88] Penrose,M., Yukich,J. (2005). Normal Approximation in Geometric Probability. *Proceedings of the Workshop 'Stein's Method and Applications', Lecture Notes Series, Singapore, Institute for Mathematical Sciences: World Scientific Press*, 5:37-58.

[89] Penrose,M. (2007). Gaussian Limits for Random Geometric Measures. *Electronic Journal of Probability*, 12:989-1035.

[90] Reed,J., Talreja,R. (2015). Distribution-valued heavy-traffic limits for the $G/GI/\infty$ queue. *Annals of Applied Probability*, 25(3):1420-1474.

[91] Schreiber,T. (2003). Large deviation principle for empirical measures generated by Cox point processes. *Colloquium Mathematicum*, 97(1):87-106.

[92] Schreiber,T., Yukich,J. (2005). Large Deviations for Functionals of Spatial Point Processes with Applications to Random Packing and Spatial Graphs. *Stochastic Processes and their Applications*, 115:1332-1356.

[93] Schreiber,T., Penrose,M., Yukich,J. (2007). Gaussian Limits for Random Sequential Packing at Saturation. *Communications in Mathematical Physics*, 272:167-183.

[94] Schreiber,T. (2010). Limit Theorems in Stochastic Geometry. *New Perspectives in Stochastic Geometry. Oxford University Press, Oxford*, 111-144.

[95] Shorack,G., Wellner,J. (1986). *Empirical Processes with Applications to Statistics.* John Wiley and Sons.

[96] Smolen,P., Baxter,D.A. and Byrne,J.H. (2000). Modeling transcriptional control in gene networks - Methods, recent results, and future directions. *Bulletin of Mathematical Biology*, 62:247-292.

[97] Stoyan,D., Penttinen,A. (2000). Recent Applications of Point Process Methods in Forestry Statistics. *Statistical Science*, 15:61-78.

[98] Stoyan,D., Kendall,W., Mecke,J. (2005). *Stochastic Geometry and its Applications 2nd Edition.* John Wiley and Sons.

[99] van der Vaart,A., Wellner,J. (2000). *Weak Convergence and Empirical Processes.* New York: Springer 2nd Edition.

[100] Varadarajan,V. (1958). Weak Convergence of Measures on Separable Metric Spaces. *Sankhyā: Indian Journal of Statistics*, 19:15-22.

[101] Varadhan,S. (1966). Asymptotic probabilities and differential equations. *Communications on Pure and Applied Mathematics*, 1:261-286.

[102] Wade,A. (2007). Explicit Laws of Large Numbers for Random Nearest-Neighbour-type Graphs. *Advances in Applied Probability*, 39:326-342.

[103] Warner,J., Vilardell,J., Sohn,J. (2001). Economics of ribosome biosynthesis. *Cold Spring Harbor Symposium on Quantatitive Biology*, 66:567-574.

[104] Whitt,W. (2002). Stochastic Process Limits. *Springer Verlag.*

[105] Wu,L. (1994). Large deviations, moderate deviations and LIL for empirical processes. *Annals of Probability*, 22:17-27.

[106] Yamasaki,Y. (1975). Kolmogorov's Extension Theorem for Infinite Measures. *Publ. RIMS, Kyoto University*, 10:381-411.

[107] Zajic,T. (1998). Rough asymptotics for tandem non-homogeneous $M/G/\infty$ queues via Poissonized empirical processes. *Queueing Systems*, 29:161-174.