Perfect Simulation, Sample-path Large Deviations, and Multiscale Modeling for Some Fundamental Queueing Systems

Xinyun Chen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Perfect Simulation, Sample-path Large Deviations, and Multiscale Modeling

for Some Fundamental Queueing Systems

Xinyun Chen

As a primary branch of Operations Research, Queueing Theory models and analyzes engineering systems with random fluctuations. With the development of internet and computation techniques, the engineering systems today are much bigger in scale and more complicated in structure than 20 years ago, which raises numerous new problems to researchers in the field of queueing theory. The aim of this thesis is to explore new methods and tools, from both algorithmic and analytical perspectives, that are useful to solve such problems.

In Chapter 2 and 3, we introduce some techniques of asymptotic analysis that are relatively new to queueing applications in order to give more accurate probabilistic characterization of queueing models with large scale and complicated structure. In particular, Chapter 2 gives the first functional large deviation result for infinite-server system with general inter-arrival and service times. The functional approach we use enables a nice description of the whole system over the entire time horizon of interest, which is important in real problems. In Chapter 3, we construct a queueing model for the so-called limit order book that is used in main financial markets worldwide. We use an asymptotic approach called multi-scale modeling to disentangle the complicated dependence among the elements in the trading system and to reduce the model dimensionality. The asymptotic regime we use is inspired by empirical observations and the resulting limit process explains and reproduces stylized features of real market data. Chapter 3

also provides a nice example of novel applications of queueing models in systems, such as the electronic trading system, that are traditionally outside the scope of queueing theory.

Chapter 4 and 5 focus on stochastic simulation methods for performance evaluation of queueing models where analytic approaches fail. In Chapter 4, we develop a perfect sampling algorithm to generate exact samples from the stationary distribution of stochastic fluid networks in polynomial time. Our approach can be used for time-varying networks with general inter-arrival and service times, whose stationary distributions have no analytic expression. In Chapter 5, we focus on the stochastic systems with continuous random fluctuations, for instance, the workload arrives to the system in continuous flow like a Lévy process. We develop a general framework of simulation algorithms featuring a deterministic error bound and an almost square root convergence rate. As an application, we apply this framework to estimate the stationary distributions of reflected Brownian motions and the performance of our algorithm is better than existing prevalent numeric methods.

Contents

List of Figures

List of Tables

Acknowledgements

Foremost, I would like to express my sincerest gratitude to my advisor Professor Jose Blanchet, for his continuous help, inspiration and encouragement. Without his guidance and support, I could not have completed this dissertation. He not only provides valuable advices on my doctoral research, but also cares about my career and personal development. He also influences me with his constant passion and curiosity towards research and life. He is the best advisor I can imagine and I am very fortunate to work with him during my doctoral study.

I am grateful to Professors Paul Glasserman, Philip Protter, Karl Sigman, Ward Whitt and David Yao for serving on my dissertation committee, writing recommendation letters and their great help in all aspects. They are among the greatest experts in stochastic systems and I am honored to have the opportunity discussing with them about my research.

I would like to thank Professor Henry Lam for his help when we were working together on the large deviation paper. He is always willing to share with me his ideas and thoughts on research and career development. I also thank Nicolas Bachelier for his research assistance. We worked together implementing the tolerance-enforced simulation algorithms.

I would like to thank all my friends and fellow Ph.D students in the IEOR department for their sincere friendship and for all the good and bad time we have shared together.

In the end, I would like to thank my parents and my husband, Yan, for their tremendous love and support and for their being with me along this journey.

To My Mother

Chapter 1

Overview

This thesis is composed of several projects that aimed at studying techniques and models applicable to modern stochastic operations research. Generally speaking, the models that are studied here can all be viewed as queueing models and they provide mathematical descriptions of actual systems in engineering and finance that are exposed to random fluctuations. Due to the complexity of the actual systems, the corresponding queueing model tends to have the following features: (1) non-Markov probability structure, (2) large scale (high dimensionality and fast transition rate), (3) time-inhomogeneous dynamics, and (4) complex dependence among elements. Having these features, the model is hard to analyze using classic queueing methods. Therefore, the purpose of this thesis is to explore and develop new methods and tools, from both algorithmic and analytical perspectives, that are useful to study queueing models with these features.

The two classes of tools that we shall mainly apply in this thesis are asymptotic analysis and stochastic simulation. Asymptotic analysis techniques are frequently used in the queueing

literature to obtain approximation, usually based on the so-called fluid or diffusion limit, of a queueing process. For instance, when studying large call centers, one can approximate the queue length process by its heavy-traffic limit given that the system has a high arrival rate of customers. As a classic example, under the Halfin-Whitt regime, the queue length process of a many-server queue converges to some diffusion process. In Chapter 2 and 3, we explore some asymptotic analysis methods in settings that are similar to the heavy-traffic setting in spirit.

Chapter 2 gives the first functional large deviation result for infinite-server systems with general inter-arrival and service times. In addition to being interesting in its own right, infinite-server systems are closely related to many-server queues in the heavy-traffic setting. The results can be used to characterize the system dynamics conditional on rare events. Besides, the functional approach enables us to describe the profile of the whole system, for instance, the arrival and departure times of all the customers, over the entire time horizon of interest conditional on the rare event, which is important in real problems such as in risk management. As two examples of applications, we use the large deviation results to analyze service systems and insurance portfolios in the occurrence of congestion or default.

In Chapter 3, we construct a queueing model for the so-called limit order book as are used in main financial markets worldwide. Intuitively, a limit order book can be viewed as two interacting multi-class queues corresponding to the buying and selling sides in the market. The asset price is determined by the number and class of orders in the queues. Ideally, one should be able to predict the price movement from the dynamics of this queueing system. However, this system is very difficult to analyze directly because it features fast arrival rates, huge number of classes, time-varying transition rates and complicated dependence between sellers and buyers.

To handle these features, we choose a special asymptotic regime inspired by empirical observations and carry out the asymptotic analysis using the so-called stochastic averaging principle. This asymptotic approach can disentangle the complicated dependence and reduce the model dimensionality. In the end, we derive the resulting asset price process in the asymptotic regime, and the price dynamics we get can explain several stylized phenomena observed in trading data.

However, when the network structure is very complicate, one usually can not evaluate the system performance analytically even using the more tractable limit processes obtained from asymptotic analysis. For example, under modest assumptions and a suitable heavy-traffic scaling, the workload process of a typical queueing network can be approximated by a so-called reflected Brownian motion in the positive orthant. The stationary distribution of the reflected Brownian motion also approximates that of the queueing network. However, for a general reflected Brownian motion, its stationary distribution has no closed-form expression and the analytic description remains an open problem. In this light, we shall pursue such evaluation problems from a different algorithmic approach using Monte Carlo simulation. Compared with other numeric methods, Monte Carlo simulation has the merits that it avoids the curse of dimensionality and can easily be adapted to parallel computing.

In Chapter 4, we develop a perfect sampling algorithm to generate exact samples from the stationary distribution of a time-varying stochastic fluid network with general inter-arrival and service times. Our algorithm is based on the coupling from the past (CFTP) technique and a novel application of importance sampling. We prove polynomial complexity for our algorithm in the number of servers. The efficiency of our algorithm is also supported by results of numerical experiments evaluating stationary performances of queueing networks.

In Chapter 5, we introduce a general framework of simulation algorithms aiming at stochastic processes with continuous random fluctuations such as the Lévy processes. These processes usually serve as the input random fluctuations that drive the stochastic dynamics of queueing models. Mathematically speaking, there is some continuous mapping (in some function space) that maps every sample path of the input process to that of the queueing process. As a result, simulation of such input processes is essential in queueing model simulation. Our framework helps to improve the precision and efficiency of the simulation algorithms of queueing models in the sense that it provides a strong control, in the supreme norm, on the path simulation error of the input process and at the same time achieves (almost) the best possible convergence rate of simulation algorithms. As an example, we design a class of algorithms under this framework to evaluate the stationary performance of different reflected Brownian motions. Numerical results indicate that our algorithms have smaller bias and are efficient compared to prevalent numerical methods.

Chapter 2

Two-parameter Sample Path Large Deviations for Infinite Server Queues

2.1 Introduction

The asymptotic analysis of queueing systems with many servers in heavy-traffic has received substantial attention, especially in recent years. Among the earliest references that come to mind in connection to this topic is the work of [39] on heavy-traffic limits for the infinite-server queue. Another highly influential paper in the area is [33] in the context of many server Markovian queues, which introduced a scaling that is now known as the "Quality and Efficiency Driven" regime. The ideas in these papers have fueled more recent results in the asymptotic analysis of many server systems such as: [55], [40], [57], [41], [42], in the setting of many server queues, and [29], [21], [52], [58], in the setting of the infinite server queue. The asymptotic analysis of queueing systems with many servers has been motivated by applications in service engineering,

in particular in the context of call centers and health-care operations. Another set of application areas that is also very relevant, but that is infrequently mentioned in the analysis of many server systems is that of insurance mathematics. It is clear, for instance, that a portfolio of insurance policies can be directly modeled as an infinite server system; casting insurance portfolios in this framework is particularly appealing in the setting of life insurance as we shall illustrate in Section 2.4.

So far most of the asymptotic analysis of many server systems has concentrated mainly on fluid and diffusion approximations on central limit scaling. Meanwhile, the literature on large deviations analysis for many server queues is not as extensive relatively; despite the fact that it is clearly of interest to understand the large deviations behavior of these types of systems. For instance, consider the consequences of dropping calls in an emergency call center or being unable to satisfy the demand for critically ill patients in the context of health-care applications. As another application, in the insurance setting, it is of interest to estimate ruin probabilities and, perhaps even more importantly, understanding the most likely path (or set of paths) to ruin. Risk theory typically concentrates on ruin probabilities for aggregated models, such as the classical ruin model (see [2]); the results in this paper, as we shall illustrate, provide a systematic way for assessing ruin probabilities for a natural class of bottom-up models.

Our main contribution in this paper is to provide the first sample-path large deviations analysis of the state descriptor of the infinite server queueing model in heavy-traffic (i.e. as the arrival rate increases to infinity without introducing any scaling on the service times). The statement of our main result, which is given in Theorem 2.3.1, features a convenient representation of a good large deviations rate function, under a strong topology. To illustrate the strength of our

result, we apply it to compute the most likely path to overflow in a loss system, and also the most likely path to ruin for a life insurance portfolio that embeds an infinite server queue with a particular service cost structure. It is important to emphasize that our result takes advantage of a convenient representation of the system's description that facilitates the representation of the rate function; detailed discussion on this system's representation is given in Section 2.2. Previous large deviations analysis of the infinite server queue has concentrated on queue-length characteristics only; see, for instance, [28] who develops large deviations for marginal quantities in the case of renewal arrivals, and [70] who develops sample path large deviations for the queue length process of infinite server queues in tandem in the case of Poisson arrivals.

Our large deviations analysis complements results on fluid analysis and diffusion approximations recently obtained for infinite server systems. For example, [52] have shown that the state descriptor of the infinite server queue, suitably parameterized in terms of a two-parameter stochastic process, converges after centering and re-scaling to a Gaussian and Markov process; see also [58] who interpret the state descriptor of the infinite server queue as a measure valued process acting on the space of tempered distributions. These recent results, in turn, extend prior work by [29] in the context of discrete and bounded service time distributions, and [21] for the case of Poisson arrivals. We also mention the growing literature on large deviations of measure valued processes, see for example, [47], and [26] for general theory. This literature is relevant as the state of the infinite server queue at time t can be represented as a measure with point masses representing the remaining service times of the customers currently in the system. This approach requires to define the right topology on the space of measures, just as in the weak convergence analysis in [21] and [58]. It appears that the resulting topologies, however, would

not be as strong as the ones that we consider here (see, for instance, the discussion on the resulting topologies in p. 3 of [47]). Our topology is basically the same as that in [52], which in turn is stronger than that in [21] and [58] (which do not include the queue length process as a continuous function, for example). In addition, the rate function would involve a different representation than the one we obtain here. We believe that our representation is more convenient for applications in queueing, as we illustrate in our examples in Section 2.4.

The analysis of the infinite server queue is important as it serves as a building block for other models of interest. For instance, in the setting of loss models one can clearly couple the loss systems with associated infinite server systems, and in the setting of many server queues [57] shows how one can precisely understand queues with multiple servers as a perturbation of infinite server queues. Furthermore, the infinite server model is a classical model in queueing theory that serves as a direct model in important applications. Of particular interest to us, as mentioned earlier, are the applications to insurance mathematics.

The rest of the paper is organized as follows. In Section 2.2 we introduce our problem setting and provide a statement of our main result. This is a fundamental section and it is divided into three parts. We first shall introduce our assumptions and define our notation. Then in Section 2.3 we provide the precise mathematical statement of our result, and, finally, we will provide a heuristic argument that allows us to gain some intuition behind it. The next two sections then provide proofs. In Section 2.4 we apply our result to computing the most likely paths to rare events in the setting of loss queueing systems and also in the setting of ruin probabilities for large life insurance portfolios. Proofs of all the Lemmas and other technical details are included in the Appendix A.

2.2 Notations, Assumptions and Function Space

We shall describe an underlying system corresponding to an arrival rate λ. We call the system with $\lambda = 1$, i.e. one customer per unit time, our "base system"; eventually we shall send λ to infinity in our asymptotic analysis. We collect our assumptions as follows.

Assumptions and notation concerning the arrival process. For the base system, we assume the interarrival times are i.i.d. positive random variables $(U_n : n \geq 1)$ with $E[U_n] = 1$ and finite exponential moments in a neighborhood of the origin; in precise words, $\kappa(\theta) := \log Ee^{\theta U_n} < \infty$ for some $\theta > 0$. Besides, we also assume that $(U_n : n \geq 1)$ are non-lattice in the sense that there does not exist any constant $\alpha > 0$ such that the value of U_n lies in $\{\alpha k : k = 0, 1, 2, ...\}$. In our λ-scaled system, the arrivals come λ times faster (i.e. the n-th interarrival times becomes U_n/λ). The associated logarithmic moment generating function of the λ-scaled service times is then $\kappa_\lambda(\theta) := \log Ee^{\theta U_n/\lambda} = \kappa(\theta/\lambda)$. Hence, following the assumptions on the base system, $\kappa_\lambda(\theta) < \infty$ for some $\theta > 0$.

The time at which the n-th arrival occurs in the base system is $A_n = U_1 + ... + U_n$ for $n \geq 1$. We simply define $A_0 := 0$ and then let $N(t) := \max\{n \geq 0 : A_n \leq t\}$ be the number of arrivals that have occurred up to time t in the base system. It is important to keep in mind that $N(\cdot)$ increases by one unit at discontinuity points since we are assuming that the U_n's are positive.

Eventually, we shall increase the arrival rate, so it is sensible to define $N_\lambda(t) := N(\lambda t)$.

Define the so-called infinitesimal logarithmic moment generating function for the arrival

process via $\psi_N(\theta) = -\kappa^{-1}(-\theta)$ (see [30]). This definition is motivated by the fact that

$$\lim_{t\to\infty} \lambda^{-1} \log E \exp\left(\theta[N_\lambda(t+\delta) - N_\lambda(t)]\right) = \psi_N(\theta)\,\delta \tag{2.2.1}$$

for any $\delta > 0$. Since the U_n's are positive with probability one we have that $\psi_N(\cdot)$ is continuous and strictly convex on the positive line. We also assume that $\psi_N(\cdot)$ is continuously differentiable throughout $\mathbb{R}$. This assumption is satisfied for most arrival processes, certainly for interarrival times that are strictly positive and such that $\sup\{\kappa(\theta) : \kappa(\theta) < \infty\} = \infty$.

Assumptions and notation concerning the service times. We assume that the n-th customer that arrives to the base system (i.e. at time A_n) brings up a service requirement of size V_n that is independent of the arrival process. The sequence $(V_n : n \geq 1)$ is assumed to be i.i.d. and is independent of the arrivals $(U_n : n \geq 1)$. We write $F(x) = P(V_n \leq x)$ to denote the associated distribution function evaluated at x, and set $\bar{F}(x) := 1 - F(x)$ to be the tail distribution. Moreover, we assume that $F(\cdot)$ is continuous.

Two-parameter representation of system status. For any fixed $0 < T < \infty$, let $\bar{Q}_\lambda(t,y)$ denote the number of customers who arrived before or at time t and leave after time y in the λ-scaled system for all $(t,y) \in [0,T] \times [0,\infty)$. In detail,

$$\bar{Q}_\lambda(t,y) = \begin{cases} \bar{Q}_\lambda(0, y-t) + \sum_{n=1}^{N_\lambda(t)} I\left(V_n + A_n/\lambda > y\right) & t \leq y, \\ \bar{Q}_\lambda(y,y) + N_\lambda(t) - N_\lambda(y) & t > y. \end{cases}$$

We shall assume that the system is initially empty at the beginning. This is done for simplicity. Since we have infinitely many servers, we can incorporate the initial configuration by keeping track of its evolution independently of what occurs subsequently. Given our assumption of an initial empty system we then have that $\bar{Q}_\lambda(0,u) = 0$ for all $u \geq 0$. Note that for all $(t,y) \in [0,T] \times [0,\infty)$,

$$\bar{Q}_\lambda(t,y) = \bar{Q}_\lambda(t \wedge y, y) + N_\lambda(t) - N_\lambda(t \wedge y). \tag{2.2.2}$$

It is worth comparing the current system representation with the more common one involving the quantity $Q_\lambda(t,u)$ defined as the number of customers in the system currently at time t who have residual service time larger than $u \geq 0$; more precisely,

$$Q_\lambda(t,u) = \bar{Q}_\lambda(t,u+t). \tag{2.2.3}$$

These two system representations are equivalent in the sense that $(Q_\lambda(t,u) : t \in [0,T], u \geq 0)$ encodes the evolution of the infinite server systems and thus, such evolution can be used in principle to retrieve $(\bar{Q}_\lambda(t,u) : t \in [0,T], u \geq 0)$. We have chosen the representation based on $\bar{Q}_\lambda$ to facilitate the representation of the rate function; a more detailed discussion is given towards the end of Section 2.3.1. In addition, the representation based on $\bar{Q}_\lambda$ allows to obtain a rich large deviations principle to which one can apply the contraction principle directly to several continuous functions of interest. For instance, it follows immediately that the arrival process $N_\lambda(t) = \bar{Q}_\lambda(t,0)$, and the departure process, $D_\lambda(t) := N_\lambda(t) - \bar{Q}_\lambda(t,t)$ are continuous functions under the topology that we consider (and that we shall discuss in the next paragraphs). More applications of the contraction principle will be discussed in Section 2.4.

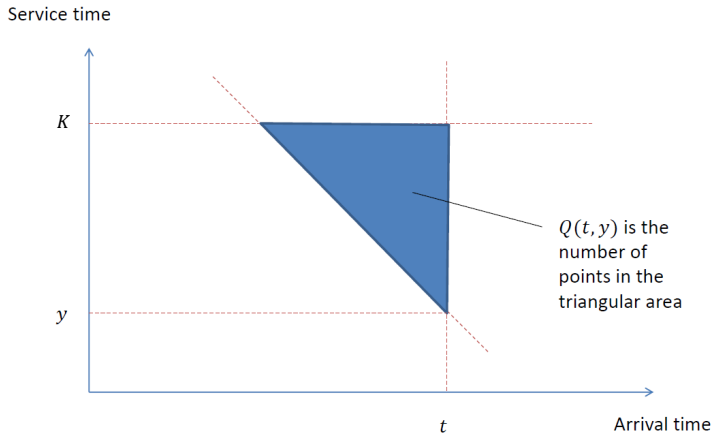


Figure 2.1: Representation of $Q(t,y)$

Discussion about the topological space. Let $\mathcal{D} = \{(t,y) : 0 \leq t \leq T, y \geq 0\}$ and let us write $||\cdot||_{\mathcal{C}}$ to denote the supremum norm over any set $\mathcal{C}$. The space of functions that we consider for our large deviations principle shall be denoted by $L_{+,\infty}(\mathcal{D})$ and it corresponds to bounded functions with domain in $\mathcal{D}$, such that $x(0,u) = 0$ for $u \geq 0$, $x(t,\cdot)$ is non increasing, and $x(t,\cdot)$ vanishes at infinity. We will develop the large deviations principle for the family of stochastic processes $(\bar{Q}_\lambda/\lambda : \lambda > 0)$ on the space $L_{+,\infty}(\mathcal{D})$ endowed with the topology generated by the supremum norm. Following [22] p. 4, the probability measures in path space in our development are assumed to have been completed.

Our large deviations principle for $\bar{Q}_\lambda/\lambda$ immediately implies in particular a large deviations principle in the Skorokhod topology in the space $D_{D_{\mathbb{R}}[0,\infty)}[0,T]$ which is the space of right-continuous-with-left-limits (RCLL) functions x, with domain on $[0,T]$, that take values on the space of RCLL functions taking values on $\mathbb{R}$. That is, on each time point t in $x = (x(t) : t \in [0,T]) \in D_{D_{\mathbb{R}}[0,\infty)}[0,T]$ is a function $x(t) \in D_{\mathbb{R}}[0,\infty)$. This is precisely the topol-

ogy considered in [52], who also provide a discussion on the benefits of using this topology relative to other natural (but weaker) alternative options (see Section 2.3 in [52]).

An alternative approach that one might consider given the available results on functional weak convergence analysis of the infinite server queue, such as [58], is to interpret the space descriptor of the infinite server queue as acting on the space of tempered distributions. We believe, however, that this approach, although elegant, has important limitations in terms of assumptions and the class of functions to which the contraction principle can be directly applied to obtain other large deviations principles of interest.

2.3 Main Result

We are now ready to state our main result. Let $\bar{q} := (\bar{q}(t,y) : (t,y) \in \mathcal{D}) \in L_{+,\infty}(\mathcal{D})$. We say that $\bar{q} \in AC_+(\mathcal{D})$ if the following conditions hold:

i) $\bar{q}$ is absolutely continuous on $\mathcal{D}$ in the sense that $\forall\ \varepsilon > 0$, $\exists\ \gamma > 0$ such that $\forall\ (t,y)$ and $(t',y') \in \mathcal{D}$, $|\bar{q}(t,y) - \bar{q}(t',y')| < \varepsilon$ if both $|t-t'|$ and $|y-y'| < \gamma$. Besides,

$$\int_0^T \int_0^\infty \left| \frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right| dy dt < \infty,$$

ii) $\partial^2 \bar{q}(t,y)/(\partial t \partial y) = 0$ almost everywhere for $(t,y) \in \{(t,y) : 0 \le y \le t \le T\}$,

For $\bar{q} \in AC_+(\mathcal{D})$, we define $I(\bar{q})$ via the following expression

$$\sup_{\theta(\cdot,\cdot) \in C_b(\mathcal{D})} \int_0^T \left[\int_t^\infty \theta(t, y-t) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right) dy - \psi_N \left(\log \left(\int_0^\infty e^{\theta(t,y)} dF(y) \right) \right) \right] dt \quad (2.3.1)$$

where $C_b(\mathcal{D})$ is the set of all bounded continuous functions on $\mathcal{D}$. On the other hand, if $\bar{q} \in L_{+,\infty}(\mathcal{D})$ fails to satisfy any of the conditions i) to ii), simply let $I(\bar{q}) = \infty$.

We now can state our main result.

Theorem 2.3.1. *Under the set of assumptions discussed in Section 2.2,* $(\bar{Q}_\lambda/\lambda : \lambda > 0)$ *satisfies a large deviations principle with good rate function* $I(\cdot)$ *on the space* $(L_{+,\infty}(\mathcal{D}), ||\cdot||_{\mathcal{D}})$. *In precise terms, for each open set O we have that*

$$\underline{\lim}_{\lambda\to\infty} \log \frac{1}{\lambda} P(\bar{Q}_\lambda/\lambda \in O) \geq - \inf_{q\in O} I(q),$$

and for each closed set C

$$\overline{\lim}_{\lambda\to\infty} \log \frac{1}{\lambda} P(\bar{Q}_\lambda/\lambda \in C) \leq - \inf_{q\in C} I(q).$$

As mentioned earlier, an immediate corollary that we can obtain is a large deviations principle for $(Q_\lambda/\lambda : \lambda > 0)$ under the Skorokhod topology in the space $D_{D_{\mathbb{R}}[0,\infty)}[0,T]$, discussed in the previous section and introduced in [52].

We shall explain the strategy behind the proof of Theorem 2.3.1. At the foremost, we shall introduce an auxiliary continuous process $\tilde{Q}_\lambda/\lambda$, defined as (2.3.7) in Section 2.3.3, that is exponentially equivalent to $\bar{Q}_\lambda/\lambda$. Then, the proof strategy composes of two parts. In the first part (Section 2.3.3), in addition to the assumptions imposed in Section 2.2 we will assume that there exists a deterministic constant $K \in (0,\infty)$ such that $P(V_n \in [0,K]) = 1$. In the second part (Section 2.3.4) of the argument we will relax this truncation assumption.

In turn, the first part of the argument (i.e. assuming truncation) is divided into several steps. The first step consists in developing the large deviations principle for $\tilde{Q}_\lambda/\lambda$ with rate $I(\cdot)$ under the topology of pointwise convergence using the Dawson-Gartner projective limit theorem (see Chapter 4.6 in [22]). The second step involves showing that $\tilde{Q}_\lambda/\lambda$ is exponentially tight as $\lambda \to \infty$ under the uniform topology on the compact set $[0,T] \times [0,K]$. The third and last step involves lifting the large deviations principle to the uniform topology.

In the second part of our argument we introduce an approximation scheme that proceeds by ignoring the customers that arrive to the system with a service time larger than K. Using a coupling argument, the process that is obtained using this scheme is shown to be a good approximation to the original system for the purpose of computing large deviations probabilities.

However, before we do this let us provide a heuristic argument in order to guess the form of the rate function. Later we will explain what are the technical difficulties that need to be addressed.

2.3.1 Heuristics: Guessing the rate function

One can take advantage of the point process representation of the input process (i.e. the arrivals and the service times represented as a marked point process). Let us start with the case of Poisson arrivals. We shall briefly explain how to adapt the development that follows to the more general case of renewal arrivals.

Consider the scaled system with arrival rate λ and suppose that $F(\cdot)$ has a density $f(\cdot)$. The amount of customers that arrive during the time interval $[t,t+dt]$ and that bring a service requirement of size $[r,r+dr]$ is denoted by the quantity $\mathcal{M}_\lambda(t+dt,r+dr)$, which is governed

by a Poisson distribution with rate $\lambda f(r)\,dtdr$. It follows then by elementary considerations involving the Poisson distribution that for a fixed value of (t,r), $\mathcal{M}_\lambda(t+dt,r+dr)/\lambda$ satisfies a large deviations principle as $\lambda\to\infty$. In particular, we formally obtain that

$$P(\mathcal{M}_\lambda(t+dt,r+dr)/\lambda\approx\mu(t,r)\,dtdr)=\exp(-\lambda J(\mu(t,r))\,dtdr),$$

with

$$J(\mu(t,r))=\sup_{\eta(t,r)}[\eta(t,r)\mu(t,r)-\psi_N(\eta(t,r))f(r)],$$

and $\psi_N(\eta)=\exp(\eta)-1$. The supremum above is obtained formally with $\eta_*(t,r):=\log(\mu(t,r)/f(r))$.

So, by pasting independent regions of the form $[t,t+dt]\times[r,r+dr]$ together one expects that the Poisson random measure $\mathcal{M}_\lambda(\cdot)/\lambda$ would satisfy a large deviations principle under a suitable topology, so that

$$P\left(\mathcal{M}_\lambda(A\times B)/\lambda\approx\int_{A\times B}\mu(t,r)\,dtdr,\text{ for a large class }A\times B\right)\approx\exp(-\lambda\mathbf{J}(\mu))$$

with

$$\begin{aligned}\mathbf{J}(\mu)&=\int_{[0,T]\times[0,\infty)}[\eta_*(t,r)\mu(t,r)-\psi_N(\eta_*(t,r))f(r)]dtdr\\&=\sup_{\eta(\cdot,\cdot)\in C_b(\mathcal{D})}\int_{\mathcal{D}}[\eta(t,r)\mu(t,r)-\psi_N(\eta(t,r))f(r)]dtdr.\end{aligned}\tag{2.3.2}$$

Now, observe that for all $y \geq t$

$$\bar{Q}_\lambda(t,y) = \int_0^t \int_{y-s}^\infty \mathcal{M}_\lambda(s+ds, r+dr), \tag{2.3.3}$$

and $\bar{Q}_\lambda(0,y) = 0$ for $y \geq 0$. If $\bar{q}(\cdot,\cdot)$ can also be expressed as

$$\bar{q}(t,y) = \int_0^t \int_{y-s}^\infty \mu(s,r)\, drds \tag{2.3.4}$$

with

$$\mu(s,r) = -\frac{\partial^2}{\partial y \partial t}\bar{q}(t,y)\Big|_{t=s, y=s+r},$$

we can develop the large deviations results for $\bar{q}(\cdot,\cdot)$ based on a similar idea as the contraction principle. In fact, for $\bar{q}(\cdot,\cdot)$ that is absolutely continuous and $\bar{q}(0,y) = 0$ for all $y \geq 0$, the representation (2.3.4) is applicable. Therefore, one can formally compute the rate function of $\bar{Q}_\lambda(\cdot,\cdot)/\lambda$ evaluated at $\bar{q}(\cdot,\cdot)$ by evaluating $\mathbf{J}(\mu)$ for $s \in [0,T]$ and $r \in [0,\infty)$. In particular, this analysis yields that $I(\bar{q})$ is equal to

$$\begin{aligned}
&\sup_{\eta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_{\mathcal{D}} [\eta(s,r)\mu(s,r) - \psi_N(\eta(s,r)) f(r)] drds \\
&= \sup_{\eta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_{\mathcal{D}} \left[\eta(s,r)\left(-\frac{\partial^2}{\partial y \partial t}\bar{q}(s,s+r)\right) - \psi_N(\eta(s,r)) f(r)\right] drds \\
&= \sup_{\eta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_{\mathcal{D}} \left[\eta(s,u-s)\left(-\frac{\partial^2}{\partial y \partial t}\bar{q}(s,u)\right) - (\exp(\eta(s,u-s)) - 1) f(u-s)\right] duds,
\end{aligned}$$

which is, of course, equivalent to (2.3.1) in the Poisson case assuming the existence of a density $f(\cdot)$ for the distribution of the service times. The previous form of the rate function was

heuristically obtained assuming that $y \geq t$. However, since all the information of the infinite server queue is contained in the evolution of the process $(Q_\lambda(t,u) : (t,u) \in \mathcal{D})$ defined in Section 2.2 with $Q_\lambda(t,u) = \bar{Q}_\lambda(t,t+u)$, we must have that the rate function should be specified only over $\bar{q}(t,y)$ for $y \geq t$. Indeed, one can check that $\partial^2 \bar{q}(t,y)/(\partial y \partial t) = 0$ for $0 \leq y < t \leq T$ as $\bar{Q}_\lambda(t+\Delta t, y+\Delta y) - \bar{Q}_\lambda(t+\Delta t, y) - \bar{Q}(t, y+\Delta y) + \bar{Q}(t,y) = 0$ for all $y < t$.

For the non-Poisson case one can argue using renewal arguments. We need to compute the log-moment generating function of the vertical strip $(\mathcal{M}_\lambda(t+dt, r_i+dr) : 1 \leq i \leq n)$, where $r_1 < r_2 < ... < r_n$ for an arbitrary partition $(r_i : 1 \leq i \leq n)$. We obtain, using elementary properties of the multinomial distribution together with an application of the key renewal theorem as in [28],

$$
\begin{aligned}
& E\left[\exp\left(\sum_{i=1}^{n} \theta(t,r_i)\, \mathcal{M}_\lambda(t+dt, r_i+dr)\right)\right] \\
=& E\left[\left(\sum_{i=1}^{n} \exp(\theta(t,r_i))\, P(V_1 \in [r_i, r_i+dr])\right)^{N(\lambda(t+dt))-N(\lambda t)}\right] \\
=& E\left[\exp\left([N(\lambda(t+dt)) - N(\lambda t)] \log\left(\sum_{i=1}^{n} \exp(\theta(t,r_i))\, P(V_1 \in [r_i, r_i+dr])\right)\right)\right] \\
=& \exp\left(\lambda \psi_N\left(\log\left(\sum_{i=1}^{n} \exp(\theta(t,r_i))\, P(V_1 \in [r_i, r_i+dr])\right)\right) + o(\lambda)\right)
\end{aligned}
$$

as $\lambda \to \infty$.

So, by pasting together vertical strips (i.e. ranging the parameter t) we obtain that the family of random measures $\mathcal{M}_\lambda(\cdot)/\lambda$ is expected to satisfy a large deviations principle under a suitable

topology with rate function

$$\mathbf{J}(\mu) = \sup_{\theta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_{[0,T]} \left[\int_0^\infty \theta(t,r)\,\mu(t,r)\,dr - \psi_N\left(\log\left(\int_0^\infty \exp(\theta(t,r))\,dF(r) \right) \right) \right] dt.$$

The rest of the formal analysis proceeds similarly as in the Poisson case.

The formal argument just outlined, even if heuristic, suggests a potential approach to developing sample path large deviations for $\bar{Q}_\lambda/\lambda$. Namely, first develop a large deviations for the random measures $\mathcal{M}_\lambda(\cdot)/\lambda$, and then apply the contraction principle to obtain the desired large deviations result for $\bar{Q}_\lambda/\lambda$. This approach, although intuitive, will not be followed in our development. We found it easier to directly work with the topology that we wish to impose. Part of the problem involved in making the argument based on random measures rigorous in the setting of the topology that is of interest to us is that indicator functions are not continuous, so the contraction principle is not directly applicable if one is to endow the space of measures with the weak convergence topology. Of course, one can proceed by trying a different topology (stronger than weak convergence) or by trying to use the extended contraction principle. However, the technical development, we believe, would end up being more involved than the direct approach that we will follow.

An additional concern that might arise at this point is our selection of $\bar{Q}_\lambda/\lambda$ in order to represent the system status; as opposed to Q_λ/λ, which might appear more natural at first sight. Let us explain why $\bar{Q}_\lambda/\lambda$ is a more convenient object to consider. Note that if $q(s,r) = \bar{q}(s, s+$

r), then

$$\begin{aligned}\frac{\partial^2}{\partial s\partial r}q(s,r) &= \frac{\partial^2}{\partial s\partial r}\bar{q}(s,s+r)+\frac{\partial^2}{\partial r^2}\bar{q}(s,s+r)\\ &= \frac{\partial^2}{\partial s\partial r}\bar{q}(s,s+r)+\frac{\partial^2}{\partial r^2}q(s,r)\end{aligned}$$

and therefore

$$\frac{\partial^2}{\partial s\partial r}\bar{q}(s,s+r) = \frac{\partial^2}{\partial s\partial r}q(s,r)-\frac{\partial^2}{\partial r^2}q(s,r).$$

Since $Q_\lambda(t,u)/\lambda = \bar{Q}_\lambda(t,u+t)/\lambda$ and our heuristic analysis suggests that the candidate rate function of $\bar{Q}_\lambda(t,y)/\lambda$ is given by

$$\sup_{\theta(\cdot,\cdot)\in C_b(\mathcal{D})}\int_0^T\left[\int_t^\infty \theta(t,y-t)\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dy-\psi_N\left(\log\left(\int_0^\infty e^{\theta(t,y)}dF(y)\right)\right)\right]dt,$$

it is then sensible to conjecture, making $y=u+t$, a representation based on $q(t,u)=\bar{q}(t,t+u)$ via

$$\begin{aligned}&\sup_{\theta(\cdot,\cdot)\in C_b(\mathcal{D})}\int_0^T\left[\int_0^\infty \theta(t,u)\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,u+t)\right)du-\psi_N\left(\log\left(\int_0^\infty e^{\theta(t,u)}dF(u)\right)\right)\right]dt\\ &= \sup_{\theta(\cdot,\cdot)\in C_b(\mathcal{D})}\int_0^T\left[\int_0^\infty \theta(t,u)\left(-\frac{\partial^2}{\partial t\partial u}q(t,u)+\frac{\partial^2}{\partial u^2}q(t,u)\right)du-\psi_N\left(\log\left(\int_0^\infty e^{\theta(t,u)}dF(u)\right)\right)\right]dt.\end{aligned} \tag{2.3.5}$$

This representation, in turn, suggests that for the rate function to be finite at $q(\cdot)$, one might need to impose as a necessary condition the existence of $\partial^2 q(t,u)/\partial^2 u$. Nevertheless, as we

shall see in our examples, one might have a finite-valued rate function even in cases in which $\partial q(t,\cdot)/\partial u$ is not even continuous for every value of $t \in (0,T)$.

2.3.2 An Auxiliary Continuous Process

In order to prove Theorem 2.3.1 we introduce an auxiliary approximating continuous process, $\tilde{Q}_\lambda$, which shall be shown to be exponentially equivalent to the process of interest $\bar{Q}_\lambda$ in the uniform norm. The construction of $\tilde{Q}_\lambda$ which is based on simple polygonal interpolations is explicitly given in the Appendix (see Section A.1). First, we show that one can construct a continuous process $(Q^*_\lambda(t,y) : t \in [0,T], y \geq 0)$ such that $Q^*_\lambda(t,\cdot)$ is non-increasing for each $t \in [0,T]$ and satisfying $\|Q^*_\lambda - Q_\lambda\| \leq 2$. Then, we define our auxiliary process $\tilde{Q}_\lambda(t,y)$ for $y \geq t$ via

$$\tilde{Q}_\lambda(t,y) = Q^*_\lambda(t,y-t), \tag{2.3.6}$$

which is the analogue of (2.2.3). Finally, we define $\tilde{Q}_\lambda(t,y)$ for $0 \leq y \leq t \leq T$ as follows. First let $\tilde{N}_\lambda(\cdot)$ be the continuous process obtained by the polygonal interpolation of $N_\lambda(\cdot)$, so that $\tilde{N}_\lambda(0) = 0$ and $\tilde{N}_\lambda(A_k/\lambda) = N_\lambda(A_k/\lambda)$ for all $k \geq 1$. Then, for $y < t$ define

$$\tilde{Q}_\lambda(t,y) = \tilde{Q}_\lambda(y,y) + \tilde{N}_\lambda(t) - \tilde{N}_\lambda(y), \tag{2.3.7}$$

analogous to (2.2.2). Observe that $\|\tilde{N}_\lambda - N_\lambda\| \leq 1$. It follows from the triangle inequality and expressions (2.2.3), (2.3.6) and (2.3.7) that

$$||\tilde{Q}_\lambda - \bar{Q}_\lambda||_{\mathcal{D}} \leq 4, \tag{2.3.8}$$

where $||\cdot||_{\mathcal{D}}$ represents the uniform norm over the set $\mathcal{D}$.

2.3.3 The Sketch of Proof: Bounded Service Time

In addition to the assumptions imposed in Section 2.2 here we also assume that $P(V_n \in [0,K]) = 1$ for $K \in (0,\infty)$.

We define $\mathcal{D}_K = \{(t,u) : 0 \leq t \leq T, 0 \leq u \leq K+T\}$ and let $C_+(\mathcal{D}_K)$ be the space of functions $(x(t,u) : (t,u) \in \mathcal{D}_K)$ such that $x(\cdot)$ is continuous in both components, $x(t,\cdot)$ is non-increasing on $[0,K+T]$, and $x(0,u) = 0$ for $u \geq 0$. Following the same notation in Section 2.2, we say that $x(\cdot,\cdot) \in AC_+(\mathcal{D}_K)$ if $x(\cdot,\cdot)$ is absolutely continuous, and $\partial^2 x(t,y)/(\partial t \partial y) = 0$ almost everywhere on $0 \leq y \leq t \leq T$.

Our initial goal is to obtain a large deviations principle for $(\tilde{Q}_\lambda/\lambda : \lambda > 0)$ as $\lambda \to \infty$ on the space $(C_+(\mathcal{D}_K), ||\cdot||_{\mathcal{D}_K})$; we then will use (2.3.8) to obtain the corresponding large deviations principle for $(\bar{Q}_\lambda/\lambda : \lambda > \infty)$.

We start by deriving a large deviations principle in the topology of pointwise convergence. The proof of this result will be given in Appendix A.2.

Lemma 2.3.2. *Let $\mathcal{X}$ consist of all the maps from $\mathcal{D}_K$ to $\mathbb{R}$, and we equip $\mathcal{X}$ with the topology of pointwise convergence on $\mathcal{D}_K$. Then $\tilde{Q}_\lambda/\lambda$ satisfies a large deviations principle with good rate*

function $I(\bar{q})$ defined by

$$\sup_{\theta(\cdot,\cdot)\in C[0,T]\times[0,K]} \int_0^T \left[\int_t^{K+t} \theta(t,y-t) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right) dy - \psi_N \left(\log \left(\int_0^K e^{\theta(t,y)} dF(y) \right) \right) \right] dt \tag{2.3.9}$$

if $\bar{q}(\cdot) \in AC_+(\mathcal{D}_K)$, and $I(\bar{q}) = \infty$ otherwise. Here $C[0,T]\times[0,K]$ denotes the set of all continuous functions on $[0,T]\times[0,K]$.

In order to lift the large deviations principle indicated in Lemma 2.3.2 to the uniform topology we need the following result on exponential tightness; we shall also give the proof of this result in Appendix A.2.

Lemma 2.3.3. *$\tilde{Q}_\lambda/\lambda$ is exponentially tight in $C_+(\mathcal{D}_K)$ equipped with the topology of uniform convergence.*

Using the previous two lemmas we are ready to state and prove the main result of this section, which is a version of Theorem 2.3.1 for the case of bounded service times.

Theorem 2.3.4. *$\tilde{Q}_\lambda/\lambda$ satisfies a large deviations principle with good rate function defined in* (2.3.9) *under the uniform topology on $\mathcal{D}_K$.*

Proof. Since the domain of $I(\cdot)$ is a subset of $C_+(\mathcal{D}_K)$, and $\tilde{Q}_\lambda/\lambda \in C_+(\mathcal{D}_K)$ with probability 1, the large deviations principle in Lemma 2.3.2 holds in the space $C_+(\mathcal{D}_K)$ with pointwise topology, (Lemma 4.1.5 (b) in [22]). Since by Lemma 2.3.3 $\tilde{Q}_\lambda/\lambda$ is exponentially tight in $\left(C_+(\mathcal{D}_K), ||\cdot||_{\mathcal{D}_K}\right)$ the same large deviations principle holds in $\left(C_+(\mathcal{D}_K), ||\cdot||_{\mathcal{D}_K}\right)$ (Corollary 4.2.6 in [22]) and the result follows. □

As a corollary of the previous theorem we obtain that $(\bar{Q}_\lambda/\lambda : \lambda > 0)$ satisfies a large deviations principle on $(L_{+,\infty}(\mathcal{D}_K), ||\cdot||_{\mathcal{D}_K})$

Corollary 2.3.5. *The process* $(\bar{Q}_\lambda/\lambda : \lambda > 0)$ *satisfies a large deviations principle on* $(L_{+,\infty}(\mathcal{D}_K), ||\cdot||_{\mathcal{D}_K})$ *with rate function* $I(\cdot)$ *defined in* (2.3.9).

Proof. First we verify that $\bar{Q}_\lambda/\lambda$ and $\tilde{Q}_\lambda/\lambda$ are exponentially equivalent according to Definition 4.2.10 in [22]). Since the laws of $(\bar{Q}_\lambda/\lambda, \tilde{Q}_\lambda/\lambda)$ are induced by a separable stochastic process and the underlying topology is induced by the uniform norm, the set

$$\{\omega : ||\bar{Q}_\lambda/\lambda - \tilde{Q}_\lambda/\lambda||_{\mathcal{D}_K} > \eta\}$$

is Borel measurable (see Remark b) following Definition 4.2.10 in [22]). Now recall that by the construction of $\tilde{Q}_\lambda$ that $\|\bar{Q}_\lambda - \tilde{Q}_\lambda\| \leq 4$ a.s. Hence for any $\eta > 0$,

$$P(||\bar{Q}_\lambda/\lambda - \tilde{Q}_\lambda/\lambda||_{\mathcal{D}_K} > \eta) = 0$$

for large enough λ. Hence

$$\limsup_{\lambda\to\infty} \frac{1}{\lambda} \log P(||\bar{Q}_\lambda/\lambda - \tilde{Q}_\lambda/\lambda||_{\mathcal{D}_K} > \eta) = -\infty.$$

The result then follows by applying Theorem 4.2.13 in [22]. □

2.3.4 The Sketch of Proof: Unbounded Service Time

In this section, we will extend our result to unbounded service times. The main intuition of the extension beyond the bounded case is to justify that we can ignore in certain sense the customers who arrive with very large service time. Let us first introduce a suitable truncation scheme. For any $K > 0$ and $\bar{q} \in AC_+(\mathcal{D})$ define

$$\phi_K(\bar{q})(t,y) = \int_0^t \int_{y-w}^K -\frac{\partial^2}{\partial s \partial z}\bar{q}(s,z)\Big|_{s=w,z=r+w} dr dw \tag{2.3.10}$$

for $t \in [0,T]$ and $y := u+t \geq 0$.

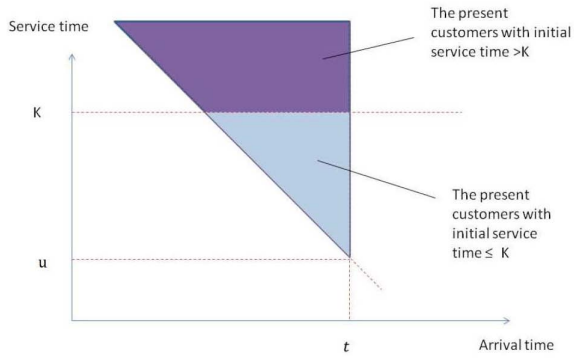


Figure 2.2: Illustration for $\phi_K(\bar{q})(t,t+u)$

Since $\bar{q}$ is absolutely continuous, $\phi_K(\bar{q})(t,u+t)$ is well defined. Moreover, the region over which the integration in (2.3.10) is performed corresponds to the triangular area depicted in light color in Figure 2.2. This region corresponds to the customers that are present at time t, have residual service time greater than y, and whose initial service time is less than K, as illustrated in Figure 2.2.

Moreover, for a sample path $\bar{Q}_\lambda$, define $\bar{Q}_{\lambda,K}$ as the two-parameter process derived from $\bar{Q}_\lambda$ by ignoring the arrivals with service time greater than K (one way to imagine is that they leave the system immediately upon arrival). Therefore, $\bar{Q}_{\lambda,K}$ is a two-parameter queue length process corresponding to an infinite server system with i.i.d. interarrival times following the law $\bar{U} = \sum_{i=1}^{G} U_i/\lambda$, where G is a geometric r.v. independent of the U_i's such that $P(G=n) = \bar{F}(K)^{n-1}F(K)$, $n \geq 1$. It is easy to check that the arrival process corresponding to $\bar{Q}_{\lambda,K}$, i.e. by ignoring the arrivals with initial service time larger than K, satisfies the conditions in Section 2.2. The service time then has the distribution function $F_K(x) = F(x)/F(K)$ for $x \in [0,K]$. We denote $(V_n^{(K)}, n = 1, \ldots)$ as the sequence of service times in this modified system.

Now recall the continuous version of $\bar{Q}_\lambda$, denoted by $\tilde{Q}_\lambda$ constructed in Section 2.3.2. Moreover, define $\tilde{Q}_{\lambda,K}$ to be the continuous approximation to $\bar{Q}_{\lambda,K}$ constructed in exactly the same fashion. In addition, for $\bar{q} \in AC_+(\mathcal{D}_K)$ define $I_K(\bar{q})$ as

$$\sup_{\theta(t,\cdot)\in C[0,T]\times[0,T+K]} \int_0^T \left[\int_t^{t+K} \theta(t, y-t) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right) dy - \psi_N^{(K)} \left(\log \left(\frac{1}{F(K)} \int_0^K e^{\theta(t,y)} dF(y) \right) \right) \right] dt, \tag{2.3.11}$$

and set $I_K(\bar{q}) = \infty$ otherwise, where $\psi_N^{(K)}$ is the infinitesimal logarithmic moment generating function corresponding to the truncated arrival process.

Theorem 2.3.4 yields that $\tilde{Q}_{\lambda,K}/\lambda$ satisfies a full large deviations principle with good rate function $I_K(\cdot)$. For $\bar{q} \in AC_+(\mathcal{D})$ we shall also evaluate $I_K(\bar{q})$ according to the expression (2.3.11).

Since the geometric r.v. G is independent of the U_i's, we can compute the associated loga-

rithmic moment generating function of the modified interarrival times

$$\kappa^{(K)}(\theta) := \kappa(\theta) + \log\left(\frac{F(K)}{1-\bar{F}(K)e^{\kappa(\theta)}}\right),$$

and from which we solve that the associated infinitesimal logarithmic moment generating function of the arrival process is

$$\psi_N^{(K)}(\theta) := \psi_N(\log(F(K)e^{\theta} + \bar{F}(K))).$$

Plugging in the above expressions into (2.3.11), we have the following expression of $I_K(\bar{q})$

$$\sup_{\theta(\cdot,\cdot)\in C(\mathcal{D}_K)} \int_0^T \left[\int_t^{T+K} \theta(t,y-t)\left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right)dy - \psi_N\left(\log\left(\bar{F}(K) + \int_0^K e^{\theta(t,y)}dF(y)\right)\right)\right]dt. \tag{2.3.12}$$

At this point our strategy involves two steps. First, we want to show that $\bar{Q}_{\lambda,K}/\lambda$ and $\tilde{Q}_{\lambda,K}/\lambda$ are exponentially good approximations as $K \nearrow \infty$ to both $\bar{Q}_\lambda/\lambda$ and $\tilde{Q}_\lambda/\lambda$ respectively. The second step consists in using this fact, together with the properties of $I_K(\bar{q})$ as $K \nearrow \infty$ and also properties of $I(\bar{q})$ to conclude the identification of the rate function of $\tilde{Q}_\lambda/\lambda$.

So, to execute the first step we first define

$$N_\lambda^{(K)}(t) = \sum_{j=1}^{N_\lambda(t)} I\left(V_j > K\right),$$

that is, $N_\lambda^{(K)}(t)$ is the number of arrivals with service time larger than K in the λ-scaled system. Then we obtain the following result, the proof of which is given in Appendix A.3.

Lemma 2.3.6. *For any* $\varepsilon > 0$,

$$\lim_{K\to\infty} \overline{\lim}_{\lambda\to\infty} \frac{1}{\lambda} \log P(N_\lambda^{(K)}(T) > \lambda\varepsilon) = -\infty. \tag{2.3.13}$$

Consequently, $\bar{Q}_{\lambda,K}/\lambda$ *and* $\tilde{Q}_{\lambda,K}/\lambda$ *are exponentially good approximations as* $K \nearrow \infty$ *to both* $\bar{Q}_\lambda/\lambda$ *and* $\tilde{Q}_\lambda/\lambda$ *respectively.*

Using the previous lemma we obtain the following result. The proof is straightforward, but following our convention we shall give it in Appendix A.3.

Lemma 2.3.7. *The family* $(\tilde{Q}_\lambda/\lambda : \lambda > 0)$ *satisfies a weak large deviations principle on* $C_+(\mathcal{D})$ *with rate function*

$$I^*(\bar{q}) := \sup_{\delta>0} \underline{\lim}_{K\to\infty} \inf_{\{z:||z-\bar{q}||_{\mathcal{D}}\le\delta\}} I_K(z).$$

We now extend the weak large deviations principle into a full large deviations principle with a good rate function using exponential tightness.

Lemma 2.3.8. *The family* $(\tilde{Q}_\lambda/\lambda : \lambda > 0)$ *is exponentially tight on* $C_+(\mathcal{D})$ *and therefore it satisfies a full large deviations principle with good rate function* $I^*(\cdot)$.

We proceed to show the identification $I^*(\bar{q}) = I(\bar{q})$. We now collect useful properties that we will need to show this identification.

Lemma 2.3.9.

i) For any $\bar{q}$ such that $I(\bar{q}) < \infty$, we have $I(\phi_K(\bar{q})) = I_K(\phi_K(\bar{q})) = I_K(\bar{q}) \nearrow I(\bar{q})$ as $K \to \infty$; the notation $I_K(\bar{q}) \nearrow I(\bar{q})$ implies that $(I_K(\bar{q}) : K > 0)$ is non-decreasing in K and convergent to $I(\bar{q})$.

ii) For any $\bar{q}$ such that $I(\bar{q}) = \infty$, and each $M > 0$, there exists a projection p_κ (following the notation introduced in the proof of Lemma 2.3.2) such that, for large enough K,

$$I_K(p_\kappa(\bar{q})) > M.$$

iii) Finally, with κ from ii) there exists $\varepsilon > 0$ such that if $\hat{q} \in C_+(\mathcal{D})$ and $||\hat{q} - \bar{q}||_{\mathcal{D}} < \varepsilon$ then

$$I_K(p_\kappa(\hat{q})) > M.$$

We now are ready to prove the following important result of this section.

Theorem 2.3.10. *$\tilde{Q}_\lambda/\lambda$ satisfies a large deviations principle with good rate function defined in (2.3.1) under the uniform topology on $[0,T] \times [0,\infty)$.*

Proof of Theorem 2.3.10. Given Lemma 2.3.8 all we need to show is that $I^*(\bar{q}) = I(\bar{q})$. Suppose that $\bar{q}$ is such that $I(\bar{q}) = \infty$. Then, parts ii) and iii) in Lemma 2.3.9 imply in particular that for every M, there exists K, a projection κ, and $\varepsilon > 0$ such that $I_K(p_\kappa(\hat{q})) > M$ for any $||\hat{q} - \bar{q}||_{\mathcal{D}} < \varepsilon$. Consequently, we conclude, by using the monotonicity of $I_K(\bar{q})$ as a function of K and taking subsequences, that

$$I^*(\bar{q}) = \sup_{\delta>0} \underline{\lim}_{K\to\infty} \inf_{\{z:||z-\bar{q}||_{\mathcal{D}} \le \delta\}} I_K(z) = \sup_{\delta>0} \sup_{K>0} \inf_{\{z:||z-\bar{q}||_{\mathcal{D}} \le \delta\}} I_K(z) = \infty.$$

If $I(\bar{q}) < \infty$, then note that

$$I^*(\bar{q}) = \sup_{\delta>0}\sup_{K>0}\inf_{\{z:||z-\bar{q}||_{\mathcal{D}}\leq\delta\}} I_K(z) = \sup_{K>0}\sup_{\delta>0}\inf_{\{z:||z-\bar{q}||_{\mathcal{D}}\leq\delta\}} I_K(z).$$

Further, observe that

$$\sup_{\delta>0}\inf_{\{z:||z-\bar{q}||_{\mathcal{D}}\leq\delta\}} I_K(z) = \sup_{\delta>0}\inf_{\{\phi_K(z):||\phi_K(z)-\phi_K(\bar{q})||_{\mathcal{D}_K}\leq\delta\}} I_K(\phi_K(z)),$$

Since $I_K(\cdot)$ is a rate function (in particular $I_K(\cdot)$ is lower semicontinuous) we have that

$$\sup_{\delta>0}\inf_{\{\phi_K(z):||\phi_K(z)-\phi_K(\bar{q})||_{\mathcal{D}_K}\leq\delta\}} I_K(\phi_K(z)) = I_K(\phi_K(\bar{q}))$$

and then by part i) of Lemma 2.3.9 we conclude that $\sup_{K>0} I_K(\phi_K(\bar{q})) = I(\bar{q})$, thus concluding that $I^*(\bar{q}) = I(\bar{q})$ as claimed. □

We finish this section with the proof of Theorem 2.3.1.

Proof of Theorem 2.3.1. All we need to show is that $\bar{Q}_\lambda/\lambda$ and $\tilde{Q}_\lambda/\lambda$ are exponentially equivalent. This follows exactly as in the proof of Corollary 2.3.5 since $\|\bar{Q}_\lambda - \tilde{Q}_\lambda\|_{\mathcal{D}} \leq 4$ a.s. The measurability issue again is dealt with using separability. The result then follows by applying Theorem 4.2.13 in [22]. □

2.4 Examples from Service and Insurance Systems

This section is devoted to two examples that apply the large deviations principle that we have developed in the previous sections. The first example is on the most likely path to overflow in a loss queue, while the second example is on the ruin of a large life insurance portfolio that embeds an infinite server queue with service cost.

Example 1. (*Finite-horizon maximum of queue length process for $M/G/\infty$*) Consider an $M/G/\infty$ queue with Poisson arrivals with rate λ. Suppose that the service times have a density $f(\cdot)$ with respect to the Lebesgue measure. The system initially starts empty.

We want to find the optimal large deviations sample path to attain the event $\{\max_{0\leq t\leq T} \bar{Q}_\lambda(t,t)/\lambda \geq x\}$, for fixed T and x, as $\lambda \to \infty$; this event corresponds precisely to the event of observing a loss in a queue with λx servers, no waiting room, starting empty. Note that $g(\bar{q}) := \max_{0\leq t\leq T} \bar{q}(t,t)$ is a continuous function under the uniform norm, so the contraction principle is directly applicable.

We impose the condition that $\int_0^T \bar{F}(t)dt < x$. This condition implies that the probability for the queue to reach λx decreases exponentially fast as $\lambda \to \infty$ (Such condition will be clear when we solve the constrained optimization in a moment).

To proceed, let us first observe that $\psi_N(\theta) = e^\theta - 1$. The maximization problem in (2.3.1) can be solved and the rate function is immediately recognized as

$$\int_0^T \int_t^\infty \left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right)\left(\log\left(\frac{-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)}{f(y-t)}\right)-1\right)dydt + T,$$

which is easily seen to be a convex function of $\partial^2\bar{q}(t,y)/\partial t\partial y$. To find the optimal sample path amounts to solving the minimization problem

$$\begin{array}{ll} \min & \int_0^T \int_t^\infty \left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)\left(\log\left(\frac{-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)}{f(y-t)}\right)-1\right)dydt+T \\ \text{subject to} & \max_{0\le u\le T}\int_0^u\int_u^\infty \left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dydt \ge x, \end{array} \tag{2.4.1}$$

which is a convex optimization problem. The integral $\int_0^s\int_s^\infty \left(-\partial^2\bar{q}(t,y)/(\partial t\partial y)\right)dydt$ is equal to $\bar{q}(s,s)$ when $\bar{q}$ is absolutely continuous and the integral is finite, and $\bar{q}(s,s)$ represents the scaled queue length process at time s.

To solve (2.4.1), we first consider a fixed u in the constraint and then optimize over u. When considering u fixed we replace the constraint in (2.4.1) by $\bar{q}(u,u)\ge x$. Under this new constraint, it suffices to look at the time 0 to s in the objective function, that is, we now solve

$$\begin{array}{ll} \min & \int_0^u \int_t^\infty \left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)\left(\log\left(\frac{-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)}{f(y-t)}\right)-1\right)dydt+u \\ \text{subject to} & \int_0^u\int_u^\infty \left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dydt \ge x. \end{array} \tag{2.4.2}$$

The solution to (2.4.1) is then the optimal sample path from (2.4.2), among $0\le u\le T$, that gives the smallest objective.

We now consider (2.4.2). Introducing a Lagrange multiplier $\mu\ge 0$, we minimize

$$\int_0^u\left(\int_t^\infty\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)\left(\log\left(\frac{-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)}{f(y-t)}\right)-1\right)dy-\mu\int_u^\infty\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dy\right)dt.$$

By a formal application of Euler-Lagrange equations, we differentiate the integrand with respect

to $-\partial^2\bar{q}(t,y)/\partial t\partial y$ to get

$$\begin{cases} \log\left(\frac{-\partial^2\bar{q}(t,y)/\partial t\partial y}{f(t-y)}\right)=0 & \text{for } t\le y\le u \\ \log\left(\frac{-\partial^2\bar{q}(t,y)/\partial t\partial y}{f(t-y)}\right)-\mu=0 & \text{for } y>u \end{cases}$$

which gives

$$-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)=\begin{cases} f(y-t) & \text{for } t\le y\le u \\ \mu f(y-t) & \text{for } y>u \end{cases}$$

for some $\mu\ge 1$ (we replace e^{μ} by another dummy μ for convenience). Complementary slackness then implies

$$\int_0^u\int_u^\infty(-\partial^2\bar{q}(t,y)/\partial t\partial y)dydt=\int_0^u\int_u^\infty\mu f(y-t)dydt=x,$$

which in turn gives

$$\mu=\frac{x}{\int_0^u\bar{F}(t)dt}$$

(note that we have assumed $\int_0^u\bar{F}(t)dt<x$ and so the condition $\mu>1$ is satisfied). As a result

$$-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)=\begin{cases} f(y-t) & \text{for } t\le y\le u \\ \frac{xf(y-t)}{\int_0^u\bar{F}(t)dt} & \text{for } y>u. \end{cases}$$

The optimal sample path $\bar{q}(t,y)$ leading to the constraint $\bar{q}(u,u)\ge x$ is given by

$$\begin{aligned}\bar{q}(t,y)&=\int_0^t\int_y^\infty\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(s,w)\right)dwds \\ &=\int_0^t\left(\int_{y\wedge u}^u f(w-s)dw+\int_{u\vee y}^\infty\frac{xf(w-s)}{\int_0^u\bar{F}(r)dr}dw\right)ds\end{aligned}$$

Transforming into $q(t,y) = \bar{q}(t,y+t)$ and some simple calculus gives the optimal sample path

$$q(t,y) = \int_y^{y+t} \bar{F}(s)ds - \int_{u-t}^{u} \bar{F}(s)ds + \frac{x\int_{u-t}^{u} \bar{F}(s)ds}{\int_0^u \bar{F}(r)dr}, \text{ for } y+t \leq u,$$

and

$$q(t,y) = \frac{x\int_y^{y+t} \bar{F}(s)ds}{\int_0^u \bar{F}(r)dr}, \text{ for } y+t > u.$$

In connection to our discussion about the direct rate function representation in terms of Q_λ/λ (see equation (2.3.5), in Section 2.3.1), one can check that $\partial q(t,y)/\partial y$ is not continuous on the line $y = t$ and therefore $\partial^2 q(t,y)/\partial y^2$ does not exists though $I(\bar{q})$ is finite.

Note also that the objective is

$$\begin{aligned}
&\int_0^u \left(-\int_t^u f(y-t)dy + \int_u^\infty \frac{xf(y-t)}{\int_0^u \bar{F}(t)dt} \left(\log\left(\frac{x}{\int_0^u \bar{F}(t)dt} \right) - 1 \right) dy \right) dt + u \\
&= \int_0^u \bar{F}(t)dt + \left(\log\left(\frac{x}{\int_0^s \bar{F}(t)dt} \right) - 1 \right) x. \qquad (2.4.3)
\end{aligned}$$

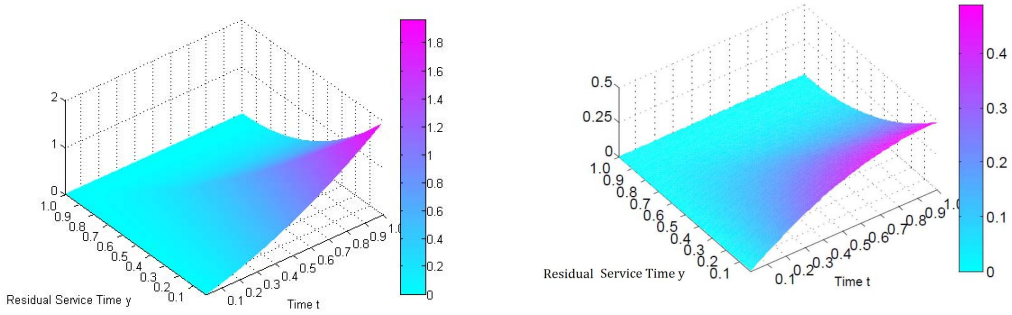
This is the rate function corresponding to the probability $P(\bar{Q}_\lambda(u,u) \geq \lambda x) = P(Q_\lambda(u,0) \geq \lambda x)$, where $Q_\lambda(u,0)$ is the queue length at time u. This rate of decay is consistent with direct calculation using the fact that $Q_\lambda(u,0)$ is a Poisson random variable with rate $\lambda\int_0^u \bar{F}(t)dt$, which gives

$$P(Q_\lambda(u,0) > \lambda x) = \sum_{n \geq \lambda x} e^{-\lambda\int_0^u \bar{F}(t)dt} \left(\lambda \int_0^u \bar{F}(t)dt \right)^n / n!.$$

For a consistency check, our result here can in fact recover the large deviations for the arrival process itself. If one changes the constraint in (2.4.2) to $\bar{q}(u,0) = \int_0^u \int_0^\infty \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right) dydt \geq$

x, the optimal value of (2.4.2) then becomes $x[\log(x/u)-1]+u$, which coincides with the exponential decay rate of $P(\text{Poisson}(\lambda u) > \lambda x)$ as $\lambda \to \infty$.

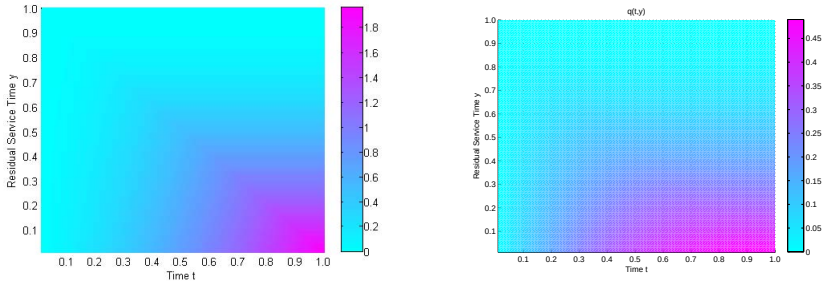
Figures 2.3 and 2.4 illustrate both the law of large numbers (i.e. the typical path) and the most likely path to the overflow event $\max_{0\le u\le T} \bar{Q}_\lambda(u,u)/\lambda \ge x$ for $T=1$, $x=2$. The underlying service time distribution is uniform in the interval $[0,1]$. We can see that the optimal path of $Q(t,y)$ increases gradually over time to overflow at time 1.



(a) $Q(t,y)$ for the most likely path to overflow (surface)

(b) $Q(t,y)$ for the unconditional path (surface)

Figure 2.3: Surface plots of the asymptotic surface $Q_\lambda(t,y)/\lambda$, as λ increases, both an optimal (most likely) path leading to overflow, and the unconditional path.



(a) $Q(t,y)$ for the most likely path to overflow (contour)

(b) $Q(t,y)$ for the unconditional path (contour)

Figure 2.4: Contour plots of the asymptotic surface $Q_\lambda(t,y)/\lambda$, as λ increases, both an optimal (most likely) path leading to overflow, and the unconditional path.

It is easy to see that since we assume $\int_0^T \bar{F}(u)du < x$, the rate function (2.4.3) is non-

decreasing in s, and as a result an optimal time horizon is T. If the service time has bounded support $[0,K]$ with $K < T$, then the selection of any time $s \in [K,T]$ will give an optimal sample path.

Example 2. (*Insurance risk process*) The net reserve of a life insurance company consists of the premium collected from policyholders, deducted by the benefit paid to policyholders in the event of deaths; often all these payments are discounted at zero in order to recognize the value of money in time. When policyholders arrive at the insurance company over time (an arrival is interpreted as the moment when a contract is signed), one can model the net assets of the insurer as a function of the underlying arrivals and death events of policyholders. Specifically, we shall assume that policyholders arrive according to a Poisson process with rate λ, and that the time-until-death upon arrival of the policyholders are independent and identically distributed. Moreover, we assume that the time-until-death upon arrival has density $f(\cdot)$, distribution function $F(\cdot)$, and tail distribution $\bar{F}(\cdot)$. The time-until-death in this setting can be thought as the service time in the queueing context. We shall assume without loss of generality that the initial net reserve of the company is zero.

It is often more convenient to work with the negative net reserve process, also known as the *aggregate loss process*, defined as the total benefit that the insurer has paid up to time t, minus the total premium received up to time t. For a policyholder who arrives at time A_i, and who dies at time $A_i + V_i < t$, the payoff by the insurer, discounted at time zero, is denoted $h_1(A_i, A_i + V_i)$; here A_i and V_i are the arrival time and time-until-death at the time of arrival of the policyholder. This quantity, $h_1(s,y)$, for $y \geq s$, captures the benefit paid at y minus the accumulated premium

collected from time s to y. On the other hand, for a policyholder who has arrived prior to t, at time A_i, and who is still alive at time t, the payoff from the insurer to the policyholder is $h_2(A_i,t)$ (typically $h_2(A_i,t)$ will be negative as it represents premium that are paid to the insurer, so the payoff is negative). Here $h_2(s,t)$, for $t \geq s$, captures the premium accumulated from s up to the present time t, discounted to obtain the net present value at time zero.

Consider, for instance, the setting of whole life insurance policies. That is, policies that pay a benefit b to the family of the policyholder, at the time of eventual death, in exchange of a premium which is paid at rate p continuously in time during all the time the policy was held, from arrival, up until the time of death. If the interest rate (or force of interest as it is known in the insurance setting) is constant equal to $\delta > 0$, then

$$h_1(s,y) = be^{-\delta y} - \int_s^y pe^{-\delta r}dr = be^{-\delta y} - p(e^{-\delta s} - e^{-\delta y})/\delta,$$

and

$$h_2(s,t) = -\int_s^t pe^{-\delta r}dr = -p(e^{-\delta s} - e^{-\delta t})/\delta.$$

The aggregate loss process, $S_\lambda(t)$, is represented as the net present value of the sum of the payoffs for all policyholders who arrive before t and it is given by

$$\begin{aligned} S_\lambda(t) &= \sum_{i=1}^{N_\lambda(t)} \left(I(A_i+V_i \leq t)\, h_1(A_i, A_i+V_i) + I(A_i+V_i > t) h_2(A_i,t)\right) \\ &= \int_0^t \int_s^t h_1(s,y) d\bar{Q}_\lambda(ds,dy) + \int_0^t \int_t^\infty h_2(s,y) d\bar{Q}_\lambda(ds,dy). \end{aligned}$$

We claim that $S_\lambda(\cdot)$ is a continuous function of $\bar{Q}_\lambda(\cdot)$ under the uniform topology on $\mathcal{D}$. In order to see this, define $D_\lambda(t)$ to be the number of departures by time t, that is,

$$D_\lambda(t) = N_\lambda(t) - \bar{Q}_\lambda(t,t) = \bar{Q}_\lambda(t,0) - \bar{Q}_\lambda(t,t). \tag{2.4.4}$$

Note that $D_\lambda(\cdot)$ and $N_\lambda(\cdot)$ are clearly continuous functions of $\bar{Q}_\lambda(\cdot)$. Moreover, we have that

$$\sum_{i=1}^{N_\lambda(t)} I(A_i + V_i \le t)\, h_1(A_i, A_i + V_i) = \int_0^t \int_0^t h_1(s,u)\, D_\lambda(du)\, N_\lambda(ds),$$

and therefore

$$S_\lambda(t) = \int_0^t \int_0^t h_1(s,u)\, D_\lambda(du)\, N_\lambda(ds) + \int_0^t \int_t^\infty h_2(s,y) \bar{Q}_\lambda(ds,dy).$$

Now, integration by parts shows that

$$\begin{aligned}
&\int_0^t \int_0^t h_1(s,u)\, D_\lambda(du)\, N_\lambda(ds) \\
&= \int_0^t \int_0^t \frac{\partial^2}{\partial u \partial s} h_1(s,u)\, D_\lambda(u)\, N_\lambda(s)\, ds du - N_\lambda(t) \int_0^t D_\lambda(u) \frac{\partial}{\partial u} h_1(t,u)\, du \\
&- D_\lambda(t) \int_0^t \frac{\partial}{\partial s} h_1(s,t)\, N_\lambda(s)\, ds + D_\lambda(t)\, N_\lambda(t)\, h_1(t,t).
\end{aligned} \tag{2.4.5}$$

A similar development yields

$$\int_0^t \int_t^\infty h_2(s,y)\bar{Q}_\lambda(ds,dy) = \int_t^\infty \int_0^t \bar{Q}_\lambda(s,y)\frac{\partial^2 h_2(s,y)}{\partial s \partial y}dsdy - \int_t^\infty \bar{Q}_\lambda(t,y)\frac{\partial h_2(t,y)}{\partial y}dy + \int_0^t \frac{\partial h_2(s,t)}{\partial s}\bar{Q}_\lambda(s,t)ds - \bar{Q}_\lambda(t,t)h_2(t,t). \quad (2.4.6)$$

It is now not difficult to see from (2.4.5) and (2.4.6) that indeed $S_\lambda(\cdot)$ is a continuous function of $Q_\lambda(\cdot)$ in the uniform topology on $[0,T]\times[0,\infty)$.

Consider the finite-horizon ruin probability that the negative net asset of the insurer rises above the level λx by time T. That is, the event $\{\max_{t\in[0,T]} S_\lambda(t)/\lambda \geq x\}$. We wish to solve for the most likely path that leads to this event and therefore, applying our theory, we must solve the following convex calculus of variations problem.

$$\begin{array}{ll} \text{min} & \int_0^T \int_t^\infty \left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)\left(\log\left(\frac{-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)}{f(y-t)}\right)-1\right)dydt + T \\ \text{subject to} & \max_{0\leq u\leq T}\int_0^u \left(\int_t^u h_1(t,y)\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dy + \int_u^\infty h_2(t,u)\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dy\right)dt \geq x \end{array}$$

Following the recipe of Example 1, we first consider

$$\begin{array}{ll} \text{min} & \int_0^u \int_t^\infty \left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)\left(\log\left(\frac{-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)}{f(y-t)}\right)-1\right)dydt + u \\ \text{subject to} & \int_0^u \left(\int_t^u h_1(t,y)\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dy + \int_u^\infty h_2(t,u)\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dy\right)dt \geq x. \end{array}$$

Introducing the Lagrange multiplier $\mu \geq 0$, we get

$$-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y) = \begin{cases} f(y-t)e^{\mu h_1(t,y)} & \text{for } t \leq y \leq u \\ f(y-t)e^{\mu h_2(t,u)} & \text{for } y > u. \end{cases}$$

When x is large, complementary slackness forces μ to satisfy

$$\int_0^u \left(\int_t^u f(y-t)e^{\mu h_1(t,y)} h_1(t,y) dy + \bar{F}(u-t)e^{\mu h_2(t,u)} h_2(t,u) \right) dt = x \tag{2.4.7}$$

for some $\mu > 0$. Denote the integration on the left hand side by $G(\mu)$, then

$$G'(\mu) = \int_0^u \left(\int_t^u f(y-t)e^{\mu h_1(t,y)} h_1^2(t,y) dy + \bar{F}(u-t)e^{\mu h_2(t,u)} h_2^2(t,u) \right) dt > 0.$$

Therefore, for given u, $G(\mu)$ is monotone in μ. Besides, $|G'(\mu)| \to \infty$ as $\mu \to \infty$. As a direct consequence, for any x large enough, equation (2.4.7) can be easily fit to many standard numerical solvers, and it admits a unique solution. Given μ, the optimal sample path is given by

$$\begin{aligned} \bar{q}(t,y) &= \int_0^t \int_y^\infty \left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(s,w) \right) dw ds \\ &= \int_0^t \left(\int_{y \wedge u}^u f(w-s)e^{\mu(u)h_1(s,w)} dw + \int_{u \vee y}^\infty f(w-s)e^{\mu(u)h_2(s,u)} dw \right) ds \end{aligned}$$

for $y \geq t$, and hence

$$\begin{aligned} q(t,y) &= \int_0^t \int_{y+t}^{\infty} \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(s,w) \right) dw ds \\ &= \int_0^t \left(\int_{(y+t)\wedge u}^{u} f(w-s) e^{\mu(u) h_1(s,w)} dw + \int_{u \vee (y+t)}^{\infty} f(w-s) e^{\mu(u) h_2(s,u)} dw \right) ds \end{aligned}$$

for $y \geq 0$. Note that here we highlight the dependence of μ on u. Moreover, the rate function for the fixed-time probability is

$$\int_0^u \left(\int_t^u f(y-t) e^{\mu(u) h_1(t,y)} (\mu(u) h_1(t,y) - 1) dy + \int_u^{\infty} f(y-t) e^{\mu(u) h_2(t,u)} (\mu(u) h_2(t,u) - 1) dy \right) dt. \tag{2.4.8}$$

The optimal time horizon u over $0 \leq u \leq T$ is chosen to minimize (2.4.8).

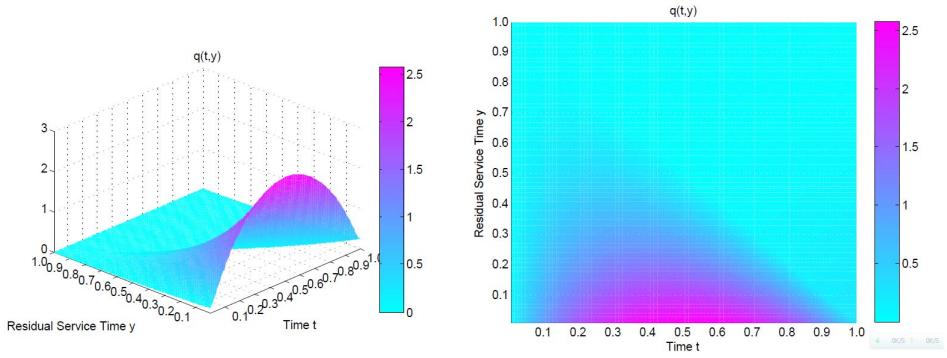
Now we consider a whole life insurance contract with benefit $b = 1.5$, continuous premium $p = 1$, zero interest rate and time-until-death which follows the uniform distribution on $[0,1]$. Our goal is to compute the optimal sample path for ruin before time $T = 1$, where we set $x = 10$ as the ruin threshold. We solve the constraint equation (2.4.7) in Matlab and obtain the optimal $u = 1$ with $\mu = 2.251$.

In this case, we can compute the optimal path

$$q(t,y) = \frac{1}{\mu^2}(e^{\mu b - \mu y} - e^{\mu b - \mu y - \mu t} - e^{\mu b - \mu + \mu t} + e^{\mu b - \mu}) + \frac{t}{\mu} e^{-\mu + \mu t} - \frac{1}{\mu^2}(e^{-\mu + \mu t} - e^{-\mu}), \text{ for } y + t \leq 1,$$

and

$$q(t,y) = e^{-\mu(2-t-y)} \left(\frac{1-y}{\mu} e^{\mu - \mu y} - \frac{1}{\mu^2}(e^{\mu - \mu y} - 1) \right), \text{ for } y + t > 1.$$



(a) $Q(t,y)$ for the most likely path to ruin (surface)

(b) $Q(t,y)$ for the most likely path to ruin (contour)

Figure 2.5: The surface and the corresponding contour plot of the asymptotic most likely path to ruin in a portfolio of life insurance policies.

These optimal paths to ruin are shown in Figure 2.5. We just show the conditional paths, as the unconditional path are identical to the figures illustrated in Example 1. The optimal path here is qualitatively very different from that of Example 1. The value of $Q(t,y)$ is the largest midway between time 0 and 1. Intuitively, it is because it requires the smallest "energy", or distortion from the law of large numbers, at such time point in contributing to a large cash outflow from the insurer.

Chapter 3

Modeling the Limit Order Book: from Order Queues to the Price-Spread Process

3.1 Introduction

Limit order book (LOB) models have recently attracted a lot of attention in the literature given their importance in modern financial markets and they are used as trading protocol in most exchanges around the world. For a brief review of worldwide financial markets that use LOB mechanism, see the first paragraph of [31]. As we shall discuss, the literature on price modeling based on LOB dynamics has mostly focused on one side of the order book, or price dynamics that are not fully informed by the LOB.

One of our main contributions is the construction of a continuous time model for the joint evolution of the mid price and the bid-ask spread (see Theorem 3.4.3 in Section 3.4). Such construction is informed by the full LOB dynamics, which we model as a multiclass queueing

system (see Section 3.2). We endow the multiclass queue with characteristics that are inspired by common stylized features which are observed empirically in order book data, such as: very fast speed of orders relative to price changes, high cancellation rates, and power-law tails (see for example Sections 3.3 and 3.5). Some of these stylized features allow us to justify the use of certain asymptotic limits and weak convergence analyses which are applied to the LOB and ultimately give rise to our continuous time pricing model.

Another contribution that it is important to highlight is that our analysis sheds light on the connection between power-law tails which are present both in the distribution of orders inside the book, and also in the realized return distributions of price processes. The connection between these features, which are documented in the statistical literature (see [10]) are explained as a result of the statements obtained in Theorem 3.3.1 and Proposition 3.3.3 in Section 3.4. Basically the power-law tails arising from the distribution of returns in the price processes can be explained as a consequence of the power-law tails in the distribution of orders inside the book, the effect of cancellation policies, and the asymptotic regime under which LOBs operate.

We establish a one-to-one correspondence between the distribution of orders inside the book and the price-return distribution assuming a specific form of the cancellation policy (see Proposition 3.3.3 for the one-to-one correspondence and Assumption 3.3.2 for the form of the cancellation policy). We argue that such cancellation policy has qualitative features observed in practice. For example, we postulate higher cancellation rates for orders that are placed closer to the spread and lower cancellation rates for orders placed away from the spread (see the discussion at the end of Section 3.3.2). Further, we also argue that the cancellation rate that we postulate is such that, in statistical equilibrium, the probability at which a given order is

executed before cancellation is roughly the same regardless of where the order is placed in book.

Although we believe that our cancellation policy is reasonable in some circumstances, more generally, our results are certainly useful to obtain insights into the form of cancellation used by market participants. This insight could be obtained by comparing the distribution of orders inside the book and the price return distribution relative to equation (3.3.1), which also contains the cancellation rates.

Furthermore, we also argue in Section 3.3.2 that a strong connection between distribution of orders inside the book and price-return distribution is to be expected in a different asymptotic regime, namely, that in which market and limit orders arrive at comparable speed, and constant cancellation rates per order across the book. We thus believe that our analysis provides significant evidence for such connection between these distributions.

We envisage our model to be useful in intra-day trading. Underlying the motivation behind the construction of our model is our belief that there is a significant amount of information in the order book which can be used to help describe the evolution of the price and bid-ask spread in the order of a few hours. At the same time, we recognize that it might be challenging in practice to keep track of the full LOB to describe price dynamics. Fortunately, as we demonstrate here, under the asymptotic regime that we utilize, implied by empirical observations, it is possible to keep track of the prices in continuous time using only a two dimensional Markov process. Our continuous-time pricing model can be calibrated, for example, from the historical data of the distribution of limit orders inside the book, and then fine tuned (through an additional parameter which we call the patience ratio, c_p) again based on historical price return distribution data (see

for example the discussion in Section 3.5 involving Empirical Observation 3). Therefore, we are able to use the information on the order book in a meaningful, yet relatively simple, way to inform the future evolution of prices. Moreover, as we shall illustrate, using simulated data, our final model captures empirical features observed in practice (see Section 3.5).

Let us discuss briefly the elements that distinguish our work from previous contributions. As we indicated earlier, we built our model from a multiclass queueing system. Queueing theory provides a natural environment for the study of LOB dynamics at a microstructure level. Consequently, it is no surprise that there is a fast growing literature that leverages off the use of queueing theory in order to analyze LOBs and the corresponding price dynamics. [16] introduces birth-death queueing model similar to our prelimit model discussed in Section 3.2. They demonstrate that this model can lead to computable conditional probabilities of interest in terms of Laplace transforms studied in queueing theory. A scaling is introduced in a subsequent paper [15] in which a continuous-time process is derived for price dynamics, but there the authors assume that price dynamics only depend on best bid and ask quotes. In contrast, we derive a two dimensional Markovian process for the best bid and best ask quotes from the whole order book model. As a result, we arrive at a limiting process which is different from that of [15]. In [49], queueing models are used to address the problem of routing of orders in a fragmented market with different books. More recently, [46]. discuss a one sided order book and track the whole process using measure-valued characteristics. Although they analyze the system in a high-frequency trading environment their scaling does not appear to highlight the role of the cancellations relative to what occurs in the markets, in which a large proportion of

the orders are actually cancelled. In contrast we not only consider the two sides of the book, but we believe our scalings better preserve the empirical features observed in practice.

As mentioned earlier, we use multiscale analysis, which allows us to replace most of the stochasticity in the book by steady-state dynamics. The paper by [64] also takes advantage of multiscale analysis, but their model is not purely derived from the microstructure characteristics at the level of arrivals of limit and market orders and therefore the ultimate model is different from the one we obtain. A recent paper by [38] provides a law of large numbers description of the order book and the price process, which is in the end deterministic and therefore also different in nature to the stochastic model we derive here. Nevertheless, we feel that the spirit of [38] is close to the work we do here.

A related literature on model building for order book dynamics relates to the use of self-exciting point processes. The approach is somewhat related to the queueing perspective, although the models are more aggregated than the ones we consider in this paper; for example, in [71] a model is discussed that considers only best bid and ask quotes but with a self-exciting mechanism with constraints (see also [12], [51] and the references therein) for more information on self-exciting processes in high-frequency trading.

The rest of Chapter 3 is organized as follows. In Section 3.2 we discuss the pre-limit model underlying the LOB dynamics. In Section 3.3 we discuss some empirical observations that inform the construction of certain approximations in our model which in particular allow to connect the price increments and the distribution of orders in the LOB. In Section 3.4 we present our asymptotic scalings and our continuous time price model. Finally, in Section 3.5 we discuss how, using simulated data, our final model captures empirical features observed in practice.

3.2 Basic Building Blocks

Ultimately, our goal is to construct a continuous time model for the joint evolution of the mid price and the bid-ask spread, which is informed by the whole order book dynamics in such a way that key stylized features are captured. In the end our model will be obtained as an asymptotic limit which is informed by stylized features observed empirically. We first discuss the building blocks of our model in the prelimit.

The building blocks of our model are consistent with prevalent limit order book models that describe the interactions between order flows, market liquidity and price dynamics, such as in [10] [16]. In most existing models, the arrival rate of limit orders corresponding, say, to a given price is given as a function of the distance between such given price and the best price of all present limit orders of the same type (buy or sell).

The best price of all limit sell (and buy) orders is called the ask (and bid) price and is usually denoted by $a(t)$ (and $b(t)$) at time t. However, as observed in recent empirical data, due to the growing popularity of algorithmic trading, limit orders are put and canceled without being executed at high frequency, especially at positions between the best ask or bid prices (fleeting orders). Therefore, the continuous observation of the best bid-ask prices may result in a process with too much "noise" due to variability caused by cancellations of such fleeting orders.

Instead, we shall construct our continuous time model by looking at the prices only at time at which an actual trade occurs; we call these quantities *prices per trade*. We believe that this is the natural time scale at which track the evolution of the LOB in order to derive a continuous time price process.

The relation between the *price-per-trade process* $(\bar{a}(\cdot),\bar{b}(\cdot))$ and $(a(\cdot),b(\cdot))$ is as follows. Suppose $\{t_k : k \geq 1\}$ are the arrival times of market orders (on both sides), then $(\bar{a}(t),\bar{b}(t)) = (a(t_k),b(t_k))$ if $t_k \leq t < t_{k+1}$. As indicated, intuitively, we can think of $(\bar{a}(\cdot),\bar{b}(\cdot))$ as a mechanism to filter out the noise made by fleeting orders in the prelimit process $(a(\cdot),b(\cdot))$. The model that we consider in the prelimit is described as follows.

Model Dynamics and Notation:

1. Limit orders or market orders arrive one at a time (i.e. there are no batch arrivals).

2. Arrivals of limit buy orders and of sell orders are modelled as two independent Poisson processes with equal rate λ.

3. Arrivals of market buy orders and of sell orders are modelled as two independent Poisson processes with equal rate μ.

4. Let $\{t_k : k \in \mathbb{Z}^+\}$ be the arrival times of the market orders (either buy or sell, so these are the arrivals of a Poisson process with rate 2μ).

5. The prices take values on the lattice $\{i\delta : i \in \mathbb{Z}^+\}$ and we observe their change at time lattice points $\{t_k : k \in \mathbb{Z}^+\}$. The parameter δ is called the tick size and in the sequel we will specify an asymptotic relationship between δ and the frequency of arrival times of orders.

6. At time t, the ask price $a(t)$ (the bid price $b(t)$) equals the minimum (maximum) of prices of all limit sell (buy) orders on the order book at time t.

7. For $t_k \leq t < t_{k+1}$, an order placed at a relative ask price equal to $i\delta$ at time t implies that the order is posted for ask price equal to $a(t_k) + i\delta$. Similarly, a relative bid price equal to $i\delta$ at time t implies an absolute bid price equal to $b(t_k) - i\delta$.

8. For $t_k \leq t < t_{k+1}$, upon its arrival at time t, a limit buy (and sell) order sits at a relative price equal to $i\delta$ with probability $p(i\delta; \bar{a}(t_k), \bar{b}(t_k))$. In particular, $p(i\delta; a, b) = 0$ for all $i\delta \geq -(a-b)/2$ so that the incoming limit buy and sell orders do not overlap with each other.

9. An order that right after time t_k sits at a relative (buy or sell) price equal to $i\delta$ is cancelled at rate $\alpha(i\delta; \bar{a}(t_k), \bar{b}(t_k))$ during the time interval $[t_k, t_{k+1})$.

10. A market order immediately transacts with any of the best matching limit orders in the order book upon its arrival.

Remark: We can actually weaken the assumption described in item 1. above to allow market orders arriving in batches as long as the size of incoming market order is less than the volume of standing limit orders at the best quote.

In our model, the ask and bid sides are two separate multi-class single-server queues with exponentially distributed times between transition of events (i.e. Markovian queues). On each side, the limit orders can be view as customers that are divided into different classes according to their relative prices (i.e. number of ticks to the best price). The class with lower relative price has higher priority. The market orders play the role of the server, as each of them causes a departure of a limit order from the best tick price. In other words, limit orders are served at

the same rate as the arrival rate of market orders and the market orders pick customers from the non-empty class with the highest priority.

It is important to note that between the arrivals of two consequent market orders, the dynamic of the limit order book is equivalent to a set of independent infinite-server systems. One such infinite-server system for each class in each of the sides (buy and sell) of the order book. The "service rate" of each such infinite server system is equal to the cancellation rate of the corresponding class.

We now proceed to develop the main ingredients of our model using stylized features that are prevalent in market data.

3.3 Empirical Observations, Price, and LOB's Distributions

We now discuss several empirical features that motivate the asymptotic regime that we consider. In particular, these observations will help us inform the asymptotic distribution of the price increments in intermediate time scales (order of several seconds).

3.3.1 Empirical Observations and Distribution of Price Increments

Empirical Observation 1: Multi-Scale Evolution of Limit Order Flows and the Occurrence of Trades . Table 3.1 is a sample from the descriptive statistics of TAQ data from [14]. In particular, we would like to hight the contrast between the daily number of updates, which include the submission, cancellation and transactions of limit orders at the best quote, and the daily number of trades (transactions). Since each market order causes a transaction of limit

orders, the fact that the daily number of updates of limit orders at the best quote is much bigger than that of trades indicate that the evolution of limit orders is much more frequent than the arrivals of market orders in the limit order book. Moreover, such difference is prevalent in both high-liquidity stocks (such as Bank of America) and low-liquidity stocks (such as CME Group). *We will adopt the relative fast speed of the limit order flows relative to market order flow (or occurrences of trade) in our asymptotic regime.*

Table 3.1: Daily average of 50 randomly chosen stocks in NYSE over 21 trading days in April 2010.

	Daily Number of best quote updates	Daily number of trades
Bank of America	1529395	15008
CME Group	38504	1412
Grand mean	223132	4552

Empirical Observation 2: Fleeting Orders and High Cancellation Rate.

Due to the prevalence of algorithmic trading in these days, the cancellation rate of limit orders has increased dramatically over recent years. Among recent studies on the Nasdaq INET data, [36] compares the data from year 1990, 1999 and 2004 and find that the cancellation rate has increased dramatically, while [37] finds that about 95% of the limit orders are canceled in the 2010 data as shown in Table 3.2. As suggested by this empirical work, the high cancellation

Table 3.2: Percentage of limit orders that are canceled without (partial) execution for 10 stocks on NASDAQ. Samples are collected in October 2010, covering 21 trading days.

GOOG	ADBE	VRTX	WFMI	WCRX	DISH	UTHR	LKQX	PTEN	STRA
97.52	92.57	93.82	95.25	92.83	94.56	95.54	96.62	91.62	95.57

rate can be attributed to the large proportion of “fleeting” limit orders which are usually put

inside the spread and then canceled immediately if not executed. For example, [37] reports that in 2010 Nasdaq data the mean inter-arrival time of market orders is about 7 seconds while the mean cancellation rate of limit orders inside the spread is less than 0.2 seconds. However, limit orders deep outside the spread are more patient. *We will assume that the arrival rates and cancellation rates are substantially higher than the arrival rates of market orders in our model.*

Recall that $\{t_k : k \geq 1\}$ is the sequence of arrival times of market orders (on both sides). We now discuss the distribution of the ask price-per-trade increment $(\bar{a}(t_{k+1}) - \bar{a}(t_k))$. Similar results on the distribution of the bid price-per-trade increment $(\bar{b}(t_{k+1}) - \bar{b}(t_k))$ follow by symmetry.

Note that the following identity of events holds

$$\{\bar{a}(t_{k+1}) - \bar{a}(t_k) \geq i\delta\} = \{\text{The queues at relative prices lower than } i\delta \text{ are all empty at time } t_{k+1}\}.$$

Let

$$\theta(i\delta; a, b) = P_\pi(\text{the queues at relative tick prices lower then } i\delta \text{ are all empty}),$$

where π is the stationary distribution of the underlying infinite server queues associated to each class (which are all independent). The i-th queue has arrival rate $\lambda p(i\delta; \bar{a}(t_k), \bar{b}(t_k))$ and cancellation rate $\alpha(i\delta; \bar{a}(t_k), \bar{b}(t_k))$. It is known that the stationary distribution of an infinite server queue with arrival rate $\lambda p(i\delta; \bar{a}(t_k), \bar{b}(t_k))$ and service rate $\alpha(i\delta; \bar{a}(t_k), \bar{b}(t_k))$ is Poisson with pa-

rameter $\lambda p(i\delta;\bar{a}(t_k),\bar{b}(t_k))/\alpha(i;\bar{a}(t_k),\bar{b}(t_k))$, therefore we have that

$$\theta(i\delta;\bar{a}(t_k),\bar{b}(t_k)) = \exp\left(-\sum_{j\delta\geq -(\bar{a}(t_k)-\bar{b}(t_k))/2}^{i-1} \frac{\lambda p(j\delta;\bar{a}(t_k),\bar{b}(t_k))}{\alpha(j\delta;\bar{a}(t_k),\bar{b}(t_k))} \right). \tag{3.3.1}$$

As $\lambda >> \mu$ is large, suggested by Empirical Observation 1, and $\alpha(\cdot) >> \mu$ as suggested by the Empirical Observation 2, we can typically approximate the distribution of the queue lengths at time t_{k+1} (given the state of the system at time t_k) by the associated steady-state distribution of the queues. More precisely, we expect the approximation $P(\bar{a}(t_{k+1}) - \bar{a}(t_k) \geq i\delta|\bar{a}(t_k),\bar{b}(t_k)) \approx \theta(i\delta;\bar{a}(t_k),\bar{b}(t_k))$ to hold. This heuristic is made rigorous in the following theorem, the proof, which is given in Appendix B.1, is based on the so-called stochastic averaging principle, see [45]. We use $D[0,\infty)$ to denote the space of right continuous with left limits functions from $[0,\infty)$ to $\mathbb{R}$ endowed with the Skorokhod J_1 topology (see [6] for reference).

Theorem 3.3.1. *Consider a sequence of LOB systems indexed by n. In the n-th system, the total number of orders in the order book is given by $q_n = q < \infty$, the distribution of these orders in the order book is assumed to satisfy $\bar{a}_n(0) = a_n(0) = \bar{a}$ and $\bar{b}_n(0) = b_n(0) = \bar{b}$. We assume that the arrival rate of market orders satisfies $\mu_n = \mu$ and the distribution of incoming limit orders is $p_n(\cdot;\cdot,\cdot) = p(\cdot;\cdot,\cdot)$ (i.e. constant along the sequence of systems). Suppose there exists a sequence of positive number $\{\xi_n : n \geq 1\}$ such that $\lambda_n = \xi_n\lambda$, $\alpha_n(\cdot) = \xi_n\alpha(\cdot)$ and $\xi_n \to \infty$ as $n \to \infty$. We also assume the regularity condition that for any $(a,b) \in \mathbb{Z}^2$*

$$\lim_{i\to\infty} \theta(i\delta;a,b) = 0. \tag{3.3.2}$$

Then the corresponding price process $(\bar{a}_n(\cdot), \bar{b}_n(\cdot))$ *converges weakly in* $D[0,\infty)$ *to a pure jump process* $(\hat{a}(\cdot), \hat{b}(\cdot))$. *The process* $(\hat{a}(\cdot), \hat{b}(\cdot))$ *jumps at time corresponding to the arrivals of a Poisson process with rate* 2μ. *The size of the jumps are not independent, in particular, if* t *is a jump time, then the jump size at* t *is a random vector following the the distribution*

$$\begin{aligned} & P\left(\hat{a}(t) - \hat{a}(t-) = i\delta, \hat{b}(t) - \hat{b}(t-) = j\delta \,|\hat{a}(t-), \hat{b}(t-)\right) \\ = \;& [\theta(i\delta; \hat{a}(t-), \hat{b}(t-)) - \theta((i+1)\delta; \hat{a}(t-), \hat{b}(t-))] \times [\theta(j\delta; \hat{a}(t-), \hat{b}(t-)) - \theta((j+1)\delta; \hat{a}(t-), \hat{b}(t-))]. \end{aligned}$$

Remark: The regularity condition (3.3.2) not only is quite natural, but it can be easily verified in terms of $p(i\delta; a, b)$ and $\alpha(i\delta; a, b)$ because there is an explicit formula for $\theta(i\delta; a, b)$ given below in equation (3.3.1).

It is important to note that in the previous result we have held the arrival rates of market orders constant, so this result simply describes the price processes at times scales corresponding to the inter-arrival times of market orders (i.e. in the order of a few seconds according to the representative date discussed earlier). In Section 3.4 we shall introduce a scaling that will allow us to consider the process in longer time scales (several minutes or longer) by increasing the arrival rate of market orders.

Theorem 3.3.1 can be extended without much complications to include more complex dynamics in the arrivals of the market orders. For instance, one way to extend our model is to allow traders to post market orders depending on the current bid-ask price. This modification can be introduced as thinning procedure to the original Poisson process with rate 2μ, the thinning parameter might depend on the observed bid-ask price $(a(\cdot), b(\cdot))$. Other examples of the

interactions between market participants that can be included in our model extensions are correlation between the buying and selling sides and dependence between arrival rate of market orders and the spread width.

3.3.2 Connecting Distribution of Price Increments and LOB's Distributions

We now will argue how Theorem 3.3.1 allows us to provide a direct connection between the increment distribution of the price-per-trade process and the distribution of orders in the LOB. We believe that this connection, although simple, is quite remarkable because it forms the basis behind our idea of using information in the LOB to predict the evolution of prices.

Let us assume the following form for the cancellation rate.

Assumption 3.3.2.

$$\alpha(i\delta;\bar{a}(t_k),\bar{b}(t_k)) = \frac{\lambda}{c_p}\cdot\frac{p(i\delta;\bar{a}(t_k),\bar{b}(t_k))}{log(1-\sum_{j\leq i-1}p(j\delta;\bar{a}(t_k),\bar{b}(t_k)))-log(1-\sum_{j\leq i}p(j\delta;\bar{a}(t_k),\bar{b}(t_k)))}, \tag{3.3.3}$$

where $c_p > 0$ is a constant we call the patience ratio of limit orders and we will see in a moment that it plays an important role in the connection between the distributions of limit order flow and price returns.

Under this assumption, by simple algebra in (3.3.1), we have the following result.

Proposition 3.3.3.

$$\theta(i\delta;\bar{a}(t_k),\bar{b}(t_k)) = (1-\sum_{j\leq i-1} p(j\delta;\bar{a}(t_k),\bar{b}(t_k)))^{c_p}, \tag{3.3.4}$$

Since $\theta(\cdot;a,b)$ is the tail of price return and $1-\sum_{j\leq\cdot} p(j;a,b)$ is the tail of the relative price of incoming limit orders. Proposition 3.3.3 indicates that the price return inherits some statistical properties of the distribution of limit order flow on the order book. In real market data, power-law tails are reported in both the relative price of limit orders and the mid-price return as discussed in the next subsection. In our model, given that the distribution of the limit order flow $p(\cdot;\bar{a},\bar{b})$ has a power-law with exponent v, in other words, $\bar{F}^{c_p}(i\delta) = \sum_{j\geq i} p(j\delta;\bar{a},\bar{b}) \approx (c_1 i\delta)^{-v}$. Then, as a direct consequence of (3.3.4), we have $P(\hat{a}(t)-\hat{a}(t-) \geq i\delta|\bar{a},\bar{b}) \approx (c_1 i\delta)^{-c_p v}$ and therefore the price returns also follows a power law but with different exponent $c_p v$.

For $c_p > 1$, our model recovers an interesting phenomenon in real market that the price return has a thinner tail than the relative price of limit orders as reported in [10]. Besides, in our model when the limit orders are more patient (c_p increases), the price return has a thinner tail. This is consistent with the fact that price volatility decreases when the market has more liquidity supply (since the limit orders stand longer in the order book as c_p increases).

We shall put some remarks on our Assumption 3.3.2 on the cancellation rate of limit orders. Although this assumption is made for mathematical tractability and because we did not find enough data to develop an empirical-based model, it is consistent with empirical observations to some level. In particular, under Assumption 3.3.2, the cancellation rate is decreasing with respect to the relative price $i\delta$ as is reported in [31] and the references therein. To see why

$\alpha(\cdot;a,b)$ is decreasing, let's assume that $p(i\delta;a,b) \leq A\delta$ for some constant $A > 0$. Since for any fixed $h > 0$ and $x > 0$ small enough, $x/(\log(a+x) - \log(a)) \approx a$, we have

$$\alpha(i\delta;a,b) \approx \frac{\lambda}{c_p}(1 - \sum_{j \leq i-1} p(j\delta;a,b))$$

when δ is small enough and hence it is decreasing in $i\delta$.

Moreover, when $c_p = 1$, $\theta(\cdot;a,b) = 1 - \sum_{j \leq i-1} p(j\delta;a,b)$. In this case, $\theta(\cdot;a,b)$ is approximately proportional to $\alpha(\cdot;a,b)$, which implies that the impatience level of a standing limit order at position $i\delta$ (namely $\alpha(i\delta;a,b)$) is proportional to its rate of execution as observed by the arriving market orders (namely $\mu\theta(i\delta;a,b)$). So, the probability that a given limit order in equilibrium at position $i\delta$ gets executed before cancellation is equal to $\mu\theta(i\delta;a,b)/(\alpha(i\delta;a,b) + \mu\theta(i\delta;a,b)) \approx \mu/(\lambda+\mu)$. Consequently, in this sense all limit orders have roughly the same probability of execution in equilibrium.

Connecting Distribution of Price Increments and LOB's Distributions in Other Regimes

We close this section by briefly discussing another asymptotic regime. Suppose that one assumes that the cancellation rate per order at relative price $i\delta$ equals $\alpha(i\delta,a,b) = \alpha$ (constant) – see for example [16]. Then one can check that the stationary distribution of the multi-class queues becomes,

$$\theta(i\delta,a,b) = \left(\sum_{l=0}^{\infty} \prod_{k=1}^{l} \frac{\lambda F(i\delta,a,b)}{\mu + k\alpha} \right)^{-1}. \tag{3.3.5}$$

So, if $\lambda_n, \mu_n \to \infty$ in such a way that $\lambda_n/\mu_n \to c'$ as $n \to \infty$ and $\alpha_n/\lambda_n \to 0$, then we obtain

$$\theta(i\delta, a, b) \approx \bar{F}(i\delta, a, b)/c',$$

and arrive at the same conclusion as in Proposition 3.3.3 which $c_p = 1$. We therefore believe that the sort of relationship that we have exposed via Proposition 3.3.3 between the return distribution and the distribution of orders in the book might be relatively robust. Under the assumption that $\lambda_n/\mu_n \to c'$, there is no stochastic averaging principle such as discussed in Theorem 3.3.1. However, one can obtain a limiting price process with price increment distributed as (3.3.5) by observing the LOB and the price in suitably chosen discrete time intervals.

3.4 Continuous Time Model

We write $\bar{s}(t) = \bar{a}(t) - \bar{b}(t)$ to denote the bid-ask spread per-trade at time t. We shall develop a stochastic model for the price-spread dynamics in longer time scale (order of several minutes or more). The model will be a jump-diffusion limit of the discrete price-spread processes as given in Section 3.2.

We will now introduce the distribution of relative prices in the LOB, $p(\cdot; \bar{a}(t_k), \bar{b}(t_k))$. We shall impose our assumptions directly on $\theta(\cdot; \bar{a}(t_k), \bar{b}(t_k))$ because we can go back and forth between $\theta(\cdot; \bar{a}(t_k), \bar{b}(t_k))$ and $p(\cdot; \bar{a}(t_k), \bar{b}(t_k))$ directly via (3.3.4). We shall consider a sequence of limit order books indexed by n and their ask-bid (per-trade) price process $\{(\bar{a}^n(\cdot), \bar{b}^n(\cdot))\}$. The dynamic of $(\bar{a}^n(\cdot), \bar{b}^n(\cdot))$ is characterized by the arrival rate of market orders and the price in-

crements. In turn, the price increments will be defined in terms of auxiliary (spread-dependent) random variables denoted by $\Delta_a^n(\bar{s}^n(t_k))$ for the ask price process and $\Delta_b^n(\bar{s}^n(t_k))$ for the buy price process. For simplicity in the notation, we often write $\Delta_a^n(t_k)$ instead of $\Delta_a^n(\bar{s}^n(t_k))$ (similarly for $\Delta_b^n(t_k)$). We will assume that both $\Delta_a^n(t_k)$ and $\Delta_b^n(t_k)$ have the same distribution given the spread-per trade, so we simply provide the description for $\Delta_a^n(t_k)$ in our following assumption which is motivated by the Empirical Observation 3.

Assumption 3.4.1. ***(Price return distribution)*** *First define,*

$$\Delta_a^n(\bar{s}^n(t_k)) = (1-I_k^a)\cdot(-1)^{R_k^a}[U_k^a/(\sqrt{n}\delta)]\delta + I_k^a[\bar{s}^n(t_k)V_k^a/(2\delta)]\delta \tag{3.4.1}$$

where:

i) I_k^a *is Bernoulli with* $P(I_k^a = 1) = q$ *for some* $q > 0$*,*

ii) U_k^a *is a random variable with support on* $[0,\xi]$ *for* $\xi \in (0,\infty)$*.*

iii) R_k^a *is Bernoulli with* $P(R_k^a = 1) = (1+2\beta/\sqrt{n})/2$ *for some* $\beta > 0$*,*

iv) V_k^a *is a continuous random variable so that* $P(V_k^a \geq -1) = 1$*.*

v) the random variables I_k^a*,* U_k^a*,* R_k^a *and* V_k^a *are independent of each other (independence is assumed to hold across* k *and for the superindices* a, b*).*

Then we let

$$\begin{cases} \bar{a}^n(t_{k+1}) = \bar{a}^n(t_k) + \Delta_a^n(\bar{s}(t_k)) \vee ([-\bar{s}^n(t_k)/(2\delta)]\delta), \\ \bar{b}^n(t_{k+1}) = \bar{b}^n(t_k) - \Delta_b^n(\bar{s}(t_k)) \vee ([-\bar{s}^n(t_k)/(2\delta)]\delta), \end{cases} \tag{3.4.2}$$

and this is equivalent to assuming

$$\theta(i\delta, \bar{a}^n(t_k), \bar{b}^n(t_k)) = P\left(\Delta_a^n(\bar{s}(t_k)) \vee ([-\bar{s}^n(t_k)/(2\delta)]\delta) \geq i\delta\right).$$

Remarks:

1. The first term $(1 - I_k^a) \cdot (-1)^{R_k^a}[U_k^a/(\sqrt{n}\delta)]\delta$ captures limit orders tend to cluster close to their respective best bid or ask prices; the parameter $(1-q) \in (0,1)$ can represents the proportion of orders that are concentrated around the best bid or ask price. Since, as observed earlier, these correspond to a substantial proportion of the total number of orders placed, we might choose $q \approx 0$.

2. The second term $I_k^a[\bar{s}^n(t_k)V_k^a/(2\delta)]$ captures the limit orders that are put far away from the current bid or ask price. In Section 3.5, we shall choose V_k^a to have a density f_V with a power-law decaying tails which are consistent with empirical observations (see Empirical Observation 3). We also postulate a multiplicative dependence on $\bar{s}^n(t_k)$ to capture the positive correlation between size of spread and variability in return distribution as reported in ([9]).

3. Recall that the most aggressive price ticks that are allowed in our pre-limit model assumptions are at the mid price; this results in the cap $\vee([-\bar{s}^n(t_k)/(2\delta)]\delta)$ appearing in (3.4.2), which consequently yields $\bar{a}^n, \bar{b}^n$.

4. The asymmetry in the distribution of R_k^a allows us to introduce a drift term in the spread, which will be useful to induce the existence of steady-state distributions. We will validate certain features of the steady-state distribution of our model vis-a-vis statistical evidence in Section 3.5.

In addition, we impose the following assumptions on time and space scalings, which are consistent with Empirical Observations 1 and 2. In order to carry out a heavy traffic approximation, we consider a sequence of LOB systems indexed by $n \in \mathbb{Z}^+$, such that in the n-th system:

Assumption 3.4.2. ***(Time and Space Scale)***

1. *The arrival rate of market orders on each side* $\mu_n = n\mu$*;*
2. *Tick size* $\delta_n \longrightarrow 0$ *so that either* $\delta_n = o\left(n^{-1/2}\right)$ *or* $\delta_n = n^{-1/2}$.
3. *We assume that* $q_n = \gamma/n$ *for some* $\gamma > 0$.

In order to explain our scaling, note that the number of jumps, corresponding to the component involving V_k^a in (3.4.1) is Poisson with rate $\gamma\mu$ so Condition 1 of Assumption 3.4.2 helps us capture jump effects in the limit. The scaling that we consider implies the existence of two types of arriving limit orders, one type that arrives more frequently than the other, see Assumption 3.4.2, part 3. in connection to Assumption 3.4.1, part i). This scaling feature, together with the fact that the probability of an order being executed is roughly constant across the book (as discussed at the end of Section 3.3.2) induces a much higher cancellation rate close to the spread, which is consistent with empirical findings.

Now we are ready to state our result on the spread and price dynamics informed by the limit order book, the proof of which is given in Appendix B.2

Theorem 3.4.3. *For the n-th system, let* $\bar{s}^n(t) = \bar{a}^n(t) - \bar{b}^n(t)$ *be the spread process and* $\bar{m}^n(t) = \bar{a}^n(t) + \bar{b}^n(t)$ *be twice of the mean price. Suppose* $(\bar{s}^n(0), \bar{m}^n(0)) = (s_0, m_0)$*. Then, under As-*

sumptions 1-4, the pair of processes $(\bar{s}^n, \bar{m}^n) \in D([0,\infty), \mathbb{R}^+ \times \mathbb{R})$ *converges weakly to* $(\bar{s}, \bar{m}) \in D([0,\infty), \mathbb{R}^+ \times \mathbb{R})$ *with* $(\bar{s}(0), \bar{m}(0)) = (s_0, m_0)$ *such that*

$$\begin{cases} d\bar{s}(t) &= -\eta dt + dW_a(t) + dW_b(t) + \bar{s}(t_-)\, dJ_1(t)/2 + \bar{s}(t_-)\, dJ_2(t)/2 + dL(t), \\ d\bar{m}(t) &= dW_a(t) - dW_b(t) + \bar{s}(t_-)\, dZ_1(t)/2 - \bar{s}(t_-)\, dZ_2(t)/2. \end{cases} \tag{3.4.3}$$

Here,

1. $\eta = 2\mu\beta E\left(\left[U_1^a\right]\right) = 2\mu\beta E\left(\left[U_1^b\right]\right)$ *if* $\delta_n = n^{-1/2}$, *and* $\eta = 2\mu\beta E\left(U_1^a\right) = 2\mu\beta E\left(U_1^b\right)$ *for* $\delta_n = o(n^{-1/2})$.

2. W_a *and* W_b *are two independent Brownian motions, each with zero mean and variance rate* $\sigma^2 = \mu E(\left[U_j^a\right]^2) = \mu E(\left[U_j^b\right]^2)$ *if* $\delta_n = n^{-1/2}$, *and* $\sigma^2 = \mu E((U_j^a)^2) = \mu E((U_j^b)^2)$ *for* $\delta_n = o(n^{-1/2})$.

3. J_1 *and* J_2 *are two i.i.d. compound Poisson processes with constant jump intensity* $\gamma\mu$ *and the jump density distribution given by the density of* V_1^a.

4. $\bar{s}(t) \geq 0$ *and* $L(t)$ *satisfies:* $L(t) = 0$, $dL(t) \geq 0$ *and* $\bar{s}(t)dL(t) = 0$ *for all* $t \geq 0$.

3.5 Simulation Results

We simulate the pair of the spread and mid-price processes according to their asymptotic approximation $(\bar{s}(\cdot), M(\cdot))$ as given by (3.4.3) under different parameters. We use the distribution

$$f_U(x) = r/\xi, \text{ for } x \in (0,\xi] \text{ and } P(U=0) = 1-r,$$

and

$$f_V(x) = \frac{(u-1)(\rho+x)^{-u}}{2(\rho^{1-u}-(c+\rho)^{1-u})} I(x \in (0,c)) + \frac{(u-1)(\rho-x)^{-u}}{2(\rho^{1-u}-(1+\rho)^{1-u})} I(x \in (-1,0)),$$

for $u \in (1,3]$, and $\rho, c > 0$. We have chosen $f_V(\cdot)$ so that in the pre-limit, the distribution of the orders inside the order book yields a power-law tail (when $i > 0$ is big) so that for fixed a, b,

$$p(i\delta; a, b) \propto \frac{1}{(c_2 + i\delta)^u},$$

for some $c_2 > 0$. This choice is justified in view of the following empirical observation.

Empirical Observation 3: Distribution of limit orders inside the order book.

Power-law decaying tails in the distribution of the relative prices of incoming limit orders inside the book have been reported in several empirical studies on order books in different financial markets (see for instance [10], [53] and [72]). Market data suggest that although incoming limit orders concentrate around the bid or ask price (according to [10], half of the limit orders have relative tick price $-1, 0$ and 1), they spread widely on the order book and the tail of

the relative price, either buy or sell, can be well approximated by a power-law with some power index $u > 0$ (i.e. the proportion of orders at i ticks away from the best quote is proportional to $1/(c_1+i)^u$ for some $c_1 > 0$). The index u varies among different financial markets as reported in [10], [53] and [72] with values $u \in (1,3]$. It is also observed that the relative price distributions are basically symmetric on the sell and buy sides. Moreover, empirical observations also show that a substantial part of the limit sell (and buy) orders is clustered close to the ask (and bid) price, as is captured by the first term involving U (see Remark 1 under Assumption 3.4.1).

We are choosing the parametric family of our price-return distribution directly to match empirical features of the distribution of orders in the book, so here implicitly we are assuming $c_p = 1$. This parameter can be adjusted to better reflect tail behavior of the empirical price-return distribution.

We proceeded to simulate the spread and mid-price processes according to their asymptotic approximation $(\bar{s}(\cdot), M(\cdot))$ as given by (3.4.3) under different parameters. We then compute the stationary distribution of the spread and the volatility process of the mid-price return from the simulation data. The computation results show that the joint jump-diffusion dynamics of the spread and mid-price (3.4.3) derived from our LOB model can capture several stylized features in real spread and price data as reported in [69], [31] and the references therein.

Stationary Distribution of the Spread: In [69], the authors study the spread size immediately before trade from Philippine Stock Exchange market data and they find that the stationary distribution of the spread is close to an exponential while in some other markets, the stationary distribution of the spread admits a power-law. We simulated the spread process

$\bar{s}(\cdot)$ according to (3.4.3) and estimate the mean $E[\bar{s}(\infty)]$ and standard deviation $std[\bar{s}(\infty)] = \sqrt{E\pi[(\bar{s}(\infty) - E[\bar{s}(\infty)])^2]}$ of the spread under its stationary distribution. In particular, we estimate expectations under the stationary distribution by the path-average of the simulated $\bar{s}(\cdot)$. The results are reported in Table 3.3 and show that the mean $E[\bar{s}(\infty)]$ is close to the standard deviation $std[\bar{s}(\infty)]$.

Figure 3.1 compares the empirical distribution of the simulated spread data of the spread $\bar{s}$ under the parameter set (b) and the exponential distribution with the same mean. Although the stationary distribution of $\bar{s}$ is roughly well fitted by the exponential distribution with the same mean as shown in Figure 3.1 (a), its tail is much heavier than exponential and resembles a power-law tail as shown in Figure 3.1 (b). Intuitively, the limit spread process (3.4.3) is a reflected Brownian motion with jumps. In this light, one could argue that the stationary distribution of the spread could be well approximated by a mixture of an exponential distribution and a power-law distribution because the reflected Brownian motion admits an exponential stationary distribution and the jump size follows some power-law distribution.

	u	β	ξ	r	$\mu\gamma$	$E[\bar{s}(\infty)]$	$std[\bar{s}(\infty)]$
(a)	2.8	0.25	0.02	0.25	6.75	0.1704	0.2068
(b)	2.3	0.25	0.02	0.25	6.75	0.1812	0.2273
(c)	2.8	0.5	0.025	0.25	6.75	0.0957	0.1056
(d)	2.8	0.25	0.02	0.5	4.5	0.1576	0.1663

Table 3.3: The mean and standard deviation of $\bar{s}$ in stationarity under different sets of parameters. $\rho = 0.02$ and $\mu = 9$ are the same for case (a) to (d).

Correlation between Spread and Volatility: We study the relation between the spread and the volatility of the mid-price return *per trade* as in [69]. In their paper, the volatility of the mid-

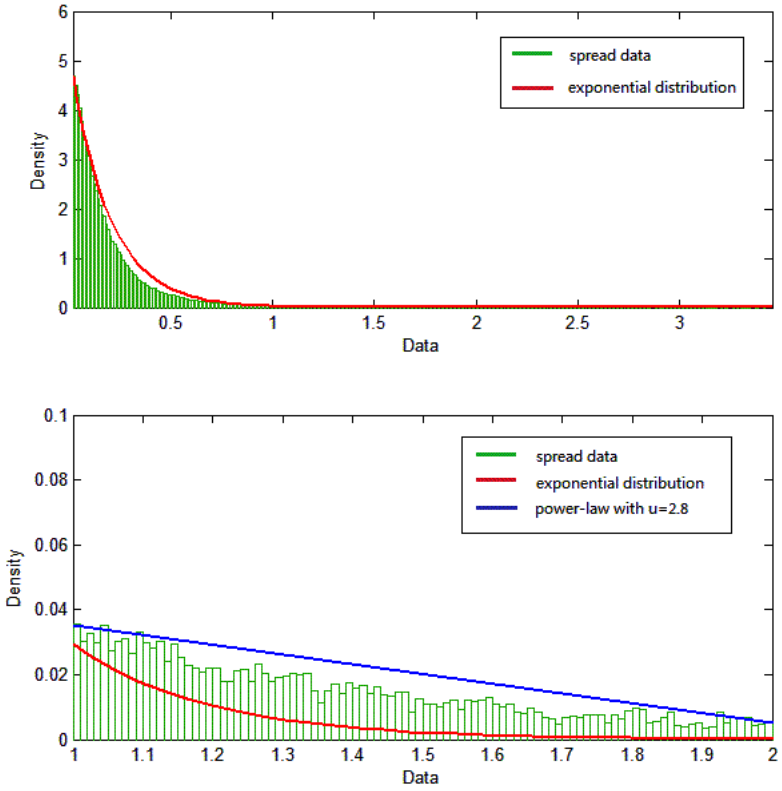


Figure 3.1: Power-law Decaying Tail of $\bar{s}$

price return per trade is computed from empirical data as $\sigma_1 = \sqrt{\sum_{i=1}^{N}(m_i - m_{i-1})^2/N}$, where N is the number of trades that has been observed and m_i is the mid-price *per trade*. They find a strong linear relationship between σ_1 and the mean spread in stationary distribution $E[\bar{s}(\infty)]$. Our model also capture the linear relationship between the volatility of the mid-price return and the spread *per trade*. To see this, we first simulate $(\bar{s}(\cdot), \bar{m}(\cdot))$ according to (3.4.3). Since $(\bar{s}(\cdot), \bar{m}(\cdot))$ is the limit of the price and spread *per trade* when the arrival rate of trades $\to \infty$, we estimate the volatility σ_1 (up to some constance) of the mid-price return per trade by

$$\hat{\sigma}_1^2 = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (\bar{m}(k\Delta t) - \bar{m}((k-1)\Delta t))^2/\Delta t,$$

and we choose $\Delta t = 0.1$ unit of time. We compute $\hat{\sigma}_1$ as the path average of the simulated path of $\bar{m}(\cdot)$. We also compute $E[\bar{s}(\infty)]$ as the path average taking at every $\Delta t = 0.1$ unit of time interval.

	u	ξ	r	μ	β	$\mu\gamma$	$E[\bar{s}(\infty)]$	$\hat{\sigma}_1$
1	2.8	0.08	0.25	12	0.25	9	0.1704	0.0822
2	2.8	0.4	0.25	12	0.25	9	0.7934	0.4074
3	2.8	0.8	0.25	12	0.25	9	1.5862	0.8169
3	2.3	0.08	0.25	12	0.25	9	0.1812	0.0885
4	2.3	0.08	0.5	6	0.25	3	0.1696	0.0848
5	2.3	0.08	0.75	4	0.25	1	0.1559	0.0812

Table 3.4: Estimation of $E[\bar{s}(\infty)]$ and $\hat{\sigma}_1$ under different parameters.

The simulation results reported in Table 3.4 indicate a linear relation between $E[\bar{s}(\infty)]$ and $\hat{\sigma}_1$ that is found in [69]. Heuristically, without the jump part in (3.4.3), $\bar{s}(\cdot)$ becomes a one dimensional reflected Brownian motion with drift $\xi\beta$ and variance coefficient $\xi^2 r\mu/3$, and the mid-price is simply a Brownian motion with variance coefficient $\xi^2 r\mu/3$. It is known that the

stationary distribution of a reflected Brownian motion is exponential and one can compute that $\hat{\sigma}_1 = \xi r\mu/(6\beta)$. Also, in the case of no jumps, we can clearly evaluate $\hat{\sigma}_1 = \xi\sqrt{r\mu/3}$. Therefore, the mean spread and the mean volatility have a linear relationship of the form $E[\bar{s}(\infty)] = l \times \hat{\sigma}_1$ with $l = \sqrt{r\mu}/(2\beta\sqrt{3})$. In Table 3.4, we choose different sets of parameters such that $\sqrt{r\mu}/(2\beta\sqrt{3}) \equiv 2$ and one can check that the estimated mean $E[\bar{s}(\infty)] \approx 2\hat{\sigma}_1$, so the effect of the jumps is actually relatively minor on the parameter ranges that we explored, for this particular performance measure.

Volatility Clustering: The jump-diffusion limit (3.4.3) also captures the volatility clustering feature in limit order book data as reported in a series of empirical studies (see Section G.2 in [31]). To see this, we measure the volatility in the mid-price process as the standard deviation of the mid-price return per 0.1 unit of time over every 10 units long time window. In detail, we compute a time series $\bar{\sigma}(t)$ from the simulated mid-price process $\bar{m}$ as

$$\bar{\sigma}(t) = \sqrt{\frac{1}{99}\sum_{i=1}^{100}(\bar{m}(10t+0.1i) - \frac{1}{100}\sum_{j=1}^{100}\bar{m}(10t+0.1j))^2}.$$

To illustrate volatility clustering, in Figure 3.2, we compare the original time series of the volatility we have computed and its random permutation. In the original time series, peaks are gathering together while in the permuted time series, peaks are uniformly distributed along the time.

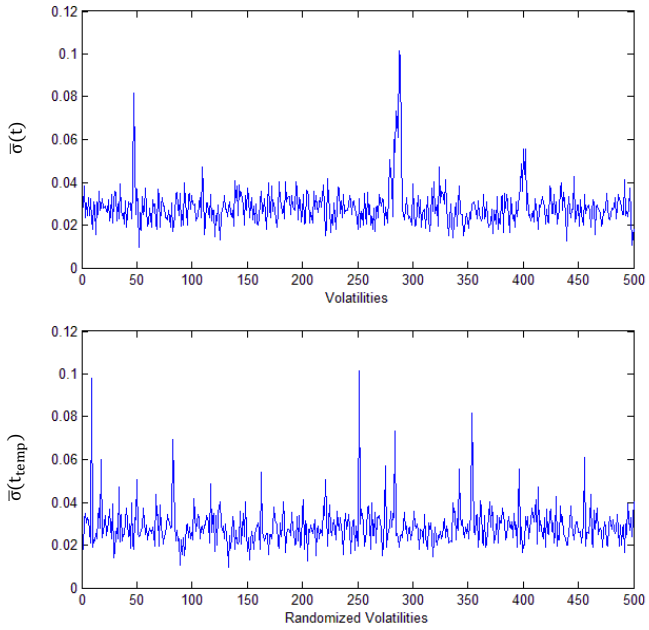


Figure 3.2: Volatility clustering of realized mid-price returns in simulation data

Chapter 4

A Perfect Sampling Algorithm for Stochastic Fluid Networks

4.1 Introduction

Stochastic fluid network is a class of queueing models that has been used to model telecommunication networks and data processing procedures. Despite its relative tractability due to the deterministic internal routing, there is no closed-form expression for the stationary distribution of a stochastic fluid network with general internal routing, even when the input process is a simple compound Poisson process. To characterize the stationary distribution of stochastic networks remains an open problem in queueing theory. Some progress has been made in recent works. For instance, in [20] computes the Laplace transformation of the stationary distribution for acyclic stochastic fluid networks with Lévy inputs and [11] does that for 2-dimensional networks with cycles.

Instead of studying the stationary distribution analytically, here we would like to pursue this problem using the computational approach. In particular, we develop a *perfect sampling* simulation algorithm that is able to generates i.i.d. samples exactly from the stationary distribution of stochastic fluid networks. The algorithm is based on the co-called Dominated Coupling From The Past (DCFTP). This technique was proposed by [44], following the introduction of the Coupling From The Past by [54]. The idea behind DCFTP is to construct a suitable pair of upper and lower bound processes that can be simulated in stationarity and backwards in time. We take the lower bound to be the process identically equal to zero. We use results from [43], see also [56], to construct an upper bound process corresponding to a network with trivial internal routing. It turns out that simulation of the stationary upper-bound process backwards involves sampling the infinite horizon maximum (component-by-component) from t to infinity of a d-dimensional compound Process with negative drift. We use importance sampling and sequential acceptance / rejection techniques to simulate from such infinite horizon maximum process.

Although we shall mainly describe and explain our perfect sampling algorithm for the case where the input process is a compound Poisson process(however, dependence between the inputs to different stations are allowed), we believe the strategy used in our algorithm can be adapted for more general cases. As an example, in Section 4.4 we extend our exact sampling algorithm to the case in which the network is exposed to a random environment. In detail, there is an independent Markov chain driving the arrival rates, the service rates, and the distribution of job sizes at the time of arrivals.

In addition, we analyze the computational complexity of our perfect sampling algorith-

m, measured in terms of expected random numbers generated, has polynomial increase in the number of dimensions, or equivalently, in the number of queues in the network. We also find our algorithm is efficient from numerical experiments.

The rest of this Chapter is organized as follows. We first describe in detail the stochastic fluid network model and the closely related Skorokhod problem in Section 4.2. In Section 4.3, we describe and explain the strategies and simulation techniques in our perfect sampling algorithm. The complexity analysis is given in Section 4.3.4. In Section 4.4 we extend our algorithm to Markov modulated networks and In Section 4.5 we report the results of the numerical experiments.

Notations: Throughout the paper we shall use boldface to write vector quantities, which are encoded as columns. For instance, we write $\mathbf{y} = (y_1, ..., y_d)^T$. All the inequalities hold component by component for vectors. For instance, by $\mathbf{x} < \mathbf{y}$ we mean that $x_i < y_i$ for all $i = 1, ..., d$. RCLL is the abbreviation for right continuous with left limit existing.

4.2 Stochastic Fluid Network Model and Skorokhod Mapping

Roughly speaking, a stochastic fluid network is a network where the external inputs are random but all internal flows are deterministic (see for instance, [43]). In this work, we will mainly consider stochastic fluid networks in which jobs arrive according to some Poisson process with general job-size distributions. In detail, we consider a network with d queueing stations indexed by $\{1, 2, ..., d\}$ and jobs arrive to the network according to a Poisson process with rate

λ, denoted by $(N(t) : t \geq 0)$. Specifically, the k-th arrival brings a vector of job requirements $\mathbf{W}(k) = (W_1(k), ..., W_d(k))^T$ which are non-negative random variables (r.v.'s) and that add to the workload at each station right at the moment of arrival. So, if the k-th arrival occurs at time t, the workload of the i-th station (for $i \in \{1, ..., d\}$) increases by $W_i(k)$ units right at time t. We assume that $\mathbf{W} = (\mathbf{W}(k) : k \geq 1)$ is a sequence of i.i.d. (independent and identically distributed) non-negative r.v.'s. For fixed k, the components of $\mathbf{W}(k)$ are not necessarily independent, however, $\mathbf{W}$ is assumed to be independent of $N(\cdot)$.

The workload is processed as a fluid by the server at a constant rate r_i and continuously in time at the i-th station. This means that if the workload in the i-th station remains strictly positive during the time interval $[t, t+dt]$ then the output from station i during this time interval equals $r_i dt$. In addition, a deterministic proportion $Q_{i,j} \geq 0$ of the fluid processed by the i-th station is circulated to the j-th server. We have that $\sum_{j=1}^{d} Q_{i,j} \leq 1$, $Q_{i,i} = 0$, and we define $Q_{i,0} = 1 - \sum_{j=1}^{d} Q_{i,j}$. The proportion $Q_{i,0}$ corresponds to the fluid that exists the network from station i.

We are interested in the *workload process* (or the virtual waiting-time process) $\mathbf{Y}(\cdot)$. In particular, $Y_i(t)$ is the amount of workload content at the i-th station at time t. Let $\mathbf{J}(t) = \sum_{k=1}^{N(t)} \mathbf{W}(k)$ be the total amount of jobs that arrive before time t. Let $R = I - Q$ and $\mathbf{r} = (r_1, ..., r_d)^T$ be the column vector corresponding to the service rates. The process $\mathbf{X} = \mathbf{J}(t) - R\mathbf{r}t$ is called the *net-input process*. Then, it is well-known that the workload process $\mathbf{Y}(\cdot)$ can be defined formally as the solution of a multi-dimensional Skorokhod problem corresponding to $\mathbf{X}(\cdot)$ and R. In detail, for any given initial content $\mathbf{Y}(0)$ and the input process $\mathbf{X}(\cdot)$ that is RCLL, there exists a unique pair of RCLL function $(\mathbf{Y}(\cdot), \mathbf{L}(\cdot))$ satisfying the following *Skorokhod Problem* Constraints:

1. for all for all $0 \leq t \leq T$,

$$\mathbf{Y}(t) \geq 0 = \mathbf{Y}(0) + \mathbf{X}(t) + R\mathbf{L}(t) \geq 0 \tag{4.2.1}$$

2. $L_i(\cdot)$ non-decreasing for each $i \in \{1, ...d\}$ and $L_i(0) = 0$,

3. $\int_0^T Y_i(s)\, dL_i(s) = 0$.

The mapping taking $(\mathbf{X}(\cdot), \mathbf{Y}(0))$ into $(\mathbf{Y}(\cdot), \mathbf{L}(\cdot))$ is called the *Skorokhod mapping*, which is well-defined and Lipschitz continuous (with Lipschitz constant ≤ 1) in the function space $D([0,T], \mathbb{R}^d)$ of all RCLL functions on $[0,T]$ equipped with the usual topology. Moreover, the workload process $\mathbf{Y}$ is a Markov process by itself.

In the our setting, the matrix R is an M-matrix. It is well known (see for instance, [43]) that, $R^{-1}\mathbf{r} < 0$ (component-by component) is a necessary and sufficient condition for the existence of a unique stationary distribution of the workload process.

4.3 Algorithm for Networks with Compound Poisson Input

First we summarize the assumptions that we shall impose on the stochastic fluid network in this section.

Assumptions:

A1) The matrix R is invertible and R^{-1} has non-negative components,

A2) $R^{-1}E\mathbf{X}(1) < 0$ (recall that inequalities apply component-by-component for vectors),

A3) There exists $\theta > 0$, $\theta \in \mathbb{R}^d$ such that

$$\log E[\exp\left(\theta^T \mathbf{W}(k)\right)] < \infty.$$

We have commented on A1) and A2) in the Introduction. Assumption A3) is important in order to simulate from the steady-state upper bound process that we shall construct to apply DCFTP.

In addition to A1) to A3) we shall assume that one can simulate from exponential tilting distributions associated to the marginal distribution of $\mathbf{W}(k)$. That is, we can simulate from $P_{\theta_i}(\cdot)$ such that

$$\begin{aligned} & P_{\theta_i}(W_1(k) \in dy_1, ..., W_d(k) \in dy_d) \\ = & \frac{\exp(\theta_i y_i)}{E\exp(\theta_i W_i(k))} P(W_1(k) \in dy_1, ..., W_d(k) \in dy_d), \end{aligned}$$

where $\theta_i \in \mathbb{R}$ and $E\exp(\theta_i W_i(k)) < \infty$. We will determine the value of θ_i through Assumption A3b) given below.

Let us briefly explain our program, which is based on DCFTP. First, we will construct a *stationary* dominating process $(\mathbf{Y}^+(s) : -\infty < s \leq 0)$ that is *coupled* with our target process, that is, a stationary version of the process process $(\mathbf{Y}(s) : -\infty < s \leq 0)$ satisfying the Skorokhod problem constraints (4.2.1). Under coupling, the dominating process satisfies

$$R^{-1}\mathbf{Y}(s) \leq R^{-1}\mathbf{Y}^+(s), \tag{4.3.1}$$

for each $s \leq 0$. We then simulate the process $\mathbf{Y}^+(\cdot)$ backwards up to a time $-\tau \leq 0$ such that $\mathbf{Y}^+(-\tau) = 0$. Following the tradition of the CFTP literature, we call a time $-\tau$ such that $\mathbf{Y}^+(-\tau) = 0$ a coalescence time. Since $\mathbf{Y}(s) \geq 0$, inequality (4.3.1) yields that $\mathbf{Y}(-\tau) = 0$. The next and final step in our strategy is to evolve the solution $\mathbf{Y}(s)$ of the Skorokhod problem (4.2.1) forwards from $s = -\tau$ to $s = 0$ with $\mathbf{Y}(-\tau)$=0, *using the same input that drives the construction of* $(\mathbf{Y}^+(s) : -\tau \leq s \leq 0)$ so that $\mathbf{Y}$ and $\mathbf{Y}^+$ are coupled. The output is therefore $\mathbf{Y}(0)$ which is, of course, stationary. The precise algorithm will be summarized in Section 4.3.3.

So, a crucial part of the whole plan is the construction of $\mathbf{Y}^+(\cdot)$ together with a coupling that guarantees inequality (4.3.1). In addition, the coupling must be such that one can use the driving randomness that defines $\mathbf{Y}^+(\cdot)$ directly as an input to the Skorokhod problem (4.2.1) that is then used to evolve $\mathbf{Y}^+(\cdot)$. We shall first start by constructing a time reversed stationary version of a suitable dominating process $\mathbf{Y}^+$.

4.3.1 Mathematical Construction of the Stationary Dominating Process

In order to construct the dominating process $\mathbf{Y}^+(\cdot)$ we first need the following result attributed to [43].

Lemma 4.3.1 (Kella and Whitt '96)**.** *There exists* $\mathbf{z}$ *such that* $E\mathbf{X}(1) < \mathbf{z}$ *and* $R^{-1}\mathbf{z} < 0$*. Moreover, if*

$$\mathbf{Z}(t) = \mathbf{X}(t) - \mathbf{z}t,$$

and $\mathbf{Y}^{+}(\cdot)$ *is the solution to the following Skorokhod problem*

$$d\mathbf{Y}^{+}(t) = d\mathbf{Z}(t) + d\mathbf{L}^{+}(t),\ \mathbf{Y}^{+}(0) = \mathbf{y}_0, \tag{4.3.2}$$

$$\mathbf{Y}^{+}(t) \geq 0,\ Y_j^{+}(t)\, dL_j^{+}(t) = 0,\ L_j^{+}(0) = 0,\ dL_j^{+}(t) \geq 0,$$

then $0 \leq R^{-1}\mathbf{Y}(t) \leq R^{-1}\mathbf{Y}^{+}(t)$ *for all* $t \geq 0$ *where* $\mathbf{Y}(\cdot)$ *solves the Skorokhod problem*

$$d\mathbf{Y}(t) = d\mathbf{X}(t) + Rd\mathbf{L}(t),\ \mathbf{Y}(0) = \mathbf{y}_0$$

$$\mathbf{Y}(t) \geq 0,\ Y_j(t)\, dL_j(t) = 0,\ L_j(0) = 0,\ dL_j(t) \geq 0.$$

We note that computing $\mathbf{z}$ from the previous lemma is not difficult, one can simply pick $\mathbf{z} = E\mathbf{X}(1) + \delta\mathbf{1}$, where $1 = (1, ..., 1)^T$ and with δ chosen so that $0 < \delta R^{-1}\mathbf{1} < -R^{-1}E\mathbf{X}(1)$. In what follows we shall assume that $\mathbf{z}$ has been selected in this form and we shall assume without loss of generality that $E[\mathbf{Z}(1)] < 0$.

The Skorokhod problem corresponding to the dominating process can be solved explicitly. Indeed, it is not difficult to verify, see for instance [35], that if $\mathbf{Y}^{+}(0) = 0$, the solution of the Skorokhod problem (4.3.2) is given by

$$\mathbf{Y}^{+}(t) = \mathbf{Z}(t) - \min_{0 \leq u \leq t} \mathbf{Z}(u) = \max_{0 \leq u \leq t} (\mathbf{Z}(t) - \mathbf{Z}(u)), \tag{4.3.3}$$

where the running maximum is obtained component-by-component.

In order to construct a stationary version of $\mathbf{Y}^{+}(\cdot)$ backwards in time, we first extend $\mathbf{Z}(\cdot)$

to a two-sided compound Poisson process with $\mathbf{Z}(0) = 0$. We define a time-reversion of $\mathbf{Z}(\cdot)$ as $\mathbf{Z}^{\leftarrow}(t) = -\mathbf{Z}(-t)$. It is easy to check that $\mathbf{Z}^{\leftarrow}(\cdot)$ has stationary and independent increments that are identically distributed as those of $\mathbf{Z}(\cdot)$.

For any given $T \leq 0$, we define a process $\mathbf{Z}_T^{\leftarrow}$ via $\mathbf{Z}_T^{\leftarrow}(t) = \mathbf{Z}^{\leftarrow}(T+t)$ for $0 \leq t \leq |T|$. And for any given $\mathbf{y} \geq 0$ we define $\mathbf{Y}_T^{+}(t,\mathbf{y})$ for $0 \leq t \leq |T|$ to be the solution to the Skorokhod problem with input process $\mathbf{Z}_T^{\leftarrow}$, initial condition $\mathbf{Y}_T^{+}(0,\mathbf{y}) = \mathbf{y}$ and reflection matrix $R = I$. In detail, $\mathbf{Y}_T^{+}(\cdot,\mathbf{y})$ solves

$$d\mathbf{Y}_T^{+}(t,\mathbf{y}) = d\mathbf{Z}_T^{\leftarrow}(t) + d\mathbf{L}_T^{+}(t,\mathbf{y}),\ \mathbf{Y}_T^{+}(0,\mathbf{y}) = \mathbf{y}, \tag{4.3.4}$$

$$\mathbf{Y}_T^{+}(t,\mathbf{y}) \geq 0,\ \mathbf{Y}_{T,j}^{+}(t,\mathbf{y})\, d\mathbf{L}_{T,j}^{+}(t,\mathbf{y}) = 0,\ \mathbf{L}_{T,j}^{+}(0,\mathbf{y}) = 0,\ d\mathbf{L}_{T,j}^{+}(t,\mathbf{y}) \geq 0.$$

According to (4.3.3), if $\mathbf{y} = 0$,

$$\mathbf{Y}_T^{+}(t,\mathbf{0}) = \max_{0 \leq u \leq t} (\mathbf{Z}_T^{\leftarrow}(t) - \mathbf{Z}_T^{\leftarrow}(u)). \tag{4.3.5}$$

Since $E[\mathbf{Z}(1)] < 0$, The process $\mathbf{Y}^{+}$ satisfying the Skorokhod problem (4.3.2) with orthogonal reflection ($R = I$) possesses a unique stationary distribution. So, we can construct a stationary version of $(\mathbf{Y}^{+}(s) : -\infty < s \leq 0)$ as

$$\mathbf{Y}_*^{+}(s) = \lim_{T \to -\infty} \mathbf{Y}_T^{+}(-T-s,\mathbf{0}). \tag{4.3.6}$$

The next proposition provides an explicitly evaluation of the limit in (4.3.6).

Proposition 4.3.2. *Given any $t \geq 0$,*

$$\mathbf{Y}_*^+(-t) = -\mathbf{Z}(t) + \max_{t \leq u < \infty} \mathbf{Z}(u), \tag{4.3.7}$$

Proof. Expression (4.3.5) together with the definition of $\mathbf{Z}_T^{\leftarrow}(\cdot)$ yields, for $T, s \leq 0$:

$$\begin{aligned}\mathbf{Y}_T^+(-T+s, \mathbf{0}) &= \max_{0 \leq u \leq -T+s} (\mathbf{Z}^{\leftarrow}(s) - \mathbf{Z}^{\leftarrow}(T+u)) \\ &= \max_{T \leq u+T \leq s} (\mathbf{Z}^{\leftarrow}(s) - \mathbf{Z}^{\leftarrow}(T+u)).\end{aligned}$$

Now we let $r = u + T$, substitute $\mathbf{Z}^{\leftarrow}(s) = -\mathbf{Z}(-s)$ and obtain

$$\begin{aligned}\mathbf{Y}_T^+(-T+s, \mathbf{0}) &= \max_{T \leq r \leq s} (\mathbf{Z}^{\leftarrow}(s) - \mathbf{Z}^{\leftarrow}(r)) \\ &= \max_{T \leq r \leq s} (-\mathbf{Z}(-s) + \mathbf{Z}(-r)) \\ &= -\mathbf{Z}(-s) + \max_{T \leq r \leq s} \mathbf{Z}(-r).\end{aligned}$$

Letting $-s = t \geq 0$ and $-r = u \geq 0$ we obtain

$$\mathbf{Y}_T^+(-T-t, \mathbf{0}) = -\mathbf{Z}(t) + \max_{t \leq u \leq -T} \mathbf{Z}(u).$$

Now send $-T \to \infty$ and arrive at (4.3.7), thereby obtaining the result. □

4.3.2 The Framework of the Perfect Sampling Algorithm

We now are ready to explain our main algorithm to simulate unbiased samples from the steady-state distribution of $\mathbf{Y}$. For this purpose, let us first define

$$\mathbf{M}(t) = \max_{t \le u < \infty} \mathbf{Z}(u),$$

for $t \ge 0$ so that $\mathbf{Y}_*^+(t) = \mathbf{M}(t) - \mathbf{Z}(t)$. Since $E[\mathbf{Z}(1)] < 0$ it follows that $(\mathbf{M}(t) : t \ge 0)$ is a well defined stochastic process. Let us for the moment assume that we can simulate $\mathbf{M}(\cdot)$ jointly with $\mathbf{Z}(\cdot)$ until the coalescence time τ. We shall explain how to perform such simulation procedures in Section 4.3.3.

Algorithm 4.1 Framework of the Perfect Sampling Algorithm

1: Simulate $(\mathbf{M}(t), \mathbf{Z}(t))$ jointly until time $\tau \ge 0$ such that $\mathbf{Z}(\tau) = \mathbf{M}(\tau)$.
2: Set $\mathbf{X}_{-\tau}^{\leftarrow}(t) = \mathbf{Z}(\tau) - \mathbf{Z}(\tau - t) + \mathbf{z} \times t$ and compute $\mathbf{Y}_{-\tau}(t, \mathbf{0})$ for $0 \le t \le \tau$ that solves the Skorokhod problem with input process $\mathbf{X}_{-\tau}^{\leftarrow}(t)$ and initial value $\mathbf{Y}_{-\tau}(0, \mathbf{0}) = 0$. In detail, $\mathbf{Y}_{-\tau}(t, \mathbf{0})$ solves

$$d\mathbf{Y}_{-\tau}(t, \mathbf{0}) = d\mathbf{X}_{-\tau}^{\leftarrow}(t) + R d\mathbf{L}_{-\tau}(t, \mathbf{0})$$
$$\mathbf{Y}_{-\tau}(t, \mathbf{0}) \ge 0,\ Y_{-\tau,j}(t, \mathbf{0})\, dL_{-\tau,j}(t, \mathbf{0}) = 0,\ L_{-\tau,j}(0, \mathbf{0}) = 0,\ dL_{-\tau,j}(t, \mathbf{0}) \ge 0,$$

for τ units of time.
3: **return** $\mathbf{Y}_{-\tau}(\tau, \mathbf{0})$.

In Step 2, The constant $\mathbf{z}$ is chosen according to Lemma 1 such that $\mathbf{Z}(t) = \mathbf{X}(t) - \mathbf{z}t$. The time is $-\tau$ precisely the coalescence time as in a DCFTP algorithm. The following proposition summarizes the validity of this algorithm.

Proposition 4.3.3. *The previous algorithm terminates with probability one and its output is an unbiased sample from the distribution of* $\mathbf{Y}(\infty)$.

Proof. It is immediate that $\mathbf{Y}^+(\infty)$ has an atom at zero, this follows from the fact that between arrivals of jumps all of the components of $\mathbf{Y}^+(\cdot)$ decrease linearly and the inter-arrival times, being exponentially distributed, have unbounded support. This implies that $\tau < \infty$ with probability one. Actually, we will show later that $E[\exp(\delta\tau)] < \infty$ for some $\delta > 0$ in Theorem 1. Let $T < 0$ and note that thanks to Lemma 5.4.3, for $t \in (0, |T|]$

$$R^{-1}\mathbf{Y}_T(t,\mathbf{0}) \leq R^{-1}\mathbf{Y}_T^+(t,\mathbf{0}). \tag{4.3.8}$$

In addition, by monotonicity of the solution to the Skorokhod problem in terms of its initial condition, see [56], we also have (using the definition of $\mathbf{Y}_T^+(t,y)$ form (4.3.4) and $\mathbf{Y}_*^+(T)$ from (4.3.6)) that

$$\mathbf{Y}_T^+(t,\mathbf{0}) \leq \mathbf{Y}_T^+(t,\mathbf{Y}_*^+(T)) = \mathbf{Y}_*^+(T+t). \tag{4.3.9}$$

So $\mathbf{Y}_*^+(T+t) = 0$ implies $\mathbf{Y}_T^+(t,\mathbf{0}) = 0$. One step further, as R^{-1} has non-negative components, equation (4.3.8), together (4.3.9) imply that $\mathbf{Y}_T(t,\mathbf{0}) = 0$. Consequently, if $-T > \tau \geq 0$

$$\mathbf{Y}_T(|T| - \tau, \mathbf{0}) = \mathbf{0},$$

which in particular yields that $\mathbf{Y}_T(-T,\mathbf{0}) = \mathbf{Y}_{-\tau}(\tau,0)$. We then obtain that

$$\lim_{T \longrightarrow -\infty} \mathbf{Y}_T(-T,0) = \mathbf{Y}_{-\tau}(\tau,0),$$

thereby concluding that $\mathbf{Y}_\tau(-\tau,0)$ follows the distribution $\mathbf{Y}(\infty)$ as claimed. □

Step 2 in Algorithm 4.1 is straightforward to implement because the process $\mathbf{X}^{\leftarrow}_{-\tau}(\cdot)$ is piecewise linear and the solution to the Skorokhod problem, namely $\mathbf{Y}_{-\tau}(\cdot,\mathbf{0})$ is also piecewise linear. The gradients are simply obtained by solving a sequence of linear system of equations which are dictated from the constraints given in (4.2.1). Therefore, the most interesting part is the simulation of the stochastic object $(\mathbf{M}(t) : 0 \leq t \leq \tau)$ in Step 1 as we will discuss in Section 4.3.3

4.3.3 Simulation Algorithm of the Stationary Dominating Process

As customary, we use the notation $E_0(\cdot)$ or $P_0(\cdot)$ to indicate the conditioning $\mathbf{Z}(0) = 0$. We define $\phi_i(\theta) = E_0[\exp(\theta Z_i(1))]$ be the moment generating function of $Z_i(1)$ and let $\psi_i(\theta) = \log(\phi_i(\theta))$. In order to simplify the explanation of the simulation procedure to sample $(\mathbf{M}(t) : t \geq 0)$, we introduce the following assumption.

Assumption: A3b) Suppose that in every dimension i there exists $\theta_i^* \in (0,\infty)$ such that

$$\psi_i(\theta_i^*) = \log E_0 \exp(\theta_i^* Z_i(1)) = 0. \tag{4.3.10}$$

This assumption is a strengthening of Assumption A3) and it is known as Cramer's condition in the large deviations literature. As we shall explain at the end of Section 4.3.3, it is possible to dispense this assumption and only work under Assumption A3). For the moment, we continue under Assumption A3b).

We wish to simulate $(\mathbf{Z}(t) : 0 \leq t \leq \tau)$ where τ is a time such that

$$\mathbf{Z}(\tau) = \mathbf{M}(\tau) = \max_{s \geq \tau} \mathbf{Z}(s), \text{ and hence } \forall 0 \leq t \leq \tau,\ \mathbf{M}(t) = \max_{t \leq s \leq \tau} \mathbf{Z}(s)$$

Recall that $-\tau$ is precisely the coalescence time since $\mathbf{Y}_*^+(-\tau) = \mathbf{0}$. We also keep in mind that our formulation at the beginning of the Introduction implies that

$$\mathbf{Z}(t) = \mathbf{J}(t) - R\mathbf{r}t - \mathbf{z}t = \sum_{k=1}^{N(t)} \mathbf{W}(k) - R\mathbf{r}t - \mathbf{z}t,$$

where $\mathbf{z}$ is selected according to Lemma 1. Define

$$\mu = R\mathbf{r} + \mathbf{z}$$

and let $\mu_i > 0$ be the i-th component of μ. In addition, we assume that we can choose a constant $m > 0$ large enough such that

$$\sum_{i=1}^{d} \exp(-\theta_i^* m) < 1. \tag{4.3.11}$$

Define

$$T_m = \inf\{t \geq 0 : Z_i(t) \geq m, \text{ for some } i\}. \tag{4.3.12}$$

Now we are ready to propose the following Algorithm 4.2 to simulate the coalescence time τ and we shall explain the key Steps 7 and 9 immediately, which are basically to simulate a path of a random walk with negative drift conditional on reaching a positive level in finite time.

Algorithm 4.2 Simulating the Coalescence Time

Input:

constant m as defined in (4.3.11)

Output:

$(\mathbf{Z}(t) : 0 \leq t \leq \tau)$ and the coalescence time τ.

Set $\tau = 0$, $\mathbf{Z}(0) = \mathbf{0}$ and $I = 1$.
while $I == 1$ **do**
  Generate an inter-arrival time U distributed Exp(λ) and sample $\mathbf{W} = (W_1, ..., W_d)$ independent of U.
  Let $\mathbf{Z}(\tau + t) = \mathbf{Z}(\tau) - t\mu$ for $0 \leq t < U$ and $\mathbf{Z}(\tau + U) = \mathbf{Z}(\tau) + \mathbf{W} - U\mu$.
  $\tau \longleftarrow \tau + U$.
  if $W_i - U\mu_i < -m$ **then**
    Sample a Bernoulli I with parameter $p = P_0(T_m < \infty)$.
    if $I == 1$ **then**
      simulate a new *conditional path* $(\mathbf{C}(t) : 0 \leq t \leq T_m)$ following the conditional distribution of $(\mathbf{Z}(t) : 0 \leq t \leq T_m)$ given that $T_m < \infty$ and $\mathbf{Z}(0) = \mathbf{0}$. Let $\mathbf{Z}(\tau + t) = \mathbf{Z}(\tau) + \mathbf{C}(t)$ for $0 \leq t \leq T_m$ and reset $\tau \longleftarrow \tau + T_m$.
    else
      return τ and the feed-in path $(\mathbf{Z}(t) : 0 \leq t \leq \tau)$.
    end if
  end if
end while

Simulating a Path Conditional on Reaching a Positive Level in Finite Time

The procedure that we shall explain now is an extension of the one dimensional procedure given in [7]; see also the related one dimensional procedure by [24]. The strategy is to use acceptance / rejection. The proposal distribution is based on importance sampling by means of exponential tilting. In order to describe our strategy we need to introduce some notation.

We think of the probability measure $P_0(\cdot)$ as defined on the canonical space of right-continuous with left-limits $\mathbb{R}^d$-valued functions, namely, the ambient space of $(\mathbf{Z}(t) : t \geq 0)$ which we denote by $\Omega = D_{[0,\infty)}\left(\mathbb{R}^d\right)$. We endow the probability space with the Borel σ-field generated by the Skorokhod J_1 topology, see [6]. Our goal is to simulate from the conditional

law of $(\mathbf{Z}(t) : 0 \leq t \leq T_m)$ given that $T_m < \infty$ and $\mathbf{Z}(0) = \mathbf{0}$, which we shall denote by P_0^* in the rest of this part.

Now let us introduce our proposal distribution, $P_0'(\cdot)$, defined on the space $\Omega' = D_{[0,\infty)}\left(\mathbb{R}^d\right) \times \{1,2,...,d\}$. We endow the probability space with the product σ-field induced by the Borel σ-field generated by the Skorokhod J_1 topology and all the subsets of $\{1,2,...,d\}$. So, a typical element ω' sampled under $P_0'(\cdot)$ is of the form ω'=$((\mathbf{Z}(t) : t \geq 0)$,$Index)$, where $Index \in \{1,2,...,d\}$. The distribution of ω' induced by $P_0'(\cdot)$ is described as follows, first,

$$P_0'(Index = i) = w_i := \frac{\exp(-\theta_i^* m)}{\sum_{j=1}^d \exp\left(-\theta_j^* m\right)}. \tag{4.3.13}$$

Now, given $Index = i$, for every set $A \in \sigma(\mathbf{Z}(s) : 0 \leq s \leq t)$,

$$P_0'(A|Index = i) = E_0[\exp(\theta_i^* Z_i(t)) I_A].$$

So, in particular, the Radon-Nikodym derivative (i.e. the likelihood ratio) between the distribution of $\omega = (\mathbf{Z}(s) : 0 \leq s \leq t)$ under $P_0'(\cdot)$ and $P_0(\cdot)$ is given by

$$\frac{dP_0'}{dP_0}(\omega) = \sum_{i=1}^d w_i \exp(\theta_i^* Z_i(t)).$$

The distribution of $(\mathbf{Z}(s) : s \geq 0)$ *under* $P_0'(\cdot)$ *is precisely the proposal distribution that we shall use to apply acceptance / rejection.* It is straightforward to simulate under $P_0'(\cdot)$. First, sample *Index* according to the distribution (4.3.13). Then, conditional on $Index = i$, the process $\mathbf{Z}(\cdot)$ also follows a compound Poisson process. Indeed, given $Index = i$, under $P_0'(\cdot)$ it follows

that $\mathbf{J}(t)$ can be represented as

$$\mathbf{J}(t) = \sum_{k=1}^{N'(t)} \mathbf{W}'(k), \tag{4.3.14}$$

where $N'(\cdot)$ is a Poisson process with rate $\lambda E[\exp(\theta_i^* W_i)]$. In addition, the distribution of W' is obtained by exponential titling such that for all $A \in \sigma(\mathbf{W})$,

$$P'(\mathbf{W}' \in A) = E[\exp(\theta_i^* W_i) I_A]. \tag{4.3.15}$$

In sum, conditional on $Index = i$, we simply let

$$\mathbf{Z}(t) = \sum_{k=1}^{N'(t)} \mathbf{W}'(k) - \mu t. \tag{4.3.16}$$

Now, note that we can write

$$\begin{aligned} E_0'(Z_{Index}(t)) &= \sum_{i=1}^{d} E_0(Z_i(t) \exp(\theta_i^* Z_i(t))) P'(Index = i) \\ &= \sum_{i=1}^{d} \frac{d\phi_i(\theta_i^*)}{d\theta} w_i > 0, \end{aligned}$$

where the last inequality follows by convexity of $\psi_k(\cdot)$ and by definition of θ_k^*. So, we have that $Z_{Index}(t) \nearrow \infty$ as $t \nearrow \infty$ with probability one under $P_0'(\cdot)$ by the Law of Large Numbers. Consequently $T_m < \infty$ a.s. under $P_0'(\cdot)$.

Recall that $P_0^*(\cdot)$ is the conditional law of $(\mathbf{Z}(t) : 0 \le t \le T_m)$ given that $T_m < \infty$ and $\mathbf{Z}(0) = \mathbf{0}$. In order to assure that we can indeed apply acceptance / rejection theory to simulate from

$P_0^*(\cdot)$, we need to show that the likelihood ratio dP_0/dP_0' is bounded.

$$\begin{aligned}\frac{dP_0^*}{dP_0'}(\mathbf{Z}(t):0\le t\le T_m) &= \frac{1}{P_0(T_m<\infty)}\times\frac{dP_0}{dP_0'}(\mathbf{Z}(t):0\le t\le T_m)\\ &= \frac{1}{P_0(T_m<\infty)}\times\frac{1}{\sum_{i=1}^d w_i\exp(\theta_i^* Z_i(T_m))}.\end{aligned} \tag{4.3.17}$$

Upon T_m, there is an index L (L may be different from *Index*) such that $\exp(\theta_L^* Z_L(T_m)) \ge \exp(\theta_L^* m)$, therefore

$$\frac{1}{\sum_{i=1}^d w_i\exp(\theta_i^* Z_i(T_m))} \le \frac{1}{w_L\exp(\theta_L^* m)} = \sum_{i=1}^d \exp(-\theta_i^* m) < 1, \tag{4.3.18}$$

where the last inequality follows by (4.3.11). Consequently, plugging (4.3.18) into (4.3.17) we obtain that

$$\frac{dP_0^*}{dP_0'}(\mathbf{Z}(t):0\le t\le T_m) \le \frac{1}{P_0(T_m<\infty)}. \tag{4.3.19}$$

We now are ready to summarize our acceptance / rejection procedure and the proof of its validity.

Proposition 4.3.4. *The probability that $I=1$ at any given call of Step 4 in Algorithm 4.3 is $P_0(T_m<\infty)$. Moreover, the output of Algorithm 4.3 follows the distribution P_0^*.*

Proof. The result follows directly from the theory of acceptance / rejection (see [3] page 39-42). According to it, since the two probability measures P_0^* and P_0' satisfy

$$\frac{dP_0^*}{dP_0'} \le c = \frac{1}{P_0(T_m<\infty)},$$

Algorithm 4.3 Simulation of Paths Conditional on $T_m < \infty$

Input:

θ^* and m as defined in C.1 and 4.3.11.

Output:

$(\mathbf{Z}(t) : 0 \le t \le T_m) \sim P_0^*$.

1: $I = 0$.
2: **while** $I == 0$ **do**
3: Sample $(\mathbf{Z}(t) : 0 \le t \le T_m)$ according to $P_0'(\cdot)$ as indicated via equations (4.3.13), (4.3.14) and (4.3.16).
4: Simulate a Bernoulli I with probability

$$\frac{1}{\sum_{i=1}^{d} w_i \exp\left(\theta_i^* Z_i(T_m)\right)}.$$

(Note that the previous quantity indeed is less than unity due to (4.3.18).)
5: **end while**
6: **return** $(\mathbf{Z}(t) : 0 \le t \le T_m)$.

as indicated by (4.3.17) and (4.3.19), one can sample exactly from P_0^* by the so-called acceptance / rejection procedure:

1. Generate i.i.d. samples $\{\omega_i\}$ from P_0' and i.i.d. random numbers $U_i \sim U[0,1]$ independent of $\{\omega_i\}$.

2. Define $N = \inf\{n \ge 1 : U_n \le c^{-1} \frac{dP_0^*}{dP_0'}(\omega_i)\}$.

3. Output ω_N.

The output w_N follows exactly the law P_0^* and N is a geometric random variable with mean c, in other words, the probability of accepting a proposal is c. In our specific case, we have $c = 1/P_0(T_m < \infty)$, and according to (4.3.17) the likelihood ration divided by constant c is

$$c^{-1} \frac{dP_0^*}{dP_0'}(\omega) = \frac{1}{\sum_{i=1}^{d} w_i \exp\left(\theta_i^* Z_i(T_m)\right)}.$$

Therefore, Algorithm 4.3 has acceptance probability $P(I=1)=P_0(T_m<\infty)$ and indeed it generates path from P_0^* upon acceptance. □

As the previous result shows, the output of the previous procedure indeed follows the distribution of $(\mathbf{Z}(t):0\leq t\leq T_m)$ given that $T_m<\infty$ and $\mathbf{Z}(0)=\mathbf{0}$. Moreover, the Bernoulli random variable I has probability $P_0(T_m<\infty)$ of success. So, this procedure actually allows to execute both Steps 4 and 5 in Algorithm 4.3 simultaneously. In detail, one simulate a path following the law of P_0' until T_m, and then, if the proposed path is accepted, one can conclude that T_m is finite and the proposed path is exactly a sample path following the law of P_0^*, otherwise one can conclude that $T=\infty$.

Remark: As mentioned earlier, Assumption A3b) is a strengthening of Assumption A3). We can carry out our ideas under Assumption A3) as follows. First, instead of $(\mathbf{M}(t):t\geq 0)$, given a vector $\mathbf{a}=(a_1,a_2,...,a_d)^T$ with non-negative components that we will explain how to choose momentarily, consider the process $\mathbf{Z_a}(\cdot)$ and $\mathbf{M_a}(\cdot)$ defined by

$$\mathbf{Z_a}(t):=\mathbf{Z}(t)+\mathbf{a}t,\ \mathbf{M_a}(t)=\max_{s\geq t}(\mathbf{Z_a}(s)).$$

Note that we can simulate $(\mathbf{M}(t):t\geq 0)$ jointly with $(\mathbf{Z}(t):t\geq 0)$ if we are able to simulate $(\mathbf{M_a}(t):t\geq 0)$ jointly with $(\mathbf{Z_a}(t):t\geq 0)$. Now, note that $\psi_i(\cdot)$ is strictly convex and that $d\psi_i(0)/d\theta<0$ so there exists $a_i>0$ large enough to force the existence of $\theta_i^*>0$ such that $E\exp(\theta_i^*Z_i(1)+a_i\theta_i^*)=1$, but at the same time small enough to keep $E(Z_i(1)+a_i)<0$; again, this follows by strict convexity of $\psi_i(\cdot)$ at the origin. So, if Assumption A3b) does not hold, but Assumption A3) holds, one can then execute the Algorithm 4.3 based on the process $\mathbf{Z_a}(\cdot)$.

4.3.4 Complexity Analysis

In this section we provide a complexity analysis of our algorithm. We first make some direct observations assuming the dimension of the network remains fixed. In particular, we note that the expected number of random variables simulated has a finite moment generating function in a neighborhood of the origin.

Theorem 4.3.5. *Suppose that A1) to A3) are in force. Let* τ *be the coalescence time and N be the number of random variables generated to terminate the overall procedure to sample* $\mathbf{Y}(\infty)$. *Then, there exists* $\delta > 0$ *such that*

$$E \exp(\delta\tau + \delta N) < \infty.$$

Proof. This follows directly from classical results about random walks (see [32]). In particular it follows that $E_0'(\exp(\delta T_m)) < \infty$. The rest of the survey follows from elementary properties of compound geometric random variables arising from the acceptance / rejection procedure. □

We are more interested, however, in complexity properties as the network increases. We shall impose some regularity conditions that allow us to consider a sequence of systems indexed by the number of dimensions d. We shall grow the size of the network in a meaningful way, in particular, we need to make sure that the network remains stable as the dimension d increases. Additional regularity will also be imposed.

Assumptions:

There exists two constants $0 < \delta < 1 < H < \infty$ independent of d satisfying the following conditions.

C1) $R^{-1}E[\mathbf{X}(1)] < -2\delta R^{-1}\mathbf{1}$ in each network.

C2) Let θ_i^* for $i = 1, ..., d$ be the tilting parameters as defined in Assumption A3b), then

$$E\exp[(\delta + \theta_i^*)W_i] \leq H < \infty,$$

and

$$H > \delta + \theta_i^* \text{ for all } 1 \leq i \leq d.$$

C3) The arrival rate $\lambda \in (\delta, H)$.

Remark: Assumption C1) indeed implies that $\mu = R\mathbf{r} + \mathbf{z} > \delta\mathbf{1}$, where $\mathbf{z}$ is defined according to Lemma 1. In detail, we choose $\mathbf{z} = E[\mathbf{X}(1)] + \delta\mathbf{1}$ and therefore, $R\mathbf{r} + \mathbf{z} = E[\mathbf{J}(1)] + \delta\mathbf{1} > \delta\mathbf{1}$. Throughout the rest part of this section, we assume that

$$\mu > \delta\mathbf{1}.$$

Besides, Assumption C2) jointly with Assumption C3) implies that $E[J_i(1)] = \lambda E[W_i]$ is uniformly bounded. In fact, one can check that $E[W_i] \leq E[\exp((\theta_i^* + \delta)W_i)]/(e\delta) < H/(e\delta)$. Therefore, we can also assume without loss of generality that

$$\mu = E[\mathbf{J}(1)] + \delta\mathbf{1} < H\mathbf{1}.$$

With a similar argument, Assumptions C2) and C3) also allow us to conclude that

$$\max_{1\leq i\leq d} E_0[Z_i(1)^2] \leq H.$$

We shall measure the complexity of Algorithm 4.1 by the expected total number of random numbers generated in a single call of the algorithm, which is denoted by $\mathcal{N}$. Then under Assumptions C1) through C3), we can prove that Algorithm 4.1 has polynomial complexity in terms of the number of dimensions (the size of the network) and the proof is given in Appendix C.1. We shall write $\mathcal{N}(d)$ instead of $\mathcal{N}$ to emphasis the dependence of $\mathcal{N}$ on the number of dimensions d.

Theorem 4.3.6. *Under assumptions C1) to C3),*

$$E[\mathcal{N}(d)] = O(d^{\gamma}) \text{ as } d \to \infty,$$

for some γ depending on δ and H.

4.4 Extension to Markov Modulated Networks

We shall briefly explain how our development in Section 4.3, specifically **Algorithm 4.1**, can be implemented beyond input with stationary and independent increments. As an example, we shall concentrate on Markov modulated stochastic fluid networks. Our extension to Markov modulated networks are first explained in the one dimensional case and later we will indicate how to treat the multidimensional setting.

Let $(I'(t) : t \geq 0)$ be an irreducible continuous time Markov chain taking values on the set $\{1,...,n\}$. We assume that, conditional on $I'(\cdot)$, the number of arrivals, $N'(\cdot)$, follows a time in-homogeneous Poisson process with rate $\lambda_{I'(\cdot)}$. We further assume that $\int_0^t \lambda_{I'(s)} ds > 0$ with positive probability. The process $N'(\cdot)$ is said to be a doubly-stochastic Poisson process with intensity $\lambda_{I(\cdot)}$. Define A'_k to be the time of the k-th arrival, for $k \geq 1$; that is, $A'_k = \inf\{t \geq 0 : N'(t) = k\}$.

We assume that the k-th arrival brings a job requirement equal to $W'(k)$. We also assume that the $W'(k)$'s are conditionally independent given the process $I'(\cdot)$. Moreover, we assume that the moment generating function $\phi_i(\cdot)$ defined via

$$\phi_i(\theta) = E\left(\exp\left(\theta W'(k)\right) | I'\left(A'_k\right) = i\right),$$

is finite in a neighborhood of the origin. In simple words, the job requirement of the k-th arrival might depend upon the environment, $I'(\cdot)$, at the time of arrival. But conditional on the environment the job sizes are independent. Finally, we assume that the service rate at time t is equal to $\mu_{I'(t)} \geq 0$.

Let $X'(t) = \sum_{k=1}^{N'(t)} W_k - \int_0^t \mu_{I'(s)} ds$, then the workload process, $(\mathbf{Y}(t) : t \geq 0)$, can be expressed as

$$Y(t) = X'(t) - \inf_{0 \leq s \leq t} X'(s),$$

assuming that $\mathbf{Y}(0) = 0$. In order for the process $Y(\cdot)$ to be stable, in the sense of having a stationary distribution, we assume that $\lim_{t\to\infty} X'(t)/t < 0$. Following the same argument as

in Section 4.3, we can construct a stationary version of the process $Y(\cdot)$ by a time reversal argument.

Since $I'(\cdot)$ is irreducible one can define its associated *stationary* time-reversed Markov chain $I(\cdot)$ with transition rate matrix $\mathcal{A}$ (for the existence and detailed description of such reversed chain see Chapter 2.5 of [1]). Let us write $N(\cdot)$ to denote a doubly stochastic Poisson process with intensity $\lambda_{I(\cdot)}$ and let $A_k = \inf\{t \geq 0 : N(t) = k\}$. We consider a sequence $(W(k) : k \geq 1)$ of conditionally independent random variables representing the service requirements (backwards in time) such that $\phi_i(\theta) = E(\exp(\theta W(k)) | I(A_k) = i)$.

We then can define $Z(t) = \sum_{k=1}^{N(t)} W(k) - \int_0^t \mu_{I(s)} ds$. Following the same arguments as in Section 4.3, we can run a stationary version Y^* of Y backwards via the process

$$Y^*(-t) = \sup_{s \geq t} (Z(s) - Z(t)).$$

Therefore, $Y^*(-t)$ can be simulated exactly as long as a convenient change of measure can be constructed for the process $(I(\cdot), Z(\cdot))$, so that a suitable adaptation of Algorithm 4.3 can be applied. Once the adaptation of Algorithm 4.3 is in place, the adaptation of Algorithm 4.2 and Algorithm 4.1 is straightforward.

In order to define such change of measure, let us define the matrix $\mathcal{M}(\theta, t) \in \mathbb{R}^{n \times n}$, for $t \geq 0$, via

$$\mathcal{M}_{ij}(\theta, t) = E_i[\exp(\theta Z(t)); I(t) = j],$$

where the notation $E_i(\cdot)$ means that $I(0) = i$. Note that $\mathcal{M}(\cdot, t)$ is well defined in a neighborhood

of the origin. In what follows we assume that θ is such that $\mathcal{M}(\theta,t)$ is finite component by component.

It is known (see for instance Chapter 11.2 and Chapter 13.8 of [1] and the references therein) that $\mathcal{M}(\theta,t) = \exp(t\mathbf{G}(\theta))$ where the matrix $\mathbf{G}$ is defined by

$$G_{ij}(\theta) = \begin{cases} \mathcal{A}_{ij} & \text{if } i \neq j, \\ \mathcal{A}_{ii} - \mu_i\theta + \lambda_i\phi_i(\theta) & \text{if } i = j. \end{cases}$$

Besides, $\mathbf{G}(\theta)$ has a unique eigenvalue $\beta(\theta)$ corresponding to a strictly positive eigenvector $(u(i,\theta) : 1 \leq i \leq n)$. The eigenvalue $\beta(\theta)$ has the following properties which follow from Proposition 2.4 and Proposition 2.10 in Chapter 11.2 of [1]:

Lemma 4.4.1.

1. $\beta(\theta)$ *is convex in* θ *and* $\beta'(\theta)$ *is well defined.*

2. $\lim_{t\to\infty} Z(t)/t = \beta'(0) = \lim_{t\to\infty} X'(t)/t < 0$*;*

3. $(M(t,\theta) : t \geq 0)$ *defined via*

$$M(t,\theta) = \frac{u(I(t),\theta)}{u(I(0),\theta)} \exp(\theta Z(t) - t\beta(\theta))$$

is a martingale.

As explained in Chapter 13.8 of [1], the martingale $M(\cdot)$ induces a change of measure for

the process $(I(\cdot), Z(\cdot))$ as we shall explain. Let P be the probability law of $(I(\cdot), Z(\cdot))$ and define a new probability measure $\tilde{P}$ for $(I(s), Z(s) : s \leq t)$ as $d\tilde{P} = M(t,\theta)\, dP$.

We now describe the law of $(I(\cdot), Z(\cdot))$ under $\tilde{P}$. The process $I(\cdot)$ is a continuous time Markov chain with rate rate matrix $\widetilde{\mathcal{A}}_{ij} = \mathcal{A}_{ij} u(j,\theta)/u(i,\theta)$ for $i \neq j$ (and $\widetilde{\mathcal{A}}_{ii} = -\sum_{j \neq i} \widetilde{\mathcal{A}}_{ij}$). In addition,

$$Z(t) \stackrel{\mathrm{d}}{=} \sum_{k=1}^{\tilde{N}(t)} \tilde{W}(k) - \int_0^t \mu_{I(s)} ds,$$

where $\tilde{N}$ is a doubly stochastic Poisson process with rate at time t equal to $\phi_{I(t)}(\theta)\lambda(I(t))$ and the $\tilde{W}(k)$'s are conditionally independent given $I(\cdot)$ with moment generating function $\widetilde{\phi}_i(\cdot)$ defined via

$$\widetilde{\phi}_i(\eta;\theta) = \widetilde{E}(\exp(\eta \tilde{W}(k)) | A_k = i) = \phi_i(\eta+\theta)/\phi_i(\eta),$$

which is finite in a neighborhood of the origin. In addition, $Z(t)/t \to \beta'(\theta)$ under $\tilde{P}$.

Because of the stability condition of the system we have that $\beta'(0) < 0$. Then, following the same argument as in the remark given at the end of Section 4.3.3, we may assume the existence of the Cramer root $\theta^* > 0$ such that $\beta(\theta^*) = 0$ and $\beta'(\theta^*) > 0$. The change of measure that allows to adapt Algorithm 4.3 is given by selecting $\theta^* > 0$ as indicated. Now, select $m > 0$ such that

$$K := \exp(-\theta^* m) \max_{i,j} \frac{u(i,\theta^*)}{u(j,\theta^*)} \leq 1. \tag{4.4.1}$$

We will use the notation $P_{0,i}(\cdot)$ to denote the law $P(\cdot)$ conditional on $Z(0) = 0$ and $I(0) = i$. Let us write $P^*_{0,i}(\cdot)$ to denote the law of $(Z(t) : 0 \leq t \leq T_m)$ (under $P_{0,i}(\cdot)$) conditional on $T_m < \infty$. Further, we write $\tilde{P}_{0,i}(\cdot)$ to denote the law of $\tilde{P}(\cdot)$, selecting $\theta = \theta^*$, conditional on $Z(0) = 0$

and $I(0) = i$. Then we have that $\tilde{P}_{0,i}(T_m < \infty) = 1$ (by Lemma 4.4.1 since $\beta'(\theta^*) > 0$) and therefore (by (4.4.1)) we have

$$\frac{dP^*_{0,i}}{d\tilde{P}_{0,i}}((I(t), Z(t)) : 0 \le t \le T_m) = \frac{u(i,\theta^*)}{u(I(T_m),\theta^*)} \times \frac{\exp(-\theta^* Z(T_m)) I(T_m < \infty)}{P_{0,i}(T_m < \infty)}$$
$$\le \frac{K}{P_{0,i}(T_m < \infty)} \le \frac{1}{P_{0,i}(T_m < \infty)}.$$

It is clear from this identity, which is completely analogous to identities (4.3.17) and (4.3.19) which are the basis for Algorithm 4.3, that the corresponding adaptation to our current setting follows easily.

For the $d-$dimensional case ($d > 1$), we first assume the existence of the Cramer root $\theta^*_j > 0$ for each dimension $j \in \{1, ..., d\}$. In this setting we also must compute the corresponding positive eigenvector $(u_j(i, \theta^*_j) : 1 \le i \le n)$ for each $j \in \{1, ..., d\}$. The desired change of measure that allows the adaptation of Algorithm 4.3 is just a mixture of changes of measures such as those described above induced by $M(\cdot, \theta^*_j)$ in each direction, just as discussed in Section 4.3.3, with weight $w_j = \exp(-\theta^*_j m)/\sum_{k=1}^m \exp(-\theta^*_k m)$. The corresponding likelihood ratio is then

$$\frac{dP^*_{0,i}}{d\tilde{P}_{0,i}}((I(t), Z(t)) : 0 \le t \le T_m)$$
$$= \frac{1}{\sum_{j=1}^d w_j \exp\left(\theta^*_j Z_j(T_m)\right) u_j\left(I(T_m), \theta^*_j\right) / u_j\left(i, \theta^*_j\right)}$$

and m must be selected so that

$$\sum_{j=1}^{d} \exp\left(-\theta_j^* m\right) \sup_{j,i,k} \frac{u_j\left(i,\theta_j^*\right)}{u_j\left(k,\theta_j^*\right)} \leq 1.$$

4.5 Numerical Experiment

We implemented Algorithm 4.1 to generate exact samples from the steady-state distribution of stochastic fluid networks. Our implementations were performed in Matlab. In all the experiments we simulated 10,000 independent replications and we display our estimates with a margin of error obtained using a 95% confidence interval based on the Central Limit Theorem.

We considered a 10-station system in tandem. In other words, $Q_{i,i+1} = 1$ for $i = 1,2,...,9$ and $Q_{10,j} = 0$ for all $j = 1,...,10$. We assume the arrival rate $\lambda = 1$ and the job sizes are exponentially distributed with unit mean. The service rates $(\mu_1,...,\mu_{10})^T$ are given by $(1.55, 1.5, 1.45, 1.4, 1.35, 1.3, 1.25, 1.2, 1.$ We are interested in computing the steady-state mean and the second moment of the workload at each station (i.e. $E[Y_i(\infty)]$ and $E[Y_i(\infty)^2]$ for $i = 1,2,...,10$). For a network of this type, it turns out, the true values of the quantities we are interested in can be computed from the corresponding Laplace transforms as given in [20].

Both the simulation results and the true values are reported in Table 1. The procedure took a few minutes (less than 5) in a desktop, which is quite a reasonable time.

Table 4.1: Unbiased estimate of $E[Y_i(\infty)]$ and $E[Y_i^2(\infty)]$ for a network with ten stations in tandem.

	$E[Y_i(\infty)]$		$E[Y_i^2(\infty)]$	
Station	Simulation Result	True Value	Simulation Result	True Value
1	1.7919±0.0521	1.8182	10.2755±0.5289	10.2479
2	0.1761±0.0068	0.1818	0.1511±0.0170	0.1642
3	0.2171±0.0083	0.2222	0.2242±0.0224	0.2382
4	0.2706±0.0102	0.2778	0.3462±0.0339	0.3610
5	0.3516±0.0131	0.3571	0.5717±0.0590	0.5778
6	0.4737±0.0171	0.4762	0.9840±0.0871	0.9921
7	0.6632±0.0233	0.6667	1.8472±0.1513	1.8715
8	1.0033±0.0345	1.0000	4.1004±0.3377	4.0300
9	1.6497±0.0542	1.6667	10.3734±0.7823	10.6065
10	3.3200±0.1040	3.3333	39.2015±2.9950	39.3631

Chapter 5

Tolerance-Enforced Simulation of Lévy Processes and Applications in Queueing Model Computation

5.1 Introduction

This paper develops a novel simulation methodology that allows to efficiently estimate sample path expectations of fully continuous processes that arise frequently in important queueing applications. The methodology actually has applications to virtually any area where continuous time stochastic models are involved, such as finance (see for instance [17]), but in this chapter we will focus primarily on the queueing applications. Our methodology, as we shall explain, combines elements that lie at the center of typical weak convergence analysis (such as the application of the continuous mapping principle) with modern computing methodology

based on Monte Carlo methods, such as multi-level Monte Carlo, via a new concept that we call *tolerance-enforced simulation*.

In order to motivate the implications of our contribution let us discuss a few examples of great importance in Operations Research that are covered by our theory. We start with Reflected Brownian Motion (RBM), which was introduced in [35] and was shown to approximate in distribution (under a suitable topology in the function space) the workload process of a d-dimensional single class queueing network in heavy traffic (i.e. as the proportion of time the system is utilized is close to 100%). The definition of RBM has been extended to accommodate approximations to multi-class networks (see [59] and [66]). By providing a rigorous way to approximate in great generality flexible and powerful queueing models, the introduction of RBM has dramatically increased the arsenal of available tools for performance analysis in diverse areas of applications such as: manufacturing systems, telecommunication systems, and service engineering(see [34]).

In order to apply RBM for performance analysis one must be able to compute associated transient and steady-state expectations. RBM has a corresponding Markov generator that allows to compute expectations involving marginal distributions by means of a parabolic partial differential equation (PDE) with constant coefficients but with 'oblique' Neumann-type boundary conditions (corresponding to non-orthogonal reflection typically arising in queueing settings). While there is a great deal of research on numerical methods for PDEs, most of these methods are developed for orthogonal (or standard) Neumann boundary conditions, not for oblique reflections. The work of [18] provides a class of numerical solvers specially designed for the steady-state distribution of RBM via the so-called Basic Adjoint Relationship (an equation that

is solved in a weak sense by the stationary density). Dai and Harrison's algorithms rest on the validity of a conjecture that has not been proved, although it is strongly believed to hold. However, all these numerical PDE solvers suffer significantly from the curse of dimensionality. This feature is particularly relevant in modern queueing networks which involve a large number of stations. In addition, it is also natural to quantify expectations of path-dependent functions (for which there might not be a characterizing PDE) and not only marginal distributions.

Throughout this paper, we will use Landou's notation for asymptotic behaviors. So, if $f(\cdot)$ and $g(\cdot)$ are positive functions, $f(x) = O(g(x))$ if $f(x) \leq cg(x)$ for some constant $c \in (0,\infty)$, similarly, $f(x) = \Omega(g(x))$ if $f(x) \geq cg(x)$.

Monte Carlo simulation provides a natural approach for computing expectations of stochastic processes. One of the main reasons, often advocated in favor of the Monte Carlo approach, is that it is relatively unsensitive to the underlying dimension. Indeed, in the presence of an unbiased estimator with finite variance, $O(1/\varepsilon^2)$ replications of the estimator suffices to guarantee an error of size ε to estimate an expectation. So the dimension does not play a direct role in the quadratic rate of the cost's increase as $\varepsilon \searrow 0$. Nevertheless, the challenge in the setting of continuous processes that are of interest in queueing, such as RBM, is that the underlying distributions (either the transition density or the steady-state distributions) are typically unknown. Moreover, RBM is described in terms of a stochastic differential equation (SDE) which involves terms of finite bounded variation but singular with respect to the Lebesgue measure (i.e. terms that behave like local times).

Of course, one cannot really simulate a full RBM path in a computer, so one has to resort to numerical solutions to SDEs that could be very costly. So, an important challenge is to find

a way to enable the power of Monte Carlo simulation to these types of continuous processes to guarantee an ε error with basically $O\left(1/\varepsilon^2\right)$ simple random variables, such as one dimensional Gaussians, simulated even in the case path-dependent expectations. In fact [3] note that:

"An open problem of considerable interest is to find good algorithms for simulating reflected Brownian motion in higher dimensional regions..."

One of our contributions consists in developing simulation methodology to estimate a large class of path-dependent transient expectations of RBM with a guaranteed $\varepsilon > 0$ error in order $O\left(1/\varepsilon^2 \log(1/\varepsilon)^3\right)$ function evaluations. A function evaluation in our setting typically correspond to the generation of a simple random variable, such as a one dimensional Gaussian, and a simple operation such as addition or multiple. Following the computer science tradition, we write $\widetilde{O}\left(1/\varepsilon^2\right)$ instead of $O\left(1/\varepsilon^2 \log(1/\varepsilon)^k\right)$ for some $k \geq 1$. Although we concentrate on the setting of [35], given that $O\left(1/\varepsilon^2\right)$ cost is the best possible practical performance in terms of simulation, we believe that our results provide a substantial step forward into addressing the problem highlighted by [3]. Moreover, our methodology allows to estimate a large class of steady-state expectations with $\tilde{O}\left(1/\varepsilon^2\right)$ complexity. Again, note that if one was able to directly sample from the steady-state density of RBM it would take $\Omega\left(1/\varepsilon^2\right)$ function evaluations to estimate a typical steady-state expectation to ensure ε error. So, the performance of our algorithm is very close to the best possible practical simulation benchmark.

Now we briefly discuss our strategy. There are three key ingredients that lie at the core of our methodology: First, the development of *torelance-enforced simulation* (TES) algorithms for the underlying free process $\{X(t) : 0 \leq t \leq 1\}$ such as a Brownian motion. A TES algorithm is a procedure such that for any given $\varepsilon > 0$ (*deterministic*) outputs a path $X_\varepsilon(\cdot)$, parameterized

by finitely many random variables, with the property $||X - X_\varepsilon|| < \varepsilon$, where $||\cdot||$ denotes the supremum norm on $[0,1]$; note that the error is deterministic *with probability one*. The second ingredient involves continuity properties of a map $\Phi(\cdot)$ that defines the process of interest. For instance, RBM can be seen as Lipschitz continuous mappings (with respect to the supremum norm) of Brownian motion. Finally, we use the first and second ingredients in combination with multi-level Monte Carlo (a technique introduced in [27]) to obtain the desired level of accuracy at an expense of $\widetilde{O}\left(1/\varepsilon^2\right)$ function evaluations.

Our methodology follows in the spirit of classical weak convergence analysis in which the continuous mapping principle transfers an approximation result of a simple process (such as a random walk) to a more complicated process (such as a queueing network). Our Monte Carlo approach is to transfer a TES algorithm for a free process (such as Brownian motion) to an SDE of interest via a well behaved mapping. We are able to isolate the main properties, in terms of computational cost, that are required for the TES algorithm with error bound ε for a free process to transfer into an algorithm that has $\widetilde{O}\left(1/\varepsilon^2\right)$ computational cost for a complex process. So, in the setting of SDEs that can be seen as suitable Lipschitz mappings of an underlying free process, everything boils down to developing suitable TES procedures.

We illustrate in this paper how to design TES procedures both for Brownian motion and Lévy processes with infinitely many jumps in any compact interval. We show that the complexity properties that we advocated earlier ($\widetilde{O}\left(1/\varepsilon^2\right)$ function evaluations) can be achieved thereby obtaining approximations not only for functionals of Brownian motion such as the RBM, but also for those driven by Léby processes. In particular, we focus on Lévy processes whose jump part has finite variation. For Lévy processes whose jump part has infinite variation, which is out

of the scope of this chapter, [4] provides a simulation algorithm approximating small jumps by some Brownian motion and gives a necessary and sufficient condition for convergence.

The rest of the paper is organized as follows: in Section 5.2, we introduce the framework of the TES algorithms and give the implementable algorithm for Brownian motion and for Lévy processes with infinite activity. In Section 5.3, we combine the TES algorithm with the Multilevel technique to obtain $\widetilde{O}\left(1/\varepsilon^2\right)$ complexity in terms of the error bound ε. In Section 5.4 we apply our results to multidimensional RBM, to estimate transient and steady-state expectations. In particular, in Section 5.4.3, we use the TES algorithm to extend the *perfect sampling* algorithm developed for stochastic fluid networks in Chapter 4 to the more complicated RBMs. Numerical experiments are also included.

5.2 Tolerance-Enforced Simulation Algorithms

In this section, we introduce our tolerance-enforced simulation algorithms which are constructed based on series representations of stochastic processes. The section is organized as follows: in Section 5.2.1 we discuss the algorithm for generic processes satisfying a suitable series representation, whereas in Sections 5.2.2 and 5.2.3 detailed algorithms for Brownian motions and some Lévy processes are presented.

5.2.1 A General Construction of a TES Procedure

Basically, the key feature of our TES algorithm is that for any given $\varepsilon > 0$ and finite time interval $[0, T]$, it can generate a path of the target stochastic process with error that is uniformly

bounded on $[0,T]$ with probability 1 by the given constant $\varepsilon > 0$. The precise definition of a TES algorithm is as follows.

Definition 5.2.1. *Given $\varepsilon > 0$, a **TES** algorithm for the stochastic process X is able to generates a stochastic process X^ε on $D([0,T],\mathbb{R}^d)$ satisfying that there exists a coupling (X, X^ε) such that $P(\|X - X^\varepsilon\| < \varepsilon) = 1$. We denote the complexity(or computational cost) of the ε-TES procedure by $C(\varepsilon)$.*

Our TES algorithm is designed for stochastic processes with series representations.

Definition 5.2.2. *A stochastic process $\{X(t) : 0 \le t \le T\}$ is said to have a series representation*

$$X(\cdot) = \sum_{n=1}^{\infty} Z_n f_n(\cdot), \tag{5.2.1}$$

if the series on the left hand side converges to $X(t)$ uniformly on $[0,T]$ with probability 1. Here $\{Z_n\}$ is a sequence of random variables and $\{f_n\}$ a sequence of functions.

The key to our TES algorithm is to find a sequence of deterministic numbers η_n such that:

$$P(|Z_n| > \eta_n \text{ infinitely often }) = 0, \tag{5.2.2}$$

and

$$\sum_{n=1}^{\infty} \eta_n f_n(t) \text{ converges uniformly.} \tag{5.2.3}$$

Given condition (5.2.2), there exists some random number $M > 0$ such that $|Z_n| < \eta_n$ for all $n > M$. On the other hand, owning to condition (5.2.3), for any $\varepsilon > 0$, there exists $m > 0$ such

that $\sum_{n=m+1}^{\infty} \eta_n f_n(t) < \varepsilon$. As a result, $|\sum_{n=M\vee m+1}^{\infty} Z_n f_n(t)| < \varepsilon$ for all $t \in [0,T]$, and as a direct consequence,

$$X^{\varepsilon}(t) := \sum_{n=1}^{M\vee m} Z_n f_n(t)$$

is a desirable output of an ε-TES algorithm for X.

For chosen $\{\eta_n\}$ and $\varepsilon > 0$, the constant m is easy to compute. The difficult part is to simulate M jointly with $\{Z_n : 1 \le n \le M\}$. To make it easier to understand, we describe how to simulate M jointly with $\{Z_n : 1 \le n \le M\}$ for two specific classes of processes, Brownian motions and Lévy processes, in Section 5.2.2 and Section 5.2.3, respectively. Although no general algorithm is given here, one can learn from these specific examples the basic ideas and apply them to other processes.

5.2.2 TES for Brownian Motion

We just need to describe the TSE for a standard one dimensional Brownian motion because any multi-dimensional Brownian motion can be expressed as the linear combination of independent one-dimensional standard Brownian motions. Due to the time-space scaling, to simulate a Brownian motion on $[0,T]$ is equivalent to simulate it on $[0,1]$ for any $T < \infty$. A recent paper [5] proposed a different method to simulate Brownian motion with a given deterministic error. In contrast to our approach, however, their strategy cannot be generalized beyond the Brownian case.

Throughout our discussion below we will use Z to denote a standard normal random variable.

A standard Brownian motion on $[0,1]$ has a special series representation called the *wavelet representation*, we now introduce the definition of this wavelet representation of Brownian motion as described in [65]. The wavelet generator that we shall use, $H(\cdot)$, is defined on $[0,1]$ as

$$H(t) = \begin{cases} 1 & \text{for } 0 \le t < \frac{1}{2} \\ -1 & \text{for } \frac{1}{2} \le t \le 1 \cdot \\ 0 & \text{otherwise.} \end{cases}$$

We set $H_0(t) = 1$ and generate a family of wavelet functions via $H_n(t) = 2^{-j/2}H(2^j t - l)$ for $n = 2^j + l$ where $j > 0$ and $0 \le l \le 2^j$. A Brownian path on $[0,1]$ then can be decomposed as a series in terms of $H_n(\cdot)$:

Proposition 5.2.3. *If $\{Z_n : 0 \le n < \infty\}$ is a sequence of independent standard normal random variables, then the series defined by*

$$B_t = \sum_{n=0}^{\infty} \left(Z_n \int_0^t H_n(s)\, ds \right)$$

converges uniformly on $[0,1]$ with probability 1. Moreover, the process $\{B_t\}$ defined by the limit is a standard Brownian motion on $[0,1]$.

Let $f_n(t) = \int_0^t H_n(s)ds$, then $B_t = \sum_0^\infty Z_n f_n(t)$, which is consistent with the series representation (5.2.1). We choose $\eta_n = \rho\sqrt{\log n}$ for $\rho > 2$. The following two properties correspond to Condition (5.2.2) and (5.2.3) in Section 5.2.2.

Proposition 5.2.4. *For any $\rho > 2$, $P(|Z_n| > \rho\sqrt{\log n}$ i.o.$) = 0$.*

Proof. It follows immediately from the Borel-Cantelli Lemma and that

$$\sum_{n=1}^{\infty} P(|Z_n| > \rho\sqrt{\log n}) \leq \sum_{n=1}^{\infty} \frac{1}{\rho\sqrt{\log n}} n^{-\rho^2/2} < \infty.$$

□

Proposition 5.2.5.

$$\sum_{n=M}^{\infty} H_n(t)|Z_n| \leq \sum_{l=L}^{\infty} \rho 2^{-l/2}\sqrt{l+1} \leq \rho Er(L),$$

where $Er(L)$ has the following expression

$$Er(L) = 2^{-(L-2)/2} \cdot \left(\frac{\sqrt{L+1}}{\log 2} + \frac{1}{(\log 2)^2\sqrt{L+1}} \right)$$

and $Er(L) \to 0$ as $L \to \infty$.

Proof. As shown in [65], for all $0 \leq t \leq 1$, given $|Z_n| \leq \rho\sqrt{\log n}$ for all $n \geq M \geq 2^L$, then

$$\sum_{n=M}^{\infty} |Z_n \int_0^t H_n(s)ds| \leq \rho \sum_{j=L}^{\infty} 2^{-j/2}\sqrt{j+1} \leq \rho \int_L^{\infty} \sqrt{u+1}\exp(-u\log 2/2)du,$$

We take $y = \sqrt{u+1}$ and obtain

$$\int_L^{\infty} \sqrt{u+1}\exp(-u\log 2/2)du = 2\sqrt{2}\int_{\sqrt{L+1}}^{\infty} \exp(-y^2\log 2/2)y^2 dy$$

By integration by parts, we have

$$\int_{\sqrt{L+1}}^{\infty} \exp(-y^2 \log 2/2) y^2 dy = -\frac{y}{\log 2} \exp(-y^2 \log 2/2)|_{\sqrt{L+1}}^{\infty} + \frac{1}{\log 2} \int_{\sqrt{l+1}}^{\infty} e^{-y^2 \log 2/2} dy$$
$$\leq \left(\frac{\sqrt{L+1}}{\log 2} + \frac{1}{\log 2^2 \sqrt{L+1}} \right) \cdot 2^{-(L+1)/2}$$

Therefore,

$$\sum_{n=M}^{\infty} |Z_n \int_0^t H_n(s) ds| \leq \rho 2\sqrt{2} \int_{\sqrt{L+1}}^{\infty} \exp(-y^2 \log 2/2) y^2 dy \leq \rho Er(L).$$

□

Since Condition 5.2.2 and 5.2.3 are satisfied, for any $\varepsilon > 0$, we can find $L > 0$ such that $Er(L) < \varepsilon$. To give an ε-TSE algorithm, it is now sufficient for us to show how to simulate M jointly with $\{Z_n\}$, which is given as Algorithm 5.1.

In Algorithm 5.1, we keep an array S, which is used to record the indices such that $|Z_n| > \rho\sqrt{\log n}$, and a number G which is the next index to be added onto S. Precisely speaking, given that the last element in array S is K, say, $\max(S) = K$, $G = \inf\{n \geq K+1 : |Z_n| > \rho\sqrt{\log n}\}$. The key part of the algorithm is to simulate a Bernoulli with success parameter $P(G < \infty)$ and sample G given $G < \infty$.

For this purpose, we keep updating two constants U and D such that $U > P(G = \infty) > D$ and $(U - D) \to 0$ as the number of iterations grows. To illustrate this point, denote the value of

Algorithm 5.1 Simulate M jointly with $\{Z_n\}$

Input:

$L := \inf\{l : Er(l) < \varepsilon\}$.

Output:

M jointly with $\{Z_n\}_{n=1}^{M}$.

Initialize $G = 2^L$ and S to be an empty array. Set $I = 1$.
while I==1 **do**
  Set $U = 1, D = 0$. Simulate $V \sim Uniform(0,1)$.
  while $U > V > D$ **do**
    set $G \leftarrow G+1$ and $U \leftarrow P(|Z_G| \le \rho\sqrt{\log G}) \times U$ and $D \leftarrow (1 - G^{1-\rho^2/2}) \times U$.
  end while
  if $V \ge U$ **then**
    add G to the end of S, i.e. $S = [S, G]$.
  else
    if $V \le D$ **then**
      $M = G$ and set $I = 0$.
    end if
  end if
end while
for n=1:M **do**
  if $n \in S$ **then**
    generate Z_n according to the conditional distribution of Z given $\{|Z| > \rho\sqrt{\log n}\}$;
  else
    generate Z_n according to the conditional distribution of Z given $\{|Z| \le \rho\sqrt{\log n}\}$.
  end if
end for

U and D in the m-th iteration by U_m and D_m respectively. Then for all $m > 0$,

$$P(G = \infty) = \prod_{k=K+1}^{\infty} P(|Z_k| \le \rho\sqrt{\log k}) < \prod_{k=K+1}^{K+m} P(|Z_k| \le \rho\sqrt{\log k}) = U_m.$$

On the other hand, for all $\rho > \sqrt{2}$ and K large enough,

$$\prod_{k=K+m+1}^{\infty} P(|Z_k| \le \rho\sqrt{\log k}) > 1 - \sum_{k=K+m+1}^{\infty} P(|Z_k| > \rho\sqrt{\log k}) \ge 1 - (K+m+1)^{1-\rho^2/2}.$$

and hence we conclude that $D_m < P(G = \infty)$. In sum, $U_m > P(G = \infty) > D_m$. Because $(1 -$

$(K+m+1)^{1-\rho^2/2}) \to 1$ as $m \to \infty$, the while loop 2 ends after a finite number of iterations and we can then decide whether $G < \infty$ or not.

Now we show that we can actually sample G simultaneously as the Bernoulli is generated. If $V < D$, we can conclude that $V < P(G = \infty)$ and hence $G = \infty$ and $M = \max(S)$. Otherwise, we have $G < \infty$. In this case, suppose the while loop 2 ends in the $(m+1)$-th iteration and $V > U$. Since $U_m = P(|Z_n| \leq \rho\sqrt{\log n}$ for $n = K+1, ..., K+m)$, $U_{m+1} \leq V < U_m$ implies nothing but that $K+m+1 = \inf\{n \geq K+1 : |Z_n| > \rho\sqrt{\log n}\}$. Therefore, by definition, $G = K+m+1$ and should be added into array S. Once S and M are generated, $\{Z_n\}$ can be generated jointly with S and M according to the for loop 15.

In the end, we prove the following result on the efficiency of the TES algorithm for Brownian motion.

Theorem 5.2.6. *The complexity of Algorithm 5.1 is* $O(\varepsilon^{-2}\log(1/\varepsilon))$

Proof. The complexity of Algorithm 5.1 is of the same order with $E[M \vee 2^L]$, so we just need to show that $E[M \vee 2^L] = O(\varepsilon^{-2}\log(1/\varepsilon))$. Note that

$$Er(L) = 2^{-(L-2)/2} \cdot \left(\frac{\sqrt{L+1}}{\log 2} + \frac{1}{(\log 2)^2\sqrt{L+1}} \right).$$

Therefore, we can choose $2^L = O(\varepsilon^{-2}\log(\frac{1}{\varepsilon}))$ such that $\rho Er(L) < \varepsilon$.

For fixed $\rho > 2$, $p_k = P(|Z_k| > \rho\sqrt{\log k}) \leq c \cdot \frac{1}{\rho\sqrt{\log k}}\, k^{-\rho^2/2}$. Recall that $M = \max\{k : |Z_k| > \rho\sqrt{\log k}\}$, then

$$E[M] = \sum_{n=1}^{\infty} P(M > n) \leq \sum_{n=1}^{\infty}\sum_{k=n}^{\infty} p_k = \sum_{k=1}^{\infty} k \cdot p_k \leq \sum_{k=1}^{\infty} c \cdot \frac{1}{\rho\sqrt{\log k}} k^{1-\rho^2/2},$$

which is finite and independent of ε. In summary, we have $E[M \vee 2^L] \leq E[2^L + M'] = O(\varepsilon^{-2}\log(1/\varepsilon))$.

□

5.2.3 TES for Lévy Processes

Series representations of Lévy processes are known under very general conditions, see [60]. There are existing simulation algorithms for Lévy processes based on their series representations, see for instance [8], [48] and [62]. (For a brief survey on simulation for Lévy processes, see Chapter 6 in [17] and Chapter 12 in [3].) Our TES algorithm for Lévy processes also uses series representation but is stronger than those existing algorithms in the sense that TES is able to generate sample pathes with deterministically bounded uniform error. Of course, to obtain this strong control on the uniform error, we need an extra condition on the Lévy process that its jump part has finite variation.

Due to the infinite divisibility of Lévy processes, to simulate a Lévy process on any finite time interval is equivalent to simulate it on $[0,1]$. Throughout this subsection, we consider a Lévy process $\{X_t\}$ for $t \in [0,1]$. The Lévy-Ito Decomposition Theorem shows that the process X can be decomposed into three parts $X^{(1)}$, $X^{(2)}$ and $X^{(3)}$ such that:

1. $X^{(1)}$ is a Brownian motion;

2. $X^{(2)}$ is a compound Poisson process;

3. $X^{(3)}$ is a square-integrable pure jump martingale.

Since $X^{(1)}$ is a Brownian motion and has already been discussed in Section 5.2.2, we only need to explain how to design TES algorithms for $X^{(2)}$ and $X^{(3)}$.

An important quantity related to Lévy process is the Lévy measure. Heuristically, if X has $\Pi(\cdot)$ as its Lévy measure, the jumps arrive to X according to a Poisson point process with density $\lambda(dx,dt) = \Pi(dx)dt$, where x is the jump 'size'. Different from the simple compound Poisson processes, $\Pi(\mathbb{R}^d)$ may equal to ∞ and hence the Lévy process X may have infinite number of jumps on any finite time interval.

In this part, we focus on those Lévy processes with finite first-order variations, in other words, the corresponding Lévy measure satisfies that

$$\int_{\mathbb{R}^d} (|x| \wedge 1)\, \Pi(dx) < \infty. \tag{5.2.4}$$

In this case, $X^{(2)} + X^{(3)}$ can be expressed as a 'mixture' of compound Poisson processes:

$$(X^{(2)} + X^{(3)})(t) = \int_0^t \int_{\mathbb{R}^d} xN(dx,ds),$$

where $N(\cdot,\cdot)$ is the Poisson point process with measure $\Pi(dx)dt$. It is natural to construct a TES algorithm based on this compound Poisson representation of Lévy processes. However, in most cases, for instance, the Gamma processes, the density of the Lévy measure $\Pi(dx) := \pi(x)dx$ is known but cannot be integrated in closed form. As a consequence, the probability p_k as in Algorithm 5.1, which in this case equals the tail of Poisson with mean $\Pi(A)$ for some set A, cannot be computed.

For this reason, we consider another series representation of the Lévy Process. Under some mild conditions, which can be deduced from the assumption (5.2.4), the following series repre-

sentation holds for a Lévy process (see Chapter 12.4 of [3]):

$$X(t) = \sum_{n=1}^{\infty} f(\xi_n, \Gamma_n) 1(U_n \leq t) \text{ for } 0 \leq t \leq 1,$$

where Γ_n is the nth epoch of a Poisson process with rate 1; $\{\xi_n\}$ are i.i.d. random seeds and $\{U_n\}$ are i.i.d. uniforms on $(0,1)$, both of which are independent of the Poisson process.

We are interested in the series representations satisfying the following condition:

$$|f(\cdot, y)| \leq c y^{-1/\alpha}, \tag{5.2.5}$$

where $c > 0$ and $0 < \alpha < 1$ are two constants.

Remark: The condition $\alpha < 1$ indicates that $X(t)$ has finite first-order variation.

Now we illustrate the series representation for Lévy processes with the following examples from and [63] and [61].

Examples:

1. **Stable process**: A stable process $S_\alpha(1, \beta, 0)$ is a pure jump process with Lévy measure:

$$\pi(x) = \begin{cases} \frac{\beta+1}{2} x^{-1-\alpha} dx & \text{for } x \in (0, \infty) \\ \frac{\beta-1}{2} |x|^{-1-\alpha} dx & \text{for } x \in (-\infty, 0) \end{cases}$$

for $0 < \alpha < 2$ and $\beta \in [-1, 1]$. Then,

$$S_\alpha(1, \beta, 0)(t) = C_\alpha^{1/\alpha} \sum_{n=1}^{\infty} \xi_n \Gamma_n^{-1/\alpha} 1(U_n \leq t),$$

where

$$C_\alpha = \left(\int_0^\infty x^{-\alpha} \sin(x) dx \right)^{-1}, \quad P(\xi_n = 1) = \frac{\beta + 1}{2} \text{ and } P(\xi_n = -1) = \frac{\beta - 1}{2}.$$

2. **CGMY process** (one-sided): The CGMY process, also known as the tempered stable process, has a Lévy measure in the form of $\pi(x) = C \exp(-Mx) x^{-1-\alpha}$ for $x > 0$ with given parameters $M, C > 0$ and $0 < \alpha < 1$. Then,

$$X(t) = \sum_{n=1}^{\infty} \left[\left(\frac{C}{\alpha \Gamma_n} \right)^{1/\alpha} \wedge \left(\frac{T_n V_n^{1/\alpha}}{M} \right) \right] 1(U_n \leq t),$$

where $\{T_n\}$ are i.i.d. exponentials with mean 1 and $\{V_n\}$ are i.i.d. uniforms on $(0, 1)$. $\{T_n\}$ and $\{V_n\}$ are independent of each other and of the Poisson process.

Without loss of generality, we assume $c = 1$ in (5.2.5). Since $\alpha < 1$, there exists some integer $l > 0$ and $0 < \beta < 1$ such that

$$\beta > \alpha, \quad l(1 - \beta) > 2.$$

We intend to choose $\eta_n = n^{-\beta/\alpha}$ and the following proposition corresponds to Condition (5.2.2) for the TES algorithm.

Proposition 5.2.7. $P(\Gamma_{(k+1)l} - \Gamma_{kl} < k^{\beta-1},\ i.o.) = 0$

Proof. Let V_1,...,V_l be i.i.d. exponentials with mean 1, then

$$P(\Gamma_{(k+1)l} - \Gamma_{kl} < k^{\beta-1}) = P(V_1 + ... + V_l < k^{\beta-1}) = O(k^{l(\beta-1)}).$$

Since $l(1-\beta) > 2$, we can conclude $\sum_{k=1}^{\infty} P(\Gamma_{(k+1)l} - \Gamma_{kl} < k^{\beta-1}) < \infty$, and the result follows immediately from the Borel-Cantelli Lemma. □

Define $M = \max\{k : \Gamma_{(k+1)l} - \Gamma_{kl} < k^{\beta-1}\}$, so M is finite w.p.1. Since $\Gamma_n/n \to 1$, w.p.1 by LLN we can assume without loss of generality that $\Gamma_{Ml} \geq M^{\beta}/\beta$. Since $\Gamma_{(k+1)l} - \Gamma_{kl} \geq k^{\beta-1}$ for all $k \geq M$, we have $\Gamma_{kl+j} > \Gamma_{kl} \geq k^{\beta}/\beta$ for all $k \geq M$ and $1 \leq j \leq l-1$. Therefore, we have the following result corresponding to Condition (5.2.3).

Proposition 5.2.8. *For any $m > M$,*

$$|\sum_{n=ml}^{\infty} f(\xi_n, \Gamma_n) 1(U_n \leq t)| \leq Er(m) := \sum_{k=m}^{\infty} \frac{l}{\beta} k^{-\beta/\alpha},$$

which converges to 0 as $m \to \infty$.

Having Proposition 5.2.7 and 5.2.8, we now just need to simulate M jointly with $\{\Gamma^n\}$. Note that $\Gamma_n = \sum_{i=1}^{n} V_i$, where V_i's are i.i.d. exponentials with mean 1. Hence the probability $P(\Gamma_{(k+1)l} - \Gamma_{kl} < k^{\beta-1}) = P(V_1 + ...V_l < k^{\beta-1}) := p(k)$ has an explicit expression in k. As a result, in Algorithm 5.2 that simulates M jointly with Γ_n is equivalent to simulate M jointly with V_n, we just modify Algorithm 5.1 slightly, say, replace the c.d.f.'s of normals by $\{p(k)\}$.

Remarks on Algorithm 5.2:

Algorithm 5.2 Simulate Simulate M jointly with $\{\Gamma^n\}$

Input:

$L := \inf\{l : Er(l) < \varepsilon\}$.

Output:

Simulate M jointly with $\{\Gamma^n\}_{n=1}^M$.

1: Initialize $G = m$ and S to be an empty array. Set $I = 1$.
2: **while** I==1 **do**
3: Set $U = 1, D = 0$. Simulate $V \sim Uniform(0,1)$.
4: **while** $U > V > D$ **do**
5: set $G \leftarrow G+1$ and $U \leftarrow p(G) \times U$ and $D \leftarrow (1-1/G) \times U$.
6: **end while**
7: **if** $V \geq U$ **then**
8: add G to the end of S, i.e. $S = [S, G]$.
9: **else**
10: **if** $V \leq D$ **then**
11: $M = G$ and set $I = 0$.
12: **end if**
13: **end if**
14: **end while**
15: **for** n=1:M **do**
16: **if** $n \in S$ **then**
17: generate $V_{kl+1}, ..., V_{(k+1)l}$ conditional on $\{V_{kl+1} + ... + V_{(k+1)l} < k^{\beta-1}\}$;
18: **else**
19: generate $V_{kl+1}, ..., V_{(k+1)l}$ conditional on $\{V_{kl+1} + ... + V_{(k+1)l} \geq k^{\beta-1}\}$.
20: **end if**
21: **end for**

1. Since $l(1-\beta) > 2$, $\sum_{k=G}^{\infty} p(k) < 1/G$ and hence $P(\Gamma_{(k+1)l} - \Gamma_{kl} \geq k^{\beta-1}$ for all $k \geq G) \geq 1 - 1/G$. This validates $D \leftarrow (1-1/G)U$ in Step 5.

2. $l(1-\beta) > 2$ also guarantees that $E[M] < \infty$.

3. As V_i's are exponentials, we can use the standard exponential tilting technique to sample from the conditional distributions in Step 19.

Now we show that the TES algorithm for Lévy processes is also efficient.

Theorem 5.2.9. *The complexity of Algorithm 5.2 is* $O(\varepsilon^{-1})$.

Proof. The proof is similar with that of Theorem 5.2.6. From Proposition 5.2.8, $Er(m) = ln^{-\beta/\alpha}/\beta$. In order to make $Er(m) < \varepsilon$, we can take $m = O(\varepsilon^{-1})$ as $\beta/\alpha > 1$. On the other hand, $E[M]$ is finite and independent of ε. Therefore, the complexity of Algorithm 5.2 is $O(\varepsilon^{-1})$ as the total computational complexity has the same order as $E[m \vee M]$. □

5.3 Application in Multilevel Simulation of Stochastic Differential Equation

In many applications of Monte Carlo method, such as in financial engineering and system performance evaluation, the goal is to efficiently estimate via simulation $\alpha := E[f(X)]$, where X is a stochastic process on $D([0,T],\mathbb{R}^d)$, the space of paths $w : [0,T] \longrightarrow \mathbb{R}^d$ that are right continuous with left limits, and $f : D([0,T],\mathbb{R}^d) \to \mathbb{R}$ is a suitable functional. We shall combine our TES algorithm with the so-called Multilevel Monte Carlo method developed in [27] to design a 'good' simulation strategy(GSS) to estimate α. We call it 'good' because it is almost as efficient as if we can simulate i.i.d. samples exactly following the same distribution of $f(X)$ (which is not obtainable in most cases). Besides, thanks to the TES algorithm, the good simulation strategy works for more a larger class of functionals then the original Multilevel Monte Carlo method.

First let's introduce some definitions and notations.

Definition 5.3.1. *The functional f is said to be **locally Lipschitz** if for all x in $D([0,T],\mathbb{R}^d)$,*

there exists some constant $\kappa = \kappa(x) > 0$ *such that*

$$|f(x) - f(y)| \leq \kappa(x)\|x-y\| \quad \textit{for all } \|x-y\| \leq \delta,$$

where $\|\cdot\|$ *is the supremum norm and* $\delta > 0$ *is independent of x. Moreover, we say that* f *is* ***nice*** *for a stochastic process* X *if* $E[\kappa(X)^2] < \infty$.

Remark: In many queueing models of interest (such as those discussed in the Introduction), the conditions in Definition 5.3.1 are satisfied with X being a Brownian motion or a stable Lévy process. This is because the diffusion limit D is usually the image of a Brownian motion under some Lipschitz mapping Φ, say $D = \Phi(X)$, and the quantity of interest can be expressed as $g(D(t_1), ..., D(t_n))$ with g smooth. For example, consider $f(B) = \Phi(B)_i(t)^2$, where B is a Brownian motion in $\mathbb{R}^d$ and $\Phi(B)_i(t)$ is the i-th component of an RBM. Then $Ef(B)$ represents the second moment of the workload amount at service station i at time t. One can check in this example that it is possible to select $\kappa(B) = 3\Phi(B)_i(t)$ and $E[\kappa(B)^2] < \infty$, so $f(\cdot)$ is a nice locally Lipschitz function.

Now we are ready to state the GSS algorithm 5.3. We now show that 5.3 is almost as efficient as if we can simulate i.i.d. samples exactly following the same distribution of $f(X)$. In the proof we also specifies the allocation of computation resource N_l among different levels and the coupling in Step 4 of GSS algorithm 5.3.

Theorem 5.3.2. *Assume the following conditions:*

1. f *is locally Lipschitz and nice for* X*;*

Algorithm 5.3 Framework of the Good Simulation Strategy

Input:

$\varepsilon > 0$.

Output:

An estimate $\hat{\alpha}$ of $\alpha = E[f(X)]$.

1: Compute $L = \lfloor \log_2 \varepsilon \rfloor$.
2: For $l = 0 : L$, compute the number of replication N_l.
3: For $l = 0$, generate N_l i.i.d. samples $Y_0^{(i)}$ of $f(X^0)$ using the TES algorithm for X with precision $\varepsilon = 1$.
4: For $l = 1 : L$, generate N_l i.i.d. samples $Y_l^{(i)}$ of $f(X^l) - f(X^{l-1})$, using the TES algorithm for X with precision $\varepsilon = 2^{-l}$. (X^{l-1} is coupled and generated simultaneously with X^{l-1}.)
5: **return** $\hat{\alpha} = \sum_{l=1}^{L} N_l^{-1} \sum_{i=1}^{N_l} Y_l^{(i)}$.

2. *For all $\varepsilon > 0$, the ε-TES algorithm for X exists and has complexity $C(\varepsilon) = O(\varepsilon^{-2}\log(1/\varepsilon)^k)$;*

3. *Given X^ε, the complexity (in terms of function evaluations) to compute $f(X^\varepsilon)$ is at most $O(\varepsilon^{-2}\log(1/\varepsilon)^k)$*

Then the good simulation strategy 5.3 for $\alpha = E[f(X)]$ has at most $O(\varepsilon^{-2}\log(1/\varepsilon)^{k+2})$ computational cost.

Remarks: The TES algorithm that we have developed in Section 5.2 Brownian motions and more general Lévy processes satisfies the second condition. So the good simulation strategy 5.3 works for 'nice' functionals of Brownian motions and Lévy processes, which include a large class of SPDEs driven by Browian motions or/and Lévy processes.

The proof follows the line of reasoning in [27]. Here is a brief and simple description of the proof:

For given $\varepsilon > 0$, the complexity to simulate a single path by the ε-TES is $O(\varepsilon^{-2}\log(1/\varepsilon)^k)$. Suppose $L = \lfloor \log_2 \varepsilon \rfloor$, then by Condition 2, for $l = 0, 1, ..., L$, there exists a corresponding ε_l-TES with $\varepsilon_l = 2^{L-l}\varepsilon$ and complexity $C_l = O(2^{2l-2L}\varepsilon^{-2}\log(1/\varepsilon)^k)$. Let us denote by X^l the

sample path generated by the ε_l-TES. It is obvious that

$$E[f(X^L)] = E[f(X^0)] + \sum_{l=1}^{L} E[f(X^l) - f(X^{l-1})].$$

The multi-level method independently estimates each of the expectation on the right hand side in a way that minimizes the computational complexity. The key point is to reduce the variance of the estimator $f(X^l) - f(X^{l-1})$ via a certain coupling. In detail, we couple X^l and X^{l-1} with the same stochastic process X so that the variance V_l is bounded:

$$\begin{aligned} V_l &\le E[(f(X^l) - f(X^{l-1}))^2] \le E[2(f(X^l) - f(X))^2 + 2(f(X^{l-1}) - f(X))^2] \\ &\le E[2\kappa(X)^2((X^l - X)^2 + (X^{l-1} - X)^2)] \le E[\kappa(X)^2] \cdot 2^{2L-2l+2}\varepsilon^2 = O(2^{2L-2l+2}\varepsilon^2). \end{aligned}$$

Such kind of coupling is intrinsic in our tolerance-enforced simulation algorithm as X^l are basically truncations of the same infinite series X. On the aspect of complexity, we have assumed that the complexity to compute $f(X^l)$ for given X^l has at most the same order as that to generate X^l, which is true in most cases in application.

Let N_l be the number of replications used to estimate $E[f(X^l) - f(X^{l-1})]$ and $Y_l^{(i)}$ for $i = 1, ..., N_l$ be the i.i.d. samples of $f(X^l) - f(X^{l-1})$. The multi-level Monte Carlo estimator can be expressed as

$$\hat{\alpha} = \sum_{l=1}^{L} N_l^{-1} \sum_{i=1}^{N_l} Y_l^{(i)}. \tag{5.3.1}$$

By a simple computation, we have that the variance of $\hat{\alpha}$ is $V = \sum_{l=0}^{L} N_l^{-1} V_l$ and that the total computational cost to generate $\hat{\alpha}$ is $C = \sum_{l=0}^{L} N_l C_l$. Recall that $V_l = O(2^{2L-2l+2}\varepsilon^2)$ and

$C_l = O(2^{2l-2L}\varepsilon^{-2}\log(1/\varepsilon)^k)$, by choosing $N_l = L\varepsilon^{-2}2^{-2l+2}$, we have

$$V = O(\sum_{l=1}^{L} L^{-1}2^{2L}\varepsilon^4) = O(2^{2L}\varepsilon^4) = O(\varepsilon^2), \text{ as } L = O(\log(1/\varepsilon))$$

and

$$C = O(\sum_{l=1}^{L} L2^{2-2L}\varepsilon^{-4}\log(1/\varepsilon)^k) = O(L^2 2^{2-2L}\varepsilon^{-4}\log(1/\varepsilon)^k) = O(\varepsilon^{-2}\log(1/\varepsilon)^{k+2}).$$

Note that the total mean squared error can be decomposed as

$$E[|\hat{\alpha}-\alpha|^2] = Var(\hat{\alpha}) + (E[\hat{\alpha}]-\alpha)^2 = V + E[f(X^L)-f(X)]^2.$$

We have already shown that V and $E[f(X^L)-f(X)]^2$ both have the same order, $O(\varepsilon^2)$, therefore, the estimator (5.3.1) gives a good simulation strategy for α.

5.4 Application in Simulation of Reflected Brownian Motions

Reflected Brownian motion(RBM) is an important class of stochastic process in operation research. In this section, we describe how our TES algorithm can be used to simulate the sample pathes of RBM, and especially to estimate the stationary distribution of RBM with an error that is bounded with probability one by some small number chosen by the user. (Note that in most Monte Carlo algorithm, the error is random and has infinite support.) In Section 5.4.1 , we discuss how to use the TES algorithm for Brownian motions, Algorithm 5.1, to design a TES

algorithm for reflected Brownian motions on finite time horizon. In Section 5.4.2, we design a heuristic algorithm to estimate the stationary expectations based on GSS framework 5.3 as discussed in Section 5.3 and give numerical results for some examples. In the end, we describe how to combine the TES algorithm with the perfect sampling algorithm we developed in Chapter 4.3 to obtain almost perfect sampling algorithm for RBM in Section 5.4.3 and the details are given in Appendix D.

5.4.1 Path Simulation

A reflected Brownian motion $\mathbf{Y}(t) = (Y_i(t))_{i=1}^d$ in space $\mathbb{R}^d$ with drift vector μ, covariance matrix Σ and reflected matrix R (we shall donate it as $RBM(\mu, \Sigma, R)$) can be defined in several different ways (see for instance [68]). Here we consider the pathwise formulation based on the Skorokhod problem as defined in 4.2.1. For the Skorokhod problem to be well posed we will restrict R to be of class M. More precisely, we assume that R^{-1} exists and it has non-negative elements. Then $\mathbf{Y}(\cdot)$ is the solution of the Skorokhod problem where the input process $\mathbf{X}$ is a multidimensional Brownian motion with drift vector μ, covariance matrix Σ. In detail, $\mathbf{Y}$ is the unique process that satisfies $\mathbf{Y}(t) = \mathbf{X}(t) + R\mathbf{L}(t) \geq 0$, and $L(t)$ is the minimum of all non-decreasing and non-negative functions that keeps $\mathbf{Y} \geq 0$. Let's denote $\mathbf{Y} = \Psi(\mathbf{X}, R)$.

In the queueing setting, the drift rate μ and variance matrix Σ of the input Brownian motion X and reflection matrix R are obtained in terms of by the arrival process, service rate and routing mechanism of the queueing network (see Chapter 6 in [13]). The following regularity properties of the Skorokhod problem will turnout to be very useful in the design of our algorithm (see for instance Theorem 1 in [35]):

Proposition 5.4.1. *Suppose that the reflection matrix R is of class M. Then, for every $x \in C[0,T]$ (the space of continuous functions on $[0,T]$ taking values on $\mathbb{R}$), there exists a unique pair $(y,l) \in C[0,T] \times C[0,T]$ that satisfies the Skorokhod Problem. Moreover, the Skorokhod mapping $\Psi : C[0,T] \to C[0,T]$ defined by $\Psi(x) = y$ is a contraction map on $C[0,T]$ under the uniform norm.*

Owning to this nice property, we can use the following scheme to generate an approximation to RBM on time interval $[0,T]$ with deterministic error size ε:

Algorithm 5.4 TSE Algorithm for Reflected Brownian Motions

Input:

$\mu \in \mathbb{R}^d$, Σ, $R \in \mathbb{R}^{d\times d}$, $\mathbf{X}(0) = \mathbf{x}_0 \geq 0 \in \mathbf{R}^d$ and $T, \varepsilon > 0$.

Output:

A piecewise linear function $\mathbf{Y}^\varepsilon$.

1: Compute the Cholesky decomposition $\Sigma = A'A$ and let $a = \max |A_{ij}|$.
2: Simulate independently d discretized Brownian path, $\mathbf{B}_i^\varepsilon$ for $i = 1, ..., d$ on $[0,T]$ with an error of size less than ε/ad using Algorithm 5.1.
3: Compute $\mathbf{X}^\varepsilon(t) = \mathbf{x}_0 + A\mathbf{B}^\varepsilon(t) - \mu t$, where $\mathbf{B}^\varepsilon = (B_1^\varepsilon, ..., B_d^\varepsilon)'$. **return** $\mathbf{Y}^\varepsilon = \Psi(\mathbf{X}^\varepsilon, R)$.

Then, according to Proposition 5.4.1, $\|\mathbf{Y}^\varepsilon - \mathbf{Y}\| < \varepsilon$ where Y is the reflected Brownian motion with parameters (μ, Σ, R).

Remark: In Step 3, the Skorokhod problem can be solved explicitly using linear programming as the input process $\mathbf{X}^\varepsilon$ is piecewise linear. Moreover, the complexity to compute $\mathbf{Y}^\varepsilon$ has the same order as the number of linear sections of $\mathbf{X}^\varepsilon(t)$ just equals $E[M \vee 2^L] \cdot T = O(\varepsilon^{-2}\log(1/\varepsilon))$.

Figure 5.1 shows a sample path of a 2-dimensional Reflected Brownian motion generated

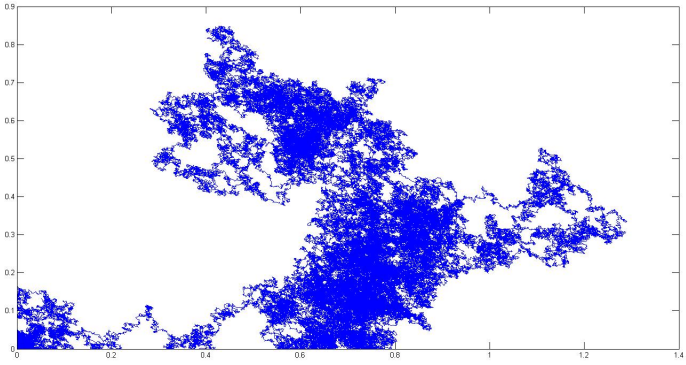


Figure 5.1: Sample path of a 2-dimensional RBM with uniform error of less than 0.01

by Algorithm 5.4 with accuracy parameter $\varepsilon = 0.01$. The parameters are as follows: $\mu = [1,1]$, $\Sigma = [1 - 0.2; -0.2\ 1]$ and $R = [1 - 0.2; -0.2\ 1]$.

5.4.2 Estimating Stationary Expectations

Suppose our goal is to estimate the quantity $a = \pi(\phi) := \int \phi(x)\pi(dx)$ for some function ϕ. Then , $E[\phi(\mathbf{Y}(T))|\mathbf{Y}(0) = 0]$ converges to $\pi(\phi)$ as $T \to \infty$. We prove in Appendix D.1 that an RBM satisfying the stability condition $R^{-1}\mu < 0$ converges exponentially fast in distribution to its unique stationary distribution. Therefore, one can approximate $\pi(\phi)$ by $E[\phi(\mathbf{Y}(n))|\mathbf{Y}(0) = 0]$ for n reasonably large. A straightforward way to estimate the value of $E[\phi(\mathbf{Y}(T))|\mathbf{Y}(0) = 0]$ is to apply the multilevel simulation technique to Algorithm 5.4. To be more efficient, we develop a new multilevel scheme such that we use pathes of different levels of not only the precision ε but also of the time t. Roughly speaking, we 'paste' the segments of i.i.d. sample paths of $\mathbf{Y}^{\varepsilon}$ with different levels of accuracy into a single 'long' path; as the path goes on, we gradually improve the accuracy 'level' of the i.i.d. segments. A detailed description of our algorithm is

given below:

Algorithm 5.5 Estimating Stationary Expectations of RBM

Input: $T > 0$ and $\varepsilon > 0$.

Output: An Estimate $\hat{a}$ of $\pi(\phi)$.

1: Compute the total number of levels $L = L(\varepsilon)$.
2: Compute the number of replications N_l of level-l pathes for $l = 0 : L$.
3: Simulate an approximated RBM $\mathbf{Y}^0$ on $[0, TN_0]$ starting at 0 using Algorithm 5.4 with precision $\varepsilon = 1$. Compute $S_0 = \sum_{i=1}^{N_0} \phi(\mathbf{Y}^0(Ti))/N_0$. Set $\mathbf{y} = \mathbf{Y}^0(TN)$.
4: For $l = 1 : L$ and $i = 1 : N_l$, simulate two approximated RBM $\mathbf{Y}_i^l$ and $\mathbf{Y}_i^{l-1}$ on $[0, T]$ starting at $\mathbf{Y}_i^l(0) = \mathbf{Y}_i^{l-1}(0) = y$ using Algorithm 5.4, which correspond to a pair of coupled approximated Brownian path $\mathbf{X}_i^l$ and $\mathbf{X}_i^{l-1}$. Compute $p_i^l = \phi(\mathbf{Y}_i^l(T)) - \phi(\mathbf{Y}_i^{l-1}(T))$ and update $\mathbf{y} = \mathbf{Y}_i^{l-1}(T)$.
5: : For each l, compute $S_l = \sum_{i=1}^{N_l} p_i^l / N_l$.
6: **return** $\hat{a} = \sum_{l=0}^{L} S_l$.

We implement Algorithm 5.5 in Matlab for $\phi(y) := y$. For simplicity, let's call it GSS (good simulation strategy). We first apply GSS to some 8-dimensional reflected Brownian motions with the true value of $\pi(\phi)$ in closed form. Comparison between the simulation results and the true values are listed in Table 5.1.

Table 5.1: Estimate of stationary expectations of 8-dimensional symmetric RBM.

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.9$	$\rho = -0.05$	$\rho = -0.1$	Complexity*
Estimated Val.	0.179	0.241	0.470	0.167	0.146	1.1×10^7
Stand. Dev.	0.044	0.061	0.042	0.043	0.040	N.A.
True Val.	0.182	0.245	0.468	0.166	0.150	N.A.

RBM parameters: $\mu_i = -1$, $\Sigma_{ii} = 1$,$\Sigma_{ij} = \Sigma_{ji} = \rho$, $R_{ii} = 1$ and $R_{ij} = R_{ji} = r = 0.1$ for all $1 \le i < j \le 8$. Algorithm parameters: $T = 5$ and $\varepsilon = 0.05$. $\varepsilon = 0.05$. * Complexity = total number of one dimensional Gaussian random variables simulated.

Then, we compare Algorithm 5.5 to a well known numerical procedure called QNET developed by [18]. We consider a RBM in 10 dimensions with the same parameters as in the

Table 5.2: Estimate of the stationary mean waiting time at each station for ten stations in series.

Station Number	1	2	3	4	5	6	7	8	9	10	Total time in waiting
Prelimit Simul.	29.55	3.21	2.02	1.79	1.58	1.50	1.44	1.36	1.32	16.36	60.12
QNET(n=4)	32.40	3.25	1.42	1.12	1.04	1.00	0.98	0.96	0.95	8.12	51.24
GSS	31.92	2.72	1.19	0.96	0.84	0.78	0.73	0.70	0.68	16.19	56.72

example shown in Table XIII in [19]. We also report the extensive simulation results of a pre-limit queueing network whose heavy-traffic limit is the RBM in consideration. All the results are reported in Table 5.2.

The QNET is based on approximating the stationary probability density using series expansion of weighted polynomials. Computing the weights of the polynomials up to degree 4, as in [19], requires at least $O(N^2)$ function evaluations with $N = 15015$(see page 78 of [18]). We implement GSS with parameters $T = 15, \varepsilon = 0.05$ so that the total number of one dimensional Gaussian random variables simulated is 1.5×10^8, which is comparable to N^2.

Given the parameters, Station 1 and 10 are heavily loaded with traffic intensities both equal to 0.9. Station 1's steady-state mean waiting time can be computed exactly for the RBM limit, resulting in 32.4, which is obtained exactly by QNET because this procedure is designed to compute exact expectations when the marginal distributions are exponentials, as in the case of Station 1. However, the distribution of Station 10 is not exponential, yet close to heavy-traffic. The fact that QNET substantially underestimate the pre-limit simulated value suggest that the 4-th degree polynomials are not enough to guarantee the convergence of QNET. Our procedure, GSS, showed considerable stability after we performed extensive experiments. Of course, more numerical experiments should be performed, but we believe that for networks of dimensions larger than 10, one might prefer to use GSS given the computational budget.

5.4.3 Perfect Sampling Algorithm

In this section, we shall explain how use the TES algorithm for Brownian motions to adapt the Perfect Sampling Algorithm 4.1 in Chapter 4 to develop an 'almost' perfect sampling algorithm for RBMs. Algorithm 4.1 fails for RBMs due to the following two problems. First, the input process $\mathbf{X}$ is a Brownian motion and hence requires a continuous path description while the computer can only encode and generate discrete objects. Second, the dominating process is a reflected Brownian motion with orthogonal reflection, therefore the hitting time τ to the origin is infinity almost surely (see [67]), which means that Algorithm 4.1 will not terminate in finite time in this case. To solve the first problem, we shall use the TES algorithm for Brownian motions to simulate a piecewise linear approximation with uniformly small (deterministic) error. To solve the second problem, we define an approximated coalescent time τ_ε as the first passage time to a small ball around the origin so that $E[\tau_\varepsilon] < \infty$ and the error caused by replacing τ with τ_ε is bounded by ε. In sum, we concede to an algorithm that is not exact but one that could give any user-defined ε precision.

Following the same notation in Lemma in Section 4.3.1, let $\mathbf{Z}(t) = \mathbf{X}(t) + \mathbf{z}t$. Following the same argument as in Proposition 4.3.7, we can construct the stationary dominating process backwards in time as $\mathbf{Y}^+(-t) = \mathbf{M}(t) - \mathbf{Z}(t)$ where $\mathbf{M}(t) = \max_{t \le u < \infty} \mathbf{Z}(u)$. Now we are ready to give the main structure of our algorithm. The details will be given in Appendix D.

The following proposition shows that the error of the above algorithm has a deterministic bound of order $O(\varepsilon)$.

Proposition 5.4.2. *Suppose* $X \in \mathbb{R}^d$. *Let* $r = \max_{i,j} R_{ij}^{-1} / \min_{i,j}\{R_{ij}^{-1} : R_{ij}^{-1} > 0\}$. *Then there*

Algorithm 5.6 Framework of the Almost Perfect Sampling Algorithm for RBM

Input:

$\mu \in \mathbb{R}^d$, Σ, $R \in \mathbb{R}^{d\times d}$ and error bound $\varepsilon > 0$.

Output:

Random vector $\mathbf{Y} \in \mathbb{R}^d$.

1: Let $\tau_\varepsilon \geq 0$ be any time for which $\mathbf{M}(\tau_\varepsilon) \leq \mathbf{Z}(\tau_\varepsilon) + \varepsilon$ and simulate, jointly with τ_ε, $\mathbf{Z}^{\leftarrow}_{-\tau_\varepsilon}(t) = -\mathbf{Z}^{\varepsilon}(\tau_\varepsilon - t)$ for $0 \leq t \leq \tau_\varepsilon$.

2: Define $\mathbf{X}^{\leftarrow}_{-\tau_\varepsilon}(t) = \mathbf{Z}^{\varepsilon}(\tau_\varepsilon) - \mathbf{Z}^{\varepsilon}(\tau_\varepsilon - t) + \mathbf{z}t$ and compute $\mathbf{Y}^{\varepsilon}_{-\tau_\varepsilon}(\cdot) = \Psi(\mathbf{X}^{\leftarrow}_{-\tau_\varepsilon}(\cdot), R)$ on $[0, \tau_\varepsilon]$.

3: **return** $\mathbf{Y} = \mathbf{Y}^{\varepsilon}_{-\tau_\varepsilon}(\tau_\varepsilon)$.

exists a stationary version $\mathbf{Y}^*$ *of* $\mathbf{Y}$ *such that in each component* i

$$|Y_i^*(0) - Y^{\varepsilon}_{\tau_\varepsilon,i}(\tau_\varepsilon)| \leq (\frac{1}{1-\alpha} + dr)\varepsilon.$$

Proof. Consider three processes on $[-\tau_\varepsilon, 0]$. The first is the coupled stationary process $\mathbf{Y}^*(\cdot)$ as constructed in Proposition 4.3.2, which is the solution to the Skorokhod problem with input process $\tilde{\mathbf{X}}(t) = \mathbf{Y}^*(-\tau_\varepsilon) + \mathbf{Z}(\tau_\varepsilon) - \mathbf{Z}(\tau_\varepsilon - t) + \mathbf{z}t$ for $t \in [0, \tau_\varepsilon]$ and reflection matrix R; the second is a process $\tilde{\mathbf{Y}}(\cdot)$, which is the solution to the Skorokhod problem with input process $\tilde{\mathbf{X}}(\cdot) - \mathbf{Y}^*(-\tau_\varepsilon)$(so its initial value is $\mathbf{0}$); the third is the process $\mathbf{Y}^{\varepsilon}_{-\tau_\varepsilon}(t)$ as we described in the algorithm, which is the solution to the Skorokhod problem with input process $\mathbf{X}^{\leftarrow}_{-\tau_\varepsilon}(t)$ as defined in Step 2 of Algorithm 5.6.

By definition, we know that for each component i, $|Y_i^+(-\tau_\varepsilon)| < \varepsilon$. Since $R^{-1}\mathbf{Y}(\tau_\varepsilon) \leq R^{-1}\mathbf{Y}^+(\tau_\varepsilon)$, the coupled process $Y_i^*(-\tau_\varepsilon) < dr\varepsilon$. Note that $\mathbf{Y}^*(\cdot)$ has the same input data as $\tilde{\mathbf{Y}}(\cdot)$ except for their initial values. According to the comparison theorem of [56], the difference between these two processes is uniformly bounded by the difference of their initial values in each component. Therefore, we can conclude $|Y_i^*(0) - \tilde{Y}_i(0)| < dr\varepsilon$.

On the other hand, $\tilde{\mathbf{Y}}(\cdot)$ and $\mathbf{Y}^{\varepsilon}_{-\tau_\varepsilon}(\cdot)$ have common initial value 0 and input processes whose

Table 5.3: Estimate of Stationary Expectation for a 2-dimensional RBM with precision $\varepsilon = 0.01$.

	Simulation Result	True Value
$E[Y_1(\infty)]$	0.4164±0.0137	0.4167
$E[Y_2(\infty)]$	0.4201±0.0131	0.4167

difference is uniformly bounded by ε. It was proved in [35] that the Skorokhod mapping is Lipschitz continuous under the uniform metric

$$d_T(Y^1(\cdot), Y^2(\cdot)) \triangleq \max_{1\le i\le d} \sup_{0\le t\le T} |Y_i^1(t) - Y_i^2(t)|$$

for all $0 < T < \infty$ and the Lipschitz constant is equal to $1/(1-\alpha)$, where $0 \le \alpha < 1$ is the spectral radius of Q. Therefore, we have that $|\tilde{Y}_i(0) - Y^{\varepsilon}_{-\tau_{\varepsilon},i}(\tau_\varepsilon)| < \varepsilon/(1-\alpha)$.

Simply applying the triangle inequality, we obtain that

$$|Y_i^*(0) - Y^{\varepsilon}_{\tau_\varepsilon,i}(\tau_\varepsilon)| \le (\frac{1}{1-\alpha} + dr)\varepsilon.$$

□

We then implemented a two dimensional RBM example. Let's denote the RBM by $\mathbf{Y}(t)$. The parameters to specify $\mathbf{Y}$ are as follows: drift vector $\mu = (-1,-1)$, covariance matrix $\Sigma = [1,0;0,1]$ and reflection matrix $R = [1,-0.2;-0.2,1]$. Since $\mathbf{Y}$ is a symmetric RBM, one could compute in close that $E[Y_1(\infty)] = E[Y_2(\infty)] = 5/12 \simeq 0.4167$. The output of our simulation algorithm is reported in Table 5.3.

Our implementations here are given with the objective of verifying empirically the validity of the algorithms proposed. We stress that a direct implementation of Algorithm 5.6, although

capable of ultimately producing unbiased estimates of the expectations of RBM, might not be practical. The simulations took substantially more time to be produced than those reported for the stochastic fluid models. Indeed the bottleneck in the algorithm is finding a time at which both stations are close to ε.

Bibliography

[1] S. Asmussen, *Applied probability and queues*, Springer-Verlag, 2003.

[2] S. Asmussen and H. Albrecher, *Ruin probabilities*, 2 ed., World Scientific, New Jercy, US., 2010.

[3] S. Asmussen and P. Glynn, *Stochastic simulation: Algorithms and analysis*, vol. 57, Springer-Verlag, 2007.

[4] S. Asmussen and J. Rosiński, *Approximations of small jumps of Lévy processes with a view towards simulation*, Annals of Applied Probability **38** (2001), 482–493.

[5] A. Beskos, S. Peluchetti, and G. Roberts, ε-*Strong simulation of the Brownian path*, Bernoulli **18** (2012), no. 4, 1223–1248.

[6] P. Billingsley, *Convergence of probability measures, 2nd edition*, Wiley, 1999.

[7] J. Blanchet and K. Sigman, *On exact sampling of stochastic perpetuities*, Journal of Applied Probability **48A** (2011), 165–182.

[8] L. Bondesson, *On simulation from infinity divisible distributions*, Advances in Applied Probabililty **14** (1982), 855–869.

[9] J. Bouchard, Y. Gefen, M. Potters, and M. Wyart, *Fluctuations and response in financial markets: the subtle nature of 'random' price changes*, Quantitative Finance **4** (2004), no. 2, 176.

[10] J. Bouchard, M. Mezard, and M. Potters, *Statistical properties of stock order books: empirical results and models*, Quantitative Finance **2** (2002), no. 4, 251.

[11] O. Boxma and Ivanovs, *Two coupled Lévy queues with independent input*, Working paper (2013).

[12] A. Cartea, S. Jaimungal, and J. Ricci, *Buy low sell high: a high frequency trading perspective*, Working paper (2011).

[13] H. Chen and D. D. Yao, *Fundamentals of queueing networks*, Springer-Verlag, 2001.

[14] R. Cont, A. Kukanov, and S. Stoikov, *The price impact of order book events*, Journal fo Financial Econometrics (2013).

[15] R. Cont and A. Larrard, *Price dynamics in a Markovian limit order book market*, SIAM Journal on Financial Mathematics **4** (2013), no. 1, 1–25.

[16] R. Cont, S. Stoikov, and R. Talreja, *A stochastic model for order book dynamics*, Operations Research **58** (2010), 549–563.

[17] R. Cont and P. Tankov, *Financial modelling with jump processes*, Chapman and Hall/ CRC Press, 2003.

[18] J. G. Dai and J. M. Harrison, *Reflected Brownian motion in an orthant: numerical methods for steady-state analysis*, Annals of Applied Probability **2** (1992), 65–86.

[19] J. G. Dai, V. Nguyen, and M. I. Reiman, *Sequential bottleneck decomposition: an approximation method for generalized jackson networks*, Operations Research **42** (1994), 119–136.

[20] K. Debicki, T. Dieker, and T. Rolski, *Quasi-product forms for Lévy-driven fluid networks*, Mathematics of Operations Research **32** (2007), 629–647.

[21] L. Decreusefond and P. Moyal, *A functional central limit theorem for theM/GI/∞ queue*, Annals of Applied Probability **18** (2008), 2156–2178.

[22] A. Dembo and O. Zeitouni, *Large deviations: Techniques and applications*, 2 ed., Springer, New York, 1998.

[23] P. Dupuis and R.J. Williams, *Lyapunov functions for semimartingale reflecting Brownian motions*, The Annals of Probability **22** (1994), 680–702.

[24] K. B. Ensor and P. W. Glynn, *Simulating the maximum of a random walk*, Journal of Statistical Planning and Inference **85** (2000), 127–135.

[25] S. Ethier and T. Kurtz, *Markov processes: Characterization and convergence*, Wiley, 1986.

[26] J. Feng and T. Kurtz, *Large deviations for stochastic processes*, vol. 131, American Mathematical Society, 2006.

[27] M. Giles, *Multilevel Monte Carlo path simulation*, Operations Research **56** (2008), 607–617.

[28] P. Glynn, *Stochastic networks*, Lecture Notes in Statistics, vol. 71, ch. Large deviations for the infinite server queue in heavy traffic, pp. 387–395, Springer, New York, 1995.

[29] P. Glynn and W. Whitt, *A new view of the heavy-traffic limit theorem for the infinite-server queue*, Annals of Applied Probability **19** (1991), 2211–2269.

[30] _____, *Large deviations behavior of counting processes and their inverses*, Queueing Systems **17** (1994), 107–128.

[31] M. Gould, M. Porter, S. Wiliams, M. McDonald, D. Fenn, and S. Howison, *Limit order book*, Working paper (2012).

[32] A. Gut, *Stopped random walks: Limit theorems and applications*, Springer-Verlag, 2009.

[33] S. Halfin and W. Whitt, *Heavy-traffic limits for queues with many exponential servers*, Operations research **29** (1981), 567–588.

[34] J. M. Harrison and V. Nguyen, *Brownian models of multiclass queueing networks: current status and open problems*, Queueing Systems **13** (1993), 5–40.

[35] J. M. Harrison and M. I. Reiman, *Reflected Brownian motion on an orthant*, Annals of Probability **9** (1981), 302–308.

[36] J. Hasbrouck and G. Saar, *Technology and liquidity provision: The blurring of traditional definitions*, Journal of Finanical Markets **12** (2009), 143–172.

[37] N. Hautsch and R. Huang, *The market impact of a limit order*, Journal of Economic Dynamics and Control **36** (2012), 501–522.

[38] U. Horst and M. Paulsen, *A law of large numbers for limit order books*, Working paper (2013).

[39] D. Iglehart, *Limit diffusion approximations for the many server queue and repairman problem*, Journal of Applied Probability **2** (1965), 429–441.

[40] P. Jelenkovic, A. Mandelbaum, and P. Momcilovic, *Heavy traffic limits for queues with many deterministic servers*, Queueing Systems **47** (2004), 53–69.

[41] H. Kaspi and K Ramanan, *Law of large numbers limits for many server queues*, Annals of Applied Probability **21** (2011), 33–114.

[42] H. Kaspi and K. Ramanan, *SPDE limit of many-server queues*, Annals of Applied Probability (in press).

[43] O. Kella and W. Whitt, *Stability and structural properties of stochastic storage networks*, Journal of Applied Probability **33** (1996), 1169–1180.

[44] W. Kendall, *Geometric ergodicity and perfect simulation*, Electronic Communications in Probability **9** (2004), 140–151.

[45] T. G. Kurtz, *Averaging for martingale problems and stochastic approximation*, Applied Stochastic Analysis (Ioannis Karatzas and Daniel Ocone, eds.), Lecture Notes in Control and Information Sciences, vol. 177, Springer Berlin Heidelberg, 1992, pp. 186–209.

[46] P. Lakner, J. Reed, and S. Stoikov, *High frequency asymptotic for the limit order book*, Working paper (2013).

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

[48] R. LePage, *Multidimensional infinitely divisidle variables and processes part ii*, Probability in Banach Spaces III (Anatole Beck, ed.), Lecture Notes in Mathematics, vol. 860, Springer Berlin Heidelberg, 1981, pp. 279–284.

[49] C. Maglaras, C.C. Moallemi, and H. Zheng, *Optimal order routing in a fragmented market*, Working paper (2012).

[50] S. Meyn and R. Tweedie, *Markov chain and stochastic stability*, Cambridge, 1993.

[51] I. Muni Toke, *Econophysics of order-driven markets*, ch. "Market making" in an order book model and its impact on the bid-ask spread, Springer-Verlag Milan, 2010.

[52] G. Pang and W. Whitt, *Two-parameter heavy-traffic limits for infinite-server queues*, Queueing Systems **65** (2010), 325–364.

[53] M. Potters and J. Bouchard, *More statistical properties of order books and price impact*, Physica A **324** (2003), no. 1-2, 133.

[54] J. G. Propp and D. B. Wilson, *Exact sampling with coupled Markov chains and applications to statistical mechanics*, Random Structures & Algorithms **9** (1996), 223–252.

[55] A. A. Puhalskii and M. I. Reiman, *The multiclass $GI/PH/N$ queues in the Halfin-Whitt regime*, Advances in Applied Probabililty **32** (2000), 564–595.

[56] S. Ramasubramanian, *A subsidy-surplus model and the Skorokhod problem in an orthant*, Mathematics of Operations Research **25** (2000), no. 3, 509–538.

[57] J. Reed, *The $G/G/N$ queues in the Halfin-Whitt regime I: infinite server queue system equations*, Annals of Applied Probability (Forthcoming).

[58] J. Reed and R. Talreja, *Distribution-valued heavy-traffic limits for the G/GI/∞ queue, working paper*, (2009).

[59] M. I. Reiman and R. J. Williams, *A boundary property of semimartingale reflecting Brownian motions*, Probability Theory and Related Fields **77** (1988), 87–97.

[60] J. Rosiński, *Series representations of lévy processes from the perspective of point processes*, Lévy Processes (OleE. Barndorff-Nielsen, SidneyI. Resnick, and Thomas Mikosch, eds.), Birkh auser Boston, 2001, pp. 401–415 (English).

[61] _____, *Tempered stable processes*, Stochastic Processes and Their Applications **117** (2007), 677–707.

[62] S. Rubenthaler and M. Wiktorsson, *Improved convergence rate for the simulation of subordinated Lévy processes*, Stochastic Processes and their Applications **103** (2003), 311–349.

[63] G. Samorodnitsky and M.L. Taqqu, *Non-Gaussian stable processes*, Chapman and Hall, 1994.

[64] R. Sowers, A. Kirilenko, and X. Meng, *A multiscale model of high-frequencey trading*, Algorithmic Finance **2** (2013), no. 1, 59–98.

[65] J. M. Steele, *Stochastic calculus and financial application*, Springer-Verlag, 2001.

[66] L. M. Taylor and R. J. Williams, *Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant*, Probability Theory and Related Fields **96** (1993), 283–317.

[67] S. R. S. Varadhan and R. J. Williams, *Brownian motion in a wedge with oblique reflection*, Communications on Pure and Applied Mathematics **38** (1985), no. 4, 405–443.

[68] R. J. Williams, *Semimartingale reflecting brownian motions in the orthant, stochastic networks, ima volumes in mathematics and its applications*, vol. 71, Springer-Verlag, New York, 1995.

[69] M. Wyart, J. Bouchard, J. Kockelkoren, M. Potters, and M. Vettorazzo, *Relation between bid-ask spread, impact and volatility in order-driven markets*, Quantitative Finance **8** (2008), no. 1, 41.

[70] T. Zajic, *Rough asymptotics for tandem non-homogeneous M/G/∞ queues via Poissonized empirical processes*, Queueing Systems **29** (1998), 161–174.

[71] B. Zheng, F. Roueff, and F. Abergel, *Modelling Bid and Ask prices using constrained Hawkes processes ergodicity and scaling limit*, Working paper (2013).

[72] I. Zovko and J. Farmer, *The power of patience: a behavioral regularity in limit order placement*, Quantitative Finance **2** (2002), no. 5, 387.

Appendix A

Appendix for Chapter 2

A.1 Construction of an Auxiliary Continuous Process

In this section we provide the explicit construction of the process $(Q^*_\lambda(t,y) : 0 \leq t \leq T, y \geq 0)$ introduced in Section 2.3.2 in order to define our exponentially equivalent continuous process, $\tilde{Q}_\lambda$.

The construction will be based on polygonal interpolations, so it will be convenient to introduce some notation.

First, given (t,y) and (t',y') where $t \neq t'$ we write $Q_\lambda(t,y) \leftrightarrow Q_\lambda(t',y')$ to denote the straight line that joins the points $(t,y,Q_\lambda(t,y))$ and $(t',y',Q_\lambda(t',y'))$ in the associated three-dimensional space.

Now, given a sample path of the process $Q_\lambda(\cdot)$, consider the set $\{t_1,...,t_m\}$ of points corresponding to either arrivals or departures in the interval $[0,T]$ (in increasing order); and put $t_0 = 0$ and $t_{m+1} = T$. First let us consider $Q_\lambda(t,\cdot)$ for a fixed time $t \in \{t_0,...,t_{m+1}\}$. Let

$\{y_1(t),...,y_{n(t)}(t)\}$ be the set of discontinuities of the function $Q_\lambda(t,\cdot)$ with $n(t)$ equal the number of customers present in the system at time t (recall again that $Q_\lambda(t,\cdot)$ is a right continuous non-increasing step function). Interpolate using straight lines forming the segments $Q_\lambda(t,0) \leftrightarrow Q_\lambda(t,y_1(t))$, $Q_\lambda(t,y_1(t)) \leftrightarrow Q_\lambda(t,y_2(t)),\ldots,Q_\lambda(t,y_{n(t)-1}(t)) \leftrightarrow Q_\lambda(t,y_{n(t)}(t))$.

The next step is to join the end points of these straight lines to the end points of adjacent (suitably matched in the time axis) end points of straight lines in order to form segments of adjacent planes. In order to do this matching note that for each successive t_i and t_{i+1}, either $Q_\lambda(t_{i+1},\cdot)$ has one less discontinuous point than $Q_\lambda(t_i,\cdot)$ (i.e. a departure occurs at t_{i+1}) or one more discontinuity point (i.e. an arrival occurs at t_{i+1}); the exception is the last segment from t_m to $t_{m+1}=T$, where there might be no difference between the number of discontinuity points between $Q_\lambda(t_m,\cdot)$ and $Q_\lambda(t_{m+1},\cdot)$. Note that batch arrivals are not possible since the interarrival times are positive.

According to the notation introduced earlier for discontinuity points, $y_1(t_i),...,y_{n(t_i)}(t_i)$ are the discontinuous points of $Q_\lambda(t_i,\cdot)$ with corresponding values $Q_\lambda(t_i,y_1(t_i))$, $Q_\lambda(t_i,y_2(t_i))$, ..., $Q_\lambda(t_i,y_{n(t_i)}(t_i))$. We will explain how to joint discontinuity points of $Q_\lambda(t_i,\cdot)$ with those from $Q_\lambda(t_{i+1},\cdot)$.

Suppose a departure occurs at time t_{i+1}. Then we can label the discontinuous points of $Q_\lambda(t_{i+1},\cdot)$ as $y_1(t_{i+1}),...,y_{n(t_{i+1})}(t_{i+1})$, with $n(t_{i+1})=n(t_i)-1$. We form a set of straight lines $Q_\lambda(t_i,0) \leftrightarrow Q_\lambda(t_{i+1},0) \leftrightarrow Q_\lambda(t_i,y_1(t_i)) \leftrightarrow Q_\lambda(t_{i+1},y_1(t_{i+1})) \leftrightarrow Q_\lambda(t_i,y_2(t_i)) \leftrightarrow Q_\lambda(t_{i+1},y_2(t_{i+1})) \leftrightarrow$ $... \leftrightarrow Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1})) \leftrightarrow Q_\lambda(t_i,y_{n(t_i)}(t_i))$ in a zig-zag manner; together with another set of straight lines $Q_\lambda(t_i,0) \leftrightarrow Q_\lambda(t_i,y_1(t_i)) \leftrightarrow \ldots \leftrightarrow Q_\lambda(t_i,y_{n(t_i)}(t_i))$, and also the set of straight

lines $Q_\lambda(t_{i+1},0) \leftrightarrow Q_\lambda(t_{i+1},y_1(t_{i+1})) \leftrightarrow \ldots \leftrightarrow Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1}))$. These three sets describe a series of adjacent triangular planar sections which jointly form a continuous surface.

Similarly, suppose that an arrival occurs at time t_{i+1}. Then we can label the discontinuous points of $Q_\lambda(t_{i+1},\cdot)$ as $Q_\lambda(t_{i+1},y_1(t_{i+1}))$, ..., $Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1}))$, with $n(t_{i+1}) = n(t_i)+1$. We then form the set of straight lines $Q_\lambda(t_{i+1},0) \leftrightarrow Q_\lambda(t_i,0) \leftrightarrow Q_\lambda(t_{i+1},y_1(t_{i+1})) \leftrightarrow Q_\lambda(t_i,y_1(t_i)) \leftrightarrow Q_\lambda(t_{i+1},y_2(t_{i+1})) \leftrightarrow ... \leftrightarrow Q_\lambda(t_i,y_{n(t_i)}(t_i)) \leftrightarrow Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1}))$. Again, together with a second set of straight lines $Q_\lambda(t_i,0) \leftrightarrow Q_\lambda(t_i,y_1(t_i)) \leftrightarrow \ldots \leftrightarrow Q_\lambda(t_i,y_{n(t_i)}(t_i))$, and a third set of straight lines, namely $Q_\lambda(t_{i+1},0) \leftrightarrow Q_\lambda(t_{i+1},y_1(t_{i+1})) \leftrightarrow \ldots \leftrightarrow Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1}))$. These three sets of straight lines, once again describe a series of adjacent triangular planar sections which jointly form a continuous surface. The last time interval from t_m to T is dealt with similarly, with perhaps one less triangle formed if $n(t_m) = n(T)$.

The continuous function $(Q^*_\lambda(t,y) : 0 \le t \le T, y \ge 0)$ is defined by concatenating all these adjacent triangular planar regions as one varies t_i and t_{i+1} for $i \in \{0,1,...,m\}$, and setting $Q^*_\lambda(t,y) = 0$ for the region where y is beyond the boundary of the last triangular plane i.e. beyond the lines $Q_\lambda(t_i,y_{n(t_i)}(t_i)) \leftrightarrow Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1})), i \in \{0,1,...,m\}$. It is immediate from the previous construction, and the fact that $Q_\lambda(t,\cdot)$ is non-increasing, that $Q^*_\lambda(t,\cdot)$ is also non-increasing for each $t \in [0,T]$.

A.2 Proofs of Technical Results in Section 2.3.3

We start with Lemma 2.3.2 which takes advantage of the Dawson-Gartner projective limit theorem and thus requires that we obtain an auxiliary large deviations principle for finite dimen-

sional objects defined via

$$\Delta_{ij}(\lambda) = \sum_{k=N_\lambda(t_{i-1})+1}^{N_\lambda(t_i)} I(y_{j-1} < A_k/\lambda + V_k \le y_j)$$

$$= \bar{Q}_\lambda(t_i, y_{j-1}) - \bar{Q}_\lambda(t_i, y_j) - \bar{Q}_\lambda(t_{i-1}, y_{j-1}) + \bar{Q}_\lambda(t_{i-1}, y_j), \tag{A.1}$$

for $t_{i-1} < t_i$, and $y_{j-1} < y_j$.

Lemma A.2.1. *For* $0 = t_0 < t_1 < t_2 < ... < t_m \le T$ *and* $0 = y_0 < y_1 < ... < y_n < y_{n+1} = T + K$, $(\Delta_{ij}(\lambda)/\lambda : 1 \le i \le m, 1 \le j \le n+1)$ *possesses a large deviations principle with a good rate function*

$$\sup_{\theta_{ij}: 1 \le i \le m, 1 \le j \le n+1} \sum_{i=1}^{m} \sum_{j=1}^{n+1} \theta_{ij}\delta_{ij} - \sum_{i=1}^{m} \int_{t_{i-1}}^{t_i} \psi_N \left(\log \sum_{j=1}^{n+1} e^{\theta_{ij}} P(y_{j-1} - u < V_1 \le y_j - u) \right) du.$$

Proof of Lemma A.2.1. We use that $\psi_N(\cdot)$ is continuously differentiable over $\mathbb{R}$. Since U_i are non-lattice, the key renewal theorem implies that for any set of $0 \le t_0 < t_1 < t_2 < \cdots < t_m$,

$$\lim_{\lambda \to \infty} \frac{1}{\lambda} \log E \exp \left\{ \sum_{i=1}^{m} \theta_i (N_\lambda(t_i) - N_\lambda(t_{i-1})) \right\} = \sum_{i=1}^{m} \psi_N(\theta_i)(t_i - t_{i-1})$$

for any $\theta_i \in \mathbb{R}$; see [30] p. 115 and [28] p. 390, and also (2.2.1). Then, from [28], the Gartner-

Ellis limit $\Lambda(\Theta)$ of $(\Delta_{ij}(\lambda) : 1 \le i \le m, 1 \le j \le n+1)$ equals

$$\begin{aligned}\Lambda(\Theta) &\triangleq \lim_{\lambda\to\infty} \frac{1}{\lambda} \log E \exp\left\{ \sum_{i=1}^{m} \sum_{j=1}^{n+1} \theta_{ij}\Delta_{ij}(\lambda) \right\} \\ &= \sum_{i=1}^{m} \int_{t_{i-1}}^{t_i} \psi_N \left(\log \sum_{j=1}^{n+1} e^{\theta_{ij}} P(y_{j-1} - u < V_1 \le y_j - u) \right) du\end{aligned}$$

which is finite for any $\Theta := (\theta_{i,j} : 1 \le i \le m, 1 \le j \le n+1)$. Moreover, for any $t_{i-1} < u \le t_i$,

$$\begin{aligned}&\left| \frac{\partial}{\partial\theta_{ij}} \psi_N \left(\log \sum_{k=1}^{n+1} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u) \right) \right| \\ &= \left| \psi_N' \left(\log \sum_{k=1}^{n+1} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u) \right) \right| \cdot \frac{e^{\theta_{ij}} P(y_{j-1} - u < V_1 \le y_j - u)}{\sum_{k=1}^{n+1} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u)} \\ &\le \max\{ \ |\psi_N'(\max\{\theta_{ik}, k = 1, \ldots, n+1\})| \ , \ |\psi_N'(\min\{\theta_{ik}, k = 1, \ldots, n+1\})| \ \}\end{aligned}$$

which is uniformly bounded over a neighborhood of θ_{ij} and $t_{i-1} < u \le t_i$, fixing all other θ_{lk}'s. Therefore,

$$\begin{aligned}&\frac{1}{h} \left| \psi_N \left(\log \left(e^{\theta_{ij}+h} P(y_{j-1} - u < V_1 \le y_j - u) + \sum_{k\ne j} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u) \right) \right) \right. \\ &\quad \left. - \psi_N \left(\log \sum_{k=1}^{n+1} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u) \right) \right|\end{aligned}$$

is also uniformly bounded on the same region. By dominated convergence theorem, we have

$$\begin{aligned}&\frac{\partial}{\partial\theta_{ij}} \Lambda(\Theta) \\ &= \int_{t_{i-1}}^{t_i} \psi_N' \left(\log \sum_{k=1}^{n+1} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u) \right) \cdot \frac{e^{\theta_{ij}} P(y_{j-1} - u < V_1 \le y_j - u)}{\sum_{k=1}^{n_i} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u)} du.\end{aligned}$$

Moreover, it is dominated by

$$(t_i - t_{i-1}) \max\{|\psi_N'(\max\{\theta_{ik}, k = 1, \ldots, n+1\})|, |\psi_N'(\min\{\theta_{ik}, k = 1, \ldots, n+1\})|\} < \infty$$

for any given $\Theta \in \mathbb{R}^{m\times(n+1)}$. Since $\Lambda(\cdot)$ is finite and differentiable everywhere on $\mathbb{R}^{m\times(n+1)}$, by the Gartner-Ellis Theorem for the case $\mathcal{D}_\Lambda = \mathbb{R}^{m\times(n+1)}$ ([22], p. 52, Ex 2.3.20 (g)), $\{\Delta_{ij}(\lambda)\}$ possesses a rate function

$$\sup_{\theta_{ij}:1\le i\le m, 1\le j\le n+1} \sum_{i=1}^{m}\sum_{j=1}^{n+1} \theta_{ij}\delta_{ij} - \sum_{i=1}^{m} \int_{t_{i-1}}^{t_i} \psi_N \left(\log \sum_{j=1}^{n+1} e^{\theta_{ij}} P(y_{j-1} - u < V_1 \le y_j - u) \right) du.$$

We argue that the rate function is good. By [22] p. 8, Lemma 1.2.18, it suffices to show that $(\Delta_{ij}(\lambda) : 1 \le i \le m, 1 \le j \le n+1)$ is exponentially tight. Denoting $\|\cdot\|_1$ as the L_1-norm, we have by Chernoff's bound

$$\overline{\lim}_{\lambda\to\infty} \frac{1}{\lambda} \log P\left(\left\|\Delta_{ij}(\lambda)/\lambda\right\|_1 > \alpha\right) \le \overline{\lim}_{\lambda\to\infty} \frac{1}{\lambda} \log P(N_\lambda(T) > \alpha\lambda) \le -\theta\alpha + \psi_N(\theta),$$

for any $\theta > 0$. Sending $\alpha \to \infty$ we then obtain

$$\overline{\lim}_{\alpha\to\infty} \overline{\lim}_{\lambda\to\infty} \frac{1}{\lambda} \log P\left(\left\|\Delta_{ij}(\lambda)/\lambda\right\|_1 > \alpha\right) = -\infty,$$

thereby obtaining exponential tightness and the goodness of the underlying rate function as claimed. □

Proof of Lemma 2.3.2. We will use the Dawson-Gartner projective limit theorem. Consider

a collection of points in the plane of the form $\kappa = \left(\left(t_i, y_j\right) : 1 \leq i \leq m, 0 \leq j \leq n\right)$, such that $0 := t_0 < t_1 < t_2 < ... < t_m \leq T$ and $0 := y_0 < y_1 < ... < y_n$. Moreover, we assume that $y_l = t_l$ if $0 \leq l \leq \min(m,n)$. Let $\mathcal{K}$ be the union of such collection of sets κ. Further, let $\{p_\kappa\}_{\kappa \in \mathcal{K}}$ be the projective system generated by $\mathcal{K}$. We will proceed to obtain a large deviations principle for the projections $\left(\bar{Q}_\lambda(t,y)/\lambda : (t,y) \in \kappa\right)$. However, we will do this by first obtaining a large deviations principle for quantities $\Delta_{ij}(\lambda)/\lambda$ and then the large deviations principle for the projections follows using the contraction principle as the $\left(\bar{Q}_\lambda(t,y)/\lambda : (t,y) \in \kappa\right)$ will be shown to be continuous functions. Set $y_{n+1} = \infty$, so that $\bar{Q}_\lambda(t, y_{n+1}) = 0$ for every $t \in [0,T]$. It is important to note, given the structure of the partition κ, that if $1 \leq i \leq m$, $1 \leq j \leq n$, and $i > j$, then $\Delta_{ij}(\lambda) = 0$. Now, similar to the definition of $\Delta_{ij}(\lambda)$ we define, for $1 \leq i \leq m$ and $1 \leq j \leq n+1$,

$$\widetilde{\Delta}_{ij}(\lambda) = \tilde{Q}_\lambda(t_i, y_{j-1}) - \tilde{Q}_\lambda(t_i, y_j) - \tilde{Q}_\lambda(t_{i-1}, y_{j-1}) + \tilde{Q}_\lambda(t_{i-1}, y_j). \tag{A.2}$$

Once again, observe that $\tilde{Q}_\lambda(t, y_{n+1}) = 0$, and also if $i > j$, for $1 \leq i \leq m$, $1 \leq j \leq n$, we have $t_{i-1} \geq y_j$ and therefore

$$\begin{aligned}
\widetilde{\Delta}_{ij}(\lambda) = {} & \tilde{Q}_\lambda(y_{j-1}, y_{j-1}) + \tilde{N}_\lambda(t_i) - \tilde{N}_\lambda\left(y_{j-1}\right) - (\tilde{Q}_\lambda(y_j, y_j) + \tilde{N}_\lambda(t_i) - \tilde{N}_\lambda\left(y_j\right)) \\
& - (\tilde{Q}_\lambda(y_{j-1}, y_{j-1}) + \tilde{N}_\lambda(t_{i-1}) - \tilde{N}_\lambda\left(y_{j-1}\right)) + (\tilde{Q}_\lambda(y_j, y_j) + \tilde{N}_\lambda(t_{i-1}) - \tilde{N}_\lambda\left(y_j\right)) \\
& = 0.
\end{aligned}$$

Moreover, clearly we have for $1 \leq i \leq m$, and $1 \leq j \leq n$

$$\tilde{Q}_\lambda(t_i, y_j) = \sum_{l=1}^{i} \sum_{r=j+1}^{n+1} \widetilde{\Delta}_{lr}(\lambda),$$

so indeed we have that $(\tilde{Q}_\lambda(t_i, y_j) : 1 \leq i \leq m, 1 \leq j \leq n+1)$ can be recovered as a continuous function of the $\widetilde{\Delta}_{lr}(\lambda)$'s. Since $\|\tilde{Q} - \bar{Q}\| \leq 4$ by (2.3.8), it follows by the triangle inequality and from (A.1) and (A.2) that $\left|\widetilde{\Delta}_{ij}(\lambda) - \Delta_{ij}(\lambda)\right| \leq 16$ and hence we have

$$\lim_{\lambda\to\infty} \frac{1}{\lambda} \log E \exp\left\{ \sum_{i=1}^{m} \sum_{j=i}^{n+1} \theta_{ij} \Delta_{ij}(\lambda) \right\} = \lim_{\lambda\to\infty} \frac{1}{\lambda} \log E \exp\left\{ \sum_{i=1}^{m} \sum_{j=i}^{n+1} \theta_{ij} \widetilde{\Delta}_{ij}(\lambda) \right\}.$$

Consequently, from Lemma A.2.1, the rate function for the projections represented by κ (these projections are denoted by $p_\kappa(\bar{q})$) can be written as

$$I(p_\kappa(\bar{q}))$$
$$= \sup_{\{\theta_{ij}: 1 \leq \iota \leq m, 1 \leq j \leq n+1\}} \sum_{i=1}^{m} \sum_{j=1}^{n} \theta_{ij} \delta_{ij}(\kappa) - \sum_{i=1}^{m} \int_{t_{i-1}}^{t_i} \psi_N \left(\log \sum_{j=i}^{n+1} e^{\theta_{ij}} P(y_{j-1} - u < V_1 \leq y_j - u) \right) du.$$

To possess a finite $I(p_\kappa(\bar{q}))$, the quantity $\delta_{ij}(\kappa) := \bar{q}(t_i, y_{j-1}) - \bar{q}(t_i, y_j) - \bar{q}(t_{i-1}, y_{j-1}) + \bar{q}(t_{i-1}, y_j)$ must satisfy that

$$\delta_{ij}(\kappa) = 0 \tag{A.3}$$

for $i > j$, and $1 \leq i \leq m$, $1 \leq j \leq n+1$; otherwise, if $\delta_{ij}(\kappa) \neq 0$, the rate function can be made arbitrarily large by picking $\theta_{ij} = c \times sgn(\delta_{ij}(\kappa))$ with arbitrarily large constant $c > 0$ for

$1 \leq j < i \leq m$, as

$$\int_{t_{i-1}}^{t_i} \psi_N \left(\log \sum_{j=i}^{n+1} e^{\theta_{ij}} P(y_{j-1} - u < V_1 \leq y_j - u) \right) du$$

is independent of θ_{ij}'s that have $j < i$. In the representation of the rate function $I(p_\kappa(\bar{q}))$ we have also used the fact that $\bar{q}(t_i, y_j) = \sum_{l \leq i, r > j} \delta_{lr}(\kappa)$, with $\bar{q}(0, y_j) = 0$, so the relation from the $\delta_{ij}(\kappa)$'s to the $\bar{q}(t_i, y_j)$ is a one-to-one, continuous function, so that the contraction principle (Theorem 4.2.1, [22]) is invoked for the above representation for $I(p_\kappa(\bar{q}))$. We want to show that $\sup_{\kappa \in \mathcal{K}} I(p_\kappa(\bar{q}))$ is equal to (2.3.9), and hence conclude the proof by Dawson-Gartner Theorem (see Theorem 4.6.1, [22]). Clearly it suffices to concentrate on functions $\bar{q}$ such that $\bar{q}(t, y) = 0$ whenever $t > T$ or $y > t + K$ given that we are assuming service times bounded by K. Note that the constraint (A.3) implies that for any $\bar{q}$, in order that $I(\bar{q}) < \infty$, we must have absolute continuity throughout $0 \leq y \leq t \leq T$ and, moreover, that

$$\partial^2 \bar{q}(t, y) / (\partial y \partial t) = 0$$

almost everywhere on $0 \leq y \leq t \leq T$ (see [22] p. 189). We now focus on $\bar{q}(t, y)$ that is absolutely continuous on $\mathcal{D}_K$ and has $\partial^2 \bar{q}(t, y) / (\partial y \partial t) = 0$ almost everywhere on $0 \leq y \leq t \leq T$. Observe that

$$\begin{aligned} \delta_{ij}(\kappa) &= \bar{q}(t_i, y_{j-1}) - \bar{q}(t_i, y_j) - \bar{q}(t_{i-1}, y_{j-1}) + \bar{q}(t_{i-1}, y_j) \\ &= - \int_{t_{i-1}}^{t_i} \int_{y_{j-1}}^{y_j} \frac{\partial^2}{\partial t \partial y} \bar{q}(t, y) dy dt. \end{aligned}$$

Regarding $\theta(\cdot,\cdot)$ as a step function with jumps at $0 = t_1 < t_2 < \cdots < t_m \leq T$ and $0 \leq y_0 < y_1 < \cdots < y_n < y_{n+1} = T + K$, and denote $S(\mathcal{C})$ as the set of all step functions on a given domain $\mathcal{C}$. We can write

$$\begin{aligned}
&\sup_{\kappa} I(p_\kappa(\bar{q})) \\
&= \sup_{\theta(\cdot,\cdot)\in S(\mathcal{D}_K)} \left\{ \int_0^T \int_t^{t+K} \theta(t,y) \left(-\frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right) dy dt - \int_0^T \psi_N \left(\log \int_t^{t+K} e^{\theta(t,y)} dF(y-t) \right) dt \right\}
\end{aligned} \tag{A.4}$$

To show that $\sup_\kappa I(p_\kappa(\bar{q})) \geq I(\bar{q})$ where $I(\bar{q})$ is as defined in (2.3.9), note first that the set of step functions $S(\mathcal{D}_K)$ is dense in $C(\mathcal{D}_K)$, the set of continuous functions equipped with the uniform metric. So for any continuous function $\theta(\cdot,\cdot) \in C(\mathcal{D}_K)$, we can find a sequence $\theta_k(\cdot,\cdot) \in S(\mathcal{D}_K)$ with $\|\theta_k - \theta\|_{\mathcal{D}_K} \to 0$. Note that since θ is continuous, it is bounded and so θ_k is also uniformly bounded i.e. $|\theta_k(t,y)| \leq C$ for all k and some $C > 0$. Consider

$$\int_0^T \int_t^{t+K} \theta(t,y) \left(-\frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right) dy dt - \int_0^T \psi_N \left(\log \int_t^{t+K} e^{\theta(t,y)} dF(y-t) \right) dt$$

with $\theta \in C(\mathcal{D}_K)$. We want to show that this can be approximated by the counterpart in $\theta_k \in S(\mathcal{D}_K)$. Note that

$$\int_0^T \int_t^{t+K} \left| \theta_k(t,y) \left(-\frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right) \right| dy dt \leq C \int_0^T \int_t^{t+K} \left| \frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right| dy dt < \infty$$

since $\bar{q}$ is absolutely continuous. By dominated convergence we have

$$\int_0^T \int_t^{t+K} \theta_k(t,y) \left(-\frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right) dy dt \to \int_0^T \int_t^{t+K} \theta(t,y) \left(-\frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right) dy dt. \tag{A.5}$$

Similarly, since, as mentioned earlier $|\theta_k(t,y)| \leq C$, by the bounded convergence theorem we have

$$\int_t^{t+K} e^{\theta_k(t,y)} dF(y-t) \to \int_t^{t+K} e^{\theta_k(t,y)} dF(y-t)$$

and so by the continuity of $\psi_N(\log(\cdot))$ we get

$$\psi_N \left(\log \int_t^{t+K} e^{\theta_k(t,y)} dF(y-t) \right) \to \psi_N \left(\log \int_t^{t+K} e^{\theta(t,y)} dF(y-t) \right)$$

for any t. Furthermore, the obvious inequality

$$e^{-C} = e^{-C} \int_0^K dF(y) \leq \int_0^K e^{\theta_k(t,y)} dF(y) \leq e^C \int_0^K dF(y) = e^C, \tag{A.6}$$

yields

$$\left| \psi_N \left(\log \int_t^{t+K} e^{\theta_k(t,y)} f(y-t) dy \right) \right| \leq \sup_{\xi \in [-C,C]} |\psi_N(\xi)|.$$

Hence yet another application of dominated convergence gives

$$\int_0^T \psi_N \left(\log \int_t^{t+K} e^{\theta_k(t,y)} dF(y-t) \right) dt \to \int_0^T \psi_N \left(\log \int_t^{t+K} e^{\theta(t,y)} dF(y-t) \right) dt. \tag{A.7}$$

Combining (A.5) and (A.7) and using the expression in (A.4), we conclude that $\sup_\kappa I(p_\kappa(\bar{q})) \geq$

$I(\bar{q})$ (note a shift of variable y in (2.3.9)). For the other direction, consider

$$\int_0^T \int_t^{t+K} \theta(t,y) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right) dy dt - \int_0^T \psi_N \left(\log \int_t^{t+K} e^{\theta(t,y)} dF(y-t) \right) dt$$

now with $\theta \in S(\mathcal{D}_K)$. Note that we can find a sequence $\theta_k \in C(\mathcal{D}_K)$ such that $\theta_k \to \theta$ pointwise almost everywhere and that θ_k is uniformly bounded; this sequence can be found, for example, by convolving θ with a sequence of mollifiers (i.e. smooth kernels with bandwidth that tends to zero as $k \to \infty$). Exactly the same argument as above would then yield $\sup_\kappa I(p_\kappa(\bar{q})) \leq I(\bar{q})$.

Now, let $\bar{q} \in C_+(\mathcal{D}_K)$ and suppose that $\bar{q}$ is not absolutely continuous. That is, it is not of bounded total variation in the sense of [22] p. 189. Then, for every $\gamma > 0$ there exists $t_1(\gamma) < \ldots < t_m(\gamma)$ and $y_0(\gamma) < \ldots < y_n(\gamma)$ such that $\sum_{i=1}^m \sum_{j=1}^n \left| \delta_{ij}^\gamma \right| \geq \gamma$, where

$$\delta_{ij}^\gamma = \bar{q}(t_i(\gamma), y_{j-1}(\gamma)) - \bar{q}(t_i(\gamma), y_j(\gamma)) - \bar{q}(t_{i-1}(\gamma), y_{j-1}(\gamma)) + \bar{q}(t_{i-1}(\gamma), y_j(\gamma)).$$

Now observe that

$$\begin{aligned} &\sup_{\kappa \in \mathcal{K}} I(p_\kappa(\bar{q})) \\ &= \sup_{\substack{\theta_{ij}: 1 \leq i \leq m, 1 \leq j \leq n \\ \kappa \in \mathcal{K}}} \sum_{i=1}^m \sum_{j=1}^n \theta_{ij} \delta_{ij}(\kappa) - \sum_{i=1}^m \int_{t_{i-1}}^{t_i} \psi_N \left(\log \sum_{j=1}^{n+1} e^{\theta_{ij}} P(y_{j-1} - u < V_1 \leq y_j - u) \right) du. \end{aligned}$$

Following [22] p. 192, we can select $\theta_{ij} = sgn\left(\delta_{ij}^\gamma\right)$ for the partition introduced earlier that defines δ_{ij}^γ, and obtain

$$\sup_{\kappa \in \mathcal{K}} I(p_\kappa(\bar{q})) \geq \sum_{i=1}^m \sum_{j=1}^n \left| \delta_{ij}^\gamma \right| - T \psi_N(1).$$

Since $\gamma > 0$ is arbitrary we conclude that

$$\sup_{\kappa \in \mathcal{K}} I(p_\kappa(\bar{q})) = \infty$$

as required. □

Proof of Lemma 2.3.3. We want to prove that for any η,

$$\lim_{\lambda \to \infty} \frac{1}{\lambda} \log P\left(\left\| \frac{\tilde{Q}_\lambda(0,0)}{\lambda} \right\|_{\mathcal{D}_K} > \eta \right) = -\infty$$

and

$$\lim_{\delta \to 0} \overline{\lim}_{\lambda \to \infty} \frac{1}{\lambda} \log P\left(w\left(\frac{\tilde{Q}_\lambda}{\lambda}, \delta \right) > \eta \right) = -\infty$$

where $w(\tilde{Q}_\lambda/\lambda, \delta)$ is the modulus of continuity of $\tilde{Q}_\lambda/\lambda$ with order δ defined by

$$w\left(\tilde{Q}_\lambda/\lambda, \delta\right) = \sup_{\substack{|t_1 - t_2| < \delta \\ |y_1 - y_2| < \delta}} |\tilde{Q}_\lambda(t_1, y_1)/\lambda - \tilde{Q}_\lambda(t_2, y_2)/\lambda|.$$

Recall that $\|\tilde{Q}_\lambda - \bar{Q}_\lambda\|_{\mathcal{D}_K} \leq 4$ a.s., and that $\bar{Q}_\lambda(t,y) = Q_\lambda(t, y-t)$ for $y > t$ and $\bar{Q}_\lambda(t,y) = \bar{Q}_\lambda(y,y) + N_\lambda(t) - N_\lambda(y)$ for $0 \leq y \leq t \leq T$. Therefore, it suffices to show that for any $\eta > 0$,

$$\lim_{\lambda \to \infty} \frac{1}{\lambda} \log P\left(\left\| \frac{Q_\lambda(0,0)}{\lambda} \right\|_{\mathcal{D}_K} > \eta \right) = -\infty, \tag{A.8}$$

also

$$\lim_{\delta \to 0} \overline{\lim}_{\lambda \to \infty} \frac{1}{\lambda} \log P\left(w\left(\frac{Q_\lambda}{\lambda}, \delta \right) > \eta \right) = -\infty, \tag{A.9}$$

and finally that

$$\lim_{\delta\to 0}\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(\sup_{0\le t_2-t_1<\delta}(N_\lambda(t_2)/\lambda - N_\lambda(t_1)/\lambda) > \eta\right) = -\infty. \tag{A.10}$$

By our assumption that the system is empty, (A.8) is obvious. Condition (A.10) will follow as a direct consequence of our analysis of (A.9). Now, to prove (A.9) consider

$$P\left(w\left(\frac{Q_\lambda}{\lambda},\delta\right) > \eta\right) \le \sum_{m=0}^{\lfloor T/\delta\rfloor}\sum_{n=0}^{\lfloor K/\delta\rfloor} P\left(\sup_{\substack{0<t_1-t_2<\delta,\ t_1\in(m\delta,(m+1)\delta]\\ |y_2-y_1|<\delta,\ y_1\in(n\delta,(n+1)\delta]}} |Q_\lambda(t_1,y_1) - Q_\lambda(t_2,y_2)| > \lambda\eta\right)$$

It is best to proceed our analysis by keeping in mind the pictorial representation that we shall

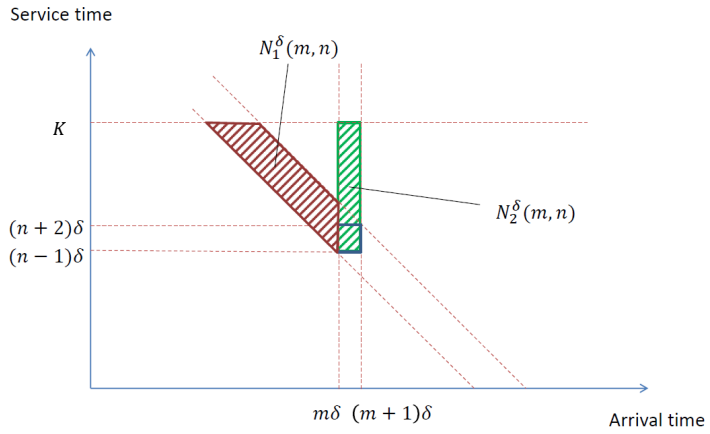


Figure A.1: Areas of $\mathcal{N}_1^\delta(m,n)$ and $\mathcal{N}_2^\delta(m,n)$

describe. One can represent the arrival and status of each customer in a two-dimensional plane, with x-axis representing the arrival time and y-axis the service time at the time of arrival. Under this representation, $Q_\lambda(t,y)$ is the number of points in the triangle formed by a vertical line and

a 45° line passing through (t,y) in its northwest direction. Consequently, we have

$$P\left(\sup_{\substack{0<t_1-t_2<\delta,\ t_1\in(m\delta,(m+1)\delta],\\ |y_2-y_1|<\delta,\ y_1\in(n\delta,(n+1)\delta]}} |Q_\lambda(t_1,y_1)-Q_\lambda(t_2,y_2)|>\lambda\eta\right)$$
$$\leq P(\mathcal{N}_1^\delta(m,n,\lambda)+\mathcal{N}_2^\delta(m,n,\lambda)>\eta\lambda)$$

where

$$\mathcal{N}_1^\delta(m,n,\lambda)=Q_\lambda(m\delta,(n-1)\delta)-Q_\lambda(m\delta,(n+2)\delta)$$

is the number of customers present at time $m\delta$ who have residual service time between $(n-1)\delta$ and $(n+2)\delta$, and

$$\mathcal{N}_2^\delta(m,n,\lambda)=\sum_{i=N_\lambda(m\delta)+1}^{N_\lambda((m+1)\delta)} I(V_i>(n-1)\delta),$$

is the number of arrivals between $m\delta$ and $(m+1)\delta$ which bring service requirements larger than $(n-1)\delta$. Figure A.2 depicts the areas under which the points are included in $\mathcal{N}_1^\delta(m,n,\lambda)$ and $\mathcal{N}_2^\delta(m,n,\lambda)$. Fixing $\delta>0$ and $\theta>0$ we take the limit as $\lambda\to\infty$ in the following display,

obtaining

$$
\begin{aligned}
&\frac{1}{\lambda}\log E e^{\theta(\mathcal{N}_1^{\delta}(m,n,\lambda)+\mathcal{N}_2^{\delta}(m,n,\lambda))} \\
=\ &\frac{1}{\lambda}\log E\exp\left\{\theta\left(\sum_{i=1}^{N_\lambda(m\delta)} I(m\delta+(n-1)\delta-A_i/\lambda<V_i\le m\delta+(n+2)\delta-A_i/\lambda)\right.\right. \\
&\left.\left.+\sum_{i=N_\lambda(m\delta)+1}^{N_\lambda((m+1)\delta)} I(V_i>(n-1)\delta)\right)\right\} \\
=\ &\frac{1}{\lambda}\log E\exp\left\{\int_0^{m\delta}\log(e^{\theta}P(m\delta+(n-1)\delta-u<V_i\le m\delta+(n+2)\delta-u)+1\right. \\
&-P(m\delta+(n-1)\delta-u<V_i\le m\delta+(n+2)\delta-u))dN_\lambda(u)+\log(e^{\theta}\bar{F}((n-1)\delta) \\
&\left.+F((n-1)\delta))[N_\lambda((m+1)\delta)-N_\lambda(m\delta)]\right\} \\
\to&\int_0^{m\delta}\psi_N(\log(e^{\theta}P(m\delta+(n-1)\delta-u<V_i\le m\delta+(n+2)\delta-u)+1 \\
&-P(m\delta+(n-1)\delta-u<V_i\le m\delta+(n+2)\delta-u)))du \\
&+\psi_N(\log(e^{\theta}\bar{F}((n-1)\delta)+F((n-1)\delta)))\delta. \qquad \text{(A.11)}
\end{aligned}
$$

Let us use $\psi_\delta(\theta;m,n)$ to denote the last expression (A.11). For fixed $\theta\ge 0$, we argue that $\psi_\delta(\theta,m,n)\to 0$ as $\delta\to 0$ uniformly over m,n. Indeed, for any m,n, the first term in (A.11)

$$
\begin{aligned}
&\int_0^{m\delta}\psi_N(\log(e^{\theta}P(m\delta+(n-1)\delta-u<V_i\le m\delta+(n+2)\delta-u)+1 \\
&-P(m\delta+(n-1)\delta-u<V_i\le m\delta+(n+2)\delta-u)))du \\
\le&\int_0^{K}\psi_N(\log(e^{\theta}P((n-1)\delta+u<V_i\le(n+2)\delta+u)+1-P((n-1)\delta+u<V_i\le(n+2)\delta+u)))du \\
\le&\ K\psi_N(\log(e^{\theta}\alpha(\delta)+1-\alpha(\delta))) \qquad \text{(A.12)}
\end{aligned}
$$

where $\alpha(\delta) := \sup_{x\in[0,K]} P(x < V_i \le x+3\delta) = o(1)$ as $\delta \to \infty$ by our assumption that the distribution of V_i is continuous and the fact that a continuous function is uniformly continuous on a compact set. On the other hand, the second term in (A.11)

$$\psi_N(\log(e^\theta \bar{F}((n-1)\delta) + F((n-1)\delta)))\delta \le \psi_N(\theta)\delta \tag{A.13}$$

for any m,n. Combining (A.12) and (A.13), we get

$$\psi_\delta(\theta,m,n) \le K\psi_N(\log(e^\theta\alpha(\delta)+1-\alpha(\delta))) + \psi_N(\theta)\delta. \tag{A.14}$$

Now fix m and n. By Chernoff's inequality we get

$$P(\mathcal{N}_1^\delta(m,n,\lambda) + \mathcal{N}_2^\delta(m,n,\lambda) > \eta\lambda) \le e^{-\eta\theta\lambda+\psi_\delta(\theta,m,n)\lambda+o(\lambda)}$$

and so

$$\begin{aligned}
&\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(\mathcal{N}_1(m,n,\lambda)+\mathcal{N}_2(m,n,\lambda) > \eta\lambda\right)\\
\le\ & -\eta\theta + \psi_\delta(\theta,m,n)\\
\le\ & -\eta\theta + K\psi_N(\log(e^\theta\alpha(\delta)+1-\alpha(\delta))) + \psi_N(\theta)\delta
\end{aligned}$$

by (A.14). Hence

$$\begin{aligned}
&\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(w\left(\frac{Q_\lambda}{\lambda},\delta\right)>\eta\right)\\
\leq\ &\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log\sum_{m=0}^{\lfloor T/\delta\rfloor}\sum_{n=0}^{\lfloor K/\delta\rfloor}P(\mathcal{N}_1^\delta(m,n,\lambda)+\mathcal{N}_2^\delta(m,n,\lambda)>\eta\lambda)\\
\leq\ &-\eta\theta+K\psi_N(\log(e^\theta\alpha(\delta)+1-\alpha(\delta)))+\psi_N(\theta)\delta
\end{aligned}$$

which gives

$$\lim_{\delta\to 0}\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(w\left(\frac{Q_\lambda}{\lambda},\delta\right)>\eta\right)\leq-\eta\theta.$$

Since θ can be arbitrarily large, we conclude (A.9). Finally, condition (A.10) follows from the analysis of $\mathcal{N}_2^\delta(m,1,\lambda)/\lambda$. □

A.3 Proofs of Technical Results in Section 2.3.4

Proof of Lemma 2.3.7. This result is a direct application of part a) in Theorem 4.2.16 in [22].

□

Proof of Lemma 2.3.8. This is similar to the case with bounded service time, but the conditions for tightness are slightly different given that our domain $\mathcal{D}$ is not compact. We must show that for any $\eta,\gamma>0$, we can choose small enough $\rho>0$, such that for $\delta<\rho$,

$$\frac{1}{\lambda}\log P(w(\tilde{Q}_\lambda/\lambda,\delta)>\eta)<-\gamma$$

when λ is large; this part is indeed basically the same as the case $\mathcal{D}_K$. In addition, however, we also must show that for all $\eta > 0$ and every $a > 0$ there exists $K > 0$ such that

$$P\left(\sup_{t\in[0,T]}\sup_{y\geq K}\tilde{Q}_\lambda(t,y)/\lambda > \eta\right) \leq \exp(-\lambda a). \tag{A.1}$$

Note that

$$\begin{aligned}
&P(w(\tilde{Q}_\lambda/\lambda,\delta) > \eta)\\
\leq &P\left(w(\tilde{Q}_\lambda/\lambda,\delta) > \eta, \|\tilde{Q}_\lambda/\lambda - \tilde{Q}_{\lambda,K}/\lambda\| \leq \frac{\eta}{2}\right) + P\left(\|\tilde{Q}_\lambda/\lambda - \tilde{Q}_{\lambda,K}/\lambda\| > \frac{\eta}{2}\lambda\right)\\
\leq &P\left(w(\tilde{Q}_{\lambda,K}/\lambda,\delta) > \frac{\eta}{2}\right) + P\left(N_\lambda^{(K)}(T) > \lambda\eta/2\right),
\end{aligned} \tag{A.2}$$

and

$$P\left(\sup_{t\in[0,T]}\sup_{y\geq K}\tilde{Q}_\lambda(t,y)/\lambda > \eta\right) \leq P\left(N_\lambda^{(K)}(T) > \lambda\eta\right).$$

By Lemma 2.3.6, for every $\gamma > 0$ we can choose K large enough such that

$$\frac{1}{\lambda}\log P\left(N_\lambda^{(K)}(T) > \lambda\eta/2\right) < -2\gamma.$$

for all λ large enough. So, condition (A.1) is enforced and the second term in the sum in (A.2) is also appropriately controlled. Now, by a similar argument as in Lemma 2.3.3 in the previous section, for a chosen K, we have, for all small enough δ,

$$\frac{1}{\lambda}\log P\left(w(\tilde{Q}_{\lambda,K}/\lambda,\delta) > \frac{\eta}{2}\right) < -2\gamma.$$

for large enough λ. In summary, we get

$$\frac{1}{\lambda}\log P(w(\tilde{Q}_\lambda/\lambda,\delta)>\eta)<-\gamma$$

for large enough λ. Therefore, exponential tightness follows. It follows immediately that a weak large deviations principle and exponential tightness imply a full large deviations principle. The goodness of the rate function then is a consequence of exponential tightness together with the weak large deviations principle; see Lemma 1.2.18, p. 8, part b) of [22]. □

Proof of Lemma 2.3.9. We start with part i), assuming that $I(\bar{q})<\infty$. Since

$$\frac{\partial^2}{\partial t\partial y}\phi_K(\bar{q})(t,y)=\frac{\partial^2}{\partial y\partial t}\phi_K(\bar{q})(t,y)=\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)1_{\{y\le K+t\}},$$

we have immediately that $I(\phi_K(\bar{q}))=I_K(\phi_K(\bar{q}))=I_K(\bar{q})$. It is obvious that $I_K(\bar{q})$ is non-decreasing in K and that $I_K(\bar{q})\le I(\bar{q})$. On the other hand, there exists $\theta^n\in C_b(\mathcal{D})$ (the space of bounded and continuous functions on $\mathcal{D}$) such that

$$I^n(\bar{q}):=\int_0^T\left[\int_t^\infty\theta^n(t,y-t)\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dy-\psi_N\left(\log\left(\int_0^\infty e^{\theta^n(t,y)}dF(y)\right)\right)\right]dt$$

converges to $I(\bar{q})$ as $n\to\infty$. Since $I(\bar{q})<\infty$, it follows easily that $\partial^2\bar{q}(\cdot)/(\partial t\partial y)$ is integrable over $\mathcal{D}$. Therefore, given that $\theta^n(\cdot)$ is bounded,

$$\int_0^T\int_K^\infty\theta^n(t,y-t)\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dydt\to 0$$

as $K \to \infty$. Besides, for given n, as ψ_N is uniformly continuous on the bounded set $[-M_n, M_n]$ where $M_n = \sup|\theta^n(t,y)| < \infty$, we have that

$$\psi_N\left(\log\left(\bar{F}(K) + \int_0^K e^{\theta^n(t,y)} dF(y)\right)\right)$$

converges to

$$\psi_N\left(\log\left(\int_0^\infty e^{\theta^n(t,y)} dF(y)\right)\right)$$

uniformly on $t \in [0,T]$. In summary, $\underline{\lim}_{K\to\infty} I_K^n(\bar{q}) = I^n(\bar{q})$ as $K \to \infty$. Therefore, there exists K_n such that $I_{K_n}^n(\bar{q}) \geq I^n(\bar{q}) - 1/n$. Recall that $I_{K_n}^n(\bar{q}) \leq I_{K_n}(\bar{q})$ and consequently we obtain

$$I^n(\bar{q}) - \frac{1}{n} \leq I_{K_n}(\bar{q}) \leq I(\bar{q}).$$

Since $I_K(\bar{q})$ increases in K, we have $I_K(q) \nearrow I(q)$ as claimed.

For part ii), when $I(\bar{q}) = \infty$, there are two cases: a) $\bar{q}$ is not absolutely continuous, and b) $\bar{q}$ is absolutely continuous. Case b) in turn is divided into two subcases: b.1) $\partial^2\bar{q}(t,y)/(\partial t \partial y)$ is not integrable over $\mathcal{D}$, and b.2) $\partial^2\bar{q}(t,y)/(\partial t \partial y)$ is integrable over $\mathcal{D}$. We shall proceed to analyze all these cases now. For Case a). We can construct a projection p_κ with $I_K(p_\kappa(\bar{q})) > M$ for K large enough, as we did in the proof of Lemma 2.3.2. For Case b), we have that $\bar{q}$ is absolutely continuous, but

$$\sup_{\theta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_0^T \left[\int_0^\infty \theta(t,y)\left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y+t)\right) dy - \psi_N\left(\log\left(\int_0^\infty e^{\theta(t,y)} dF(y)\right)\right)\right] dt = \infty,$$

so we proceed to study case b.1): Assume that $\partial^2\bar{q}(t,y)/\partial t\partial y$ is not integrable on $\mathcal{D}$. We shall assume that

$$\int_0^T \int_t^\infty \left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)^+ dydt = \infty \tag{A.3}$$

(if this integral is finite, then integral of the negative part must diverge and the analysis that follows next is identical). As in the proof of Lemma 2.3.2, given a projection κ induced by $0 \le t_1 < t_2 < ... < t_m \le T$ and $0 \le y_0 < y_1 < ... < y_{n+1}$, define

$$\begin{aligned}\delta_{ij}(\kappa) &:= \bar{q}(t_i,y_{j-1}) - \bar{q}(t_i,y_j) - \bar{q}(t_{i-1},y_{j-1}) + \bar{q}(t_{i-1},y_j) \\ &= -\int_{t_{i-1}}^{t_i}\int_{y_{j-1}}^{y_j} \frac{\partial^2}{\partial y\partial t}\bar{q}(t,y)dydt,\end{aligned} \tag{A.4}$$

as long as $y_{j-1} \ge t_{i-1}$ (otherwise, if $y_{j-1} < t_{i-1}$, $\delta_{ij}(\kappa) = 0$). Then, from (A.3) and (A.4), it follows easily that for any M, there exists a partition κ such that

$$\sum_{i,j}\delta_{ij}(\kappa) > M + T\psi_N(1).$$

Therefore, for large enough K,

$$I_K(p_\kappa(\bar{q})) \ge \sum_{i,j}\delta_{ij}(\kappa) - T\psi_N(1) > M.$$

Now, for case b.2) suppose that $\partial^2\bar{q}(t,y)/\partial t\partial y$ is integrable on $\mathcal{D}$. We can find $\theta(t,y)$ such that

$$3M < \int_0^T \int_t^\infty \theta(t,y-t)\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dy - \psi_N\left(\log\left(\int_0^\infty e^{\theta(t,y)}dF(y)\right)\right)dt < \infty.$$

Following the same line of reasoning as in the proof of part i) we can conclude that there exists $K > 0$ such that $I_K(\bar{q}) > 2M$. According to Dawson-Gartner Theorem, $I_K(\bar{q}) = \sup I_K(p_\kappa(\bar{q}))$ where the supremum is taken over all projections restricted to $\{t \in [0,T], 0 \leq y \leq t+K\}$. As a result, there exists some projection p_κ such that $I_K(p_\kappa(\bar{q})) > M$ and hence we are done.

Now we turn to part iii). So far we proved that for any $\bar{q}$ and $M > 0$, we can find a projection p_κ such that $I_K(p_\kappa(\bar{q})) > 2M$. As discussed in the proof of Lemma 2.3.2, we have

$$\sup_{\{\theta_{ij}:1\leq i\leq m, 1\leq j\leq n+1\}} \sum_{i=1}^{m}\sum_{j=1}^{n+1} \theta_{ij}\delta_{ij}(\kappa) - \sum_{i=1}^{m}\int_{t_{i-1}}^{t_i} \psi_N^{(K)}\left(\log \sum_{j=1}^{n+1} e^{\theta_{ij}} P(y_{j-1}-u < V_1^{(K)} \leq y_j - u)\right) du > 2M \tag{A.5}$$

where $\delta_{ij}(\kappa)$ is induced by the projection p_κ. From (A.5), and by the definition of supremum, there exists some θ_{ij} such that

$$\sum_{i=1}^{m}\sum_{j=1}^{n+1} \theta_{ij}\delta_{ij}(q) - \sum_{i=1}^{m}\int_{t_{i-1}}^{t_i} \psi_N^{(K)}\left(\log \sum_{j=1}^{n+1} e^{\theta_{ij}} P(y_{j-1}-u < V_1^{(K)} \leq y_j - u)\right) du > 3M/2.$$

For all $\varepsilon > 0$ and $\hat{q} \in B_\varepsilon(p)$, we have

$$|\delta_{ij}(q) - \delta_{ij}(\hat{q})| \leq 4\varepsilon.$$

Hence for $\varepsilon = M/(8\sum_{i,j}|\theta_{ij}|)$ and all $\hat{q} \in B_\varepsilon(q)$, we have

$$\begin{aligned} I(p_\kappa(\hat{q})) &\geq \sum_{i=1}^{m}\sum_{j=1}^{n+1}\theta_{ij}\delta_{ij}(\hat{q}) - \sum_{i=1}^{m}\int_{t_{i-1}}^{t_i}\psi_N^{(K)}\left(\log\sum_{j=1}^{n+1}e^{\theta_{ij}}P(y_{j-1}-u<V_1^{(K)}\leq y_j-u)\right)du \\ &\geq \sum_{i=1}^{m}\sum_{j=1}^{n+1}\theta_{ij}\delta_{ij}(q) - \sum_{i=1}^{m}\int_{t_{i-1}}^{t_i}\psi_N^{(K)}\left(\log\sum_{j=1}^{n+1}e^{\theta_{ij}}P(y_{j-1}-u<V_1^{(K)}\leq y_j-u)\right)du - 4\varepsilon\sum_{i=1}^{m}\sum_{j=1}^{n+1}|\theta_{ij}| \\ &> M. \end{aligned}$$

Thus we conclude the result. □

Proof of Lemma 2.3.6. Let $N_\lambda^{(K)}(T)$ be the total number of arrivals from time 0 up to T with service time longer than K, under the λ-scaled system. Then following [28] (or as in the proof of Lemma 2.3.3) we have that

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\log Ee^{\theta N_\lambda^{(K)}(T)} = T\psi_N(\log(e^\theta\bar{F}(K)+F(K))).$$

Chernoff's bound yields

$$P(N_\lambda^{(K)}(T)>\lambda\varepsilon) \leq \exp\{-\theta\lambda\varepsilon+\lambda T\psi_N(\log(e^\theta\bar{F}(K)+F(K)))+o(\lambda)\}.$$

Hence

$$\begin{aligned} &\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P_\lambda(N_\lambda^{(K)}(T)>\lambda\varepsilon) \\ \leq &-\theta\varepsilon+T\psi_N(\log(e^\theta\bar{F}(K)+F(K))) \end{aligned}$$

Letting $K \to \infty$ gives

$$\lim_{K\to\infty} \overline{\lim}_{\lambda\to\infty} \frac{1}{\lambda} \log P_\lambda(N_\lambda^{(K)}(T) > \lambda\varepsilon) \leq -\theta\varepsilon.$$

Since θ can be arbitrarily large, the result follows. $\square$

Appendix B

Appendix for Chapter 3

B.1 The Proof of Theorem 3.3.1

Proof of Theorem 3.3.1. In this proof, we will follow the notations as used in [45]. Define $X_n(\cdot) = (\bar{a}_n(\cdot), \bar{b}_n(\cdot)) \in \mathbb{Z}^2$. Define $Y_n(\cdot) = (Y_n^i(\cdot) : i \in \mathbb{Z}) \in \mathbb{Z}^\infty$, where $Y_n^i(t)$ equals the number of limit sell orders (or minus the number of limit buy orders) on price tick i in the n-th LOB system at time t. We denote by E_1 and E_2 respectively the space of $\mathbb{Z}^2$ and $\mathbb{Z}^\infty$ that are endowed with the discrete topology. Then, $\{(X_n(\cdot), Y_n(\cdot))\}$ is a sequence of stochastic processes living in the product space $E_1 \times E_2$.

Note that for each n, $(X_n(\cdot), Y_n(\cdot))$ is a continuous time Markov Chain with countable number of states. Let $\{\mathfrak{F}_t^n\}$ be the natural filtration associated with $(X_n(\cdot), Y_n(\cdot))$. Then, one can check that for any $f \in C(E_1)$,

$$f(X_n(t)) - \int_0^t 2\mu(f(a(Y_n(s)), b(Y_n(s))) - f(X_n(s)))ds \triangleq f(X_n(t)) - \int_0^t \mathcal{A}f(X_n(s), Y_n(s))ds$$

is a martingale with respect to $\{\mathfrak{F}_t^n\}$. Here the functional $a:\mathbb{R}^\infty\to\mathbb{R}$ is just the ask price of the LOB at time s, more precisely,

$$a(Y_n(s))=\min\{i:Y_n^i(s)>0\}.$$

The functional $b(\cdot)$ can be defined in a similar way. One can also check that for any $g\in C(E_2)$,

$$\begin{aligned}
&g(Y_n(t))-\int_0^t\sum_i\xi_n\Big[\lambda p(i;X_n(s))\Big(g(Y_n(s)+e_{\bar a_n(s)+i})+g(Y_n(s)+e_{\bar b_n(s)-i})-2g(Y_n(s))\Big)\\
&+\alpha(i;X_n(s))\Big(Y_n^{\bar a_n(s)+i}(s)(g(Y_n(s)-e_{\bar a_n(s)+i})-g(Y_n(s)))+Y_n^{\bar b_n(s)-i}(s)(g(Y_n(s)-e_{\bar b_n(s)-i})-g(Y_n(s)))\Big)\Big]ds\\
&\triangleq g(Y_n(t))-\int_0^t\xi_n\mathcal{B}g(X_n(s),Y_n(s))ds
\end{aligned}$$

is also a martingale. Due to the regularity condition imposed (3.3.2), $\{X_n(\cdot)\}$ is tight. Therefore, each subsequence of $\{X_n(\cdot)\}$ admit a sub-subsequence that converges weakly to some sub-limit process $X(\cdot)$. Then according to Theorem 2.1 and the subsequent Example 2.4 in [45], each sub-limit process $X(\cdot)=(\hat a(\cdot),\hat b(\cdot))$ is a solution to the martingale problem

$$f(X(t))-\int_0^t\int_{E_2}\mathcal{A}f(X(s),y)\pi_{X(s)}(dy)ds, \tag{B.1}$$

in the sense that the stochastic process defined by (B.1) is a martingale. Moreover, in the expression (B.1), $\pi_{X(s)}(\cdot)$ is the unique stationary distribution of a stochastic process $Y\in E_2$

which satisfies the martingale problem

$$g(Y(t)) - \int_0^t \mathcal{B}g(x,Y(u))du.$$

In our case, $\pi_x(\cdot)$ is simply the stationary distribution of the LOB system under the parameters $\{p(i;x),\alpha(i;x)\}$.

Now we compute that in (B.1),

$$\begin{aligned}
&\int_{E_2} \mathcal{A}f(X(s),y)\pi_{X(s)}(dy) \\
=&\sum_{i,j}\mu(f(\hat{a}(\cdot)+i,\hat{b}(\cdot)+j)-f(\hat{a}(\cdot),\hat{b}(\cdot)))\pi_{X(s)}(\{a(y)-\hat{a}_n(s)=i,b(y)-\hat{b}_n(s)=-j\}) \\
=&\sum_{i,j}\mu(f(\hat{a}(\cdot)+i,\hat{b}(\cdot)+j)-f(\hat{a}(\cdot),\hat{b}(\cdot)))[\theta(i;\hat{a}(t),\hat{b}(t))-\theta(i+1;\hat{a}(s),\hat{b}(s))] \\
&\times[\theta(j;\hat{a}(s),\hat{b}(s))-\theta(j+1;\hat{a}(s),\hat{b}(s))].
\end{aligned}$$

One can check that the martingale problem (B.1) has a unique solution $X(\cdot)$, see for instance Chapter 4.4 in [25]. In particular, $X(\cdot)=(\hat{a}(\cdot),\hat{b}(\cdot))$ is equivalent in distribution to a jump process with jump intensity 2μ and jump size distribution

$$\begin{aligned}
&P\left(\hat{a}(t)-\hat{a}(t-)=i,\hat{b}(t)-\hat{b}(t-)=j|\hat{a}(t-),\hat{b}(t-)\right) \\
=\ &[\theta(i;\hat{a}(t-),\hat{b}(t-))-\theta(i+1;\hat{a}(t-),\hat{b}(t-))]\times[\theta(j;\hat{a}(t-),\hat{b}(t-))-\theta(j+1;\hat{a}(t-),\hat{b}(t-))].
\end{aligned}$$

Since $\{X_n(\cdot)\}$ is tight and each of its convergent subsequence admits the same limit $X(\cdot)$, we can conclude that $\{X_n(\cdot)\}$ weakly converges to $X(\cdot)$. □

B.2 The Proof of Theorem 3.4.3

Now let us describe the roadmap for the proof of Theorem 3.4.3. We first construct some auxiliary process $(\tilde{S}^n(\cdot),\tilde{M}^n(\cdot))$ living in the same probability space as the underlying process $(\bar{s}^n(\cdot),\bar{m}^n(\cdot))$. The auxiliary process is a convenient Markov process whose generator can be analyzed to conclude weak convergence to the postulated limiting jump diffusion (3.4.3). The auxiliary process has the same dynamics as the target process except when it is on the boundary-layer set $[0,2/\sqrt{n}]\times\mathbb{R}$. We also show the time spent by the two processes on the boundary-layer is small and as a result their difference caused by their different dynamics on the boundary is also small. Actually, such difference is negligible as $n\to\infty$ and therefore the target process converges to the same limit process.

First, we define the auxiliary process coupled with the target process in a path by path fashion. Recall that by Assumption 3.4.1 and (3.4.2), we can write

$$\begin{cases} \bar{s}^n(t_{k+1}) & = \bar{s}^n(t_k)+\Delta_a^n(\bar{s}^n(t_k))\vee([-\bar{s}^n(t_k)/2\delta]\delta)+\Delta_b^n(\bar{s}^n(t_k))\vee([-\bar{s}^n(t_k)/2\delta]\delta), \\ \bar{m}^n(t_{k+1}) & = \bar{m}^n(t_k)+\Delta_a^n((\bar{s}^n(t_k)))\vee([-\bar{s}^n(t_k)/2\delta]\delta)-\Delta_b^n(\bar{s}^n(t_k))\vee([-\bar{s}^n(t_k)/2\delta]\delta). \end{cases}$$

Now we define the auxiliary process $(\tilde{S}^n(\cdot),\tilde{M}^n(\cdot))$ coupled with $(\bar{s}^n(\cdot),\bar{m}^n(\cdot)))$ as

$$\begin{cases} \tilde{S}^n(t_{k+1}) & = \tilde{S}^n(t_k)+(\Delta_a^n(\tilde{S}^n(t_k))+\Delta_b^n(\tilde{S}^n(t_k)))\vee(-\tilde{S}^n(t_k)), \\ \tilde{M}^n(t_{k+1}) & = \tilde{M}^n(t_k)+\Delta_a^n((\tilde{S}^n(t_k)))-\Delta_b^n(\tilde{S}^n(t_k)), \end{cases} \tag{B.1}$$

with the initial condition $\tilde{S}^n(0)=\bar{s}^n(0)$ and $\tilde{M}^n(0)=\bar{m}^n(0)$.

Then Theorem 3.4.3 is an immediate corollary of the following two propositions.

Proposition B.2.1. *The auxiliary process* $(\tilde{S}^n(\cdot), \tilde{M}^n(\cdot))$ *converges weakly to the limit process given by (3.4.3).*

Proposition B.2.2. *The difference process* $\left(\bar{s}^n(\cdot) - \tilde{S}^n(\cdot), \bar{m}^n(\cdot) - \tilde{M}^n(\cdot)\right)$ *converges weakly to* $(0,0)$ *on* $D_{\mathbb{R}^2}[0,t]$ *for any* $t < \infty$.

Proof of Proposition B.2.2. For simplicity, we assume that $\xi = 1$, otherwise we can divide $\tilde{S}$, $\tilde{M}$ and $\bar{s}$, $\bar{m}$ by the constant ξ. Assume also that $V_k^a \leq c$ for some $c \geq 1$. The general case can be dealt with using truncation because there are only a Poisson number of jumps that arise in $O(1)$ time.

Now, let us first give a bound for the difference $\bar{s}^n(\cdot) - \tilde{S}^n(\cdot)$. For fixed n, we define $N(t) = \sum_{\{j:t_j \leq t\}} (I_j^a + I_j^b)$, intuitively $N(t)$ corresponds to the number of jumps of the limiting process from time 0 to time t. Now we prove by induction that

$$0 \leq \bar{s}^n(t_k) - \tilde{S}^n(t_k) \leq \left((1+c)^{N(t_k)} - 1\right) \cdot \frac{2}{\sqrt{n}}. \tag{B.2}$$

At $t_0 = 0$, we have $\tilde{S}^n(t_0) = \bar{s}(t_0)$. Now suppose the relation (B.2) holds at time t_{k-1}, there are two cases at time t_k, case 1): $N(t_k) = N(t_{k-1})$, and case 2): $N(t_k) > N(t_{k-1})$.

First let us consider the case when $N(t_k) = N(t_{k-1})$. In this case, we know that $\Delta_a^n(\bar{s}^n(t_{k-1})) = \Delta_a^n(\tilde{S}(t_{k-1})) := \Delta_a^n$ is independent of $\bar{s}^n(t_{k-1})$ and $\tilde{S}^n(t_{k-1})$. Also, keep in mind that $|\Delta_a^n(t_k)| \leq$

$1/\sqrt{n}$. Now we can write the increment of the difference process

$$\begin{aligned}
&(\vec{s}^n(t_k) - \tilde{S}^n(t_k)) - (\vec{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1})) \\
=\ & \Delta_a^n \vee ([-\vec{s}^n(t_k)/(2\delta)]\delta) + \Delta_b^n \vee ([-\vec{s}^n(t_k)/(2\delta)]\delta) - (\Delta_a^n + \Delta_b^n) \vee (-\tilde{S}^n(t_{k-1})). \qquad \text{(B.3)}
\end{aligned}$$

Therefore, if $\vec{s}^n(t_{k-1}) \geq \tilde{S}^n(t_{k-1}) \geq 2/\sqrt{n}$, we have

$$(\vec{s}^n(t_k) - \tilde{S}^n(t_k)) - (\vec{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1})) = \Delta_a^n + \Delta_b^n - (\Delta_a^n + \Delta_b^n) = 0$$

and as a result $\vec{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}) = \vec{s}^n(t_k) - \tilde{S}^n(t_k)$. If $\vec{s}^n(t_{k-1}) \geq 2/\sqrt{n} \geq \tilde{S}^n(t_{k-1}) \geq 0$, We have

$$\begin{aligned}
&(\vec{s}^n(t_k) - \tilde{S}^n(t_k)) - (\vec{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1})) \\
=\ & \Delta_a^n + \Delta_b^n - (\Delta_a^n + \Delta_b^n) \vee (-\tilde{S}^n(t_{k-1})) = -(\tilde{S}^n(t_{k-1}) + \Delta_a^n + \Delta_b^n)^- \leq 0.
\end{aligned}$$

Therefore,

$$\vec{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}) \geq \vec{s}^n(t_k) - \tilde{S}^n(t_k) \geq \frac{2}{\sqrt{n}} - (\tilde{S}^n(t_{k-1}) + \Delta_a^n + \Delta_b^n) \geq 0.$$

Otherwise, we have $0 \leq \tilde{S}^n(t_{k-1}) \leq \vec{s}^n(t_{k-1}) \leq 2/\sqrt{n}$. In this case, one can check that for any fixed $\tilde{S}^n(t_k) = \tilde{s}$ and $\vec{s}^n(t_k) = s$, the increment of the difference process (B.3) reaches its maximum at $\Delta_a^n = -1//\sqrt{n}$ and $\Delta_b^n = -\tilde{s} + /\sqrt{n}$ and its minimum at $\Delta_a^n = \Delta_b^n = -[s/2\delta]\delta$.

Hence,

$$\tilde{s} - s \leq \left(\vec{s}^n(t_k) - \tilde{S}^n(t_k)\right) - \left(\vec{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1})\right) \leq 0 \vee \left(\frac{1}{\sqrt{n}} - \frac{s}{2} - \tilde{s}\right).$$

Plugging in $\tilde{S}^n(t_k) = \tilde{s}$ and $\vec{s}^n(t_k) = s$, we have

$$0 \leq \vec{s}^n(t_k) - \tilde{S}^n(t_k) \leq (s - \tilde{s}) \vee \left(\frac{1}{\sqrt{n}} + \frac{s}{2}\right) \leq (\vec{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1})) \vee \frac{2}{\sqrt{n}}.$$

The last inequality holds as $s = \vec{s}^n(t_{k-1}) \leq 2/\sqrt{n}$. In summary, we have proved that when $N(t_k) = N(t_{k-1})$, if the relation (B.2) holds at time t_{k-1}, so does it at time t_k.

Now if $N(t_k) \geq N(t_{k-1}) + 1$, intuitively, at least one jump occurs in Δ_a^n and Δ_b^n. If $I_k^a = 1$ we have

$$\Delta_a^n(\tilde{S}^n(t_{k-1})) = I_k^a V_k^a [\tilde{S}^n(t_{k-1})/(2\delta)]\delta, \text{ and } \Delta_a^n(\vec{s}^n(t_{k-1})) = I_k^a V_k^a [\vec{s}^n(t_{k-1})/(2\delta)]\delta.$$

If in addition $I_k^b = 1$, then

$$\Delta_b^n(\tilde{S}^n(t_{k-1})) = I_k^b V_k^b [\tilde{S}^n(t_{k-1})/(2\delta)]\delta, \text{ and } \Delta_b^n(\vec{s}^n(t_{k-1})) = I_k^b V_k^b [\vec{s}^n(t_{k-1})/(2\delta)]\delta,$$

and therefore,

$$\vec{s}^n(t_k) - \tilde{S}^n(t_k) \leq (\vec{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}))(I_k^a V_k^a + I_k^b V_k^b + 1)/2.$$

As

$$0 \le V_k^a + V_k^b + 1 \le 2c+1 \text{ and } 0 \le \vec{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}) \le ((c+1)^{N(t_{k-1})} - 1) \cdot \frac{2}{\sqrt{n}},$$

by the induction assumption we have

$$0 \le \vec{s}^n(t_k) - \tilde{S}^n(t_k) \le ((c+1)^{N(t_{k-1})} - 1)(2c+1) \cdot \frac{2}{\sqrt{n}} \le ((c+1)^{N(t_k)} - 1) \cdot \frac{2}{\sqrt{n}}.$$

If $I_k^b = 0$, then following a similar argument as in the case when $N(t_k) = N(t_{k-1})$, we have

$$\begin{aligned} \vec{s}^n(t_k) - \tilde{S}^n(t_k) &\le (\vec{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}))(V_k^a + 1) + (\vec{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1})) \vee \frac{2}{\sqrt{n}} \\ &= ((c+1)^{N(t_k)} - 1) \cdot \frac{2}{\sqrt{n}} \quad \text{as } c \ge 1. \end{aligned}$$

In summary, we have proved the relation (B.2) of $\tilde{S}^n(\cdot)$ and $\vec{s}^n(\cdot)$ by induction.

Now let us turn to the difference $\bar{m}^n(\cdot) - \tilde{M}^n(\cdot)$. Actually, $\bar{m}^n(t) - \tilde{M}^n(t)$ can be decomposed into two parts,

$$\begin{aligned} &\bar{m}^n(t) - \tilde{M}^n(t) \\ \le &\sum_{0 \le k \le [nt]: N(t_{k+1}) = N(t_k)} [\Delta_a^n(t_k) \vee ([\vec{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta_b^n(t_k) \vee ([\vec{s}^n(t_{k-1})/(2\delta)]\delta) - (\Delta_a^n(t_k) - \Delta_b^n(t_k))] \\ &+ \sum_{i=1}^{N(t)} ([\vec{s}^n(t_{k-1})/(2\delta)]\delta - [\tilde{S}^n(t_{k-1})/(2\delta)]\delta)(I_{k_i}^a V_{k_i}^a + I_{k_i}^b V_{k_i}^b), \end{aligned}$$

where $\{t_{k_i}\}$ are the jump times. We denote the two summation parts as

$$\bar{m}^n(t) - \tilde{M}^n(t) = \varepsilon_0^n(t) + \varepsilon_1^n(t).$$

Intuitively, $\varepsilon_0^n(t)$ is the error corresponding to the diffusion part when $I_k^a = I_k^b = 0$ and $\varepsilon_1^n(t)$ is the error corresponding to the jumps. In the summation part $\varepsilon_0^n(t)$, we write $\Delta_a^n(\vec{s}^n(t_k)) = \Delta_a^n(\tilde{S}^n(t_k)) = \Delta_a^n(t_k)$, because they are independent of $\vec{s}^n(t_k)$ and $\tilde{S}^n(t_k)$ when when $I_k^a = I_k^b = 0$.

Following a same induction argument as for $\vec{s}^n - \tilde{S}^n$, we can show that the error caused by jumps $\varepsilon_1^n(t)$ satisfies that

$$\varepsilon_1^n(t) \leq ((1+2c)^{N(t)} - 1) \cdot \frac{2}{\sqrt{n}}.$$

On the other hand, note that $\varepsilon_0^n(t)$ equals

$$\sum_{0 \leq k \leq [nt]: N(t_{k+1}) = N(t_k)} [\Delta_a^n(t_k) \vee ([\vec{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta_b^n(t_k) \vee ([\vec{s}^n(t_{k-1})/(2\delta)]\delta) - (\Delta_a^n(t_k) - \Delta_b^n(t_k))]$$

$$= \sum_{0 \leq k \leq [nt]: N(t_{k+1}) = N(t_k)} [(\Delta_a^n(t_k) \vee ([\vec{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta_a^n(t_k)) - (\Delta_b^n(t_k) \vee ([\vec{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta_b^n(t_k))].$$

Since $\Delta_a^n(t_k)$ and $\Delta_b^n(t_k)$ are independent and identically distributed, we have that for any $k \geq 1$

$$\begin{aligned} E[\varepsilon_0^n(t_k) - \varepsilon_0^n(t_{k-1}) | \mathcal{F}_{t_k}^n] = \; & P(N(t_k) = N(t_{k-1})) \cdot (E[\Delta_a^n(t_k) \vee ([\vec{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta_a^n(t_k)] \\ & - E[\Delta_b^n(t_k) \vee ([\vec{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta_b^n(t_k)]) = 0, \end{aligned}$$

where $\mathcal{F}_{t_k}^n$ is the σ-field generated by $\{\Delta_a^n(t_i), \Delta_b^n(t_i), \tilde{S}^n(t_i)\}_{i=1}^k$. Therefore, the process $\varepsilon_0^n(\cdot)$ is a

martingale under the filtration $\mathcal{F}^n$. Besides, as $|\Delta_a^n(t_k)| \le 1/\sqrt{n}$ when $N(t_k) = N(t_{k-1})$, we have

$$|\varepsilon_0^n(t_k) - \varepsilon_0^n(t_{k-1})| \le \frac{2}{\sqrt{n}}\, I(\bar{s}^n(t_{k-1}) < \frac{2}{\sqrt{n}}).$$

The quadratic variation

$$[\varepsilon_0^n](t) \ \le \frac{4}{n}\sum_{i=0}^{[nt]} I(\bar{s}^n(t_i) < \frac{2}{\sqrt{n}}).$$

Recall that we have proved $\bar{s}^n(\cdot) \ge \tilde{S}^n(\cdot)$,

$$[\varepsilon_0^n](t) \ \le \frac{4}{n}\sum_{i=0}^{[nt]} I(\tilde{S}^n(t_i) < \frac{2}{\sqrt{n}}).$$

Since $2/\sqrt{n} \to 0$, for any $\zeta > 0$ we have

$$\lim_{n\to\infty}[\varepsilon_0^n](t) \ \le \lim_{n\to\infty} 4\int_0^t I(\tilde{S}^n(u) < \zeta)du \le \lim_{n\to\infty} 4\int_0^t f^{\zeta}(\tilde{S}^n(u))du,$$

where $f^{\zeta}(\cdot)$ is a smooth function on $\mathbb{R}^+$ and satisfies $f(x) = 1$ for all $0 \le x \le \zeta$, $0 \le f(x) \le 1$ for $\zeta \le x \le 2\zeta$ and $f(x) = 0$ for $x > 2\zeta$. (Such function can be constructed, for instance, by convolution.) Since $f^{\zeta}(\cdot)$ is bounded and $\tilde{S}^n(\cdot)$ converges weakly to the limit process (3.4.3), we have

$$\lim_{n\to\infty} E[[\varepsilon_0^n](t)] \ \le \lim_{n\to\infty} 4E[\int_0^t f^{\zeta}(\tilde{S}^n(u))du] = 4E[\int_0^t f^{\zeta}(\bar{s}(u))du] \le 4E[\int_0^t I(\bar{s}(u) \le 2\zeta)],$$

As the limit process $\bar{s}(\cdot)$ has the same dynamics as a reflected Brownian motion except when at the finite time of jumps on $[0,t]$, we have $E[\int_0^t I(\bar{s}(u) \le 2\zeta)] \to 0$ as $\zeta \to 0$. Since ζ can be

arbitrarily small, we conclude that the expected quadratic variation $E[[\varepsilon_0^n](t)] \to 0$ as $n \to 0$ for any $t < \infty$. By Doob's Inequality, we have that for all fixed $\zeta > 0$,

$$P(\max_{0\le u\le t} |\varepsilon_0^n(u)| > \zeta) \le \frac{E[[\varepsilon_0^n](t)]}{\zeta^2} \to 0.$$

Therefore, $\varepsilon_0^n(\cdot)$ converges weakly to $x(\cdot) \equiv 0$ in space $D[0,t]$ for all $t < \infty$.

In the end, it is given in Assumption 3.4.2 that $nP(I_k^a V_k^a + I_k^b V_k^b \neq 0) = 2\gamma + o(1)$, so the counting process $N(\cdot)$ converges to a Poisson process with rate $2\mu\gamma$. Therefore, for any $t < \infty$

$$E[\max_{0\le u\le t} |((2c+1)^{N(u)} - 1)\frac{2}{\sqrt{n}}|] = E[((2c+1)^{N(t)} - 1)\frac{2}{\sqrt{n}}] = O(\frac{1}{\sqrt{n}}).$$

As a result, the process $((2c+1)^{N(\cdot)} - 1)\frac{2}{\sqrt{n}}$ converges weakly to $x(\cdot) \equiv 0$ in space $D[0,t]$. Recall that we have proved that $((2c+1)^{N(\cdot)} - 1)\frac{2}{\sqrt{n}}$ is an upper bound of $|\bar{s}^n(\cdot) - \tilde{S}^n(\cdot)|$ and the 'jump part' of $|\bar{m}^n(\cdot) - \tilde{M}^n(\cdot)|$. As a consequence, we can conclude that the difference process $(\bar{s}^n(\cdot) - \tilde{S}^n(\cdot), \bar{m}^n(\cdot) - \tilde{M}^n(\cdot))$ converges weakly to $(0,0)$ on any compact interval $[0,t]$. □

Proof of Proposition B.2.1. Define $N_a(t) = \sum_{\{j:t_j\le t\}} I_j^a$, $N_b(t) = \sum_{\{j:t_j\le t\}} I_j^b$, and note that $N_a(\cdot)$, $N_b(\cdot)$ are two independent Poisson processes with rate $\gamma\mu$ each. Next, define $\tilde{S}^n(0) = \bar{s}^n(0) \ge 0$, and $\tilde{M}^n(0) = \bar{m}^n(0)$, and set

$$S_a^n(t) = \sum_{\{j:t_j\le t\}} (-1)^{R_j^a}[U_j^a/(\sqrt{n}\delta_n)]\delta_n, \;\; S_b^n(t) = \sum_{\{j:t_j\le t\}} (-1)^{R_k^b}[U_k^b/(\sqrt{n}\delta_n)]\delta_n).$$

We will also define

$$\begin{aligned} S^n(t) &= S_a^n(t) + S_b^n(t) + \bar{s}^n(0), \\ M^n(t) &= S_a^n(t) - S_b^n(t) + \bar{m}^n(0), \end{aligned}$$

(so by convention we set $t_0 = 0$ and $S^n(0) = \bar{s}^n(0)$). Also, we define

$$R_1^n(t) = S^n(t) - \min(S^n(u) : u \leq t, 0).$$

Let $A = \inf\{t \geq 0 : N_a(t) \geq 1\}$ and $B = \inf\{t \geq 0 : N_b(t) \geq 1\}$ be the first arrival times of $N_a(\cdot)$ and $N_b(\cdot)$, respectively. Since

$$\tilde{S}^n(t_k) = (\Delta_a^n(\tilde{S}^n(t_k)) + \Delta_b^n(\tilde{S}^n(t_k)) + \tilde{S}^n(t_k))^+,$$

we have that on $\min(A,B) > t$

$$\tilde{S}^n(t) = R^n(t).$$

The strategy proceeds as follows.

Step 1): Show that if $(\bar{s}^n(0), \bar{m}^n(0)) \Rightarrow (\bar{s}(0), \bar{m}(0))$, the processes $(S_a^n(t), S_b^n(t) : t \geq 0)$ converges weakly in $D[0,\infty)$ to the process $(X(t) : t \geq 0)$ defined via

$$\begin{aligned} X_1(t) &= \bar{s}(0) - \eta t + W_a(t) + W_b(t), \\ X_2(t) &= \bar{m}(0) + W_a(t) - W_b(t). \end{aligned}$$

Step 2): Once Step 1) has been executed we can directly apply the continuous mapping principle to conclude joint weak convergence on $[0, \min(A,B))$ of the processes

$$\begin{aligned} R^n(\cdot) &\Rightarrow R(\cdot) := X_1(\cdot) - \min\left(X_1(u) : 0 \le u \le \cdot\right), \\ M^n(\cdot) &\Rightarrow X_2(\cdot). \end{aligned}$$

Step 3): By invoking the Skorokhod embedding theorem, we can assume that the joint weak convergence in Step 2) occurs almost surely. We can add the jump right at time $\min(A,B)$ without changing the distribution of $X_1(t)$ and $\tilde{S}^n(t)$ for $t < \min(A,B)$. More precisely, define

$$D = I(A < B) R(A) V^a/2 + I(B \le A) R(B) V^b/2,$$

where V^b and V^a are i.i.d. copies of V_k^b and V_k^a respectively, and we also define

$$D^n = I(A < B) [R^n(A) V^a/(2\delta)]\delta + I(B \le A) [R^n(B) V^b/(2\delta)]\delta.$$

Then put on $t \in [0, \min(A,B)]$

$$\begin{aligned}
\tilde{S}^n(t) &= R^n(t) I(t < \min(A,B)) + I(t = \min(A,B)) (R^n(\min(A,B)) + D^n), \\
\bar{s}(t) &= R(t) I(t < \min(A,B)) + I(t = \min(A,B)) (R(\min(A,B)) + D), \\
\tilde{M}^n(t) &= M^n(t) I(t < \min(A,B)) \\
&\quad + I(t = \min(A,B)) I(A < B) [R^n(A) V^a/(2\delta)]\delta \\
&\quad - I(t = \min(A,B)) I(B \leq A) [R^n(B) V^b/(2\delta)]\delta, \\
\bar{m}(t) &= X_2(t) I(t < \min(A,B)) \\
&\quad + I(t = \min(A,B)) I(A < B) R(A) V^a/2 \\
&\quad - I(t = \min(A,B)) I(B \leq A) R(B) V^b/2\delta.
\end{aligned}$$

So, assuming Step 2) and using Skorokhod embedding we then conclude that

$$\sup_{0 \leq t \leq \min(A,B)} \left| \tilde{S}^n(t) - \bar{s}(t) \right| + \sup_{0 \leq t \leq \min(A,B)} \left| \tilde{M}^n(t) - \bar{m}(t) \right| \to 0$$

almost surely.

Step 4): Finally, note that the convergence extends throughout the interval $[0,t]$ by repeatedly applying Steps 1) to 3) given that there are only finitely many jumps in $[0,t]$. Clearly then this procedure completes the construction to the solution of the SDE (3.4.3).

So, we see that everything rests on the execution of Step 1), and for this we invoke the

martingale central limit theorem (see [25], Theorem 7.1.4). Define

$$Z_k^a(n) = (-1)^{R_k^a}[U_k^a/(\sqrt{n}\delta_n)]\delta_n, \quad Z_k^b(n) = (-1)^{R_k^b}[U_k^b/(\sqrt{n}\delta_n)]\delta_n.$$

We have that

$$EZ_k^a(n) = E([U_k^a/(\sqrt{n}\delta_n)]\delta_n)(-\beta/(\sqrt{n})) = \beta EZ_k^b(n) = \begin{cases} \frac{-\beta E(U_1^a)}{n} + o(1/n) & \text{if } \delta_n = o(1/\sqrt{n}), \\ \frac{-\beta E([U_1^a])}{n} + o(1/n) & \text{if } \delta_n = 1/\sqrt{n}, \end{cases}$$

and

$$Var(Z_k^a(n)) = E([U_k^a/(\sqrt{n}\delta_n)]^2\delta_n^2) - (EZ_k^a(n))^2 = E([U_k^a/(\sqrt{n}\delta_n)]^2\delta_n^2) - O(1/n).$$

We write

$$S_a^n(t) = M_a^n(t) + \mu t EZ_k^b(n) = M_a^n(t) - \eta t + o(1),$$

where $M_a^n(t)$ is a martingale, and we have that

$$\sup_{0\le t\le T} |M_n^a(t) - M_n^a(t_-)| \le (\theta/\sqrt{n} + \delta_n),$$

therefore

$$E \sup_{0\le t\le T} |M_n^a(t) - M_n^a(t_-)| + E \sup_{0\le t\le T} |M_n^a(t) - M_n^a(t_-)|^2 = o(1)$$

as $n \to \infty$, which verifies conditions a) and b.1) from [25],Theorem 7.1.4. Moreover, we have

that

$$[M_a^n, M_a^n](t) = \sum_{j=1}^{N(nt)} [U_j^a/(\sqrt{n}\delta_n)]^2 \delta_n^2 \to t\sigma = \begin{cases} t\mu E(\left[U_j^a\right]^2) & \text{if } \delta_n = 1/\sqrt{n}, \\ t\mu E(\left(U_j^a\right)^2) & \text{if } \delta_n = o(1/\sqrt{n}). \end{cases}$$

Furthermore, we have that

$$E \max_{t \le T} |[M_a^n, M_a^n](t) - [M_a^n, M_a^n](t_-)| \le (\theta/\sqrt{n} + \delta_n)^2 = o(1),$$

which corresponds to condition b.2) in [25],Theorem 7.1.4. Hence, we conclude that

$$M_a^n(\cdot) \Rightarrow W_a(\cdot)$$

under the uniform topology on compact sets. A completely analogous strategy is applicable to conclude $M_a^n(\cdot) \Rightarrow W_b(\cdot)$. The convergence holds jointly due to independence and therefore we obtain the conclusion required in Step 1). As indicated earlier, Steps 2) to 4) now follow directly. □

Appendix C

Appendix for Chapter 4

C.1 Proof of Theorem 4.3.6

First, we rewrite Algorithm 4.2 as following:

Then $\mathcal{N}(d) = O(d \cdot N)$. Again we use $N(d)$ instead of N to emphasize the dependence on the number of dimensions d. The following result shows that our algorithm has polynomial complexity with respect to d:

Theorem C.1.1. *Under assumptions C*1) *to C*3)*,*

$$E[N(d)] = O(d^{\gamma}) \text{ as } d \to \infty,$$

for some γ *depending on* δ *and* H*.*

Denote the number of Bernoulli's generated in Step 9 by N_b and the number of random variables generated before entering Step 6 in a single iteration by N_a. By Wald's identity, we

Algorithm C.1 Simulating the Coalescence Time

Input:

constant m as defined in (4.3.11)

Output:

$(\mathbf{Z}(t) : 0 \leq t \leq \tau)$ and the coalescence time τ.

Set $\tau = 0$, $\mathbf{Z}(0) = \mathbf{0}$, $I = 1$ and $N = 1$.
while $I == 1$ **do**
  Generate an inter-arrival time U distributed Exp(λ) and sample $\mathbf{W} = (W_1, ..., W_d)$ independent of U.
  Let $\mathbf{Z}(\tau + t) = \mathbf{Z}(\tau) - t\mu$ for $0 \leq t < U$ and $\mathbf{Z}(\tau + U) = \mathbf{Z}(\tau) + \mathbf{W} - U\mu$.
  $\tau \longleftarrow \tau + U$ and $N = N + 1$.
  if $W_i - U\mu_i < -m$ **then**
    Otherwise, simulate a random walk $\{\mathbf{C}(n)\}$ such that $\mathbf{C}(0) = 0$ and $\mathbf{C}(n) = \mathbf{C}(n-1) + \mathbf{W}'(n) - U'(n)\mu$, where $\mathbf{W}'(n) - U'(n)\mu$ are independent and identically distributed as $\mathbf{W}' - U'\mu$ under the tilted measure P' defined in Section 2.3.1 through (4.3.14) to (4.3.16). Perform the simulation until $N_m = \inf\{n \geq 0 : C_i(n) > m \text{ for some } i\}$.
    Reset $N \leftarrow N + N_m$.
    Compute $p = 1/\sum_{k=1}^{d} w_k \exp(\theta_k^* C_k(N_m))$ and sample a Bernoulli I with probability p.
    if $I == 1$ **then**
      $\mathbf{Z}(\tau + \sum_{k=1}^{N_m} U'(k)) = \mathbf{Z}(\tau) + \mathbf{C}(N_m)$ and $\tau = \tau + \sum_{k=1}^{N_m} U'(k)$.
    else
      return τ, $N$and the feed-in path $(\mathbf{Z}(t) : 0 \leq t \leq \tau)$.
    end if
  end if
end while

can conclude:

$$E[N(d)] = E[N_b](E[N_a] + E[N_m]).$$

The following proposition gives an estimate for $E[N_m]$.

Proposition C.1.2. *Under our Assumptions C1) to C3),*

$$E[N_m] = O(\log d)$$

and the coefficient in the bound depends only on δ *and* H.

Proof. First, let us consider the cases in which W_i are uniformly bounded from above by some constant B.

Recall that $\phi_i(\theta) = E_0[\exp(\theta Z_i(1))]$. Given $Index = i$, one can check that $E_0'[C_i(1)] = \phi_i'(\theta_i^*)/(\lambda E[\exp(\theta_i^* W_i)]) \geq \phi_i'(\theta_i^*)/(\lambda H)$. N_m is a stopping time and $C_i(N_m) < m + B$. By the Optional Sampling Theorem, we have

$$E[N_m] = \sum_{i=1}^{d} \omega_i \frac{E_0'[C_i(N_m)]}{E_0'[C_i(1)]} \leq \sum_{i=1}^{d} \omega_i \frac{\lambda H(m+B)}{\phi_i'(\theta_i^*)}.$$

For each $1 \leq i \leq d$, we are going to estimate a lower bound for $\phi'(\theta_i^*)$. Using Taylor's expansion around 0, we have

$$\phi_i(\theta_i^*) = \phi_i(0) + \theta_i^* \phi_i'(0) + \frac{(\theta_i^*)^2}{2} \phi_i''(u_1 \theta_i^*),$$

for some $u_1 \in [0,1]$. As $\phi_i(\theta_i^*) = \phi_i(0) = 1$, we have,

$$\theta_i^* \phi_i'(0) + \frac{(\theta_i^*)^2}{2} \phi_i''(u_1 \theta_i^*) = 0.$$

As $\theta_i^* > 0$,

$$\phi_i'(0) + \frac{\theta_i^*}{2} \phi_i''(u_1 \theta_i^*) = 0. \tag{C.1}$$

Under Assumption C1), $\phi_i'(0) = E_0[Z_i(1)] < -\delta$. Under Assumption C2), we have that

$$E_0[\exp((\delta + \theta_i^*) Z_i(1))] \leq \exp(\lambda \log(E[\exp((\delta + \theta_i^*) W_i)])) \leq H^{\lambda} \leq H^H \triangleq H_1 < \infty.$$

As a result,

$$\begin{aligned}
\phi_i^{''}(u_1\theta_i^*) &= E[Z_i(1)^2\exp(u_1\theta_i^* Z_i(1))] \\
&\leq E[Z_i(1)^2 I(Z_i(1)\leq 0)] + E[Z_i(1)^2\exp(\theta_i^* Z_i(1))I(Z_i(0)>0)] \\
&\leq E[Z_i(1)^2] + E[Z_i(1)^2\exp(\theta_i^* Z_i(1))I(Z_i(0)>0)] \\
&\leq E[Z_i(1)^2] + E[Z_i(1)^2\exp(-\delta Z_i(1))\cdot\exp((\delta+\theta_i^*)Z_i(1))]
\end{aligned}$$

Besides, one can check that for any $x>0$, $x^2\exp(-\delta x)\leq 4e^{-2}/\delta^2$. Therefore,

$$\begin{aligned}
\phi_i^{''}(u\theta_i^*) &\leq E[Z_i(1)^2] + \frac{4}{\delta^2}e^{-2}E[\exp((\delta+\theta_i^*)Z_i(1))] \\
&\leq H + \frac{4}{\delta^2}e^{-2}H_1.
\end{aligned}$$

Plug this result into equation (C.1) and use that $\phi_i'(0) < -\delta$ to conclude the inequality:

$$\theta_i^* \geq \frac{2\delta}{H + 4e^{-2}H_1/\delta^2}. \tag{C.2}$$

On the other hand, by a Taylor's expansion of $\phi_i(\cdot)$ around θ_i^*, we can conclude that

$$\phi_i'(\theta_i^*) = \frac{\theta_i^*}{2}\phi''(u_2\theta_i^*), \tag{C.3}$$

for some $u_2 \in [0,1]$. Note that

$$\begin{aligned}\phi_i''(u_2\theta_i^*) &= E_0[Z_i(1)^2\exp(u_2\theta_i^* Z_i(1))] \geq E_0[Z_i(1)^2\exp(u_2\theta_i^* Z_i(1))I(U>1)] \\ &\geq E[\mu_i^2\exp(-\theta_i^*\mu_i)I(U>1)] \geq \mu_i^2\exp(-H\mu_i)\exp(-\lambda) \geq \delta^2\exp(-H^2-H).\end{aligned}$$

Thus, (C.2) together with (C.3) imply

$$\phi_i'(\theta_i^*) \geq \frac{1}{2}\theta_i^*\delta^2 e^{-\lambda} \geq \frac{\delta^3 e^{-H^2-H}}{H+4e^{-2}H_1/\delta^2}. \tag{C.4}$$

Note that for the lower bound (C.4) to hold we do not require W_i to be bounded.

Therefore,

$$E[N_m] \leq \sum_{i=1}^{d}\omega_i\frac{\lambda H(m+B)}{\phi_i'(\theta_i^*)} \leq \frac{\lambda H(m+B)(H+4e^{-2}H_1/\delta^2)}{\delta^3 e^{-H^2-H}},$$

as $\omega_i > 0$ and $\sum_i \omega_i = 1$.

By (C.2), we have that θ_i^* are all uniformly bounded away from 0, so we can choose $m = O(\log d/\min_i\theta_i^*) = O(\log d)$ to satisfy equation (4.3.11). Now, we can conclude that $E[N_m] = O(\log d)$ as B, H and δ are all constants independent of d.

Now, let's consider the more general cases when the W_i's are not bounded from above. Recall that $\mathbf{W}'$ is derived from $\mathbf{W}$ by exponential tilting, see (4.3.15). For any $B > 0$, define $\tilde{\mathbf{W}}'$ by $\tilde{W}_i' = W_i' I(W_i' \leq B)$ be the truncation of $\mathbf{W}'$ and define the random walk $\tilde{C}_i(n) = \tilde{C}_i(n-1) + \tilde{W}_i'(n) - U'(n)\mu_i$. Let $\tilde{N}_m = \inf\{n : \tilde{C}_i(n) > m \text{ for some } i\}$. Since $\tilde{C}_i(n) \leq C_i(n)$, we have $\tilde{N}_m \leq$

N_m. Our goal is to show that one can choose a proper value for B such that $E[\tilde{N}_m] = O(\log d)$ and hence so is $E[N_m]$.

Since $\tilde{W}_i'$ is bounded from above by B, by the Optimal Stopping Theorem we have:

$$E[\tilde{N}_m] \leq \sum_{i=1}^{d} \omega_i \frac{m+B}{E[\tilde{C}_i(1)]}.$$

By definition,

$$E[\tilde{C}_i(1)] = E[(W_i I(W_i \leq B) - U\mu_i)\exp(\theta_i^*(W_i I(W_i \leq B) - U\mu_i))],$$

Since $U\mu_i \geq 0$, we have

$$\begin{aligned} & E[(W_i I(W_i \leq B) - U\mu_i)\exp(\theta_i^*(W_i I(W_i \leq B) - U\mu_i))] \\ \geq\ & E[(W_i - U\mu_i)\exp(\theta_i^*(W_i - U\mu_i)] - E[W_i \exp(\theta_i^* W_i) I(W_i > B)]. \end{aligned}$$

By Assumption C2), δ and $H > 0$ are constants independent of d such that

$$E[\exp((\delta + \theta_i^*)W_i)] \leq H < \infty.$$

As a consequence,

$$\begin{aligned} E[W_i \exp(\theta_i^* W_i) I(W_i > B)] &\leq E[W_i \exp(-\delta W_i) I(W_i > B)\exp((\delta + \theta_i^*)W_i)] \\ &\leq \max_{w > B}\{w\exp(-\delta w)\} E[\exp((\delta + \theta_i^*)W_i)] \leq B\exp(-\delta B)H \end{aligned}$$

for all $B > 1/\delta$. Recall that by (C.4)

$$E[(W_i - U\mu_i)\exp(\theta_i^*(W_i - U\mu_i)] = E[C_i(1)] \geq \phi_i'(\theta_i^*)/(\lambda H) \geq \frac{\delta^3 e^{-H^2-H}}{\lambda H(H + 4e^{-2}H_1/\delta^2)},$$

where $H_1 = H^H$. Therefore, we can take $B = O(-\frac{1}{\delta}\log(\frac{\delta^3 e^{-H^2-H}}{2\lambda H^2(H+4\delta e^{-2}H_1/\delta^2)}))$ independent of d such that

$$B\exp(-\delta B)H < \frac{\delta^3 e^{-H^2-H}}{2\lambda H(H + 4e^{-2}H_1/\delta^2)}, \text{ and hence, } E[\tilde{C}_i(1)] \geq \frac{\delta^3 e^{-H^2-H}}{2\lambda H(H + 4e^{-2}H_1/\delta^2)}.$$

In the end, since $m = O(\log(d))$ we have

$$E[N_m] \leq E[\tilde{N}_m] \leq \frac{2\lambda H(m+B)(2H + 8e^{-2}H_1/\delta^2)}{\delta^3 e^{-H^2-H}} = O(\log d).$$

□

Now, we give the proof of Theorem 4.3.6.

Proof to Theorem 2. Recall that

$$E[N] = E[N_b](E[N_a] + E[N_m]).$$

Since N_b is the number of trials required to obtain $I = 0$, $E[N_b] = 1/P(I = 0)$. As discussed in

section 2.3.1, $P(I=0) \geq 1-\sum_{i=1}^{d} \exp(-\theta_i^* m)$ and hence

$$E[N_b] \leq \frac{1}{1-\sum_{i=1}^{d} \exp(-\theta_i^* m)} \leq \frac{1}{1-\frac{1}{d}}.$$

if we take $m = 2\log d / \min_i \theta_i^*$.

Similarly, we have $E[N_a] = 1/P(U > (m+W_i)/\mu_i,\ \forall\ i)$. For any $K > 0$,

$$P(U > \frac{m+W_i}{\mu_i},\ \forall\ i) \geq P(U > \frac{m+K}{\min_i \mu_i}; W_i \leq K \text{ for all } i).$$

Under Assumption C2), we have

$$P(W_i \leq K \text{ for all } i) \geq 1 - \sum_{i=1}^{d} P(W_i > K) \geq 1 - dH\exp(-K\delta).$$

Under Assumption C3), we have

$$P(U > \frac{m+K}{\min_i \mu_i}) \geq \exp(-\frac{H(m+K)}{\min_i \mu_i}).$$

As U and $\mathbf{W}$ are independent,

$$P(U > \frac{m+W_i}{\mu_i},\ \forall\ i) \geq \exp\left(-\frac{H(m+K)}{\min_i \mu_i}\right)(1-dH\exp(-K\delta)).$$

Choosing $K = (2\log d + \log H)/\delta$ and plugging in $m = 2\log d / \min_i \theta_i^*$, we get

$$E[N_a] \leq \frac{1}{1-\frac{1}{d}} \, d^{(2H/(\min_i \mu_i \min_i \theta_i^*) + 2H/(\delta \min_i \mu_i))} \, H^{H/(\delta \min_i \mu_i)}$$

By Proposition 4 we have $E[N_m] = O(\log d)$. In summary, we have:

$$\begin{aligned} E[N] = E[N_b](E[N_a] + E[N_m]) &= O\left((\frac{1}{1-\frac{1}{d}})^2 \, \log d \, d^{\frac{2H}{\min_i \mu_i \min_i \theta_i^*}} \right) \\ &= O(d^{1+\frac{2H}{\min_i \mu_i \min_i \theta_i^*}}). \end{aligned}$$

As discussed in the Proof of Proposition 4, $\theta_i^* \geq \delta/(H + 4e^{-2}H_1/\delta^2)$ and $\mu_i \geq \delta$ are uniformly bounded away from 0, therefore,

$$E[N] = O(d^{1+\frac{2H(H+4e^{-2}H_1/\delta^2)}{\delta^2}}).$$

□

Appendix D

Appendix for Chapter 5

D.1 Exponential Ergodicity of RBM

Consider a reflected Brownian motion Y with parameters (μ, Σ, R). Define $\mathcal{P}(x, A) = P(Y(1) \in A | Y(0) = x)$ be the transition probability of the reflected Brownian motion. We prove that $\mathcal{P}^n$ converges to the stationary distribution π in an exponential rate under the stability condition that $R^{-1}\mu < 0$.

Theorem D.1.1. *If $R^{-1}\mu < 0$, $\mathcal{P}$ admits a unique invariant measure π^*. Moreover, there exist some constants $C > 0$ and $\rho \in (0,1)$ such that*

$$\|\mathcal{P}^n \phi - \pi^*(\phi)\| \leq C\rho^n \|\phi - \pi^*(\phi)\|$$

holds for every measurable function $\phi : \mathbb{R}_+^d \to \mathbb{R}$ *such that*

$$\sup_x \frac{|\phi(x)|}{1+V(x)} < \infty,$$

where $V(x)$ *is a chosen positive function.*

Proof. It is proved in [23] that for any $x \in \mathbb{R}_+^d$ and any closed set $A \subset \mathbb{R}_+^d$ with positive Lebesgue measure, $\mathcal{P}(x,A) > 0$. As a result, the Markov chain $Y(\cdot)$ is aperiodic and $\psi-$irreducible. According to [50], it is sufficient to find a function $W : \mathbb{R}_+^d \to \mathbb{R}$ satisfying the following conditions:

(a) For any $K > 0$, $\{x \in \mathbb{R} \leq K\}$ is compact (bounded).

(b) There exists a compact set F and $\varepsilon, R > 0$ such that

$$\mathcal{P}W(x) \leq \begin{cases} W(x) - \varepsilon & \text{if } x \notin F \\ R & \text{if } x \in F. \end{cases}$$

(c) There exists $\theta, \Theta > 0$ independent of x, such that $\int \exp(\theta(W(y) - W(x)))\mathcal{P}(x,dy) < \Theta$ and $\int (W(y) - W(x))^2 \mathcal{P}(x,dy) < \Theta$.

Given the above conditions, we can choose $V(x) = \exp(\delta W(x))$ for some $\delta > 0$ so that $\mathcal{P}V(x) \leq rV(x) + B1(x \in F)$ and Theorem D.1 follows immediately by Theorem 16.3.1 in [50].

We shall use the Lyapunov function constructed in [23] as the function W we desired. Then W is $C^2 \in \mathbb{R} - \{0\}$. Besides, W has the following properties as proved in [23]:

(i) For any $M > 0$, there exists $N > 0$ such that $W(x) > M$ for all $\|x\| > N$.

(ii) There exists $r_0, \varepsilon_0 > 0$ such that for all $\|x\| > r_0$, $\lim_{t\to 0} \frac{E[W(Y(t))|Y(0)=x]-W(x)}{t} < -\varepsilon_0$.

(iii) For any $\eta > 0$, there exists $N > 0$ such that $\|D^2 W(x)\| < \eta$ for all $\|x\| > N$.

Then Condition (a) follows immediately after Condition (i). To check Condition (b), lets define $F = \{x : x \leq \max_{\|x\|\leq r_0} W(x) + \varepsilon_0\}$. Then, according to Condition (a) (or (i)), F is compact. Using Condition (ii) and the continuity of $W(Y(\cdot))$, it is easy the check that $\mathcal{P}W(x) \leq W(x) - \varepsilon$ for some fixed $\varepsilon > 0$ and all $x \notin F$. In the end, we choose $R = \max_{\|x\|\leq r_0} W(x) + \varepsilon_0$ and Condition (b) is satisfied.

As to Condition (c), according to Condition (i) and (iii), for any $\eta > 0$ there exists $K > 0$ such that $\|W(y) - W(x)\| \leq \eta\|y - x\|^2 + K$. Note that

$$\int \exp(\theta(W(y) - W(x)))\mathcal{P}(x, dy) = E[\exp(\theta(W(Y(1)) - W(x)))] \leq e^K E[\exp(\eta\theta\|Y(1) - x\|)].$$

Let $Z(\cdot)$ be the input free Brownian motion. Then, $\|Y(1) - x\| \leq l \max_{0\leq t\leq 1} \|Z(t) - x\|$ for some constant $l > 0$ by the Lipschitz continuity of the Skorokhod mapping and hence

$$\int \exp(\theta(W(y) - W(x)))\mathcal{P}(x, dy) \leq e^K E[\exp(\eta\theta l \max_{0\leq t\leq 1} \|Z(t) - x\|)].$$

On the other hand, as Σ is non-degenerate, there exists $\eta_0 > 0$ such that $P(\max_{0\leq t\leq 1} Z(t) \in dz) = O(\exp(-\eta_0\|x - z\|^2))$ as $\|x - z\| \to \infty$. As a result, we can check that $E[\exp(\eta\theta l \max_{0\leq t\leq 1} \|Z(t) - x\|)] < \Theta$ for some $\Theta > 0$, $\theta < \eta_0/l\eta$ and for all x. Similarly, we can check that $\int (W(y) - W(x))^2 \mathcal{P}(x, dy) < \Theta$ holds for some $\theta, \Theta > 0$.

In summary, Condition (a), (b) and (c) are all satisfied for the Lyapunov function W constructed in [23] and according to Theorem 16.3.1, Theorem D.1 is proved. □

Remark: Suppose π is the stationary distribution of the EBM, then π^* is also an invariant measure for $\mathcal{P}$. By uniqueness, we have $\pi^* = \pi$ and hence $\mathcal{P}^n$ converges to π in an exponential rate.

D.2 Details on Algorithm 5.6

D.2.1 A Conceptual Framework for the Joint Simulation of τ_ε and $\mathbf{Z}^\varepsilon$

Our goal now is to develop an algorithm for simulating τ_ε and $(\mathbf{Z}^\varepsilon(t) : 0 \le t \le \tau_\varepsilon)$ jointly. In detail, we want to simulate $\mathbf{Z}^\varepsilon(t)$ forwards in time and stop at a random time τ_ε such that for any time $s > \tau_\varepsilon$, $Z_i(s) \le Z_i(\tau_\varepsilon) + \varepsilon$ for $1 \le i \le d$.

Because of the special structure of the wavelet representation used in the TES algorithm simulating the process $\mathbf{Z}^\varepsilon(\cdot)$, the time $T_m \triangleq \inf\{t \ge 0 : Z_i^\varepsilon(t) > m \text{ for some } 1 \le i \le d\}$ is no longer a stopping time with respect to the filtration generated by $\mathbf{Z}(\cdot)$. As a consequence, we cannot directly carry out importance sampling as in Algorithm 4.3. To remedy this problem, we decompose the process $\mathbf{Z}^\varepsilon(t)$ into two parts, say, a random walk $\{\mathbf{Z}^\varepsilon(n) : n \ge 0\}$ with Gaussian increment and a series of independent Brownian bridges $\{\bar{\mathbf{B}}_n(s) \triangleq \mathbf{Z}^\varepsilon(n+s) - \mathbf{Z}^\varepsilon(n) : s \in [0,1], n \ge 0\}$. Our strategy is to first carry out the importance sampling as in Algorithm 4.3 to the random walk $\{\mathbf{Z}^\varepsilon(n) : n \ge 0\}$ to find its upper bound, and next develop a new scheme to

control the upper bounds attained in the intervals $\{(n,n+1) : n \geq 0\}$ for the i.i.d. Brownian bridges $\{\bar{\mathbf{B}}_n(s) : s \in [0,1], n \geq 0\}$.

The whole procedure is based on the wavelet representation of Brownian motion. Let $\{W_n^k(i) : n,k \in \mathbb{N}, i = 1,2,...,d\}$ be a sequence of i.i.d. standard normal random variables. According to the expression given in 5.2.3, for any $t = n+s$, $s \in [0,1]$:

$$Z_i(t) = Z_i(n) + s(Z_i(n+1) - Z_i(n)) + \sum_{j=1}^{d} A_{ij}(\sum_{k=1}^{\infty} W_n^k(j) \int_0^s H_k(u)du). \tag{D.1}$$

Let us put (D.1) in matrix form:

$$\mathbf{Z}(t) = \mathbf{Z}(n) + s(\mathbf{Z}(n+1) - \mathbf{Z}(n)) + A \sum_{k=1}^{\infty} \mathbf{W}_n^k \cdot \int_0^s H_k(u)du.$$

For all $n \geq 0$ and $s \in [0,1]$, $\bar{\mathbf{B}}_n(s) = A\sum_{k=1}^{\infty} \mathbf{W}_n^k \cdot \int_0^s H_k(u)du$. It is obvious that the sequence $\{\bar{\mathbf{B}}_n(\cdot) : n \geq 0\}$ is i.i.d. Note that $(Z_i(n+1) - Z_i(n))$ is independent of $\{W_n^k(i) : k \geq 1\}$. We can split the simulation into two independent parts:

1. Simulate the discrete-time random walk $\{\mathbf{Z}(n) : n \geq 0\}$ with i.i.d. Gaussian increments and $\mathbf{Z}(0) = 0$. That is $Z_i(0) = 0$ and $Z_i(n+1) = Z_i(n) + \sum_{j=1}^{d} A_{ij} W_{n+1}^0(j) - \mu_i$, where $\{W_n^0(j) : n \geq 0\}$ are i.i.d. standard normals.

2. For each n, simulate $\bar{\mathbf{B}}_n(s)$ to do bridging between $\mathbf{Z}(n)$ and $\mathbf{Z}(n+1)$.

Now, any time $t_0 > 0$ is an approximate coalescence time τ_ε if there exists some positive constant $\zeta > 0$ such that the following two conditions hold for all $n \geq t_0$: Condition 1), $\mathbf{Z}(n) \leq \mathbf{Z}(t_0) - \zeta(n - \lceil t_0 \rceil)\mathbf{1} + \varepsilon$, and Condition 2), $\max\{\bar{\mathbf{B}}_n(s) : s \in [0,1]\} \leq \zeta(n - \lceil t_0 \rceil)\mathbf{1}$. Based

on these observations, we develop an algorithm to simulate the approximate coalescence time τ_ε jointly with $\{\mathbf{Z}^\varepsilon(t) : 0 \le t \le \tau_\varepsilon\}$.

Since $\mu < 0$.We can find$\zeta > 0$ and let $\mathbf{S}(n) = \mathbf{Z}(n) + n\zeta\mathbf{1}$ such that $\{\mathbf{S}(n) : n \ge 0\}$ is a random walk with strictly negative drift. Therefore, Condition 1) can be checked by carrying out the importance sampling procedure as in Algorithm 4.3 for the random walk $\{\mathbf{S}(n) : n \ge 0\}$. More precisely, since $S_i(n)$ has Gaussian increments, we can compute explicitly that $\theta_i^* = 2(\mu_i - \zeta)/\sigma_i$ and choose $m > 0$ satisfying (4.3.11) in order to carry out the importance sampling procedure for the random walk $\{\mathbf{S}(n) : n \ge 0\}$. Suppose we use the importance sampling procedure and find t_0 such that $\mathbf{S}(n) \le \mathbf{S}(t_0)$ for all $n \ge t_0$ and hence Condition 1) is satisfied for t_0.

About Condition 2), recall that $\bar{\mathbf{B}}_n(\cdot)$'s are i.i.d. linear combinations of Brownian bridges and let M be a random time, finite almost surely, such that

$$M \ge \max\{n \ge t_0 : \max_{0 \le s \le 1} (\bar{B}_{n,i}(s) - \zeta(n - t_0)) > 0 \text{ for some } i\}. \tag{D.2}$$

Observe that for t_0 to be an approximate coalescence time, Condition 1) and Condition 2) must hold simultaneously. If for time t_0, for example, Condition 1) is satisfied while Condition 2) is not, we need to continue the testing procedure and simulation of the process for $t > t_0$. Then, however, the random walk $\{\mathbf{S}(n) : n \ge \lceil t_0 \rceil\}$ should be conditioned on that $\mathbf{S}(n) \le \mathbf{S}(t_0)$ for the fact that Condition 1) holds for t_0 reveals 'additional information' on the random walk for $n \ge t_0$. Therefore, such 'additional information' or 'conditioning event' must be incorporated and tracked when Conditions 1) and 2) are sequentially tested. All of these conditioning events

are described and accounted for in Section D.2.3 , which also includes the overall procedure to sample τ_ε jointly with $\mathbf{Z}^\varepsilon$.

Now, let us first provide a precise description of M and explain the simulation algorithm for M in Section D.2.2.

D.2.2 Simulating M and $\{\bar{\mathbf{B}}_n^\varepsilon(\cdot) : 1 \leq n \leq M\}$

Recall that $\bar{\mathbf{B}}_n(t) = A\sum_{k=1}^{\infty} \mathbf{W}_n^k \cdot \int_0^t H_k(u)du$, where $\{W_n^k(i) : n \geq 0, k \geq 1, 1 \leq i \leq d\}$ are i.i.d. standard normals. Note that

$$\sum_{n=0}\sum_{k=1} P\left(|W_n^k(i)| \geq 4\sqrt{\log(n+1)} + 4\sqrt{\log k}\right) \leq \sum_{n=0}\sum_{k=1} \frac{1}{((n+1)k)^4} < \infty.$$

By the Borel-Cantelli Lemma, we can conclude that for each $i \in \{1,...,d\}$ there exists $M^i < \infty$ such that for all $(n+1)k > M^i$, $|W_n^k(i)| \leq 4\sqrt{\log(n+1)} + 4\sqrt{\log k}$. Clearly, $\sqrt{\log t} = o(t)$ as $t \to \infty$, so we can select a m_0 large enough such that for any $n > m_0$

$$(n+1)\zeta - ad(4\sqrt{\log(n+1)} - \sum_{j=1}^{\infty} 2^{-j}\sqrt{j}) \geq 0.$$

Note that M^i can be simulated jointly with $(W_n^k(i) : n \geq 0, k \geq 1, 1 \leq i \leq d, (n+1)k \leq M^i)$ by easily adapting Algorithm 5.1 and M^i's are independent of each other. Then, for any $n >$

$\max_{i=1}^{d} M^i \vee m_0$,

$$\begin{aligned}\bar{\mathbf{B}}_n(t) &= A\sum_{k=1}^{\infty}\mathbf{W}_n^k\cdot\int_0^t H_k(u)du \\ &\le ad(4\sqrt{\log(n+1)}+\sum_{j=1}^{\infty}2^{-j/2}\sqrt{j}) \le (n+1)\zeta,\end{aligned}$$

where, $j=\lceil\log_2 k\rceil$. Therefore, we can choose $M=\max_i M^i\vee m_0$.

Now we introduced a variation of Algorithm 5.1 that will be used in the procedure to simulate M and $\{\bar{B}_n^{\varepsilon}(\cdot):1\le n\le M\}$ jointly as Algorithm D.1. In this algorithm, a sequence of 'conditioning events' of the form $|W^k|\le\beta_k$, for some given constants $\{\beta^k:\beta^k>4\sqrt{\log k}\}$, is in force. Let $\Phi(a)=P(|W|<a)$ for all $a>0$, where W is a standard normal. The random number $K=\max\{k:|W^k|>4\sqrt{\log k}\}$.

The main difference between Algorithm D.1 and the original Algorithm 5.1 is that U and V are now computed from the conditional probability, however, the relations that $U>V>D$ and $U-D\to 0$ still hold and hence Algorithm D.1 is valid. Based on this, we can now give the main procedure Algorithm D.2 to simulate M and $\{\bar{B}_n^{\varepsilon}(\cdot):1\le n\le M\}$ jointly.

In Step 1 of Algorithm D.2, we use a similar procedure as Algorithm D.1 to impose conditioning events of form $|W_n^k(i)|\le\beta_n^k(i)$ while simulating M_i's jointly with $W_n^k(i)$'s. In this way, Algorithm D.2 is able to simulate M jointly with $\{\bar{\mathbf{B}}_n^{\varepsilon}(\cdot):1\le n\le M\}$ conditional on $|W_n^k(i)|\le\beta_n^k(i)$ for all $n\ge 0$, $k\ge 1$ and $1\le i\le d$ for any given sequence of $\{\beta_n^k(i)\}$ such that $\beta_n^k(i)>4(\sqrt{\log(n+1)}+\sqrt{\log k})$. As a result, it can be used to keep track of 'conditioning events' corresponding to Condition 2).

Algorithm D.1 Simulate K jointly with $\{W^k : 1 \le k \le K\}$ conditional on $|W^k| \le \beta^k$ for all $k \ge 1$

Input:

$K_0 := \inf\{l : Er(l) < \varepsilon\}$.

Output:

K jointly with $\{W^k : 1 \le k \le K\}$.

Initialize $G = 2^{K_0}$ and S to be an empty array. Set $I = 1$.
while I==1 **do**
  Set $U = 1, D = 0$. Simulate $V \sim Uniform(0,1)$.
  while $U > V > D$ **do**
    set $G \leftarrow G+1$ and $U \leftarrow \frac{\Phi(4\sqrt{\log G})}{\Phi(\beta^k)} \times U$ and $D \leftarrow (1-G^{-7}) \times U$.
  end while
  if $V \ge U$ **then**
    add G to the end of S, i.e. $S = [S, G]$.
  else
    if $V \le D$ **then**
      $K = G$ and set $I = 0$.
    end if
  end if
end while
for n=1:K **do**
  if $n \in S$ **then**
    generate W^k according to the conditional distribution of Z given $\{4\sqrt{\log k} < |W| \le \beta^k\}$;
  else
    generate W^k according to the conditional distribution of W given $\{|W| \le 4\sqrt{\log k}\}$.
  end if
end for

D.2.3 Keeping Track of the Conditioning Events

As we have discussed just prior to the beginning of Section D.2.2, we need to keep track of several conditioning events introduced by Conditions 1) and 2). First, let us explain how to deal with the conditioning event corresponding to Condition 1). These conditioning events involves only the random walk $\mathbf{S}(\cdot)$. Now we split $\mathbf{S}(\cdot)$ according to the sequences of $\{\Gamma_l : l \ge 1\}$ and $\{\Delta_l : l \ge 1\}$ of random times defined as follows:

1. Set $\Delta_1 = \min\{n : S_i(n) \le -2m \text{ for every } i\}$.

Algorithm D.2 Simulating of M and $\{\bar{\mathbf{B}}^{\varepsilon}_n(\cdot) : 1 \le n \le M\}$ jointly conditional on $|W^k_n(i)| \le \beta^k_n(i)$ for all $n \ge 0$, $k \ge 1$ and $1 \le i \le d$

1: For each component i, simulate M_i and $(W^k_n(i) : n \ge 0, k \ge 1, nk < M)$ conditional on $|W^k_n(i)| \le \beta^k_n(i)$ using a similar procedure as Algorithm D.1. Compute $M = \max_i M^i \vee m_0$.
2: For each $0 \le n \le M$ and each component i, $\{W^k_n(i) : k < M^i/n\}$ are already given in Step 1. For $k \ge M^i/n$, use Algorithm D.1 to simulate K^i_n jointly with $\{W^k_n(i) : M^i/n \le k \le K\}$ conditional on $|W^k_n(i)| \le 4(\sqrt{\log(n+1)} + \sqrt{\log k})$. (Note that $\beta^k_n(i) > 4(\sqrt{\log(n+1)} + \sqrt{\log k}) > 4\sqrt{\log k}$ and hence this step is well-defined.)
3: For any $0 \le n \le M$, compute and output

$$\bar{B}^{\varepsilon}_{n,i}(t) = \sum_{i=1}^{d} A_{ij} \Big(\sum_{k=1}^{K^i_n} W^k_n(i) \int_0^t H_k(u)du \Big). \tag{D.3}$$

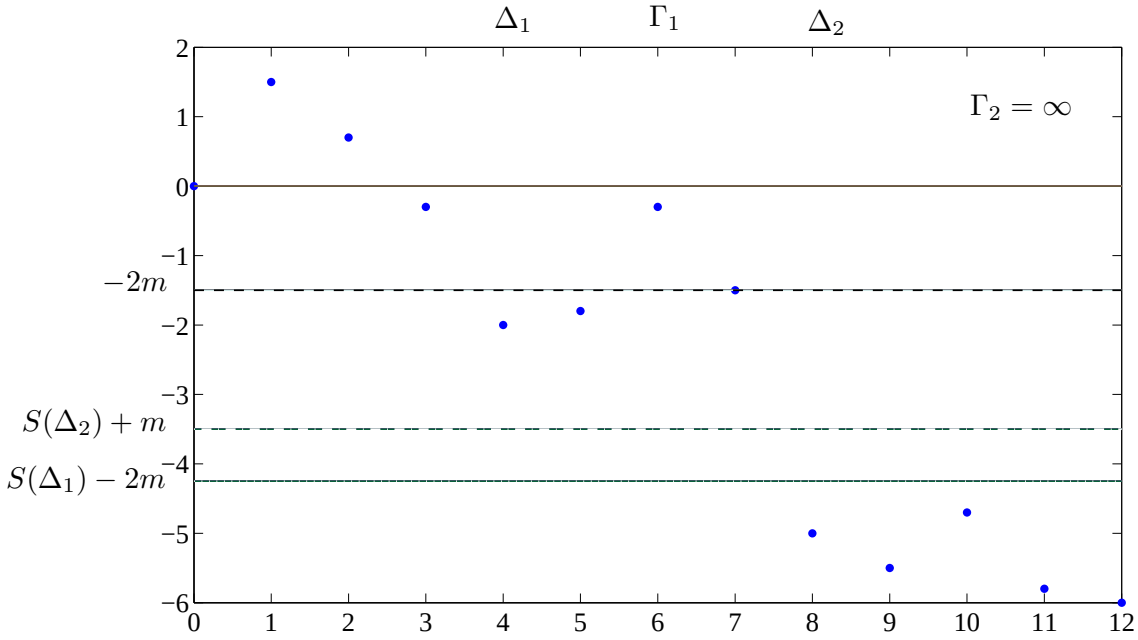
2. Define $\Gamma_l = \min\{n \ge \Delta_l : S_i(n) > S_i(\Delta_l) + m \text{ for some } i\}$.

3. Put $\Delta_{l+1} = \min\{n \ge \Gamma_l I(\Gamma_l < \infty) \vee \Delta_l : S_i(n) < S_i(\Delta_l) - 2m \text{ for every } i\}$.

Figure D.1 illustrates a sample path of the random walk with the sequence of random times $\{\Gamma_l : l \ge 1\}$ and $\{\Delta_l : l \ge 1\}$ in one dimension. The message is that the joint simulation of $\{\mathbf{S}(n) : n \ge 0\}$ with $\{\Gamma_l : l \ge 1\}$ and $\{\Delta_l : l \ge 1\}$ allows us to keep track of the process $\{\max_{m \ge n} \mathbf{S}(m) : n \ge 0\}$, which includes the 'additional information' introduced by Condition 1). The main steps in the simulation of $\{\mathbf{S}(n) : n \ge 0\}$ jointly with $\{\Gamma_l : l \ge 1\}$ and $\{\Delta_l : l \ge 1\}$ are explained in Lemma 2 through Lemma 4 in [7]. The approach of [7], which works in one dimension, could be modified for multidimensional cases using the change-of-measure as described in Section 4.3.3.

Now, regarding the verification of Condition 2) involving M and the Brownian bridges. Given the discussion in Section D.2.2, we just need to keep track of certain deterministic $\beta^k_n(i)$ for each $|W^k_n(i)|$, that is to condition on the events of the form $|W^k_n(i)| \le \beta^k_n(i)$ which are related to the sequential construction of the random variable M in each time of testing Condition 2) as

Figure D.1: Illustration of $\{\Gamma_l : l \geq 1\}$ and $\{\Delta_l : l \geq 1\}$



described in Section D.2.2. Now, we can write down the integrated Algorithm D.3for sampling τ_ε and $\{\mathbf{Z}^\varepsilon(t) : 0 \leq t \leq \tau_\varepsilon\}$ jointly.

Algorithm D.3 Simulating τ_ε and $\{\mathbf{Z}^\varepsilon(t) : 0 \le t \le \tau_\varepsilon\}$.

Input:

$\mu \in \mathbb{R}^d$, Σ, $R \in \mathbb{R}^{d\times d}$ and error bound $\varepsilon > 0$

Output:

$\{\mathbf{Z}^\varepsilon(t) : 0 \le t \le \tau_\varepsilon\}$ and the approximation coalescence time τ_ε

set $\beta_n^k(i) = \infty$ for all $n \ge 1$, $k \ge 1$ and $1 \le i \le d$, $L = 0$, $\tau_\varepsilon = 0$ and $I = 1$
while I==1 **do**
  simulate $\mathbf{S}(n)$ until Δ_l, where $l = \min\{j : \Gamma_j = \infty, \Delta_j > \tau_\varepsilon\}$
  $\mathbf{Z}^\varepsilon(n) = \mathbf{S}(n) - n\zeta$
  $\forall n \in [\tau_\varepsilon, \Delta_l] \bigcap \mathbb{Z}_+$ and $1 \le i \le d$, compute the i.i.d. bridges $\{\bar{\mathbf{B}}_n^\varepsilon(\cdot)\}$ using (D.3), in which K_n^i is jointly simulated with $(W_n^k(i) : 1 \le k \le K_n^i)$ conditional on that $|W_n^k(i)| \le \beta_n^k(i)$ for all $k \ge 1$ using Algorithm D.1
  if $\exists t \ge \Gamma_{l-1}$ s.t. $\forall t \le s \le \Delta_l$, $Z_i^\varepsilon(t) \ge Z_i^\varepsilon(s) - 2\varepsilon$ and $Z_i^\varepsilon(t) \ge Z_i^\varepsilon(\Delta_l) + m - 2\varepsilon$ **then**
    $\tau_\varepsilon \leftarrow t$.
    use Algorithm D.2 to simulate M jointly with $(\bar{\mathbf{B}}_{\tau_\varepsilon+n}^\varepsilon(\cdot) : 0 \le n \le M)$ conditional on $|W_{\tau_\varepsilon+n}^k(i)| \le \beta_{\tau_\varepsilon+n}^k(i)$ for all $n \ge 0$, $k \ge 1$ and $1 \le i \le d$.
    $\beta_{\tau_\varepsilon+n}^k(i) \leftarrow 4\sqrt{\log(n+1)} + 4\sqrt{\log k}$, $\forall n \cdot k \ge M^i$.
    simulate $\mathbf{S}(n)$ until $n = \Delta_l + M$ and compute $\{\mathbf{Z}^\varepsilon(t) : t \in [\Delta_l, \Delta_l + M]\}$.
    if $\exists t, i$ s.t. $Z_i^\varepsilon(t) > Z_i^\varepsilon(\tau_\varepsilon) + \varepsilon$ **then**
      $\tau_\varepsilon \leftarrow t$.
    else
      $I \leftarrow 0$.
    end if
  else
    $\tau_\varepsilon \leftarrow \Delta_l$.
  end if
end while
return τ_ε and $(\mathbf{Z}^\varepsilon(t) : 0 \le t \le \tau_\varepsilon)$.