Studies in Stochastic Networks: Efficient Monte-Carlo Methods, Modeling and Asymptotic Analysis

Jing Dong

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

Studies in Stochastic Networks: Efficient Monte-Carlo Methods, Modeling and Asymptotic Analysis

Jing Dong

This dissertation contains two parts. The first part develops a series of bias reduction techniques for: point processes on stable unbounded regions, steady-state distribution of infinite server queues, steady-state distribution of multi-server loss queues and loss networks and sample path of stochastic differential equations. These techniques can be applied for efficient performance evaluation and optimization of the corresponding stochastic models. We perform detailed running time analysis under heavy traffic of the perfect sampling algorithms for infinite server queues and multi-server loss queues and prove that the algorithms achieve nearly optimal order of complexity. The second part aims to model and analyze the load-dependent slowdown effect in service systems. One important phenomenon we observe in such systems is bi-stability, where the system alternates randomly between two performance regions. We conduct heavy traffic asymptotic analysis of system dynamics and provide operational solutions to avoid the bad performance region.

Table of Contents

List of Figures

List of Tables

Acknowledgments

The completion of this dissertation would have never been possible without the support and inspiration of many people. I would like to acknowledge their contribution to make this endeavor a wonderful experience.

First and foremost, I owe my deepest gratitude to my advisor, Professor Jose Blanchet, for his guidance and support throughout the course of my doctoral study. His knowledge, sharpness, dedication and most importantly passion for research has made the experience both stimulating/challenging and enjoyable. The example he set, as a researcher and mentor, will be invaluable to me throughout my career.

It is my genuine fortune and pleasure to learn from and inspired by many great teachers at Columbia. I am extremely grateful to Professor Guillermo Gallego, Professor Philip Protter, Professor Karl Sigman, Professor Ward Whitt and Professor David Yao for all the insightful discussions, help and encouragement they provided me with at various stage of my academic journey.

I am also fortunate to encounter mentors and collaborators outside Columbia. Chapter 6 of the thesis is based on joint work with Professor Pnina Feldman and Professor Galit Yom-Tov. I would like to thank them for the fun exploring new subjects together and their help with my personal and professional development. I am also very grateful to Professor Josh Reed for serving on my thesis committee.

My most sincere appreciation goes for all my fellow Ph.D classmates at IEOR and the good friends I met along the journey. It is impossible for me to quantify how much I gained from their friendship. I am really grateful to them for sharing with me the happiness and supporting me through the hardship. I would also like to thank

all the IEOR faculty and staff members for creating such a supportive community for us.

The unconditional love, trust and support from my parents, Bing and Weidong, has been a constant source of strength and inspiration for me. No words can express my gratitude to them. Lastly, I want to thank Changyao, for taking very good care of me and brightening up my day. I would like to dedicate this work to him.

Part I

Bias Reduction Techniques for Stochastic Networks

Chapter 1

Introduction to Part I

The first part of the dissertation focuses on the algorithmic development and theoretical analysis of bias reduction techniques for several stochastic models that arise in various engineering and business applications. Sampling based computational methods are a fundamental part of the numerical toolset for performance evaluation and optimization of these models.

When evaluating the performance of a sampling scheme, the most analytically tractable measure of estimator quality is the mean square error (MSE). Suppose we are interested in estimating the mean of a random quantity Z, denoted as $\alpha = EZ$. Our sampling scheme outputs an estimator of α, denoted as $\hat{Z}$. Then

$$\begin{aligned} \mathrm{MSE}(\hat{Z}) &= E[(\hat{Z}-\alpha)^2] \\ &= \mathrm{Bias}(\hat{Z})^2 + \mathrm{Var}(\hat{Z}) \end{aligned}$$

where $\mathrm{Bias}(Z) = E[\hat{Z}] - \alpha$.

We can reduce the variance of the estimator by sampling i.i.d. copies of $\hat{Z}$ and take the average. Specifically,

$$\mathrm{Var}\left(\frac{1}{N}\sum_{k=1}^{N}\hat{Z}_k\right) = \frac{1}{N}\mathrm{Var}(\hat{Z}_1)$$

When applying this Monte Carlo method, the variance the estimator converges to zero at rate $1/N$. In the contrast, $\mathrm{Bias}(Z)$, cannot be eliminated through sampling

i.i.d. copies of $\hat{Z}$, i.e. $E[1/N \sum_{k=1}^{N} \hat{Z}_k] = E[\hat{Z}_1]$. Bias measures the systematic error of $\hat{Z}$.

In this dissertation, we develop algorithms to eliminate or reduce the bias of the estimator, thus improve the efficiency of Monte Carlo methods. In Chapter 2, we develop simulation schemes to sample point processes on stable unbounded regions. We also develop perfect sampling algorithms for infinite server queues. Perfect sampling consists of simulating without any bias from the steady-state distribution of a given ergodic process. In Chapter 3, we constructed perfect sampling algorithms for multi-server loss queues and loss networks. We conduct the running time analysis of our algorithms under heavy-traffic. Lastly in Chapter 4, we propose and analyze a class of algorithms that would allow us approximate the sample path of multi-dimensional stochastic differential equations (SDE) with any desired level of accuracy (ϵ-strong simulation).

The rest of the introduction is organized as follows. We first give an overview of perfect sampling (Section 1.1) and sample path simulation of SDEs (Section 1.2). We then introduce the idea of "record breakers", which are used in the development of both our perfect simulation algorithms and our SDE sampling schemes.

Throughout the discussion, we refer to a unit of computational effort (computational cost) as the simulation of a random variable or the evaluation of a simple function.

1.1 Perfect sampling

The steady state distribution in a particular set, measures the long run proportion of time the stochastic process spends in that set [1]. Specifically, for an ergodic process.

$$\pi(\mathcal{C}) := \lim_{t\to\infty} \frac{1}{t} \int_0^t 1\{X(s) \in \mathcal{C}\} ds$$

It is widely used for system performance evaluation. As the steady-state distribution is defined as the limiting distribution of the process as time goes to infinity, most naive

forward simulation algorithms suffer from the bias induced by the initial transient. This is because the process is in general initialized from an arbitrary state that does not follow the steady-state distribution [2].

The most common perfect sampling protocol, known as coupling from the past (CFTP), was proposed in the ground breaking paper by Propp and Wilson [3]. The theoretical concepts underlying is the following. Suppose that the process starts operating from the infinite past at an arbitrary state. It would be at stationarity at time zero. If we could recover the state of such system at time zero, then we get an unbiased sample from its steady-state distribution. There has since been a lot of work involving implementation of this idea for various applications and improving the efficiency of the algorithms (see for example [4],[5],[6] et al.). Kendall once said *"The topic of perfect simulation is made up of a variety of interacting ideas rather than a single grand theory: more of an orchestra of complementary techniques than a virtuoso prima donna of a Big Theory."* It still remains an active research area. Chapter 2 of this dissertation develops the perfect sampling algorithm for infinite sever queues with general interarrival time and service time distributions using the idea of CFTP. The main difficulty in applying CFTP to the general infinite server queueing models is that the state space, which consists of the age process of the renewal arrival process and a measured value process for remaining service times, is infinite dimensional and the transition kernel of the underlying Markov process is not directly accessible.

Foss and Tweedie [7] proved that CFTP can be applied if and only if the underlying process is uniformly ergodic. Kendall [8] proposed a variation of CFTP, called Dominated Coupling From the Past (DCFP), which allows one to obtain samples from the steady-state distribution of ergodic process without requiring uniform ergodicity. A nice summary of DCFTP is given in [9]. The idea is to construct a stationary process which suitably dominates the process of interest and can be simulated backwards in time from a stationary state at time zero. Then, a suitable lower bound process,

coupled with the upper bound, must also be simulated in stationarity and backwards in time. A typical application of DCFTP involves the construction of the upper and lower bounds up to a time in the past when they both meet. Then one says that the coalescence occurs. The process of interest is reconstructed forward in time from the coalescence position up to time zero, using the same input sequence that was used to simulate the coupled upper and lower bounds. The state of the process of interest at time zero must then follow the corresponding steady-state distribution. Chapter 3 of this dissertation develops perfect sampling algorithms for multi-server loss queues and loss networks with general interarrival time and service time distributions based on this idea.

The majority of the available perfect sampling algorithms for queues involve exponential distributional assumptions (on service times and/or interarrival times) and very few of such algorithms are applicable in the context of queueing networks. None of them, up to date, have been designed and analyzed in the setting of many server systems in heavy-traffic.

The paper [10] is one of the earliest to consider DCFTP in the setting of geometrically ergodic Harris recurrent Markov chains. General DCFTP algorithms have been developed more recently in [8] and [11] for Harris recurrent chains, although there are important practical limitations as outlined on p.788 in [11]. In particular, their algorithms assume that one has analytical access to the transition kernel of the underlying Markov chain after several transitions. A recent paper by Sigman [12] provides an implementable DCFTP algorithm for multi-server queues with Poisson arrivals, but the algorithm requires rather strong conditions on stability; in [13] the conditions are relaxed (also in the setting of Poisson arrivals), using a regenerative technique, but the expected termination time of the algorithm is infinite. In connection to loss queueing systems, Murdoch and Takahara [14] applied CFTP in the context of queueing models with bounded state space. For instance, they consider loss queues with renewal arrivals but with bounded service times. In this case, CFTP

can be easily implemented. The papers [15], [16] and [17] are close in spirit to the main ideas of our development, as we take a point process approach to the problem. However, their approach requires the use of spatial birth and death processes (generally of poisson type) as the dominating processes and as pointed out in Section 8 of [18], the algorithms appear to significantly increase in complexity as the arrival rate increases.

In Chapter 2 & 3, we provide a practical simulation procedure that works under the assumption of renewal arrivals with finite mean and i.i.d. service time distribution with finite mean (although in our running time analysis in heavy traffic we impose additional moment conditions for service times, but we still are able to cover distributions such as log-normal, which have been observed to accurately fit service time distributions in many server applications [19]). In order to implement DCFTP strategy in the setting of loss queues, we simulate a stationary infinite server queue backwards in time as our dominating process. A variation from the standard DCFTP protocol just explained is that we use the upper bound process itself to detect coalescence, thereby bypassing the need for a lower bound process and improving the running time of the algorithm. Basically we detect coalescence over a time interval in which all customers initially present in the infinite server system leave and no loss of customers occurs during that time interval. We perform running time analysis of the algorithms under heavy traffic and prove that they achieve nearly optimal order of complexity.

1.2 Sample path simulation of SDEs

Consider an SDE of the form

$$dX(t) = \mu(X(t))dt + \sigma(X(t))dB(t)\ ,\ X(0) = x(0) \tag{1.1}$$

where $B(\cdot)$ is a d'-dimensional Brownian motion, and $\mu(\cdot) : R^d \to R^d$ and $\sigma(\cdot) : R^d \to R^{d\times d'}$ satisfy suitable regularity conditions. We assume, in particular, that

both $\mu(\cdot)$ and $\sigma(\cdot)$ are Lipschitz continuous so that a strong solution to the SDE is guaranteed to exist [20]. SDEs occur in a variety of applications from physics to financial engineering. These applications require characterizing complex path-dependent functionals, such as the first passage time $\inf\{t \geq 0 : X(t) \in \mathcal{C}\}$, or the mean performance measure $Ef(X)$ where $X = \{X(t) : 0 \leq t \leq T\}$. In most cases, explicit analytical solutions are not available or suffer greatly from the curse of dimensionality (inefficient). In this context, simulation-based methods become attractive.

In general, the sample path X can not be generated and stored exactly/completely, because it is infinite dimensional. A natural approximation would be to first simulate the process on discrete skeletons and then construct the rest of the process by linear interpolation between grid points or treat them as piecewise constants.

One simplest such approximation is the Euler Scheme, where we simulate $\tilde{X}^h(t)$ for $t \in \{0, h, 2h, 3h, \dots\}$ sequentially,

$$\tilde{X}_i^h(t+h) = \tilde{X}_i^h(t) + \mu_i(\tilde{X}^h(t))h + \sum_{j=1}^{d'} \sigma_{i,j}(\tilde{X}^h(t))(B_j(t+h) - B_j(t))$$

for $i = 1, 2, \dots, d$.

Under regularity conditions on the drift and diffusion function: 1) lipschitz conditions [2], $|\mu(x) - \mu(y)| \leq K|x-y|$ and $|\sigma(x) - \sigma(y)| \leq K(x-y)$, 2) linear growth condition $|\mu(x)| + |\sigma(x)| \leq K(1+|x|)$, we have

$$E[(X(1) - \tilde{X}^h(1))^2] = O(h).$$

This implies $E|X(1) - \tilde{X}^h(1)| = O(h^{1/2})$. Every replication of the sample path takes $O(1/h)$ units of computational effort. Thus, if we want to achieve a MSE of order ϵ^2, it would take $O(\epsilon^{-4})$ units of computational effort. This is far from the optimal rate of convergence of such estimator, $O(\epsilon^{-2})$, when we have access to an unbiased estimator for which each sample requires $O(1)$ units of computational effort.

If we view Euler scheme as a first order expansion, then by applying a second order expansion on the diffusion term, we get the Milstein scheme [2], where we use

the recursion

$$\begin{aligned}\hat{X}_i^h(t+h) &= \hat{X}_i^h(t) + \mu_i(\hat{X}^h(t))h + \sum_{j=1}^{d'} \sigma(\hat{X}^h(t))(B_j(t+h) - B_j(t)) \\ &\quad + \sum_{j=1}^{d'} \sum_{l=1}^{d} \sum_{l=1}^{m} \sum_{m=1}^{d'} \partial_l \sigma'_{i,j}(\hat{X}^h(t)) \sigma_{l,m}(\hat{X}^h(t) \int_t^{t+h} (B_m(s) - B_m(t))\, dB_j(s)\end{aligned}$$

for $i = 1, 2, \ldots, d$, to simulate $\hat{X}^h(t)$ for $t \in \{0, h, 2h, 3h, \ldots\}$ sequentially.
Under Lipschitz conditions and the linear growth condition on the drift and diffusion functions as in the Euler scheme [2], we have,

$$E[(X(1) - \hat{X}^h(1))^2] = O(h^2).$$

This implies $E|X(1) - \hat{X}^h(1))| = O(h)$. Under this scheme, if we want to achieve a MSE of order ϵ^2, it would take $O(\epsilon^{-3})$ units of computational effort, still worse than the $O(\epsilon^{-2})$ rate of convergence. The term $\int_t^{t+h} (B_m(s) - B_m(t))\, dB_j(s)$ in the recursion for the Milstein scheme in the multi-dimensional setting is called the Lévy area. One main obstacle in implementing the Milstein scheme and other higher order approximating schemes are that we do not know how to simulate the Lévy area and other higher order iterated integrals exactly for multi-dimensional SDEs. We believe our work in Chapter 4 provides some insights to this problem.

1.2.1 Multilevel Monte Carlo methods

Heinrich [21] introduced the idea of Multilevel Monte Carlo methods. Giles [22] applied the idea to sample path simulation, which substantially improve the rate of convergence of the above mentioned SDE estimation schemes (Euler scheme and Milstein scheme). The idea goes as follows. If we write the estimator as a telescoping sum of estimators across different grid size levels, assuming we can evaluate $E[f(X^{h_0})]$ explicitly,

$$\hat{Z} = E[f(X^{h_0})] + \sum_{l=1}^{L} \bar{Y}_{h_l}$$

where $\bar{Y}_{h_l} = \frac{1}{N_l}\sum_{k=1}^{N_l} Y_{h_l}(k)$ are independent for different l's, and $Y_{h_l}(k)$'s are i.i.d. samples distributed as $f(X^{h_l}) - f(X^{h_{l-1}})$.

then

1)
$$\text{Bias}(\hat{Z}) = \text{Bias}(f(X^{h_L}))$$
The bias is only determined by the finest grid size (largest l).

2)
$$\text{Var}(\hat{Z}) = \sum_{l=1}^{L} \frac{1}{N_L}\text{Var}(Y_{h_l})$$
If the variance decreases as the grid size decreases, we would allocate more computational budget N_l to smaller l's and less to larger l's to achieve an overall optimal budget allocation.

Specifically, under Euler scheme, if f is Lipschitz continuous, we have

$$\text{Bias}(\hat{Z}) = O(h_L^{1/2})$$

and

$$\begin{aligned}
\text{Var}(Y_{h_l}) &\leq E[(f(X^{h_l}) - f(X^{h_{l-1}}))^2] \\
&\leq 2E[(f(X^{h_l}) - f(X))^2] + 2E[(f(X) - f(X^{h_{l-1}}))^2] \\
&= O(h_l).
\end{aligned}$$

Then
$$\text{MSE} = O(h_L) + O(\sum_{l=1}^{L} \frac{1}{N_l} h_{l-1})$$

If we pick $h_l = 2^{-l}$, then to achieve a MSE of order ϵ^2, we need to set $L = O(\log(\epsilon))$. As for N_l's, by solving the optimization problem

$$\begin{aligned}
&\min \sum_{l=1}^{L} \frac{N_l}{h_l} \\
&\text{s.t. } \sum_{l=1}^{L} \frac{h_l}{N_l} \leq \epsilon^2
\end{aligned}$$

we have $N_l = O(\epsilon^{-2}\log(1/\epsilon)h_l)$, and the total computational cost is

$$\sum_{l=1}^{L}\frac{N_l}{h_l} = O(\epsilon^{-2}(\log(1/\epsilon))^2).$$

Recently, McLeish [23] and Rhee and Glynn [24] (around the same time, independently) developed a randomized multilevel method to eliminate the bias of the estimator completely. The idea again use the telescoping sum. For f Lipschitz continuous we write,

$$f(X) = f(X^{h_0}) + \sum_{l=1}^{\infty}(f(X^{h_l}) - f(X^{h_{l-1}})).$$

We next introduce a random variable N independent of everything else and write

$$\begin{aligned} Ef(X) &= Ef(X^{h_0}) + \sum_{l=1}^{\infty} E[f(X^{h_l}) - f(X^{h_{l-1}})]\frac{E[I(N \geq l)}{P(N \geq l)} \\ &= Ef(X^{h_0}) + \sum_{l=1}^{\infty} E\left[\frac{f(X^{h_l}) - f(X^{h_{l-1}})}{P(N \geq l)} I(N \geq l\right] \\ &= E\left[f(X^{h_0}) + \sum_{l=1}^{N}\frac{f(X^{h_l}) - f(X^{h_{l-1}})}{P(N \geq l)}\right] \end{aligned}$$

Then $\tilde{Z} := f(X^{h_0}) + \sum_{l=1}^{N}(f(X^{h_l}) - f(X^{h_{l-1}}))/P(N \geq l)$ is an unbiased estimator of $Ef(X)$.

If we simulate $f(X^{h_l}) - f(X^{h_{l-1}})$'s independent of each other, then

$$\begin{aligned} \mathrm{Var}(Z) &\leq E[\tilde{Z}^2] \\ &\leq E[f(X^{h_0})^2] + CE[f(X^{h_0})]|\mathrm{Bias}(X^{h_0})| \\ &\quad + \sum_{l=1}^{\infty}\frac{E[(f(X^{h_l}) - f(X^{h_{l-1}}))^2]}{P(N \geq l)} + C\sum_{l=1}^{\infty}\frac{\mathrm{Bias}(X^{h_{l-1}})^2}{P(N \geq l)} \end{aligned}$$

and the expected computation cost is

$$O(\sum_{l=1}^{\infty} h_l^{-1}P(N \geq l)).$$

When applying Euler scheme in this setting, to guarantee finite variance of the estimator, the expected computational cost is infinity. The Milstein scheme will

ensure both finite variance and finite mean computational cost, but as we pointed out before, for multi-dimensional diffusion process, we do not know how to simulate the Lévy area exactly.

1.2.2 ϵ-strong simulation and our contributions

ϵ-strong simulation of stochastic processes is a very recent research area. In Chapter 4 of this dissertation, we develop and analyze a simulation scheme that would allow us to construct a family of processes $X_\varepsilon = \{X_\varepsilon(t) : t \in [0,1]\}$, for each $\varepsilon \in (0,1)$, supported on a probability space $(\Omega, \mathcal{F}, P)$, and such that the following properties hold:

(T1) The process X_ε is piecewise constant, with finitely many discontinuities in $[0,1]$.

(T2) The process X_ε can be simulated exactly and, since it takes only finitely many values, its path can be fully stored.

(T3) We have that with P-probability *one*

$$\sup_{t\in[0,1]} ||X_\varepsilon(t) - X(t)||_\infty < \varepsilon. \tag{1.2}$$

(T4) For any $m > 1$ and $0 < \varepsilon_m < ... < \varepsilon_1 < 1$ we can simulate X_{ε_m} conditional on X_{ε_1},...,$X_{\varepsilon_{m-1}}$.

We refer to the family of procedures that achieve the construction of such family $\{X_\varepsilon : \varepsilon \in (0,1)\}$ as ε-strong simulation methods or *Tolerance-Enforced Simulation* (TES).

The paper of Chen and Huang [25] provides the construction of X_ε satisfying only (T1) to (T3). In particular, bound (4.2) is satisfied for a given fixed $\varepsilon_0 = \varepsilon > 0$, but it is not clear how to jointly simulate $\{X_{\varepsilon_m}\}_{m\geq 1}$ as $\varepsilon_m \searrow 0$ when applying the techniques in [25]. Chen and Huang [25] extended the applicability of an algorithm introduced by Beskos and Roberts [26]. The procedure of Beskos and Roberts [26], applicable only

to one dimensional diffusions, imposed strong boundedness assumptions on the drift coefficient and its derivative. The technique in [25] enabled the extension by using a localization technique; see also [27] for another extension. All these algorithms assume $\sigma(\cdot)$ constant. The assumption of a constant diffusion coefficient comes at basically no cost in the context of one dimensional diffusions, because one can always apply Lamperti (one-to-one) transformation to recast the simulation problem to one involving a diffusion with constant $\sigma(\cdot)$. However, such transformation cannot be generally applied in higher dimensions.

The major obstacle involved in developing exact sampling algorithms for multidimensional diffusions is the fact that $\sigma(\cdot)$ cannot be assumed to be constant. Moreover, even in the case of multidimensional diffusions with constant $\sigma(\cdot)$, the one dimensional algorithms developed so far can only be extended to the case in which the drift coefficient $\mu(\cdot)$ is the gradient of some function, that is, if $\mu(x) = \nabla v(x)$ for some $v(\cdot)$. The reason is that in this case one can represent the likelihood ratio $L(t)$, between the solution to (4.1) and Brownian motion (assuming $\sigma = I$ for simplicity) involving a Riemann integral of the form

$$\begin{aligned} L(t) &= \exp\left(\int_0^t \mu(X(s))\, dX(s) - \frac{1}{2}\int_0^t \|\mu(X(s))\|_2^2\, ds\right) \\ &= \frac{\exp(v(X(t)))}{\exp(v(X(0)))} \exp\left(-\frac{1}{2}\int_0^t \lambda(X(s))\, ds\right), \end{aligned}$$

for $\lambda(x) = \Delta v(x) + ||\nabla v(x)||_2^2$.

The fact that the stochastic integral can be transformed into a Riemann integral facilitates the execution of the acceptance/rejection method, because one can interpret (up to a constant and using localization as in [25]) the exponential of the integral of $\lambda(\cdot)$ as the probability that no arrivals occur in a Poisson process with a stochastic intensity. Such event (i.e. no arrivals) can be simulated by the thinning property of Poisson processes.

The paper of [28] extended the work of [26] in that their algorithms satisfy (T1) to (T4). The paper [29] not only provides an additional extension which allows to deal

with one dimensional SDEs with jumps, but also contains a comprehensive discussion on exact and ε-strong simulation for SDEs. Property (T4) in the definition of TES is desirable because it provides another approach to construct unbiased estimators for expectations of the form $Ef(X)$. In order to see this, let us assume for simplicity that $f(\cdot)$ is positive and Lipschitz continuous in the uniform norm with Lipschitz constant K. Then, let T be a positive random variable, independent of everything else, with a strictly positive density $g(\cdot)$ on $[0,\infty)$ and define

$$Z := I(f(X) > T)/g(T). \tag{1.3}$$

Observe that

$$EZ = E(E(Z|X)) = E\int_0^\infty I(f(X) > t)\frac{g(t)}{g(t)}dt = Ef(X),$$

so EZ is an unbiased estimator for $Ef(X)$. If Properties (T1) to (T4) hold, it is possible to simulate Z by noting that $f(X_\varepsilon) > T + K\varepsilon$ implies $f(X) > T$ and $f(X_\varepsilon) < T - K\varepsilon$ implies $f(X) \leq T$. Since (T4) allows to keep simulating as ε becomes smaller and T is independent of X_ε with a positive density $g(\cdot)$, then one eventually is able to simulate Z exactly. It is noted in [28] that the expected number of random variables required to simulate Z is typically infinite. The recent paper by Pollock et. al. [29] discusses via numerical examples the practical limitations of these types of estimators.

Our motivation in Chapter 4 of this dissertation is to investigate a novel approach using the theory of rough path that allows to study ε-strong simulation for multidimensional diffusions in substantial generality, without imposing the assumption that $\sigma(\cdot)$ is constant or that a Lamperti-type transformation can be applied. Given the previous discussion on the connections between exact sampling and ε-strong simulation, and the limitations of the current techniques, we believe that our results here provide an important step in the development of exact sampling algorithms for general multidimensional diffusions. Bayer et. al. [30] also use rough path analysis for

Monte Carlo estimation, but their focus is on connections to multilevel techniques and not on ε-strong simulation.

Finally, we note that in order to build our Tolerance-Enforced Simulation procedure we had to obtain new tools for the analysis of Lévy areas and associated conditional large deviations results conditional on the increments of Brownian motion. We believe that these results might be of independent interests.

1.3 The idea of record breakers

One common strategy used throughout the development of the first part of this dissertation is the use of "record breakers" to control the contribution of "future" information. In this section, we introduce how this idea arises in the simulation of infinite server queue and multi-dimensional SDEs.

1.3.1 Infinite server queue

We start with a point process description of the infinite server queue. In Figure 1.1, the point $Z_n = (A_n, V_n)$ denotes the n-th customer (counting backward in time), whose arrival time is A_n and service requirement is V_n, $n = 1, \ldots, 4$. One important feature of infinite server queue is that every customer starts service immediately upon arrival (there is no queue). If we project Z_n to the horizontal axis by drawing a -45^o line. The intersection of this line with the horizontal axis is the departure time of such n-th customer. We follow the technical tradition that an arrival at time t is counted in the system at time t (closed circle) while a departure at time t is not counted (open circle), so to make the process Càdlàg. We can also draw a vertical line at any $t \in \mathbb{R}$. The height of the intersection of the -45^o lines emanating from the points Z_n with $A_n \leq t$ and such vertical line, if positive, represents the corresponding remaining service time of that customer at time t.

We notice from the point process description in Figure 1.1 that customer $Z_n =$

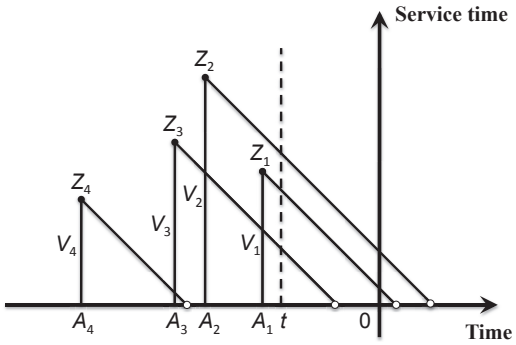


Figure 1.1: Point process description of an infinite server queue

$\{A_n, V_n\}$, with $V_n \leq |A_n|$ will have left the system by time 0. Thus if we can find a random number κ such that

$$V_n \leq |A_n| \text{ for all } n \geq \kappa,$$

then we can simulate the arrival stream backwards in time up to κ (i.e. $\{Z_n : 1 \leq n \leq \kappa\}$) to recover the state of the system at time zero.

The challenge here is that κ defined above depends on future customer information, i.e. $\{Z_n : n > \kappa\}$, and simulating this future information takes infinite amount of time. We overcome this difficulty by defining a sequence of "record breakers". Then instead of simulating all the customer information in the future we only ask future a yes/no question defined as "are there any more record breakers". In simulation, answering this yes/no question is equivalent to sampling a Bernoulli random variable with probability of success p, which equals to the probability that there are no more record breakers. If the Bernoulli trial is a success, then we are done. Otherwise we find the next record breaker, move to that time point and ask future the same yes/no question again. We repeat the above process until the Bernoulli trial returns a success. At that time, we know that there are no more record breakers in the future. We also locate the position (time) of all the record breakers. In Chapter 2 &3, we shall explain how to use this "record breaker" idea to simulate the infinite server queue.

1.3.2 Stochastic differential equations

The diffusion process

$$dX(t) = \mu(X(t))dt + \sigma(X(t))dB(t)$$

can be view as a mapping of the underlining Brownian motion. The general strategy would be if we can construction the Brownian path with certain error bound, then we can use some continuous mapping property to pass this error bound to the solution of the SDE. However, the diffusion mapping is not in general continuous under the uniform topology unless the drift term $\sigma(\cdot)$ is a constant. The way to solve this continuity problem is to lift the space up to the space of rough paths endowed with some suitable α-Hölder metric. The new mapping from the space of rough paths to the solution of SDE is then continuous. However, the space of rough paths contains not only the path itself, but also the iterated integral of the path. The theory of rough path applies to more general settings. In the specific case of Brownian motion, the space of rough path consists of the Brownian path $\{B(t), 0 \leq t \leq T\}$ and the Lévy area $\{\int_s^t (B_i(u) - B_i(s))\, dB_j(u), 0 \leq s < t \leq T\}$. Thus, in order to control the error of the approximation to the SDE in this case, we not only need to control the error of the approximated Brownian path but also the approximated Lévy area.

In Chapter 4, we use a wavelet construction of Brownian motion, known as the Lévy-Ciesielski Construction,

$$B\left(t\right) = W_0^0 \Lambda_0^0(t) + \sum_{n=1}^{\infty} \sum_{k=1}^{2^{n-1}} \left(W_k^n \Lambda_k^n(t)\right)$$

where $\Lambda_k^n(\cdot)$ is a sequence base functions.

Let $t_k^n = k/2^n$. We can also write the Lévy area as the following infinite sum.

$$\begin{aligned} & \int_{t_k^n}^{t_{k+1}^n} \left(B_i(u) - B_i(t_k^n)\right)\, dB_j(u) \\ = & \sum_{h=n+1}^{\infty} \sum_{l=1}^{2^{h-n-1}} [B_i(t^h_{2^{h-n}k+2l-1}) - B_i(t^h_{2^{h-n}k+2l-2})][B_j(t^h_{2^{h-n}k+2l}) - B_j(t^h_{2^{h-n}k+2l-1})]. \end{aligned}$$

We call n in the infinite sum "level n" and k in the sum from 0 to $2^n - 1$ the "kth term in level n". We then use the strategy of simulating the infinite sum up to a random but finite level N, such that the contribution of the higher level terms are under control.

The challenge here is that N is not a stopping time with respect to the filtration generated by $\{W_k^n : 0 \leq n \leq N, 0 \leq k \leq 2^n - 1\}$. We again use the idea of "record breakers", where by asking future yes/no questions we find all the "record breakers" and by knowing that there are no more "record breakers", the contribution of the terms in higher levels that are not simulated yet are well under control. See Chapter 4 for details in term of how to implement this idea to simulate the Brownian path and the Lévy areas with desired level of accuracy.

The main technicality in implementing the idea of "record breakers" is the simulation of the Bernoulli random variable with unknown/uncomputable probability of success, which is used to answer the yes/no question. To accomplish this, we use techniques from rare-event simulation.

Chapter 2

Sampling Point Processes on Stable Unbounded Regions and Perfect Sampling for Infinite Server Queues

Given a marked renewal point process (assuming that the marks are i.i.d.) we say that an unbounded region is stable if it contains finitely many points of the point process with probability one. In this chapter we provide algorithms that allow to sample these finitely many points efficiently. We explain how exact simulation of the steady-state measure valued state descriptor of the infinite server queue follows as a simple corollary of our algorithms. We provide numerical evidence supporting that our algorithms are not only theoretically sound but also practical. Finally, we also apply our results to gradient estimation of steady-state performance measures.

2.1 Problem formulation and main contributions

Let $N = \{N(t) : t \in (-\infty, \infty)\}$ be a two sided time stationary renewal point process. We write $\{A_n : n \in \mathbb{Z}_0\}$ for the times at which the process N jumps, where $\mathbb{Z}_0 = \mathbb{Z}\backslash\{0\}$ denotes the set of integers removing zero, and with $A_1 > 0 > A_{-1}$. For simplicity we assume that $A_n < A_{n+1}$ for every n. Further, we define $X_n = A_{n+1} - A_n$.

Now let $\{V_n : n \in \mathbb{Z}_0\}$ be a sequence of independent and identically distributed (i.i.d.) random variables (r.v.'s) which are independent of the process N. Define $Z_n = (A_n, V_n)$ and consider the marked point process $\mathcal{M} = \{Z_n : n \in \mathbb{Z}_0\}$ which forms a subset of $\mathbb{R}^2$. We say that a (Borel measurable) set $\mathcal{B}$ is stable if $|\mathcal{M} \cap \mathcal{B}| < \infty$ almost surely (where $|\mathcal{C}|$ is used to denote the cardinality of the set $\mathcal{C}$).

Under natural assumptions on the inter-arrival times underlying N and on the distribution of the V_n's (stated in Section 3.2) we propose and study a class of algorithms that allow to sample exactly (i.e. without any bias) a realization of the set $\mathcal{M} \cap \mathcal{B}$ for a large class of unbounded, stable sets $\mathcal{B}$.

Our method is based on a construction that is being used in [31]; see also [32] for related ideas. The method involves the technique of simulating the maximum of a negative drift random walk and the last passage time of independent and identically distributed random variables to an increasing boundary.

As an application of the class of algorithms that we study here, we provide a procedure that allows to sample from the steady-state measure valued descriptor of an infinite server queue without any bias (i.e. perfect sampling). Such a procedure, for instance, is obtained by considering the particular case in which $\mathcal{B}$ takes the form $\mathcal{B} = \{(t, v) : v > |t|, t \leq 0\}$. Given that point processes constitute a natural way of constructing queueing models in great generality, we believe that the class of algorithms that we propose here have the potential to be applicable to the design of exact sampling algorithms of more general queueing models.

We argue empirically that it is cheaper to run our exact sampling procedure to fully delete the initial bias than it is to do a burn-in period that reduces the bias to

a reasonable size, say 5%, when talking about, for instance, the steady-state queue length.

Finally, we apply our exact sampling algorithms for infinite server queues to perform steady-state sensitivity analysis. For instance, we consider quantities such as the derivative of the steady-state average remaining service time with respect to the arrival rate or service rate. These quantities are of great interests in stochastic optimization via simulation.

So, in summary, our contributions are as follows:

i) We provide the first exact sampling algorithm for stationary marked renewal processes on unbounded and stable sets, see Section 3.2.

ii) As a corollary of i) we explain how to obtain an exact sampling algorithm for the steady-state measure valued descriptor of the infinite server queue. We also show empirically that this algorithm is *practical* in the sense of being both easy to code and fast to run, see Section 3.1.2.

iii) Finally, we provide new procedures for the sensitivity analysis of steady-state performance measures of the infinite server queue, see Section 2.4.

2.2 Sampling from stable unbounded regions

We start by discussing the assumptions behind our development.

Assumptions:

A1) Assume that $E\left|V_n\right|^{1/\alpha} < \infty$ for some $\alpha > 0$, we also write $F\left(\cdot\right) = P\left(V_n \leq \cdot\right)$ for the cumulative distribution function (CDF) of V_n and put $\overline{F}\left(\cdot\right) = 1 - F\left(\cdot\right)$ for the tail CDF.

A2) We assume that $F\left(\cdot\right)$ is known and easily accessible either in closed form or via efficient numerical procedures. Moreover, we can simulate V_n conditional

on $V_n \in [a,b]$ with $P(V_n \in [a,b]) > 0$. Finally we can find $u(k)$ such that $u(k) \geq \int_k^\infty P(|V_1|^{1/\alpha} > \nu) d\nu$ and $u(k) \to 0$ as $k \to \infty$.

A3) Recall that $X_n = A_{n+1} - A_n > 0$. Define $\psi(\theta) = \log E \exp(\theta X_n)$ and assume that there exists $\delta > 0$ such that $\psi(\delta) < \infty$. Finally, let us write $\mu = EX_n$.

A4) Define $G(\cdot) = P(X_n \leq \cdot)$ and $\overline{G}(\cdot) = 1 - G(\cdot)$. Suppose that $G(\cdot)$ is known and that it is possible to simulate from $G_{eq}(\cdot) := \mu^{-1} \int_\cdot^\infty \overline{G}(t)\, dt$. Moreover, let $G_\theta(\cdot) = E \exp(\theta X_n - \psi(\theta)) I(X_n \leq \cdot)$ be the associated exponentially tilted distribution with parameter θ for $\psi(\theta) < \infty$. We assume that we can simulate from $G_\theta(\cdot)$.

Consider the class of sets $\mathcal{B} \subset \mathbb{R}^2$ that are Borel measurable and such that

$$\mathcal{B} \subset \mathcal{C}_\alpha = \{(t,v) : |v| \geq |t|^\alpha\}.$$

Our goal in this section is to develop an algorithm that allows to sample without any bias the random set $\mathcal{M} \cap \mathcal{C}_\alpha$, and therefore $\mathcal{M} \cap \mathcal{B}$. We will discuss extensions that follow immediately from our formulation at the end of this section. Figure 2.1 illustrates the different shapes that the set $\mathcal{C}_\alpha$ can take depending on the values of $\alpha > 0$.

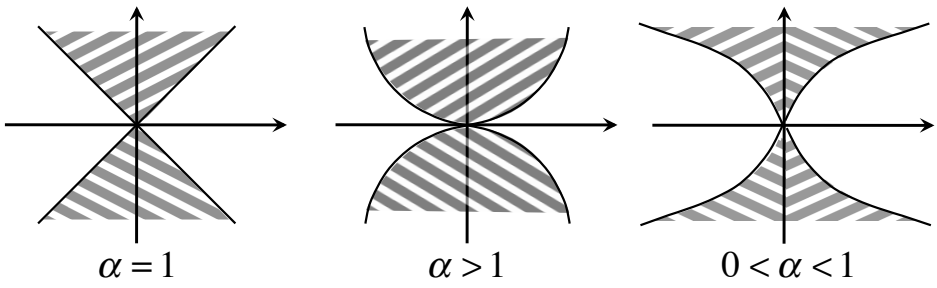


Figure 2.1: The area of $\mathcal{C}_\alpha$. The horizontal axis corresponds to the t coordinate while the vertical axis represents the v coordinate

We now proceed to explain our construction. As the stationary renewal point process is time reversible, starting at 0 the distribution of the forward process $\{Z_n : n > 0\}$ and the backward process $\{Z_n : n < 0\}$ are the same. In what follows we limit our discussion to the construction of the forward process and the simulation of the backward process is completely analogous.

Let $\epsilon \in (0, \mu)$. Consider any random time κ, finite with probability one but large enough such that

$$A_{n+1} \geq n(\mu - \epsilon) \text{ and } |V_{n+1}| \leq (n(\mu - \epsilon))^{\alpha}$$

for all $n \geq \kappa$.

If such random time κ is well defined, we only need to simulate the stationary process up to κ to get a sample from the unbounded region.

Proposition 2.2.1 *The random time κ defined above exists and it is finite with probability one.*

Proof. By Chebyshev's inequality,

$$P(A_{n+1} < n(\mu - \epsilon)) \leq E[\exp(\theta(n(\mu - \epsilon) - A_{n+1}))] \leq \exp(-n(-\theta(\mu - \epsilon) - \psi(-\theta)))$$

for any $\theta \geq 0$.

Let

$$I(-\epsilon) = \max_{\theta \geq 0}\{-\theta(\mu - \epsilon) - \psi(-\theta)\}$$

As $\psi(0) = 0$, $\psi'(0) = \mu$ and $\psi''(0) = Var(X) > 0$, $I(-\epsilon) > 0$. Then

$$P(A_{n+1} < n(\mu - \epsilon)) \leq \exp(-nI(-\epsilon))$$

and

$$\sum_{n=1}^{\infty} P(A_{n+1} < n(\mu - \epsilon)) \leq \frac{\exp(-I(-\epsilon))}{1 - \exp(-I(-\epsilon))} < \infty$$

By Borel-Cantelli lemma, $\{A_{n+1} \geq n(\mu-\epsilon)\}$ eventually almost surely.
Similarly and independently we have

$$\begin{aligned} & \sum_{n=1}^{\infty} P(|V_{n+1}| > (n(\mu-\epsilon))^{\alpha}) \\ = & \sum_{n=1}^{\infty} P(|V_1|^{1/\alpha} > n(\mu-\epsilon)) \\ \leq & \frac{1}{\mu-\epsilon}\int_0^{\infty} P(|V_1|^{1/\alpha} > \nu)d\nu < \infty \end{aligned}$$

Thus, again by Borel-Cantelli lemma, $\{|V_{n+1}| \leq (n(\mu-\epsilon))^{\alpha}\}$ eventually almost surely.
Therefore, $P(\kappa < \infty) = 1$ □

As $\{A_n : n \geq 1\}$ and $\{V_n : n \geq 1\}$ are independent of each other, we consider the following construction. Let $\kappa(A)$ be a random time satisfying that $A_{n+1} \geq n(\mu-\epsilon)$ for $n \geq \kappa(A)$, and $\kappa(V)$ be a random time satisfying that $V_{n+1} \leq n(\mu-\epsilon)$ for $n \geq \kappa(V)$. Clearly $\kappa(A)$ and $\kappa(V)$ are *not* stopping times and this makes the simulation of these times challenging. However, we will explain how to sample these times and then we can set $\kappa = \max\{\kappa(A), \kappa(V)\}$. Our construction will allow us to simulate $\{A_n : n \geq 1\}$ and $\{V_n : n \geq 1\}$ separately.

2.2.1 Simulation of $\{A_k : 1 \leq k \leq \max\{n, \kappa(A)\} + 1\}$

In this subsection we will introduce a method to simulate $\kappa(A)$ together with $\{A_k : k \geq 1\}$.

First, define A_1 according to the distribution $G_{eq}(\cdot)$. Sampling A_1 can be done according to A4).

Now, observe that $A_{n+1} = A_1 + X_1 + ... + X_n$ and define

$$\tilde{S}_n = n(\mu-\epsilon) - (A_{n+1} - A_1) = \sum_{i=1}^{n} Y_i,$$

where $Y_i = (\mu-\epsilon) - X_i$. Note that the Y_i's are i.i.d. with $EY_i = -\epsilon$. If we set $\tilde{S}_0 = 0$, then $\{\tilde{S}_n : n \geq 0\}$ is a random walk with negative drift. We are interested in sampling up to the *last time* n at which $\tilde{S}_n > 0$.

We define the following sequence of random times:

$$\Delta_1 = 0,\ \Gamma_1 = \inf\{n \geq \Delta_1 : \tilde{S}_n - \tilde{S}_{\Delta_1} > 0\},$$

and for $j \geq 2$

$$\begin{aligned}\Delta_j &= \inf\{n \geq \Gamma_{j-1}\mathbf{1}\{\Gamma_{j-1} < \infty\} \vee \Delta_{j-1} : \tilde{S}_n \leq 0\},\\ \Gamma_j &= \inf\{n \geq \Delta_j : \tilde{S}_n - \tilde{S}_{\Delta_j} > 0\}.\end{aligned}$$

Now, let $\gamma = \inf\{j \geq 1 : \Gamma_j = \infty\}$ and note that $\Delta_{\gamma+1} = \Delta_\gamma$ and that $\tilde{S}_n \leq 0$ for $n \geq \Delta_\gamma$, which in particular implies that $A_{n+1} \geq n(\mu - \epsilon)$ for $n \geq \Delta_\gamma$. Therefore, we have that $\Delta_\gamma = \kappa\,(A)$.

In what follows we will explain how to simulate the Δ_j's and Γ_j's sequentially and jointly with the underlying random walk until time Δ_γ. One important observation is that for every $j \geq 1$, $\Delta_j < \infty$ almost surely by the strong law of large numbers.

Let us write $\mathcal{F}_n = \sigma\{Y_1, Y_2, ..., Y_n\}$ for the σ-field generated by the Y_j's up to time n. Let $\xi \geq 0$ and define

$$T_\xi := \inf\{n \geq 0 : \tilde{S}_n > \xi\},$$

then by the strong Markov property we have that for $j \leq \gamma$,

$$P(\Gamma_j = \infty|\mathcal{F}_{\Delta_j}) = P(\Gamma_j = \infty|\tilde{S}_{\Delta_j}) = P(T_0 = \infty) > 0,$$

where we use $P\,(\cdot)$ to denote the nominal probability measure under which $\tilde{S}_0 = 0$.

It is important then to note that

$$P\,(\gamma = k) = P\,(T_0 < \infty)^{k-1}\,P\,(T_0 = \infty)$$

for $k \geq 1$. In other words, γ is geometrically distributed. The procedure that we have in mind is to simulate Δ_γ in time intervals, and the number of time intervals is precisely γ.

Let $\psi_Y(\theta) = \log E \exp(\theta Y_i)$. As the moment generating function of X_i is finite in a neighborhood of the zero, $\psi_Y(\cdot)$ is also finite in a neighborhood of zero and

$EY_i = \psi'_Y(0) = -\epsilon$, $\mathrm{Var}(Y_i) = \psi''_Y(0) > 0$. Then by the convexity of $\psi_Y(\cdot)$, one can always select $\epsilon > 0$ sufficiently small so that there exists $\eta > 0$ with $\psi_Y(\eta) = 0$ and $\psi'_Y(\eta) > 0$. The root η allows us to define a new measure P_η based on exponential tilting so that

$$\frac{dP_\eta}{dP}(Y_i) = \exp(\eta Y_i).$$

Moreover, under P_η, $\tilde{S}_n$ is random walk with positive drift equal to $\psi'_Y(\eta)$ ([1] P. 365). Therefore $P_\eta(T_0 < \infty) = 1$ and $P(T_0 < \infty) = E_\eta(\exp(-\eta \tilde{S}_{T_0}))$. More generally, $P_\eta(T_\xi < \infty) = 1$ and

$$q(\xi) := P(T_\xi < \infty) = E_\eta(\exp(-\eta \tilde{S}_{T_\xi}))$$

for each $\xi \geq 0$. Based on the above analysis we now introduce a convenient representation to simulate a Bernoulli random variable $J(\xi)$ with parameter $q(\xi)$ namely,

$$J(\xi) = I(U \leq \exp(-\eta \tilde{S}_{T_\xi})). \tag{2.1}$$

where U is a uniform random variable independent of everything else under P_η.

Identity (2.1) provides the basis for an implementable algorithm to simulate a Bernoulli with success probability $q(\xi)$. Sampling $\{\tilde{S}_1, ..., \tilde{S}_{T_0}\}$ conditional on $T_0 < \infty$, as we shall explain now, corresponds to basically the same procedure. First, let us write $P^*(\cdot) = P(\cdot | T_0 < \infty)$. The following result provides an expression for the likelihood ratio between P^* and P_η.

Lemma 2.2.2 *We have that*

$$\frac{dP^*}{dP_\eta}(\tilde{S}_1, ..., \tilde{S}_{T_0}) = \frac{\exp(-\eta \tilde{S}_{T_0})}{P(T_0 < \infty)} \leq \frac{1}{P(T_0 < \infty)}.$$

Proof.

$$\begin{aligned} & P(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_0} \in H_{T_0} | T_0 < \infty) \\ = & \frac{P(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_0} \in H_{T_0}, T_0 < \infty)}{P(T_0 < \infty)} \\ = & \frac{E_\eta[\exp(-\eta \tilde{S}_{T_0}) I(\tilde{S}_0 \in H_0, ..., \tilde{S}_{T_0} \in H_{T_0})]}{P(T_0 < \infty)}. \end{aligned}$$

□

The previous lemma provides the basis for a simple acceptance / rejection procedure to simulate $\{\tilde{S}_1, ..., \tilde{S}_{T_0}\}$ conditional on $T_0 < \infty$. More precisely, we propose $(\tilde{S}_1, ..., \tilde{S}_{T_0})$ from $P_\eta(\cdot)$. Then one generates a uniform random variable U independent of everything else and accept the proposal if

$$U \leq \frac{1}{1/P(T_0 < \infty)} \times \frac{dP^*}{dP_\eta}(\tilde{S}_1, ..., \tilde{S}_{T_0}) = \exp(-\eta \tilde{S}_{T_0}).$$

This criterion coincides with $J(0)$ according to (2.1). So, the procedure above simultaneously obtains both a Bernoulli r.v. $J(0)$ with parameter $q(0)$, and the corresponding path $\{\tilde{S}_1, ..., \tilde{S}_{T_0}\}$ conditional on $T_0 < \infty$.

Algorithm 2.1 (Outputs $(\tilde{S}_0, ..., \tilde{S}_{\Delta_\gamma})$)

S0. Set $K = 0$, and $S_0 = 0$

S1. Simulate $(\tilde{S}_1, ..., \tilde{S}_{T_0})$ from P_η and compute $J := J(0)$ according to (2.1).

S2. If $J = 1$, then let $S_{K+j} = \tilde{S}_j$ for $j = 1, ..., T_0$ and update $K \longleftarrow K + T_0$. Then, go back to S1.
Otherwise, $J = 0$ (i.e. $\Delta_\gamma = K$), stop and output $(S_0, ..., S_K)$

Remark: We will show in Section 3.3.1 of Chapter 3 that the expected number of times we need to repeat Step 1 does not change with the system scale (i.e. the arrival rate).

We noted earlier that $\Delta_\gamma = \kappa(A)$ and Algorithm 1 together with the initial procedure to sample A_1 allows us to simulate $(A_{j+1} : 0 \leq j \leq \kappa(A))$, and we know that $A_{n+1} \geq n(\mu - \epsilon)$ for $n \geq \kappa(A)$. We need to simulate A_{n+1} for $n \leq \kappa = \max\{\kappa(A), \kappa(V)\}$, and $\kappa(V)$ is independent of $\kappa(A)$. So, there might be cases for which we will have to sample A_{n+1} for $n > \kappa(A)$. Since $A_{n+1} = A_1 - \tilde{S}_n + n(\mu - \epsilon)$ it suffices to explain how to simulate $\tilde{S}_n$ for $n > \Delta_\gamma$. In turn, it suffices to explain

how to simulate $(\tilde{S}_n : n \geq 0)$ with $\tilde{S}_0 = 0$ conditional on $T_0 = \infty$. We will once again apply an acceptance/rejection procedure but this time we will use the original (nominal) distribution as the proposal distribution. Define

$$P'(\cdot) = P(\cdot|T_0 = \infty).$$

The following result provides an expression for the likelihood ratio between P' and P.

Lemma 2.2.3 *We have that*

$$\frac{dP'}{dP}(\tilde{S}_1, ..., \tilde{S}_l) = \frac{I(T_0 > l)(1 - q(-\tilde{S}_l))}{P(T_0 = \infty)} \leq \frac{1}{P(T_0 = \infty)}.$$

Proof.

$$\begin{aligned} & P(\tilde{S}_1 \in H_1, ..., \tilde{S}_l \in H_l | T_0 = \infty) \\ = & \frac{P(\tilde{S}_1 \in H_1, ...\tilde{S}_l \in H_l, T_0 = \infty)}{P(T_0 = \infty)} \\ = & \frac{E[I(\tilde{S}_1 \in H_1, ..., \tilde{S}_l \in H_l) I(T_0 > l) P(T_0 = \infty | \tilde{S}_0, ..., \tilde{S}_l)]}{P(T_0 = \infty)}. \end{aligned}$$

The result then follows from the strong Markov property and homogeneity of the random walk. □

We are in good shape now to apply acceptance/rejection to sample from P'. The previous lemma indicates that to sample $\{\tilde{S}_0, ..., \tilde{S}_l\}$ given $T_0 = \infty$ we can propose from the original (nominal) distribution and accept with probability $q(-\tilde{S}_l)$ as long as $\tilde{S}_j \leq 0$ for all $0 \leq j \leq l$. So, in order to perform the acceptance test we need to sample a Bernoulli with parameter $q(-\tilde{S}_l)$, but this is easily done using identity (2.1). Thus we obtain the following procedure.

Algorithm 2.2 (Given $n \geq 0$ outputs $\{A_1, A_2, ..., A_{\max\{n, \kappa(A)\}+1}\}$)

S1. Run Algorithm 2.1 and obtain $\{S_0, S_1, ..., S_K\}$.

S2. If $K = \kappa(A) \geq n$, jump to S6. Otherwise, $K < n$, let $l = n - K \geq 1$.

S3. Simulate $\{\tilde{S}_0, \tilde{S}_1, ..., \tilde{S}_l\}$ from the original (nominal) distribution with $\tilde{S}_0 = 0$.

S4. If $\tilde{S}_j \leq 0$ for all $0 \leq j \leq l$ then sample a Bernoulli $J(-\tilde{S}_l)$ with parameter $q(-\tilde{S}_l)$ using (2.1) and continue to S5. Otherwise (i.e. $\tilde{S}_j > 0$ for some $1 \leq j \leq l$) go back to S3.

S5. If $J(-\tilde{S}_l) = 1$, go back to S3. Otherwise, $J(-\tilde{S}_l) = 0$, let $S_{K+i} = S_K + \tilde{S}_i$ for $i = 1, 2, ..., l$

S6. Let $m = \max\{n, \kappa(A)\}$. Simulate A_1 with CDF $G_{eq}(\cdot) = \mu^{-1} \int_{\cdot}^{\infty} \bar{G}(t)dt$. Set $A_{n+1} = A_1 - S_n + n(\mu - \epsilon)$ for $n = 1, ..., m$. Output $\{A_1, ..., A_{m+1}\}$.

2.2.2 Simulation of $\{V_n : 1 \leq n \leq \kappa(V) + 1\}$

In this section we will introduce a method to simulate $\kappa(V)$ together with the $\{V_n : n \geq 1\}$.

Let $p(n) = P(|V_1| > (n(\mu - \epsilon))^\alpha)$. We define $\Upsilon_0 = 0$ and $\Upsilon_i = \inf\{n > \Upsilon_{i-1} : |V_{n+1}| > (n(\mu - \epsilon))^\alpha\}$ for $i = 1, 2, ...$ We also define two independent sequences of random variables, $\{\hat{V}_{n+1} : n \geq 1\}$, and $\{\bar{V}_{n+1} : n \geq 1\}$ as follows. The elements in each sequence are i.i.d., $\hat{V}_{n+1}$ is distributed as V_{n+1} conditional on $|V_{n+1}| > (n(\mu - \epsilon))^\alpha$, and $\bar{V}_{n+1}$ follows the distribution of V_{n+1} conditional on $|V_{n+1}| \leq (n(\mu - \epsilon))^\alpha$. We simulate V_1 following its nominal distribution independent of everything else.

Let $\sigma = \inf\{i \geq 0 : \Upsilon_i = \infty\}$. Then $V_{n+1} \leq (n(\mu - \epsilon))^\alpha$ for $n \geq \Upsilon_{\sigma-1} + 1$. We next introduce a method to sample $\Upsilon_1, \Upsilon_2, ...$ sequentially and jointly with the V_n's up until $\Upsilon_{\sigma-1}$.

The following lemma provides the basis to guarantee the termination of our procedure.

Lemma 2.2.4 *If* $E|V_1|^{1/\alpha} < \infty$, *then*

$$P(\Upsilon_1 = \infty) = \prod_{i=1}^{\infty}(1 - p(i)) \geq \exp(-2E|V_1|^{1/\alpha}/(\mu - \epsilon)) > 0,$$

consequently $E\sigma \leq \exp(2E|V|^{1/\alpha}/(\mu - \epsilon)) < \infty$.

Remark: The bound on $E\sigma$ can be improved. This improvement is important for the theoretical asymptotic analysis of GI/GI/∞ application, see Section 3.3.1 in Chapter 3 for details.

Proof.

$$\begin{aligned} P(\Upsilon_1 = \infty) &= \prod_{n=1}^{\infty}(1 - p(n)) \\ &\geq \prod_{n=1}^{\infty}\exp(-2p(n)) \\ &\geq \exp(-\frac{2}{\mu - \epsilon}\int_0^{\infty} P(|V_1|^{1/\alpha} > \nu)d\nu) \\ &= \exp(-\frac{2E|V_1|^{1/\alpha}}{\mu - \epsilon}). \end{aligned}$$

For $i = 2, 3, ...$ conditional on $\Upsilon(i-1) = k$:

$$\begin{aligned} & P(\Upsilon_i = \infty | \Upsilon_{i-1} = k) \\ = & \prod_{n=k+1}^{\infty}(1 - p(n)) \\ \geq & \exp(-\frac{2\int_k^{\infty} P(|V_1|^{1/\alpha} > \nu)d\nu}{\mu - \epsilon} \geq \exp(-\frac{2E|V_1|^{1/\alpha}}{\mu - \epsilon}) \end{aligned}$$

Thus σ is stochastically dominated by a geometric random variable with parameter $p = \exp(-2E|V_1|^{1/\alpha}/(\mu - \epsilon))$, the result then follows. □

Notice that

$$\begin{aligned} & \prod_{i=k+1}^{l}(1 - p(i)) \\ \geq & P(\Upsilon_i = \infty | \Upsilon_{i-1} = k) \\ \geq & \prod_{i=k+1}^{l}(1 - p(i)) \times \exp(-\frac{2\int_l^{\infty} P(|V_1|^{1/\alpha} > \nu)d\nu}{\mu - \epsilon}) \end{aligned} \tag{2.2}$$

for $l \geq k+1$.

Thus if we are simulating $I \sim \text{Bernoulli}(r_i)$ with $r_i := P(\Upsilon_i = \infty | \Upsilon_{i-1})$, then with probability one we can check whether $U \leq P(\Upsilon_i = \infty | \Upsilon_{i-1})$ for $U \sim \text{Unif}[0,1]$ by making l sufficiently large without calculating the infinite product in the definition of $P(\Upsilon_i = \infty | \Upsilon_{i-1})$.

On the other hand, if we define $\prod_{j=1}^{0}(1-p(j)) := 1$, then

$$P(\Upsilon_1 = n | \Upsilon_1 < \infty) = p(n) \frac{\prod_{j=1}^{n-1}(1-p(j))}{P(\Upsilon_1 < \infty)} \leq p(n) \frac{1}{P(\Upsilon_1 < \infty)}.$$

Consider a random variable N with the following probability density function

$$P(N = n) = cp(n)$$

for $n = 1, 2, ...$, where $c = (\sum_{n=1}^{\infty} p(n))^{-1}$. Then $P(\Upsilon_1 = n | \Upsilon_1 < \infty)/P(N = n) \leq 1/(cP(\Upsilon_1 < \infty))$.

So we can simulate Υ_1 given $\Upsilon_1 < \infty$ using acceptance / rejection with N as the proposal random variable. Generalizing the idea to Υ_i, we can obtain the following algorithm

Algorithm 2.3 (Given $\Upsilon_{i-1} = k$, outputs Υ_i conditional on $\Upsilon_i < \infty$)

S1. Let $c = (\sum_{n=k+1}^{\infty} p(n))^{-1}$. Simulate N with probability density function $P(N = n) = cp(n)$ for $n = k+1, k+2, ...$

S2. Simulate $U \sim \text{Unif}[0,1]$ independently. If $U \leq \prod_{j=k+1}^{N-1}(1-p(j))$, set $\Upsilon_i = N$ and stop. Otherwise go back to S1

We conclude this section with our procedure to simulate $\{V_1, V_2, ...V_{\kappa(V)+1}\}$.

Algorithm 2.4 (Outputs $\{V_1, V_2, ...V_{\kappa(V)+1}\}$)

S0. Set $\Upsilon_0 = 0$, $i = 1$. Simulate V_1 from its nominal distribution.

S1. Simulate $I \sim \text{Bernoulli}(r_i)$ with $r_i := P(\Upsilon_i = \infty | \Upsilon_{i-1})$ (see (2.2)).

S2. If $I = 1$, set $\kappa(V) = \Upsilon_{i-1} + 1$. Simulate $V_{\kappa(V)+1}$ by sampling from $\bar{V}_{\kappa(V)+1}$ and stop. Otherwise $I = 0$, sample Υ_i conditional on $\Upsilon_i < \infty$ and the value of Υ_{i-1} using Algorithm 2.3. Simulate the process between $\Upsilon_{i-1} + 2$ and $\Upsilon_i + 1$ by sampling from $\bar{V}_n$ for $\Upsilon_{i-1} + 2 \leq n \leq \Upsilon_i$ and $\hat{V}_n$ for $n = \Upsilon_i + 1$. Set $i = i + 1$ and then go back to S1.

2.3 Application to the infinite-server queue

As a direct application of the ideas discussed in the previous section we study steady-state simulation for the infinite server queue. The following diagram indicates how to construct the steady-state measure valued descriptor assuming that we can sample all the points inside the set

$$\mathcal{C} = \{(t, v) : v \geq |t|, t \leq 0\}.$$

Let $Q(t, y)$ denote the number of people in the system at time t with residual service time strictly greater than y and $E(t)$ denote the time elapsed since the previous arrival at time t (i.e. $E(\cdot)$ is the age process associated with $N(\cdot)$). Figure 2.2 below depicts the region $\mathcal{C}$. Every point in $|\mathcal{M} \cap \mathcal{C}|$ is projected to the vertical line at time zero by drawing a -45^0 line. The final position in the vertical line if positive, represents the corresponding remaining service time. Since the underlying point process is time stationary, the whole configuration of points obtained by this procedure at time zero is a snap shot of the steady-state distribution of the infinite server queue.

2.3.1 Algorithm for the infinite server queue

As depicted in Figure 2.2 after projecting into the vertical line at $t = 0$, we obtain the stationary remaining service requirements of the customers at time zero. We shall use $R_1, R_2, ..., R_{Q(0,0)}$ to denote the remaining service times. The labeling is arbitrary

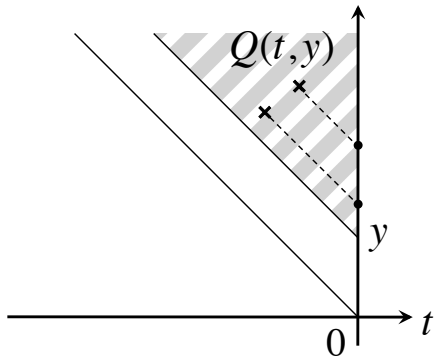


Figure 2.2: The points lies in the shaded area correspond to people who are still in the system at time 0 with remaining service time greater than y

although we will assign smaller indexes to customers that have spent less time in the system. Our algorithm proceeds as follows.

Algorithm 2.5 (Outputs $\{R_1, R_2, ..., R_{Q(0,0)}\}$ and $E(0)$)

S1. Use Algorithm 2.4 to simulate the $\{V_n, 1 \leq n \leq \kappa(V) + 1\}$.

S2. Use Algorithm 2.2 to simulate the $\{A_1, A_2, ..., A_{\max\{\kappa(V),\kappa(A)\}+1}\}$.

S3. Set $\kappa = \max(\kappa(V), \kappa(A))$. If $\kappa > \kappa(V)$, simulate V_n by sampling from $\bar{V}_n$ for $n = \kappa(V) + 2, ..., \kappa + 1$.

S4. Set $q = 0$, $i = 0$ and repeat the following procedure until $i = \kappa$:
set $i = i + 1$*; if* $V_i > A_i$*, set* $q = q + 1$ *and* $R_q = V_i - A_i$.
Output $\{R_1, R_2, ...R_q\}$ and A_1.

2.3.2 Numerical results

Let $Y = \{Y(t) : t \geq 0\}$ be a continuous time Markov process on the state space Ω and f is a real-valued function defined on Ω. The ergodic theorem guarantees in great

generality (assuming a unique stationary distribution $\pi(\cdot)$) that

$$\frac{1}{t}\int_0^t f(Y(s))ds \to \int_\Omega f(y)\pi(dy)$$

as $t \to \infty$ almost surely for every positive, measurable function $f(\cdot)$. In the setting of the infinite server queue such a stationary distribution exists if $EV_n < \infty$ and $EX_n < \infty$. The most natural estimator for $E_\pi f(Y) := \int_\Omega f(y)\pi(dy)$ is therefore

$$\Phi(t, Y(0)) := \frac{1}{t}\int_0^t f(Y(s))ds,$$

where $Y(0)$ is the initial state. The estimator $\Phi(t, Y(0))$ is generally biased unless $Y(0)$ is sampled from the stationary distribution $\pi(\cdot)$ ([2] P. 97). Our algorithm has the obvious advantage of removing the initial transient.

In what follows we conduct some simulation experiment to evaluate the practical performance of our algorithm. The idea is to fix a reasonable tolerance error, say 10%, for a given performance measure. Then we want to empirically find how large a burn-in period one would need in practice to reduce the initial transient bias to about 10%. In order to effectively quantify the error we select a class of systems for which $\pi(\cdot)$ can be explicitly evaluated.

We consider an infinite server queue with Poisson arrivals and Lognormal service times. As we are interested in the efficiency of our algorithm for relatively large systems, we set the arrival rate $\lambda = 100$ and the service time $V_n \sim \text{Lognormal}(-0.25, 0.5)$ (i.e. V_n has the same distribution as $\exp(-.25 + .5 \times N(0,1))$, where $N(0,1)$ denotes a standard Gaussian random variable).

Let $Y(t) = (Q(t,\cdot), E(t)) \in \mathcal{D}[0,\infty) \times \mathbb{R}_+$, then $Y(t)$ is a Markovian measure valued descriptor of the infinite server queue (of course in the Poisson arrival case one does not need to keep track of $A(\cdot)$).

We first compare the performance of our algorithm to the burn-in period defined as the period needed to reduce the initial transient as indicated earlier. Let $f(Y(t)) = Q(t,0)$, i.e. the number of people in the system at time t. We measure the computation effort of the algorithm in terms of the number of arrivals (we call

this the number of steps) simulated. Given $\epsilon > 0$ we let $n(\epsilon)$ denote the minimum number of steps required so that $|E\Phi(A_{n(\epsilon)}, (\phi, 0)) - E_\pi Q(0,0)|/E_\pi Q(0,0) \leq \epsilon$, where $(\phi, 0)$ denotes a system that starts empty with $E(0) = 0$ (recall that $E(\cdot)$ is the age process associated with $N(\cdot)$, i.e. when $E(0) = x$, A_1 is distributed as X_n conditional on $X_n > x$). Table 2.1 shows the relation between ϵ and $n(\epsilon)$, obtained empirically based on the average of 10^4 independent replications

Table 2.1: Bias of $\Phi(S_{n(\epsilon)})$

ϵ	$n(\epsilon)$	computer time (s)
10.26%	6×10^2	0.0310
5.71%	1×10^3	0.0382
1.17%	5×10^3	0.1367

Compared to the results in Table 2.1, our algorithm is unbiased. The average number of steps involved is $n = 592.6369$ based on the average of 10^4 independent replications and the average computer time needed for a single replication is 0.0249 s.

In addition, in Table 2.2 we compare the performance of the estimators $\Phi(A_n, (\phi, 0))$ and $\Phi(A_{n'}, (Q(0,\cdot), A_1))$, where $Q(0,\cdot)$ and A_1 are sampled according to Algorithm 5. n and n' are calibrated so that the computation budget is basically the same in both estimators. Under our procedure, $E\kappa$, the average number of arrivals required to terminate is approximately equal to 600. So for instance, the first row in Table 2 corresponds to $n = 10^4$. This means that $n' \approx 9.4 \times 10^3 = 10^4 - 600$. The true value of $E_\pi Q(0,0)$ is 88.2497. The sample mean and sample standard deviation are calculated using the method of Batch means. The result in Table 2.2 shows that our mixed method performs better than the batch means with relatively small computation budget, while with large budget, the two methods are about the same.

Table 2.2: Simulation result with different initial states.

	$(\phi, 0)$		$(Q(0,\cdot), A_1)$	
n	Sample Mean	Sample Std	Sample Mean	Sample Std
1×10^4	86.1274	1.0104	88.1713	0.6018
5×10^4	89.0893	0.4587	88.2956	0.3770
1×10^5	88.5151	0.3531	88.1270	0.2976
5×10^5	88.3022	0.1481	88.3581	0.1402

2.4 Application to sensitivity analysis of the infinite-server queue

In this section, we apply our algorithm to sensitivity analysis of the infinite server queue. We consider a sequence of systems indexed by (λ, ν), $\lambda > 0$, $\nu > 0$. Given (λ, ν), the interarrival times are multiplied by $1/\lambda$, obtaining X_n/λ for all n, and the service times are multiplied by $1/\nu$, thus we have V_n/ν for all n. We assume that $EV_n < \infty$ and $EX_n < \infty$. We will use the notation $Q_{\lambda,\nu}(\cdot)$ to denote the infinite server queue descriptor for the (λ, ν)-system. Our strategy rests on the application of Infinitesimal Perturbation Analysis (IPA), see for instance [33] P. 386. We assume here that the interarrival times have a continuous distribution.

We illustrate the methodology by computing the sensitivity of the steady-state average remaining service time, which we denote by $E_\pi \bar{R}(\lambda, \nu)$; namely,

$$E_\pi \bar{R}(\lambda, \nu) = E_\pi \frac{1}{Q_{\lambda,\nu}(0,0)} \int_0^\infty y Q_{\lambda,\nu}(0, dy).$$

We also consider

$$E_\pi R^\infty(\lambda, \nu) = E_\pi(\inf\{y \geq 0 : Q_{\lambda,\nu}(0, y) = 0\}),$$

in words, the steady-state maximum remaining service time. In order to apply IPA we need to define a few quantities.

First, let us define $\bar{\Xi}(\lambda,\nu)$ to be the average elapsed service time of the customers that are present at time zero (given the construction of the stationary process $\{Q_{\lambda,\nu}(t,\cdot) : t \in (-\infty,\infty)\}$, see Figure 2.2). That is,

$$\bar{\Xi}(\lambda,\nu) = \frac{1}{Q_{\lambda,\nu}(0,0)} \sum_{n=-1}^{-\infty} \frac{|A_n|}{\lambda} I\left(\frac{|A_n|}{\lambda} < \frac{V_n}{\nu}\right)$$

Likewise, define $\bar{V}(\lambda,\nu)$ as the average of the total service requirement of the customers that are present at time zero, namely

$$\bar{V}(\lambda,\nu) = \frac{1}{Q_{\lambda,\nu}(0,0)} \sum_{n=-1}^{-\infty} \frac{V_n}{v} I\left(\frac{|A_n|}{\lambda} < \frac{V_n}{\nu}\right).$$

Next, we define $\Xi^{(\infty)}(\lambda,\nu)$ as the elapsed service time of the customer with the maximum remaining service time at time zero and $V^{(\infty)}(\lambda,\nu)$ as his total service time requirement. Specifically, if we let $m = \arg\max\{n : V_n/\nu - |A_n|/\lambda\}$ then

$$\Xi^{(\infty)}(\lambda,\nu) = \frac{|A_m|}{\lambda} \text{ and } V^{(\infty)}(\lambda,\nu) = \frac{V_m}{\nu}$$

We then obtain the following representation for the derivatives of $E_\pi \bar{R}(\lambda,\nu)$ and $E_\pi R^\infty(\lambda,\nu)$ with respect to λ and ν.

Lemma 2.4.1 *We have that*

i)

$$\frac{\partial}{\partial\lambda} E_\pi \bar{R}(\lambda,\nu) = \frac{1}{\lambda} E_\pi \bar{\Xi}(\lambda,\nu) \text{ and } \frac{\partial}{\partial\nu} E_\pi \bar{R}(\lambda,\nu) = -\frac{1}{\nu} E_\pi \bar{V}(\lambda,\nu);$$

ii)

$$\frac{\partial}{\partial\lambda} E_\pi R^\infty(\lambda,\nu) = \frac{1}{\lambda} E_\pi \Xi^{(\infty)}(\lambda,\nu) \text{ and } \frac{\partial}{\partial\nu} E_\pi R^\infty(\lambda,\nu) = -\frac{1}{\nu} E_\pi V^{(\infty)}(\lambda,\nu).$$

Proof. We only give a proof of part i) here as the proof of part ii) is entirely analogous. Let R_n denote the remaining service time of the nth customer at time zero and V_n as his total service time requirement, then $R_n \le V_n$. Thus if $EV_n < \infty$, we have

$$E_\pi \bar{R}(\lambda,\nu) < \infty$$

for any $\lambda > 0$,$\nu > 0$.

For a fixed sample path ω constructed backward in time, let $R_n(\lambda, \nu, \omega)$, $n < 0$, denote the remaining service time of customer n (counting backward in time) at time 0 in system (λ, ν). Then $R_n(\lambda, \nu, \omega) = (V_n(\omega)/\nu - |A_n(\omega)|/\lambda)^+$ and

$$\lim_{h\to 0} \frac{R_n(\lambda + h, \nu, \omega) - R_n(\lambda, \nu, \omega)}{h} = \frac{|A_n(\omega)|}{\lambda^2} 1\{\frac{V_n(\omega)}{\nu} \geq \frac{|A_n(\omega)|}{\lambda}\}$$

$$\lim_{h\to 0} \frac{R_n(\lambda, \nu + h, \omega) - R_n(\lambda, \nu, \omega)}{h} = -\frac{V_n(\omega)}{\nu^2} 1\{\frac{V_n(\omega)}{\nu} \geq \frac{|A_n(\omega)|}{\lambda}\}$$

Thus the derivative $\frac{\partial}{\partial\lambda}\bar{R}(\lambda, \nu)$ and $\frac{\partial}{\partial\nu}\bar{R}(\lambda, \nu)$ exists.

Let Ξ_n denote the elapsed service time of the nth customer at time zeros and define $\Xi_n = V_n$ if he is no longer in the system at time zero, then $\Xi_n \leq V_n$. Therefore $E_\pi \frac{\partial}{\partial\lambda}\bar{R}(\lambda, \nu) < \infty$ and $E_\pi \frac{\partial}{\partial\nu}\bar{R}(\lambda, \nu) < \infty$.

As

$$|(\bar{R}_n(\lambda + h, \nu) - \bar{R}_n(\lambda, \nu))/h| \leq \max_{\kappa_{\lambda+h,\nu} < n < 0} V_n/\lambda^2$$

and

$$|(\bar{R}_n(\lambda, \nu + h) - \bar{R}_n(\lambda, \nu))/h| \leq \max_{\kappa_{\lambda,\nu+h} < n < 0} V_n/\nu^2,$$

by Lebesgue Dominated Convergence Theorem, we have

$$\frac{\partial}{\partial\lambda} E_\pi \bar{R}(\lambda, \nu) = E_\pi \frac{\partial}{\partial\lambda}\bar{R}(\lambda, \nu) \text{ and } \frac{\partial}{\partial\nu} E_\pi \bar{R}(\lambda, \nu) = E_\pi \frac{\partial}{\partial\nu}\bar{R}(\lambda, \nu)$$

As the interarrival times have a continuous distribution, $P(V_n/\nu = |A_n|/\lambda) = 0$ for $n < 0$.

Combining the change of limit and the sample path analysis we have

$$\frac{\partial}{\partial\lambda} E_\pi \bar{R}(\lambda, \nu) = \frac{1}{\lambda} E_\pi \bar{\Xi}(\lambda, \nu) \text{ and } \frac{\partial}{\partial\nu} E_\pi \bar{R}(\lambda, \nu) = -\frac{1}{\nu} E_\pi \bar{V}(\lambda, \nu)$$

□

Table 2.3 shows the simulated results of an infinite server queue with base (i.e. $\lambda = 1$) interarrival times distributed as Gamma$(2, 2)$ and base (i.e. $\nu = 1$) service times distributed as Lognormal$(-0.25, 0.5)$.

Table 2.3: Simulation result from exact sampling.

(λ, ν)	$\frac{\partial}{\partial\lambda} E_\pi \bar{R}(\lambda, \nu)$	$\frac{\partial}{\partial\nu} E_\pi \bar{R}(\lambda, \nu)$	$\frac{\partial}{\partial\lambda} E_\pi R^\infty(\lambda, \nu)$	$\frac{\partial}{\partial\nu} E_\pi R^\infty(\lambda, \nu)$
$(80, 1)$	7.0741×10^{-3}	-1.1320	6.1022×10^{-3}	-2.8389
$(100, 1)$	5.6470×10^{-3}	-1.1316	4.9379×10^{-3}	-2.9495
$(120, 1)$	4.7236×10^{-3}	-1.1337	4.2337×10^{-3}	-3.0684

Chapter 3

Perfect Sampling for Loss Systems

In this chapter we present the first class of perfect sampling algorithms for the steady-state distribution of non-Markovian multi-server loss queues and loss networks. The running time of our algorithms is analyzed in the context of many server systems in heavy-traffic; corresponding both to the so-called Quality-Driven (QD) regime, and the Quality-and-Efficiency-Driven (QED, also known as Halfin-Whitt) regime. In both cases, we show that our algorithm achieves sub-exponential complexity as the number of servers and the arrival rate increase. Moreover, in the QD regime, our algorithm achieves a nearly optimal rate of convergence.

In order to implement our strategy in the setting of loss queues, we simulate a stationary infinite server queue backwards in time from the stationary distribution at time zero as our dominating process. In Chapter 2, we explain how to simulate the steady-state measure valued system descriptor of the infinite server queue at a single time point (time zero). In this chapter, we introduce an extension on that, which allows us to simulate the infinite serve queue backwards in time. We also propose a novel application of the DCFTP protocol. Specifically, we use the upper bound process itself to detect coalescence, thereby bypassing the need for a lower bound process and improving the running time of the algorithm. Basically we detect coalescence over a time interval in which all customers initially present in the infinite

server system leave and no loss of customers occurs during that time interval, so that at the end the of time interval, the two systems (coupled infinite server queue and loss queue) have the same set of customers.

3.1 Basic strategy and main results

In this section we introduce the basic strategy to simulate the systems. We also present some results about the efficiency of our algorithms. We leave the details of the algorithms and proofs of the results to subsequent sections. We start with the strategy to simulate the many-server loss queue in steady state and then generalize our strategy to cover loss networks.

3.1.1 Coalescence time with an $GI/GI/C/C$ queue

To facilitate our explanation, we restate the Markovian descriptor of the infinite server (GI/GI/∞) queue which was introduced in Chapter 2. The Markovian description of the state of the multi-server loss queue (GI/GI/C/C) follows the same rationale.

Let $N = \{N(t) : t \in (-\infty, 0]\}$ be a one sided time stationary renewal point process. We write $\{A_n : n \geq 1\}$ for the times at which the process N jumps counting backwards in time from time zero with $A_{n+1} < A_n < 0$. Furthermore, we define $X_n = |A_{n+1} - A_n|$. Now let $\{V_n : n \geq 1\}$ be a sequence of i.i.d. random variables (r.v.'s) which are independent of the process N. Define $Z_n = (A_n, V_n)$ and consider the marked point process $\mathcal{M} = \{Z_n : n \geq 1\} \in \mathbb{R}^2$ which we call the "arriving customer stream". More specifically, we consider customers arriving to the system according to a renewal process with i.i.d. interarrival times X_n's. Independent of the arrival process, their service requirements V_n's are also i.i.d..

We write $G(\cdot) = P(X_n \leq \cdot)$ for the cumulative distribution function (CDF) of X_n and put $\bar{G}(\cdot) = 1 - G(\cdot)$ for its tail CDF. Similarly, we write $F(\cdot) = P(V_n \leq \cdot)$ as the CDF of V_n and $\bar{F}(\cdot) = 1 - F(\cdot)$ as its tail CDF.

The following assumption is imposed throughout our discussion:

Assumption 3.1.1 *$EX_n < \infty$ and $EV_n < \infty$.*

We next introduce a Markovian description of the states of the systems. Let $Q(t, y)$ denote the number of people in the system at time t with residual service time strictly greater than y. Notice that for fixed t, $Q(t, \cdot)$ is a piecewise constant step function. If we denote $\{r_1(t), .., r_m(t)\}$ as the ordered (positive) remaining service times of customers in the system at time t. Then $Q(t, 0) = m$ and $Q(t, y) = \sum_{i=1}^{m} I(r_i(t) > y)$. We also let $E(t)$ denote the time elapsed since the previous arrival at time t (i.e. $E(t) = t - \max\{A_n : A_n \leq t\}$) and $W(t) = (E(t), Q(t, \cdot)) \in \mathbb{R}^+ \times \mathcal{D}[0, \infty)$. Then $\{W(t) : t \in R\}$ forms a Markov process which describes the state of the infinite server queue.

Similarly, we denote $W^L(t) = (E^L(t), Q^L(t, \cdot)) \in \mathbb{R}^+ \times \mathcal{D}[0, \infty)$ as the state of the loss system with C servers at time t, where $E^L(t) = t - \max\{A_n : A_n \leq t\}$ denotes the time elapsed since the previous arrival, and $Q^L(t, y)$ counts the number of people in the loss system at time t with residual service time strictly greater than y. Only costumers who see less than C servers busy at arrival are admitted to the system and all admitted customers start service immediately upon arrival. If we let $(r^L_{(1)}(t), \ldots, r^L_{(m^L)}(t))$ denote the ordered (positive) remaining service times of customers in the system at time t, then $Q^L(t, 0) = m^L$ and $Q^L(t, y) = \sum_{i=1}^{m^L} I(r_{(i)}(t) > y)$.

We now provide a coupling between $W(\cdot)$ and $W^L(\cdot)$ such that $E^L(t) = E(t)$ and $Q^L(t, y) \leq Q(t, y)$ for all $y \geq 0$. In this sense, we say that $W^L(t) \leq W(t)$. The coupling proceeds as follows: we use same stream of customers, $\mathcal{M}$ (same arrival times and service requirements), to update both systems. One can label the servers in the infinite server system, assign customers to the empty server with the smallest label, and by tracking only the state of the first C servers in the infinite server system one automatically tracks the state of the loss system. Based on this coupling, we have

that if $W^L(s) = W(s)$, then $W^L(t) \leq W(t)$ for $t \geq s$.

Definition 3.1.2 *A coalescence time is a time $T < 0$ at which the state of the loss system is identified from the coupled infinite server system, i.e. $W^L(T) = W(T)$.*

As discussed earlier the infinite server system imposes an upper bound on the loss system. A natural way to construct the coalescence (or coupling) time would be to define the coalescence time as the first time (going backwards in time) the infinite server queue empties (assuming, say, unbounded interarrival time distribution, this will occur). However, this coalescence time generally grows exponentially with the arrival rate [34]. So, to detect the coalescence in a more efficient manner, we consider the following construction. Let $R(t)$ denote the maximum remaining service time among all customers in the system at time t. And consider a random time $\tau < 0$ satisfying

1) $R(\tau) < |\tau|$;

2) $\inf_{\tau \leq t \leq \tau + R(\tau)}\{C - Q(t, 0)\} \geq 0$ where C is the number of servers in the loss queue.

As we will show in Section 3.3.2, τ is well defined and our coalescence time is $T := \tau + R(\tau)$. In simple words, Everyone who was present at time τ in the infinite server queue will have left at time $\tau + R(\tau)$. And since the infinite sever queue has less than C customers on $[\tau, \tau + R(\tau)]$, the loss queue is also operating below capacity C on that interval. Thus the infinite serve queue and the loss queue must have the same set of customers present in the system by $\tau + R(\tau)$. From then on we can recover the state of the loss queue at time zero using the same stream of customers as for the infinite server queue on $[\tau + R(\tau), 0]$.

3.1.2 Basic strategy and main results for the GI/GI/∞ queue

Simulating the infinite server queue in backwards in time from stationarity at time zero is not trivial, so we first need to explain how to do this task. There are two cases to be considered.

Case 1 The interarrival time has finite exponential moment in a neighborhood of the origin. More specifically, define $\psi(\theta) = \log E \exp(\theta X_n)$. There exists $\theta > 0$ such that $\psi(\theta) < \infty$.

Case 2 The interarrival time does not have finite exponential moment, i.e. it has heavy-tail distribution.

As we shall explain, we can always reduce the second case to the first one by defining yet another coupled upper bound process trough truncation. Specifically, denote $X_n \wedge b = \min\{X_n, b\}$. We then fix a suitably large constant b and define a coupled infinite server queue with truncated interarrival times: $\{X_n \wedge b : n \geq 1\}$. This truncation essentially speed up the arrival process. By coupling we mean we use the same stream of customers to update both the original system and the truncated one, i.e., We use (X_n, V_n) to update the original system and $(X_n \wedge b, V_n)$ to update the truncated one. We also define the event times as the arrival time and the departure time of the n-th customer, $n \geq 1$ (counting backwards in time). Then the infinite server queue with truncated interarrival times imposes an upper bound, in terms of the number of customers in the system, on the original infinite server queue at the corresponding event times. Precisely, the event times are defined as $A_n = \sum_{i=1}^{n} X_i$ and $A_n + V_n$, $n \geq 1$, for the infinite server system, and $A_n(b) := \sum_{i=1}^{n}(X_i \wedge b)$ and $A_n(b) + V_n$ for the truncated infinite server system. Notice that the actual time of the events (such as arrivals and departures) may be different for the two systems because of the truncation. But from the simulation point of view, we simulate the same amount of information to get the corresponding event times in both systems.

In what follows, we shall first concentrate our discussion on Case 1 which also includes the infinite server queue with truncated interarrival times. We then explain how to extend the result to the heavy-tailed case.

3.1.2.1 Simulating the stationary $GI/GI/\infty$ queue backwards in time.

In this subsection, we introduce a procedure to simulate states of the stationary infinite server system backwards in time for time intervals of any specified length. The construction is similar to the single time point (i.e. time zero) case explained in Chapter 2.

We write $\mu = EX_n$ and fix an $\epsilon \in (0, \mu)$. Define $\kappa_0 := 1$. We consider a sequence of random times κ_j, $j = 1, 2, \cdots$, finite with probability one but large enough such that

$$|A_n - A_{\kappa_{j-1}}| \geq (n - \kappa_{j-1})(\mu - \epsilon) \text{ and } V_n \leq (n - \kappa_{j-1})(\mu - \epsilon) \text{ for all } n \geq \kappa_j. \quad (3.1)$$

Notice that $V_n \leq |A_n - A_{\kappa_{j-1}}|$ for $n \geq \kappa_j$. This implies that a customer who arrives before A_{κ_j} will not be in the system at time $A_{\kappa_{j-1}}$. Thus, using $\{Z_n : 1 \leq n \leq \kappa_j\}$, we can recover the system descriptor $W(t)$ for $t \in [A_{\kappa_{j-1}}, 0]$.

Figure 3.1 gives more details about the construction. Every point Z_n, with $n > \kappa_j$, will not land into the upper triangle defined by the vertical line at $A_{\kappa_{j-1}}$ and the -45^o line intersecting it at the time axis (x axis).

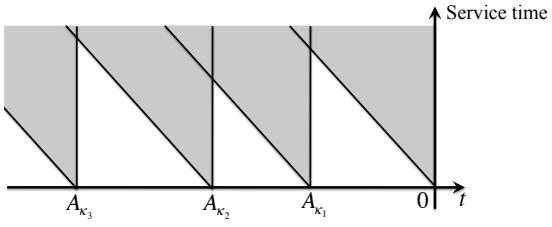


Figure 3.1: Coupling times of the infinite server queue

The κ_j's give us some flexibility to separate the simulation of the two processes.

We first simulate the service times and then conditional on the sample path of the service time we simulate the arrival process jointly with κ_j's.

Define $J_1(0) := 1$ and let

$$\begin{aligned} J_k(l) &= \inf\{n > J_k(l-1) : V_n > (n - J_k(0))(\mu - \epsilon)), \\ \gamma_k &= \inf\{l \geq 0 : J_k(l) = \infty\}, \\ J_{k+1}(0) &= J_k(\gamma_k - 1) \end{aligned}$$

for $k = 1, 2, \cdots$ and $l = 1, 2, \cdots, \gamma_k$.
We first simulate the random time: $J_k(l)$'s for $k = 1, 2, \cdots$ and $l = 1, 2, \cdots, \gamma_k$, and then simulate $\{V_n : n \geq 1\}$ conditional on $J_k(l)$'s; see Algorithm 3.1 in Section 3.2.1 for details.

Given the sample path of $\{V_n : n \geq 1\}$ and $J_k(l)$'s, we next simulate $\{A_n : n \geq 1\}$ and κ_j's. This is done by simulating the negative-drift random walk jointly with its running time maximum. Specifically, we define

$$\tilde{S}_n = n(\mu - \epsilon) - (A_{n+1} - A_1) = \sum_{i=1}^{n} Y_i,$$

where $Y_i = (\mu - \epsilon) - X_{i+1}$. Note that Y_i's are i.i.d. with $EY_i = -\epsilon$. $A_{n+1} = A_1 - \tilde{S}_n + n(\mu - \epsilon)$. We then define $\Delta_1(0) := 0$ and $\Gamma_1(0) := 0$. Fix $m > 0$ and let

$$\begin{aligned} \Delta_j(l) &= \inf\{n \geq \Gamma_j(l-1) : \tilde{S}_n - \tilde{S}_{\Delta_j(0)} \leq -m\}, \\ \Gamma_j(l) &= \inf\{n \geq \Delta_j(l) : \tilde{S}_n - \tilde{S}_{\Delta_j(l)} \geq m\}, \\ \alpha_j &= \inf\{l \geq 1 : \Gamma_j(l) = \infty\}, \\ \kappa_j &= \min\{J_k(0) : J_k(0) \geq \Delta_j(\alpha_j) + 1\}, \\ \Delta_{j+1}(0) &= \kappa_j - 1, \\ \Gamma_{j+1}(0) &= \Delta_{j+1}(0) \end{aligned}$$

for $j = 1, 2, ...$ and $l = 1, 2, \cdots, \alpha_j$.
Notice that the process $\tilde{S}_n$ will never go above $\tilde{S}_{\Delta_j(0)}$ from $\Delta_j(\alpha_j)$ on. This implies

that $|A_n - A_{\kappa_{j-1}}| \geq (n - \kappa_{j-1})(\mu - \epsilon)$ for $n \geq \kappa_j$. Under the light-tail assumption (Case 1), we simulate the random times $\Delta_j(l)$ and $\Gamma_j(l)$ for $j = 1, 2, ..., l = 1, 2, ..., \alpha_j$ and $\{\tilde{S}_n : n \geq 0\}$ by the exponential tilting and acceptance-rejection method. The details are explained in Algorithm 3.2 in Section 3.2.2.

For the heavy-tailed case (Case 2), we can choose the truncation parameter b such that $E[X_n \wedge b] = \int_0^b \bar{G}(x)dx = \mu - 1/2\epsilon$. This is doable because we assume $EX_n = \int_0^\infty \bar{G}(x)dx < \infty$. Denote $A_n(b)$ as the backwards renewal times of the truncated arrival process. Then the $\kappa_j(b)$'s we constructed for the truncated system must automatically satisfy the conditions characterizing κ_j's in (3.1) for the original system as well.

Our algorithm works only under the mild condition in Assumption 3.1.1. But we do impose stronger conditions on the service time distribution to rigorously show good algorithmic performance, especially in heavy traffic (i.e. as the arrival rate increases).

We consider a sequence of systems indexed by $s \in \mathbb{N}^+$. We shall say that s *is the scale of the system.* We speed up the arrival rate of the s-th system by scale s. That is, the interarrival times of the s-th system are given by $X_n^{(s)} = X_n/s$. We keep the service time distribution fixed for all systems, i.e. the service times do not scale with s. The following theorem summarizes the performance of the procedure we proposed for simulating stationary infinite server queue.

Theorem 3.1.3 *Assume $E[X_n] < \infty$, and*

(1) if $EV_n^q < \infty$ for some $q > 2$, then

$$E_\pi^s \kappa_1 = O(s^{q/(q-1)});$$

(2) if we further assume $E[\exp(\theta V_n)] < \infty$ for some $\theta > 0$, then

$$E_\pi^s \kappa_1 = O(s \log s).$$

Its proof is given in Section 3.3.1.

3.1.3 Basic strategy and main results for the $GI/GI/C/C$ system

Once we simulate the customer streams backwards in time and construct the states of the dominating stationary infinite server queue accordingly, we can check and find the coalescence time $T = \tau + R(\tau)$ where τ is defined in Section 3.1.1 backwards in time. Use the state of the infinite server queue at time T as the state of the many-server loss queue at the same time and go forwards in time using the same stream of customers to construct the state of the loss queue up to time 0.

Like in the infinite server queue case, we again consider a sequence of systems indexed by $s \in \mathbb{N}^+$ where the arrival rate of the s-th system is scaled by s and the service rate is kept fixed. Let $\rho = E[V_n]/E[X_n]$ (the ratio of the mean service time and mean interarrival time of the base system). We analyze the system in two heavy-traffic asymptotic regimes. One is the quality driven (QD) regime where $\rho < 1$ and the number of servers in the s-th system, C_s, is s. The other is the quality and efficiency driven (QED) regime where $\rho = 1$ and the number of servers in the s-th system, C_s, is $s + b\sqrt{s}$ with $b > 0$.

Theorem 3.1.4 summarizes the performance of the coalescence time in the QD regime.

Theorem 3.1.4 *Assume $EX_n < \infty$ and X_n's are non-lattice and strictly positive. We also assume that $EV_n^q < \infty$ for any $q > 0$ and the cumulative distribution function (CDF) of V_n is continuous. Then*

$$E_\pi^s \tau = o(s^\delta)$$

for any $\delta > 0$.

Remark 3.1.1 *The existence of all moments assumption on the service time distribution covers a range of heavy tailed distributions, such as Weibull and log-normal, which are known to fit well data in applications [19].*

Theorem 3.1.5 analyzes the performance of the coalescence time in the QED regime.

Theorem 3.1.5 *Assume $EX_n^2 < \infty$. We also assume $EV_n^q < \infty$ for any $q > 0$ and the CDF of V_n is continuous. Then for b large enough, we have*

$$\log E_\pi^s \tau = o(s^\delta)$$

for any $\delta > 0$.

The main difficulty in the proof of Theorem 3.1.4 and Theorem 3.1.5 is that it involves the state of the system on an interval rather than a single point. In Section 3.3.2, we prove Theorem 3.1.4 by using the sample path large deviation results [35] of infinite server queue. For Thereom 3.1.5, we prove it by applying Borel-TIS inequality [36] to the diffusion limit process of infinite server queue [37]. The details is also given in Section 3.3.2.

3.1.4 Extensions and main results for the loss network

Following the definition in [34], we consider a generalized loss network with J stations, labeled $1, 2, \cdots, J$ and suppose that station j comprises C_j servers. We have L possible routes, labeled $1, 2, ..., L$ and for each route l, a J dimensional routing vector P_l. P_l is consist of 1's and 0's, where $P_l(j) = 1$ means route l requires a server at station j. A routing request l is blocked and thus lost if any station j with $P_l(j) = 1$ is full at the arrival time of the request. Customers requesting route l form a renewal process with i.i.d. interarrival times $\{X_n^{(l)} : n \geq 1\}$. The CDF of $X_n^{(l)}$ is G_l. Independent of the arrival process, the service times $\{V_n^{(l)} : n \geq 1\}$ are also i.i.d. with CDF F_l. We assume that G_l's and F_l's satisfy Assumption 3.1.1.

Following the same strategy as in the many-server loss queue case, we first couple the loss network with a network of infinite-server stations. Notice that no customer is blocked or lost in the infinite server system, thus it imposes an upper bound on the number of jobs in the loss system. Let $Q_j(t, y)$ denote the number of jobs in the j-th station with remaining service time strictly greater than y at time t. Note that a class l job with remaining service time greater than y in the system will be counted in all $Q_j(t, y)$'s with $P_l(j) = 1$. Let $R_j(t)$ denote the longest remaining service time among all customers in station j at time t. Let $R(t) = \max_{1\le j\le J}\{R_j(t)\}$. Then similar to the many server loss queue, we define a random time τ' satisfying the following conditions:

1) $R(\tau') \le |\tau'|$,

2) $\inf_{\tau'\le t\le \tau'+R(\tau')} \inf_{1\le j\le J}\{C_j - Q_j(t, 0)\} \ge 0$,
   i.e. all links are operating below capacity on the interval $[\tau', \tau' + R^m(\tau')]$.

At time $\tau' + R(\tau')$, everyone in the network of infinite-server stations will be in the loss network as well. Thus from then on (forwards in time), we can update the loss system using the inputs of the infinite-server system.

In order to simulate the network of infinite-server stations with L types of routing requests, we simulate L independent networks of infinite-server stations; each dealing with a single type of routing request. Then we do a superposition of them. The simulation of each independent network of infinite-server stations are exactly the same as what we have described in Section 3.1.2, as a type l routing request occupies a server from each station j with $P_l(j) = 1$ simultaneously and for the same amount of time. For the l-th system, let $Z_n^{(l)} = (A_n^{(l)}, V_n^{(l)})$ represent the arrival time and service time of the n-th routing request counting backwards in time and $\kappa^{(l)}$ be a random time satisfying that $V_n^{(l)} \le |A_n^{(l)}|$ for all $n \ge \kappa^{(l)}$. Then following the procedure described in Section 3.1.2, we will be able to simulate $\kappa^{(l)}$ as the maximum of two random times associated the arrival process and service time process respectively.

We now consider a sequence of systems indexed by $s \in \mathbb{N}^+$. We speed up the the arrival rate of the s-th system by s, i.e. $X_n^{(l,s)} = X_n^{(l)}/s$, and keep the service rate fixed. The same result as in Theorem 3.1.3 will still be holding here. Specifically,

Theorem 3.1.6 (Theorem 3.1.3') *Assume $EX_n^{(l)} < \infty$ (C).*

(1) if $E[(V_n^{(l)})^q] < \infty$ for some $q > 2$, then

$$E_\pi^s \kappa^{(l)} = O(s^{q/(q-1)});$$

(2) if we further assume $E[\exp(\theta V_n^{(l)})] < \infty$ for some $\theta > 0$, then

$$E_\pi^s \kappa^{(l)} = O(s \log s)$$

for $l = 1, 2, \cdots, L$.

The proof of Theorem 3.1.6 is the same as that of Theorem 3.1.3 except for a few notational changes, thus we shall omit it here.

If we held the number of routing request types, L, fixed, as we shall explain below, similar results as in Theorem 3.1.4 and Theorem 3.1.5 for the coalescence time will be holding here as well. We again run L independent networks of infinite-server stations as described above. Network l serves routing request of type l only, for $l = 1, 2, ..., L$. Let $Q^{(l)}(t, 0)$ denote the number of jobs in network l at time t and $R^{(l)}(t)$ denote the maximum remaining service time among all jobs in the network at time t. Then we have $R(t) = \max\{R^{(l)}(t) : 1 \le l \le L\}$.

We consider two asymptotic regimes. One is the QD regime where for the base system we have

$$\sum_{l=1}^{L} \frac{EV_n^{(l)}}{EX_n^{(l)}} P_j(l) < C_j. \tag{3.2}$$

For the s-th system, the number of servers in the j-th station is $C_j^s = sC_j$ for $j = 1, 2, ..., J$.

Assign a fixed number H_l to each route l. H_l is well chosen such that $E[V_n^{(l)}]/E[X_n^{(l)}] < H_l$ and $\sum_{l=1}^{L} H_l P_l(j) \le C_j$. This is doable because of (3.2). Let $H_l^s = sH_l$. Define a random time $\bar{\tau}'$ satisfying the following two conditions:

1) $R^{(l)}(\bar{\tau}') \leq |\bar{\tau}'|$ for $l = 1, 2, \cdots, L$,

2) $\inf_{\bar{\tau}' \leq t \leq \bar{\tau}' + R(\bar{\tau}')}\{H_l - Q^l(t, 0)\} \geq 0$ for $l = 1, 2, \cdots, L$.

Notice that $\bar{\tau}'$ is an upper bound on τ'. As the number of types of routing request is fixed at L (it does not scale with s), using the construction outlined in Section 3.3.2.1, we can show that the result in Theorem 3.1.4 holds for $\bar{\tau}'$ as well.

Theorem 3.1.7 (Theorem 3.1.4') *Assume $EX_n^{(l)} < \infty$ and $X_n^{(l)}$'s are non-lattice and strictly positive. We also assume $E[(V_n^{(l)})^q] < \infty$ for any $q > 0$ and F_l is continuous. Then*

$$E_\pi^s \tau' = o(s^\delta)$$

for any $\delta > 0$.

The other asymptotic regime is the QED regime where for the base system we have

$$\sum_{l=1}^{L} \frac{EV_n^{(l)}}{EX_n^{(l)}} P_j(l) = C_j$$

and the number of servers in the j-th station of the s-th system is $C_j^s = sC_j + \beta_j\sqrt{s}$ for $j = 1, 2, \cdots, J$

We then let $I_l = E[V_n^{(l)}]/E[X_n^{(l)}]$ and $I_l^s = sI_l + a_l\sqrt{s}$ where a_l's are well chosen such that $\sum_{l=1}^{L} a_l P_j(l) \leq \beta_j$.

We define a random time $\tilde{\tau}'$ that satisfies the following two conditions:

1) $R^{(l)}(\tilde{\tau}') \leq |\tilde{\tau}'|$ for $l = 1, 2, \cdots, L$,

2) $\inf_{\tilde{\tau}' \leq t \leq \tilde{\tau}' + R(\tilde{\tau}')}\{I_l - Q^{(l)}(t, 0)\} \geq 0$ for $l = 1, 2, \cdots, L$.

As before, $\tilde{\tau}'$ is an upper bound on τ'. It is easy to check using the construction outlined in Section 3.3.2.2 that the result in Theorem 3.1.5 holds for $\tilde{\tau}'$ as well.

Theorem 3.1.8 (Theorem 3.1.5') *Assume* $E[(X_n^{(l)})^2] < \infty$. *We also assume* $E[(V_n^{(l)})^q] < \infty$ *for any* $q > 0$. *Then for* b_j*'s large enough, we have*

$$\log E_\pi^s \tau' = o(s^\delta)$$

for any $\delta > 0$.

We shall omit the proof of Theorem 3.1.7 and Theorem 3.1.8 as it is the same as the proof of Theorem 3.1.4 and Theorem 3.1.5 with the introduction $\bar{\tau}'$ and $\tilde{\tau}'$ except for a few notational changes.

The rest of this chapter is organized as follows. In Section 3.2 we provide the details required to implement our general strategy outlined in this section for the infinite server queue backward in time from the steady state at time zero. In Section 3.3 we study the running time of our algorithms under heavy traffic. Some technical results in the development of Section 3.3 are given in Section 3.4.

3.2 Detailed simulation algorithms

In order to provide the details of our simulation algorithms outlined in Section 3.1, we shall first work under the light-tailed case (Case 1) where we assume there exists $\theta > 0$ such that $\psi(\theta) < \infty$. The extension to the heavy-tailed case (Case 2) was introduced in Section 3.1 and we shall provide more details in Section 3.2.3.

We further impose the following assumptions on our ability to simulate the service times and interarrival times.

Assumption 3.2.1 *We assume that for each* $x \geq 0$, $F(x)$ *is easily computable, either in closed form or via efficient numerical procedures. Moreover, we can simulate* V_n *conditional on* $V_n \in (a, b]$ *with* $P(V_n \in (a, b]) > 0$. *The sampling time of* V_n *conditional on* $V_n \in (a, b]$ *is assumed to be independent of* a *and* b.

Assumption 3.2.2 *Suppose that $G(\cdot)$ is known and that it is possible to simulate from $G_{eq}(\cdot) := \mu^{-1}\int_{\cdot}^{\infty}\overline{G}(t)\,dt$. Moreover, let $G_\theta(\cdot) = E\exp(\theta X_n - \psi(\theta))I(X_n \le \cdot)$ be the associated exponentially tilted distribution with parameter θ for $\psi(\theta) < \infty$. We assume that we can simulate from $G_\theta(\cdot)$.*

Remark 3.2.1 *Assumption 3.2.1 can be applied to virtually any model used in practice, including distributions such as Gamma, phase-type, Pareto, Weibull, Lognormal, and mixtures of them. Knowledge of the underlying distribution is required in Procedure A below. Note that the required simulation procedure is not restricted to the inversion method. One can use, for example, the acceptance/rejection method, but a good proposal distribution for the conditional distribution given $V_n \in (a, b]$ might have to be constructed based on knowledge of the density function to increase efficiency. Assumption 3.2.2 is applicable to models for which the moment generating function is finite, these include distributions such as Gamma, phase type, hyperexponential, and other mixtures of them.*

We next introduce our algorithm to simulate $\{V_n : n \ge 1\}$. Conditional on the sample path of $\{V_n : n \ge 1\}$, we then explain how to to simulate $\{A_n : n \ge 1\}$ and κ_j's.

3.2.1 Simulation of $\{V_n : n \ge 1\}$ and $J_k(l)$'s for $k = 1, 2, \cdots$, $l = 1, 2, \cdots, \gamma_k$

We will first introduce the procedure to simulate $J_1(l)$ for $l = 1, 2, \cdots, \gamma_1$. Recall that $J_1(l)$'s record the position of all the record breakers. Let $p(n) = P(V_1 > n(\mu - \epsilon))$. Then $P(J_1(l) = \infty | J_1(l-1) = k) = \prod_{n=k+1}^{\infty}(1 - p(n))$, which is the probability of success (there are no more record breakers) of the Bernoulli trial. It involves the evaluation of the product of infinite terms. In Procedure A, we introduce a sandwiching approximation scheme to sampling the Bernoulli trail together with $J_1(l)$ if $J_1(l) < \infty$.

The following lemma guarantees the termination of our procedure.

Lemma 3.2.3 *If $EV_1 < \infty$, then*

$$P(J_1(1) = \infty) = \prod_{n=1}^{\infty}(1 - p(n)) \geq \exp(-\frac{cEV_1}{\mu - \epsilon}) > 0 \tag{3.3}$$

for some constant c depending on the value of $p(1)$, and consequently

$$E\gamma_1 \leq \exp(cEV_1/(\mu - \epsilon)) < \infty.$$

Proof.

$$P(J_1(1) = \infty) = \prod_{n=1}^{\infty}(1 - p(n)) \geq \prod_{n=1}^{\infty}\exp(-cp(n))$$
$$\geq \exp(-\frac{c}{\mu - \epsilon}\int_0^{\infty} P(V_1 > \nu)d\nu) = \exp(-\frac{cEV_1}{\mu - \epsilon}).$$

For $l = 2, 3, \cdots$, conditional on $J_1(l-1) = k$:

$$P(J_1(l) = \infty | J_1(l-1) = k) = \prod_{n=k+1}^{\infty}(1 - p(n))$$
$$\geq \exp(-\frac{c\int_k^{\infty} P(V_1 > \nu)d\nu}{\mu - \epsilon}) \geq \exp(-\frac{cEV_1}{\mu - \epsilon}),$$

thus γ_1 is stochastically dominated by a geometric random variable with parameter $p = \exp(-cEV_1/(\mu - \epsilon))$. The result then follows. □

We next introduce our sandwiching approximation scheme. Notice that

$$\prod_{i=k+1}^{h}(1 - p(i)) \geq P(J_1(l) = \infty | J_1(l-1) = k)$$
$$\geq \prod_{i=k+1}^{h}(1 - p(i)) \times \exp(-\frac{2\int_h^{\infty} P(V_1 > \nu)d\nu}{\mu - \epsilon}) \tag{3.4}$$

for $h > k$.

Another important observation is that if we let $\prod_{i=k+1}^{k}(1 - p(i)) = 1$, then

$$\prod_{i=k+1}^{h-1}(1 - p(i)) - \prod_{i=k}^{h}(1 - p(i)) = p(h)\prod_{i=k}^{h-1}(1 - p(i)) = P(J_1(l) = h | J_1(l-1) = k)$$

for $h > k$.

Let

$$u(h) = \exp(-\frac{2\int_h^\infty P(V_1 > \nu)d\nu}{\mu - \epsilon}).$$

We now propose the following procedure to simulate the value of $J_1(l)$ conditional on $J_1(l-1) = k$.

Procedure D (Simulate $J_1(l)$ given $J_1(l-1) = k$)

S1. Initialize $h = k+1$, $g = 1 - p(h)$ and $f = gu(h)$. Simulate $U \sim \text{Unif}[0,1]$

S2. While $f < U < g$,
set $h = h+1$, $g = g(1-p(h))$ and $f = gu(h)$
end while

S3. If $U \le f$, then $J_1(l) = \infty$. Otherwise, $J_1(l) = h$.

The simulation of $J_k(l)$ for $k = 1, 2, \ldots$, $l = 1, 2, \ldots, \gamma_k$ follows the same rationale. We let $p_k(n) = P(V_1 > n(\mu-\epsilon)|V_1 \le (n + J_k(0) - J_{k-1}(0))(\mu-\epsilon))$. Then following the same argument leading to (3.3) and (3.4), we have correspondingly

$$P(J_k(1) = \infty) > 0,$$

and

$$\begin{aligned}
&\prod_{i=n+1}^{h}(1-p_k(i)) \\
\ge &P(J_k(l) - J_k(0) = \infty | J_k(l-1) - J_k(0) = n) \\
\ge &\prod_{i=n+1}^{h}(1-p_k(i)) \times \exp(-\frac{2\int_h^\infty P(V_1 > \nu | V_1 \le \nu + (J_k(0) - J_{k-1}(0))(\mu-\epsilon))d\nu}{\mu-\epsilon})
\end{aligned}$$

for $h > n$.

Let

$$u_k(h) = \exp(-\frac{2\int_h^\infty P(V_1 > \nu | V_1 \le \nu + (J_k(0) - J_{k-1}(0))(\mu-\epsilon))d\nu}{\mu-\epsilon}).$$

We now propose a modification of Procedure D that allows us to simulate $J_k(l)$ conditional on $J_k(l-1) - J_k(0) = n$.

Procedure D1 (Simulate $J_k(l)$ given $J_k(l-1) - J_k(0) = n$)

S1. Initialize $h = n+1$, $g = 1 - p_k(h)$ and $f = gu_k(h)$. Simulate $U \sim \text{Unif}[0,1]$.

S2. While $f < U < g$,
set $h = h+1$, $g = g(1-p_k(h))$ and $f = gu_k(h)$
end while

S3. If $U \leq f$, then $J_k(l) = \infty$. Otherwise, $J_k(l) = J_k(0) + h$.

Based on Procedure D1 and our previous analysis we have:

Algorithm 3.1 (Sample V_n's jointly with $J_k(l)$'s)

S0. Set $J_0(0) = -\infty$, $J_1(0) = 1$, $k = 1$, $l = 1$. Simulate V_1 according to its nominal distribution.

S1. Simulate $J_k(l)$ conditional on the value of $J_k(l-1)$ using Procedure D1.

S2. If $J_k(l) = \infty$, set $\gamma_k = l$, $J_{k+1}(0) = J_k(\gamma_k - 1)$, $k = k+1$, $l = 1$ and go back to Step 1. Otherwise, go to S3.

S3. Simulate V_n for $J_k(l-1) < n < J_k(l)$ by conditioning on $V_n \leq (n - J_k(0))(\mu - \epsilon)$ and simulate $V_{J_k(l)}$ by conditioning on $(J_k(l) - J_k(0))(\mu - \epsilon) < V_{J_k(l)} \leq (J_k(l) - J_{k-1}(0))(\mu - \epsilon)$. Set $l = l+1$ and go back to S1.

When running the above algorithm, we specify K as the number of intervals ($[J_k(0), J_k(\gamma_k - 1)]$) we want to simulate. We then run Algorithm I from $k = 1$ till $k = K$. The program will give us $\{V_n : 1 \leq n \leq J_K(\gamma_K - 1)\}$ and $J_k(l)$'s for $k = 1, 2, \cdots, K$, $l = 1, 2, \cdots, \gamma_k$.

3.2.2 Simulation of $\{A_n : n \geq 1\}$ and $\Delta_j(l)$'s, $\Gamma_j(l)$'s for $j = 1, 2, ...,$ $l = 1, 2, ..., \alpha_j$

Given the sample path of $\{V_n : n \geq 1\}$, we will first explain how to simulate the $\Delta_j(l)$'s and $\Gamma_j(l)$'s sequentially and jointly with the underlying random walk $\{\tilde{S}_n : n \geq 1\}$. We then simulate A_1 according to $G_{eq}(\cdot)$ and set $A_{n+1} = A_1 + n(\epsilon - \mu) - \tilde{S}_n$. The analysis and methodology in this subsection closely follows those in [32] and [31]. The same procedure can be used to simulate a negative drifted random walk, $\tilde{S}_n$, together with its running time maximum defined as $\max_{k \geq n}\{\tilde{S}_k\}$.

Let $\mathcal{F}_n = \sigma\{Y_1, Y_2, \cdots, Y_n\}$, the σ-field generated by the Y_j's up to time n. Let $\xi \geq 0$ and define

$$T_\xi = \inf\{n \geq 0 : \tilde{S}_n > \xi\}.$$

Then by the strong Markov property we have that for $1 \leq l \leq \alpha_j$,

$$P(\Gamma_j(l) = \infty | \mathcal{F}_{\Delta_j(l)}) = P(\Gamma_j(l) = \infty | \tilde{S}_{\Delta_j(l)}) = P(T_m = \infty) > 0,$$

where we use $P(\cdot)$ to denote the nominal probability measure.

It is important then to notice that

$$P(\alpha_j = k) = P(T_m < \infty)^{k-1} P(T_m = \infty)$$

for $k \geq 1$. In other words, α_j is geometrically distributed. The procedure that we have in mind is to simulate each stage $\Delta_j(\alpha_j)$ in time intervals, and the number of time intervals is precisely α_j.

Let $\psi_Y(\theta) = \log E \exp(\theta Y_i)$ be the log moment generating function of Y_i. As we assume $\psi_X(\theta)$ is finite in a neighborhood of zero, $\psi_Y(\cdot)$ is also finite in a neighborhood of zero. Moreover $EY_i = \psi_Y'(0) = -\epsilon$ and $\mathrm{Var}(Y_i) = \psi_Y''(0) > 0$. Then by the convexity of $\psi_Y(\cdot)$, one can always select $\epsilon > 0$ sufficiently small so that there exists $\eta > 0$ with $\psi_Y(\eta) = 0$ and $\psi_Y'(\eta) \in (0, \infty)$. The root η allows us to define a new measure P_η based on exponential tilting so that

$$\frac{dP_\eta}{dP}(Y_i) = \exp(\eta Y_i).$$

Moreover, under P_η, $\tilde{S}_n$ is random walk with positive drift equal to $\psi'_Y(\eta)$ [1]. Therefore $P_\eta(T_\xi < \infty) = 1$ and

$$q(\xi) := P(T_\xi < \infty) = E_\eta \exp(-\eta \tilde{S}_{T_\xi})$$

for each $\xi \geq 0$. Based on the above analysis we now introduce a convenient representation to simulate a Bernoulli random variable $J(\xi)$ with parameter $q(\xi)$, namely,

$$J(\xi) = I(U \leq \exp(-\eta \tilde{S}_{T_\xi})), \tag{3.5}$$

where U is a uniform random variable independent of everything else under P_η.

Identity (3.5) provides the basis for an implementable algorithm to simulate a Bernoulli random variable with success probability $q(\xi)$. Sampling $\{\tilde{S}_1, \cdots, \tilde{S}_{T_\xi}\}$ conditional on $T_\xi < \infty$, as we shall explain now, corresponds to basically the same procedure. First, let us write

$$P^*(\cdot) = P(\cdot | T_\xi < \infty).$$

The following result provides an expression for the likelihood ratio between P^* and P_η.

Lemma 3.2.4 *We have that*

$$\frac{dP^*}{dP_\eta}(\tilde{S}_1, ..., \tilde{S}_{T_\xi}) = \frac{\exp(-\eta \tilde{S}_{T_\xi})}{P(T_\xi < \infty)} \leq \frac{\exp(-\eta\xi)}{P(T_\xi < \infty)}.$$

Proof.

$$\begin{aligned}
&P(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_\xi} \in H_{T_\xi} | T_\xi < \infty) \\
=&\frac{P(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_\xi} \in H_{T_\xi}, T_\xi < \infty)}{P(T_\xi < \infty)} \\
=&\frac{E_\eta[\exp(-\eta \tilde{S}_{T_\xi}) I(\tilde{S}_1 \in H_0, ..., \tilde{S}_{T_\xi} \in H_{T_\xi})]}{P(T_\xi < \infty)}.
\end{aligned}$$

□

The previous lemma provides the basis for a simple acceptance / rejection procedure to simulate $\{\tilde{S}_1, ..., \tilde{S}_{T_\xi}\}$ conditional on $T_\xi < \infty$. More precisely, we propose $\{\tilde{S}_1, ..., \tilde{S}_{T_\xi}\}$ from $P_\eta(\cdot)$. Then one generates a uniform random variable U independent of everything else and accept the proposal if

$$U \leq \frac{P(T_\xi < \infty)}{\exp(-\eta\xi)} \times \frac{dP^*}{dP_\eta}(\tilde{S}_1, ..., \tilde{S}_{T_\xi}) = \exp(-\eta(\tilde{S}_{T_\xi} - \xi)).$$

This criterion coincides with $J(\xi)$ according to (3.5). So, the procedure above simultaneously obtains both a Bernoulli random variable $J(\xi)$ with parameter $q(\xi)$, and the corresponding path $\{\tilde{S}_1, ..., \tilde{S}_{T_\xi}\}$ conditional on $T_\xi < \infty$ under $P(\cdot)$ if $J(\xi) = 1$.

As $E[Y_i] = -\epsilon < 0$, by strong law of large numbers we have $\Delta_j(l) < \infty$ almost surely for $j = 1, 2, ...$ and $l = 1, 2, ..., \alpha_j$. We next define

$$\bar{q}(\xi) = 1 - q(\xi) = P(T_\xi = \infty)$$

and

$$P'(\cdot) = P(\cdot | T_\xi = \infty).$$

The following result provides an expression for the likelihood ratio between P' and P.

Lemma 3.2.5 *We have that*

$$\frac{dP'}{dP}(\tilde{S}_1, ..., \tilde{S}_n) = \frac{I(T_\xi > l)\bar{q}(\xi - \tilde{S}_n)}{P(T_\xi = \infty)} \leq \frac{1}{P(T_\xi = \infty)}.$$

Proof.

$$\begin{aligned}
&P(\tilde{S}_1 \in H_1, ..., \tilde{S}_n \in H_n | T_\xi = \infty) \\
=&\frac{P(\tilde{S}_1 \in H_1, ...\tilde{S}_n \in H_n, T_\xi = \infty)}{P(T_\xi = \infty)} \\
=&\frac{E[I(\tilde{S}_1 \in H_1, ..., \tilde{S}_n \in H_n) I(T_\xi > n) P(T_\xi = \infty | \tilde{S}_1, ..., \tilde{S}_n)]}{P(T_\xi = \infty)}.
\end{aligned}$$

The result then follows from the strong Markov property and homogeneity of the random walk. □

We are in good shape now to apply acceptance/rejection to sample from P'. The previous lemma indicates that to sample $\{\tilde{S}_1, ..., \tilde{S}_n\}$ given $T_\xi = \infty$. We can propose from the original (nominal) distribution and accept with probability $\bar{q}(\xi - \tilde{S}_n)$ as long as $\tilde{S}_j \leq \xi$ for all $0 \leq j \leq n$. And in order to perform the acceptance test we need to sample a Bernoulli with parameter $\bar{q}(\xi - \tilde{S}_n)$, but this is easily done using identity (2.1).

Now consider $0 \leq \xi_1 < \xi_2$, we define

$$P^o(\cdot | T_{\xi_1} < \infty, T_{\xi_2} = \infty).$$

The following result provides an expression for the likelihood ratio between P^o and P_η.

Lemma 3.2.6 *We have that*

$$\frac{dP^o}{dP_\eta}(\tilde{S}_1, ...\tilde{S}_{T_{\xi_1}}) = \frac{\exp(-\eta \tilde{S}_{T_{\xi_1}}) \bar{q}(\xi_2 - \tilde{S}_{T_{\xi_1}})}{P(T_{\xi_1} < \infty, T_{\xi_2} = \infty)} \leq \frac{\exp(-\eta \xi_1)}{P(T_{\xi_1} < \infty, T_{\xi_2} = \infty)}.$$

Proof.

$$\begin{aligned} & P(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_{\xi_1}} \in H_{T_{\xi_1}} | T_{\xi_1} < \infty, T_{\xi_2} = \infty) \\ =& \frac{E_\eta[I(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_{\xi_1}} \in H_{T_{\xi_1}}) \exp(-\eta \tilde{S}_{T_{\xi_1}}) P(T_{\xi_2} = \infty | \tilde{S}_1, ..., \tilde{S}_{T_{\xi_1}})]}{P(T_{\xi_1} < \infty, T_{\xi_2} = \infty)}. \end{aligned}$$

□

We again use acceptance/rejection to sample $\{\tilde{S}_1, ..., \tilde{S}_{T_{\xi_1}}\}$ given $T_{\xi_1} < \infty$ and $T_{\xi_2} = \infty$. We propose $\{\tilde{S}_1, ..., \tilde{S}_{T_{\xi_1}}\}$ from $P_\eta(\cdot)$. Then we simulate a uniform random variable U independent of all else and accept the proposal if

$$U \leq \frac{P(T_{\xi_1} < \infty, T_{\xi_2} = \infty)}{\exp(-\eta \xi_1)} \times \frac{dP^o}{dP_\eta}(\tilde{S}_1, ..., \tilde{S}_{T_{\xi_1}}) = \exp(-\eta(\tilde{S}_{T_{\xi_1}} - \xi_1)) q(\xi_2 - \tilde{S}_{T_{\xi_1}}).$$

Based on the above analysis we propose the following algorithm.

Algorithm 3.2 (Given V_n's and $J_k(l)$'s, sample $\tilde{S}_n$'s together with $\Delta_j(l)$'s, $\Gamma_j(l)$'s and κ_j's)

S0. Set $\Delta_1(0) = \Gamma_1(0) = 0$, $\tilde{S}_0 = 0$, $j = 1$, $l = 1$, $\xi = \infty$, $\gamma = -m$. Sample A_1 according to $G_{eq}(\cdot)$.

S1. Simulate $S_1, ..., S_{T_\gamma}$ from the original (nominal) distribution.

S2. If $S_i \leq \xi$ for all $1 \leq i \leq T_\gamma$ then sample a Bernoulli $J(\xi - S_{T_\gamma})$ with parameter $q(\xi - S_{T_\gamma})$ using (3.5) and continue to S3. Otherwise (i.e. $S_i > \xi$ for some $1 \leq i \leq T_\gamma$) go back to S1.

S3. If $J(\xi - S_{T_\gamma}) = 1$, go back to S1. Otherwise $J(\xi - S_{T_\gamma}) = 0$, let $\Delta_j(l) = \Gamma_j(l-1) + T_\gamma$ and $\tilde{S}_{\Gamma_j(l-1)+i} = \tilde{S}_{\Gamma_j(l-1)} + S_i$ for $i = 1, ..., T_\gamma$. If $j \geq 2$, set $\xi = \tilde{S}_{\Delta_{j-1}(\alpha_{j-1})} + m - \tilde{S}_{\Delta_j(l)}$.

S4. Simulate $S_1, ..., S_{T_m}$ from $P_\eta(\cdot)$. Sample a Bernoulli $J(\xi - S_{T_m})$ with parameter $q(\xi - S_{T_m})$ using (3.5) and $U \sim \text{Unif}[0,1]$. Let $J^* = I(U \leq \exp(-\eta(S_{T_m} - m)) \times (1 - J(\xi - S_{T_m}))$.

S5. If $J^* = 1$, let $\Gamma_j(l) = \Delta_j(l) + T_m$ and $\tilde{S}_{\Delta_j(l)+i} = \tilde{S}_{\Delta_j(l)} + S_i$ for $1 \leq i \leq T_m$. Set $\gamma = \min\{0, \tilde{S}_{\Delta_j(0)} - m - \tilde{S}_{\Gamma_j(l)}\}$. If $j \geq 2$, set $\xi = \tilde{S}_{\Delta_{j-1}(\alpha_{j-1})} + m - S_{\Gamma_j(l)}$. Set $l = l + 1$ and go back to step 1. Otherwise $J^* = 0$, set $\alpha_j = l$, $\kappa_j = \inf\{J_k(0) : J_k(0) \geq \Delta_j(\alpha_j) + 1\}$, $\Delta_{j+1}(0) = \kappa_j - 1$, $\xi = m$ and continue to S6.

S6. Let $h = \Delta_{j+1}(0) - \Delta_j(\alpha_j)$. Sample $S_1, ..., S_h$ from the original distribution.

S7. If $S_i \leq \xi$ for all $1 \leq i \leq h$ then sample a Bernoulli $J(\xi - S_h)$ with parameter $q(\xi - S_h)$ using (3.5) and continue to S8. Otherwise (i.e. $S_i > \xi$ for some $1 \leq i \leq h$), go back to S6.

S8. If $J(\xi - S_h) = 1$, go back to S6. Otherwise $J(\xi - S_h) = 0$, let $\tilde{S}_{\Delta_j(\alpha_j)+i} = \tilde{S}_{\Delta_j(\alpha_j)} + S_i$ for $i = 1, ..., h$. Set $A_{n+1} = A_1 + n(\epsilon - \mu) - \tilde{S}_n$ for $n = \Delta_j(0) + 1, ..., \Delta_{j+1}(0)$. Set $j = j + 1$, $l = 1$, $\xi = \tilde{S}_{\Delta_{j-1}(\alpha_{j-1})} + m - \tilde{S}_{\Delta_j(0)}$, $\gamma = -m$ and go back to S1.

When running the above algorithm, we specify K as the number of intervals ($[\kappa_{j-1}, \kappa_j]$) we want to simulate and then repeat the above process from $j = 1$ till $j = K$. The program will give us $\{A_n : 1 \le n \le \kappa_K\}$ and $\{\kappa_j : 1 \le j \le K\}$.

3.2.3 Coupled infinite server queue with truncated interarrival times

In this subsection, we provide some additional details for simulating the coupled truncated system together with the original system.

We first explain how to simulate A_1 jointly with $A_1(b)$. The equilibrium distribution of X_n is $G_{eq}(x) = \int_0^x \bar{G}(u)du/EX_n$ and the equilibrium distribution of $X_n \wedge b$ is

$$G^b_{eq}(x) = \frac{\int_0^x \bar{G}(u)du}{E[X_n \wedge b]} I\{x \le b\}.$$

Thus we simulate A_1 with CDF $G_{eq}(x)$, if $A_1 \le b$, we set $A_1(b) = A_1$. Otherwise if $A_1 > b$, we keep simulating X_e with CDF $G_{eq}(x)$ until $X_e \le b$ and set $A_1(b) = X_e$. In particular we have $A_1(b) \le A_1$.

When simulating $X_n \wedge b$'s from the nominal distribution, we first simulate X_n with CDF $G(\cdot)$ and set $X_n \wedge b = \min\{X_n, b\}$. Denote $Y_n(b) = (E[X_n \wedge b] - \epsilon') - X_n \wedge b$ and let η_b be chosen such that $\log E \exp(\eta_b Y_n(b)) = 0$. When simulating $X_n \wedge b$'s under exponential tilting $P_{\eta_b}(\cdot)$, we first simulate $Y_n(b)$ under $P_{\eta_b}(\cdot)$ and set $X_n \wedge b = (E[X_n \wedge b] - \epsilon') - Y_n(b)$. If $X_n \wedge b < b$, set $X_n = X_n \wedge b$, otherwise ($X_n \wedge b = b$), sample X_n conditional on $X_n \ge b$ under the nominal distribution $P(\cdot)$.

3.3 Performance analysis

In the previous section, we provide our simulation algorithm and show that our algorithm works in the sense that the termination time is finite with probability one. In this section, we conduct some further asymptotic analysis on the performance of

our algorithm. We first analyze the algorithm for the infinite server system and then conduct some analysis on the coalescence time for the many-server loss system.

3.3.1 Termination time for the infinite server system (Proof of Theorem 3.1.3)

Theorem 3.1.3 provides the relationship between the moment of the service times and $E_\pi^s\kappa$. We next give a proof of it. We shall omit the subscription π and s when there is no confusion for notational convinience. We first give a proof of the light tailed case. Let $\kappa(V) = \inf\{k > 1 : V_{n+1} \le n(\mu-\epsilon)/s \text{ for all } n \ge k\}$ and $\kappa(A) = \inf\{k > 1 : A_{n+1} \ge n(\mu-\epsilon)/s \text{ for all } n \ge k\}$. Then $\kappa_1 = \max\{\kappa(V), \kappa(A)\}$ We prove the theorem by establishing the bounds for $\kappa(V)$ (Lemma 3.3.1) and $\kappa(A)$ (Lemma 3.3.2) respectively.

Lemma 3.3.1 *If $EV_n^q < \infty$ for some $q > 2$, then*

$$E\kappa(V) = O(s^{q/(q-1)}).$$

Proof. Let $p(n) = P(V_1 > n(\mu-\epsilon)/s)$. For k sufficiently large, we have

$$\begin{aligned} P(\kappa(V) > k) &= 1 - \prod_{n=k+1}^{\infty}(1-p(n)) \\ &\le 1 - \exp(-\frac{2s}{\mu-\epsilon}\int_{k(\mu-\epsilon)/s}^{\infty} P(V > \nu)d\nu). \end{aligned}$$

By Chebyshev's inequality

$$P(V_n > \nu) \le \frac{EV_n^q}{\nu^q}.$$

Let $\delta = 1/(q-1)$, then for s sufficiently large, we have

$$\begin{aligned} \sum_{k=s^{1+\delta}}^{\infty} P(\kappa(V) > k) &\le \sum_{k=s^{1+\delta}}^{\infty} \frac{2s}{\mu-\epsilon}\int_{k(\mu-\epsilon)/s}^{\infty} P(V > \nu)d\nu \\ &\le \frac{2EV_n^q s^q}{(q-1)(q-2)(\mu-\epsilon)^q}\sum_{k=s^{1+\delta}}^{\infty}\frac{1}{k^{q-1}} \\ &= O(s^{q-(1+\delta)(\delta-2)}). \end{aligned}$$

As $q-(1+\delta)(q-2)=1+\delta$,

$$\begin{aligned} E\kappa(V) &= \sum_{k=0}^{\infty} P(\kappa(V)>k) \\ &= \sum_{k=0}^{s^{1+\delta}-1} P(\kappa(V)>k) + \sum_{k=s^{1+\delta}}^{\infty} P(\kappa(V)>k) \\ &\le s^{1+\delta} + O(s^{1+\delta}). \end{aligned}$$

□

Notice that when $E\exp(\theta V_n)<\infty$ for some $\theta>0$,

$$P(V_n>\nu) \le E\exp(\theta(V_n-\nu)) = E\exp(\theta V_n)\exp(-\theta\nu).$$

Similarly as above, for s sufficiently large we have

$$\sum_{k=\lceil \frac{2}{\theta(\mu-\epsilon)} s\log s\rceil}^{\infty} P(\kappa(V)>k) \le \frac{2E\exp(\theta V_n)}{(\mu-\epsilon)^2\theta^2}$$

and

$$E\kappa(V) = \sum_{k=0}^{s\log s-1} P(\kappa(V)>k) + \sum_{k=s\log s}^{\infty} P(\kappa(V)>k) \le s\log s + O(1).$$

Thus if $E\exp(\theta V)<\infty$ for some $\theta>0$, then

$$E\kappa(V) = O(s\log s).$$

Lemma 3.3.2 *Assume there exist $\theta>0$, such that $\psi(\theta)<\infty$, then*

$$E\kappa(A) = O(s).$$

Proof. Based on the algorithm proposed in Section 3.2.2, we divide the proof into two parts. We first prove that the expected number of iterations is $O(1)$. We then prove that the expected number of steps to reach $-m$ or m is $O(s)$.

Let $T_\xi = \inf\{n\ge 0 : \tilde{S}_n > \xi\}$. Recall that for the base system there exist $\eta>0$ with $\psi_Y(\eta)=0$ and $\psi_Y'(\eta)>0$. And the number of iterations is distributed as a geometric random variable with probability of success $P(T_m=\infty) = 1-E_\eta\exp(-\eta\tilde{S}_{T_m})$

Then for the sth system with $Y_i^s = Y_i/s$ we have $\tilde{S}_n/s > m$ is equivalent to $\tilde{S}_n > sm$. Thus the number of iterations is a Geometric random variable with probability of success

$$P(T_{sm} = \infty) = 1 - E_\eta \exp(-\eta \tilde{S}_{T_{sm}}) \geq 1 - \exp(-\eta sm).$$

Similarly, let $T'_\xi = \inf\{n \geq 0 : \tilde{S}_n < \xi\}$. Define $M_n = \tilde{S}_n + n\epsilon$, then M_n is a martingale with respect to the filtration generated by $\{Y_1, Y_2, ..., Y_n\}$. As $EY_i = -\epsilon < 0$, $P(T_{-m} < \infty) = 1$. By the Optional Sampling Theorem, $EM_{T'_{-m}} = E\tilde{S}_{T'_{-m}} + \epsilon ET'_{-m} = 0$. Thus

$$ET'_{-m} = \frac{m}{\epsilon} - \frac{E[m - S_{T'_{-m}}]}{\epsilon}.$$

Then for the sth system we have

$$ET'_{-sm} = \frac{sm}{\epsilon} - \frac{E[sm - S_{T'_{-sm}}]}{\epsilon}.$$

$(sm - S_{T'_{-sm}})$ converges to the ladder hight Y^- distribution as $s \to \infty$ and $\sup_m E[(sm - S_{T'_{-sm}})^p] < \infty$ yields $E[(Y^-)^p] < \infty$ for $p > 1$ [1]. Therefore, $ET'_{-sm} = O(s)$. □

For the heavy-tailed case, we select the truncation parameter b such that $E[X_n \wedge b] = \mu - 1/2\epsilon$. Then we set $\epsilon' = 1/2\epsilon$ and define $\kappa(A(b))$ as a random time satisfying that $|A_{n+1}| \geq n(E[X_n \wedge b] - \epsilon') = n(\mu - \epsilon)$ for $n \geq \kappa(A(b))$. As $|A_{n+1}| \geq |A_{n+1}(b)|$ under our coupling scheme, we can set $\kappa(A) = \kappa(A(b))$. By Lemma 3.3.2, we have $E\kappa(A) = E\kappa(A(b)) = O(s)$. $\kappa(V)$ is defined as before, a random time satisfying that $V_n \leq n(\mu - \epsilon)$ for $n \geq \kappa(V)$. Then $E\kappa(V) = O(s \log s)$ by Lemma 3.3.1.

As $\kappa_1 = \max\{\kappa(V), \kappa(A(b)\}$, we have $E\kappa = O(s \log s)$. This concludes the proof of Theorem 3.1.3.

3.3.2 Coalescence time for the many-server loss system (Proof of Theorem 3.1.4 and Theorem 3.1.5)

As we are simulating the process backwards in time, it is natural to define the following filtration

$$\overleftarrow{\mathcal{H}}_t = \sigma\{W(-u) : 0 \leq u \leq t\},$$

for which $\overleftarrow{\mathcal{H}}_u \subset \overleftarrow{\mathcal{H}}_t$ for $0 \leq u \leq t$. τ is a stopping time with respect to $\overleftarrow{\mathcal{H}}_t$. We next try to draw connections between the backward process and some forward process. Define

$$\tau^* = \inf\{t + R(t) : \sup_{t \leq u \leq t+R(t)} \{Q(u,0)\} < s, t \geq 0\}.$$

τ^* is a stopping time with respect to $\mathcal{H}_t$ where $\mathcal{H}_t = \sigma\{M(u) : 0 \leq u \leq t\}$. The stochastic process $\{Q(t,0) : t \in \mathbb{R}\}$ has a piecewise constant sample path with a finite number of points of discontinuity on any finite length intervals almost surely. Thus for any fixed $T > 0$, we have

$$\begin{aligned}
P_\pi(\tau > T) &= P_\pi(\bigcap_{-T \leq t \leq 0} (\{R(t) > -t\} \bigcup (\bigcup_{t \leq u \leq (t+R(t)) \wedge 0} (\{Q(u,0) > s\}))) \\
&= P_\pi(\bigcap_{-T \leq t \leq 0} (\{R(T+t) > -t\} \bigcup (\bigcup_{T+t \leq u \leq (T+t+R(T+t)) \wedge T} \{Q(u,0) > s\}))) \\
&= P_\pi(\bigcap_{0 \leq w \leq T} (\{R(w) > T - w\} \bigcup (\bigcup_{w \leq u \leq (w+R(w)) \wedge T} \{Q(u,0) > s\}))) \\
&= P_\pi(\tau^* > T).
\end{aligned}$$

The second equality holds by stationarity; this gives us $E_\pi \tau = E_\pi \tau^*$.

Next, we use a special construction similar to that in Section 4 of [38] to prove the results for $E_\pi^s \tau^*$. The idea is to use a geometric trial argument. We divide the time frame into blocks that are roughly independent. And if the process is well-behaved (staying around its measure-valued fluid limit) on one block, then τ^* is reached before the end of that block.

Let $\bar{Q}(t,y)$ denote the number of customers in the infinite server queue that starts empty at time zero with remaining service time greater than y at time $t \geq 0$. For convenience, we also define $\bar{Q}_u(t,y) = \bar{Q}(u+t,y) - \bar{Q}(u,t+y)$ as the number of customers who arrive after u with remaining service time larger than y at time $u+t$.

3.3.2.1 Proof of Theorem 3.1.4.

We first prove the theorem for the light-tailed case. The heavy-tail case proceeds by selecting the truncation parameter c sufficiently large.

For the QD regime, by "well-behaved", we mean that the process does not deviate δs, for some $\delta > 0$, from its fluid limit. The following lemma states that the probability of not being "well-behaved" decays exponentially fast with the system scale.

Lemma 3.3.3 *Assume $\psi(\theta) < \infty$ for some $\theta > 0$ and X_n's are non-lattice and strictly positive. We also assume the CDF of V_n is continuous. Then for any $\delta > 0$, there exist $I^*(\delta) > 0$, such that*

$$\begin{aligned} & P(\bar{Q}(t,y) > (1+\delta)\lambda s \int_y^{t+y} \bar{F}(u)du \text{ for some } t \in [0,1], y \in [0,\infty)) \\ = \quad & \exp(-sI^*(\delta) + o(s)). \end{aligned}$$

The proof of Lemma 3.3.3 follows form the tow-parameter sample path large deviation result for infinite server queues in [35]. We shall omit it here.

We next introduce our construction of "blocks". Let $l(s) = \inf\{y : (1+\delta)s \int_y^\infty \bar{F}(u)du \leq 1/2\}$, we define the following sequence of random times Ξ_i's: $\Xi_0 := 0$. Given Ξ_{i-1} for $i = 1, 2, \cdots$, define

$$\begin{aligned} r_i &= \inf\{k : k \geq R(\Xi_{i-1}), k = 1, 2, \cdots\}, \\ z &= \inf\{k : k \geq l(s), k = 1, 2, \cdots\}, \\ \Xi_i &= \Xi_{i-1} + r_i + z. \end{aligned}$$

We define a Bernoulli random variable ξ_i, with $\xi_i = 1$ if and only if

$$\bar{Q}_{\Xi_{i-1}+(k-1)t_0}(t,y) \leq (1+\delta)\lambda s \int_y^{t+y} \bar{F}(u)du$$

for all $t \in [0,1]$, $y \in [0,\infty)$ and every $k = 1, 2, \cdots, r_i + z$.

Choose $\delta < 1/\rho - 1$. We first check that $\xi_i = 1$ implies that τ^* is reached before Ξ_i. Since $r_i \geq R(\Xi_{i-1})$, all the customers in the system at time $\Xi_{i-1} + r_i$ will be those who arrive after Ξ_i. Then $\xi_i = 1$ implies that

$$\begin{aligned} Q(\Xi_{i-1} + r_i, y) &\leq \sum_{k=1}^{r_i/t_0} \int_{(k-1)t_0+y}^{kt_0+y} \bar{F}(u)du \\ &= (1+\delta)\lambda s \int_y^{r_i+y} \bar{F}(u)du \\ &\leq (1+\delta)\lambda s \int_y^{\infty} \bar{F}(u)du, \end{aligned}$$

thus, $R(\Xi_{i-1} + r_i) \leq l(s)$.

And for every $t \in (k-1, k]$, $k = 1, 2, ..., z$

$$\begin{aligned} Q(\Xi_{i-1} + r_i + t, y) &\leq (1+\delta)\lambda s \int_y^{r_i+t+y} \bar{F}(u)du \\ &\leq (1+\delta)\lambda s \int_y^{\infty} \bar{F}(u)du, \end{aligned}$$

thus $Q(\Xi_{i-1} + r_i + t, 0) \leq (1+\delta)\rho s \leq s$ for $t \in [0, R(\Xi_{i-1} + r_i)]$.

Now let $N = \inf\{i \geq 1 : \xi_i = 1\}$, then

$$E\tau^* \leq E\sum_{i=1}^{N}(r_i + z).$$

We now show a bound for $E\sum_{i=1}^{N}(r_i + z)$. The proof is given in the Section 3.4.

Lemma 3.3.4 *Assume $\psi(\theta) < \infty$ for some $\theta > 0$ and $\psi_N(\theta)$ is continuously differentiable throughout $\mathbb{R}$. We also assume the CDF of V_n is continuous and $EV_n^q < \infty$ for any $q > 0$. Then*

$$E[\sum_{i=1}^{N}(r_i + z)] = o(s^{\delta})$$

for any $\delta > 0$.

This concludes the proof of the light tailed case. We next extend the theorem to the heavy-tailed case. We prove it by drawing connection to the truncated system.

Here we delicately choose the truncation parameter b so that the truncated system still operating the QD regime. More specifically, we choose b such that

$$\int_b^\infty \bar{G}(x)dx < 1/\rho - 1.$$

This can be achieved since $EX_n = \int_0^\infty \bar{G}(x)dx < \infty$. Then for fixed such b we have

$$\rho_b = \frac{E[V_n]}{E[X_n \wedge b]} = \frac{EV_n}{EX_n - \int_b^\infty \bar{G}(x)dx} < 1$$

and

$$E_\pi^s \tau(b) = o(s^\delta)$$

for any $\delta > 0$, where $\tau(b)$ denote the coalescence time in the truncated system.

We next prove by contradiction that the coalescence in the truncated system implies the coalescence in the original system with the same amount of information simulated. Recall that $\tau(b)$ is a random time satisfying that the system has less than s customers at $\tau(b)$. The maximum remaining service time among all customers in the system at time τ is denoted as $R(\tau(b))$. $R(\tau(b)) \leq |\tau(b)|$ and during $R(\tau(b))$ unites of time from $\tau(b)$ on the system always has less than s customers. We can look for $\tau(b)$ at departure times of customers. We assume the process $Q(t, y)$ is right continuous with left limit, so customers departure at time t will not counted in $Q(t, 0)$. Suppose $\tau(b)$ equals to the departure time of the n-th customer. Then every customer arriving between $\tau(b)$ and $\tau(b) + R(\tau(b))$ sees strictly less than s customers (excluding himself) when he enters the system. We set τ equal to the departure time of the n-th customer in the original system and $R(\tau)$ by definition equals to the maximum remaining service time among all customers in the system at time τ. We have $R(\tau) \leq R(\tau(b))$. We claim that every customer arriving between τ and $\tau + R(\tau)$ must see less than s customers (excluding himself) when he enters the system. Suppose this is not the case. Then there exist a customer m, $1 \leq m \leq n$ who arrives between τ and $\tau + R(\tau)$ and finds at least s customers in the system already. The customer with the same index m must have arrived between $\tau(b)$ and

$\tau(b) + R(\tau(b))$ in the truncated system and $Q(A_m(b)-) \geq Q(A_m-) \geq s$. We get a contradiction. Therefore, we must have seen the coalescence in the original system as well with the same amount of information simulated.

3.3.2.2 Proof of Theorem 3.1.5.

For QED regime, by "well-behaved", we mean that the process does not deviate $C\sqrt{s}$, for some $C > 0$, from its fluid limit. The following lemma states that the probability of both being "well-behaved" and not "well-behaved" are bounded away from zero.

Lemma 3.3.5 *Fix any $\eta > 0$. Let $\nu(y) = (\int_y^\infty \bar{F}(u)du)^{1/(2+\eta)}$. Assume $EX_n^2 < \infty$ and $EV_n^q < \infty$ for any $q > 0$. Then for any large enough C, there exists $\zeta_1(C) > 0$ and $\zeta_2(C) > 0$, such that*

$$P(\bar{Q}(t,y) \leq \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) \geq \zeta_1(C) \quad (3.6)$$

and

$$P(\bar{Q}(t,y) > \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \text{ for some } t \in [0,1], y \in [0,\infty)) \geq \zeta_2(C). \quad (3.7)$$

The proof of Lemma 3.3.5 follows form the proof of Lemma 9 in [38]. Our case is actually simpler, as we are dealing with a one sided bound (upper bound) as appose to the two sided bound in [38]. This simplification allows us to remove the light-tail assumption on interarrival time distribution required in [38]. We shall only briefly outline the procedure here.

For Inequality (3.6), the idea is to consider the diffusion limit of $Q(t,y)$ as a two dimensional Gaussian random field [37], and then invoke Borell-TIS inequality [36].

Assume $EX_n^2 < \infty$, $EV_n < \infty$ and the CDF of V_n is continuous. Pang and Whitt [37] has proved that for the $GI/GI/\infty$ queue with any given initial age $E(0)$,

$$\frac{\bar{Q}(t,y) - \lambda s \int_t^{t+y} \bar{F}(u)du}{\sqrt{s}} \Rightarrow R(t,y) \text{ in } D_{D[0,\infty)}[0,\infty),$$

where $R(t,y) = R_1(t,y)+R_2(t,y)$ is a Gaussian random field with $R_1(t,y) = \lambda \int_0^t \int_0^\infty I(u+x > t+y)dK(u,x)$ and $R_2(t,y) = \lambda c_a^2 \int_0^t \bar{F}(t+y-u)dB(u)$, where $K(u,x) = W(\lambda u, F(x)) - F(x)W(\lambda u, 1)$ in which $W(\cdot,\cdot)$ is a standard Brownian sheet on $[0,\infty)\times[0,1]$ and $B(\cdot)$ is a standard Brownian motion independent of $W(\cdot,\cdot)$. The constant c_a is coefficient of variation of the interarrival times, i.e. $c_a = \sqrt{\mathrm{Var}(X_n)}/EX_n$. We denote

$$\tilde{R}_i(t,y) = \frac{R_i(t,y)}{v(y)}$$

and define the d-metric (a pseudo-metric) for $i = 1, 2$

$$d_i((t,y),(t',y')) = E[(\tilde{R}_1(t,y) - \tilde{R}_2(t',y'))^2]$$

We then invoke the Borell-TIS inequality. We shall skip the verification of the conditions for such invocation here as it is tedious and detailedly proved in [38]. Let $S = [0,1]\times[0,\infty)$. it is shown in [38] that, there exist constants $M_{i,1} > 0$ and $M_{i,2} > 0$, such that $E[\sup_S \tilde{R}_i(t,y)] \le M_{i,1} < \infty$ and $\sup_S E[\tilde{R}_i(t,y)^2] \le M_{i,2} < \infty$. And for $C_i \ge E[\sup_S \tilde{R}_i(t,y)]$, for $i = 1, 2$,

$$P(\sup_S \tilde{R}_i(t,y) \ge C_i) \le \exp\{-\frac{1}{2\sup_S E[\tilde{R}_i(t,y)^2]}(C_i - E[\sup_S \tilde{R}_i(t,y)])^2\}.$$

Let $C \ge 2\max\{E[\sup_S \tilde{R}_1(t,y)], E[\sup_S \tilde{R}_2(t,y)]\}$. Then

$$\begin{aligned} &P(R(t,y) \le C\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) \\ \ge &P(\sup_S \tilde{R}_1(t,y) + \sup_S \tilde{R}_2(t,y) \le C) \\ \ge &P(\sup_S \tilde{R}_1(t,y) \le \frac{C}{2}) P(\sup_S \tilde{R}_2(t,y) \le \frac{C}{2}) > 0. \end{aligned}$$

Let X_0 denote the interarrival time of the first customer and V_0 denote its service time. We also denote $\bar{Q}^0(t,y)$ as an independent infinite server process starting empty

and with $E(0) = 0$. Then for s large enough, we have

$$\begin{aligned}
&P(\bar{Q}(t,y) \le \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) \\
=&P(\bar{Q}^0(t - X_0, y) + 1\{V_0 > t + y\} \le \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \\
&\text{for all } t \in [X_0, 1], y \in [0,\infty)) \\
\ge&P(\bar{Q}^0(t,y) + 1\{V_0 > t + X_0 + y\} \le \lambda s \int_y^{t+X_0+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \\
&\text{for all } t \in [0, 1 - X_0], y \in [0,\infty)) \\
\ge&P(\bar{Q}^0(t,y) \le \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) \\
=&P(\frac{\bar{Q}^0(t,y) - \lambda s \int_y^{t+y} \bar{F}(u)du}{\sqrt{s}} \le C\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)).
\end{aligned}$$

It is easy to check that the set $\{f : |f(t,y)| \le C\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)\}$ is a continuity set, thus by the Functional Central Limit Theorem result in [37], we have

$$\begin{aligned}
&P(\frac{\bar{Q}^0(t,y) - \lambda s \int_y^{t+y} \bar{F}(u)du}{\sqrt{s}} \le C\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) \\
\to&P(R(t,y) \le C\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) > 0.
\end{aligned}$$

Inequality (3.7) is easy to prove as we can always isolate a point (t^*, y^*) inside S. The projection of the process on that point posses Gaussian distribution. More specifically,

$$\begin{aligned}
&P(\bar{Q}(t,y) > \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \text{ for some } t \in [0,1], y \in [0,\infty)) \\
\ge&P(\bar{Q}(t^*,y^*) > \lambda s \int_{y^*}^{t^*+y^*} \bar{F}(u)du + C\sqrt{s}\nu(y^*)) \\
=&P(\frac{\bar{Q}(t^*,y^*) - \lambda s \int_{y^*}^{t^*+y^*} \bar{F}(u)du}{\sqrt{s}} > C\nu(y^*)),
\end{aligned}$$

and by Fatou's lemma

$$\liminf_{s\to\infty} P(\frac{\bar{Q}(t^*,y^*) - \lambda s \int_{y^*}^{t^*+y^*} \bar{F}(u)du}{\sqrt{s}} > C\nu(y^*)) \ge P(R(t^*,y^*) > C\nu(y^*)) > 0.$$

Let $m(s) = \inf\{y : C\sqrt{s}(v(y) + \int_y^\infty v(s)ds) \le \frac{1}{2}\}$. Following the same construction as for the QD regime, we define the sequence of random times Ξ_i's as follows: $\Xi_0 := 0$. Given Ξ_{i-1} for $i = 1, 2, \cdots$,

$$\begin{aligned} r_i &= \inf\{k : k \ge R(\Xi_{i-1}), k = 1, 2, ...), \\ z &= \inf\{k : k \ge m(s), k = 1, 2, ...\}, \\ \Xi_i &= \Xi_{i-1} + r_i + z. \end{aligned}$$

We introduce a Bernoulli random variable ξ_i with $\xi_i = 1$ if and only if

$$\bar{Q}_{\Xi_{i-1}+(k-1)t_0}(t, y) \le \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y)$$

for all $t \in [0, 1]$, $y \in [0, \infty)$ and every $k = 1, 2, ..., r_i + z$.

We next show that $\xi_i = 1$ implies that τ^* is reached before Ξ_i. Since $r_i \ge R(\Xi_{i-1})$, all the customers at time $\Xi_{i-1} + r_i$ will be those arrive after Ξ_i. Thus we have $\xi_i = 1$ implies that

$$\begin{aligned} Q(\Xi_{i-1} + r_i, y) &\le \sum_{k=1}^{r_i} \{\lambda s \int_{(k-1)t_0+y}^{kt_0+y} \bar{F}(u)du + C\sqrt{s}\nu((k-1)+y))\} \\ &\le \lambda s \int_y^\infty \bar{F}(u)du + C\sqrt{s}(\nu(y) + \int_y^\infty \nu(u)du). \end{aligned}$$

As $\int_y^\infty \bar{F}(u)du$ decays faster than $\nu(y)$ as y grows large, for s large enough, we have $R(\Xi_{i-1} + r_i) < m(s)$.

Likewise for every $t \in (k-1, k]$ and $k = 1, 2, \cdots, z$,

$$Q(\Xi_{i-1} + r_i + t, y) \le \lambda s \int_y^\infty \bar{F}(u)du + C\sqrt{s}(\nu(y) + \int_y^\infty \nu(u)du).$$

Thus when $\beta > C(\nu(0) + \int_0^\infty \nu(u)du)$, for $t \in [0, R(\Xi_{i-1} + r_i)]$, we have

$$Q(\Xi_{i-1} + r_i + t, 0) \le s + C(\nu(0) + \int_0^\infty \nu(u)du)\sqrt{s} \le s + \beta\sqrt{s}$$

Now let $N = \inf\{i \ge 1 : \xi_i = 1\}$. Then $E\tau^* \le E[\sum_{i=1}^N (r_i + z)]$. We next show a bound for $E\sum_{i=1}^N (r_i + z)$. The proof is given in the Section 3.4.

Lemma 3.3.6 *Assume* $EX_n^2 < \infty$ *and* $EV_n^q < \infty$ *for any* $q > 0$. *Then*

$$\log E[\sum_{i=1}^{N}(r_i + z)] = o(s^\delta)$$

for any $\delta > 0$.

Notice that our proof of Theorem 3.1.5 only requires the existence of the second moment of the interarrival time distribution. We thus conclude the proof of Theorem 3.1.5.

3.3.2.3 Numerical experiment

In this subsection, we run some numerical experiments aimed at verifying the running time of our algorithm measured by $E_\pi^s[\tau]$ for different values of s. The algorithms appear to have substantially better performance in practice. In the QD regime, our numerical experiments suggest that $E_\pi^s[\tau]$ is almost bounded as apposed to grow sub-linearly with s indicated by Theorem 3.1.4. This is because in the QD regime, the stationary probability that the queue length process is above C_s decays exponentially with the system scale s. In the QED regime, our numerical experiments suggest a growth rate of $O(\sqrt{s})$ as apposed to the sub-exponentially growth rate in Theorem 3.1.5. This empirical bound is intuitive, as in the QED regime, the situation when coalescence occurs is similar to the case when a mean zero random walk spends s units of time below 0. If the increments of the random walk have finite variance, this situation occurs with probability $O(1/\sqrt{s})$.

The performance was tested using a wide range of distributions and the overall conclusions are similar. The numbers displayed (Table 3.1 and Table 3.2) are obtained assuming that a generic base interarrival time, X_n, follows a Gamma distribution with shape parameter 2 and rate parameter 2 ($\Gamma(2, 2)$). For the s-th system, the interarrival is distributed as X_n/s, and a generic service time, V_n, follows lognormal distribution, where $\log V_n \sim N(-1/2, 1/2)$. We use 1000 replications for each value of s.

Table 3.1: Simulation results for τ (QD: $\lambda = s, C_s = 1.2s$)

s	mean	95% confidence interval
100	22.6297	$[21.3381, 23.9213]$
500	15.6162	$[15.1791, 16.0533]$
1000	15.8816	$[15.4559, 16.3073]$

Table 3.2: Simulation results for τ (QED: $\lambda = s, C_s = s + 2\sqrt{s}$)

s	mean	95% confidence interval
100	22.6297	$[21.3381, 23.9213]$
500	37.0449	$[32.7770, 41.3128]$
1000	42.0704	$[37.9622, 46.1786]$

3.4 Proof of Lemma 3.3.4 and Lemma 3.3.6

We first prove the following two lemmas (Lemma 3.4.1 and Lemma 3.4.2) as a preparation.

Lemma 3.4.1 *If $EV_n^q < \infty$ for any $q > 0$, then for any fixed $p > 0$,*

$$E[(\max_{k=1,2,\ldots n} V_k)^p] = o(n^\delta)$$

for any $\delta > 0$.

Proof. For any fixed $\delta > 0$ we can find $\delta' \in (0, \delta)$. Let $q = 1/\delta' + p$. By Chebyshev's inequality we have

$$\bar{F}(u) \leq \frac{EV^q}{u^q}.$$

Let $\bar{F}_n(u) = P(\max_{k=1,2,\ldots,n} V_k > u)$ then

$$\begin{aligned} E[(\max_{k=1,2,\ldots n} V_k)^p] &= p\int_0^\infty u^{p-1}\bar{F}_n(u)du \\ &\leq n^{1/(q-p)} + np\int_{n^{1/(q-p)}}^\infty u^{p-1}\bar{F}(u)du \\ &\leq n^{1/(q-p)} + np\int_{n^{1/(q-p)}}^\infty \frac{EV^q}{u^{q-p+1}}du \\ &= n^{\delta'} + \frac{p}{q-p}EV^q. \end{aligned}$$

□

$$\begin{aligned} E[\sum_{i=1}^N (r_i+z)] &= E[\sum_{i=1}^\infty (r_i+z)I\{N\geq i\}] \\ &\leq \sum_{i=1}^\infty E[(r_i+z)^2]^{1/2}P(N\geq i)^{1/2} \text{ by Holder's inequality.} \end{aligned}$$

Lemma 3.4.2 *If $EX_n < \infty$ and $EV_n^q < \infty$ for any $q > 0$, then for any $p \geq 1$ we have*

$$E[(r_i+z)^p]^{1/p} = o(s^\delta)$$

for any $\delta > 0$.

Proof. By Minkowski inequality, $E[(r_i+z)^p]^{1/p} \leq E[r_i^p]^{1/p} + z$. Using similar argument as in the proof of Lemma 3.4.1, we can show that $l(s) = o(s^\delta)$ for any $\delta > 0$, thus $z = o(s^\delta)$ for any $\delta > 0$.

For fixed $\delta > 0$, we can find $\delta' \in (0, p\delta/(1+p\delta))$, such that

$$\begin{aligned} E[r_i^p] &\leq E[E[(\max_{k=1,\ldots,N_s(\Xi_{i-1})-N_s(\Xi_{i-2})} V_k)^p | N_s(\Xi_{i-1}) - N_s(\Xi_{i-2})]] \\ &\leq CE[(N_s(\Xi_{i-1}) - N_s(\Xi_{i-2}))^{\delta'}] \text{ Lemma 3.4.1} \\ &\leq C(E[N_s(\Xi_{i-1}) - N_s(\Xi_{i-2})])^{\delta'} \text{ Jensen's inequality for concave function} \\ &\leq C\tilde{\lambda}^{\delta'}s^{\delta'}E[r_{i-1}+z]^{\delta'} \text{ Key Renewal Theorem.} \end{aligned}$$

Let $w_i = r_i + z$ for $i = 1, 2, \cdots$. As z it is a constant that only depends on s and $z = o(s^{\delta'})$, then $Ew_i \geq z \geq 1$ and $Ew_i = Er_i + z \leq C\tilde{\lambda}^{\delta'}s^{\delta'}(Ew_{i-1})^{\delta'} + z \leq \tilde{C}s^{\delta'}(Ew_{i-1})^{\delta'}$,

where $\tilde{C} = C\tilde{\lambda}^{\delta'} + 1$. As $E[r_1^p] = E_\pi[R(0)^p] = o(s^{\delta'})$. By iteration we have

$$Ew_i \leq \tilde{C}^{1/(1-\delta')} s^{\delta'/(1-\delta')} \text{ for } i = 1, 2, \cdots$$

Thus $Er_i^p = o(s^{p\delta})$ and $E[(r_i + z)^p]^{1/p} = o(s^\delta)$. □

Proof.[Proof of Lemma 3.3.4] We first notice that $P(\xi_i = 0) \leq E[w_1] \exp(-sI^*(\delta) + o(s))$ by Lemma 3.3.3.

$$P(N \geq 1) = 1.$$

$$P(N \geq 2) = P(\xi_1 = 0) \leq E[w_1] \exp(-sI^*(\delta) + o(s))$$

Recall that $w_i = r_i + z$ for $i = 1, 2, \cdots$.

$$\begin{aligned} P(N \geq 3) &= P(N \geq 1)P(N \geq 3|N \geq 2) \\ &= P(\xi_1 = 0)P(\xi_2 = 0|\xi_1 = 0) \\ &\leq P(\xi_1 = 0)E[w_2|\xi_1 = 0] \exp(-sI^*(\delta) + o(s)) \\ &\leq E[w_1]E[w_2|\xi_1 = 0] \exp(-2sI^*(\delta) + o(s)). \end{aligned}$$

We next prove that $E[w_2|\xi_1 = 0] = \exp(o(s))$. Notice that $P(\xi_i = 0) \geq \exp(-sI^*(\delta) + o(s))$ by Lemma 3.3.3. Then for any $p > 0$, $q > 0$ and $1/p + 1/q = 1$,

$$\begin{aligned} E[w_2|\xi_1 = 0] &= \frac{E[w_2 I\{\xi_1 = 0\}]}{P(\xi_1 = 0)} \\ &\leq \frac{E[w_2^p]^{1/p} P(\xi_1 = 0)^{1/q}}{P(\xi_1 = 0)} \text{ Holder's inequality} \\ &\leq E[w_2^p]^{1/p} E[w_1]^{1/q} \exp(\frac{1}{p} sI^*(\delta) + o(s)), \end{aligned}$$

thus

$$\frac{1}{s} \log E[w_2|\xi_1 = 0] \leq \frac{1}{s}(\frac{1}{p} \log E[w_2^p] + \frac{1}{q} \log E[w_1] + o(s)) + \frac{1}{p} I^*(\delta).$$

By sending p to infinity, we have $E[w_2|\xi_1 = 0] = \exp(o(s))$.

Similarly by iteration,

$$P(N \geq k) = \exp(-ksI^*(\delta) + o(s)) \text{ for } k = 4, 5, \ldots$$

Then $\sum_{i=1}^{\infty} P(N \geq i)^{1/2} = O(1)$. As $E[\sum_{i=1}^{N}(r_i+z)] \leq \sum_{i=1}^{\infty} E[(r_i+z)^2]^{1/2} P(N \geq i)^{1/2}$ and $E[(r_i+z)^2]^{1/2} = o(s^\delta)$ for any $\delta > 0$, we have $E[\sum_{i=1}^{N}(r_i+z)] = o(s^\delta)$. □

Proof.[Proof of Lemma 3.3.6]

$$P(N \geq 1) = 1.$$

$$\begin{aligned} P(N \geq 2) &= P(\xi_1 = 0) \\ &\leq 1 - E[\zeta_1(C)^{w_1}] \text{ Lemma 3.3.5} \\ &\leq 1 - \zeta_1(C)^{E[w_1]} \text{ Jensen's inequality} \\ &= 1 - b\exp(-o(s^\delta)). \end{aligned}$$

Moreover,

$$\begin{aligned} P(N \geq 3) &= P(N > 2|N > 1)P(N > 1) \\ &= P(\xi_2 = 0|\xi_1 = 0)P(\xi_1 = 0) \\ &\leq E[1 - \zeta_1(C)^{w_2}|\xi_1 = 0]P(\xi_1 = 0) \\ &\leq (1 - \zeta_1(C)^{E[w_2|\xi_1=0]})P(\xi_1 = 0). \end{aligned}$$

We next show that $E[w_2|\xi_1 = 0] = o(s^\delta)$ for any $\delta > 0$. Notice that $P(\xi_i = 0) \geq \zeta_2(C)$ by Lemma 3.3.5, then

$$E[w_2|\xi_1 = 0] = \frac{E[w_2 I\{\xi_1 = 0\}]}{P(\xi_1 = 0)} \leq \frac{Ew_2}{\zeta_2(C)}$$

Similarly by iteration we have

$$P(N \geq k) \leq (1 - b\exp(-o(s^\delta)))^k \text{ for any } \delta > 0 \text{ and } k = 4, 5, \ldots$$

Then for any $\delta > 0$, $\log \sum_{i=1}^{\infty} P(N \geq i)^{1/2} = o(s^\delta)$. As $E[\sum_{i=1}^{N}(r_i+z)] \leq \sum_{i=1}^{\infty} E[(r_i+z)^2]^{1/2} P(N \geq i)^{1/2}$ and $E[(r_i+z)^2]^{1/2} = o(s^\delta)$ for any $\delta > 0$, we have $\log E[\sum_{i=1}^{N}(r_i+z)] = o(s^\delta)$. □

Chapter 4

ε-Strong Simulation for Multidimensional Stochastic Differential Equations via Rough Path Analysis

Consider the a multi-dimensional Stochastic Differential Equation (SDE)

$$dX(t) = \mu(X(t))dt + \sigma(X(t))dZ(t)\ ,\ X(0) = x(0) \tag{4.1}$$

where $Z(\cdot)$ is a d'-dimensional Brownian motion, and $\mu(\cdot) : R^d \to R^d$ and $\sigma(\cdot) : R^d \to R^{d\times d'}$ satisfy suitable regularity conditions. We shall assume, in particular, that both $\mu(\cdot)$ and $\sigma(\cdot)$ are Lipschitz continuous so that a strong solution to the SDE is guaranteed to exist. Additional assumptions on the first and second order derivatives of $\mu(\cdot)$ and $\sigma(\cdot)$ which are standard in the theory of rough paths will be discussed in the sequel.

Our contribution in this chapter is the joint construction of $X = \{X(t) : t \in [0,1]\}$ and a family of processes $X_\varepsilon = \{X_\varepsilon(t) : t \in [0,1]\}$, for each $\varepsilon \in (0,1)$, supported on a probability space $(\Omega, \mathcal{F}, P)$, and such that the following properties hold:

(T1) The process X_ε is piecewise constant, with finitely many discontinuities in $[0,1]$.

(T2) The process X_ε can be simulated exactly and, since it takes only finitely many values, its path can be fully stored.

(T3) We have that with P-probability *one*

$$\sup_{t\in[0,1]} ||X_\varepsilon(t) - X(t)||_\infty < \varepsilon. \tag{4.2}$$

(T4) For any $m > 1$ and $0 < \varepsilon_m < ... < \varepsilon_1 < 1$ we can simulate X_{ε_m} conditional on X_{ε_1},...,$X_{\varepsilon_{m-1}}$.

We refer to the family of procedures that achieve the construction of such family $\{X_\varepsilon : \varepsilon \in (0,1)\}$ as ε-strong simulation methods.

Our construction requires a detailed study of continuity estimates of the Ito map using Lyon's theory of rough paths. We approximate the underlying Brownian motion, jointly with the Lévy areas with a deterministic ε error in the underlying rough path metric.

4.1 Main Results

Our approach consists in studying the process X as a transformation of the underlying Brownian motion Z. Such transformation is known as the Ito-Lyons map and its continuity properties are studied in the theory of rough paths, pioneered by T. Lyons, in [39]. A rough path is a trajectory of unbounded variation. The theory of rough paths allows to define the solution to an SDE such as (4.1) in a path-by-path basis (free of probability) by imposing constraints on the regularity of the iterated integrals of the underlying process Z. Namely, integrals of the form

$$A_{i,j}(s,t) = \int_s^t (Z_i(u) - Z_i(s))\, dZ_j(u). \tag{4.3}$$

The theory results in different interpretations of the solution to (4.1) depending on how the iterated integrals of Z are interpreted. In this paper, we interpret the integral in (4.3) in the sense of Ito.

It turns out that the Ito-Lyons map is continuous under a suitable α-Hölder metric defined in the space of rough paths. In particular, such metric can be expressed as the maximum of the following two quantities:

$$||Z||_\alpha := \sup_{0\le s<t\le 1} \frac{||Z(t)-Z(s)||_\infty}{|t-s|^\alpha}, \tag{4.4}$$

$$||A||_{2\alpha} := \sup_{0\le s<t\le 1} \max_{1\le i,j\le d'} \frac{|A_{i,j}(s,t)|}{|t-s|^{2\alpha}}. \tag{4.5}$$

In the case of Brownian motion, as we consider here, we have that $\alpha \in (1/3, 1/2)$. It is shown in [40], that under suitable regularity conditions on $\mu(\cdot)$ and $\sigma(\cdot)$, which we shall discuss momentarily, the Euler scheme provides an almost sure approximation in uniform norm to the solution to the SDE (4.1). Our first result provides an explicit characterization of all of the (path-dependent) quantities that are involved in the final error analysis (such as $||Z||_\alpha$ and $||A||_{2\alpha}$),the difference between our analysis and what has been done in previous developments is that ultimately we must be able to implement the Euler scheme jointly with the path-dependent quantities that are involved in the error analysis. So, it is not sufficient to argue that there exists a path-dependent constant that serves as a bound of some sort, we actually must provide a suitable representation that can be simulated in finite time.

In order to provide our first result, we introduce some notations. Let D_n denote the dyadic discretization of order n and Δ_n denote the mesh of the discretization. Specifically, $D_n := \{t_0^n, t_1^n, \ldots, t_{2^n}^n\}$ where $t_k^n = k/2^n$ for $k = 0, 1, 2, \ldots, 2^n$ and $\Delta_n = 1/2^n$. Suppose we have a discretized approximation scheme.

Given $\hat{X}^n(0) = x(0)$, define $\{\hat{X}^n(t) : t \in D_n\}$ by the following recursion:

$$\hat{X}_i^n(t_{k+1}^n) = \hat{X}_i^n(t_k^n) + \mu_i(\hat{X}^n(t_k^n))\Delta_n + \sigma_i(\hat{X}^n(t_k^n))(Z(t_{k+1}^n) - Z(t_k^n)), \tag{4.6}$$

and let $\hat{X}^n(t) = \hat{X}^n(\lfloor t \rfloor)$ where $\lfloor t \rfloor = \max\{t_k^n : t_k^n \leq t\}$ for $t \in [0,1]$. We denote

$$R_{i,j}^n(t_l^n, t_m^n) := \sum_{k=l+1}^{m} A_{i,j}(t_{k-1}^n, t_k^n).$$

and for fixed $\beta \in (1-\alpha, 2\alpha)$, write

$$\Gamma_R := \sup_n \sup_{0\leq s<t\leq 1, s,t\in D_n} \max_{1\leq i,j\leq d'} \frac{|R_{i,j}^n(s,t)|}{|t-s|^\beta \Delta_n^{2\alpha-\beta}}.$$

We also redefine $||Z||_\alpha$ and $||A||_{2\alpha}$ as

$$||Z||_\alpha := \sup_n \sup_{0\leq s<t\leq 1, s,t\in D_n} \frac{||Z(t)-Z(s)||_\infty}{|t-s|^\alpha},$$
$$||A||_{2\alpha} := \sup_n \sup_{0\leq s<t\leq 1, s,t\in D_n} \max_{1\leq i,j\leq d'} \frac{|A_{i,j}(s,t)|}{|t-s|^{2\alpha}}.$$

The new definitions are equivalent to (4.4) and (4.5) since both Z and A are continuous processes.

Theorem 4.1.1 *Suppose that there exists a constant M such that $||\mu||_\infty \leq M$, $||\mu'||_\infty \leq M$ and $||\sigma^{(i)}||_\infty \leq M$ for $i = 0,1,2,3$. Then it is well known that a solution to X can be constructed path-by-path (see [40] and Section 4.3). Moreover, if $||Z||_\alpha \leq K_\alpha < \infty$, $||A||_{2\alpha} \leq K_{2\alpha} < \infty$, and $\Gamma_R < K_R$, we can compute G explicitly in terms of M, K_α, $K_{2\alpha}$ and K_R, such that*

$$\sup_{t\in[0,1]} ||\hat{X}^n(t) - X(t)||_\infty \leq G\Delta_n^{2\alpha-\beta}.$$

Remark 4.1.1 *A recipe that explains step-by-step how to compute G is given in the appendix to this section.*

Using Theorem 4.1.1, we can proceed to state the main contribution of this paper.

Theorem 4.1.2 *In the context of here Theorem 4.1.1, there is an explicit Monte Carlo procedure that allows to simulate random variables K_α, $K_{2\alpha}$and K_R jointly with $\{Z(t) : t \in D_n\}$ for any $n \geq 1$. In turn, given any deterministic $\varepsilon > 0$ we can select N_0 sufficiently large, such that for $n \geq N_0$*

$$\sup_{t\in[0,1]} ||\hat{X}^n(t) - X(t)||_\infty \leq \varepsilon, \tag{4.7}$$

with probability one. Moreover, conditional on $\hat{X}^n(\cdot)$ we can subsequently refine our approximation to produce $\hat{X}^{n'}(\cdot)$ with the property that $\sup_{t\in[0,1]} ||\hat{X}^{n'}(t)-X(t)||_\infty \leq \varepsilon'$ for any $\varepsilon' < \varepsilon$.

Remark 4.1.2 *An explicit description of the algorithm involved in the Monte Carlo procedure of Theorem 4.1.2 is given in Algorithm II at the end of Section 4.2.5 and the discussion that follows it. The discussion in the remark that follows Algorithm II explains how to further refine the discretization to obtain $\hat{X}^{n'}(\cdot)$.*

4.1.1 On Relaxing Boundedness Assumptions

The construction of $\hat{X}^n(\cdot)$ in order to satisfy (4.7) assumes that $||\mu||_\infty \leq M$, $||\mu^{(1)}||_\infty \leq M$ and $||\sigma^{(i)}||_\infty \leq M$ for $i = 0, 1, 2, 3$. While these assumptions are strong we can relax them. In particular, as we shall argue now. Theorem 4.1.2 extends directly to the case in which μ and σ are Lipschitz continuous, with μ differentiable and σ is three times differentiable. Since μ and σ are Lipschitz continuous we know that $X(\cdot)$ has a strong solution which is non-explosive.

We can always construct μ_M and σ_M so that $\mu^{(i)}(x) = \mu_M^{(i)}(x)$ for $\|x\|_\infty \leq c_M$ and $i = 0, 1$, and $\sigma^{(i)}(x) = \sigma_M^{(i)}(x)$ for $\|x\|_\infty \leq c_M$ for $i = 0, 1, 2, 3$. Also we can choose $c_M \to \infty$, and $||\mu_M||_\infty \leq M$, $||\mu_M^{(1)}||_\infty \leq M$ and $||\sigma_M^{(i)}||_\infty \leq M$ for $i = 0, 1, 2, 3$.

For $M \geq 1$ we consider the SDE (4.1) with μ_M and σ_M as drift and diffusion coefficients, respectively, and let $X_M(\cdot)$ be the corresponding solution to (4.1). We start by picking some $M_0 \geq 1$ such that $\varepsilon < c_{M_0}$ and let $M = M_0$. Then run Algorithm II to produce $\{\hat{X}_M^n(t) : t \in [0, 1]\}$, which according to Theorem 4.1.2 satisfies,

$$\sup_{t\in[0,1]} ||\hat{X}_M^n(t) - X_M(t)||_\infty \leq \varepsilon.$$

Note that only Steps 5 to 8 in Algorithm II depend on the SDE (4.1), through the evaluation of G, which depends on M and so we write $G_M := G$. If

$$\sup_{t\in[0,1]} ||\hat{X}_M^n(t)||_\infty \leq c_M - \varepsilon,$$

then we must have that $X(t) = X_M(t)$ for $t \in [0,1]$ and we are done. Otherwise, we let $M \longleftarrow 2M$ and run again only Steps 5 to 8 of Algorithm II. We repeat doubling M and re-running Steps 5 to 8 (updating G_M) until we obtain a solution for which $\sup_{t\in[0,1]} ||\hat{X}^n_M(t)||_\infty \leq c_M - \varepsilon$. Eventually this must occur because

$$\lim_{M\to\infty} \sup_{t\in[0,1]} ||X_M(t) - X(t)||_\infty = 0$$

almost surely and $X(\cdot)$ is non explosive.

The rest of the chapter is organized as follows. Section 4.2 is divided into three subsections and it builds the elements behind the proof of Theorem 4.1.2. As it turns out, one needs to simulate bounds on the so-called Hölder norms of the underlying Brownian motion and the corresponding Lévy areas. So, Section 4.2 first studies some basic estimates of for Brownian motion obtained out of its wavelet synthesis. Section 4.3 is also divided in several parts, corresponding to the elements of rough path theory required to analyze the SDE described in (4.1) as a continuous map of Brownian motion under a suitable metric (described in Section 4.1). While the final form of the estimates in Section 4.3 might be somewhat different than those obtained in the literature on rough path analysis, the techniques that we use there are certainly standard in that literature. We have chosen to present the details because the techniques might not be well known to the Monte Carlo simulation community and also because our emphasis is in finding explicit constants (i.e. bounds) that are amenable to simulation.

4.1.2 The Evaluation of G.

We next summarize the way to calculate G in terms of M, $||Z||_\alpha$, $||A||_{2\alpha}$ and Γ_R.

Procedure A.

S1. Find δ and $C_i(\delta)$ for $i = 1, 2, 3$ that satisfies the following relations:

$$\begin{aligned}
C_1(\delta) &\geq C_3(\delta)\delta^{2\alpha} + M\delta^{1-\alpha} + dM||Z||_\alpha + d^3M^2||A||_{2\alpha}\delta^\alpha \\
C_2(\delta) &\geq C_3(\delta)\delta^\alpha + d^3M^2||A||_{2\alpha} \\
C_3(\delta) &\geq \frac{2}{1-2^{1-3\alpha}}\big(MC_1(\delta) + dMC_1(\delta)^2||Z||_\alpha + d^2MC_2(\delta)||Z||_\alpha \\
&\qquad +2d^3M^2C_1(\delta)||A||_\alpha\big)
\end{aligned}$$

S2. Set $C_1 = \frac{2}{\delta}C_1(\delta)$, $C_2 = \frac{2}{\delta}(C_2(\delta) + MC_1 + dMC_1||Z||_\alpha)$ and

$$C_3 = \frac{2}{1-2^{1-3\alpha}}(MC_1 + dMC_1^2||Z||_\alpha + d^2MC_2||Z||_\alpha + 2d^3M^2C_1||A||_\alpha)$$

S3. Find δ' and $B_i(\delta')$ for $i = 1, 2, 3$ that satisfies the following relations:

$$\begin{aligned}
B_1(\delta') &> B_3(\delta')\delta'^{2\alpha} + 2M\delta'^{1-\alpha} + 2M||Z||_\alpha + 4M^2||A||_{2\alpha}\delta'^\alpha \\
B_2(\delta') &> B_3(\delta')\delta'^\alpha + 4M^2||A||_{2\alpha} \\
B_3(\delta') &> \frac{4}{1-2^{1-3\alpha}}\big(MB_1(\delta') + MB_1(\delta'^2||Z||_\alpha + MB_2(\delta')||Z||_\alpha \\
&\qquad +2M^2B_1(\delta')||A||_\alpha\big)
\end{aligned}$$

S4. Set $B = \frac{2}{\delta'}B_1(\delta')$

S5. Set $G_1 = (1 + B)C_3$

S6. Find δ'' and $C_4(\delta'')$ such that

$$\begin{aligned}
B\delta''^\alpha &\leq 2^{\alpha+\beta} - 2 \\
C_4(\delta'') &\geq 2(1 - \frac{2 + B\delta''^\alpha}{2^{\alpha+\beta}})^{-1}(Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)
\end{aligned}$$

S7. Set $C_4 = (1 + B\delta''^\alpha)C_4(\delta''^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)/\delta''$

S8. Set $G_2 = C_4 + d^3M^2\Gamma_R$

S9. Set $G = G_1 + G_2$.

4.2 Use of Wavelets to Bound α-Hölder Norms, Tolerance-Enforced Simulation, and The Proof of Theorem 4.1.2.

Our goal in this section is to simulate the upper bounds for $||Z||_\alpha$, $||A||_{2\alpha}$ and Γ_R respectively. We will first recall Lévy-Ciesielski's construction of Brownian motion and provide a high level picture of the approach that we will follow based on "record breakers".

4.2.1 Wavelet Synthesis of Brownian Motion and Record Breakers

We do so by implementing the Lévy-Ciesielski construction of Brownian motion which is explained next following [41] pages 34-39. First we need to define a step function $H(\cdot)$ on $[0,1]$ by

$$H(t) = I\left(0 \leq t < 1/2\right) - I\left(1/2 \leq t \leq 1\right).$$

Then define a family of functions

$$H_k^n(t) = 2^{(n-1)/2} H(2^{n-1}t - (k-1))$$

for all $n \geq 0$ and $1 \leq k \leq 2^{n-1}$. Set $H_0^0(t) = 1$ and then one obtains the following.

Theorem 4.2.1 (Lévy-Ciesielski Construction) *If $\{W_k^n : 0 \leq k < 2^n, n \geq 0\}$ is a sequence of independent standard normal random variables, then the series defined by*

$$Z(t) = W_0^0 \int_0^t H_0^0(s)\, ds + \sum_{n=1}^{\infty} \sum_{k=1}^{2^{n-1}} \left(W_k^n \int_0^t H_k^n(s)\, ds \right) \tag{4.8}$$

converges uniformly on $[0,1]$ with probability one. Moreover, the process $\{Z(t) : t \in [0,1]\}$ is a standard Brownian motion on $[0,1]$.

Eventually we will simulate the series up to a random but finite level N which can be viewed as the order of dyadic discretization. The level N is selected so that the contribution of the remaining terms (terms beyond level N) can be guaranteed to be bounded by a user defined tolerance error. We think of simulating the discretization levels sequentially, so we often refer to "time" when discussing levels.

Once we have simulated up to time (or level) N, we further decompose the analysis into two parts. One dealing with finding the upper bound for $||Z||_\alpha$ (Section 4.2.2), the other dealing with finding the upper bound for Γ_R (Section 4.2.3). We then combine these two parts to obtain an upper bound for $||A||_{2\alpha}$.

For both parts, we use a strategy based on a suitably defined sequence of "record breakers". We ask a "yes or no" question to the future (i.e. to higher order discretization levels). The question, which corresponds to the simulation of a Bernoulli random variable, is "will there be a new record breaker?" The definition of record breakers need to satisfy the following two conditions.

Conditions:

1. The following event happens with probability one: beyond some random but finite time, there will be no more record breakers.

2. By knowing that there are no more record breakers, the contribution of the terms that we have not simulated yet are well under control (i.e. bounded by a user defined tolerance error).

Now we explain for each part, how the above strategy is applied. We have d' independent Brownian motions and we will use $W^n_{i,k}$ for $i \in \{1, ..., d'\}$ to denote the (n, k) coefficient in the expansion (4.8) for the i-th Brownian motion.

For $||Z||_\alpha$ we say a record is broken at (i,n,k), for $1 \le i \le d'$, $n \ge 0$ and $0 \le k < 2^n$, if

$$|W^n_{i,k}| > 4\sqrt{n+1}.$$

Let $N_1 := \max\{n \ge 1 : |W^n_{i,k}| > 4\sqrt{n+1} \text{ for some } 1 \le k \le 2^{n-1}\}$. Lemma 4.2.3 shows that $E[N_1] < \infty$. Thus Condition 1 for "record breaker" is satisfied. We then check Condition 2. By Lemma 4.2.2, we have $||Z||_\alpha \le 2^{2\alpha} \sum_{n=0}^{\infty} 2^{-n(1/2-\alpha)} V^n$ where $V^n = \max_{1\le k\le 2^{n-1}} |W^n_k|$. Once we found N_1, we have

$$\begin{aligned} ||Z||_\alpha &\le \sum_{n=1}^{\lceil \log_2 N_1 \rceil} 2^{-n(1/2-\alpha)} V^n + \sum_{\lceil \log_2 N_1 \rceil + 1}^{\infty} 2^{-n(1/2-\alpha)} \sqrt{n+1} \\ &\le \sum_{n=1}^{\lceil \log_2 N_1 \rceil} 2^{-n(1/2-\alpha)} V^n + \frac{(\lceil \log_2 N_1 \rceil + 1)^{-1/2(1/2-\alpha)}}{1 - 2^{-1/2(1/2-\alpha)}}. \end{aligned}$$

For Γ_R, we first define a sequence of random walks

$$\begin{aligned} &L^n_{i,j}(0) := 0, \\ &L^n_{i,j}(k) := L^n_{i,j}(k-1) + \left(Z_i(t^n_{2k-1}) - Z_i(t^n_{2k-2})\right)\left(Z_j(t^n_{2k}) - Z_j(t^n_{2k-1})\right), \end{aligned}$$

for $k = 1, 2, \cdots, 2^{n-1}$.

We then say a record is broken at (n,k,k'), for $n \ge 1, 0 \le k < k' < 2^{n-1}$, if

$$|L^n_{i,j}(k') - L^n_{i,j}(k)| > (k'-k)^\beta \Delta_n^{2\alpha}.$$

Let $N_2 := \max\{n \ge 1 : |L^n_{i,j}(k') - L^n_{i,j}(k)| > (k'-k)^\beta \Delta_n^{2\alpha} \text{ for some } 0 \le k < k' \le 2^{n-1}\}$. Lemma 4.2.5 proves that $N_2 < \infty$ with probability 1, which justifies Condition 1. Once we found N_2, by Lemma 4.2.7, we have

$$\Gamma_R \le \frac{2^{-(2\alpha-\beta)}}{1 - 2^{-(2\alpha-\beta)}} \max_{n\le N_2} \max_{1\le i,j\le d'} \max_{0\le k<k'\le 2^{n-1}} \left\{ \frac{|L^n_{i,j}(k') - L^n_{i,j}(k)|}{(k'-k)^\beta \Delta_n^{2\alpha}} \right\}.$$

Thus, Condition 2 is satisfied as well.

Once we established the bounds for $||Z||_\alpha$ and Γ_R, by Lemma 4.2.7, we have

$$||A||_{2\alpha} \leq \Gamma_R \frac{2}{1-2^{-2\alpha}} + ||Z||_\alpha^2 \frac{2^{1-\alpha}}{1-2^{-\alpha}}.$$

In Section 4.2.5, we will explain how to simulate the random numbers (N_1 and N_2) jointly with the wavelet construction using the "record breaker" strategy introduced above. Specifically, we first find all the record breakers in sequence and then simulate the rest of the process conditional on the information of the record breakers.

4.2.2 ϵ-Strong Simulation of Bounds on α-Hölder Norms of Brownian Path

In this section, we will explain how to use the wavelet synthesis to approximate a single Brownian motion, Z, in the α-Hölder norm, (4.4). Of course, since we have d' Brownian motion, ultimately the algorithm that we shall describe for such an approximation (see Algorithm I below) will be run d' independent times.

Let us define $V^n = \max_{0\leq k<2^n} |W_k^n|$. We have the following auxiliary lemma.

Lemma 4.2.2

$$\|Z\|_\alpha \leq 2^{2\alpha} \sum_{n=0}^{\infty} 2^{-n(\frac{1}{2}-\alpha)} V^n.$$

Proof. For any interval $[t, t+\delta] \subset [0,1]$, suppose $2^{-m+2} \leq \delta \leq 2^{-m+1}$, then there exists two level n dyadic points t_k^m and t_{k+1}^m such that $[t, t+\delta] \subset [t_k^m, t_{k+1}^m]$. Using the Lévy-Ciesielski construction, one can check that

$$|Z(t+\delta) - Z(t)| \leq \sum_{n=0}^{m} 2^{-m+\frac{n}{2}} V^n + \sum_{n=m+1}^{\infty} 2^{-\frac{n}{2}} V^n.$$

Since $\delta > 2^{-m+2}$, we have

$$\begin{aligned}\frac{|Z(t+\delta)-Z(t)|}{\delta^\alpha} &\leq 2^{2\alpha}\Big(\sum_{n=0}^{m} 2^{-(1-\alpha)m+\frac{n}{2}}V^n + \sum_{n=m+1}^{\infty} 2^{-\frac{n}{2}+\alpha m}V^n\Big)\\ &\leq 2^{2\alpha}\Big(\sum_{n=0}^{m} 2^{-(1-\alpha)n+\frac{n}{2}}V^n + \sum_{n=m+1}^{\infty} 2^{-\frac{n}{2}+\alpha n}V^n\Big)\\ &\leq 2^{2\alpha}\sum_{n=0}^{\infty} 2^{-n(\frac{1}{2}-\alpha)}V^n.\end{aligned}$$

As the interval $[t, t+\delta]$ is arbitrarily chosen, we obtain the result. □

Owing to Lemma 4.2.2 we can now find a bound on $||Z||_\alpha$. Let

$$N_1 = \max\{n \geq 1 : |W_k^n| > 4\sqrt{n+1} \text{ for some } 1 \leq k \leq 2^{n-1}\}.$$

Lemma 4.2.3

$$E(N_1) < \infty$$

Proof. We note that

$$E(N_1) \leq \sum_{n=1}^{\infty}\sum_{k=1}^{2^{n-1}} P(|W_k^n| > 4\sqrt{n+1}) \leq \sum_{n=1}^{\infty} 2^{n-1}\exp(-8n) < \infty.$$

□

The strategy is then to simulate N_1 jointly with the sequence $\{W_k^n\}$. It is important to note that N_1 is not a stopping time with respect to the filtration generated by $\{(W_k^n : 1 \leq k \leq 2^{n-1}) : n \geq 1\}$. Note that if N_1 is simulated jointly with $\{W_k^n\}$, then for $2^n + k \geq N_1 + 1$, $|W_k^n| \leq 4\sqrt{n+1}$ and thus we can compute

$$K_\alpha = \sum_{n=1}^{[\log_2 N_1]} 2^{-n(\frac{1}{2}-\alpha)}V^n + \sum_{n=\log_2 N_1+1}^{\infty} 2^{-n(\frac{1}{2}-\alpha)}\sqrt{n+1} < \infty. \tag{4.9}$$

We call a pair (n,k) such that $|W_k^n| > 4\sqrt{n+1}$ a broken-record-pair. All pairs (both broken-record-pairs and non broken-record-pairs) can be totally ordered lexicographically. The distribution of subsequent pairs at which records are broken is not difficult to compute (because of the independence of the W_k^n's). So, using a sequential acceptance / rejection procedure we can simulate all of the broken-record-pairs.

Conditional on these pairs the distribution of the $\{(W_k^n : 1 \leq k \leq 2^n) : n \geq 1\}$ is straightforward to describe. Precisely, if (k, n) is a broken-record-pair, then W_k^n is conditioned on $|W_k^n| > 4\sqrt{n+1}$ and thus is straightforward to simulate. Similarly, if (k, n) is not a broken-record-pair, then W_k^n is conditioned on $|W_k^n| \leq 4\sqrt{n+1}$ and also can be easily simulated.

The simulation of the broken-record-pairs has been studied in [42], see Algorithm 2W. We synthesize their algorithm for our purposes next.

Algorithm 4.1: Simulate N_1 jointly with the broken-record-pairs

Input: A positive parameter $\rho > 4$.

Output: A vector S which gives all the indices $l = 2^n + k$ such that (n, k) is a broken-record-pair.

S0. Initialize $R = 0$ and S to be an empty array.

S1. Set $U = 1$, $D = 0$. Simulate $V \sim \text{Uniform}(0, 1)$.

S2. While $U > V > D$, set $R \leftarrow R + 1$ and $U \leftarrow P(|W_k^n| \leq \varrho\sqrt{\log R}) \times U$ and $D \leftarrow (1 - R^{1-\varrho^2/2}) \times U$.

S3. If $V \geq U$, add R to the end of S, i.e. $S = [S, R]$, and return to Step 1.

S4. If $V \leq D$, $N = \max(S)$.

S5. Output S.

Remark 4.2.1 *Observe that for every $l = 2^n + k \in S$, we can generate W_k^n conditional on the event $\{|W_k^n| > \varrho\sqrt{\log l}\}$; for other $1 \leq l \leq N$ (i.e. $l \notin S$) generate W_k^n given $\{|W_k^n| \leq \varrho\sqrt{\log l}\}$. Note that at the end of Algorithm 1 and after simulating W_k^n for $l = 2^n + k \leq N$ one can compute quantities such as K_α according to (4.9).*

4.2.3 Analysis and Bounds of α-Hölder Norms of Lévy Areas

We shall start by stating the following representation of the Lévy area $A_{i,j}(t_k^n, t_{k+1}^n)$, which we believe is of independent interest.

Lemma 4.2.4

$$\begin{aligned} & A_{i,j}(t_k^n, t_{k+1}^n) \\ = & \sum_{h=n+1}^{\infty} \sum_{l=1}^{2^{h-n-1}} [Z_i(t^h_{2^{h-n}k+2l-1}) - Z_i(t^h_{2^{h-n}k+2l-2})][Z_j(t^h_{2^{h-n}k+2l}) - Z_j(t^h_{2^{h-n}k+2l-1})]. \end{aligned}$$

The inner summation inside the expression of $A_{i,j}(t_k^n, t_{k+1}^n)$ motivates the definition of the following family of processes $(L_{i,j}^n(k) : k = 0, 1, ..., 2^{n-1})$, for $n \geq 1$:

$$\begin{aligned} & L_{i,j}^n(0) := 0 \\ & L_{i,j}^n(k) := L_{i,j}^n(k-1) + (Z_i(t_{2k-1}^n) - Z_i(t_{2k-2}^n))(Z_j(t_{2k}^n) - Z_j(t_{2k-1}^n)), \end{aligned}$$

for $k = 1, 2, ..., 2^{n-1}$.

Using this definition and Lemma 4.2.4 we can succinctly write $A_{i,j}(t_k^n, t_{k+1}^n)$ as

$$A_{i,j}(t_k^n, t_{k+1}^n) = \sum_{h=n+1}^{\infty} (L_{i,j}^h(2^{h-n}(k+1)) - L_{i,j}^h(2^{h-n}k)). \tag{4.10}$$

Moreover, the following result allows to control the behavior of the terms in the previous infinite series.

Lemma 4.2.5 *There exists $N_2 < \infty$ such that for all $n \geq N_2$ and all $l < m < 2^{n-1}$ we have*

$$|L_{i,j}^n(m) - L_{i,j}^n(l)| \leq (m-l)^{\beta} \Delta_n^{2\alpha}.$$

Now, recall that

$$R_{i,j}^n(t_l^n, t_m^n) := \sum_{k=l+1}^{m} A_{i,j}(t_{k-1}^n, t_k^n).$$

A direct application of Lemmas 4.2.4 and 4.2.5 yields the next corollary.

Corollary 4.2.6

$$R^n_{i,j}(t^n_l, t^n_m) = \sum_{h=n+1}^{\infty} (L^h_{i,j}(2^{h-n}m) - L^h_{i,j}(2^{h-n}l)).$$

We conclude this section with a proposition which summarizes the bounds that we will simulate.

Lemma 4.2.7 *Suppose that N_2 is chosen according to Lemma 4.2.5. We define*

$$\Gamma_L := \max_{1\le i,j\le d'} \max_{n<N_2} \max_{0\le l<m\le 2^{n-1}} \left\{ \frac{|L^n_{i,j}(m) - L^n_{i,j}(l)|}{(m-l)^\beta \Delta_n^{2\alpha}} \right\}.$$

Then

$$\Gamma_R \le \frac{2^{-(2\alpha-\beta)}}{1-2^{-(2\alpha-\beta)}} \Gamma_L$$

and

$$||A||_{2\alpha} \le \Gamma_R \frac{2}{1-2^{-2\alpha}} + ||Z||^2_\alpha \frac{2^{1-\alpha}}{1-2^{-\alpha}}.$$

4.2.4 Elements of ϵ-Strong Simulation for Bounds on α-Hölder Norms of Lévy Areas

There is some resemblance between the problem of sampling N_1 in Section 4.2.2, which involves a sequence of i.i.d. random variables (W^n_k's), and sampling of N_2 introduced in Section 4.2.3. However, simulation of N_2, which is basically our main goal here, is a lot more complicated because there is fair amount of dependence on the structure of the $L^n_{i,j}(k)$'s as one varies n. Let us provide a general idea of our simulation procedure in order to set the stage for the definitions and estimates that must be studied first.

Suppose we have simulated $\{(W^m_{i,k} : 0 \le k < 2^m) : m \le N\}$ for some N (to be discussed momentarily) and define

$$\tau_1(N) := \inf\{n \ge N+1 : |L^n_{i,j}(m) - L^n_{i,j}(l)| > (m-l)^\beta \Delta_n^{2\alpha} \text{ for some } 0 \le l < m < 2^{n-1}\}.$$

Because of Lemma 4.2.5 we have that the event $\{\tau_1(N) = \infty\}$ has positive probability. We will explain how to simulate a Bernoulli random variable with success parameter

$P(\tau_1(N) = \infty|\mathcal{F}_N)$. If such Bernoulli represents a success, then we have that $N_2 = N$ and we would have basically concluded the difficult part of the simulation procedure (the rest would be simulating under a series of conditioning events whose probability increases to one n grows). If the Bernoulli in question represents a failure (i.e its value is zero), then we will try again until obtaining a successful Bernoulli trial.

Now, part of the problem is that Algorithm 4.1 has been already executed, so $N \geq N_1$, in other words, while the random variables $\{W_{i,k}^n : 0 \leq k < 2^n\}$ are independent (for fixed $n > N$), they are no longer identically distributed. Instead, $W_{i,k}^n$ is standard Gaussian conditioned on the event $\{|W_{i,k}^n| < 4\sqrt{n+1}\}$.

Nevertheless, if n is large enough, all of the events $\{|W_{i,k}^n| < 4\sqrt{n+1}\}$ will occur with high probability. So, we shall first proceed to explain how to simulate a Bernoulli random variable with probability of success $P(\tau_1(n') = \infty|\mathcal{F}_{n'})$ assuming n' is a deterministic number. The procedure actually will produce both the outcome of the Bernoulli trial and if such outcome is a failure (i.e. $\tau_1(n') < \infty$), also

$$\{W_{i,k}^m : 1 \leq k \leq 2^m, n' < m \leq \tau_1(n')\}.$$

Our procedure is based on acceptance / rejection using a carefully chosen proposal distribution for the $W_{i,k}^m$'s based on exponential tilting of the $L_{i,j}^{n'}(k)$'s, conditional on $\mathcal{F}_{n'}$. To this end, we will need to compute the associated (conditional on $\mathcal{F}_{n'}$) moment generating function of $L_{i,j}^n(k)$ and the family of distributions induced over the $W_{i,k}^n$'s and $W_{j,k}^n$'s when exponentially tilting $L_{i,j}^{n'}(k)$, this will be done in Section 4.2.4.1. Then, we need some large deviations estimates in order to enforce the feasibility of a certain randomization procedure, these estimates are given in Section 4.2.4.2. These are all the elements needed for the simulation procedure of α-Hölder Norms of Lévy Areas given in Section 4.2.5.

4.2.4.1 Basic Notation, Conditional Moment Generating Functions, and Associated Exponential Tilting

First, we recall the wavelet synthesis discussed in Section 4.2.1, which was explained for a single Brownian motion. Since we will work with d' Brownian motions here we need to adapt the notation. For each $i \in \{1, ..., d'\}$ let $\{(W_{i,k}^n : 1 \leq k \leq 2^n) : n \geq 1\}$ be the sequence of i.i.d. N(0,1) random variables arising in the wavelet synthesis (4.8) for $Z_i(\cdot)$.

Now, define

$$\mathcal{F}_n = \sigma\{\{(W_{i,k}^m : 0 \leq k < 2^m) : m \leq n\}\}.$$

and for the conditional expectation given $\mathcal{F}_n$ we write

$$E_n(\ \cdot\ ) := E(\ \cdot\ |\ \mathcal{F}_n).$$

In order to reduce the length of some of the equations that follow, we write, for each $r \in \{1, 2, ..., 2^n\}$,

$$\Lambda_i^n(t_r^n) := (Z_i(t_r^n) - Z_i(t_{r-1}^n)). \tag{4.11}$$

Then, using the following very useful pair of equations (for $k = 1, 2, ..., 2^{n-1}$)

$$\begin{aligned} \Lambda_i^n(t_{2k-1}^n) &= \frac{1}{2}\Lambda_i^{n-1}(t_k^{n-1}) + \Delta_n^{1/2} W_{i,k}^n. \\ \Lambda_i^n(t_{2k}^n) &= \frac{1}{2}\Lambda_i^{n-1}(t_k^{n-1}) - \Delta_n^{1/2} W_{i,k}^n, \end{aligned} \tag{4.12}$$

we can see that

$$\mathcal{F}_n = \sigma\{\cup_{m \leq n}(Z(t) - Z(s)) : 0 \leq s < t \leq 1, t, s \in D_m\}.$$

and we have that (for $0 \leq k \leq 2^{n-1}$)

$$L_{i,j}^n(k) = \sum_{r=1}^{k} \Lambda_i^n\left(t_{2r-1}^n\right) \Lambda_j^n\left(t_{2r}^n\right).$$

Assume that $k < k'$, we will iteratively compute

$$\begin{aligned} &E_n\{\exp\left(\theta_0\{L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\}\right)\} \\ &= E_n[E_{n+1}[...E_{n+m-1}[\exp\left(\theta_0\{L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\}\right)]...]]. \end{aligned} \tag{4.13}$$

We first start from inner expectation.

Corollary 4.2.8

$$\begin{aligned}
&E_{n+m-1}\exp(\theta_0\Lambda_i^{n+m}\left(t_{2r-1}^{n+m}\right)\Lambda_j\left(t_{2r}^{n+m}\right))\\
=&\left(1-\theta_0^2\Delta_{n+m}^2\right)^{-1/2}\exp\left(\theta_1\Lambda_j^{n+m-1}(t_r^{n+m-1})\Lambda_i^{n+m-1}(t_r^{n+m-1})\right)\\
&\exp\left(\eta_1\Lambda_j^{n+m-1}(t_r^{n+m-1})^2+\eta_1\Lambda_i^{n+m-1}(t_r^{n+m-1})^2\right),
\end{aligned}$$

where

$$\theta_1:=\theta_0(1-\theta_0^2\Delta_{n+m}^2)^{-1}/4,\quad \eta_1:=\theta_0^2\left(1-\theta_0^2\Delta_{n+m}^2\right)^{-1}\Delta_{n+m}/8.$$

Moreover, define

$$\begin{aligned}
&P'_{n+m,t_r^{n+m}}\left(W_{i,r}^{n+m}\in A,W_{j,r}^{n+m}\in B\right)\\
=&\frac{E_{n+m-1}\left(I(W_{i,r}^{n+m}\in A,W_{j,r}^{n+m}\in B)\exp(\theta_0\Lambda_i^{n+m}\left(t_{2r-1}^{n+m}\right)\Lambda_j^{n+m}\left(t_{2r}^{n+m}\right))\right)}{E_{n+m-1}\exp(\theta_0\Lambda_i^{n+m}\left(t_{2r-1}^{n+m}\right)\Lambda_j^{n+m}\left(t_{2r}^{n+m}\right))},
\end{aligned}$$

then under $P'_{n+m,t_r^{n+m}}$, *and given* $\mathcal{F}_{n+m-1}$, *we have that* $(W_{i,r}^{n+m},W_{j,r}^{n+m})$ *follows a Gaussian distribution with covariance matrix*

$$\Sigma_{n+m}^{i,j}\left(t_r^{n+m}\right)=\frac{1}{1-\theta_0^2\Delta_{n+m}^2}\begin{pmatrix}1 & -\theta_0\Delta_{n+m}\\ -\theta_0\Delta_{n+m} & 1\end{pmatrix},$$

and mean vector

$$\mu_{n+m}^{i,j}\left(t_r^{n+m}\right)=\Sigma_{n+m}^{i,j}\left(t_r^{n+m}\right)\begin{pmatrix}\theta_0\Delta_{n+m}^{1/2}\Lambda_j^{n+m-1}(t_r^{n+m-1})/2\\ -\theta_0\Delta_{n+m}^{1/2}\Lambda_j^{n+m-1}(t_r^{n+m-1})/2\end{pmatrix}.$$

So, from Corollary 4.2.8 we conclude that

$$\begin{aligned}
&E_{n+m-1}[\exp(\theta_0\sum_{r=k+1}^{k'}\Lambda_i^{n+m}\left(t_{2r-1}^{n+m}\right)\Lambda_j\left(t_{2r}^{n+m}\right))]\\
&=\left(1-\theta_0^2\Delta_{n+m}^2\right)^{-(k'-k)/2}\exp(\theta_1\sum_{r=k+1}^{k'}\Lambda_j^{n+m-1}(t_r^{n+m-1})\Lambda_i^{n+m-1}(t_r^{n+m-1}))\\
&\exp(\eta_1\sum_{r=k+1}^{k'}\Lambda_j^{n+m-1}(t_r^{n+m-1})^2+\eta_1\sum_{r=k+1}^{k'}\Lambda_i^{n+m-1}(t_r^{n+m-1})^2).
\end{aligned}\tag{4.14}$$

If $m \geq 2$, we can continue taking the corresponding conditional expectation given $\mathcal{F}_{n+m-2}$. Due to the recursive nature of (4.13) and the linear quadratic terms that arise in (4.14) it is convenient to consider

$$\begin{aligned} &\sum_{r=1}^{2^{n+m-1}} \theta_1\left(t_r^{n+m-1}\right) \Lambda_j^{n+m-1}(t_r^{n+m-1})\Lambda_i^{n+m-1}(t_r^{n+m-1}) \\ &+ \sum_{r=1}^{2^{n+m-1}} \eta_1\left(t_r^{n+m-1}\right) (\Lambda_j^{n+m-1}(t_r^{n+m-1})^2 + \Lambda_j^{n+m-1}(t_r^{n+m-1})^2), \end{aligned} \tag{4.15}$$

where

$$\theta_1\left(t_r^{n+m-1}\right) = \theta_1 \times I\left(r \in \{k+1, ..., k'\}\right), \ \eta_1\left(t_r^{n+m-1}\right) = \eta_1 \times I\left(r \in \{k+1, ..., k'\}\right).$$

Then, recursively define for $l = 2, ..., m$

$$\begin{aligned} \theta_+^l\left(t_r^{m+n-l}\right) &= \theta_{l-1}\left(t_{2r-1}^{m+n-l+1}\right) + \theta_{l-1}\left(t_{2r}^{m+n-l+1}\right) \\ \theta_-^l\left(t_r^{m+n-l}\right) &= \theta_{l-1}\left(t_{2r-1}^{m+n-l+1}\right) - \theta_{l-1}\left(t_{2r}^{m+n-l+1}\right) \\ \eta_+^l\left(t_r^{m+n-l}\right) &= \eta_{l-1}\left(t_{2r-1}^{m+n-l+1}\right) + \eta_{l-1}\left(t_{2r}^{m+n-l+1}\right) \\ \eta_-^l\left(t_r^{m+n-l}\right) &= \eta_{l-1}\left(t_{2r-1}^{m+n-l+1}\right) - \eta_{l-1}\left(t_{2r}^{m+n-l+1}\right) \\ \rho_l\left(t_r^{m+n-l}\right) &= \frac{\Delta_{n+m-l+1}\theta_+^l\left(t_r^{m+n-l}\right)}{1 - 2\Delta_{n+m-l+1}\eta_+^l\left(t_r^{m+n-l}\right)}, \\ h_l\left(t_r^{m+n-l}\right) &= \frac{\Delta_{n+m-l+1}}{\left(1 - 2\Delta_{n+m-l+1}\eta_+^l\left(t_r^{m+n-l}\right)\right)\left(1 - \rho_l\left(t_r^{m+n-l}\right)^2\right)}, \end{aligned} \tag{4.16}$$

and set

$$\begin{aligned}
\eta_l\left(t_r^{m+n-l}\right) =& \frac{\eta_+^l\left(t_r^{m+n-l}\right)}{4} \\
&+ \frac{h_l\left(t_r^{m+n-l}\right)}{8}\left(\theta_-^l\left(t_r^{m+n-l}\right)^2 + 4\eta_-^l\left(t_r^{m+n-l}\right)^2\right. \\
&\qquad \left. +4\theta_-^l\left(t_r^{m+n-l}\right)\eta_-^l\left(t_r^{m+n-l}\right)\rho_l\left(t_r^{m+n-l}\right)\right), \\
\theta_l\left(t_r^{m+n-l}\right) =& \frac{\theta_+^l\left(t_r^{m+n-l}\right)}{4} \\
&+ h_l\left(t_r^{m+n-l}\right)\left(\theta_-^l\left(t_r^{m+n-l}\right)\eta_-^l\left(t_r^{m+n-l}\right) + \frac{1}{4}\theta_-^l\left(t_r^{m+n-l}\right)^2 g_l\left(t_r^{m+n-l}\right)\right. \\
&\qquad \left. +\eta_-^l\left(t_r^{m+n-l}\right)^2\rho_l\left(t_r^{m+n-l}\right)\right).
\end{aligned}$$

Finally, define

$$\begin{aligned}
A\left(t_r^{n+m-l}\right) =& \theta_{l-1}\left(t_{2r-1}^{n+m-l+1}\right)\Lambda_j^{n+m-l+1}(t_{2r-1}^{n+m-l+1})\Lambda_i^{n+m-l+1}(t_{2r-1}^{n+m-l+1}) \\
&+ \theta_{l-1}\left(t_{2r}^{n+m-l+1}\right)\Lambda_j^{n+m-l+1}(t_{2r}^{n+m-l+1})\Lambda_i^{n+m-l+1}(t_{2r}^{n+m-l+1}), \\
B\left(t_r^{n+m-l}\right) =& \eta_{l-1}\left(t_{2r-1}^{n+m-l+1}\right)(\Lambda_j^{n+m-l+1}(t_{2r-1}^{n+m-l+1})^2 + \Lambda_j^{n+m-l+1}(t_{2r-1}^{n+m-l+1})^2) \\
&+ \eta_{l-1}\left(t_{2r}^{n+m-l+1}\right)(\Lambda_j^{n+m-l+1}(t_{2r}^{n+m-l+1})^2 + \Lambda_j^{n+m-l+1}(t_{2r}^{n+m-l+1})^2),
\end{aligned}$$

and

$$C\left(t_r^{n+m-l}\right) = \left(1 - 2\Delta_{n+m-l}\eta_+^l\left(t_r^{m+n-l}\right)\right)^{-1}\left(1 - \rho_l\left(t_r^{m+n-l}\right)^2\right)^{-1/2}.$$

So, in particular we can write (4.15) as

$$\sum_{r=1}^{2^{n+m-2}}\left(A\left(t_r^{n+m-2}\right) + B\left(t_r^{n+m-2}\right)\right),$$

and the following result is key in evaluating (4.13).

Corollary 4.2.9 *For $l = 2, 3, ..., m$ and $r = 1, 2, ..., 2^{n+m-l}$*

$$\begin{aligned}
&E_{n+m-l}\exp(A\left(t_r^{n+m-l}\right) + B\left(t_r^{n+m-l}\right)) \\
=& C\left(t_r^{n+m-l}\right)\exp\left(\theta_l\left(t_r^{m+n-l}\right)\Lambda_i\left(t_r^{m+n-l}\right)\Lambda_j\left(t_r^{m+n-l}\right)\right) \\
&\exp\left(\eta_l\left(t_r^{m+n-l}\right)\left(\Lambda_i\left(t_r^{m+n-l}\right)^2 + \Lambda_j\left(t_r^{m+n-l}\right)^2\right)\right).
\end{aligned}$$

Moreover, define

$$\begin{aligned}&P'_{n+m-l+1,t_r^{n+m-l+1}}\left(W_{i,r}^{n+m-l+1}\in A, W_{j,r}^{n+m-l+1}\in B\right)\\=&\frac{E_{n+m-l}\left(I(W_{i,r}^{n+m-l+1}\in A, W_{j,r}^{n+m-l+1}\in B)\exp(A\left(t_r^{n+m-l}\right)+B\left(t_r^{n+m-l}\right))\right)}{E_{n+m-l}\exp(A\left(t_r^{n+m-l}\right)+B\left(t_r^{n+m-l}\right))},\end{aligned}$$

then under $P'_{n+m-l+1,t_r^{n+m-l+1}}$*, and given* $\mathcal{F}_{n+m-l}$*, we have that* $(W_{i,r}^{n+m-l+1}, W_{j,r}^{n+m-l+1})$ *follows a Gaussian distribution with covariance matrix*

$$\begin{aligned}&\Sigma_{n+m-l+1}^{i,j}\left(t_r^{n+m-l+1}\right)\\=&\frac{1}{1-\rho_l\left(t_r^{m+n-l}\right)^2}\\&\times\begin{pmatrix}\left(1-2\Delta_{n+m-l}\eta_+^l\left(t_r^{m+n-l}\right)\right)^{-1} & g_l\left(t_r^{m+n-l}\right)\\ g_l\left(t_r^{m+n-l}\right) & \left(1-2\Delta_{n+m-l}\eta_+^l\left(t_r^{m+n-l}\right)\right)^{-1}\end{pmatrix}\end{aligned}$$

and mean vector

$$\begin{aligned}&\mu_{n+m}^{i,j}\left(t_r^{n+m-l+1}\right)\\=&\Delta_{n+m-l}^{1/2}\Sigma_{n+m-l+1}^{i,j}\left(t_r^{n+m-l+1}\right)\\&\times\begin{pmatrix}\Lambda_i\left(t_r^{n+m-l}\right)\eta_-^l\left(t_r^{n+m-l}\right)+\frac{1}{2}\Lambda_j\left(t_r^{n+m-l}\right)\theta_-^l\left(t_r^{n+m-l}\right)\\ \Lambda_j\left(t_r^{n+m-l}\right)\eta_-^l\left(t_r^{n+m-l}\right)+\frac{1}{2}\Lambda_i\left(t_r^{n+m-l}\right)\theta_-^l\left(t_r^{n+m-l}\right)\end{pmatrix}.\end{aligned}$$

Using Corollary 4.2.9 we conclude that

$$\begin{aligned}&E_{n+m-l}\exp(\sum_{r=1}^{2^{n+m-l}}\left(A\left(t_r^{n+m-l}\right)+B\left(t_r^{n+m-l}\right)\right))\\=&\prod_{r=1}^{2^{n+m-l}}C\left(t_r^{n+m-l}\right)\times\exp\left(\sum_{r=1}^{2^{n+m-l-1}}\left(A\left(t_r^{n+m-l-1}\right)+B\left(t_r^{n+m-l-1}\right)\right)\right).\end{aligned}$$

Therefore, combining Corollary 4.2.8 and repeatedly iterating the previous expression

we conclude that

$$\begin{aligned}
&E_n \exp(\theta_0\{L_{i,j}^{n+m}(k) - L_{i,j}^{n+m}(k')\}) \\
&= \left(1 - \theta_0^2 \Delta_{n+m}^2\right)^{-(k'-k)/2} \prod_{l=2}^{m} \prod_{r=1}^{2^{n+m-l}} C\left(t_r^{n+m-l}\right) \\
&\quad \times \exp(\sum_{r=1}^{2^n} \theta_m\left(t_r^n\right) \Lambda_i\left(t_r^n\right) \Lambda_j\left(t_r^n\right) + \sum_{r=1}^{2^n} \eta_m\left(t_r^n\right) \{\Lambda_i\left(t_r^n\right)^2 + \Lambda_j\left(t_r^n\right)^2\}). \qquad (4.17)
\end{aligned}$$

4.2.4.2 Conditional Large Deviations Estimates for $L_{i,j}^n(k)$

We wish to estimate, for $k' > k$ and $k', k \in \{0, 1, ..., 2^{n+m-1}\}$,

$$\begin{aligned}
&P_n\left(|L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)| \ > (k'-k)^\beta \Delta_{n+m}^{2\alpha}\right) \\
&\leq \exp(-\theta_0 (k'-k)^\beta \Delta_{n+m}^{2\alpha}) \times \{E_n[\exp(\theta_0\{L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\})] \\
&+ E_n[\exp(-\theta_0\{L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\})]\}.
\end{aligned}$$

We borrow some intuition from the proof of Lemma 4.2.5 and select

$$\theta_0(m, k', k) := \theta_0 = \frac{\gamma}{(k'-k)^{1/2} \Delta_n^{2\alpha'} \Delta_m}. \qquad (4.18)$$

We will drop the dependence on (m, k', k) for brevity. In addition, we pick $\gamma \leq 1/4$ and $\alpha' \in (\alpha, 1/2)$ so that

$$\exp(-\theta_0 (k'-k)^\beta \Delta_{n+m}^{2\alpha}) = \exp(-\gamma (k'-k)^{\beta-1/2} \Delta_n^{2(\alpha-\alpha')} \Delta_m^{2\alpha-1})$$

Our next task is to control the $E_n \exp(\theta_0\{L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\})$, which is the purpose of the following result, proved in the appendix to this section.

Lemma 4.2.10 *Suppose that θ_0 is chosen according to (4.18), and n is such that for $\varepsilon_0 \in (0, 1/2)$*

$$\max_{r \leq 2^n}\{|\Lambda_i\left(t_r^n\right)|, |\Lambda_j\left(t_r^n\right)|\} \leq \Delta_n^{\alpha'} \qquad (4.19)$$

and

$$\left|\sum_{r=l+1}^{m} \Lambda_i(t_r^n)\Lambda_j(t_r^n)\right| \leq \varepsilon_0 (m-l)^\beta \Delta_n^{2\alpha'} \text{ for all } 0 \leq l < m \leq 2^n \qquad (4.20)$$

with $\alpha' \in (\alpha, 1/2)$*, then*

$$E_n[\exp(\theta_0\{L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\})] \leq 4\exp\left(\varepsilon_0\gamma(k'-k)^{\beta-1/2}\right).$$

Remark 4.2.2 *It is very important to note that due to Lemma 4.2.3 we can always continue simulating the* $W_{i,k}^m$*'s (maybe conditional on* $\{|W_{i,k}^m| < 4\sqrt{m+1}\}$ *in case* $m > N_1$*) to make sure that (4.19) holds for some* n*. Similarly, condition (4.20) can be simultaneously enforced with (4.19) because of Lemma 4.2.5. Actually, Lemmas 4.2.3 and Lemma 4.2.5 indicate that conditions (4.19) and (4.20) will occur eventually for all* n *larger than some random threshold enough. Our simulation algorithms will ultimately detect such threshold, but Lemma 4.2.10 does not require that we know that threshold.*

As a consequence of Lemma 4.2.10, using Chernoff's bound, we obtain the following proposition.

Proposition 4.2.11 *If* n *is such that (4.19) and (4.20) hold, then*

$$\begin{aligned} & P_n\left(|L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)| > (k'-k)^{\beta}\,\Delta_{n+m}^{2\alpha}\right) \\ \leq\ & 8\exp\left(-\frac{1}{2}\gamma\,(k'-k)^{\beta-1/2}\,\Delta_n^{2(\alpha-\alpha')}\Delta_m^{2\alpha-1}\right). \end{aligned}$$

4.2.5 Joint Tolerance-Enforced Simulation for α-Hölder Norms and Proof of Theorem 4.1.2.

Define

$$\mathcal{C}_n(m) = \{|L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)| > (k'-k)^{\beta}\Delta_{n+m}^{2\alpha} \text{ for some } 0 \leq k < k' < 2^{n+m-1}\},$$

and put $\tau_1(n) = \inf\{m \geq 1 : \mathcal{C}_n(m) \text{ occurs}\}$. We write $\bar{\mathcal{C}}_n(m)$ for the complement of $\mathcal{C}_n(m)$, so that

$$P_n(\tau_1(n) < \infty) = \sum_{m=1}^{\infty} P\left(\mathcal{C}_n(m) \cap \cap_{l=1}^{m-1}\bar{\mathcal{C}}_n(m)\right).$$

To facilitate the explanation, we next introduce a few more notations. Let

$$\omega_{n:n+m} := \{W_{i,k}^l : 1 \leq k \leq 2^n, 1 \leq i \leq d', n < l \leq n+m\}.$$

In addition, define

$$\begin{aligned} v_n(k,k'|m) :=& 6\exp\left(-\frac{1}{2}\gamma\left(k'-k\right)^{\beta-1/2}\Delta_n^{2(\alpha-\alpha')}\Delta_m^{2\alpha-1}\right) \\ & \times I\left(0 \leq k < k' \leq 2^{n+m-1}\right) I\left(m \geq 1\right) \\ b_n(m) :=& \sum_{0\leq k<k'\leq 2^{m+n-1}} v_n(k,k'|m) \\ q_n(k,k'|m) :=& \frac{v_n(k,k'|m)}{b_n(m)} \end{aligned}$$

and

$$P_{n,m}^{i,j,k,k'}\left(\omega_{n:n+m} \in \cdot\right) = \frac{E_n I\left(\omega_{n:n+m} \in \cdot\right)\exp\left(\theta_0\{L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\}\right)}{E_n \exp\left(\theta_0\{L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\}\right)}.$$

We also denote

$$\psi_n(m,i,j,k,k') := \log E_n \exp\left(\theta_0\{L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k')\}\right)$$

Observe that

$$\begin{aligned} b_n\left(m\right) &= \sum_{0\leq k<k'\leq 2^{n+m-1}} 6\exp\left(-\frac{1}{2}\gamma\left(k'-k\right)^{\beta-1/2}\Delta_n^{2(\alpha-\alpha')}\Delta_m^{2\alpha-1}\right) \\ &\leq 2^{2(m+n)}\exp\left(-\frac{1}{2}\gamma\Delta_n^{2(\alpha-\alpha')}\Delta_m^{2\alpha-1}\right). \end{aligned}$$

Thus, $b_n(m) \to 0$ as $n \to \infty$. Then we can select any probability mass function $\{g(m) : m \geq 1\}$, e.g. $g(m) = e^{-1}/(m-1)!$ for $m \geq 1$, by assuming that n is sufficiently large, such that

$$g(m) \geq d'^2 b_n(m)$$

Now consider the following procedure, which we called Procedure Aux, for "auxiliar", which is given for pedagogical purposes because as we shall see shortly is not directly applicable but just useful to understand the nature of the method that we

shall ultimately use.

Procedure Aux

Input: We assume that we have simulated $\{(W_{i,k}^n : 0 \leq k < 2^l) : l \leq n\}\}$.

Output: A Bernoulli F with parameter $P_n(\tau_1(n) < \infty)$, and if $F = 1$, also

$$\omega_{n:n+M} = \{W_{i,k}^l : 1 \leq k \leq 2^n, 1 \leq i \leq d', n < l \leq n + M\}$$

conditional on the event $\tau_1(n) < \infty$.

S1. Sample M according to $g(m)$.

S2. Given $M = m$ sample I and J i.i.d. from the uniform distribution over the set $\{1, 2, ..., d'\}$.Then, sample K', K from $q_n(k, k'|m)$.

S3. Given $M = m$, $I = i$,$J = j$,$K = k$, and $K' = k'$, simulate $\omega_{n:n+m}$ from $P_{n,m}^{i,j,k,k'}(\cdot)$. Note that simulation from $P_{n,m}^{i,j,k,k'}(\cdot)$ can be done according to Corollary 4.2.9.

S4. Compute

$$\begin{aligned} & \Xi_n(m, i, j, k, k', \omega_{n:n+m}) \\ = & \frac{1}{g(m)d'^{-2}q_n(k, k'|m)\exp\left(\theta_0\{L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\} - \psi_n(m, i, j, k, k')\right)}, \end{aligned}$$

and

$$\mathcal{N}_n(m) = \sum_{1 \leq i,j \leq d'} \sum_{1 \leq h < h' \leq 2^{n+m-1}} I\left(\left|L_{i,j}^{n+m}(h') - L_{i,j}^{n+m}(h)\right| > (h - h')^\beta \Delta_{n+m}^{2\alpha}\right).$$

S5. Simulate U uniformly distributed on $[0, 1]$ independent of everything else and output

$$\begin{aligned} F = & I\left(U < I\left(\left\{\left|L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\right| > (k - k')^\beta \Delta_{n+m}^{2\alpha}\right\} \cap \cap_{l=1}^{m-1} \bar{\mathcal{C}}_n(l)\right)\right. \\ & \left. \times \Xi_n(m, i, j, k, k', \omega_{n:n+m}) / \mathcal{N}_n(m)\right). \end{aligned}$$

If $F = 1$, also output $\omega_{n:n+m}$.

We claim that the output F is distributed as a Bernoulli random variable with parameter $P_n\left(\tau_1(n)<\infty\right)$. Moreover, we claim that if $F=1$, then, $\omega_{n:n+M}$ is distributed according to $P_n\left(\omega_{n:n+\tau_1(n)}\in\cdot \mid \tau_1(n)<\infty\right)$. We first verify the claim that the outcome in Step 5 follows a Bernoulli with parameter $P_n\left(\tau_1(n)<\infty\right)$. In order to see this, let Q_n denote the distribution induced by Procedure Aux. Note that

$$\begin{aligned}
&Q_n(U<I\left(\left\{\left|L_{i,j}^{n+M}(K')-L_{i,j}^{n+M}(K)\right|>(K-K')^{\beta}\Delta_{n+M}^{2\alpha}\right\}\cap\cap_{l=1}^{M-1}\bar{\mathcal{C}}_n(l)\right)\\
&\quad\times\Xi_n(M,I,J,K,K',\omega_{n:n+M})/\mathcal{N}_n\left(m\right))\\
=&E^{Q_n}[I\left(\left\{\left|L_{i,j}^{n+M}(K')-L_{i,j}^{n+M}(K)\right|>(K-K')^{\beta}\Delta_{n+M}^{2\alpha}\right\}\cap\cap_{l=1}^{M-1}\bar{\mathcal{C}}_n(l)\right)\\
&\quad\times\Xi_n(M,I,J,K,K',\omega_{n:n+M})/\mathcal{N}_n\left(m\right)]\\
=&\sum_{m=1}^{\infty}\sum_{1\le i,j\le d'}\sum_{1\le k<k'\le 2^{n+m-1}}E^{Q_n}\left[I\left(\left\{\left|L_{i,j}^{n+m}(k')-L_{i,j}^{n+m}(k)\right|>(k-k')^{\beta}\Delta_{n+m}^{2\alpha}\right\}\right.\right.\\
&\qquad\left.\left.\cap\cap_{l=1}^{m-1}\bar{\mathcal{C}}_n(l)\right)\times\frac{dP_n}{dP_{n,m}^{i,j,k,k'}}(\omega_{n:n+m})\times\frac{1}{\mathcal{N}_n\left(m\right)}\right]\\
=&\sum_{m=1}^{\infty}\sum_{1\le i,j\le d'}\sum_{1\le k<k'\le 2^{n+m-1}}\\
&\qquad E_n\left(\frac{I\left(\left\{\left|L_{i,j}^{n+m}(k')-L_{i,j}^{n+m}(k)\right|>(k-k')^{\beta}\Delta_{n+m}^{2\alpha}\right\}\cap\cap_{l=1}^{m-1}\bar{\mathcal{C}}_n(l)\right)}{\mathcal{N}_n\left(m\right)}\right)\\
=&\sum_{m=1}^{\infty}P_n\left(\mathcal{C}_n(m)\cap\cap_{l=1}^{m-1}\bar{\mathcal{C}}_n(l)\right)=P_n(\tau_1\left(n\right)<\infty).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&Q_n\left(\omega_{n:n+M}\in A\mid U<I\left(\mathcal{C}_n(M)\cap\cap_{l=1}^{M-1}\bar{\mathcal{C}}_n(l)\right)\Xi_n(M,I,J,K,K',\omega_{n:n+M})\right)\\
=&\sum_{m=1}^{\infty}E^{Q_n}\left(\omega_{n:n+m}\in A\ ,\ \frac{dP_{n,m}^{I,J,K,K'}}{dP_n}\left(\omega_{n:n+m}\right)\frac{I\left(\mathcal{C}_n(m)\cap\cap_{l=1}^{M-1}\bar{\mathcal{C}}_n(l)\right)}{P_n(\tau_1\left(n\right)<\infty)}\right)\\
=&\sum_{m=1}^{\infty}P_n\left(\omega_{n:n+m}\in A\ ,\ \tau_1\left(n\right)=m\right)/P_n(\tau_1\left(n\right)<\infty)\\
=&P_n\left(\omega_{n:n+\tau_1(n)}\in A\mid\tau_1(n)<\infty\right)
\end{aligned}$$

The deficiency of Procedure Aux is that it does not recognize that $n>N_1$. Let us now account for this fact and note that conditional on $\mathcal{F}_{N_1}$ we have that $W_{i,k}^n$'s are

i.i.d. $N(0,1)$ but conditional on $\{|W_{i,k}^n| < 4\sqrt{n+1}\}$ for all $n > N_1$. Define

$$\mathcal{H}_m^n = \{|W_{i,k}^r| < 4\sqrt{r+1} : 1 \le k \le 2^r, n < r \le n+m\}.$$

In order to simulate $P_{N_1}(\tau_1(N_1) < \infty)$ we modify step 3 of Procedure Aux. Specifically, we have

Procedure B

Input: We assume that we have simulated $\{(W_{i,k}^l : 0 \le k < 2^l) : l \le n\}$. So, the $W_{i,k}^m$'s are i.i.d. $N(0,1)$ but conditional on $\{|W_{i,k}^m| < 4\sqrt{m+1}\}$ for all $m > n$. We also assume that conditions (4.19) and (4.20) hold in Lemma 4.2.10; note the discussion following Lemma 4.2.10 which notes that this can be assumed at the expense of simulating additional $W_{i,k}^m$'s (with $\{|W_{i,k}^m| < 4\sqrt{m+1}\}$ if $m > N_1$).

Output: A Bernoulli F with parameter $P_n(\tau_1(n) < \infty, \mathcal{H}_\infty^n)$, and if $F = 1$, also

$$\omega_{n:n+\tau_1(n)} = \{W_{i,k}^l : 1 \le k \le 2^n, 1 \le i \le d', n < l \le n + \tau_1(n)\}$$

conditional on $\tau_1(n) < \infty$ and on $\mathcal{H}_\infty^n$.

S1. Sample M according to $g(m)$.

S2. Given $M = m$ sample I and J i.i.d. from the uniform distribution over the set $\{1, 2, ..., d'\}$.Then, sample K', K from $q_n(k, k'|m)$.

S3. Given $M = m$, $I = i$,$J = j$,$K = k$, and $K' = k'$, simulate $\omega_{n:n+m}$ from $P_{n,m}^{i,j,k,k'}(\cdot)$. Note that simulation from $P_{n,m}^{i,j,k,k'}(\cdot)$ can be done according to Corollary 4.2.9.

S4. Compute

$$\begin{aligned} &\Xi_n(m, i, j, k, k', \omega_{n:n+m}) \\ = \; &\frac{1}{g(m)d'^{-2}q_n(k, k'|m)\exp\left(\theta_0\{L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\} - \psi_n(m, i, j, k, k')\right)}, \end{aligned}$$

and

$$\mathcal{N}_n(m) = \sum_{1\le i,j\le d'} \sum_{1\le k<k'\le 2^{n+m-1}} I\left(\left|L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\right| > (k-k')^{\beta}\Delta_{n+m}^{2\alpha}\right).$$

S5. Simulate U uniformly distributed on $[0,1]$ independent of everything else and output

$$\begin{aligned} &F \\ =&I\left(U < \frac{I\left(\mathcal{H}_m^n \cap \left\{\left|L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\right| > (k-k')^{\beta}\Delta_{n+m}^{2\alpha}\right\} \cap \cap_{l=1}^{M-1}\bar{\mathcal{C}}_n(l)\right) P\left(\mathcal{H}_\infty^{n+m}\right)}{P\left(\mathcal{H}_\infty^n\right)}\right. \\ &\left.\qquad \times \Xi_n(m,i,j,k,k',\omega_{n:n+m})/\mathcal{N}_n(m)\right) \end{aligned}$$

(Notice that $P(\mathcal{H}_\infty^{n+m})/P(\mathcal{H}_\infty^n) = P(\mathcal{H}_{n+m}^n)$ and can be computed in finite steps.)

If $F=1$, also output $\omega_{n:n+m}$.

Let $\tilde{Q}_n$ denote the distribution induced by Procedure $\tilde{B}$. Following the same analysis as that given for Procedure B, we can verify that

$$\begin{aligned} \tilde{Q}_n(U <& \frac{I\left(\mathcal{H}_m^n \cap \left\{\left|L_{i,j}^{n+m}(k') - L_{i,j}^{n+m}(k)\right| > (k-k')^{\beta}\Delta_{n+m}^{2\alpha}\right\} \cap \cap_{l=1}^{M-1}\bar{\mathcal{C}}_n(l)\right) P\left(\mathcal{H}_\infty^{n+m}\right)}{P\left(\mathcal{H}_\infty^n\right)} \\ &\times \Xi_n(m,i,j,k,k',\omega_{n:n+m})/\mathcal{N}_n(m))) = P_n\left(\tau_1(n) < \infty | \mathcal{H}_\infty^n\right). \end{aligned}$$

And if the Bernoulli trial is a success, then, $\omega_{n:n+M}$ is distributed according to

$$P_n\left(\omega_{n:n+\tau_1(n)} \in \cdot \mid \tau_1(n) < \infty, \mathcal{H}_\infty^n\right).$$

Finally, if $\tau_1(n) = \infty$, we still need to simulate $\omega_{n:n+m}$ for any $m \ge 1$. But now, conditional on $\{\tau_1(n) = \infty, \mathcal{H}_\infty^n\}$. Note that

$$\begin{aligned} &P_n\left(\omega_{n:n+m} \in A \mid \tau_1(n) = \infty, \mathcal{H}_\infty^n\right) \\ =&\frac{P_n\left(\omega_{n:n+m} \in A\ ,\ \tau_1(n) = \infty, \mathcal{H}_\infty^n\right)}{P_n\left(\tau_1(n) = \infty, \mathcal{H}_\infty^n\right)} \\ =&\frac{E_n I(\omega_{n:n+m} \in A, \tau_1(n) > m, \mathcal{H}_m^n) P_{n+m}(\tau_1(n+m) = \infty, \mathcal{H}_\infty^{n+m})}{P_n\left(\tau_1(n) = \infty, \mathcal{H}_\infty^n\right)}. \end{aligned}$$

We do this by sampling $\omega_{n:n+m}$ from $P_n(\cdot)$ and accept the path with probability

$$I(\tau_1(n) > m, \mathcal{H}_m^n) P_{n+m}(\tau_1(n+m) = \infty, \mathcal{H}_\infty^{n+m}).$$

This clearly can be done since we can easily simulate Bernoullis with probability

$$P_{n+m}(\tau(n+m) = \infty, \mathcal{H}_\infty^{n+m}) = P_{n+m}(\tau_1(n+m) = \infty \mid \mathcal{H}_\infty^{n+m}) P_{n+m}(\mathcal{H}_\infty^{n+m}).$$

We summarize the algorithm as follows:

Algorithm 4.2: Simulate N_1 and N_2 jointly with $W_{i,k}^n$'s for $1 \leq n \leq N_0$, where N_0 is chosen such that $\sup_{t\in[0,1]} ||\hat{X}^{N_0}(t) - X(t)||_\infty \leq \varepsilon$

Input: The parameters required to run Algorithm 4.1, and Procedures A and B. These are the tilting parameters θ_0's.

S1. Simulate N_1 jointly with $W_{i,k}^m$'s for $0 \leq m \leq N_1$ using Algorithm 4.1 (see the remark that follows after Algorithm I). Let $n = N_1$.

S2. If any of the conditions (4.19) and (4.20) from Lemma 4.2.10 are not satisfied keep simulating $W_{i,k}^m$'s for $m > n$ until the first level $m > n$ for which conditions (4.19) and (4.20) are satisfied. Redefine n to be such first level m.

S3. Run Procedure B and obtain as output F and if $F = 1$ also obtain $\omega_{n:n+\tau(n)}$.

S4. If $\tau(n) < \infty$ (i.e. $F = 1$) set $n \longleftarrow \tau(n)$ and go back to Step 2. Otherwise, go to Step 4.

S5. Calculate G according to Procedure A and solve for N_0 such that $G\Delta_{N_0}^{2\alpha-\beta} < \varepsilon$.

S6. If $N_0 > n$ sample $\omega_{n:N_0}$ from $P_n(\cdot)$ and sample a Bernoulli random variable, I with probability of success $P_{N_0}(\tau(N_0) = \infty, \mathcal{H}_\infty^{N_0})$.

S7. If $I = 0$, go back to Step 6.

S8. Output $\omega_{0:N_0}$.

We obtain $\{W_{i,k}^l : 0 \leq k < 2^l, l \leq N_0, 1 \leq i \leq d\}$ from Algorithm II. We have from recursions (4.11) and (4.12) how to obtain

$$\{(Z_i(t_r^l) - Z_i(t_{r-1}^l)) : 1 \leq r \leq 2^l, 1 \leq l \leq N_0, 1 \leq i \leq d\} \tag{4.21}$$

and then we can compute $\{\hat{X}^{N_0}(t) : t \in D_{N_0}\}$ using equation (4.6).

Remark 4.2.3 *Observe that after completion of Algorithm 4.2, one can actually continue the simulation of increments in order to obtain an approximation with an error $\varepsilon' < \varepsilon$. In particular, this is done by repeating Steps 4 to 8. Start from Step 4 with $n = N_0$. The value of G has been computed, it does not depend on ε. However, one needs to recompute $N_0 := N_0(\varepsilon')$ such that $G\Delta_{N_0}^{2\alpha-\beta} < \varepsilon'$. Then we can implement Steps 5 to 8 without change. One obtains an output that, as before, can be transformed into (4.21) via the recursions (4.11), yielding $\{\hat{X}^{N_0(\varepsilon')}(t) : t \in D_{N_0(\varepsilon')}\}$ with a guaranteed error smaller than ε' in uniform norm with probability 1.*

4.3 Rough Path Differential Equations, Error Analysis, and The Proof of Theorem 4.1.1

The analysis in this section follows closely the discussion from [40] Section 3 and Section 7; see also [43] Chapter 10. We made some modifications to account for the drift of the process and also to be able to explicitly calculate the constant G. Let us start with the definition of a solution to (4.1) using the theory of rough differential equations.

Definition 4.3.1 *$X(\cdot)$ is a solution of (4.1) on $[0,1]$ if $X(0) = x(0)$ and*

$$|X_i(t) - X_i(s) - \mu_i(X(s))(t-s) - \sum_{j=1}^{d'} \sigma_{i,j}(X(s))(Z_j(t) - Z_j(s))$$
$$- \sum_{j=1}^{d'} \sum_{l=1}^{d} \sum_{m=1}^{d'} \partial_l \sigma_{i,j}(X(s)) \sigma_{l,m}(X(s)) A_{m,j}(s,t)| = o(t-s)$$

for all i *and* $0 \leq s < t \leq 1$, *where* $A_{i,j}(\cdot)$ *satisfies*

$$A_{i,j}(r,t) = A_{i,j}(r,s) + A_{i,j}(s,t) + (Z_i(s) - Z_i(r))(Z_j(t) - Z_j(s)) \tag{4.22}$$

for $0 \leq r < s < t \leq 1$.

The previous definition is motivated by the following Taylor-type development,

$$\begin{aligned}
X_i(t+h) =& X_i(t) + \int_t^{t+h} \mu_i(X(u))du + \sum_{j=1}^{d'} \int_t^{t+h} \sigma_{i,j}(X(u))dZ_j(u) \\
\approx& X_i(t) + \int_t^{t+h} \mu_i(X(u))du \\
&+ \sum_{j=1}^{d'} \int_t^{t+h} \sigma_{i,j}\left(X(t) + \mu(X(t))(u-t) + \sigma(X(t))(Z(u) - Z(t))\right) dZ_j(u) \\
\approx& X_i(t) + \mu_i(X(t))h + \sum_{j=1}^{d'} \sigma_{i,j}(X(t))(Z_j(t+h) - Z_j(t)) \\
&+ \sum_{j=1}^{d'} \sum_{l=1}^{d} \sum_{m=1}^{d'} \partial_l \sigma_{i,j}(X(t))\sigma_{l,m}(X(t)) \int_t^{t+h} (Z_m(u) - Z_m(t))dZ_j(u).
\end{aligned}$$

The previous Taylor development suggests defining $A_{i,j}(s,t) := \int_s^t (Z_i(u) - Z_i(s))dZ_j(u)$. Depending on how one interprets $A(s,t)$, e.g. via Ito or Stratonovich integrals, one obtains a solution $X(\cdot)$ which is interpreted in the corresponding context.

In order to obtain the Ito interpretation of the solution to equation (4.1) via definition (4.3.1) we shall interpret the integrals in the sense of Ito. In addition, as we shall explain, some technical conditions (in addition to the standard Lipschitz continuity typically required to obtain a strong solution a la Ito) must be imposed in order to enforce the existence of a unique solution to (4.3.1).

There are two sources of errors when using $\hat{X}^n$ in equation (4.6) to approximate X. One is the discretization on the dyadic grid, but assuming that $A_{i,j}\left(t_k^n, t_{k+1}^n\right)$ is known; this type of analysis is the one that is most common in the literature on rough paths (see [40]). The second source of error arises precisely accounting for the fact

that $A_{i,j}\left(t_k^n, t_{k+1}^n\right)$ is not known. Thus we divide the proof of Theorem 4.1.1 into two steps (two propositions), each dealing with one source of error.

Similar to $\hat{X}^n(t)$, we define $\{X^n(t) : t \in D_n\}$ by the following recursion: given $X^n(0) = X(0)$,

$$\begin{aligned} X_i^n(t_{k+1}^n) =& X_i^n(t_k^n) + \mu_i(x^n(t_k^n))\Delta_n + \sum_{j=1}^{d'} \sigma_{i,j}(X_i^n(t_k^n))(Z_j(t_{k+1}^n) - Z_j(t_k^n)) \\ &+ \sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'} \partial_l\sigma_{i,j}(X_i^n(t_k^n))\sigma_{l,m}(X_i^n(t_k^n))A_{m,j}(t_k^n, t_{k+1}^n), \end{aligned} \tag{4.23}$$

and for $t \in [0,1]$, we let $X^n(t) = X^n(\lfloor t \rfloor)$, where in this context $\lfloor t \rfloor = \max\{s \in D_n : s \leq t\}$.

Proposition 4.3.2 *Under the conditions of Theorem 4.1.1, we can compute a constant G_1 explicitly in terms of M, $||Z||_\alpha$ and $||A||_{2\alpha}$, such that for n large enough*

$$||X^n(t) - X(t)||_\infty \leq G_1\Delta_n^{3\alpha-1}.$$

The proof of Proposition 4.3.2 will be given after introducing some definitions and key auxiliary results. We denote

$$I_i^n(r,t) := X_i^n(t) - X_i^n(r) - \mu_i(X^n(r))(t-r) - \sum_{j=1}^{d'} \sigma_{i,j}(X^n(r))(Z_j(t) - Z_j(r))$$

and

$$J_i^n(r,t) := I_i^n(r,t) - \sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'} \partial_l\sigma_{i,j}(X^n(r))\sigma_{l,m}(X^n(r))A_{m,j}(r,t).$$

The following lemmas introduce the main technical results for the proof of Proposition 4.3.2.

Lemma 4.3.3 *Under the conditions of Theorem 4.1.1, there exist constants C_1, C_2 and C_3 that depend only on M, $||Z||_\alpha$ and $||A||_{2\alpha}$, such that for any large enough n and $r, t \in D_n$,*

$$\begin{aligned} ||X^n(t) - X^n(r)||_\infty &\leq C_1|t-r|^\alpha, \\ |I^n(r,t)||_\infty &\leq C_2|t-r|^{2\alpha}, \end{aligned}$$

and

$$||J^n(r,t)||_\infty \leq C_3|t-r|^{3\alpha}.$$

Proof. We follow For $r \leq s \leq t$, $r,s,t \in D_n$, we have the following important recursions:

$$\begin{aligned} I_i^n(r,t) &= I_i^n(r,s) + I_i^n(s,t) + (\mu_i(X^n(s)) - \mu_i(X^n(r)))(t-s) \\ &\quad + \sum_{j=1}^{d'} (\sigma_{i,j}(X^n(s)) - \sigma_{i,j}(X^n(r)))(Z_j(t) - Z_j(s)) \end{aligned}$$

and

$$\begin{aligned} & J_i^n(r,t) \\ =& J_i^n(r,s) + J_i^n(s,t) + (\mu_i(X^n(s)) - \mu_i(X^n(r)))\,(t-s) \\ &+ \sum_{j=1}^{d'} [\sigma_{i,j}(X^n(s)) - \sigma_{i,j}(X^n(r)) - \sum_{l=1}^{d} \partial_l \sigma_{i,j}(X^n(r))(X_l^n(s) - X_l^n(r)) \\ &+ \sum_{l=1}^{d} \partial_l \sigma_{i,j}(X^n(r)) I_l^n(r,s)](Z_j(t) - Z_j(s)) \\ &+ \sum_{j=1}^{d'} \sum_{l=1}^{d} \sum_{m=1}^{d'} [\partial_l \sigma_{i,j}(X^n(s))\sigma_{l,m}(X^n(s)) - \partial_l \sigma_{i,j}(X^n(r))\sigma_{l,m}(X^n(r))]\, A_{m,j}(s,t) \end{aligned} \tag{4.24}$$

We next divide the proof into two parts. We first prove that there exists a small enough constant $\delta > 0$ and three large enough constants $C_1(\delta)$, $C_2(\delta)$ and $C_3(\delta)$, all independent of n, such that for $|t-r| < \delta$, $||X^n(t) - X^n(r)||_\infty \leq C_1(\delta)|t - r|^\alpha$, $||I^n(r,t)||_\infty \leq C_2(\delta)|t-r|^{2\alpha}$ and $||J^n(r,t)||_\infty \leq C_3(\delta)|t-r|^{3\alpha}$. We prove it by induction. First we have $J^n(r,r) = 0$ and $J^n(r, r+\Delta_n) = 0$. Suppose the result hold for all pairs of $r_0, t_0 \in D_n$ with $|t_0 - r_0| < |t-r|$. We then pick $s \in D_n$ as the largest point between r and t such that $|s-r| \leq |t-r|/2$. Then we also have $|s + \Delta_n - r| > |t-r|/2$ and $|t - (s+\Delta_n)| < |t-r|/2$.

As

$$X_i^n(t) - X_i^n(s) = J_i^n(s,t) + \mu_i(X^n(s))(t-s) + \sum_{j=1}^{d'} \sigma_{i,j}(X^n(s))(Z_j(t) - Z_j(s)) + \sum_{j=1}^{d'} \sum_{l=1}^{d} \sum_{m=1}^{d'} \partial_l \sigma_{i,j}(X^n(s)) \sigma_{l,m}(X^n(s)) A_{m,j}(s,t),$$

we have

$$\begin{aligned} & |X_i^n(t) - X_i^n(s)| \\ \leq & C_3(\delta)|t-s|^{3\alpha} + M|t-s| + dM||Z||_\alpha |t-s|^\alpha + d^3M^2||A||_{2\alpha}|t-s|^{2\alpha} \\ \leq & (C_3(\delta)\delta^{2\alpha} + M\delta^{1-\alpha} + dM||Z||_\alpha + d^3M^2||A||_{2\alpha}\delta^\alpha)|t-s|^\alpha \\ \leq & C_1(\delta)|t-s|^\alpha \end{aligned}$$

for $C_1(\delta) > C_3(\delta)\delta^{2\alpha} + M\delta^{1-\alpha} + dM||Z||_\alpha + d^3M^2||A||_{2\alpha}\delta^\alpha$.

And as

$$I_i^n(s,t) = J_i^n(s,t) + \sum_{j=1}^{d'} \sum_{l=1}^{d} \sum_{m=1}^{d'} \partial_l \sigma_{i,j}(X^n(s)) \sigma_{l,m}(X^n(s)) A_{m,j}(s,t),$$

we have

$$\begin{aligned} |I_i^n(s,t)| &\leq C_3(\delta)|t-s|^{3\alpha} + d^3M^2||A||_{2\alpha}|t-s|^{2\alpha} \\ &\leq (C_3(\delta)\delta^\alpha + d^3M^2||A||_{2\alpha})|t-s|^{2\alpha} \leq C_2(\delta)|t-s|^{2\alpha} \end{aligned}$$

for $C_2(\delta) > C_3(\delta)\delta^\alpha + d^3M^2||A||_{2\alpha}$.

We now analyze the recursion (4.24) term by term. First,

$$|\mu_i(X^n(s)) - \mu_i(X^n(r))| \leq MC_1(\delta)|s-r|^\alpha,$$

$$|\sigma_{i,j}(X^n(s)) - \sigma_{i,j}(X^n(r)) - \sum_{l=1}^{d} \partial_l \sigma_{i,j}(X^n(r))(X_l^n(s) - X_l^n(r))| \leq MC_1(\delta)^2|s-r|^{2\alpha},$$

$$|\sum_{l=1}^{d} \partial_l \sigma_{i,j}(X^n(r)) I_l^n(r,s)| \leq dMC_2(\delta)|s-r|^{2\alpha},$$

and

$$|\partial_l\sigma_{i,j}(X^n(s))\sigma_{l,m}(X^n(s)) - \partial_l\sigma_{i,j}(X^n(r))\sigma_{l,m}(X^n(r))| \leq 2M^2C_1(\delta)|s-r|^\alpha.$$

Then

$$\begin{aligned}
&|J_i^n(r,t)| \\
\leq&|J_i^n(r,s)| + |J_i^n(s,t)| \\
&+ (MC_1(\delta) + dMC_1(\delta)^2||Z||_\alpha + d^2MC_2(\delta)||Z||_\alpha + 2d^3M^2C_1(\delta)||A||_\alpha)|t-r|^{3\alpha}
\end{aligned}$$

Likewise, we have

$$\begin{aligned}
&|J_i^n(s,t)| \\
\leq&|J_i^n(s,s+\Delta_n)| + |J_i^n(s+\Delta_n,t)| \\
&+ (MC_1(\delta) + dMC_1(\delta)^2||Z||_\alpha + d^2MC_2(\delta)||Z||_\alpha + 2d^3M^2C_1(\delta)||A||_\alpha)|t-s|^{3\alpha} \\
=&|J_i^n(s+\Delta_n,t)| \\
&+ (MC_1(\delta) + dMC_1(\delta)^2||Z||_\alpha + d^2MC_2(\delta)||Z||_\alpha + 2d^3M^2C_1(\delta)||A||_\alpha)|t-s|^{3\alpha}.
\end{aligned}$$

Then

$$\begin{aligned}
&|J_i^n(r,t)| \\
\leq&|J_i^n(r,s)| + |J_i^n(s+\Delta_n,t)| \\
&+ 2\{MC_1(\delta) + dMC_1(\delta)^2||Z||_\alpha + d^2MC_2(\delta)||Z||_\alpha + 2d^3M^2C_1(\delta)||A||_\alpha\}|t-s|^{3\alpha} \\
\leq&\{2^{1-3\alpha}C_3(\delta) + 2(MC_1(\delta) + dMC_1(\delta)^2||Z||_\alpha + d^2MC_2(\delta)||Z||_\alpha \\
&+ 2d^3M^2C_1(\delta)||A||_\alpha)\}|t-s|^{3\alpha} \\
\leq&C_3(\delta)|t-s|^{3\alpha},
\end{aligned}$$

for

$$\begin{aligned}
&(1-2^{1-3\alpha})C_3(\delta) \\
>&2(MC_1(\delta) + dMC_1(\delta)^2||Z||_\alpha + d^2MC_2(\delta)||Z||_\alpha + 2d^3M^2C_1(\delta)||A||_\alpha).
\end{aligned}$$

Therefore, if we deliberately choose δ, $C_1(\delta)$, $C_2(\delta)$ and $C_3(\delta)$ such that

$$\begin{aligned}
C_1(\delta) &> C_3(\delta)\delta^{2\alpha} + M\delta^{1-\alpha} + dM||Z||_\alpha + d^3M^2||A||_{2\alpha}\delta^\alpha \\
C_2(\delta) &> C_3(\delta)\delta^\alpha + d^3M^2||A||_{2\alpha} \\
C_3(\delta) &> \frac{2}{1-2^{1-3\alpha}}\big(MC_1(\delta) + dMC_1(\delta)^2||Z||_\alpha + d^2MC_2(\delta)||Z||_\alpha \\
&\qquad +2d^3M^2C_1(\delta)||A||_\alpha\big)
\end{aligned}$$

Then we have

$$\begin{aligned}
||X^n(t) - X^n(r)||_\infty &\le C_1(\delta)|t-r|^\alpha, \\
||I^n(r,t)||_\infty &\le C_2(\delta)|t-r|^{2\alpha}, \\
||J^n(r,t)||_\infty &\le C_3(\delta)|t-r|^{3\alpha},
\end{aligned}$$

for $|t-r| < \delta$.

We now extend the analysis to the case when $|t-r| > \delta$. For n large enough ($\Delta_n < \delta/2$), if $|t-r| > \delta$, we can always find points $s_i \in D_n$ and $r = s_0 < s_1 < \cdots < s_k = t$ such that $\max_{1\le i\le k}|s_i - s_{i-1}| < \delta$ and $\min_{1\le i\le k}|s_i - s_{i-1}| > \delta/2$. Then

$$|X_i^n(t) - X_i^n(r)| \le \sum_{l=1}^{k}|X_i^n(s_l) - X_i^n(s_{l-1})| \le kC_1(\delta)|t-r|^\alpha \le \frac{2}{\delta}C_1(\delta)|t-r|^\alpha$$

Let $C_1 = \frac{2}{\delta}C_1(\delta)$ and we can write $||X^n(t) - X^n(r)||_\infty < C_1|t-r|^\alpha$. Next,

$$\begin{aligned}
|I_i^n(r,t)| \le& \sum_{l=1}^{k}\{|I_i^n(s_{l-1},s_l)| + |(\mu_i(X^n(s_l)) - \mu_i(X^n(s_0)))(s_l - s_{l-1})| \\
&+ |(\sigma_i(X^n(s_l)) - \sigma_i(X^n(s_0)))(Z(s_{l+1}) - Z(s_l))|\} \\
\le& k[C_2(\delta)|t-r|^{2\alpha} + MC_1|t-r|^{1+\alpha} + dMC_1||Z||_\alpha|t-r|^{2\alpha}] \\
\le& \frac{2}{\delta}(C_2(\delta) + MC_1 + dMC_1||Z||_\alpha)|t-r|^{2\alpha}
\end{aligned}$$

By setting $C_2 = \frac{2}{\delta}(C_2(\delta) + MC_1 + dMC_1||Z||_\alpha)$, we have $||I^n(r,t)||_\infty < C_2|t-r|^{2\alpha}$. Now following the same induction analysis on $J_i^n(s,t)$ as we did in the case $|t-s| < \delta$,

we have

$$\begin{aligned}&|J_i^n(r,t)|\\ \leq&\frac{2}{2^{3\alpha}}C_3|t-r|^{3\alpha}+2(MC_1+dMC_1^2||Z||_\alpha+d^2MC_2||Z||_\alpha+2d^3M^2C_1||A||_\alpha)|t-r|^{3\alpha}\end{aligned}$$

If we choose

$$C_3=\frac{2}{1-2^{1-3\alpha}}(MC_1+dMC_1^2||Z||_\alpha+d^2MC_2||Z||_\alpha+2d^3M^2C_1||A||_\alpha),$$

then $||J^n(r,t)||_\infty\leq C_3|t-s|^{3\alpha}$.

□

Lemma 4.3.4 *Let $x(0)$ and $\tilde{x}(0)\in\mathbb{R}^d$ be two different vectors. We denote $X^n(t)$ and $\tilde{X}^n(t)$ for $t\in D_n$ as the n-th dyadic approximation defined by (4.23) with initial value $x(0)$ and $\tilde{x}(0)$ respectively. Under the conditions of Theorem 4.1.1, there exists a constant B, independent of n, such that for $t\in D_n$,*

$$||X^n(t)-\tilde{X}^n(t)-(X^n(0)-\tilde{X}^n(0))||_\infty\leq Bt^\alpha||X^n(0)-\tilde{X}^n(0)||_\infty.$$

Moreover,

$$||X^n(t)-\tilde{X}^n(t)||_\infty\leq(1+B)||X^n(0)-\tilde{X}^n(0)||_\infty.$$

Proof. Let

$$Y_{i,h}^n(t)=\frac{X_i^n(t)-\tilde{X}_i^n(t)}{||X_h^n(0)-\tilde{X}_h^n(0)||_\infty}$$

We define $0/0 = 0$.

Then following the recursion (4.23), we have

$$\begin{aligned}
&Y^n_{i,h}(t^n_{k+1}) \\
=&Y^n_{i,h}(t^n_k) + \frac{\mu_i(X^n(t^n_k)) - \mu_i(\tilde{X}^n(t^n_k))}{||X^n_h(0) - \tilde{X}^n_h(0)||_\infty}\Delta_n \\
&+ \sum_{j=1}^{d'} \frac{\sigma_{i,j}(X^n(t^n_k)) - \sigma_{i,j}(\tilde{X}^n(t^n_k))}{||X^n_h(0) - \tilde{X}^n_h(0)||_\infty}(Z_j(t^n_{k+1}) - Z_j(t^n_k)) \\
&+ \sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'}\{\frac{\partial_l\sigma_{i,j}(X^n(t^n_k))\sigma_{l,m}(X^n(t^n_k)) - \partial_l\sigma_{i,j}(\tilde{X}^n(t^n_k))\sigma_{l.m}(\tilde{X}^n(t^n_k))}{||X^n_h(0) - \tilde{X}^n_h(0)||_\infty} \\
&\qquad \times A_{m,j}(t^n_k, t^n_{k+1})\} \qquad (4.25)
\end{aligned}$$

Then (4.23) and (4.25) together define an recursion to generate X^n, $\tilde{X}^n$ and Y^n. Following Lemma 4.3.3, there exists a constant B that depends only on M, $||Z||_\alpha$ and $||A||_{2\alpha}$, such that

$$||Y^n(t) - Y^n(0)||_\infty \leq Bt^\alpha.$$

Thus,

$$||X^n(t) - \tilde{X}^n(t) - (X^n(0) - \tilde{X}^n(0))||_\infty \leq Bt^\alpha||X^n(0) - \tilde{X}^n(0)||_\infty,$$

and

$$||X^n(t) - \tilde{X}^n(t)||_\infty \leq (1 + B)||X^n(0) - \tilde{X}^n(0)||_\infty.$$

□

We are now ready to prove Proposition 4.3.2.

Proof. [Proof of Proposition 4.3.2]From Lemma 4.3.3 we have $||X^n(t) - X^n(r)||_\infty \leq C_1|t - r|^\alpha$. By Arzela-Ascoli Theorem, there exits a subsequence of $\{X^n\}$ that converges uniformly to some continuous function X on $[0, 1]$. Moreover we have

$||X(t) - X(r)||_\infty \leq C_1|t-r|^\alpha$ and

$$|X_i(t) - X_i(r) - \mu_i(X(r) - \sum_{j=1}^{d'} \sigma_{i,j}(X(r))(Z_j(t) - Z_j(r))$$
$$- \sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'} \partial_l \sigma_{i,j}(X(r))\sigma_{l.m}(X(r))A_{m,j}(r,t)| < C_2|t-r|^{3\alpha}$$

Therefore, the limit X is a solution to the SDE.

Let $X^{n,(s)}(t; X(s)) := X^n(t-s)|X^n(0) = X(s)$. Specifically, we have $X^{n,(0)}(t; X(0)) = X^n(t)$ with $X^n(0) = X(0)$, and $X^{n,(t)}(t; X(t)) = X(t)$. Then we can write

$$X^n(t_m^n) - X(t_m^n) = \sum_{k=1}^{m} \left(X^{n,(t_k^n)}(t_m^n; X(t_k^n)) - X^{n,(t_{k-1}^n)}(t_m^n; X(t_{k-1}^n))\right)$$

By Lemma 4.3.4, $||X^{n,(t_k^n)}(t_m; X(t_k^n)) - X^{n,(t_{k-1}^n)}(t_m; X(t_{k-1}^n))||_\infty \leq (1+B)||X(t_k^n) - X^{n,t_{k-1}^n}(t_k^n; X(t_{k-1}^n))||_\infty$. We also have

$$\begin{aligned}
&|X_i(t_k^n) - X_i^{n,(t_{k-1}^n)}(t_k^n; X(t_{k-1}^n))| \\
=&|X_i(t_k^n) - X_i(t_{k-1}^n) - \mu_i(X(t_{k-1}^n)(t_k^n - t_{k-1}^n) - \sum_{j=1}^{d'} \sigma_{i,j}(X(t_{k-1}^n))(Z_j(t_k^n) - Z_j(t_{k-1}^n)) \\
&- \sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'} \partial_l \sigma_{i,j}(X(t_{k-1}^n))\sigma_{l,m}(X(t_{k-1}^n))A_{m,j}(t_{k-1}^n, t_k^n)| \\
\leq& C_3|t_k^n - t_{k-1}^n|^{3\alpha}
\end{aligned}$$

Thus,

$$\begin{aligned}
||X^n(t_m^n) - X(t_m^n)||_\infty &\leq \sum_{k=1}^{m} ||X^{n,(t_k^n)}(t_m^n; X(t_k^n)) - X^{n,(t_{k-1}^n)}(t_m^n; X(t_{k-1}^n))||_\infty \\
&\leq m(1+B)C_3\Delta_n^{3\alpha} \\
&\leq (1+B)C_3\Delta_n^{3\alpha-1}.
\end{aligned}$$

□

Next we turn to the analysis of the error induced by approximating the Lévy area.

Proposition 4.3.5 *Under the conditions of Theorem 4.1.1, we can compute a constant G_2 explicitly in terms of M, $||Z||_\alpha$, $||A||_{2\alpha}$ and Γ_R, such that for n large enough*

$$||\hat{X}^n(t) - X^n(t)||_\infty \leq G_2 \Delta_n^{2\alpha-\beta}.$$

The proof of Proposition 4.3.5 uses a similar technique as the proof of Proposition 4.3.2 and also relies on some auxiliary results. Let

$$U_i^n(s,t) := \hat{X}_i^n(t) - X_i^{n,(s)}(t; \hat{X}^n(s)) + \sum_{j=1}^{d'} \sum_{l=1}^{d} \sum_{m=1}^{d'} \partial_l \sigma_{i,j}(\hat{X}^n(s)) \sigma_{l,m}(\hat{X}^n(s)) R_{m,j}^n(s,t).$$

We first prove the following technical result.

Lemma 4.3.6 *Under the conditions of Theorem 4.1.1, there exists a constant C_4, that depends only on M, $||Z||_\alpha$, $||A||_{2\alpha}$ and Γ_R, such that*

$$||U^n(r,t)||_\infty \leq C_4 |t-r|^{\alpha+\beta} \Delta_n^{2\alpha-\beta}$$

Proof. For $0 \leq r < s < t \leq 1$, $r, s, t \in D_n$, we have

$$\begin{aligned}
&U_i^n(r,t) \\
=&U_i^n(r,s) + U_i^n(s,t) \\
&+ \left[X_i^{n,(s)}(t; \hat{X}^n(s)) - X_i^{n,(r)}(t; \hat{X}^n(r)) - (\hat{X}_i^n(s) - X_i^{n,(r)}(s; \hat{X}^n(r))) \right] \\
&- \sum_{j=1}^{d'} \sum_{l=1}^{d} \sum_{m=1}^{d'} \left(\partial_l \sigma_{i,j}(\hat{X}^n(s)) \sigma_{l,m}(\hat{X}^n(s)) - \partial_l \sigma_{i,j}(\hat{X}^n(r)) \sigma_{l,m}(\hat{X}^n(r)) \right) R_{m,j}^n(s,t)
\end{aligned}$$

From Lemma 4.3.4,

$$\begin{aligned}
&|X_i^{n,(s)}(t; \hat{X}^n(s)) - X_i^{n,(r)}(t; \hat{X}^n(r)) - \left(\hat{X}_i^n(s) - X_i^{n,(r)}(s; \hat{X}^n(r)) \right)| \\
\leq& B|t-s|^\alpha ||\hat{X}^n(s) - X^{n,(r)}(s; \hat{X}^n(r))||_\infty
\end{aligned}$$

From Lemma 4.3.3,

$$\begin{aligned}
&\left| \left(\partial_l \sigma_{i,j}(\hat{X}^n(s)) \sigma_{l,m}(\hat{X}^n(s)) - \partial_l \sigma_{i,j}(\hat{X}^n(r)) \sigma_{l,m}(\hat{X}^n(r)) \right) R_{m,j}^n(s,t) \right| \\
\leq& 2M^2 C_1 |s-r|^\alpha \Gamma_R |t-s|^\beta \Delta_n^{2\alpha-\beta} \\
\leq& 2M^2 C_1 \Gamma_R |t-r|^{\alpha+\beta} \Delta_n^{2\alpha-\beta}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&||U^n(r,t)||_\infty \\
\leq&||U^n(r,s)||_\infty + ||U^n(s,t)||_\infty + B|t-s|^\alpha||\hat{X}^n(s) - X^{n,(r)}(t_p^n;\hat{X}^n(r))||_\infty \\
&+ 2d^3M^2C_1\Gamma_R|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta} \\
\leq&||U^n(r,s)||_\infty + ||U^n(s,t)||_\infty + B|t-s|^\alpha||U^n(r,s)||_\infty \\
&+ B|t-s|^\alpha \max_i\{|\sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'}\partial_l\sigma_{i,j}(\hat{X}^n(r))\sigma_{l,m}(\hat{X}^n(r))R^n_{m,j}(r,s)|\} \\
&+ 2d^3M^2C_1\Gamma_R|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta} \\
\leq&(1+B|t-s|^\alpha)||U^n(r,s)||_\infty + ||U^n(s,t)||_\infty \\
&+ (Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta} \qquad (4.26)
\end{aligned}$$

Like the proof of Lemma 4.3.3, we divide the proof into two parts. We first prove that there exist a small enough constant $\delta > 0$ and a large enough constant $C_4(\delta)$, both independent of n, such that for $|t-r| < \delta$, $|U^n(r,t)| \leq C_4(\delta)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta}$. And we prove it by induction. First we have $U^n_{t_k^n,t_k^n} = 0$ and $U^n_{t_k^n,t_{k+1}^n} = 0$. Suppose the bound holds for all pairs $r_0, t_0 \in D_n$ with $|t_0 - r_0| < |t-r|$. We pick $s \in D_n$ as the largest point between r and t such that $|s-r| \leq 1/2|t-r|$. Then we also have $|(s+\Delta_n) - r| > 1/2|t-r|$ and $|t-(s+\Delta_n)| < 1/2|t-r|$.

$$\begin{aligned}
||U^n(r,t)||_\infty \leq&(1+B|t-s|^\alpha)||U^n(r,s)||_\infty + ||U^n(s,t)||_\infty \\
&+ (Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta}
\end{aligned}$$

and

$$\begin{aligned}
&||U^n(s,t)||_\infty \\
\leq&(1+B\Delta_n^\alpha)||U^n(s,s+\Delta_n)||_\infty + ||U^n(s+\Delta_n,t)||_\infty \\
&+ (Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)|t-s|^{\alpha+\beta}\Delta_n^{2\alpha-\beta} \\
\leq&||U^n(s+\Delta_n,t)||_\infty + (Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&||U^n(r,t)||_\infty \\
\leq&(1+B\delta^\alpha)||U^n(r,s)||_\infty + ||U^n(s+\Delta_n,t)||_\infty \\
&+ 2(Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta} \\
\leq&\frac{2+B\delta^\alpha}{2^{\alpha+\beta}}C_4(\delta)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta} + 2(Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta}.
\end{aligned}$$

If we pick δ and $C_4(\delta)$ such that

$$B\delta^\alpha \leq 2^{\alpha+\beta} - 2$$

and

$$(1-\frac{2+B\delta^\alpha}{2^{\alpha+\beta}})C_4(\delta) \geq 2(Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R),$$

Then $||U^n(r,t)||_\infty \leq C(\delta)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta}$. We next extend the result to the case when $|t-r| > \delta$. We can always divide the interval $[r,t]$ into smaller intervals of length less than δ, specifically, for n large enough, we consider $r = s_0 < s_1 < \cdots < s_k = t$ where $s_i \in D_n$ and $1/2\delta < |s_i - s_{i-1}| < \delta$ for $i = 1, 2, \ldots, k$. Then $k < 2|t-r|/\delta \leq 2/\delta$ and

$$\begin{aligned}
&||U^n(r,t)||_\infty \\
\leq&(1+B|s_1-s_0|^\alpha)||U^n(s_0,s_0)||_\infty + ||U^n(s_1,s_2)||_\infty \\
&+ (Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta} \\
\leq&\sum_{i=1}^{k}(1+B\delta^\alpha)||U^n(s_{i-1},s_i)||_\infty + k(Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta} \\
\leq&(1+B\delta^\alpha)C_4(\delta)\Delta_n^{2\alpha-\beta}\sum_{i=1}^{k}|s_i - s_{i-1}|^{\alpha+\beta} \\
&+ k(Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta} \\
\leq&(1+B\delta^\alpha)C_4(\delta)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta} + \frac{2}{\delta}(Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)|t-r|^{\alpha+\beta}\Delta_n^{2\alpha-\beta} \\
\leq&C_4|m-k|^{\alpha+\beta}\Delta_n^{2\alpha-\beta}
\end{aligned}$$

for $C_4 \geq (1+B\delta^\alpha)C_4(\delta) + 2(Bd^3M^2\Gamma_R + 2d^3M^2C_1\Gamma_R)/\delta$.

□

We are now ready to prove Proposition 4.3.5.

Proof. [Proof of Proposition 4.3.5]From Lemma 4.3.6, we have

$$||U^n(0,t)||_\infty \leq C_4 t^{\alpha+\beta}\Delta_n^{2\alpha-\beta}.$$

Then

$$\begin{aligned}|\hat{X}_i^n(t) - X_i^n(t)| &\leq |U_i^n(0,t)| + \sum_{j=1}^{d}\sum_{l=1}^{d}\sum_{m=1}^{d}|\partial_l\sigma_{i,j}(X(0))\sigma_{l,m}(X(0))||R_{m,j}^n(0,t)| \\ &\leq C_4 t^{\alpha+\beta}\Delta_n^{2\alpha-\beta} + d^3M^2\Gamma_R t^\beta \Delta_n^{2\alpha-\beta} \\ &\leq (C_4 + d^3M^2\Gamma_R)\Delta_n^{2\alpha-\beta}.\end{aligned}$$

□

4.4 Proof of Technical Results

4.4.1 Proof of Technical Results in Section 4.2.3

We now provide the proofs of the results in the order in which they were presented in the Section 4.2.3. We start by recalling the following algebraic property of the Lévy areas: for each $0 \leq r < s < t$

$$A_{i,j}(r,t) = A_{i,j}(r,s) + A_{i,j}(s,t) + (Z_i(s) - Z_i(r))(Z_j(t) - Z_j(s)). \tag{4.27}$$

Using this property and a simple use of the Borel-Cantelli lemma we can obtain the proof of Lemma 4.2.4.

Proof. [Proof of Lemma 4.2.4] We use (4.27) repeatedly. First, note that

$$\begin{aligned}A_{i,j}\left(t_k^n, t_{k+1}^n\right) &= A_{i,j}\left(t_{2k}^{n+1}, t_{2k+1}^{n+1}\right) + A_{i,j}\left(t_{2k+1}^{n+1}, t_{2k+2}^{n+1}\right) \\ &\quad + \left(Z_i\left(t_{2k+1}^{n+1}\right) - Z_i\left(t_{2k}^{n+1}\right)\right)\left(Z_j\left(t_{2k+2}^{n+1}\right) - Z_j\left(t_{2k+1}^{n+1}\right)\right).\end{aligned}$$

We continue, this time splitting $A_{i,j}\left(t_{2k}^{n+1}, t_{2k+1}^{n+1}\right)$ and $A_{i,j}\left(t_{2k+1}^{n+1}, t_{2k+2}^{n+1}\right)$, thereby obtaining

$$\begin{aligned}
&A_{i,j}\left(t_k^n, t_{k+1}^n\right) \\
&= \left(Z_i\left(t_{2k+1}^{n+1}\right) - Z_i\left(t_{2k}^{n+1}\right)\right)\left(Z_j\left(t_{2k+2}^{n+1}\right) - Z_j\left(t_{2k+1}^{n+1}\right)\right) \\
&\quad + A_{i,j}\left(t_{2^2k}^{n+2}, t_{2^2k+1}^{n+2}\right) + A_{i,j}\left(t_{2^2k+1}^{n+2}, t_{2^2k+2}^{n+2}\right) \\
&\quad + \left(Z_i\left(t_{2^2k+1}^{n+2}\right) - Z_i\left(t_{2^2k}^{n+2}\right)\right)\left(Z_j\left(t_{2^2k+2}^{n+2}\right) - Z_j\left(t_{2^2k+1}^{n+2}\right)\right) \\
&\quad + A_{i,j}\left(t_{2^2k+2}^{n+2}, t_{2^2k+3}^{n+2}\right) + A_{i,j}\left(t_{2^2k+3}^{n+2}, t_{2^2k+4}^{n+2}\right) \\
&\quad + \left(Z_i\left(t_{2^2k+3}^{n+2}\right) - Z_i\left(t_{2^2k+2}^{n+2}\right)\right)\left(Z_j\left(t_{2^2k+4}^{n+2}\right) - Z_j\left(t_{2^2k+3}^{n+2}\right)\right).
\end{aligned}$$

Iterating m times the previous splitting procedure we conclude that

$$\begin{aligned}
&A_{i,j}\left(t_k^n, t_{k+1}^n\right) \\
&= \sum_{h=n+1}^{m} \sum_{l=1}^{2^{h-n-1}} [Z_i(t_{2^{h-n}k+2l-1}^h) - Z_i(t_{2^{h-n}k+2l-2}^h)][Z_j(t_{2^{h-n}k+2l}^h) - Z_j(t_{2^{h-n}k+2l-1}^h)] \\
&\quad + \sum_{l=1}^{2^{m-n-1}} A_{i,j}\left(t_{2^{m-n}k+2l-2}^m, t_{2^{m-n}k+2l-1}^m\right) + \sum_{l=1}^{2^{m-n-1}} A_{i,j}\left(t_{2^{m-n}k+2l-1}^m, t_{2^{m-n}k+2l}^m\right).
\end{aligned} \tag{4.28}$$

We claim that

$$\sum_{l=1}^{2^{m-n-1}} A_{i,j}\left(t_{2^{m-n}k+2l-2}^m, t_{2^{m-n}k+2l-1}^m\right) + \sum_{l=1}^{2^{m-n-1}} A_{i,j}\left(t_{2^{m-n}k+2l-1}^m, t_{2^{m-n}k+2l}^m\right) \to 0 \tag{4.29}$$

almost surely as $m \to \infty$. To see this note that

$$\begin{aligned}
&P\left(\left|\sum_{l=1}^{2^{m-n-1}} A_{i,j}\left(t_{2^{h-n}k+2l-2}^h, t_{2^{h-n}k+2l-1}^h\right)\right| > 1/m\right) \\
&\leq m^2 \sum_{l=1}^{2^{m-n-1}} E(A_{i,j}^2\left(t_{2^{m-n}k+2l-2}^m, t_{2^{m-n}k+2l-1}^m\right)) = m^2 2^{m-n+1} E\int_0^{\Delta_m} Z_i^2(s)\, ds \\
&= m^2 2^{m-n}\Delta_m^2 = 2^n m^2 \Delta_m.
\end{aligned}$$

Since $\sum_{m=1}^{\infty} m^2\Delta_m < \infty$ we conclude by Borel-Cantelli's lemma that, almost surely, for m large enough

$$\left| \sum_{l=1}^{2^{m-n-1}} A_{i,j}\left(t^h_{2^{h-n}k+2l-2}, t^h_{2^{h-n}k+2l-1}\right) \right| < 1/m$$

and thus indeed we have (4.29) holds almost surely and therefore, from (4.28), sending $m \to \infty$ we obtain the conclusion of the lemma. □

Then we present the proof of Lemma 4.2.5.

Proof. [Proof of Lemma 4.2.5] Define

$$\mathcal{C}_n = \{|L^n_{i,j}(m) - L^n_{i,j}(l)| > (m-l)^\beta \Delta_n^{2\alpha} \text{ for some } 0 \le l < m < 2^{n-1}\}.$$

We will show that the events $\{\mathcal{C}_n : n \ge 0\}$ occur finitely many times. Note that

$$P(\mathcal{C}_n) \le \sum_{0\le l<m<2^{n-1}} 2P\left(\left(L^n_{i,j}(m) - L^n_{i,j}(l)\right) > (m-l)^\beta \Delta_n^{2\alpha}\right). \tag{4.30}$$

Also observe that for fixed m and n, $L^n_{i,j}(m)$ is the sum of m i.i.d. random variables, each of which is distributed as $(Z_i(t^n_1) - Z_i(t^n_0))(Z_j(t^n_2) - Z_j(t^n_1))$ and we easily evaluate

$$E\exp\left(\theta(Z_i(t^n_1) - Z_i(t^n_0))(Z_j(t^n_2) - Z_j(t^n_1))\right) = \left(1 - \theta^2\Delta_n^2\right)^{-1/2}.$$

We apply Chernoff's bound concluding

$$\begin{aligned} & P\left(\left(L^n_{i,j}(m) - L^n_{i,j}(l)\right) > (m-l)^\beta \Delta_n^{2\alpha}\right) \\ \le\ & \exp\left(-\theta\left(m-l\right)^\beta \Delta_n^{2\alpha} - \frac{1}{2}(m-l)\log\left(1-\theta^2\Delta_n^2\right)\right). \end{aligned}$$

Select $\theta = \theta'\left(m-l\right)^{-1/2}\Delta_n^{-1}$ for $\theta' \in (0, 1/4)$

$$P\left(\left(L^n_{i,j}(m) - L^n_{i,j}(l)\right) > (m-l)^\beta \Delta_n^{2\alpha}\right) \le \exp\left(-\theta'\left(m-l\right)^{\beta-1/2}\Delta_n^{2\alpha-1} + 1\right).$$

Hence,

$$P(\mathcal{C}_n) \le \sum_{0\le l<m\le 2^{n-1}} 2\exp\left(-\theta'\left(m-l\right)^{\beta-1/2}\Delta_n^{2\alpha-1} + 1\right) \le 2^{2n}\exp\left(-\theta' 2^{n(1-2\alpha)}\right).$$

Since $2\alpha < 1$ we clearly have that

$$\sum_{n=1}^{\infty} P\left(\mathcal{C}_n\right) < \infty$$

and by Borel-Cantelli's lemma we conclude that $P(\mathcal{C}_n$ infinitely often$) = 0$ which in turns yields the existence of such N_2. □

The proof of Corollary 4.2.6 follows directly from Lemma 4.2.4 and Lemma 4.2.5.

Proof. [Proof of Corollary 4.2.6] Using Lemma 4.2.4 we obtain that

$$R^n_{i,j}(t^n_l, t^n_m) = \sum_{k=l+1}^{m} \sum_{h=n+1}^{\infty} (L^h_{i,j}(2^{h-n}(k+1)) - L^h_{i,j}(2^{h-n}k)). \tag{4.31}$$

On the other hand, due to Lemma 4.2.5 if $n \geq N_2$

$$\begin{aligned} &\sum_{k=l+1}^{m} \sum_{h=n+1}^{\infty} |L^h_{i,j}(2^{h-n}(k+1)) - L^h_{i,j}(2^{h-n}k)| \\ \leq &\sum_{k=l+1}^{m} \sum_{h=n+1}^{\infty} (2^{-n}(k+1) - 2^{-n}k)^{\beta} \Delta_h^{2\alpha-\beta} < \infty \end{aligned}$$

because $\beta < 2\alpha$. Thus (by Fubini's theorem) the order of the summations in (4.31) can be exchanged and we obtain the result. □

Finally, the last proof of the section.

Proof. [Proof of Lemma 4.2.7]We start by showing the bound on Γ_R. By the definition of N_2 and Γ_L, for any n

$$|L^n_{i,j}(m) - L^n_{i,j}(l)| \leq \Gamma_L (m-l)^{\beta} \Delta_n^{2\alpha}.$$

Consequently, for any $0 \leq l < m \leq 2^{n-1}$,

$$\begin{aligned} |R^n_{i,j}(t^n_l, t^n_m)| &\leq \sum_{h=n+1}^{\infty} \left| L^h_{i,j}(2^{h-n}m) - L^h_{i,j}(2^{h-n}l) \right| \\ &\leq \sum_{h=n+1}^{\infty} \Gamma_L (m-l)^{\beta} 2^{(h-n)\beta} \Delta_h^{2\alpha} = \Gamma_L (m-l)^{\beta} \Delta_n^{\beta} \sum_{h=n+1}^{\infty} \Delta_h^{2\alpha-\beta} \\ &= \Gamma_L (t^n_l - t^n_m)^{\beta} \Delta_n^{2\alpha-\beta} \frac{2^{-(2\alpha-\beta)}}{1 - 2^{-(2\alpha-\beta)}}. \end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}\Gamma_R &:= \max_{1\le i,j\le d'} \sup_{n\ge 0} \sup_{0\le s<t\le 1, s,t\in D_n} \frac{|R^n_{i,j}(s,t)|}{|t-s|^\beta \Delta_n^{2\alpha-\beta}} \\ &\le \Gamma_L \frac{2^{-(2\alpha-\beta)}}{1-2^{-(2\alpha-\beta)}}.\end{aligned}$$

Let $r = \min\{h : |t^n_m - t^n_l| \ge \Delta_h\}$. For simplicity of notation, we define the following sequence of operators of time:

$$\underline{s}^h(t^n_l) = \min\{t^h_k : t^h_k \ge t^n_l\}$$

$$\bar{s}^h(t^n_m) = \max\{t^h_k : t^h_k \le t^n_m\}$$

for $r \le h \le n$.

Then

$$\begin{aligned}&|A_{i,j}(t^n_l, t^n_m)| \\ \le&|A_{i,j}(t^n_l, \underline{s}^{n-1}(t^n_l))| + |A_{i,j}(\underline{s}^{n-1}(t^n_l), \bar{s}^{n-1}(t^n_m))| + |A_{i,j}(\bar{s}^{n-1}(t^n_m), t^n_m)| \\ &+ |Z_i(\underline{s}^{n-1}(t^n_l)) - Z_i(t^n_l)||Z_j(\bar{s}^{n-1}(t^n_m)) - Z_j(\underline{s}^{n-1}(t^n_l))| \\ &+ |Z_i(\bar{s}^{n-1}(t^n_m)) - Z_i(t^n_l)||Z_j(t^n_m) - Z_j(\bar{s}^{n-1}(t^n_m))|\end{aligned}$$

By iterating the above procedure up to level r, we have

$$\begin{aligned}&|A_{i,j}(t^n_l, t^n_m)| \\ \le& \sum_{h=r+1}^{n} |A_{i,j}(\underline{s}^h(t^n_l), \underline{s}^{h-1}(t^n_l))| + A_{i,j}(\underline{s}^r(t^n_l), \bar{s}^r(t^n_m)) + \sum_{h=r+1}^{n} |A_{i,j}(\bar{s}^h(t^n_m), \bar{s}^{h-1}(t^n_m))| \\ &+ \sum_{h=r+1}^{n} |Z_i(\underline{s}^h(t^n_l)) - Z_i(\underline{s}^{h-1}(t^n_l)||Z_j(\bar{s}^{h-1}(t^n_m)) - Z_j(\underline{s}^{h-1}(t^n_l))| \\ &+ \sum_{h=r+1}^{n} |Z_i(\bar{s}^{h-1}(t^n_m)) - Z_i(\underline{s}^h(t^n_l))||Z_j(\bar{s}^h(t^n_m)) - Z_j(\bar{s}^{h-1}(t^n_m))|\end{aligned}$$

We make the following important observations,

$$\underline{s}^{h-1}(t^n_l) - \underline{s}^h(t^n_l) = \begin{cases} 0 & \text{if } \underline{s}^{h-1}(t^n_l) = \underline{s}^h(t^n_l) \\ \Delta_h & \text{otherwise} \end{cases}$$

$$\bar{s}^h(t_m^n) - \bar{s}^{h-1}(t_m^n) = \begin{cases} 0 & \text{if } s^{h-1}(t_m^n) = \bar{s}^h(t_m^n) \\ \Delta_h & \text{otherwise} \end{cases}$$

$$\bar{s}^r(t_m^n) - \underline{s}^r(t_l^n) = \begin{cases} 0 & \text{if } \underline{s}^r(t_l^n) = \bar{s}^r(t_m^n) \\ \Delta_r & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} &\frac{|A_{i,j}(t_l^n, t_m^n)|}{(t_m^n - t_l^n)^{2\alpha}} \\ \leq &\sum_{h=r+1}^{n} \Gamma_R \frac{\Delta_h^{2\alpha}}{\Delta_r^{2\alpha}} + \Gamma_R + \sum_{h=r+1}^{n} \Gamma_R \frac{\Delta_h^{2\alpha}}{\Delta_r^{2\alpha}} + \sum_{h=r+1}^{n} ||Z||_\alpha^2 \frac{\Delta_h^{\alpha}}{\Delta_r^{\alpha}} + \sum_{h=r+1}^{n} ||Z||_\alpha^2 \frac{\Delta_h^{\alpha}}{\Delta_r^{\alpha}} \\ \leq &\Gamma_R \frac{2}{1 - 2^{-2\alpha}} + ||Z||_\alpha^2 \frac{2^{1-\alpha}}{1 - 2^{-\alpha}}. \end{aligned}$$

Therefore,

$$\begin{aligned} ||A||_{2\alpha} &:= \max_{1 \leq i \leq j \leq d'} \sup_{n \geq 1} \sup_{0 \leq s < t \leq 1; s,t \in D_n} \frac{|A_{i,j}(s)|}{|t - s|^{2\alpha}} \\ &\leq \Gamma_R \frac{2}{1 - 2^{-2\alpha}} + ||Z||_\alpha^2 \frac{2^{1-\alpha}}{1 - 2^{-\alpha}}. \end{aligned}$$

□

4.4.2 Proof of Technical Results in Section 4.2.4

4.4.2.1 Proof of Technical Results in Section 4.2.4.1

We first provide the following auxiliary result which summarizes basic computations of moment generating functions of quadratic forms of bivariate Gaussian random variables.

Lemma 4.4.1 *Suppose that Y and Z are i.i.d. $N(0,1)$ random variables, then for any numbers $a_1, a_2, b, c_1, c_2 \in R$ define*

$$\phi(a, b, c) := E \exp\left(a_1 Y + a_2 Z + bYZ + c_1 Y^2 + c_2 Z^2\right),$$

then we have that if $|2c_i| < 1$ *for* $i = 1, 2$, *and* $|b| < (1 - 2c_1)(1 - 2c_2)$

$$\begin{aligned}\phi(a,b,c) =& (1-2c_1)^{-1/2}(1-2c_2)^{-1/2}\left(1-(b(1-2c_1)^{-1/2}(1-2c_2)^{-1/2})^2\right)^{-1/2}\\ &\times\exp\left(\frac{a_1^2(1-2c_1)^{-1}+a_2^2(1-2c_2)^{-1}+2a_1a_2b(1-2c_1)^{-1}(1-2c_2)^{-1}}{2(1-b^2(1-2c_1)^{-1}(1-2c_2)^{-1})}\right)\end{aligned}$$

Moreover, if we let

$$P'(Y\in dy, Z\in dz) = P(Y\in dy, Z\in dz)\frac{\exp(a_1y+a_2z+byz+c_1y^2+c_2z^2)}{\phi(\theta;a,b,c)},$$

then under $P'(\cdot)$ *we have that* (Y, Z) *are distributed bivariate Gaussian with covariance matrix*

$$\begin{aligned}&\Sigma(a,b,c)\\ =&\frac{1}{1-b^2(1-2c_1)^{-1}(1-2c_2)^{-1}}\\ &\times\begin{pmatrix}(1-2c_1)^{-1} & b(1-2c_1)^{-1}(1-2c_2)^{-1}\\ b(1-2c_1)^{-1}(1-2c_2)^{-1} & (1-2c_2)^{-1}\end{pmatrix},\end{aligned}$$

and mean vector

$$\mu(a,b,c) = \Sigma(a,b,c)\begin{pmatrix}a_1\\ a_2\end{pmatrix}.$$

Proof. First it follows easily that $E\exp(c_1Y^2+c_2Z^2) = (1-2c_1)^{-1/2}(1-2c_2)^{-1/2}$, and under the probability measure

$$P_1(Y\in dy.Z\in dz) = \frac{\exp(c_1y^2+c_2z^2)}{E\exp(c_1Y^2+c_2Z^2)}P(Y\in dy)P(Z\in dz)$$

Y and Z are independent with distributions $N(1,(1-2c_1)^{-1})$ and $N(1,(1-2c_2)^{-1})$, respectively. Therefore,

$$\begin{aligned}\phi(a,b,c) =&(1-2c_1)^{-1/2}(1-2c_2)^{-1/2}E_1\exp(a_1Y+a_2Z+bYZ)\\ =&(1-2c_1)^{-1/2}(1-2c_2)^{-1/2}\\ &\times E\exp\left(a_1Y(1-2c_1)^{-1/2}+a_2Z(1-2c_2)^{-1/2}\right.\\ &\left.+b(1-2c_1)^{-1/2}(1-2c_2)^{-1/2}YZ\right).\end{aligned}$$

Now, given $|\theta| < 1$ define $P_2(\cdot)$ via

$$P_2(Y \in dy, Z \in dz) = \frac{P(Y \in dy, Z \in dz)\exp(\chi yz)}{E\exp(\chi YZ)}.$$

Observe that

$$P(Y \in dy, Z \in dz)\exp(\chi yz) = \frac{1}{2\pi}\exp\left(-y^2/2 - z^2/2 + \chi yz\right)$$

and

$$-y^2/2 - z^2/2 + \chi yz = -(y,z)\Sigma^{-1}\begin{pmatrix} y \\ z \end{pmatrix}/2,$$

where

$$\Sigma^{-1} = \begin{pmatrix} 1 & -\chi \\ -\chi & 1 \end{pmatrix},$$

and thus

$$\Sigma = \frac{1}{1-\chi^2}\begin{pmatrix} 1 & \chi \\ \chi & 1 \end{pmatrix}.$$

Therefore, under $P_2(\cdot)$, (Y, Z) is distributed bivariate normal with mean zero and covariance matrix Σ, with

$$\chi = b(1-2c_1)^{-1/2}(1-2c_2)^{-1/2}$$

and we also must have that if $|\chi| < 1$,

$$E\exp(\phi YZ) = \left(1-\chi^2\right)^{-1/2} = \left(1 - (b(1-2c_1)^{-1/2}(1-2c_2)^{-1/2})^2\right)^{-1/2}.$$

Consequently, we conclude that

$$\begin{aligned}\phi(a,b,c) = {} & (1-2c_1)^{-1/2}(1-2c_2)^{-1/2}\left(1 - (b(1-2c_1)^{-1/2}(1-2c_2)^{-1/2})^2\right)^{-1/2} \\ & \times E_2\exp(a_1Y(1-2c_1)^{-1/2} + a_2Z(1-2c_2)^{-1/2}).\end{aligned}$$

The final expression for $\phi(a,b,c)$ is obtained from the fact that

$$\begin{aligned}& E_2\exp(a_1Y(1-2c_1)^{-1/2} + a_2Z(1-2c_2)^{-1/2}) \\ & = \exp\left(Var_2(a_1Y(1-2c_1)^{-1/2} + a_2Z(1-2c_2)^{-1/2})/2\right).\end{aligned}$$

And $P'(\cdot)$ is equivalent to a standard exponentially tilting to the measure $P_2(\cdot)$ using as the natural parameter the vector

$$\left(a_1(1-2c_1)^{-1/2}, a_2(1-2c_2)^{-1/2}\right),$$

and thus under $P'(\cdot)$ the covariance matrix is the same as under $P_2(\cdot)$ and the mean vector is equal to $\mu(a,b,c)$. □

We now are ready to provide the proof of Corollary 4.2.8.

Proof. [Proof of Corollary 4.2.8] Let us examine a term of the form $\Lambda_i^{n+m}\left(t_{2r-1}^{n+m}\right)\Lambda_j\left(t_{2r}^{n+m}\right)$,

$$\begin{aligned}
&\Lambda_i^{n+m}\left(t_{2r-1}^{n+m}\right)\Lambda_j\left(t_{2r}^{n+m}\right)\\
&=(\Lambda_i^{n+m-1}(t_r^{n+m-1})/2+\Delta_{n+m}^{1/2}W_{i,r}^{n+m})(\Lambda_j^{n+m-1}(t_r^{n+m-1})/2-\Delta_{n+m}^{1/2}W_{j,r}^{n+m})\\
&=\Lambda_i^{n+m-1}(t_r^{n+m-1})\Lambda_j^{n+m-1}(t_r^{n+m-1})/4-\Delta_{n+m}W_{i,r}^{n+m}W_{j,r}^{n+m}\\
&+\Delta_{n+m}^{1/2}W_{i,r}^{n+m}\Lambda_j^{n+m-1}(t_r^{n+m-1})/2-\Delta_{n+m}^{1/2}W_{j,r}^{n+m}\Lambda_i^{n+m-1}(t_r^{n+m-1})/2.
\end{aligned}$$

Then, we have that Corollary 4.2.8 follows immediately from Lemma 4.4.1. □

Finally, we provide the proof of Corollary 4.2.9.

Proof. [Proof of Corollary 4.2.9] Recall that for each $r\in\{1,2,...,2^n\}$,

$$\Lambda_i^n(t_r^n):=(Z_i(t_r^n)-Z_i(t_{r-1}^n)).$$

So

$$\begin{aligned}
\Lambda_i^n(t_{2r-1}^n)&=\Lambda_i^n(t_r^{n-1})/2+\Delta_n^{1/2}W_{i,r}^n,\\
\Lambda_i^n(t_{2r}^n)&=\Lambda_i^n(t_r^{n-1})/2-\Delta_n^{1/2}W_{i,r}^n.
\end{aligned}$$

We perform the first iteration in full detail, the rest are immediate just adjusting the notation. From Corollary 4.2.8 we obtain that

$$\begin{aligned}
&E_{n+m-1}\exp\left(\theta_0[L_{i,j}^{n+m}(k')-L_{i,j}^{n+m}(k)]\right)\\
=&\exp(\frac{1}{2}\sum_{r=k+1}^{k'}\frac{\theta_0^2\Delta_{n+m}}{4\left(1-\theta_0^2\Delta_{n+m}^2\right)}\Lambda_i\left(t_r^{n+m-1}\right)^2+\frac{1}{2}\sum_{r=k+1}^{k'}\frac{\theta_0^2\Delta_{n+m}}{4\left(1-\theta_0^2\Delta_{n+m}^2\right)}\Lambda_j\left(t_r^{n+m-1}\right)^2)\\
&\times\exp(\sum_{r=k+1}^{k'}\frac{\theta_0\Delta_{n+m}}{4\left(1-\theta_0^2\Delta_{n+m}^2\right)}\Lambda_i\left(t_r^{n+m-1}\right)\Lambda_j\left(t_r^{n+m-1}\right))\times\left(1-\theta_0^2\Delta_{n+m}^2\right)^{-(k'-k)/2}.
\end{aligned}$$

Using the definitions in (4.16) we have that

$$\frac{1}{2}\sum_{r=k+1}^{k'}\frac{\theta_0^2\Delta_{n+m}}{4\left(1-\theta_0^2\Delta_{n+m}^2\right)}\Lambda_i\left(t_r^{n+m-1}\right)^2+\frac{1}{2}\sum_{r=k+1}^{k'}\frac{\theta_0^2\Delta_{n+m}}{4\left(1-\theta_0^2\Delta_{n+m}^2\right)}\Lambda_j\left(t_r^{n+m-1}\right)^2$$
$$+\sum_{r=k+1}^{k'}\frac{\theta_0\Delta_{n+m}}{4\left(1-\theta_0^2\Delta_{n+m}^2\right)}\Lambda_i\left(t_r^{n+m-1}\right)\Lambda_j\left(t_r^{n+m-1}\right)$$

is equal to

$$\sum_{r=1}^{2^{n+m-2}}\left(\eta_1\left(t_{2r-1}^{n+m-1}\right)\Lambda_i\left(t_{2r-1}^{n+m-1}\right)^2+\eta_1\left(t_{2r}^{n+m-1}\right)\Lambda_i\left(t_{2r}^{n+m-1}\right)^2\right)$$
$$+\sum_{r=1}^{2^{n+m-2}}\left(\eta_1\left(t_{2r-1}^{n+m-1}\right)\Lambda_j\left(t_{2r-1}^{n+m-1}\right)^2+\eta_1\left(t_{2r}^{n+m-1}\right)\Lambda_j\left(t_{2r}^{n+m-1}\right)^2\right)$$
$$+\sum_{r=1}^{2^{n+m-2}}\left(\theta_1\left(t_{2r-1}^{n+m-1}\right)\Lambda_i\left(t_{2r-1}^{n+m-1}\right)\Lambda_j\left(t_{2r-1}^{n+m-1}\right)+\theta_1\left(t_{2r}^{n+m-1}\right)\Lambda_i\left(t_{2r}^{n+m-1}\right)\Lambda_j\left(t_{2r}^{n+m-1}\right)\right).$$

We now expand each of the terms; to simplify the notation write

$$x=W_{i,r}^{n+m-1}\quad\text{and}\quad y=W_{j,r}^{n+m-1}.$$

Define $\sqrt{\Delta}=\Delta_{n+m-1}^{1/2}$, put $u=\Lambda_i\left(t_r^{n+m-2}\right)$ and $v=\Lambda_j\left(t_r^{n+m-2}\right)$

$$\Lambda_i\left(t_{2r-1}^{n+m-1}\right)=u/2+\sqrt{\Delta}x,\quad \Lambda_i\left(t_{2r}^{n+m-1}\right)=u/2-\sqrt{\Delta}x,$$
$$\Lambda_j\left(t_{2r}^{n+m-1}\right)=v/2+\sqrt{\Delta}y,\quad \Lambda_j\left(t_{2r}^{n+m-1}\right)=v/2-\sqrt{\Delta}y.$$

Now, for brevity let us write $\eta_o=\eta_1\left(t_{2r-1}^{n+m-1}\right)$ and $\eta_e=\eta_1\left(t_{2r}^{n+m-1}\right)$ ('o' is used for odd, and 'e' for even)

$$\begin{aligned}&\Big(\eta_1\left(t_{2r-1}^{n+m-1}\right)\Lambda_i\left(t_{2r-1}^{n+m-1}\right)^2+\eta_1\left(t_{2r}^{n+m-1}\right)\Lambda_i\left(t_{2r}^{n+m-1}\right)^2\\&\quad+\eta_1\left(t_{2r-1}^{n+m-1}\right)\Lambda_j\left(t_{2r-1}^{n+m-1}\right)^2+\eta_1\left(t_{2r}^{n+m-1}\right)\Lambda_j\left(t_{2r}^{n+m-1}\right)^2\Big)\\&=\left(\eta_o\left(u/2+\sqrt{\Delta}x\right)^2+\eta_e\left(u/2-\sqrt{\Delta}x\right)^2+\eta_o\left(v/2+\sqrt{\Delta}y\right)^2+\eta_e\left(v/2-\sqrt{\Delta}y\right)^2\right)\\&=\frac{1}{4}u^2(\eta_e+\eta_o)+\frac{1}{4}v^2(\eta_e+\eta_o)+u(\eta_o-\eta_e)\sqrt{\Delta}x+v(\eta_o-\eta_e)\sqrt{\Delta}y\\&\quad+(\eta_e+\Delta\eta_o)\Delta x^2+(\eta_e+\eta_o)\Delta y^2.\end{aligned}$$

Likewise, put $\theta_o = \theta_1\left(t_{2r-1}^{n+m-1}\right)$ and $\theta_e = \theta_1\left(t_{2r}^{n+m-1}\right)$

$$\begin{aligned}
&\left(\theta_1\left(t_{2r-1}^{n+m-1}\right)\Lambda_i\left(t_{2r-1}^{n+m-1}\right)\Lambda_j\left(t_{2r-1}^{n+m-1}\right) + \theta_1\left(t_{2r}^{n+m-1}\right)\Lambda_i\left(t_{2r}^{n+m-1}\right)\Lambda_j\left(t_{2r}^{n+m-1}\right)\right)\\
=&\theta_o\left(u/2+\sqrt{\Delta}x\right)\left(v/2+\sqrt{\Delta}y\right) + \theta_e\left(u/2-\sqrt{\Delta}x\right)\left(v/2-\sqrt{\Delta}y\right)\\
=&\frac{1}{4}uv(\theta_e+\theta_o) + (\theta_e+\theta_o)\Delta xy + \frac{1}{2}v(\theta_o-\theta_e)\sqrt{\Delta}x + \frac{1}{2}u(\theta_o-\theta_e)\sqrt{\Delta}y
\end{aligned}$$

We then collect the terms free of x and y and obtain

$$\frac{u^2}{4}(\eta_e+\eta_o) + \frac{v^2}{4}(\eta_e+\eta_o) + \frac{uv}{4}(\theta_e+\theta_o).$$

Now the coefficients of x, y, x^2, y^2, and xy

$$\begin{aligned}
&\{u(\eta_o-\eta_e) + \frac{1}{2}v(\theta_o-\theta_e)\}\sqrt{\Delta}x + \{v(\eta_o-\eta_e) + \frac{1}{2}u(\theta_o-\theta_e)\}\sqrt{\Delta}y\\
&+ (\eta_e+\eta_o)\Delta x^2 + (\eta_e+\eta_o)\Delta y^2\\
&+ (\theta_e+\theta_o)\Delta xy.
\end{aligned}$$

And finally we can apply Lemma 4.4.1 to get the corresponding results. □

4.4.2.2 Proof of Technical Results in Section 4.2.4.2

We now provide the proof of Lemma 4.2.10.

Proof. [Proof of Lemma 4.2.10] Recalling expression (4.17), we establish the bound for

$$E_n \exp\left(\theta_0\{L_{i,j}^{n+1}(k') - L_{i,j}^{n+m}(k)\}\right)$$

by controlling the contribution of the term

$$\prod_{l=2}^{m}\prod_{r=1}^{2^{n+m-l}} C\left(t_r^{n+m-l}\right). \tag{4.32}$$

and the exponential term

$$\exp\left(\sum_{r=1}^{2^n}\theta_m(t_r^n)\Lambda_i(t_r^n)\Lambda_j(t_r^n) + \sum_{r=1}^{2^n}\eta_m(t_r^n)\left(\Lambda_i(t_r^n)^2 + \Lambda_j(t_r^n)^2\right)\right) \tag{4.33}$$

separately.

We start by analyzing θ_l and η_l. From Corollary 4.2.8, we have

$$\theta_1 = \frac{\theta_0}{4\,(1-\theta_0^2\Delta_{n+m}^2)} \text{ and } \eta_1 = \frac{\theta_0^2\Delta_{n+m}}{8\,(1-\theta_0^2\Delta_{n+m}^2)}.$$

We notice that $2\eta_1 \leq \theta_1^2\Delta_{n+m} \leq (5/2)\eta_1$.

Let

$$u = \max\{h : k' - k > 2^h\}.$$

We also denote

$$\underline{b} := \min\{r : \theta_l(t_r^{n+m-l}) > 0\}$$

and

$$\bar{b} := \max\{r : \theta_l(t_r^{n+m-l}) > 0\}.$$

The strategy throughout the rest of the proof proceeds as follows. We have that the $\theta_l(t_r^{n+m-l})$'s and $\eta_l(t_r^{n+m-l})$'s, $r = 1, 2, \ldots, 2^{n+m-l}$, are nonnegative. We also have that for $l \leq u \wedge m$, the number of positive $\theta_l(t_r^{n+m-l})$'s and $\eta_l(t_r^{n+m-l})$'s reduces by about a half at each step l and also the actual value of the positive $\theta_l(t_r^{n+m-l})$'s and $\eta_l(t_r^{n+m-l})$'s shrinks by at least 1/2. We will establish that if $m > u$, for $u < l \leq m$, there are at most two positive $\theta_l(t_r^{n+m-l})$'s and two positive $\eta_l(t_r^{n+m-l})$'s and at each step l, their values shrink by more than $2^{-3/2}$. Using these observations we will establish some facts and then use them to estimate (4.32) and finally (4.33). We now proceed to carry out this strategy.

We first verify the following claims.

Claim 1: For $l \leq u$, we claim that $\theta_l(t_r^{n+m-l}), \eta_l(t_r^{n+m-l}) \geq 0$ for all $r = 1, 2, \ldots, 2^{n+m-l}$ and $\theta_l(t_r^{n+m-l})$'s are equal for $r \in (\underline{b}, \bar{b})$ and we denote their values as θ_l. So, following the recursion in (4.16) we have that $\theta_l = \Delta_{l-1}\theta_1$. If $\theta_l(t_{\underline{b}}^{n+m-l}) \neq \theta_l(t_{\underline{b}+1}^{n+m-l})$, then $\theta_l(t_{\underline{b}}^{n+m-l}) < \theta_l(t_{\underline{b}+1}^{n+m-l}) = \theta_l$, and if $\theta_l(t_{\bar{b}}^{n+m-l}) \neq \theta_l(t_{\bar{b}-1}^{n+m-l})$, then $\theta_l(t_{\bar{b}}^{n+m-l}) <$

$\theta_l(t_{\bar{b}-1}^{n+m-l}) = \theta_l$.

Likewise, $\eta_l(t_r^{n+m-l})$'s are equal for $r \in (\underline{b}, \bar{b})$; we denote their common values as η_l and we have from (4.16) that $\eta_l = \Delta_{l-1}\eta_1$. If $\eta_l(t_{\underline{b}}^{n+m-l}) \neq \eta_l(t_{\underline{b}+1}^{n+m-l})$, then $\eta_l(t_{\underline{b}}^{n+m-l}) < \eta_l(t_{\underline{b}+1}^{n+m-l})$, and if $\eta_l(t_{\bar{b}}^{n+m-l}) \neq \eta_l(t_{\bar{b}-1}^{n+m-l})$, then $\eta_l(t_{\bar{b}}^{n+m-l}) < \eta_l(t_{\bar{b}-1}^{n+m-l})$. In other words, at each step, l for $l < u$, $\theta_l(t_r^{n+m-l})$ and $\eta_l(t_r^{n+m-l})$ decay at rate $1/2$ if it is not at the boundary ($r \in (\underline{b}, \bar{b})$), and the boundary ones ($\theta_l(t_{\underline{b}}^{n+m-l})$, $\theta_l(t_{\bar{b}}^{n+m-l})$ and $\eta_l(t_{\underline{b}}^{n+m-l})$, $\eta_l(t_{\bar{b}}^{n+m-l})$), may decay at a faster rate.

We now prove the claim by induction using the recursive relation in (4.16). The claim is immediate for θ_1 and η_1. Now suppose it holds for $\theta_l(t_r^{n+m-l})$ and $\eta_l(t_r^{n+m-l})$, $r = 1, 2, \ldots, 2^{n+m-l}$. We next show that the claim holds for $\theta_{l+1}(t_r^{n+m-l-1})$, $r = 1, 2, \ldots, 2^{n+m-l-1}$, as well. We omit the proof of $\eta_{l+1}(t_r^{n+m-l-1})$ here, as it follows exactly the same line of analysis as $\theta_{l+1}(t_r^{n+m-l-1})$.

We next divide the analysis into five cases.

Case 1. $\theta_l\left(t_{2r-1}^{m+n-l}\right) = \theta_l\left(t_{2r}^{m+n-l}\right)$ and $\eta_l\left(t_{2r-1}^{m+n-l}\right) = \eta_l\left(t_{2r}^{m+n-l}\right)$. Then $\theta_+^{l+1}\left(t_r^{m+n-l}\right) = 2\theta_l\left(t_{2r-1}^{m+n-l+1}\right)$ and $\theta_-^{l+1}\left(t_r^{m+n-l}\right) = 0$. Likewise $\eta_+^{l+1}\left(t_r^{m+n-l}\right) = 2\eta_l\left(t_{2r-1}^{m+n-l+1}\right)$ and $\eta_-^{l+1}\left(t_r^{m+n-l}\right) = 0$. From (4.16), we have $\theta_l\left(t_r^{m+n-l-1}\right) = \theta_{l-1}\left(t_{2r-1}^{m+n-l+1}\right)/2$ and $\eta_l\left(t_r^{m+n-l-1}\right) = \eta_{l-1}\left(t_{2r-1}^{m+n-l+1}\right)/2$.

Case 2. $\theta_l\left(t_{2r-1}^{m+n-l}\right) = 0$, $\theta_l\left(t_{2r}^{m+n-l}\right) > 0$ and $\eta_l\left(t_{2r-1}^{m+n-l}\right) = 0$, $\eta_l\left(t_{2r}^{m+n-l}\right) > 0$. Then we know that $2r = \underline{b}$. We also have $\theta_+^{l+1}\left(t_r^{m+n-l-1}\right) = \theta_l\left(t_{2r}^{m+n-l}\right)$ and $\theta_-^{l+1}\left(t_r^{m+n-l-1}\right) = -\theta_l\left(t_{2r}^{m+n-l}\right)$. Likewise, $\eta_+^{l+1}\left(t_r^{m+n-l-1}\right) = \eta_l\left(t_{2r}^{m+n-l}\right)$ and $\eta_-^{l+1}\left(t_r^{m+n-l-1}\right) = -\eta_l\left(t_{2r}^{m+n-l}\right)$. We rewrite the expression for $\theta_{l+1}(t_r^{n+m-l-1})$ in

(4.16) as

$$\begin{aligned}\theta_{l+1}\left(t_r^{m+n-l-1}\right) =&\theta_+^{l+1}(t_r^{m+n-l-1})\frac{1}{4}+|\theta_-^{l+1}(t_r^{m+n-l-1})|\{h_{l+1}(t_r^{m+n-l-1})|\eta_-^{l+1}(t_r^{m+n-l-1})|\\ &+\frac{1}{4}h_{l+1}(t_r^{m+n-l-1})|\theta_-^{l+1}(t_r^{m+n-l-1})|\rho_{l+1}(t_r^{m+n-l-1})\\ &+h_{l+1}(t_r^{m+n-l-1})\eta_-^{l+1}(t_r^{m+n-l-1})^2\frac{\rho_{l+1}(t_r^{m+n-l-1})}{|\theta_-^{l+1}(t_r^{m+n-l-1})|}\}\\ =&\theta_l(t_{2r}^{m+n-l})\{\frac{1}{4}+h_{l+1}(t_r^{m+n-l-1})\eta^l(t_{2r}^{m+n-l})\\ &+\frac{1}{4}h_{l+1}(t_r^{m+n-l-1})\theta_l(t_{2r}^{m+n-l})\rho_{l+1}(t_r^{m+n-l-1})\\ &+h_{l+1}(t_r^{m+n-l-1})\eta_l(t_{2r}^{m+n-l})^2\frac{\rho_{l+1}(t_r^{m+n-l-1})}{\theta_l(t_{2r}^{m+n-l})}\}\end{aligned}$$

As

$$\theta_l\Delta_{n+m-l}\le\theta_1\Delta_{n+m-1}\le\frac{1}{4}$$

and

$$\eta_l\Delta_{n+m-l}\le\eta_1\Delta_{n+m-1}\le\frac{1}{48},$$

it is then easy to check that

$$\frac{1}{4}\theta_l\left(t_{2r}^{m+n-l}\right)<\theta_{l+1}\left(t_r^{m+n-l-1}\right)<\frac{3}{10}\theta_l\le\frac{3}{5}\theta_{l+1}.$$

Case 3. $\theta_l\left(t_{2r-1}^{m+n-l}\right)>0$, $\theta_l\left(t_{2r}^{m+n-l}\right)=0$ and $\eta_l\left(t_{2r-1}^{m+n-l}\right)>0$, $\eta_l\left(t_{2r}^{m+n-l}\right)=0$. Then we know that $2r-1=\bar{b}$. Following the same line of analysis as in Case 2, we have

$$\frac{1}{4}\theta_l\left(t_{2r}^{m+n-l}\right)<\theta_{l+1}\left(t_r^{m+n-l-1}\right)<\frac{3}{10}\theta_l\le\frac{3}{5}\theta_{l+1}.$$

Case 4. $0<\theta_l\left(t_{2r-1}^{m+n-l}\right)<\theta_l\left(t_{2r}^{m+n-l}\right)$ and $0<\eta_l\left(t_{2r-1}^{m+n-l}\right)<\theta_l\left(t_{2r}^{m+n-l}\right)$. Then we know that $2r-1=\underline{b}$. There exist $\xi<1$, such that $\theta_l\left(t_{2r-1}^{m+n-l}\right)\le\xi\theta_l\left(t_{2r}^{m+n-l}\right)=$

$\xi\Delta_{l-1}\theta_1$ and $\eta_l\left(t_{2r-1}^{m+n-l}\right) \le \xi\eta_l\left(t_{2r}^{m+n-l}\right) = \xi\Delta_{l-1}\eta_1$. From (4.16), we have

$$\begin{aligned}\theta_{l+1}\left(t_r^{m+n-l-1}\right) \le & \theta_+^{l+1}(t_r^{m+n-l-1})\left\{\frac{1}{4} + h_{l+1}(t_r^{m+n-l-1})\eta_-^{l+1}(t_r^{m+n-l-1})^2\frac{\rho_{l+1}(t_r^{m+n-l-1})}{\theta_+^{l+1}(t_r^{m+n-l-1})}\right\} \\ & + |\theta_-^{l+1}(t_r^{m+n-l-1})|\{h_{l+1}(t_r^{m+n-l-1})|\eta_-^{l+1}(t_r^{m+n-l-1})| \\ & + \frac{1}{4}h_{l+1}(t_r^{m+n-l-1})|\theta_-^{l+1}(t_r^{m+n-l-1})|\rho_{l+1}(t_r^{m+n-l-1})\}.\end{aligned}$$

As $|\theta_-^{l+1}(t_r^{m+n-l-1})| \le \theta_l$ and $|\eta_-^{l+1}(t_r^{m+n-l-1})| \le \eta_l$, it is easy to check that

$$\theta_{l+1}\left(t_r^{m+n-l-1}\right) < \theta_+^{l+1}(t_r^{m+n-l-1})\left(\frac{1}{4} + 0.01\right) + |\theta_-^{l+1}(t_r^{m+n-l-1})| \times 0.05.$$

Since $\theta_{l+1}\left(t_r^{m+n-l-1}\right) + |\theta_-^{l+1}(t_r^{m+n-l-1})| = \theta_l$, we have

$$\begin{aligned}\theta_{l+1}\left(t_r^{m+n-l-1}\right) &< \theta_l\left(\left(\frac{1}{4} + 0.01 - 0.05\right)(1+\xi) + 0.05\right) \\ &= \frac{1}{2}\theta_l\left(\frac{1}{2} + 0.02 + 0.42\xi\right) < \frac{\theta_l}{2} = \theta_{l+1}.\end{aligned}$$

Case 5. $\theta_l\left(t_{2r-1}^{m+n-l}\right) > \theta_l\left(t_{2r}^{m+n-l}\right) > 0$ and $0 < \eta_l\left(t_{2r-1}^{m+n-l}\right) > \theta_l\left(t_{2r}^{m+n-l}\right) > 0$. Then we know that $2r = \bar{b}$. Following the same line of analysis as in Case 4, we have

$$\theta_{l+1}\left(t_r^{m+n-l-1}\right) < \theta_{l+1}.$$

We thus prove that the claim holds for $\theta_{l+1}(t_r^{m+n-l-1})$, $r = 1, 2, \ldots, 2^{n+m-l-1}$, as well.

We have established Claim 1. We now continue with a second claim.

Claim 2: For $u < l < m$, we have at most two positive $\theta_l(t_r^{m+n-l})$'s, namely $\theta_l(t_{\underline{b}}^{m+n-l})$ and $\theta_l(t_{\bar{b}}^{m+n-l})$. Notice that it is possible that $\underline{b} = \bar{b}$. We then claim that if $\underline{b} \neq \bar{b}$,$\theta_l(t_{\underline{b}}^{m+n-l}) \le \Delta_{l-1}\theta_1 2^{-(l-u-1)/2}$ and $\theta_l(t_{\bar{b}}^{m+n-l}) \le \Delta_{l-1}\theta_1 2^{-(l-u-1)/2}$. Similarly $\eta_l(t_{\underline{b}}^{m+n-l}) \le \Delta_{l-1}\eta_1 2^{-(l-u-1)/2}$ and $\eta_l(t_{\bar{b}}^{m+n-l}) \le \Delta_{l-1}\eta_1 2^{-(l-u-1)/2}$. If $\underline{b} = \bar{b}$, $\theta_l(t_{\underline{b}}^{m+n-l}) \le \Delta_{l-1}\theta_1 2^{-(l-u-2)/2}$, $\theta_l(t_{\bar{b}}^{m+n-l}) \le \Delta_{l-1}\theta_1 2^{-(l-u-2)/2}$ and $\eta_l(t_{\underline{b}}^{m+n-l}) \le$

$\Delta_{l-1}\eta_1 2^{-(l-u-2)/2}$, $\eta_l(t_{\bar{b}}^{m+n-l}) \leq \Delta_{l-1}\eta_1 2^{-(l-u-2)/2}$.

We prove the claim by induction. We shall give the proof of $\theta_l(t_r^{m+n-l})$ only, as the proof of $\eta_l(t_r^{m+n-l})$ follows exactly the same line of analysis. For $l = u$, we have the following cases.

i) $\bar{b} = \underline{b} + 2$, $\underline{b}$ is odd. In this case, $\theta_{l+1}(t_{(\underline{b}+1)/2}^{m+n-l-1}) < \Delta_l\theta_1$, which follows from the analysis in Case 4 for $l \leq u$. And $\theta_{l+1}(t_{(\bar{b}+1)/2}^{m+n-l-1}) < (3/5)\Delta_l\theta_1$, following the analysis in Case 3 for $l \leq u$.

ii) $\bar{b} = \underline{b} + 2$, $\underline{b}$ is even. In this case, $\theta_{l+1}(t_{\underline{b}/2}^{m+n-l-1}) < (3/5)\Delta_l\theta_1$, which follows from the analysis in Case 2 for $l \leq u$. And $\theta_{l+1}(t_{\bar{b}/2}^{m+n-l-1}) < \Delta_l\theta_1$, following the analysis in Case 5, for $l \leq u$.

iii) $\bar{b} = \underline{b} + 1$, $\underline{b}$ is odd. In this case, let $\bar{\theta}_l = \max\{\theta_l(t_{\underline{b}}^{m+n-l}), \theta_l(t_{\bar{b}}^{m+n-l})\}$, Then following the same analysis as in Case 4 or Case 5 for $l \leq u$ (depending on which one of $\theta_l(t_{\underline{b}}^{m+n-l})$ and $\theta_l(t_{\bar{b}}^{m+n-l})$ is smaller), we have $\theta_{l+1}(t_{\bar{b}/2}^{m+n-l-1}) < \bar{\theta}_l/2 \leq \Delta_l\theta_1$.

iv) $\bar{b} = \underline{b} + 1$, $\underline{b}$ is even. In this case, $\theta_{l+1}(t_{\underline{b}/2}^{m+n-l-1}) < (3/5)\Delta_l\theta_1$, which follows from the analysis in Case 2 for $l \leq u$. And $\theta_{l+1}(t_{(\bar{b}+1)/2}^{m+n-l-1}) < (3/5)\Delta_l\theta_1$, following the analysis in Case 3 for $l \leq u$.

Therefore, the claim holds for $u+1$. Suppose the claim holds for $l \geq u+1$. Then when moving from level l to level $l+1$, one of the following three cases can happen.

a) $\bar{b} = \underline{b} + 1$ and $\underline{b}$ is even. In this case, following the analysis in Case 2 and Case 3 for $l \leq u$, we have

$$\theta_{l+1}(t_{\underline{b}/2}^{m+n-l-1}) \leq \frac{3}{10}\theta_l(t_{\underline{b}}^{m+n-l}) \leq \Delta_l\theta_1 2^{-(l-u)/2}$$

and

$$\theta_{l+1}(t^{m+n-l-1}_{(\bar{b}+1)/2}) \leq \frac{3}{10}\theta_l(t^{m+n-l}_{\bar{b}}) \leq \Delta_l\theta_1 2^{-(l-u)/2}.$$

b) $\bar{b} = \underline{b}$. In this case, following the analysis in Case 2 or Case 3 for $l \leq u$ (depending on whether $\underline{b}$ is odd or even), we have

$$\theta_{l+1}(t^{m+n-l-1}_{\lceil \underline{b}/2 \rceil}) \leq \frac{3}{10}\theta_l(t^{m+n-l}_{\underline{b}}) \leq \Delta_l\theta_1 2^{-(l-u-1)/2}.$$

c) $\bar{b} = \underline{b} + 1$ and $\underline{b}$ is odd. In this case, we let $\bar{\theta}_l = \max\{\theta_l(t^{m+n-l}_{\underline{b}}), \theta_l(t^{m+n-l}_{\bar{b}})\}$, Then we can use the same analysis as in Case 4 or Case 5 for $l \leq u$ (depending on which one of $\theta_l(t^{m+n-l}_{\underline{b}})$ and $\theta_l(t^{m+n-l}_{\bar{b}})$ is smaller) to conclude that

$$\theta_{l+1}(t^{m+n-l-1}_{\bar{b}/2}) < \frac{1}{2}\bar{\theta}_l \leq \Delta_l\theta_1 2^{-(l-u-1)/2}.$$

We notice that case c) can happen only once.

We are now ready to control the contribution of the term (4.32). As

$$\Delta_{n+m-l+1}\eta^l_+(t^{n+m-l}_r) \leq 1/30$$

and

$$\rho_l(t^{n+m-l}_r) < 1/7,$$

we have when $m \leq u$

$$\begin{aligned}
&\prod_{l=2}^{m}\prod_{r=1}^{2^{n+m-l}} C(t^{n+m-l}_r) \\
\leq &\prod_{l=2}^{m}\prod_{r=1}^{2^{n+m-l}} \exp\left(4\Delta_{n+m-l+1}\eta^l_+(t^{n+m-l}_r) + \rho_l(t^{n+m-l}_r)^2\right) \\
\leq &\prod_{l=2}^{m}\exp\left(\left(16\Delta_{n+m}\eta_1 + \frac{(4\Delta_{n+m}\theta_1)^2}{(1-8\Delta_{n+m}\eta_1)^2}\right)((k'-k)\Delta_l + 2)\right) \\
\leq &\prod_{l=2}^{m}\exp\left(\left(\frac{11}{5}\frac{\gamma^2}{k'-k}\Delta_n^{1-2\alpha'} + \frac{6}{5}\frac{\gamma^2}{k'-k}\Delta_n^{2-4\alpha'}\right)((k'-k)\Delta_l + 2)\right).
\end{aligned}$$

The last inequality follows from Corollary 2 that $\theta_1 = \theta_0/4(1-\theta_0^2\Delta^2_{n+m})$, $\eta_1 = \theta_0^2\Delta_{n+m}/2(1-\theta_0^2\Delta^2_{n+m})$, and our choice of $\theta_0 = \gamma/((k'^{1/2}\Delta_n^{2\alpha'}\Delta_m)$. Then, as $(k'-$

$k)^{-1} \leq 2^{-m}$,

$$\begin{aligned}
&\prod_{l=2}^{m}\prod_{r=1}^{2^{n+m-l}} C(t_r^{n+m-l}) \\
\leq &\exp\left(\frac{11}{5}\gamma^2\left(\sum_{l=2}^{m}\Delta_l + 2(m-1)\Delta_m\right) + \frac{6}{5}\gamma^2\left(\sum_{l=2}^{u}\Delta_l + 2(m-1)\Delta_m\right)\right) \\
\leq &\exp\left(\frac{8}{25}\right).
\end{aligned}$$

When $m > u$,

$$\begin{aligned}
&\prod_{l=2}^{m}\prod_{r=1}^{2^{n+m-l}} C(t_r^{n+m-l}) \\
\leq &\prod_{l=2}^{m}\prod_{r=1}^{2^{n+m-l}} \exp\left(4\Delta_{n+m-l+1}\eta_+^l(t_r^{n+m-l}) + \rho_l(t_r^{n+m-l})^2\right) \\
\leq &\prod_{l=2}^{u}\exp\left(\left(\frac{11}{5}\frac{\gamma^2}{k'-k}\Delta_n^{1-2\alpha'} + \frac{6}{5}\frac{\gamma^2}{k'-k}\Delta_n^{2-4\alpha'}\right)((k'-k)\Delta_l + 2)\right) \\
&\times\prod_{l=u+1}^{m}\exp\left(\frac{11}{5}\frac{\gamma^2}{k'-k}\Delta_n^{1-2\alpha'}\Delta_{l-u-2}^{1/2} + \frac{6}{5}\frac{\gamma^2}{k'-k}\Delta_n^{2-4\alpha'}\Delta_{l-u-2}\right).
\end{aligned}$$

As $(k'-k)^{-1} \geq 2^{-u}$,

$$\begin{aligned}
&\prod_{l=2}^{m}\prod_{r=1}^{2^{n+m-l}} C(t_r^{n+m-l}) \\
\leq &\exp\left(\frac{11}{5}\gamma^2\left(\sum_{l=2}^{u}\Delta_l + 2(u-1)\Delta_u + \sum_{l=u+1}^{m}\Delta_{l-2}^{1/2}\right)\right. \\
&\left.+\frac{6}{5}\gamma^2\left(\sum_{l=2}^{u}\Delta_l + 2(u-1)\Delta_u + \sum_{l=u+1}^{m}\Delta_{l-2}\right)\right) \\
\leq &\exp\left(\frac{1}{2}\right).
\end{aligned}$$

For (4.33), we notice that under condition (4.19) and (4.20), we have

$$\begin{aligned}
\left|\sum_{r=1}^{2^n}\theta_m(t_r^n)\Lambda_i(t_r^n)\Lambda_j(t_r^n)\right| &\leq \theta_1\Delta_{m-1}\varepsilon_0((k'-k)\Delta_m)^\beta\Delta_n^{2\alpha'} + 2\theta_1\Delta_{m-1}\Delta_n^{2\alpha'} \\
&\leq \varepsilon_0\gamma(k'-k)^{\beta-1/2} + 2\gamma,
\end{aligned}$$

and

$$\left|\sum_{r=1}^{2^n}\eta_m(t_r^n)\left(\Lambda_i(t_r^n)^2+\Lambda_j(t_r^n)^2\right)\right| \le ((k'-k)\Delta_m+2)\,\eta_1\Delta_{m-1}2\Delta_n^{2\alpha'}$$
$$\le 2\gamma^2.$$

Combining the analysis for (4.32) and (4.33), we have

$$\begin{aligned}&E_n\exp(\theta_0\{L_{i,j}^{n+m}(k')-L_{i,j}^{n+m}(k)\})\\ \le&\exp\left(\theta_0^2\Delta_{n+m}^2(k'-k)+\frac{1}{2}+\varepsilon_0\gamma(k'-k)^{\beta-1/2}+2\gamma+2\gamma^2\right)\\ \le&4\exp\left(\varepsilon_0\gamma(k'-k)^{\beta-1/2}\right).\end{aligned}$$

□

Part II

Load-Dependent Slowdown Services

Chapter 5

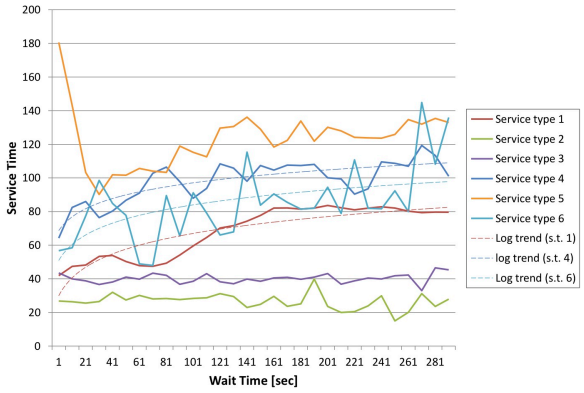
Introduction to Part II

A central assumption in the operations management literature is that service times are independent of the load of the system. However, empirical and anecdotal evidence suggest that in many service systems the two are correlated (see for example [44; 45; 46] and [47]). Depending on the service environment, heavily-loaded systems may experience service speedups or slowdowns. While speedup was theoretically investigated in [48], slowdown was so far been neglected.

Slowdown of service rate, when the system is congested, is a widely spread phenomenon, which is contributed to several psychological, physiological and technical reasons.High congestion levels may induce pressure on agents, which according to the psychology literature (see for example [49]) may impact human perception, information processing and decision making. All of these aspects may influence operational performance. While a relatively low level of arousal may increase productivity, high levels of pressure hurt performance [50]. High congestion levels may also require individuals to conduct multiple tasks in parallel which involves a cognitive switching cost [44]. At the same time, high congestion levels may lead staff to work longer hours without proper rest, causing fatigue. Empirical studies provide evidence that fatigue leads to deterioration in productivity (e.g. [46; 51]). Service rate may also deteriorate due to external capacity limitations, for exam-

ple, IT systems perform slower when heavily loaded, and hence, the service times of the workers who use them may increase [44]. On the customer side, it is well established that patients' condition may deteriorate if treatment is delayed in health care facilities, causing a service slowdown [52]. Customers may also demand a longer and more personalized service following a long wait. For example, agents might need to take some extra time to mollify irritated customers who experience long waits. In call centers, the service time could notably increase when the system is congested. This is illustrated by Figure 5.1, where the average service time of service type 1 doubles itself from 40 to 80 seconds [47]. Changes of service rate in such a manner have a significant influence on the companies' revenue.

Figure 5.1: Service time as a function of waiting time in a call center of an Israeli Bank (by service type)



Motivated by these empirical findings, we investigate how the dependence between service rate and workload affects the operational performance of the system measured by delay and abandonment, and how service providers can cope with the consequences of this dependence by adjusting staffing or routing.

Generally, there are two objectives that play opposing roles in the design of service systems. On one hand, to increase efficiency and reduce operational costs, system designers aim to increase resource utilization. On the other hand, high utilization leads to increased level of delay and abandonment, thereby reducing quality of service.

A common approach to design a service system is to balance the tradeoff between system performance, measured by the probability of waiting and the probability of abandonment experienced by customers, and resource utilization, measured by the fraction of time an agent or a resource is occupied. The Quality-and-Efficiency-Driven (QED) regime in a many-server asymptotic analysis suggests a Square-Root Staffing (SRS) rule to balance this tradeoff. According to the SRS rule the number of servers, s, is set such that $s = R + \beta\sqrt{R}$, where $R = \lambda/\mu$ is the offered load of the system, and β, the SRS parameter, is set to achieve certain performance measures. For the SRS rule in an exponential type multi-server queueing model with abandonment (commonly referred to as the *Erlang-A* model), β is determined using the Garnett functions [53]. Applying the SRS rule to the Erlang-A model implies that a significant proportion of customers (e.g. 30%–80%) gets served immediately upon arrival and the probability of abandonment is small (e.g. $< 5\%$) [53]. Other operating regimes considered in the literature include Efficiency-Driven (ED) regime and Quality-Driven (QD) regime, where the staffing level and the offered load grow in fixed proportion. ED staffing is used when the staffing cost is very high. In this case the staffing level is set to $s = R - \alpha R$ for $0 < \alpha < 1$, where α is typically selected in the range 0.1–0.25 [54]. This results in 100% occupancy, probability of waiting close to 1 and very high abandonment rate (5%–30%) [53]. A QD regime is used when the system requires a very high level of service quality. In this case, the staffing level is set to $s = R + \alpha R$ for $\alpha > 0$, where the typical range of α is as in the ED regime. This staffing level results in very low abandonment (almost 0) and negligible waiting, but also in an agent occupancy which is far below 100% [53].

In this part of the dissertation, we modify the Erlang-A model to account for the slowdown effect and analyze the performance of the modified Erlang-A model when staffing according the SRS rule. We use the term *load sensitivity* to describe the rate of service rate deterioration as a response to increased workload. We show that staffing to operate in the QED regime may not be a good enough solution in

some systems with load-sensitive service rates. Depending on the model parameters, we observe that systems designed to operate in the QED regime may have unstable performance, alternating between being overloaded and underloaded, or even end up being constantly heavily overloaded. This results in a very high probability of waiting (close to 1) and a significant proportion of customer abandonment (e.g. 10%–20%). Hence, a QED regime staffing rule, or even a QD regime staffing rule, may result in an undesirable performance, typically found when using ED regime staffing rules. We therefore propose alternative staffing rules and admission control policies that can be applied in the presence of service slowdowns.

5.1 Literature Review

Palm [55] introduced the Erlang-A ($M/M/s+M$) model to incorporate abandonment in the traditional Erlang-C ($M/M/s$) queue. [56] showed that abandonment is a significant factor in modeling service systems and making staffing decisions. [53] conducted a heavy traffic asymptotic analysis of the Erlang-A model in the QED regime. They derived approximations for the probability of waiting and abandonment and provided guidance for the design of large service systems. In this part, we study a modified Erlang-A model that accounts for a load-dependent service rates.

A few papers consider state-dependent service rates but most of them are in the single server queue setting without abandonment. [57] and [58] study the steady-state behavior of the delay process (waiting time distribution) of a $G/G/1$ queue, where both the service rate and the arrival rate depend linearly on the delay process. [59] derived the fluid and diffusion limits of a network of single server queues with state-dependent arrival rate, service rate and routing probability. [60] studies the fluid and diffusion approximation of $G/M/n + GI$ queues with state-dependent service rate, but under a different scaling on the effect of workload (queue length process) on the service rate function. The bi-stability phenomenon that does not arise in their

models.

The bi-stability phenomenon is studied in different contexts: ICU flows [48], communication networks [61], multi-class stochastic networks [62] and retrials [63]. The phenomenon is also studied in Statistical Physics (e.g. [64] [65]). The conjectured trajectory of the system under bi-stability is that it fluctuates within one stable region for a long time and then, due to some rare event it reaches the other stable region and remains there for a while [62]. In this paper, we study the bi-stability phenomenon through asymptotic analysis of the stationary distribution and sensitivity analysis of system parameters (§6.4). We impose exponential assumptions on the service time and patience time distributions for tractability reasons. More general service time and patience time distributions would require a different set of analysis. However, it is known from the statistical physics literature that a rigorous characterization of the dynamics of systems with bi-stability is in general very difficult to obtain unless we assume some very specific system structures (e.g. reversibility of the Markov process) [65].

In order to avoid bi-stability, we propose in §6.3 three policies based on adjusting staffing, abandonment, and arrival according to system state respectively. Indeed, the state-dependent abandonment rate has similar effect on the queue length process as the state-dependent service rate studied in this paper. Motivated by the way that delay announcements and observable queues change customer patience, there have also been works that study state-dependent abandonment rate [66]. They considered delay announcements as a control policy in the ED regime, while we show its potential advantages for stabilizing system performance in the QED regime. Our work is also different from [67] and [68] who analyzed how delay announcements affect system performance by changing the strategic behavior of customers. This was done by combining game theory with queueing models.

In terms of staffing and admission control policies, [69] studied the optimal admission control of an $M/G/1$ queue with service rate that is first increasing and then

decreasing as a function of the workload. Their objective is to optimize throughput and they show that under certain conditions a threshold policy is optimal. Likewise, in §6.4.3, we also consider a threshold admission control policy, but our objective is to maintain a certain performance level. Admission control in the QED regime has also been studied in [70], [71] and references therein. We also consider adjusting the staffing level as a possible solution to stabilize system performance. A few papers considered dynamic staffing (e.g., [72], [73]) to cope with time-varying arrivals. They allow the staffing level to change over time according to a predictable offered load function. In our model, the fluctuations in performance arise because of the bi-stability phenomenon. The system alternates between two equilibria in an *unpredictable* stochastic way. Therefore, we cannot propose a predetermined policy whereby the staffing levels change in a predictable fashion. Instead, we propose static policies that mitigate the effect of the unpredictable system behavior.

5.2 Main Contributions

We make the following key contributions:

1) We show that the effect of load sensitivity on system performance is nonlinear. Systems with low sensitivity may exhibit only a modest deterioration in performance, whereas when the sensitivity increases beyond a threshold, the performance deteriorates drastically. We prove that the threshold that separates the two cases is derived from the *relative* relation between the service rate sensitivity level around zero and the abandonment rate (§6.2.2).

2) When the *load sensitivity* is relatively low (i.e., the service rate does not decrease significantly with the load placed on the system), the SRS rule leads to a QED performance. However, for a fixed square-root staffing parameter, β, the performance deteriorates with the load sensitivity level. We develop new approximation functions in the presence of load sensitivity, which can be used

when making staffing decisions (§6.3). To derive these approximations, it is sufficient to accurately estimate the service rate function around zero.

3) When the *load sensitivity* is relatively high, the system alternates between two performance levels, a phenomenon we refer to as bi-stability: one provides a QED performance while the other has an ED performance (§6.2). Therefore, in such cases, applying the SRS rule does not consistently result in QED performance. We investigate how the system scale and other parameters influence the occurrence of bi-stability, and the proportion of time the system spends around each performance level (§6.4). We show that while a higher load sensitivity increases the occurrence of ED performance, a higher abandonment rate decreases such occurrences. We also show that large systems converge to the ED performance with an exponential rate; sensitivity increases the rate of convergence, and abandonment rate decreases this rate of convergence. Two interesting observations follow from our analysis. Firstly, the modified Erlang-A queue exhibits unusual dis-economies-of-scale effect. In particular, as the system scale grows, the system performance deteriorates dramatically. Secondly, firms should encourage customers to abandon when having load-sensitive service rate. This can be done by, for example, providing delay announcements. To overcome the bi-stability phenomenon, we propose three operational solutions: increasing staffing, increasing abandonment rate, and admission control (§6.4.3).

4) We show, using numerical examples (§6.5), that the bi-stability phenomenon remains when considering a large class of models. This includes settings in which the service rate deterioration is customer-driven (i.e., longer waiting results in longer service requirement for that customer), in which it is agent-driven (i.e., agents change their service rate according to queue length), or in situations where there is a delay in the slowdown effect on service rate (e.g., slowdown is

caused by agent fatigue).

Chapter 6

Slowdown Services: Potential Failures and Proposed Solutions

6.1 Model Setup

6.1.1 Load dependent Erlang-A model

We analyze a modified Erlang-A ($M/M/s+M$) model which incorporates the dependence of service rate on workload through the queue length process. Specifically, we consider an $M/M_Q/s+M$ queue, with s identical servers. Each server can serve only one customer at a time. Customers arrive to the system according to a Poisson process with rate λ. If a customer arrives and finds a server free, she starts service with that server immediately. Otherwise, she waits in the queue. Customers are served on a First-Come-First-Served basis. The service requirement is exponentially distributed with a state-dependent rate function $\mu(\cdot) \in C^2$. We assume customers have finite patience. The patience time of each customer is exponentially distributed with rate θ, which we refer to as the abandonment rate. If a customer does not get into service before her patience time expires, she abandons the queue.

We denote the queue length process by $Q \equiv \{Q(t) : t \geq 0\}$, where $Q(t)$ counts the

number of customers in the system (waiting and in service) at time t. Motivated by the empirical findings on slowdowns, we assume that the service rate of each server is a function of the scaled queue length process (as denoted by "M_Q" in our model), $\mu((Q-s)^+/s)$, where $(x)^+ \max\{0,x\}$. This scaling makes the workload process $((Q(t)-s)^+/s)$ of the same order as the delay process (waiting time of an imaginary arrival at time t) [54]. It is essential when considering scaling for approximations.

We are interested in service systems in which the service rate deteriorates as the congestion level grows. We measure the level of load sensitivity by $\mu'(x)$ and let $\mu^{(i)}(0) := \lim_{x\to 0+} \mu^{(i)}(x)$ for $i = 1, 2$. We further assume that the service rate function exhibits a diminishing decreasing rate and a minimum positive level. Formally:

Assumption 6.1.1 *$\mu'(x) \leq 0$ and $\mu''(x) \geq 0$ for all $x \geq 0$. $lim_{x\to\infty}\mu(x) = \mu(\infty) > 0$.*

In our numerical demonstrations, we use a specific form of the service rate function: $\mu(x) = c + a\exp(-bx)$ with parameters $a, b, c > 0$, which clearly satisfies Assumption 6.1.1. To demonstrate changes in load sensitivity, we change the values of b while keeping all other parameters fixed. We refer to b as the load sensitivity parameter.

Under our assumptions on the service rate function, $Q(t)$ is a Markov jump process. More specifically, $Q(t)$ is a Birth-and-Death (B&D) process with birth rate λ and state dependent death rate $\mu((Q-s)^+/s)(Q\wedge s)+\theta(Q-s)^+$, where $x\wedge y = \min\{x, y\}$. As $\theta > 0$, $Q(t)$ admits a unique steady-state distribution. We denote

$$\pi(q) := P(Q(\infty) = q).$$

$\pi(q)$ measures the long run average amount of time the system spends at q.

6.1.2 The QED heavy-traffic regime

To balance the quality of service with system efficiency, we aim to operate the queue in the QED regime. For an $M/M/s+M$ queue, the QED regime is obtained by holding

the service rate and abandonment rate fixed while letting the aggregate arrival rate λ and number of servers s grow to infinity in such a way that the utilization rate $\rho = \lambda/(s\mu)$ approaches 1 with a certain rate. Specifically, consider a sequence $M/M/n+M$ queue indexed by n with arrival rate $\lambda_n \to \infty$ as $n \to \infty$. Set $\rho_n := \lambda_n/(n\mu)$. It is assumed that

$$\sqrt{n}(1 - \rho_n) \to \beta \text{ as } n \to \infty \tag{6.1}$$

for some $\beta \in \mathbb{R}$, or equivalently, that the number of servers is set by the square root formula – $n = R_n + \beta\sqrt{R_n}$, where $R_n = \lambda_n/\mu$.

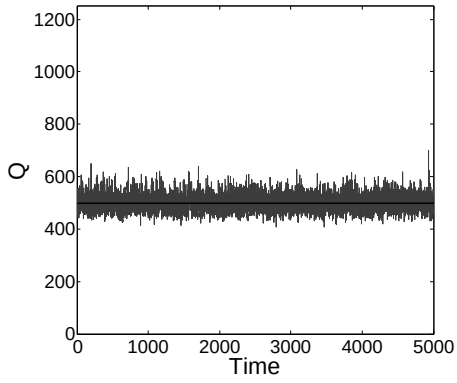
Garnett et. al. [53] proved that when a sequence of Erlang-A systems satisfies Equation (6.1) (i.e., operates in the QED regime), The probability of waiting, $P(W)$, is non-degenerate and the probability of abandonment, $P(Ab)$, converges to zero at rate $1/\sqrt{n}$. Thus, systems that operate in this regime achieve both good performance and high efficiency. However, as Figure 6.1 illustrate, in the modified Erlang-A model, SRS does not guarantee similar performance.

In the absence of workload sensitivity, (i.e., $b = 0$), the system with the same parameters as in Figure 6.1 operates in the QED regime, with $\beta = 0.3$, $P(W) = 0.1882$ and $P(Ab) = 0.0018$. Figure 6.1 illustrates that this is not necessarily the case when the systems exhibit load sensitivity. In the first case ($b = 1$), the system still operates in the QED regime with low probability of waiting, $P(W) = 0.2050$, and abandonment, $P(Ab) = 0.0023$. But the performance is worse than the one obtained without sensitivity. In the second case ($b = 2$), we observe the phenomenon of bi-stability. There are two peaks in the stationary distribution: a lower level where the performance is good ($P(W) \approx 0.2$ and $P(Ab) \approx 0.02$), and a high leveler where the service level is poor ($P(W) \approx 1$ and $P(Ab) \approx 0.2$). The average performance yields $P(W) = 0.9090$ and $P(Ab) = 0.2008$.

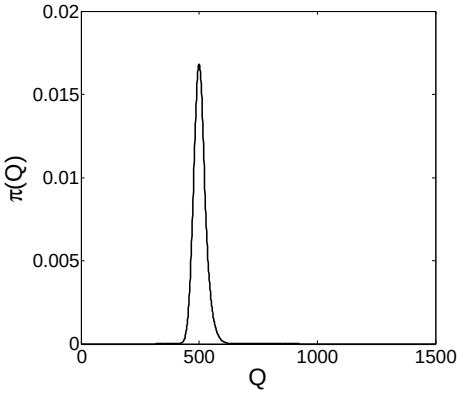
Figure 6.2 shows how the probability of waiting and the probability of abandonment change as functions of the load sensitivity parameter, b. We observe that the

Figure 6.1: Sample path and stationary distribution of the number of people in the system for $M/M_Q/s + M$ queues with different load sensitivity parameter values, b ($s = 512$, $\lambda = 500$, $\mu = 0.6 + 0.4\exp(-b(q-s)^+/s)$ and $\theta = 0.3$)

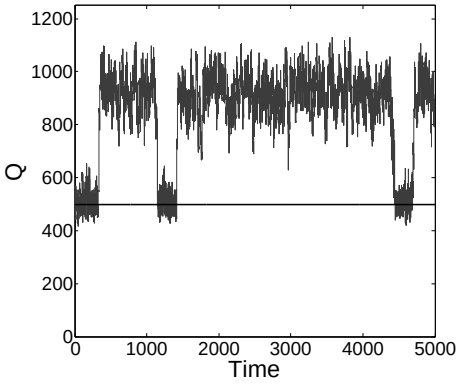
(a) $b = 0.5$: Sample path



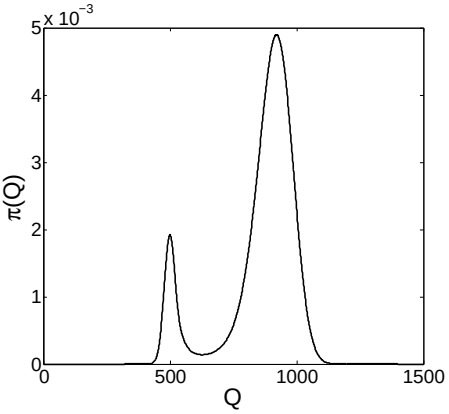
(b) $b = 0.5$: Stationary distribution



(c) $b = 1.5$: Sample path



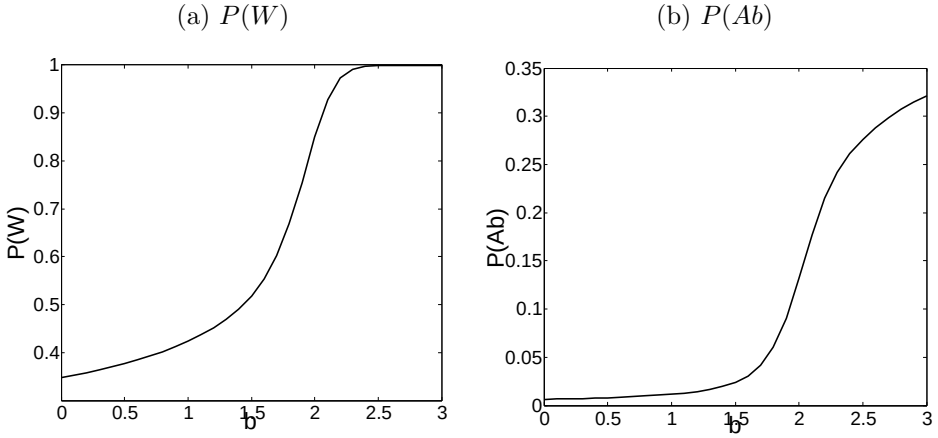
(d) $b = 1.5$: Stationary distribution



effect of load sensitivity is nonlinear. The performance deteriorates drastically as the sensitivity parameter grows beyond a certain level (e.g., at around $b = 1.5$ for the parameters in Figure 6.2).

These demonstrations imply that it is SRS may not be enough to achieve QED regime performance in service systems with load-sensitive slowdown effect. In the next section, we use the many-server heavy traffic analysis to analyze the dynamics of such systems.

Figure 6.2: Performance measures for $M/M_Q/s+M$ queues as a function of the load sensitivity parameter, b ($s = 512$, $\lambda = 500$, $\mu = 0.6 + 0.4\exp(-b(q-s)^+/s)$ and $\theta = 0.5$)



6.2 Fluid analysis

In this section, we establish the fluid limit of the queue length process of the modified Erlang-A model. This deterministic model serves as an approximation for the corresponding stochastic system when the system scale is large. We then conduct an equilibrium analysis of the fluid model. That provides important characterization of the stationary performance of the original system.

6.2.1 Fluid approximation

To develop the fluid limit, we consider a sequence of $M/M_Q/n+M$ queues indexed by n, where the arrival rate $\lambda_n \to \infty$ as $n \to \infty$. For the n-th system, we denote $Q_n \equiv \{Q_n(t) : t \geq 0\}$ as the queue length process (number of people in the system). The abandonment rate does not scale with n and the service rate function takes the same form when applied to the scaled queue length process. As we are interested in the QED asymptotic regime, we assume that there exists a β such that $\lim_{n\to\infty} \sqrt{n}(1-\lambda_n/(n\mu(0))) = \beta$.

Let $A \equiv \{A(t) : t \geq 0\}$, $S \equiv \{S(t) : t \geq 0\}$ and $R \equiv \{R(t) : t \geq 0\}$ be three independent Poisson processes, each with unit rate. A, S and R generate the arrival, service completion and abandonment processes, respectively. Then, the pathwise construction of Q_n is:

$$\begin{aligned} Q_n(t) &= Q_n(0) + A(\lambda_n t) - S\left(\int_0^t \mu\left(\frac{(Q_n(u)-n)^+}{n}\right)(Q_n(u)\wedge n)\,du\right) \\ &\quad -R\left(\int_0^t \theta(Q_n(u)-n)^+ du\right). \end{aligned}$$

We define the fluid-scaled process

$$\bar{Q}_n(t) = \frac{Q_n(t)}{n}$$

Let $\mathcal{D} := D([0,\infty),\mathbb{R})$ denote the function space of all right-continuous real-valued functions on the interval $[0,\infty)$ with left limit everywhere in $(0,\infty)$, endowed with Skorohod (J_1) topology.

Theorem 6.2.1 *If $\bar{Q}_n(0) \Rightarrow \bar{Q}(0)$ in $\mathbb{R}$, then $\bar{Q}_n \Rightarrow \bar{Q}$ in $\mathcal{D}$ as $n \to \infty$. The limit process $\bar{Q}$ is the unique solution satisfying the following integral equation*

$$\bar{Q}(t) = \bar{Q}(0) + \mu(0)t - \int_0^t \mu\left(\left(\bar{Q}(u)-1\right)^+\right)\left(\bar{Q}(u)\wedge 1\right)du - \int_0^t \theta\left(\bar{Q}(u)-1\right)^+ du.$$

The proof of Theorem 6.2.1 and all subsequent results can be found in Appendix 6.7.

Let $f(q)$ be the flow rate function of the fluid system at state q. That is $f(q) = \lambda - \mu((q-s)^+/s)(q\wedge s) - \theta(q-s)^+$. Then we can write $\bar{Q}(t)$ as the solution to the following autonomous differential equation with initial value $\bar{Q}(0)$:

$$\dot{\bar{Q}} = f(\bar{Q}) \tag{6.2}$$

where $\dot{\bar{Q}}$ denotes the derivative of $\bar{Q}$ with respect to t.

6.2.2 Equilibrium analysis

Next, we analyze the long term behavior of the fluid model, i.e., the state of the system as $t \to \infty$. To make the dependence of the flow, $\bar{Q}(t)$, on its initial value, $\bar{Q}(0)$, explicit, we write $\Phi(q_0, t) = \bar{Q}(t)$ with an initial value q_0.

Definition 6.2.2 (Equilibrium) *A point $\bar{q}$ is an* **equilibrium** *of the dynamic system (6.2) if*

$$\Phi(\bar{q}, t) = \bar{q}, \text{ for all } t \geq 0.$$

By Definition 6.2.2, $\bar{q}$ is an equilibrium of a system if when the trajectory of the flow defined by (6.2) starts at $\bar{q}$, it stays there. In our model, $\bar{q}$ can be computed by solving $f(q) = 0$. However, it is unclear where the trajectories of the flow converge to if the initial value $q_0 \neq \bar{q}$. We therefore analyze the stability of the equilibrium points.

Definition 6.2.3 (Stability of equilibrium) *Let $\bar{q}$ be an equilibrium point of the dynamic system. $\bar{q}$ is said to be* **stable** *if for any $\epsilon > 0$, there exist $\delta > 0$, such that if $|q - \bar{q}| < \delta$, $|\Phi(q, t) - \bar{q}| < \epsilon$ for any $t \geq 0$. Otherwise, $\bar{q}$ is* **unstable**. *If δ can be chosen such that not only $\bar{q}$ is stable, but also $\lim_{t\to\infty} \Phi(q, t) = \bar{q}$ for $|q - \bar{q}| < \delta$, then $\bar{q}$ is said to be* **asymptotically stable**.

By Definition 6.2.3, $\bar{q}$ is asymptotically stable if starting close enough to $\bar{q}$, trajectories defined by (6.2) converge to $\bar{q}$ as $t \to \infty$. An equilibrium may also be **semistable**. In a semistable equilibrium, trajectories that start on one side of the equilibrium converge to it, whereas trajectories that start on the other side do not. Note that a semistable equilibrium is unstable by Definition 6.2.3.

To characterize the equilibria of the fluid model in (6.2), we analyze the function $f(q)$. When $q \leq 1$, $f(q) = \mu(0) - \mu(0)q$ is a linearly decreasing function that starts at $f(0) = \mu(0) > 0$ and ends at $f(1) = \mu(0) - \mu(0) = 0$. When $q \geq 1$, under Assumption 6.1.1 $f'(q) = -\mu'(q-1) - \theta$ and $f''(q) = -\mu''(q-1) \leq 0$. Therefore, $f(q)$ is concave

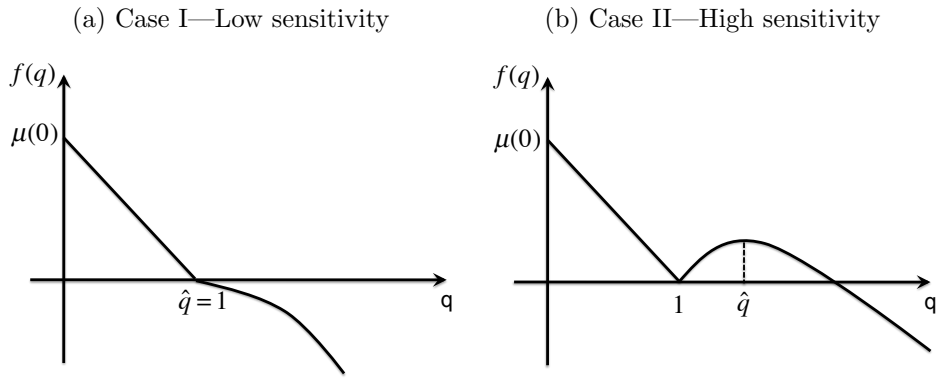
on $[1, \infty)$. Let $\hat{q} = \arg\max_{q \in [s,\infty)} f(q)$. We refer to $\hat{q}$ as the *critical point* of the system. Depending on the actual form of $f(q)$, we distinguish between the following two cases (as shown in Figure 6.3):

Case I (Low Sensitivity): $-\mu'(0) \leq \theta$.

Case II (High Sensitivity): $-\mu'(0) > \theta$.

Under Case I, the case with low sensitivity, we have $\hat{q} = 1$ and under Case II, the case with high sensitivity, $\hat{q}$ is the root of $f'(q) = 0$ for $q \geq 1$. The following theorem summarizes the stability analysis of the equilibria for the two cases.

Figure 6.3: Flow rate function under two cases



Theorem 6.2.4 *Assume $\lambda = s\mu(0)$ and Assumption 6.1.1.*

(i) If $-\mu'(0) \leq \theta$ (Low Sensitivity), there is a unique equilibrium, $\bar{q}$, with $\bar{q} = 1$. Furthermore, $\bar{q}$ is asymptotically stable.

(ii) If $-\mu'(0) > \theta$ (High Sensitivity), there are two equilibria, $\bar{q}_1$ and $\bar{q}_2$, with $\bar{q}_1 = 1$ and $\bar{q}_2 > \hat{q}$. $\bar{q}_1$ is a semistable equilibrium and $\bar{q}_2$ is an asymptotically stable equilibrium.

In the low sensitivity case, $\bar{q} = 1$ is the unique and asymptotically stable equilibrium of the fluid model. Therefore, the fluid model will converge to that value. In

the stochastic level, we sill see the trajectory of the queue length process fluctuates around n for the n-th system. We analyze its performance in more details in §6.3.

In the high sensitivity case, there are two equilibria, $\bar{q}_1$ and $\bar{q}_2$. The fluid model may converge to either one, depending on the starting point. In the stochastic level, the queue length process may alternate between the two equilibrium levels. This drives the bi-stability phenomenon observed in Figure 6.1c. However, $\bar{q}_1$ is a semistable equilibrium. Therefore, in the stochastic level, we expect the queue length process to eventually spend most of the time around the higher equilibrium level as the system scale grows large. We explore how the scale parameter n and other system parameters affect the bi-stability phenomenon under High Sensitivity in §6.4.

6.3 Performance Analysis under Low Sensitivity

In this section, we conduct asymptotic analysis for the modified Erlang-A model under low load sensitivity ($-\mu'(0) < \theta$). We establish closed-form approximations for the performance measures ($P(W)$ and $P(Ab)$), which can be used to determine the corresponding square-root staffing parameters. We then present numerical results to demonstrate the quality of the approximations.

Let Y_n denote the normalized steady-state queue length process. In particular,

$$Y_n = \frac{Q_n(\infty) - n}{\sqrt{n}}.$$

We then have the following result about the limiting distribution of Y_n.

Theorem 6.3.1 *Under low sensitivity ($-\mu'(0) < \theta$) and SRS with parameter β, Y_n converges weakly to a distribution with the following probability density function*

$$g(y) = \begin{cases} \frac{C_1}{\sqrt{2\pi}} \exp\left(-\frac{(y+\beta)^2}{2}\right) & \text{if } y \leq 0 \\ \frac{C_2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y+\beta\sigma^2)^2}{2\sigma^2}\right) & \text{if } y > 0, \end{cases}$$

where

$$\sigma = \sqrt{\frac{\mu(0)}{\mu'(0)+\theta}}\ ,\ C_1 = \frac{h(\beta\sigma)}{\sigma\phi(\beta)}\left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}\ ,\ C_2 = \frac{h(\beta\sigma)}{\phi(\beta\sigma)}\left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}$$

and $h(\cdot)$ denotes the hazard rate function of the standard normal distribution. Specifically, $h(z) = \phi(z)/\bar{\Phi}(z)$, where $\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2)$ and $\bar{\Phi}(z) = \int_z^\infty \phi(z)dz$ is the complementary cumulative distribution function.

Theorem 6.3.1 shows that the limiting distribution of the scaled process has normal tails but it is not symmetric around zero unless $(\mu'(0)+\theta)/\mu(0) = 1$, and the left tail decays slower as the sensitivity level $|\mu'(0)|$ increases.

From Theorem 6.3.1, we have the following asymptotic results about the performance measures.

Corollary 6.3.2 *Under low sensitivity ($-\mu'(0) < \theta$) and SRS with parameter β,*

$$\lim_{n\to\infty} P_n(W) = \left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}$$

and

$$\lim_{n\to\infty} \sqrt{n}P_n(Ab) = \left(\frac{h(\beta\sigma)}{\sigma} - \beta\right)\left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1} \frac{\theta}{\mu'(0)+\theta},$$

where $\sigma = \sqrt{\mu(0)/(\mu'(0)+\theta)}$.

Corollary 6.3.2 implies that the performance measures deteriorate with the load sensitivity level $\mu'(0)$, and it leads to the following approximations of the system performance measures:

$$P_n(W) \approx \left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1} \tag{6.3}$$

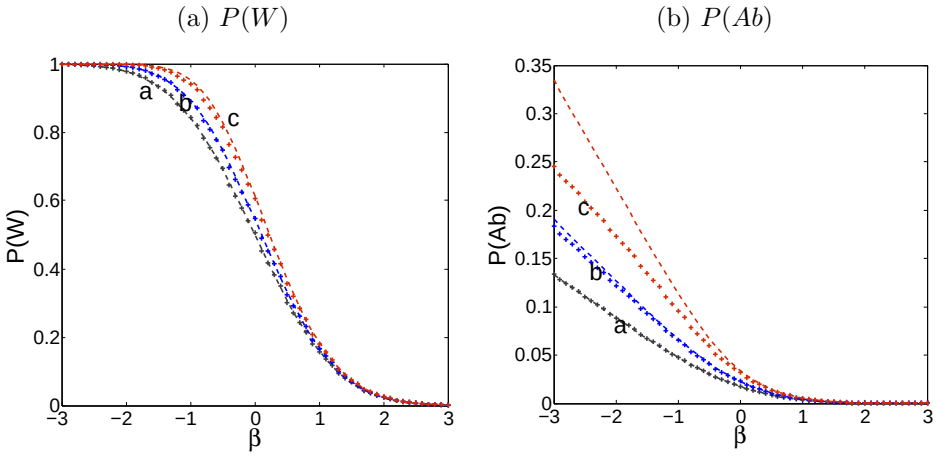
and

$$P_n(Ab) \approx \left(1 - \frac{h(\beta\sigma)}{h(\beta\sigma + 1/(\sigma\sqrt{n}))}\right)\left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1} \frac{\theta}{\mu'(0)+\theta}. \tag{6.4}$$

Figure 6.4 demonstrates the quality of these approximations (denoted by dashed lines) compared to the actual performance measures (marked by '+' signs), derived by simulation for different system parameters. Specifically, we choose three evenly spaces values of load sensitivity, measured by $\mu'(0)$, and the values of the square-root staffing parameter β between -3 to 3.

We observe that (6.3) provides a good approximation for $P(W)$ for a wide range of load sensitivity levels and β values. On the other hand, (6.4) provides a good

Figure 6.4: Approximations for $P(W)$ and $P(Ab)$ at three different load sensitivity levels: a: $\mu'(0) = 0$, b: $\mu'(0) = -0.3$, c: $\mu'(0) = -0.6$



approximation of $P(Ab)$ for only lower levels of load sensitivity ($|\mu'(0)| \leq 0.3$). In other words, the quality of (6.4) deteriorates as the load sensitivity level approaches abandonment rate, $|\mu'(0)| \to \theta$; in that case the approximation tends to overestimate the system performance measures. Practically speaking, however, because the QED regime aims for less than 10% abandonments, we can restrict attention to the range of β's which result in $P(Ab) < 10\%$. For example, $\beta > -1$ when $\mu'(0) = -0.6$. In that range, (6.4) works as a good approximation for the probability of abandonment, where the maximum gap between the two is 0.025.

We also observe that for a fixed β, system performance ($P(W)$ and $P(Ab)$) deteriorates with the load sensitivity level $\mu'(0)$. Therefore, neglecting to account for load sensitivity would *underestimate* system performance. Put differently, fixing a target system performance, a load sensitive service system requires more staffing to achieve the same level of performance. One can use (6.3) and (6.4) to find the appropriate square-root staffing parameter to achieve certain performance measures in the QED regime.

Remark 6.3.1 *We conclude this section by drawing some connections to the ordinary*

Erlang-A model. We notice that the limiting distribution of Y_n, in Theorem 6.3.1, is the same as the limiting distribution of the normalized queue length process of a sequence of ordinary Erlang-A model with the same arrival rate λ_n, constant service rate $\mu(0)$ and reduced abandonment rate $\mu'(0)+\theta$ in stationarity [53]. This is because, under the low sensitivity conditions, the two systems have the same arrival rates and very similar death rates. When $q < n$, the death rates of the two systems are equal; when $q > n$, the death rate of the modified Erlang-A queue is:

$$\begin{aligned} & \mu\left(\frac{q-n}{n}\right)n + \theta(q-n) \\ = \;& \mu(0)n + (\mu'(0)+\theta)(q-n) + \mu''(\eta)\frac{(q-n)^2}{n} \\ & \text{for some } \eta \in (0,(q-n)/n) \\ \approx \;& \mu(0)n + (\mu'(0)+\theta)(q-n) \\ & \text{when } (q-n)^2/n \text{ is small, i.e. when } q-n = O(\sqrt{n}). \end{aligned}$$

The reduced abandonment rate in the corresponding ordinary Erlang-A model suggests that the load sensitivity effectively lessens the stabilizing effect of abandonment. We also notice that as $\mu''(\cdot) \geq 0$, $\mu(\frac{q-n}{n})n + \theta(q-n) \geq \mu(0)n + (\mu'(0)+\theta)(q-n)$, the ordinary Erlang-A model with reduced abandonment rate is stochastically larger than the modified model (Lemma 6.4.6). Therefore, the stationary queue length process, $Q_n(\infty)$, of our modified model is within $n \pm O(\sqrt{n})$ with high probability.

6.4 Bi-Stability Analysis: Performance Analysis under High Sensitivity

In this section, we analyze the system dynamics when sensitivity is high and in particular, the factors that affect the bi-stability phenomenon. We start with the scale parameter n. We then keep n fixed and analyze the effect of other system parameters, specifically, the square-root staffing parameter β, the sensitivity of the service rate

function and the abandonment rate θ. We also propose some staffing and admission control policies to eliminate the bi-stability effect and avoid ED performance.

6.4.1 The effect of the scale parameter n

We begin by characterizing the local maximums of the stationary distribution, i.e., its peaks. When bi-stability occurs there are two peaks, as was shown in Figure 6.1d. Naturally, there is a one-to-one correspondence between these peaks and the (semi-)stable equilibria of the fluid model. Recall that $Q_n = \{Q_n(t) : t \geq 0\}$ is a B&D process with birth rate λ_n and state-dependent death rate $\mu((q-n)^+/n)(q \wedge n) + \theta(q-n)^+$, where $Q_n(t) = q$. Let $\pi_n(\cdot)$ denote the steady-state distribution of Q_n.

From the detailed balance equation

$$\lambda_n \pi_n(q) = \left(\mu((q-n)^+/n)(q \wedge n) + \theta(q-n)^+\right) \pi_n(q+1),$$

we get

$$\pi_n(q+1) - \pi_n(q) = \left(\frac{\lambda_n}{\mu((q-n)^+/n)(q \wedge n) + \theta(q-n)^+} - 1\right) \pi_n(q).$$

As a result, when $\lambda_n \geq \mu((q-n)^+/n)(q \wedge n) + \theta(q-n)^+$, $\pi_n(q+1) \geq \pi_n(q)$; otherwise, $\pi_n(q+1) < \pi_n(q)$. Hence, we find the value of peaks of $\pi_n(\cdot)$ by analyzing the sign of $f_n(q) := \lambda_n - \mu((q-n)^+/n)(q \wedge n) - \theta(q-n)^+$. When $q < n$, $f_n(q) = \lambda_n - \mu(0)q$ is a linearly decreasing function. When $q \geq n$, if we let $x_n = (q-n)/n$, then we have

$$\frac{f_n(q)}{n} = \frac{\lambda_n}{n} - \mu(x_n) - \theta x_n.$$

To simplify notation, let $\nu(x) := \mu(x) + \theta x$ for $x \geq 0$ and $\hat{x} > 0$ denote the root of $\nu'(x) = 0$. Under Assumption 6.1.1 and High Sensitivity, $\nu(\cdot)$ is convex and attains its minimum at $\hat{x}$. We also denote $\bar{x}_n$ as the root of $\lambda_n/n - \nu(x) = 0$ on $(\hat{x}, \infty)$. The next theorem characterizes the peaks of $\pi_n(\cdot)$.

Theorem 6.4.1 *Let $R_n = \lambda_n/\mu(0)$. Under High Sensitivity ($-\mu'(0) > \theta$) and for $n = R_n + \beta\sqrt{R_n}$ (SRS):*

i) when $\beta < 0$, $\pi_n(\cdot)$ has a unique peak, $\bar{q}_{n,2} = \lfloor(\bar{x}_n + 1)n\rfloor$;

ii) when $\beta > 0$,

(a) if $\lambda_n/n \leq \nu(\hat{x})$, $\pi_n(\cdot)$ has a unique peak, $\bar{q}_{n,1} = \lfloor\lambda_n/\mu(0)\rfloor$;

(b) if $\lambda_n/n > \nu(\hat{x})$, $\pi_n(\cdot)$ has two peaks, $\bar{q}_{n,1} = \lfloor\lambda_n/\mu(0)\rfloor$ and $\bar{q}_{n,2} = \lfloor(\bar{x}_n + 1)n\rfloor$.

As $\lim_{n\to\infty} \lambda_n/n = \mu(0)$, when $\beta > 0$, the stationary distribution, $\pi_n(\cdot)$, may have a unique peak for small values of n, but will eventually have two peaks as n grows large.

Let $\bar{x}$ be the root of $\mu(0) - \nu(x) = 0$ on $(\hat{x}, \infty)$. As $\nu(\cdot)$ is continuously increasing on $(\hat{x}, \infty)$, $\bar{x}_n \to \bar{x}$ as $n \to \infty$. It is also easy to check that $\bar{x} = \bar{q}_2 - 1$, i.e. $\bar{x}$ measures the distance between the two fluid equilibria. The next theorem characterizes the relative *magnitude* of the two peaks.

Theorem 6.4.2 *Under High Sensitivity ($-\mu'(0) > \theta$) and SRS with $\beta > 0$*

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{\pi_n(\bar{q}_{n,2})}{\pi_n(\bar{q}_{n,1})} = I(\bar{x}),$$

where

$$I(\bar{x}) = \int_0^{\bar{x}} \log \frac{\mu(0)}{\nu(x)} dx \geq 0.$$

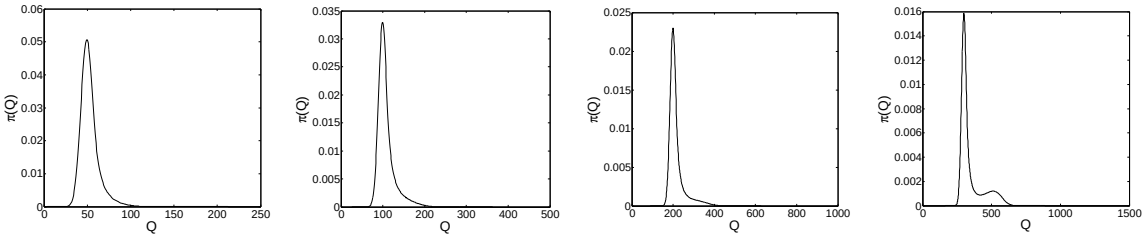
Theorem 6.4.2 indicates that $\pi_n(\bar{q}_{n,2}) \approx \pi_n(\bar{q}_{n,1}) \exp(nI(\bar{x}))$. This means that the difference in magnitude between the two peaks ($\pi_n(\bar{q}_{n,1})$ and $\pi_n(\bar{q}_{n,2})$) grows exponentially in n. Figure 6.5 demonstrates how the stationary distribution of the system with $\beta > 0$ evolves with the scale parameter, n. For small values of n ($n \leq 200$), $\pi_n(\cdot)$ has a unique peak ($\bar{q}_{n,1}$). As n increases, a "second peak" ($\bar{q}_{n,2}$) emerges and its magnitude compared to the first peak increases. For very large n, only $\bar{q}_{n,2}$ remains effective.

Remark 6.4.1 *Theorem 6.4.2 suggests that systems with high service rate sensitivity will have the dis-economies-of-scale effect. Unlike traditional Erlang-A model using*

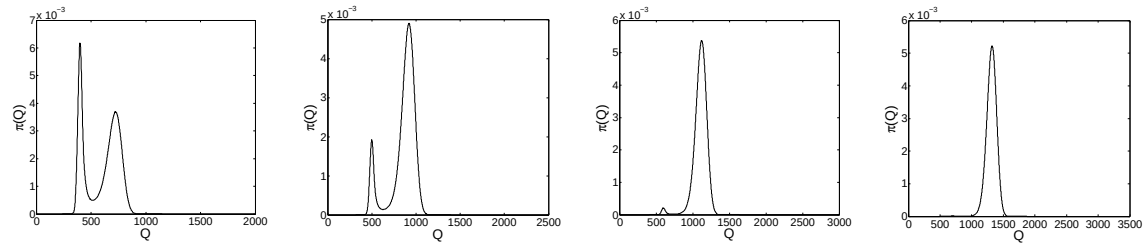
SRS, where larger system provide better performance levels, the performance of our modified model deteriorates as the system scale grows.

Figure 6.5: Approximated stationary distribution of the number of people in the system for $M/M_Q/n+M$ queues with scale parameter values n ($n = \lceil R_n + \sqrt{R_n} \rceil$, $\mu = 0.6 + 0.4\exp(-1.5(q-n)^+/n)$ and $\theta = 0.3$).

(a) $n = 58$ ($R_n = 50$) (b) $n = 110$ ($R_n = 100$) (c) $n = 215$ ($R_n = 200$) (d) $n = 318$ ($R_n = 300$)



(e) $n = 420$ ($R_n = 400$) (f) $n = 523$ ($R_n = 500$) (g) $n = 625$, ($R_n = 600$) (h) $n = 727$ ($R_n = 700$)



We next analyze factors that affect the value of $I(\bar{x})$. To facilitate the comparison, we restrict our analysis to the following ordering of load sensitivity.

Definition 6.4.3 *For two service rate function $\mu_1(\cdot)$ and $\mu_2(\cdot)$, with $\mu_1(0) = \mu_2(0)$, we say that μ_2 is more load-sensitive than μ_1, if $\mu_2(x) \leq \mu_1(x)$ for all $x > 0$.*

The next lemma looks on the effect of the system parameters on the value of the higher level fluid equilibrium, $\bar{q}_2$.

Lemma 6.4.4 *Under High Sensitivity ($-\mu'(0) > \theta$) and SRS with $\beta > 0$*

i) the more load-sensitive the service rate function , the larger the value of $\bar{q}_2$;

ii) the larger the abandonment rate θ, the smaller the value of $\bar{q}_2$.

Based on Lemma 6.4.4, the next lemma summarizes the effect of the load sensitivity of the service rate function and the abandonment rate on the value of $I(\bar{x})$.

Lemma 6.4.5 *Under High Sensitivity ($-\mu'(0) > \theta$) and SRS with $\beta > 0$*

i) the more load-sensitive the service rate function, the larger the value of $I(\bar{x})$;

ii) the larger the abandonment rate θ, the smaller the value of $I(\bar{x})$.

Remark 6.4.2 *Lemmas 6.4.4 and 6.4.5 indicate that as load sensitivity increases the distance between the two equilibria increases, and with it the rate of convergence to the upper equilibria ($I(\bar{x})$). Hence, we will observe bi-stability for smaller systems only. Abandonment rate has the opposite effect—as θ increases, the convergence to the upper equilibria is slower, hence, we will observe bi-stability in larger systems.*

6.4.2 The effect of other system parameters

For a fixed system scale parameter, n, in this section, we analyze the effect of the square-root staffing parameter, β, the service rate sensitivity and the abandonment rate, θ, on the bi-stability phenomenon (the magnitude of the two peaks). Theorem 6.4.1 shows that bi-stability (the existence of the two peaks) only arises for $\beta > 0$ and large n. We therefore concentrate on these parameter ranges.

We start by giving a formal definition for the time *around* the lower/upper equilibrium level by defining the threshold point $\tilde{q}_n$. This threshold outlines the region around the lower equilibrium level as $[0, \tilde{q}_n]$ and the region around the upper equilibrium level as $(\tilde{q}_n, \infty)$. Let $\tilde{x}_n$ be the root of $\lambda_n/n - \nu(x) = 0$ on $[0, \hat{x})$ and $\tilde{q}_n := \lfloor(\tilde{x}_n + 1)n\rfloor$. As $f_n(q) < 0$ for $q \in (\bar{q}_{n,1}, \tilde{q}_n)$, $q \in \mathbb{Z}^+$, $\pi_n(q)$ is decreasing on $(\bar{q}_{n,1}, \tilde{q}_n)$; and as $f_n(q) > 0$ for $q \in (\tilde{q}_n, \bar{q}_{n,2})$, $q \in \mathbb{Z}^+$, $\pi_n(q)$ is increasing on $(\tilde{q}_n, \bar{q}_{n,2})$. Thus, $\tilde{q}_n$ is the valley of $\pi_n(q)$ (see for example the valley around 600 in Figure 6.5f), and hence a good threshold to outline the two regions.

The next lemma provides the basis for the main comparison results (Theorem 6.4.7) in this subsection.

Lemma 6.4.6 *For two positive recurrent B&D processes, $Y^{(1)}$ and $Y^{(2)}$, defined on the same state space $\mathbb{Z}^+$, denote γ_i and $\xi_i(\cdot)$ as the birth rate and state-dependent death rate of $Y^{(i)}$, for $i = 1, 2$. If $\gamma_1 = \gamma_2$ and $\xi_1(y) \geq \xi_2(y)$ for every $y \in \mathbb{Z}^+$, then*

$$P(Y^{(1)}(\infty) > y) \leq P(Y^{(2)}(\infty) > y).$$

From Lemma 6.4.6, we have the following theorem that studies the effect of the system parameters on the proportion of time the system spends around each equilibrium level.

Theorem 6.4.7 *Under High Sensitivity ($-\mu'(0) > \theta$) and square-root Staffing with $\beta > 0$,*

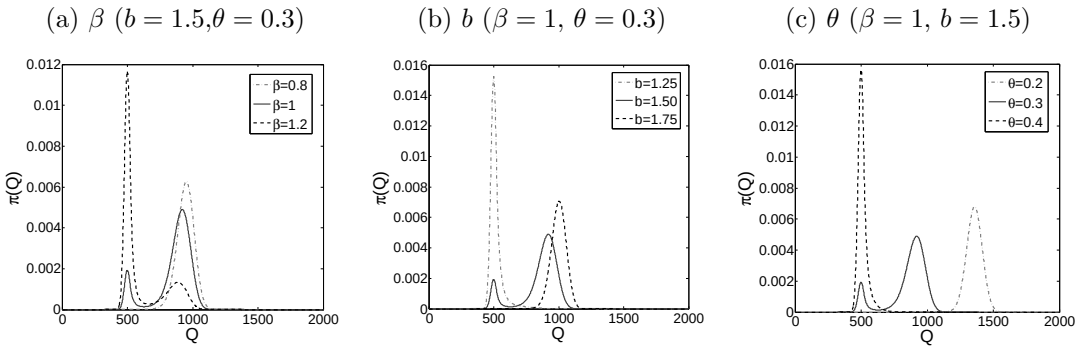
i) if $\mu(\infty) \geq \theta$, then the proportion of time the system spends around the upper equilibrium decreases with the square-root staffing parameter, β;

ii) under Definition 6.4.3, the more load-sensitive the service rate function, the larger proportion of time the system spends around the upper equilibrium;

iii) the proportion of time the system spends around the upper equilibrium decreases with the abandonment rate, θ.

Figure 6.6 demonstrates how the value of the peaks and proportion of time the system spends around each peak changes with the square-root staffing parameter β, the sensitivity parameter, b, and the abandonment rate, θ. We notice that the value of the second peak (the larger one) increases with the load sensitivity parameter b and decreases with the abandonment rate θ, as was proved in Lemma 6.4.4. In addition, we notice that the value of the second peak decreases with the square-root staffing parameter β, but the difference is much smaller when compared to the effect of b and θ. (This change is not apparent in the fluid level and, hence, less significant.)

Figure 6.6: Approximated stationary distribution of the number of people in the system for $M/M_Q/n+M$ queues with different system parameters ($n = \lambda + \beta\sqrt{\lambda}$, $\lambda = 500$, $\mu = 0.6 + 0.4\exp(-b(q-s)^+/s)$ and θ).



The effect of β on the performance measures of our load-sensitive model is similar to that of the traditional/nonsensitive Erlang-A model; $P(W)$ and $P(Ab)$ both decrease with β. The effect of load sensitivity, is also straightforward. The system with a less sensitive service rate function has on average a higher service rate. The performance measures hence improve. In contrast, the effect of the abandonment rate, θ, is quite counterintuitive. In the traditional Erlang-A model, it is well established if customers are less patient (i.e., θ increases), $P(W)$ decreases but $P(Ab)$ increases [53], while in our modified Erlang-A model with high sensitivity, both the probability of waiting and the probability of abandonment decrease with θ. This is because the load-sensitive system reaches the high equilibrium less frequently as θ increases.

The analysis implies that the abandonment rate and the load sensitivity of the service rate function affect system performance differently when service rates exhibit slowdowns due to congestion. While a high load sensitivity level negatively affects system performance, a high abandonment rate may actually improve performance by alleviating the deterioration in service rate. Hence, managers are advised to *encourage* customers to abandon in a load-sensitive environment. This can be done, for example, by providing delay announcements when the system is loaded, as it is known that announcements increase abandonment rate [56; 66].

6.4.3 Policies to avoid bi-stability under High Sensitivity

We next propose and analyze three policies to eliminate the bi-stability phenomenon and avoid ED regime performance under High Sensitivity. The first approach, *Policy A*, is to increase the staffing level sufficiently so that the stationary distribution has only one peak around the lower level. The second approach, *Policy B*, is to adjust the abandonment rate once a threshold is reached so that the stationary distribution again has only one peak around the lower level. The third approach, *Policy C*, is to block the incoming arrivals or reroute them to other service facilities once a threshold is reached, thereby preventing the system from reaching the higher level peak (equilibrium). In the remaining part of this section, we expand on each policy.

A) *Increase staffing.* In this policy, we eliminate the higher-level equilibrium by increasing staffing level. The new staffing level, $\bar{n}$, should be such that $\lambda_n/\bar{n} \leq \nu(\hat{x})$. Suppose we set $\bar{n} = \lambda_n/\nu(\hat{x})$. Then we have

$$\frac{\bar{n}-n}{n} = \frac{\lambda_n/n}{\nu(\hat{x})} - \left(\frac{\lambda_n/n}{\mu(0)} + \frac{\beta}{\sqrt{n}}\sqrt{\frac{\lambda_n/n}{\mu(0)}}\right) \to \frac{\lambda}{\nu(\hat{x})} - \frac{\lambda}{\mu(0)} \text{ as } n \to \infty.$$

This implies that we need to increase staffing by $O(n)$ servers. Thus, a potential drawback of this approach is that by raising the staffing level to $\bar{n}$, a service provider may "overstaff" the system to operate in the QD regime.

B) *Increase abandonment.* Under this policy, we eliminate the higher level equilibrium by adjusting the abandonment rate. This is doable using, for example, delay announcement [66]. In what follows, we consider a special intervention, under which we increase the abandonment rate to $\bar{\theta}$ when the queue length process pass a certain threshold h_n. The h_n and $\bar{\theta}$ need to be choose appropriately, such that under the Policy B, the stationary distribution has only one peak around the lower level. Interestingly, the value of theses parameters could be derived from our bi-stability analysis earlier.

Recall that $\tilde{x}_n$ is the root of $\lambda_n/n - \nu(x) = 0$ on $(0, \hat{x})$. From the proof of Theorem 6.4.1, to eliminate the higher level peak (equilibrium), we need to set $h_n \in (n, (\tilde{x}_n + 1)n]$ and the corresponding $\bar{\theta} \geq |\mu'((h_n - n)/n)|$. This means that for this approach to be effective the delay announcement has to increase customer impatience by at least $|\mu'((h_n - n)/n)| - \theta$ at the threshold level h_n.

We also have the following asymptotic characterization of $\tilde{x}_n$.

Lemma 6.4.8 *Assume that* $\sqrt{n}(1 - \rho_n) \to \beta$ *as* $n \to \infty$ *with* $\beta > 0$. *Under High Sensitivity*

$$\lim_{n\to\infty} \sqrt{n}\tilde{x}_n = -\frac{\beta\mu(0)}{\mu'(0) + \theta}.$$

We notice from Lemma 6.4.8 that $\tilde{x}_n = O(\sqrt{n})$, thus $h_n - n = O(\sqrt{n})$ and $|\mu'((h_n - n)/n)| = \mu'(0) + O(1/\sqrt{n})$. We also note that the increased level of abandonment rate only need to hold for $h_n < q < \bar{q}_{n,2}$; it can go back to θ for higher values of queue length.

C) *Admission Control.* In a different approach, Policy C avoids bi-stability by constraining demand for services. Here we block customers (or reroute them to a different service group/facility) once a certain threshold, c, is reached. Under Policy C, the system becomes an $M/M_Q/n/c + M$ queue. To implement this policy, the system provider needs to characterize the appropriate threshold level, and the cost that such a policy entails on the system in terms of the proportion of customers blocked/rerouted. The "right" threshold could again be chosen based on our bi-stability analysis.

From the proof of Theorem 6.4.1, any choice of c_n, satisfying $n < c_n \leq (\tilde{x}_n + 1)n$, eliminates bi-stability, but the choice presents a tradeoff between the level of performance and the proportion of customers blocked: Setting a small c_n improves performance ($P(W)$ and $P(Ab)$ are low), but increases the proportion of customers that are blocked ($P(Bl)$). To find the optimal threshold within

this range, a service provider should evaluate the costs associated with each performance measure and strike a balance between them.

To gain more insight on the performance of this policy, we consider a sequence of $M/M_Q/n/c_n + M$ queues indexed by n. System n has arrival rate λ_n, state-dependent service rate $\mu((q-n)^+/n)$, abandonment rate θ and a finite system capacity c_n, so that incoming customers are blocked once the number of customers in the system reaches c_n. We denote the queue length process of the n-th system by $Q_n^c(\cdot)$.

We next develop diffusion approximations for Q_n^c for $c_n \leq (\tilde{x}_n + 1)n$. The pathwise construction of Q_n^c is

$$\begin{aligned} Q_n^c(t) &= Q_n^c(0) + A(\lambda_n t) - S\left(\int_0^t \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)(Q_n^c(u)\wedge n)du\right) \\ &\quad -R\left(\theta\int_0^t (Q_n^c(u)-n)^+ du\right) - L_n(t), \end{aligned}$$

where $L_n(t) = \int_0^t 1\{Q_n^c(s) = c_n\} dA(\lambda_n t)$. L_n counts the number of arrivals that are blocked from the system in $[0,t]$. We define the diffusion-scaled process

$$\hat{Q}_n^c(t) := \frac{Q_n^c(t) - n}{\sqrt{n}}.$$

Theorem 6.4.9 *Assume $\sqrt{n}(1-\rho_n) \to \beta$ as $n \to \infty$, where $\rho_n = \lambda_n/(n\mu(0))$ and $c_n/\sqrt{n} \to c \leq -\beta\mu(0)s/(\mu'(0)+\theta)$ as $n \to \infty$. If $\hat{Q}_n^c(0) \Rightarrow \hat{Q}^c(0)$ in $\mathbb{R}$ as $n \to \infty$, then $\hat{Q}_n^c \Rightarrow \hat{Q}^c$ in $\mathcal{D}$ as $n \to \infty$. The limit process $\hat{Q}^c$ is the unique process satisfying the stochastic integral equation:*

$$\begin{aligned} \hat{Q}^c(t) &= \hat{Q}^c(0) - \beta\mu(0)t + \sqrt{2\mu(0)}B(t) \\ &\quad - \int_0^t \left[\mu(0)(\hat{Q}^c(u)\wedge 0) + (\mu'(0)+\theta)\hat{Q}^c(u)^+\right] du - \hat{L}(t), \quad (6.5) \end{aligned}$$

where $\{B(t) : t \geq 0\}$ is a standard Brownian motion. $\hat{L}$ is the unique nondecreasing nonnegative process in $\mathcal{D}$ satisfying equation (6.5) *and*

$$\int_0^\infty 1\left\{\hat{Q}^c(t) < c\right\} d\hat{L}(t) = 0.$$

Q_n^c is an irreducible Markov chain with a finite state space. Thus, $\hat{Q}^c$ admits a unique stationary distribution, π. As $E_\pi[Q_n(t)] = E_\pi[Q_n(0)]$, by Theorem 6.4.9 and the Basic Adjoint Relation [74],

$$E_\pi[\hat{L}(t)] = \left(-\beta\mu(0) - \mu(0)E_\pi[\hat{Q}^c(0) \wedge 0] - (\mu'(0) + \theta)E_\pi[\hat{Q}^c(0)^+]\right) t$$

and the proportion of customers that are blocked from the n-th system, $P_n(Bl)$, satisfies

$$\begin{aligned} P_n(Bl) &\approx \frac{\sqrt{n}E_\pi[\hat{L}(t)]}{\lambda_n t} \\ &= \frac{1}{\sqrt{n}} \frac{\left(-\beta\mu(0) - \mu(0)E_\pi[\hat{Q}^c(0) \wedge 0] - (\mu'(0) + \theta)E_\pi[\hat{Q}^c(0)^+]\right)}{\mu(0)}. \end{aligned}$$

The probability of blocking is of $O(1/\sqrt{n})$. This implies that for large systems, the proportion of customers blocked and the proportion of time the system is blocked are very small. As the system is restricted to fluctuate around the lower equilibrium $\bar{q}_1$, we expect QED regime performance for $P(W)$ and $P(Ab)$, i.e. non-degenerate probability of waiting and $O(1/\sqrt{n})$ probability of abandonment.

We next compare the performance of the three policies numerically. Specifically, we compare how each policy improves the service level for different load sensitivity parameter values, and report on the implementation "cost" of each policy, i.e., the amount of added staffing or the proportion of customers abandoned/blocked. In Table 6.1, we compare three modified Erlang-A models with different levels of the load sensitivity parameter, b. In the Base Case table we present the performance measures of the load-sensitive system. In the Policy A table, we present the performance measures that correspond to an increased staffing level of $\bar{n} = \lambda/\nu(\hat{x})$. We denote by Δ_n the percentage increase in staffing. In the Policy B table, we present the performance measures of the system that operates under the policy of increasing the abandonment rate to $\bar{\theta} = -\mu'(0)$ once the queue length process pass the threshold

level $h_n = \tilde{q}_n$. Finally, in the Policy C table, we present the performance measures of the system that operates under the threshold control policy with the threshold level $c_n = \tilde{q}_n$.

For Policy A, as the sensitivity increases, to eliminate bi-stability, the additional staffing required increases. In particular, the percentage of extra staffing needed, Δ_n, increases from 3.89% (for $b = 1.25$) to 17.53% (for $b = 2.75$). These levels of extra staffing yield a QD regime performance for highly sensitive systems (e.g. $b \geq 1.75$). While the policy results in very good performance, adding so much extra staffing (as high as 17.53%) may lead to low server utilization, and can potentially be a costly solution. Alternatively, under Policy B, while the performance measures deteriorate with the sensitivity parameter, the performance remains within the QED regime. The column, $P(Q(\infty) > h_n)$, represents the proportion of time the increased abandonment intervention is active. We observe that it is small and increasing with the load sensitivity. This is consistent with our previous analysis, as the proportion of time the system is pushed towards the upper equilibrium increases with the load sensitivity. Lastly, under Policy C, the proportion of customers blocked is relatively low (at most 1.51% for the parameters in Table 6.1). The admission control policy keeps the performance measures within the QED regime characteristics for all sensitivity parameters tested. Hence, Policy B & C achieve a good service level while keeping the staffing according to the SRS rule.

6.5 Extensions

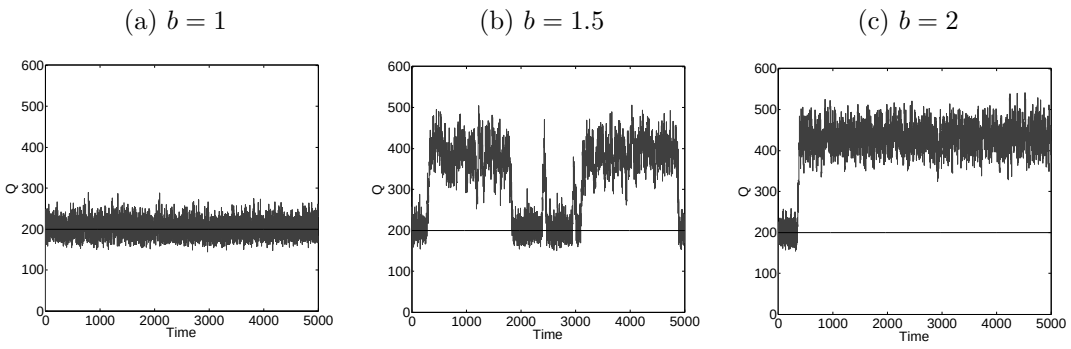
The model in Section 6.1 is the most befitting to explain agent-driven slowdowns, where the effect of the load of the system on service rates is applied instantly and to all agents simultaneously. Such a model fits situations where agents observe the current load and adjust their working rates accordingly. The exact same model cannot be directly applied to capture all various sources of the slowdown effect described in

the introduction. In this section, we modify the base model to better fit other sources of slowdown effects. In particular, we analyze: a) Customer-driven slowdowns, where each customer's waiting time affects only her own service rate, and b) Agent-driven slowdowns with a time lag, which can explain slowdowns caused by fatigue and hence takes time to take effect. Utilizing numerical approaches, we find that the primary insights (the existence of the two equilibria and the stochastic fluctuations between them) from our original model remain.

6.5.1 Customer-driven slowdown

In this section, we assume that a customer's service time is positively correlated with his own waiting time. In particular, the service rate of customer i is a function of her waiting time (not the queue length): $\mu_i = \mu(w_i)$ where w_i is the waiting time of customer i. [75] developed a performance approximation for this model based on theoretical bounds, a method which did not detect bi-stability in performance levels. Figure 6.7 demonstrates that the bi-stability phenomenon still exists in this setting and the system spends a larger proportion of time around the higher equilibrium level as the sensitivity parameter increases.

Figure 6.7: Sample paths of the number of people in the system with different sensitivity parameters, b ($n = 214$, $\lambda = 200$, $\theta = 0.3$ and $\mu_i = 0.6 + 0.4\exp(-bw_i)$.)



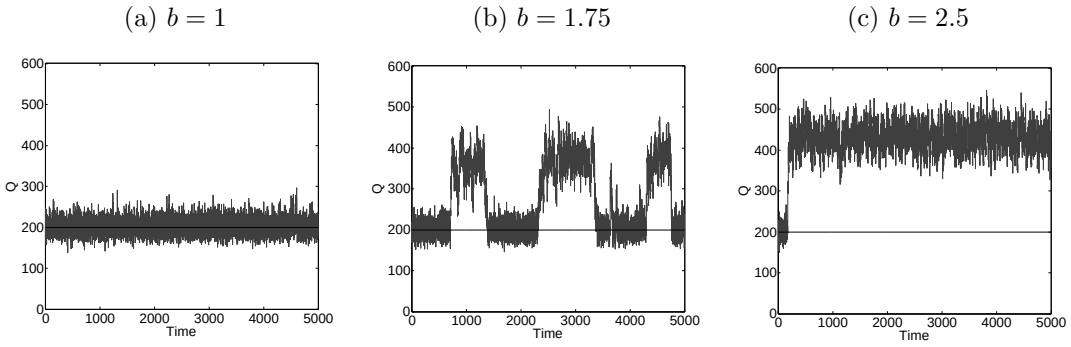
6.5.2 Agent-driven slowdown effect with delay

Another possible cause for the slowdown effect is due to fatigue of agents. Under high congestion levels, agents are working under pressure and without proper rest, which may eventually lead to deterioration in productivity. One way to model a slowdown effect caused by fatigue is to incorporate a time lag between the occurrence of high congestion levels and the deterioration in productivity. To capture this, we set the service rate as a function of the average queue length process over a time interval of length l. Specifically, the service rate at time t is

$$\mu\left(\int_{t-l}^{t} \frac{(Q(u)-n)^+}{n} du\right).$$

We observe that bi-stability still exists in this case. We find that the length of the time lag, l, affects the frequency at which the system moves between the two equilibria. If l is of the same order as the service time, the system moves "easily" from one equilibria to the other; as the time lag increases, it becomes "harder" for the system to transfer between them. Figure 6.8 illustrates how the bi-stability phenomenon evolves as the sensitivity parameter, b, increases, for a relatively small time lag (e.g., $l = 5$). As before, the proportion of time the system spends around the higher equilibrium level grows with b. In contrast, for a large time lag (e.g., $l = 30$), the trajectory of the system depends largely on its initial position, and tends to stay around the initial equilibrium level for a very long period of time (potentially forever), regardless of the value of b. Figure 6.9 demonstrates the sample paths commonly observed in this case. If the system starts around the lower equilibrium, it keeps fluctuating around that level, whereas if the system starts around the higher equilibrium, then it stays there. These observations has operational implications—whereas systems with short l will move from the upper equilibrium to the lower equilibrium by itself, systems with long l may need external intervention to move from the upper equilibrium to the lower one.

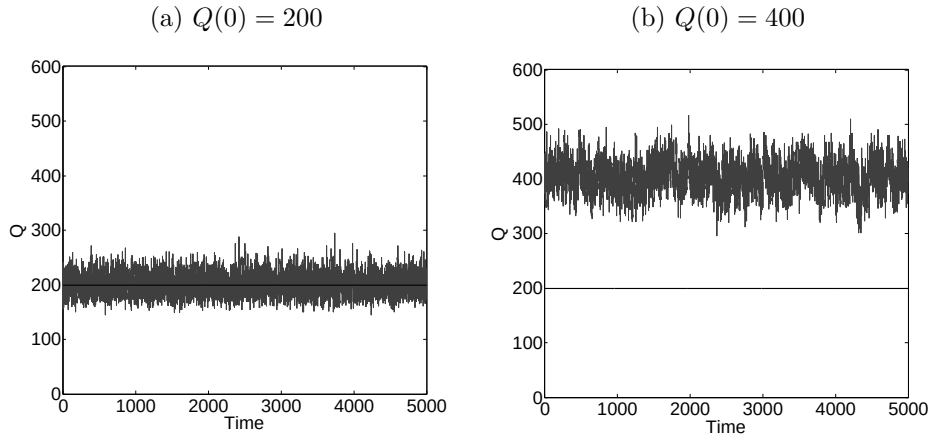
Figure 6.8: Sample paths of the number of people in the system with time lag of length $l = 5$ and different levels of the sensitivity parameter, b ($n = 214$, $\lambda = 200$, $\theta = 0.3$ and $\mu\left(\int_{t-l}^{t}(Q(u)-n)^{+}/n\,du\right) = 0.6 + 0.4\exp\left(-b\int_{t-l}^{t}(Q(u)-s)^{+}/s du\right)$.)



6.6 Concluding Remarks

Motivated by empirical findings in service systems, we modified the Erlang-A model to account for the effect of the workload-dependent service rate. When the load sensitivity is low, with relation to the abandonment rate, we observe a small gap between the performance of the standard Erlang-A model and the load-sensitive model. The latter has lower quality of service. We show that this reduction in quality measures can be fixed by adjusting the square-root staffing rule parameter. When the load sensitivity is high, we observe a bi-stability phenomenon where the system alternates between two equilibria: one equilibrium results in a QED performance and another equilibrium results in an ED performance. We conduct a sensitivity analysis of the proportion of time the system spends around each equilibrium and propose three policies to avoid the occurrence of bi-stability: A) a permanent increase of staffing, B) increasing the abandonment rate once the queue length reaches a certain threshold and C) admission control, where customers are blocked as soon as the queue length reaches the threshold level. Policy A may "overstaff" to a QD performance while Policy B & C provide a QED performance at a "low cost". Lastly, we illustrate via numerical experiments that the bi-stability phenomenon remains in a larger class of

Figure 6.9: Sample paths of the number of people in the system with time lag of length $l = 30$ and different initial queue lengths ($s = 214$, $\lambda = 200$, $\theta = 0.3$ and $\mu\left(\int_{t-l}^{t}(Q(u)-s)^{+}/sdu\right) = 0.6 + 0.4\exp\left(-2\int_{t-l}^{t}(Q(u)-s)^{+}/sdu\right)$.)



load-sensitive service systems with different sources of the slowdown effect.

We would like to conclude with some remarks regarding the construction of the model. First, for the sake of simplicity, throughout the manuscript, we assume a decreasing and convex service rate function and a constant abandonment rate. The former derived the bi-stability results. We notice that meta-stability (multiple (semi-)stable equilibria) can arise for more general forms of the service rate function and load-dependent abandonment rate. However, most of the analyses in this paper (the fluid analysis and the asymptotic analysis of the stationary distribution) can be applied to the more general cases as well. Our second remark is on the practical estimation of the service rate function. From our analyses, it is apparent that to design service systems with a load-dependent slowdown effect, it is sufficient to accurately estimate the service rate function around zero, for most purposes. The derivative of the service rate function at zero is all that is needed to distinguish between the low and the high sensitivity cases, and to approximate the performance measures in the low sensitivity case. To implement Policy B and Policy C in the high sensitivity case,

it is sufficient to estimate the service rate function up to $O(1/\sqrt{n})$.

6.7 Proofs of the technical results

Proof.[Proof of Theorem 6.2.1.] The proof follows from the method outlined in [76]. We write

$$\begin{aligned} Q_n(t) &= Q_n(0) + A(\lambda_n t) - S\left(\int_0^t \mu\left(\frac{(Q_n(u)-n)^+}{n}\right)(Q_n(u)\wedge n)du\right) \\ &\quad -R\left(\theta\int_0^t (Q_n(u)-n)^+du\right) \\ &= Q_n(0) + M_{n,1}(t) - M_{n,2}(t) - M_{n,3}(t) \\ &\quad +\lambda_n t - \int_0^t \mu\left(\frac{(Q_n(u)-n)^+}{n}\right)(Q_n(u)\wedge n)du - \theta\int_0^t (Q_n(u)-n)^+du \end{aligned}$$

where

$$\begin{aligned} M_{n,1} &= A(\lambda_n t) - \lambda_n t \\ M_{n,2} &= S\left(\int_0^t \mu\left(\frac{(Q_n(u)-n)^+}{n}\right)(Q_n(u)\wedge n)du\right) \\ &\quad -\int_0^t \mu\left(\frac{(Q_n(u)-n)^+}{n}\right)(Q_n(u)\wedge n)du \\ M_{n,3} &= R\left(\theta\int_0^t (Q_n(u)-n)^+du\right) - \theta\int_0^t (Q_n(u)-n)^+du. \end{aligned}$$

Let $\bar{Q}_n(t) = Q_n(t)/n$ and $\bar{M}_{n,i} = M_{n,i}/n$ for $i = 1, 2, 3$. Then

$$\begin{aligned} \bar{Q}_n(t) &= \bar{Q}_n(0) + \bar{M}_{n,1}(t) - \bar{M}_{n,2}(t) - \bar{M}_{n,3}(t) \\ &\quad +\frac{\lambda_n}{n}t - \int_0^t \mu\left((\bar{Q}_n(u)-1)^+\right)\left(\bar{Q}_n(u)\wedge 1\right)du - \theta\int_0^t \left(\bar{Q}_n(u)-1\right)^+du. \end{aligned}$$

Let $d(q) = -\mu((q-1)^+)(q\wedge 1) - \theta(q-1)^+$. As $\mu'(\cdot) \le 0$ and $\mu''(\cdot) \ge 0$, $|\mu'(x)| \le |\mu'(0)|$. It is easy to check that

$$|d(q_1) - d(q_2)| \le \max\{\mu(0), |\mu'(0)| + \theta\}|q_1 - q_2|.$$

Thus $d(\cdot)$ is Lipschitz. This implies that

$$q(t) = b + x(t) + \int_0^t d(q(u))du$$

has a unique solution and constitutes a function $\phi : \mathcal{D} \times R \to \mathcal{D}$ that is continuous (see Theorem 4.1 in [76]).

Let $\eta(t) \equiv 0$. We next show that $\bar{M}_{n,i} \to \eta$ in $\mathcal{D}$ w.p. 1 as $n \to \infty$ for $i = 1, 2, 3$. Applying the Functional Strong Law of Large Numbers to Poisson processes, we have $\sup_{0\le t\le T}\left\{\frac{A(nt)}{n} - t\right\} \to 0$, $\sup_{0\le t\le T}\left\{\frac{S(nt)}{n} - t\right\} \to 0$ and $\sup_{0\le t\le T}\left\{\frac{R(nt)}{n} - t\right\} \to 0$ w.p. 1 as $n \to \infty$ for any $T > 0$. We thus have

$$\bar{M}_{n,1} \to \eta \text{ in } D \text{ w.p. } 1 \text{ as } n \to \infty.$$

As $Q_n(t) < Q_n(0) + A(\lambda_n t)$, $\int_0^t Q_n(u)du \le t\,(Q_n(0) + A(\lambda_n t))$. This implies that for any fixed $T > 0$ there exists $\tau > 0$, such that

$$P\left(\frac{\mu(0)}{n}\int_0^T Q_n(u)du > \tau\right) \to 0 \text{ as } n \to \infty.$$

Then,

$$P\left(\left\|\bar{M}_{n,2}\right\|_T > \epsilon\right) \le P\left(\frac{\mu(0)}{n}\int_0^T Q_n(u)du > \tau\right) + P\left(\left\|\frac{S(nt)}{n} - t\right\|_\tau > \frac{\epsilon}{2}\right).$$

This leads to

$$\bar{M}_{n,2} \to \eta \text{ in } \mathcal{D} \text{ w.p. } 1 \text{ as } n \to \infty.$$

Similarly we can show that

$$\bar{M}_{n,3} \to \eta \text{ in } \mathcal{D} \text{ w.p. } 1 \text{ as } n \to \infty.$$

By the Continuous Mapping Theorem (CMT) we have the fluid limit in Theorem 6.2.1. □

Proof.[Proof of Theorem 6.2.4.] We prove asymptotic stability by the Lyapunov method. Specifically, a function $V(q) : \mathbb{R}^+ \to \mathbb{R}^+$ is called a Lyapunov function of (6.2) about its equilibrium $\bar{q}$ if $V(\bar{q}) = 0$ and $V(q) > 0$, $0 < |q - \bar{q}| < \delta$ for some $\delta > 0$. We denote $\dot{V}$ as the derivative of $V(\cdot)$ with respect to q. $\bar{q}$ is **locally asymptotically stable**, if there exists a Lyapunov function $V(q)$, such that $\dot{V}(q) < 0$ for all $0 < |q - \bar{q}| < \delta$ for some $\delta > 0$. $\bar{q}$ is **globally asymptotically stable**, if the locally asymptotically stable conditions hold for all $\delta \in \mathbb{R}^+$.

For the low sensitivity case, we use the following Lyapunov function

$$V(q) = |q - \bar{q}|,$$

where $\bar{q}$ is the specified equilibrium. Hence,

$$\dot{V}(q) = \text{sign}(q - \bar{q}) f(q).$$

Recall that $f(q) = \mu(0) - \mu((q-1)^+)(q \wedge 1) - \theta(q-1)^+$.
Under Assumption 6.1.1 and the assumptions of the low sensitivity case, $f(\cdot)$ is decreasing. $\bar{q} = \mu(0)/\mu(0) = 1$ and

$$\dot{V}(q) = \begin{cases} -\mu(0) + \mu(0)q < -\mu(0) + \mu(0)\bar{q} = 0, & q < \bar{q}; \\ \mu(0) - \mu(q-1) - \theta(q-1) < \mu(0) - \mu(0) = 0, & q > \bar{q}. \end{cases}$$

Therefore, $\bar{q}$ is a globally asymptotically stable equilibrium.

Under Assumption 6.1.1 and the assumptions of the high sensitivity case, $f(1) = 0$; thus $\bar{q}_1 = 1$. $f(q)$ is increasing on $[\bar{q}_1, \hat{q})$ and decreasing on $[\hat{q}, \infty)$. Since $f(\hat{q}) > 0$ and $\lim_{q\to\infty} f(q) = -\infty$, there exists $\bar{q}_2 > \hat{q}$ such that $f(\bar{q}_2) = 0$.

As $f(q) > 0$ for $q < 1$ and $f(q) > 0$ for $1 < q < \hat{q}$, $\bar{q}_1$ is semistable.

Let

$$V_2(q) = |q - \bar{q}_2|.$$

For $q \in (\bar{q}_1, \infty)$,

$$\dot{V}_2(q) = \begin{cases} -\mu(0) + \mu(q-1) + \theta(q-1) < -\mu(0) + \mu(0)\bar{q}_1 = 0, \\ \text{when } \bar{q}_1 < q \le \hat{q}; \\ -\mu(0) + \mu(q-1) + \theta(q-1) < -\mu(0) + \mu(\bar{q}_2 - 1) + \theta(\bar{q}_2 - 1) = 0, \\ \text{when } \hat{q} < q < \bar{q}_2; \\ \mu(0) - \mu(q-1) - \theta(q-1) < \mu(0) - \mu(\bar{q}_2 - 1) - \theta(\bar{q}_2 - 1) = 0, \\ \text{when } q > \bar{q}_2. \end{cases}$$

Therefore, $\bar{q}_2$ is a locally asymptotically stable equilibrium. □

In order to prove Theorem 6.3.1, we start with the following lemma.

Lemma 6.7.1 *Assume* $\sqrt{n}(1-\lambda_n/(n\mu(0))) \to \beta$ *as* $n \to \infty$. *For any* $0 < y_1 < y_2 < \infty$,

$$\lim_{n\to\infty} \log \frac{\pi_n(\lfloor n+\sqrt{n}y_2 \rfloor)}{\pi_n(\lfloor n+\sqrt{n}y_1 \rfloor)} = -\int_{y_1}^{y_2} \beta + \frac{\mu'(0)+\theta}{\mu(0)} y \, dy$$

and

$$\lim_{n\to\infty} \log \frac{\pi_n(\lfloor n-\sqrt{n}y_1 \rfloor)}{\pi_n(\lfloor n-\sqrt{n}y_2 \rfloor)} = -\int_{-y_2}^{-y_1} \beta + y \, dy.$$

Proof. From the detailed balance equation of the B&D process, we have

$$\frac{\pi_n(\lfloor n+\sqrt{n}y_2 \rfloor)}{\pi_n(\lfloor n+\sqrt{n}y_1 \rfloor)} = \prod_{k=\lfloor n+\sqrt{n}y_1 \rfloor+1}^{\lfloor n+\sqrt{n}y_2 \rfloor} \frac{\lambda_n}{\mu\left((k-n)/n\right) n + \theta(k-n)}.$$

Then,

$$\begin{aligned}
& \log \frac{\pi_n(\lfloor n+\sqrt{n}y_2 \rfloor)}{\pi_n(\lfloor n+\sqrt{n}y_1 \rfloor)} \\
= \;& \lfloor (y_2-y_1)\sqrt{n} \rfloor \log \rho_n - \sum_{k=\lfloor \sqrt{n}y_1 \rfloor+1}^{\lfloor \sqrt{n}y_2 \rfloor} \log\left(1 + \frac{\mu(\frac{k}{n}) - \mu(0) + \theta\frac{k}{n}}{\mu(0)}\right) \\
= \;& -\lfloor (y_2-y_1)\sqrt{n} \rfloor (1-\rho_n) - \sum_{k=\lfloor \sqrt{n}y_1 \rfloor+1}^{\lfloor \sqrt{n}y_2 \rfloor} \frac{\mu'(0)+\theta}{\mu(0)} \frac{k}{\sqrt{n}} \frac{1}{\sqrt{n}} + O\left(\frac{1}{n}\right) \\
\to \;& -(y_2-y_1)\beta - \int_{y_1}^{y_2} \frac{\mu'(0)+\theta}{\mu(0)} y dy.
\end{aligned}$$

Likewise,

$$\begin{aligned}
& \log \frac{\pi_n(\lfloor n-\sqrt{n}y_1 \rfloor)}{\pi_n(\lfloor n-\sqrt{n}y_2 \rfloor)} \\
= \;& \lfloor (y_2-y_1)\sqrt{n} \rfloor \log \rho_n - \sum_{k=\lfloor \sqrt{n}y_1 \rfloor+1}^{\lfloor \sqrt{n}y_2 \rfloor} \log(1-\frac{k}{n}) \\
= \;& -\lfloor (y_2-y_1)\sqrt{n} \rfloor (1-\rho_n) - \sum_{k=\lfloor \sqrt{n}y_1 \rfloor+1}^{\lfloor \sqrt{n}y_2 \rfloor} -\frac{k}{\sqrt{n}} \frac{1}{\sqrt{n}} + O\left(\frac{1}{n}\right) \\
\to \;& -(y_2-y_1)\beta - \int_{y_1}^{y_2} -y dy.
\end{aligned}$$

□

Proof.[Proof of Theorem 6.3.1.] The technique used in this proof follows from [77]. We denote G_n as the cumulative distribution function (CDF) of the scaled process

Y_n. We first prove the relative compactness of G_n by a sandwich argument using stochastic comparison.

Let $\{Q_n^l(t)\}$ and $\{Q_n^u(t)\}$ denote the queue length processes of two sequences of ordinary Erlang-A queues: both have n servers and arrival rate λ_n, which are the same as the original process $Q_n(t)$. We keep the service rate and the abandonment rate fixed regardless of the system scale. The service rates of both systems are fixed at $\mu(0)$. The abandonment rate of $Q_n^l(t)$ is θ whereas the abandonment rate of $Q_n^u(t)$ is $\theta + \mu'(0)$.

As

$$\mu\left(\frac{q-n}{n}\right)n + \theta(q-n) \leq \mu(0)n + \theta(q-n)$$

and

$$\begin{aligned} & \mu\left(\frac{q-n}{n}\right)n + \theta(q-n) \\ = \;& \mu(0)n + (\mu'(0)+\theta)(q-n) + \mu''(\eta)\frac{(q-n)^2}{n} \text{ for some } \eta \in (0,(q-n)/n) \\ \geq \;& \mu(0)n + (\mu'(0)+\theta)(q-n), \end{aligned}$$

based on Lemma 6.4.6, we have

$$P(Q_n(\infty) > q) \geq P(Q_n^l(\infty) > q)$$

and

$$P(Q_n(\infty) > q) \leq P(Q_n^u(\infty) > q).$$

Following the definition of Y_n, we let $Y_n^l := (Q_n^l(\infty) - n)/\sqrt{n}$ and $Y_n^u := (Q_n^u(\infty) - n)/\sqrt{n}$. We also denote G_n^l and G_n^u as the CDFs of the scaled processes X_n^l and X_n^u, respectively. Then both G_n^u and G_n^l converge uniformly to some limiting distributions [77]. We denote their limits as G^l and G^u respectively. Since $G_n^u(y) \leq G_n(y) \leq G_n^l(y)$, we have that for any $\epsilon > 0$, there exists a small enough y such that

$$\limsup_{n\to\infty} G_n(y) \leq \lim_{n\to\infty} G^l(y) < \epsilon$$

and

$$1 \geq \lim_{y\to\infty} \liminf_{n\to\infty} G_n(y) \geq \lim_{y\to\infty} \lim_{n\to\infty} G_n^u(y) = 1.$$

Thus, G_n is relatively compact. The limit of G_n exists and is a well-defined CDF. From Lemma 7.2.3, the distribution G is absolutely continuous with probability density function of the form

$$g(y) = \begin{cases} \frac{C_1}{\sqrt{2\pi}} \exp\left(-\frac{(y+\beta)^2}{2}\right) & \text{if } y < 0, \\ \frac{C_2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y+\beta\sigma^2)^2}{2\sigma^2}\right) & \text{if } y \geq 0, \end{cases}$$

where $\sigma = \sqrt{\mu(0)/\left(\mu'(0)+\theta\right)}$, and C_1 and C_2 are the normalizing constants. Using the fact that $\int_{-\infty}^{\infty} g(y)\,dy = 1$ and $g(y)$ is continuous at 0, we have

$$C_1 = \frac{h(\beta\sigma)}{\sigma\phi(\beta)} \left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1},$$

and

$$C_2 = \frac{h(\beta\sigma)}{\phi(\beta\sigma)} \left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}.$$

□

Proof.[Proof of Corollary 6.3.2.] As $P_n(W) = P(Q_n(\infty) \geq n) = P(Y_n \geq 0)$, and

$$\lim_{n\to\infty} P(Y_n \geq 0) = C_2\bar{\Phi}(\beta\sigma) = \left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1},$$

where $\sigma = \sqrt{\mu(0)/\left(\mu'(0)+\theta\right)}$. We thus have the desired limit for $P_n(W)$.
For $P_n(Ab)$, we have

$$\begin{aligned} \sqrt{n}P_n(Ab) &= E[(Q_n(\infty) - n)^+]\frac{\theta\sqrt{n}}{\lambda_n} \\ &= E[Y_n | Y_n \geq 0]P_n(W)\frac{\theta n}{\lambda_n}. \end{aligned}$$

As $\lim_{n\to\infty} E[Y_n|Y_n \geq 0] = \sigma h(\beta\sigma) - \beta\sigma^2$,

$$\begin{aligned} \lim_{n\to\infty} \sqrt{n}P_n(Ab) &= \left(\sigma h(\beta\sigma) - \beta\sigma^2\right)\left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}\frac{\theta}{\mu(0)} \\ &= \left(\frac{h(\beta\sigma)}{\sigma} - \beta\right)\left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}\frac{\theta}{\mu'(0)+\theta}. \end{aligned}$$

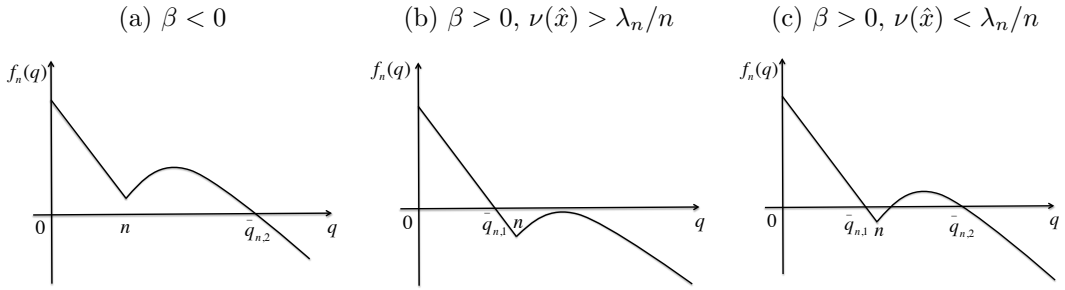
□

Proof.[Proof of Theorem 6.4.1.] We first analyze the roots of $f_n(q) = \lambda_n - \mu((q-n)^+/n)(q \wedge n) - \theta(q-n)^+ = 0$. We divide the analysis into two regions: $[0, n]$ and $[n, \infty)$. $f_n(q)$ is linearly decreasing on $[0, n]$ with $f_n(0) = \lambda_n > 0$, $f_n(n) = \lambda_n - \mu(0)n$. When $\beta < 0$, $f_n(n) > 0$. When $\beta > 0$, $f_n(n) < 0$ and $f_n(\bar{q}_{n,1}) = 0$, where $\bar{q}_{n,1} = \lambda_n/\mu(0) < n$. For $q > n$, we analyze the scaled function

$$\frac{f_n(q)}{n} = \frac{\lambda_n}{n} - \nu(x_n)$$

where $\nu(x_n) = \mu(x_n) + \theta x_n$, and $x_n = (q-n)/n$. Recall that $\nu(\cdot)$ is convex and attains its minimum at $\hat{x}$. Specifically, $\nu(\cdot)$ is decreasing on $[0, \hat{x}]$ and increasing on $(\hat{x}, \infty)$ with $\nu(x) \to \infty$ as $x \to \infty$. When $\beta < 0$, as $\nu(0) < \lambda_n/n$, there exists a unique $\bar{x}_n > \hat{x}$, such that $\lambda_n/n = \nu(\bar{x}_n)$. This implies that $f_n((\bar{x}_n + 1)n) = 0$. When $\beta > 0$, since $\nu(0) > \lambda_n/n$, we have two cases: if $\nu(\hat{x}) > \lambda_n/n$, then $f_n(q) < 0$ for all $q > n$; otherwise, $\nu(\hat{x}) < \lambda_n/n$ and there exists a unique $0 < \tilde{x}_n < \hat{x}$, such that $\lambda_n/n = \nu(\tilde{x}_n)$, and a unique $\bar{x}_n > \hat{x}$, such that $\lambda_n/n = \nu(\bar{x}_n)$. This implies that $f_n((\tilde{x}_n+1)n) = f_n((\bar{x}_n+1)n) = 0$. See Figure 6.10 for a graphical illustration. Based

Figure 6.10: $f_n(q)$ with positive or negative βs



on the above analysis, we have:

1. When $\beta < 0$, $f_n(q) \geq 0$ for $q \leq (\bar{x}_n + 1)n$, and $f_n(q) < 0$ for $q > (\bar{x}_n + 1)n$. Therefore, $\pi_n(\cdot)$ has only one peak, $\bar{q}_{n,2} = \lfloor (\bar{x}_n + 1)n \rfloor$.

2. When $\beta > 0$ and $\nu(\hat{x}) > \lambda_n/n$, $f_n(q) > 0$ for $q \leq \lambda_n/\mu(0)$, and $f_n(q) < 0$ for $q > \lambda_n/\mu(0)$. Therefore, $\pi_n(\cdot)$ has only one peak, $\bar{q}_{n,1} = \lfloor \lambda_n/\mu(0) \rfloor$.

3. When $\beta > 0$ and $\nu(\hat{x}) < \lambda_n/n$: a.) $f_n(q) \geq 0$ on $[0, \lambda_n/\mu(0)]$, $f_n(q) < 0$ on $(\lambda_n/\mu(0), (\hat{x}_n+1)n)$, therefore, $\pi_n(\cdot)$ has the first peak at $\bar{q}_{n,1} = \lfloor \lambda_n/\mu(0) \rfloor$; and b.) as $f_n(q) \geq 0$ on $[(\hat{x}_n+1)n, (\bar{x}_n+1)n]$, $f_n(q) \geq 0$ on $((\bar{x}_n+1)n, \infty)$, $\pi_n(\cdot)$ has the second peak at $\bar{q}_{n,2} = \lfloor (\bar{x}_n+1)n \rfloor$.

□

Proof.[Proof of Theorem 6.4.2.] We first establish some asymptotic results about the value of the peaks, $\bar{q}_{n,1}$ and $\bar{q}_{n,2}$. From Theorem 6.4.1, we have

$$\frac{n-\bar{q}_{n,1}}{\sqrt{n}} = \sqrt{n}\left(1 - \frac{\lambda_n}{n\mu(0)}\right) + O\left(\frac{1}{\sqrt{n}}\right) \to \beta \text{ as } n \to \infty.$$

As $\lim_{n\to\infty} \lambda_n/n = \mu(0)$ and $\nu(x)$ is continuously decreasing on $(\hat{x}, \infty)$,

$$\frac{\bar{q}_{n,2}-n}{n} = \bar{x}_n + O(\frac{1}{n}) \to \bar{x} \text{ as } n \to \infty.$$

Using the detailed balance equation of the B&D process, we have

$$\begin{aligned}
& \pi_n(\bar{q}_{n,2}) \\
= \; & \pi_n(\bar{q}_{n,1}) \prod_{k=\bar{q}_{n,1}+1}^{\bar{q}_{n,2}} \frac{\lambda_n}{\mu((k-n)^+/n)(k \wedge n) + \theta(k-n)^+} \\
= \; & \pi_n(\bar{q}_{n,1}) \exp\left((\bar{q}_{n,2} - \bar{q}_{n,1}) \log \frac{\lambda_n}{n} - \sum_{k=\bar{q}_{n,1}+1}^{n-1} \log\left(\mu(0)\frac{k}{n}\right) - \sum_{k=n}^{\bar{q}_{n,2}} \log \nu\left(\frac{q-n}{n}\right)\right).
\end{aligned}$$

Then,

$$\begin{aligned}
& \frac{1}{n} \log \frac{\pi_n(\bar{q}_{n,2})}{\pi_n(\bar{q}_{n,1})} \\
= \; & \frac{\bar{q}_{n,2} - \bar{q}_{n,1}}{n} \log \frac{\lambda_n}{n} - \frac{n - \bar{q}_{n,1}}{n} \log \mu(0) + O\left(\frac{1}{\sqrt{n}}\right) - \frac{n}{n} \sum_{k=0}^{\bar{x}_n n} \log \nu(\frac{k}{n}) \frac{1}{n} \\
\to \; & \bar{x} \log \mu(0) - \int_0^{\bar{x}} \log \nu(x) dx.
\end{aligned}$$

□

Proof.[Proof of Lemma 6.4.4.] By Theorem 6.2.4, under high sensitivity conditions, $\bar{q}_2 > 1$ is the unique root of $f(q) = \mu(0) - \mu((q-1)^+) - \theta(q-1)^+$ on $(1, \infty)$. We

establish the results of this lemma by comparing pairs of systems, (1) and (2); we denote the higher level equilibrium as $\bar{q}_2^{(1)}$ and $\bar{q}_2^{(2)}$ for the two systems, respectively. For each part of the lemma we differ the two systems by two values of a specific system parameter.

i) Keep all other system parameters equal and vary the service rate function $\mu(\cdot)$, such that $\mu_{(2)}(\cdot)$ is more sensitive than $\mu_{(1)}(\cdot)$. Then, we have

$$0 = \mu_{(1)}(0) - \mu_{(1)}\left(\bar{q}_2^{(1)} - 1\right) - \theta\left(\bar{q}_2^{(1)} - 1\right) \leq \mu_{(2)}(0) - \mu_{(2)}\left(\bar{q}_2^{(1)} - 1\right) - \theta\left(\bar{q}_2^{(1)} - 1\right).$$

As $\mu_{(2)}(0) - \mu_{(2)}(q-1) - \theta(q-1)$ is nonnegative on $[1, \bar{q}_2^{(2)}]$ and strictly negative on $(\bar{q}_2^{(2)}, \infty)$, $\bar{q}_2^{(2)} \geq \bar{q}_2^{(1)}$.

ii) Keep all other system parameters equal and vary the abandonment rate θ, such that $\theta_{(1)} < \theta_{(2)}$. Then,

$$0 = \mu(0) - \mu\left(\bar{q}_2^{(2)} - 1\right) - \theta_{(2)}\left(\bar{q}_2^{(2)} - 1\right) < \mu(0) - \mu\left(\bar{q}_2^{(2)} - 1\right) - \theta_{(1)}\left(\bar{q}_2^{(2)} - 1\right).$$

Following the same rationale as in part i), we have $\bar{q}_2^{(1)} > \bar{q}_2^{(2)}$.

□

Proof.[Proof of Lemma 6.4.5.] The proof of Lemma 6.4.5 follows the same strategy as the proof of Lemma 6.4.4. Specifically, we compare pairs of systems, (1) and (2). For each part of the lemma, we differ the two system by two values of a specific system parameter.

i) Keep all other system parameters equal and vary the service rate function $\mu(\cdot)$, such that $\mu_{(2)}(\cdot)$ is more sensitive than $\mu_{(1)}(\cdot)$. From Lemma 6.4.4, we have $\bar{x}^{(1)} \leq \bar{x}^{(2)}$. As $\mu_{(1)}(x) \geq \mu_{(2)}(x)$,

$$I_{(1)}(\bar{x}^{(1)}) = \int_0^{\bar{x}^{(1)}} \log \frac{\mu_{(1)}(0)}{\mu_{(1)}(x) + \theta x}\, dx \leq \int_0^{\bar{x}^{(2)}} \log \frac{\mu_{(2)}(0)}{\mu_{(2)}(x) + \theta x}\, dx = I_{(2)}(\bar{x}^{(2)})$$

ii) Keep all other system parameters equal and vary the abandonment rate θ, such that $\theta_{(1)} < \theta_{(2)}$. From Lemma 6.4.4, we have $\bar{x}^{(1)} > \bar{x}^{(2)}$. Then we have

$$I_{(1)}(\bar{x}^{(1)}) = \int_0^{\bar{x}^{(1)}} \log \frac{\mu(0)}{\mu(x) + \theta_{(1)}x}\, dx > \int_0^{\bar{x}^{(2)}} \log \frac{\mu(0)}{\mu(x) + \theta_{(2)}x}\, dx = I_{(2)}(\bar{x}^{(2)}).$$

□

Proof.[Proof of Lemma 6.4.6.] We prove the theorem by first introducing a coupling, under which the entire sample path of $Y^{(1)}$ and $Y^{(2)}$ are ordered, i.e.

$$P(Y^{(1)}(t) \leq Y^{(2)}(t) \text{ for all } t \geq 0) = 1.$$

Fix $\tilde{Y}^{(1)}(0) = \tilde{Y}^{(2)}(0) = y_0$ for any $y_0 \in \mathbb{Z}^+$. The coupling argument uses the thinning property of Poisson process and goes as follows. When $(\tilde{Y}^{(1)}(t), \tilde{Y}^{(2)}(t)) = (y_1, y_2)$ We generate the next potential transition by an exponential random variable with rate $\gamma_1 + \xi_1(y_1) \vee \xi_2(y_2)$. We then generate a uniform random variable independent of everything else. If $U \leq \gamma_1/(\gamma_1 + \xi_1(y_1) \vee \xi_2(y_2))$, we treat it as an arrival to both $\tilde{Y}^{(1)}$ and $\tilde{Y}^{(2)}$; else if $U \leq (\gamma_1 + \xi_1(y_1) \wedge \xi_1(y_2))/(\gamma_1 + \xi_1(y_1) \vee \xi_2(y_2))$, we treat it as a departure for both processes; else we impose a departure on $\tilde{Y}^{(i)}$ with the larger departure rate only. As when $y_1 = y_2$, we always have $\xi_1(y_1) \geq \xi_2(y_2)$, under this coupling $\tilde{Y}^{(1)}(t) \leq \tilde{Y}^{(2)}(t)$, for all $t \geq 0$, path by path. Let $P_{y_0}(\cdot) := P(\cdot | Y^{(1)} = y_0, Y^{(2)} = y_0)$. Then we have

$$\begin{aligned} P_{y_0}(Y^{(1)}(t) > y) &= P_{y_0}(\tilde{Y}^{(1)}(t) > y, \tilde{Y}^{(1)}(t) < \tilde{Y}^{(2)}(t)) \\ &\leq P_{y_0}(\tilde{Y}^{(2)}(t)) > y) = P_{y_0}(Y^{(2)}(t) > y) \end{aligned}$$

for any $t \geq 0$.

As $\lim_{t\to\infty} P_{y_0}(Y^{(i)}(t) > y) = P(Y^{(i)}(\infty) > y)$,$i = 1, 2$, for all $y_0 \in \mathbb{Z}^+$, is well-defined, and $Y^{(1)}$ and $Y^{(2)}$ live on the same state space, $P(Y^{(1)}(\infty) > y) \leq P(Y^{(2)}(\infty) > y)$.

□

Before we prove Theorem 6.4.7, we first prove the following lemma as a preparation.

Lemma 6.7.2 *Under High Sensitivity and SRS with $\beta > 0$,*

i) the larger the value of the SRS parameter β is, the larger the value of $\tilde{q}_n$;

ii) under Definition 6.4.3, the more load sensitive the service rate function is, the smaller the value of $\tilde{q}_n$;

iii) the larger the abandonment rate θ is, the smaller the value of $\tilde{q}_n$.

Proof. The proof of Lemma 6.7.2 follows the same strategy as the proof of Lemma 6.4.4. Specifically, we compare pairs of systems, (1) and (2). For each part of the lemma, we differ the two system by two values of a specific system parameter.

i) Keep all other system parameters equal and vary the staffing parameter β, such that $\beta_{(1)} < \beta_{(2)}$. Denote $n_{(1)} = R_n + \beta_{(1)}\sqrt{R_n}$ and $n_{(2)} = R_n + \beta_{(2)}\sqrt{R_n}$. Then $n_{(1)} < n_{(2)}$ and

$$0 = \lambda_n/n_{(1)} - \mu(\tilde{x}_n^{(1)}) - \theta\tilde{x}_n^{(1)} > \lambda_n/n_{(2)} - \mu(\tilde{x}_n^{(1)}) - \theta\tilde{x}_n^{(1)}.$$

As $\lambda_n/n_{(2)} - \mu(x) - \theta x$ is increasing on $[0, \hat{x}]$ and is nonpositive on $[0, \tilde{x}_n^{(2)}]$, $\tilde{x}_n^{(2)} > \tilde{x}_n^{(1)}$. Thus, $\tilde{q}_n^{(2)} = \lfloor(\tilde{x}_n^{(2)} + 1)n_{(2)}\rfloor > \tilde{q}_n^{(1)} = \lfloor(\tilde{x}_n^{(1)} + 1)n_{(1)}\rfloor$

ii) Keep all other system parameters equal and vary the service rate function $\mu(\cdot)$, such that $\mu_{(2)}(\cdot)$ is more sensitive than $\mu_{(1)}(\cdot)$. Then, we have

$$0 = \frac{\lambda_n}{n} - \mu_{(2)}\left(\tilde{x}_n^{(2)}\right) - \theta\tilde{x}_n^{(2)} \geq \frac{\lambda_n}{n} - \mu_{(1)}\left(\tilde{x}_n^{(2)}\right) - \theta\tilde{x}_n^{(2)}.$$

Following the same rationale as in part i), we have $\tilde{q}_n^{(1)} \geq \tilde{q}_n^{(2)}$.

ii) Keep all other system parameters equal and vary the abandonment rate θ, such that $\theta_{(1)} < \theta_{(2)}$. Then,

$$0 = \frac{\lambda_n}{n} - \mu\left(\tilde{x}_n^{(1)}\right) - \theta_{(1)}\tilde{x}_n^{(1)} > \frac{\lambda_n}{n} - \mu\left(\tilde{x}_n^{(1)}\right) - \theta_{(2)}\tilde{x}_n^{(1)}.$$

Following the same rationale as in part i), we have $\tilde{q}_n^{(2)} > \tilde{q}_n^{(1)}$.

□

Proof.[Proof of Theorem 6.4.7.] We prove Theorem 6.4.7 by comparing the death rates of pairs of systems denoted by $Q^{(1)}$ and $Q^{(2)}$.

i) Keeping all other parameters equal, for $\beta_{(1)} < \beta_{(2)}$, we denote $n_{(1)} = R+\beta_{(1)}\sqrt{R}$, $n_{(2)} = R + \beta_{(2)}\sqrt{R}$ where $R = \lambda/\mu(0)$. Then when $q \le n_{(1)}$, the death rates of the two systems are equal; when $n_{(1)} < q \le n_{(2)}$,

$$\mu(0)q - \left(\mu\left(\frac{q}{n_{(1)}} - 1\right) n_{(1)} + \theta(q - n_{(1)})\right) \ge (\mu(0) - \theta)(q - n_{(1)}) \ge 0;$$

when $q > n_{(2)}$

$$\begin{aligned} & \left(\mu\left(\frac{q}{n_{(2)}} - 1\right) n_{(2)} + \theta(q - n_{(2)})\right) - \left(\mu\left(\frac{q}{n_{(1)}} - 1\right) n_{(1)} + \theta(q - n_{(1)})\right) \\ = \; & \left(\mu\left(\frac{q}{n_{(2)}} - 1\right) - \mu\left(\frac{q}{n_{(1)}} - 1\right)\right) n_{(2)} \\ & +\mu\left(\frac{q}{n_{(1)}} - 1\right)(n_{(2)} - n_{(1)}) - \theta(n_{(2)} - n_{(1)}) \\ \ge \; & -\mu'\left(\frac{q}{n_{(1)}} - 1\right)\frac{(n_{(2)} - n_{(1)})q}{n_{(1)}} + (\mu(\infty) - \theta)(n_{(2)} - n_{(1)}) \ge 0. \end{aligned}$$

Then

$$P(Q^{(1)}(\infty) > \tilde{q}_n^{(1)}) \ge P(Q^{(2)}(\infty) > \tilde{q}_n^{(1)}) \ge P(Q^{(2)}(\infty) > \tilde{q}_n^{(2)}),$$

where the first inequality follows from Lemma 6.4.6 and the second inequality follows from Lemma 6.7.2.

ii) Keeping all other parameters equal, for system (2) more sensitive than system (1), we have $\mu_{(1)}((q-n)^+/n)(q\wedge n)+\theta(q-n)^+ \ge \mu_{(2)}((q-n)^+/n)(q\wedge n)+\theta(q-n)^+$ for all $q \ge 0$. Then $P(Q^{(1)}(\infty) > \tilde{q}_n^{(1)}) \le P(Q^{(2)}(\infty) > \tilde{q}_n^{(1)}) \le P(Q^{(2)}(\infty) > \tilde{q}_n^{(2)})$.

iii) Keeping all other parameters equal, for $\theta_{(1)} < \theta_{(2)}$, we have $\mu((q-n)^+/n)(q \wedge n) + \theta_{(1)}(q-n)^+ \le \mu((q-n)^+/n)(q \wedge n) + \theta_{(2)}(q-n)^+$ for all $q \ge 0$. Then $P(Q^{(1)}(\infty) > \tilde{q}_n^{(1)}) \ge P(Q^{(2)}(\infty) > \tilde{q}_n^{(1)}) \ge P(Q^{(2)}(\infty) > \tilde{q}_n^{(2)})$.

□

Proof.[Proof of Lemma 6.4.8.] Let $\psi_n(x) = \lambda_n/n - \mu(x) - \theta x$ for $x \geq 0$. Then $\tilde{x}_n$ is the unique root of $\psi_n(x) = 0$ on $[0, \hat{x}]$. Since $\lambda_n/n \to \mu(0)$ as $n \to \infty$ and $\psi_n(\cdot)$ is continuous and monotonically increasing on $[0, \hat{x}]$, $\tilde{x}_n \to 0$ as $n \to \infty$. Applying Taylor expansion to $\mu(\cdot)$, we have

$$\psi_n(\tilde{x}) = \lambda_n/n - \mu(0) - (\mu' + \theta)\tilde{x}_n + O(\tilde{x}_n^2) = 0.$$

Then

$$\frac{\mu(0) - \lambda_n/n}{\tilde{x}_n} \to -\mu'(0) + \theta \text{ as } n \to \infty.$$

As $\sqrt{n}(1 - \lambda_n/(n\mu(0))) \to \beta$, $\tilde{x}_n = O(1/\sqrt{n})$. We then have

$$\sqrt{n}\tilde{x}_n = -\frac{\sqrt{n}(\mu(0) - \lambda_n/n)}{\mu'(0) + \theta} + O\left(\frac{1}{\sqrt{n}}\right).$$

Thus, $\sqrt{n}\tilde{x}_n \to -\mu(0)\beta/(\mu'(0) + \theta)$ as $n \to \infty$. □

Proof.[Proof of Theorem 6.4.9.] The proof of Theorem 6.4.9 also follows from the method outlined in [76]. We use both the Functional Central Limit Theorem (FCLT) and CMT. We again write

$$\begin{aligned}
Q_n^c(t) &= Q_n^c(0) + A(\lambda_n t) - S\left(\int_0^t \mu\left(\frac{(Q_n^c(u) - n)^+}{n}\right)(Q_n^c(u) \wedge n)du\right) \\
&\quad -R\left(\theta\int_0^t (Q_n^c(u) - n)^+ du\right) - L_n(t) \\
&= Q_n^c(0) + M_{n,1}(t) - M_{n,2}(t) - M_{n,3}(t) - L_n(t) \\
&\quad +\lambda_n t - \int_0^t \mu\left(\frac{(Q_n^c(u) - n)^+}{n}\right)(Q_n^c(u) \wedge n)du - \theta\int_0^t (Q_n^c(u) - n)^+ du
\end{aligned}$$

where

$$\begin{aligned}
M_{n,1} &= A(\lambda_n t) - \lambda_n t \\
M_{n,2} &= S\left(\int_0^t \mu\left(\frac{(Q_n^c(u) - n)^+}{n}\right)(Q_n^c(u) \wedge n)du\right) \\
&\quad - \int_0^t \mu\left(\frac{(Q_n^c(u) - n)^+}{n}\right)(Q_n^c(u) \wedge n)du \\
M_{n,3} &= R\left(\theta\int_0^t (Q_n^c(u) - n)^+ du\right) - \theta\int_0^t (Q_n^c(s) - n)^+ du.
\end{aligned}$$

Let $\hat{Q}_n^c(t) = (Q_n(t)-n)/\sqrt{n}$, $\hat{Y}_n(t) = Y_n(t)/\sqrt{n}$ and $\hat{M}_{n,i} = M_{n,i}/\sqrt{n}$ for $i = 1, 2, 3$. As $\hat{Q}_n^c(\cdot) < c_n$, $\hat{Q}_n^c(t) = O(\sqrt{n})$. Applying Taylor expansion, we have

$$\begin{aligned}\hat{Q}_n^c(t) &= \hat{Q}_n^c(0) + \hat{M}_{n,1}(t) - \hat{M}_{n,2}(t) - \hat{M}_{n,3}(t) - L_n(t) \\ &\quad + \frac{\lambda_n - \mu(0)n}{\sqrt{n}}t - \int_0^t \mu(0)(\hat{Q}_n^c(u) \wedge 0)du - \int_0^t \mu'(0)\hat{Q}_n^c(u)^+ du \\ &\quad - \int_0^t \theta \hat{Q}_n^c(u)^+ du + O\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Let $d(q) = -\mu'(0)(q \wedge 0) - (\mu'(0) + \theta)q^+$. Consider the integral representation

$$q(t) = b + x(t) + \int_0^t d(q(s))ds - l(t), \tag{6.6}$$

where $l(t)$ is a nondecreasing nonnegative function in D such that (6.6) holds and $\int_0^\infty 1\{q(t) < c\}dl(t) = 0$. As $d(\cdot)$ is Lipschitz, the integration (6.6) has a unique solution (q, y) and it constitutes a Bonafide function $(\phi_1, \phi_2) : \mathcal{D} \times R \to \mathcal{D} \times \mathcal{D}$ mapping (b, x) into (q, y). Moreover (ϕ_1, ϕ_2) is continuous (see Theorem 7.3 in [76]).

$\hat{M}_{n,i}$ are square-integrable martingales with respect to the filtration

$$\begin{aligned}\mathcal{F}_{n,t} := \sigma\{ \quad & Q_n(0), A(\lambda_n s), S\left(\int_0^s \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)(Q_n^c(u) \wedge n)du\right), \\ & R\left(\theta \int_0^t (Q_n^c(u)-n)^+ du\right) : 0 \le s \le t\}\end{aligned}$$

augmented by including all null sets. Also

$$\begin{aligned}\langle M_{n,1}\rangle(t) &= \frac{\lambda_n t}{n} \\ \langle M_{n,2}\rangle(t) &= \int_0^t \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)\frac{Q_n^c(u) \wedge n}{n}du \\ \langle M_{n,3}\rangle(t) &= \frac{\theta}{n}\int_0^t (Q_n^c(u)-n)^+ du.\end{aligned}$$

As

$$\frac{\lambda_n t}{n} \to \mu(0)t \text{ as } n \to \infty \text{ w.p. } 1,$$

$\{\langle M_{n,1}\rangle\}$ is stochastically bounded. By the crude bound $Q_n^c(s) < Q_n^c(0) + A(\lambda_n t)$, we have

$$\int_0^t \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)\frac{Q_n^c(u) \wedge n}{n}du \le \mu(0)t\left(\frac{Q_n^c(0)}{n} + \frac{A(\lambda_n t)}{n}\right).$$

Since $\{Q_n^c(0)/n\}$ and $\{A(\lambda_n t)/n\}$ are stochastically bounded, $\{\langle M_{n,2}\rangle\}$ is stochastically bounded.

Similarly, we can show that $\{\langle M_{n,3}\rangle\}$ is also stochastically bounded. This implies that $\{M_{n,i}\}$'s for $i = 1, 2, 3$ are stochastically bounded, which in turn implies the stochastic boundedness of $\{\hat{Q}_n^c\}$ in $\mathcal{D}$. Thus,

$$\hat{Q}_n^c/\sqrt{n} \Rightarrow \eta \text{ in } \mathcal{D} \text{ as } n \to \infty$$

where η is the zero function defined above.

By FCLT for Poisson processes and CMT with composition map, we have

$$(M_{n,1}, M_{n,2}, M_{n,3}) \Rightarrow (B_1 \circ \lambda\omega, B_2 \circ s\mu(0)\omega, B_3 \circ \eta)$$

where $\omega(t) \equiv 1$ for any t.

Finally, applying the CMT with the integral representation (6.6), we get the result in Theorem 6.4.9. □

Table 6.1: Performance comparison of systems with different load sensitivity parameter, b.
($\mu(q) = 0.6 + 0.4\exp(-b_i(q-s)^+/s)$, $\lambda = 500$, $n = 511$ and $\theta = 0.3$)

(a) Base Case

b	$P(W)$	$P(Ab)$
1.25	0.9830	0.2021
1.75	1	0.3199
2.25	1	0.3562
2.75	1	0.3718

(b) Policy A - Increase staffing

b	Staffing level (Δ_n)	$P(W)$	$P(Ab)$
1.25	520 (3.89%)	0.4102	0.0234
1.75	546 (9.09%)	0.0295	0.0003
2.25	569 (13.66%)	0.0016	0
2.75	588 (17.53%)	0.0001	0

(c) Policy B - Increase abandonment

b	h_n	$\bar{\theta}$	$P(W)$	$P(Ab)$	$P(Q(\infty) > h_n)$
1.25	579	0.5	0.2408	0.0043	0.0260
1.75	541	0.7	0.2712	0.0069	0.0399
2.25	530	0.9	0.3098	0.0109	0.0591
2.75	525	1.1	0.3577	0.0170	0.0849

(d) Policy C - Admission control

b	c_n	$P(W)$	$P(Ab)$	$P(Bl)$
1.25	579	0.4873	0.0090	0.0057
1.75	541	0.3426	0.0031	0.0107
2.25	530	0.2583	0.0016	0.0135
2.75	525	0.2056	0.0010	0.0151

Chapter 7

Bi-stability Analysis of the Modified Erlang-A Model in the Quality-Driven Regime

This chapter is a short extension of the bi-stability analysis in Chapter 6 for Quality-and-Efficiency Driven regime to Quality Driven (QD) regime.

7.1 Fluid analysis in QD regime

We denote the queue length process by $Q \equiv \{Q(t) : t \geq 0\}$, where $Q(t)$ counts the number of customers in the system (waiting and in service) at time t.

Assumption 7.1.1 *$\mu \in C^2$ with $\mu'(x) \leq 0$ and $\mu''(x) \geq 0$ for all $x \geq 0$. $lim_{x\to\infty}\mu(x) = \mu(\infty) > 0$.*

To conduct the heavy-traffic analysis, we consider a sequence of systems indexed by n, where both the arrival rate and the number of servers grows with n. For the n-th system, we denote $Q_n \equiv \{Q_n(t) : t \geq 0\}$ as the queue length process (number of people in the system). We denote the arrival rate as λ_n and the number of servers is

n. The abandonment rate does not scale with n and the service rate function takes the same form when applied to the scaled queue length process, $(Q_n - n)^+/n$. We consider the QD asymptotic regime. Without loss of generality, we assume $\mu(0) = 1$ and $\lambda_n = \rho n$ for $\rho < 1$.

Let $A \equiv \{A(t) : t \geq 0\}$, $S \equiv \{S(t) : t \geq 0\}$ and $R \equiv \{R(t) : t \geq 0\}$ be three independent Poisson processes, each with unit rate. A, S and R generate the arrival, service completion and abandonment processes, respectively. Then, the pathwise construction of Q_n is:

$$\begin{aligned} Q_n(t) &= Q_n(0) + A(\lambda_n t) - S\left(\int_0^t \mu\left(\frac{(Q_n(u)-n)^+}{n}\right)(Q_n(u)\wedge n)\,du\right) \\ &\quad -R\left(\int_0^t \theta(Q_n(u)-n)^+ du\right), \end{aligned}$$

where $(x)^+ = \max(0, x)$ and $(x \wedge y) = \min(x, y)$.

We define the fluid-scaled process

$$\bar{Q}_n(t) = \frac{Q_n(t)}{n}$$

Theorem 7.1.2 *If $\bar{Q}_n(0) \Rightarrow q(0)$ in $\mathbb{R}$, then $\bar{Q}_n \Rightarrow q$ in $\mathcal{D}$ as $n \to \infty$. The limit process q is the unique solution satisfying the following integral equation*

$$q(t) = q(0) + \rho t - \int_0^t \mu\left((q(u)-1)^+\right)(q(u)\wedge 1)\,du - \int_0^t \theta\,(q(u)-1)^+\,du.$$

The proof or Theorem 7.1.2 follows from the Proof of Theorem 6.2.1 in Chapter 6.

Let $f(q) = \rho - \mu\,(q-1)^+)\,(q(u)\wedge 1) - \theta(q-1)^+$, be the flow rate function of the fluid system at state q. Then we can write $q(t)$ as the solution to the following autonomous differential equation with initial value $q(0)$,

$$\dot{q} = f(q)$$

Let $\nu(x) = \mu(x) + \theta x$ for $x \geq 0$ and $\hat{x} = \arg\max_x\{\nu(x)\}$ on $[0, \infty)$. To enforce bi-stability, we impose the following assumptions on the service rate function in addition to Assumption 7.1.1.

Assumption 7.1.3 *$-\mu'(0) > \theta$ and $\rho > \mu(\hat{x}) + \theta\hat{x}$.*

Under Assumption 7.1.3, $\hat{x}$ is the root of $\nu'(x) = 0$ on $(0, \infty)$. Let $\hat{q} = \hat{x} + 1$. $\hat{q}$ is the point where $f(q)$ attains its maximum on $[1, \infty)$.

Lemma 7.1.4 *Under Assumption 7.1.1 and 7.1.3, the fluid model has three equilibrium points, denoted as $\bar{q}_1$, $\bar{q}_2$, and $\bar{q}_3$, with $\bar{q}_1 = \rho$,$1 < \bar{q}_2 < \hat{q}$ and $\bar{q}_3 > \hat{q}$. $\bar{q}_1$ and $\bar{q}_2$ are asymptotically stable, while $\bar{q}_2$ is unstable.*

The proof of Lemma 7.1.4 follows exactly the same line of analysis as Theorem 6.2.4 in Chapter 6. We shall omit it here.

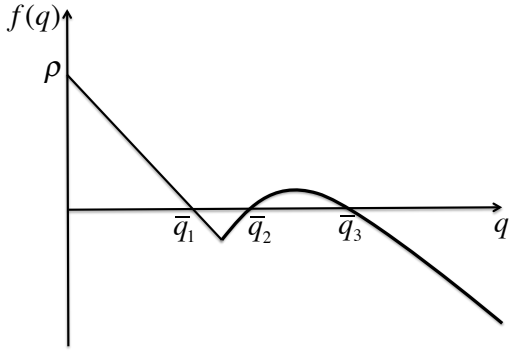


Figure 7.1: Flow rate function

7.2 Analysis of stationary distribution

Let π_n denote the stationary distribution of the n-th system, then we have the following detailed balance equation for Birth-and-Death process.

$$\lambda_n \pi_n(q-1) = \left(\mu\left(\frac{(q-n)^+}{n}\right)((q) \wedge n) + \theta(q-n)^+\right)\pi_n(q)$$

Then we have when $\lambda_n \geq \mu\left((q-n)^+/n\right)(q \wedge n) + \theta(q-n)^+$, $\pi_n(q) \geq \pi_n(q-1)$. Under Assumption 7.1.3, let $\tilde{x}$ denote the root of $\rho - \nu(x) = 0$ on $(0, \hat{x})$ and $\bar{x}$ denote the root of $\rho - \nu(x) = 0$ on $(\hat{x}, \infty)$. Then we have

Lemma 7.2.1 *Under Assumption 7.1.3, $\pi_n(\cdot)$ has two peaks, one at $\bar{q}_{n,1} = \lfloor \lambda_n \rfloor$, the other at $\bar{q}_{n,2} = \lfloor (\bar{x}+1)n \rfloor$. The minimum point between the two peaks (valley) is, $\tilde{q}_n = \lfloor (\tilde{x}+1)n \rfloor$.*

Proof. Let $f_n(q) = \lambda_n - \mu\left((q-n)^+/n\right)(q \wedge n) + \theta(q-n)^+$. For $q < \lambda_n$, $f_n(q) = \lambda_n - \mu(0)q > 0$. For $\lambda_n < q < n$, $f_n(q) = \lambda_n - \mu(0)q < 0$. When $q > n$, let $x_n = (q-n)/n$. Then $f_n(q)/n = \rho - \nu(x_n)$. As $\rho - \nu(x_n) < 0$ for $0 \le x_n < \tilde{x}$, $\rho - \nu(x) \ge 0$ for $\tilde{x} \le x_n \le \bar{x}$ and $\rho - \mu(x) < 0$ for $x_n > \bar{x}$, we have $f_n(q) < 0$ for $s_n \le q < (\tilde{x}+1)n$, $f_n(q) \ge 0$ for $(\tilde{x}+1)n \le q \le (\bar{x}+1)n$ and $f_n(q) < 0$ for $q > (\bar{x}+1)n$. □

The following theorem characterize the relationship among the values of $\bar{q}_{n,1}$, $\bar{q}_{n,2}$ and $\tilde{q}_n$.

Theorem 7.2.2

$$\frac{1}{n}\log\frac{\pi_n(q_{n,1})}{\pi_n(\tilde{q}_n)} = I_1$$

and

$$\frac{1}{n}\log\frac{\pi_n(q_{n,2})}{\pi_n(\tilde{q}_n)} = I_2$$

where

$$I_1 = (1-\rho)\log\mu(0) + \int_\rho^1 \log x\,dx + \int_0^{\tilde{x}} \log\nu(x)\,dx - (\tilde{x}+1-\rho)\log\rho.$$

and

$$I_2 = -\int_{\tilde{x}}^{\bar{x}} \log\nu(x)\,dx + (\bar{x}-\tilde{x})\log\rho.$$

Proof. As

$$\begin{aligned}
&\pi_n(\bar{q}_{n,1}) \\
= &\prod_{q=\bar{q}_{n,1}+1}^{\tilde{q}_n} \frac{\mu\left((q-n)^+/n\right)(q\wedge n) + \theta(q-n)^+}{\lambda_n}\pi_n(\tilde{q}_n) \\
= &\exp\left(\sum_{q=\bar{q}_{n,1}+1}^{n} \log(\mu(0)\frac{q}{n}) + \sum_{q=n+1}^{\tilde{q}_n} \log\nu\left(\frac{q-n}{n}\right) - (\tilde{q}_n - \bar{q}_{n,1})\log\frac{\lambda_n}{n}\right)\pi_n(\tilde{q}_n),
\end{aligned}$$

then under our scaling parameters ($n = n$, $\lambda_n = \rho n$), we have

$$\begin{aligned}
& \frac{1}{n}\log\frac{\pi_n(\bar{q}_{n,1})}{\pi_n(\tilde{q}_n)} \\
= \quad & \frac{n-\rho n}{n}\log\mu(0) + \frac{1}{n}\sum_{q=\rho n+1}^{n}\log\left(\frac{q}{n}\right) + \frac{1}{n}\sum_{k=1}^{\tilde{x}n}\log\left(\nu\left(\frac{k}{n}\right)\right) \\
& -\frac{(\tilde{x}+1)n-\rho n}{n}\log\rho \\
\to \quad & (1-\rho)\log\mu(0) + \int_{\rho}^{1}\log(x)dx + \int_{0}^{\tilde{x}}\log\nu(x)dx - (\tilde{x}+1-\rho)\log\rho
\end{aligned}$$

as $n \to \infty$. Likewise, we have

$$\pi_n(\bar{q}_{n,2}) = \exp\left(-\sum_{\tilde{q}_n+1}^{\bar{q}_{n,2}}\log\nu\left(\frac{q-n}{n}\right) + (\bar{q}_{n,2}-\tilde{q}_n)\log\frac{\lambda_n}{n}\right).$$

Then

$$\begin{aligned}
\frac{1}{n}\log\frac{\pi_n(\bar{q}_{n,2})}{\pi_n(\tilde{q}_n)} \quad &= \quad -\frac{1}{n}\sum_{\tilde{x}n+1}^{\bar{x}n}\log\nu\left(\frac{k}{n}\right) + \frac{(\bar{x}+1)n-(\tilde{x}+1)n}{n}\log\rho \\
&\to \quad -\int_{\tilde{x}}^{\bar{x}}\log\nu(x)\,dx + (\bar{x}-\tilde{x})\log\rho.
\end{aligned}$$

□

Lemma 7.2.3 *For any fixed* $0 < y_1 < y_2 < \infty$

$$\lim_{n\to\infty}\log\frac{\pi_n(\lfloor\rho n+\sqrt{n}y_2\rfloor)}{\pi_n(\lfloor\rho n+\sqrt{n}y_1\rfloor)} = -\int_{y_1}^{y_2}\frac{1}{\rho}y\,dy$$

and

$$\lim_{n\to\infty}\log\frac{\pi_n(\lfloor\rho n-\sqrt{n}y_1\rfloor)}{\pi_n(\lfloor\rho n-\sqrt{n}y_2\rfloor)} = -\int_{-y_2}^{-y_1}\frac{1}{\rho}y\,dy$$

Proof. We prove the first equation only, as the proof of the second equation follows exactly the same line of analysis.

$$\begin{aligned}
\log\frac{\pi_n(\lfloor\rho n+\sqrt{n}y_2\rfloor)}{\pi_n(\lfloor\rho n+\sqrt{n}y_1\rfloor)} \quad &= \quad -\sum_{k=\lfloor\sqrt{n}y_1\rfloor+1}^{\lfloor\sqrt{n}y_2\rfloor}\log(1+\frac{1}{\rho}\frac{k}{n}) \\
&= \quad -\sum_{k=\lfloor\sqrt{n}y_1\rfloor+1}^{\lfloor\sqrt{n}y_2\rfloor}\frac{1}{\rho}\frac{k}{\sqrt{n}}\frac{1}{\sqrt{n}} + O(\frac{1}{\sqrt{n}}) \\
&\to \quad -\int_{y_2}^{y_1}\frac{1}{\rho}y\,dy.
\end{aligned}$$

□

Lemma 7.2.4 *Recall that $\bar{x} > 0$ is the strictly positive root of $\rho - \mu(x) - \theta x = 0$. For any fixed $0 < y_1 < y_2 < \infty$,*

$$\lim_{n\to\infty} \log \frac{\pi_n(\lfloor(\bar{x}+1)n + \sqrt{n}y_2\rfloor)}{\pi_n(\lfloor(\bar{x}+1)n + \sqrt{n}y_1\rfloor} = -\int_{y_1}^{y_2} \frac{\mu'(\bar{x})+\theta}{\rho} x\, dx$$

and

$$\lim_{n\to\infty} \log \frac{\pi_n(\lfloor(\bar{x}+1)n - \sqrt{n}y_1\rfloor)}{\pi_n(\lfloor(\bar{x}+1)n - \sqrt{n}y_2\rfloor)} = -\int_{-y_1}^{-y_2} \frac{\mu'(\bar{x})+\theta}{\rho} x\, dx$$

Proof. We prove the first equation only, as the proof of the second equation follows exactly the same line of analysis.

$$\begin{aligned}
& \log \frac{\pi_n(\lfloor(\bar{x}+1)n + \sqrt{n}y_2\rfloor)}{\pi_n(\lfloor(\bar{x}+1)n + \sqrt{n}y_1\rfloor)} \\
= & -\sum_{k=\lfloor\sqrt{n}y_1\rfloor+1}^{\lfloor\sqrt{n}y_2\rfloor} \log\left(1 + \frac{1}{\rho}\left(\mu\left(\bar{x}+\frac{k}{n}\right) - \mu(\bar{x}) + \theta\frac{k}{n}\right)\right) \\
= & -\sum_{k=\lfloor\sqrt{n}y_1\rfloor+1}^{\lfloor\sqrt{n}y_2\rfloor} \frac{\mu'(\bar{x})+\theta}{\rho}\frac{k}{\sqrt{n}}\frac{1}{\sqrt{n}} + O(\frac{1}{\sqrt{n}}) \\
\to & -\int_{y_2}^{y_1} \frac{\mu'(\bar{x})+\theta}{\rho} y\, dy.
\end{aligned}$$

□

Part III

Bibliography

Bibliography

[1] S. Asmussen, *Applied Probability and Queues*. New York: Spinger, 2 ed., 2003.

[2] S. Asmussen and P. W. Glynn, *Stochastic Simulation: Algorithms and Analysis*, vol. 57. Springer-Verlag, 2007.

[3] J. Propp and D. Wilson, "Exact sampling with coupled Markov chains and applications to statistical mechanics," *Random Structures and Algorithms*, vol. 9, pp. 223–252, 1996.

[4] P. Green and D. Murdoch, "Exact sampling for Baysian inference: towards general purppose algorithms (with discussion)," *Bayesian Statistics*, vol. 6, pp. 301–321, 1999.

[5] D. Wilson, "How to couple from the past using a read-once source of randomness," *Random Structures and Algorithms*, vol. 16, no. 1, pp. 85–113, 2000.

[6] J. Corcoran and U. Schneider, "Shift and scale coupling methods for perfect simulation," *Probability in the Engineering and Informational Sciences*, vol. 17, no. 3, pp. 277–303, 2003.

[7] S. Foss and R. Tweedie, "Perfect simulation and backward coupling," *Stochastic Models*, vol. 14, pp. 187–203, 1998.

[8] W. Kendall, "Geometric ergodicity and perfect simulation," *Electron. Comm. Probab.*, vol. 9, pp. 140–151, 2004.

[9] J. Fill and M. Huber, "Perfect simulation of Vervaat perpetuities," *Electron. Comm. Probab.*, vol. 15, pp. 96–109, 2010.

[10] J. Corcoran and R. Tweedie, "Perfect sampling of ergodic Harris chains," *Ann. of Appl. Probab.*, vol. 11, no. 2, pp. 438–451, 2001.

[11] S. Connor and W. Kendall, "Perfect simulation for a class of positive recurrent Markov chains," *Ann. Appl. Probab.*, vol. 3, pp. 781–908, 2007.

[12] K. Sigman, "Exact simulation of the stationary distribution of the FIFO M/G/c queue," *Journal of Applied Probability*, vol. 48A, pp. 209–216, 2011.

[13] K. Sigman, "Exact simulation of the stationary distribution of the FIFO M/G/c queue: The general case of $\rho < c$," *Queueing Systems: Theory and Applications*, vol. 70, 2012.

[14] D. Murdoch and G. Takahara, "Perfect sampling for queues and network models," *ACM Transactions of Modeling and Computer Simulation*, vol. 16, pp. 76–92, 2006.

[15] W. Kendall, "Perfect simulation for area-interaction point processes," in *Probability Towards 2000* (L. Accardi and C. Heyde, eds.), pp. 218–234, New York: Spinger, 1998.

[16] W. Kendall and J. Møller, "Perfect simulation using dominating processes on ordered spaces, with application to locally stable point pocesses," *Adv. Appl. Prob.*, vol. 32, pp. 844–865, 2000.

[17] R. Fernandez, P. Ferrari, and N. Garcia, "Perfect simulation for interacting point processes, loss networks and ising models," *Stoch. Process. Appl.*, vol. 102(1), pp. 63–88, 2002.

[18] K. Berthelsen and J. Møller, "A primer on perfect simulation for spatial point process," *Bull Braz Math Soc*, vol. 33(3), pp. 351–367, 2002.

[19] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, "Statistical analysis of a telephone call center: a queueing-science perspective," *Preprint*, 2002.

[20] P. Protter, *Stochastic Integration and Differential Equations*. New York: Springer, 2 ed., 2005.

[21] S. Heinrich, "Multilevel Monte Carlo methods," in *Lecture Notes in Computer Science*, vol. 2179, pp. 58–67, Berlin: Springer-Verlag, 2001.

[22] M. Giles, "Multilevel Monte Carlo path simulation," *Operations Research*, vol. 56, no. 3, pp. 607–617, 2008.

[23] D. McLeish, "A general method for debiasing a Monte Carlo estimator," *Monte Carlo Methods and Applications*, vol. 17, no. 4, pp. 301–315, 2011.

[24] C. Rhee and P. Glynn, "A new approach to unbiased estimation of sdes." http://arxiv.org/pdf/1207.2452.pdf, 2012.

[25] N. Chen and Z. Huang, "Localization and exact simulation of brownian motion-driven stochastic differential equations," *Mathematics of Operations Research*, vol. 38, pp. 591–616, 2013.

[26] A. Beskos and G. Roberts, "Exact simulation of diffusions," *Annals of Applied Probability*, vol. 15, pp. 2422–2444, 2005.

[27] A. Beskos, O. Papaspiliopoulos, and G. Roberts, "Retrospective exact simulation of diffusion sample paths with applications," *Bernoulli*, vol. 12, no. 6, pp. 1077–1098, 2006.

[28] A. Beskos, S. Peluchetti, and G. Roberts, "ϵ-strong simulation of the Brownian path," *Bernoulli*, vol. 18, no. 4, pp. 1223–1248, 2012.

[29] M. Pollock, A. Johansen, and G. Roberts, "On the exact and ε-strong simulation of (jump) diffusions." http://arxiv.org/pdf/1302.6964v2.pdf, 2014.

[30] C. Bayer, P. Friz, S. Riedel, and J. Schoenmakers, "From rough path estimates to multilevel Monte Carlo." http://arxiv.org/pdf/1305.5779v1.pdf, 2013.

[31] J. Blanchet and K. Sigman, "On exact sampling of stochastic perpetuities," *J. Appl. Probab.*, vol. 48A, pp. 165–182, 2011.

[32] K. Ensor and P. Glynn, "Simulating the maximum of a random walk," *Journal of Statistical Planning and Inference*, vol. 85, pp. 127–135, 2000.

[33] P. Glasserman, *Monte Carlo Methods in Financial Engineering*. New York: Spinger, 2003.

[34] F. Kelly, "Loss networks," *Annals of Applied Probability*, no. 1, pp. 319–378, 1991.

[35] J. Blanchet, X. Chen, and H. Lam, "Two-parameter sample path large deviation for infinite server queues," *working paper*, 2012.

[36] R. J. Adler, "An introduction to continuity, extrema, and related topics for general Gaussian processes," *IMS Lecture Notes: Monograph Series*, vol. 12, 1990.

[37] G. Pang and W. Whitt, "Two-parameter heavy-traffic limits for infinite-server queues," *Queueing Systems: Theory and Applications*, no. 65, pp. 325–264, 2010.

[38] J. Blanchet and H. Lam, "Rare-event simulation for many-sever queues," *working paper*, 2012.

[39] T. Lyons, "Differential equations driven by rough signals," *Rev. Mat. Iberoamericana*, vol. 14, no. 2, pp. 215–310, 1998.

[40] A. Davie, "Differential equations driven by rough paths: An approach via discrete approximation." http://arxiv.org/abs/0710.0772.

[41] J. Steele, *Stochastic Calculus and Financial Application*. Springer-Verlag, 2001.

[42] J. Blanchet and X. Chen, "Steady-state simulation for reflected Brownian motion and related networks." http://arxiv.org/pdf/1202.2062.pdf, 2013.

[43] P. Fritz and N. Victoir, *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*, vol. 120. Cambridge University Press, 2010.

[44] R. Batt and C. Terwiesch, "Doctors under load: An empirical study of state-dependent service times." Working Paper, 2012.

[45] M. Gerla and L. Kleinrock, "Flow control: A comparative survey," *IEEE Transactions on Communications*, vol. 28, no. 4, pp. 553–574, 1980.

[46] D. KC and C. Terwiesch, "Impact of work load on service time and patient safety: An econometric analysis of hospital operations," *Management Science*, vol. 55, no. 9, pp. 1486–1498, 2009.

[47] P. Feldman, J. Li, G. Yom-Tov, and E. Yom-Tov, "Service time sensitivity to load: Who is to "blame"?." Working Paper, 2014.

[48] C. Chan, G. Yom-Tov, and G. Escobar, "When to use speedup: An examination of service systems with returns." To appear in Operations Research, 2014.

[49] J. Bertrand and H. van Ooijen, "Workload based order release and productivity: a missing link," *Production Planning and Control*, vol. 12, no. 7, pp. 665–678, 2002.

[50] C. Wickens, J. Hollands, R. Parasuraman, and S. Banbury, *Engineering psychology and human performance*. Pearson, 4 ed., 2012.

[51] J. Caldwell, "The impact of fatigue in air medical and other types of operations: a review of fatigue facts and potential countermeasures," *Air Medical Journal*, vol. 20, no. 1, pp. 25–32, 2001.

[52] D. Chalfin, S. Trzeciak, A. Likourezos, B. Baumann, and R. Dellinger, "Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit," *Critical Care Medicine*, vol. 35, pp. 1477–1483, 2007.

[53] O. Garnett, A. Mandelbaum, and M. Reiman, "Designing a call center with impatient customers," *Manufacturing and Service Operations Management*, vol. 4, no. 3, pp. 208–227, 2002.

[54] W. Whitt, "Efficiency-driven heavy traffic approximations for many-server queues with abandonments," *Management Science*, vol. 50, no. 10, pp. 1449–1461, 2004.

[55] C. Palm, "Research on telephone traffic carried by full availability groups," *Tele*, vol. 1, p. 107, 1957.

[56] A. Mandelbaum and S. Zeltyn, "Service engineering in action: The Palm/Erlang-A queue with applications to call centers," in *Advances in Service Innovations* (D. Spath and K. P. Fahnrich, eds.), pp. 17–48, Springer-Verlag, 2007.

[57] W. Whitt, "Queues with service times and interarrival times depending linearly and randomly upon waiting times," *Queueing Systems*, vol. 6, pp. 335–352, 1990.

[58] O. Boxma and M. Vlasiou, "On queues with service and interarrival times depending on waiting times," *Queueing Systems*, vol. 56, pp. 121–132, 2007.

[59] A. Mandelbaum and G. Pats, "State-dependent stochastic networks. part I: Approximations and applications with continuous diffusion limits," *The Annals of Applied Probability*, vol. 8, no. 2, pp. 569–646, 1998.

[60] A. Weerasinghe, "Diffusion approximation for G/M/n+GI queues with state-dependent service rate," *Mathematics of Operations Research*, 2013.

[61] R. Gibbens, P. Hunt, and F. Kelly, "Bistability in communication networks," in *Disorder in Physical Systems*, pp. 113–128, Oxford University Press, 1990.

[62] N. Antunes, C. Fricker, P. Robert, and D. Tibi, "Stochastic networks with multiple stable points," *The Annal of Probability*, vol. 36, no. 1, pp. 255–278, 2009.

[63] A. Janssen and J. van Leeuwaarden, "Staffing many-server systems with admission control and retrials." Working paper, 2014.

[64] F. D. Hollander, "Metastability under stochastic dynamics," *Stochastic Process. Appl.*, vol. 114, no. 1, pp. 1–26, 2004.

[65] E. Olivieri and E. Vares, *Large Deviation and Metastability.* Cambridge Univ. Press, 2005.

[66] J. Huang, A. Mandelbaum, H. Zhang, and J. Zhang, "Refined models for efficiency-driven queues with applications to delay announcements and staffing." Working Paper, 2014.

[67] E. Zohar, A. Mandelbaum, and N. Shimkin, "Adaptive behavior of impatient customers in tele-queues: theory and empirical support," *Management Science*, vol. 48, pp. 566–583, 2002.

[68] M. Armony, N. Shimkin, and W. Whitt, "The impact of delay announcements in many-server queues with abandonment," *Operations Research*, vol. 57, no. 1, pp. 68–81, 2009.

[69] R. Bekker and S. Borst, "Optimal admission control in queues with workload-dependent service rates," *Probability in the Engineering and Informational Sciences*, vol. 20, pp. 543–570, 2006.

[70] A. Janssen, J. van Leeuwaarden, and J. Sanders, "Scaled control in the QED regime." Working paper, 2013.

[71] D. Daley, J. van Leeuwaarden, and Y. Nazarathy, "BRAVO for many-server QED systems with finite buffers." Working paper, 2013.

[72] L. Green, P. Kolesar, and W. Whitt, "Coping with time-varying demand when setting staffing requirements for a service system," *Production and Operations Management*, vol. 26, no. 1, pp. 13–39, 2007.

[73] G. Yom-Tov and A. Mandelbaum, "Erlang-R: A time-varying queues with reentrant customers, in support of healthcare staffing," *To appear in MSOM*, 2014.

[74] H. Chen and D. Yao, *Fundamentals of queueing networks: performance, asymptotics and optimization*. No. 46, Springer-Verlag, 2001.

[75] C. Chan, V. Farias, and G. Escobar, "The impact of delays on service times in the intensive care unit." Working Paper, 2013.

[76] G. Pang, R. Talreja, and W. Whitt, "Martingale proofs of many server heavy-traffic limits for Markovian queues," *Probability Surveys*, vol. 4, pp. 193–267, 2007.

[77] P. Fleming, A. Stolyar, and B. Simon, "Heavy traffic limit for a mobile phone system loss model," in *Proc. 2nd Internat. Conf. Telecommunication Systems, Modeling, and Analysis*, (Nashville, TN), pp. 158–176, 1994.