Extinction time of stochastic SIRS models: criticality, ODE and diffusion approximation

Jingran Zhai

A thesis submitted for the degree of
Doctor of Philosophy

School of Mathematical Sciences
Queen Mary, University of London
United Kingdom

November 2020

Declaration

I, JINGRAN ZHAI, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: JINGRAN ZHAI   Date: 16.11.2020

Abstract

Stochastic epidemic models are useful in modelling the duration of epidemic outbreaks. It has been observed that the behaviour of the extinction time of epidemics changes across some point (or domain in multi-dimensional spaces) in the parameter space, known as the 'criticality': generally speaking, epidemics in the subcritical regime tend to end quickly, whereas epidemics in the supercritical regime tend to prevail around the quasi-stationary state for a long time before extinction. In recent years, there has been substantial interest in the phase transition window around the criticality, called the 'critical regime'. We expect to observe the critical behaviour not only at the criticality point, but across the entire critical regime, and the boundary of the critical regime is expected to be approaching the criticality as the population size tends to infinity. However, while this phenomenon is well-discussed for one-dimensional epidemic models like SIS, there is little work done on two or higher-dimensional models.

This thesis is concerned with the scaling behaviour in and around the phase transition window of the extinction time of a class of two-dimensional stochastic epidemic models named SIRS. The stochastic SIRS model is a continuous-time Markov chain modelling the spread of infectious diseases with temporary immunity, in a homogeneously-mixing population of fixed size N. More specifically, we study the asymptotic distributions of the extinction time of SIRS models as N tends to infinity, with both the parameter space and the initial state of the model treated as functions of N. Our results provide a comprehensive picture of various possible scalings and the corresponding limit distributions within the subcritical and the critical regimes. Our approach also provides us with descriptions of the entire trajectory of SIRS epidemics. Simulations are implemented to verify our results.

Acknowledgements

I would like to thank:

My supervisor Prof. Alexander Gnedin for his guidance, constant encouragement, kindness and patience;

Prof. Andrey A. Dorogovtsev, for the valuable discussions, and for inviting me to visit the Institute of Mathematics of NASU;

Everyone I met at the Institute of Mathematics of NASU, including Alexey, Ekaterina, Georgii, Mykola and Vitalii, for their generosity, for making me feel at home during my short visit, and for all the inspiring conversations online and in person;

Prof. Malwina Luczak, for introducing me to the topic of extinction times of epidemic models;

The staff and my fellow PhD students of the School of Mathematics, QMUL, for a friendly and supportive environment in which to work and play;

Queen Mary University of London and School of Mathematics, for funding me with a scholarship.

I would also like to take this rare opportunity to express my gratitude to:

My secondary school teachers Ms. Gao and Ms. Lang, for being the mentors that any teenager would feel lucky to have;

My friend CY, with whom I can always have deep conversations; and Xiaowen, who went to a lot of trouble to ship a valuable care package to me during the hardship of 2020;

My parents, for unconditionally supporting my life choices, and my family-in-law, for being supportive, loving and fun.

Last but not least, I would like to thank my husband, who accompanied me throughout this journey with generous love and support.

Contents

List of Figures

List of Symbols

$\mathbb{R}^d$ – the real coordinate d-dimensional space, where $|x|$, $x \in \mathbb{R}^d$ denotes the Euclidean norm, and $\langle x, y \rangle$ denotes the scalar product. Also, let $\mathbb{R}_+ := (0, \infty)$.

$\mathbf{S}^d$ – the set of symmetric, non-negative definite, $d \times d$ real matrices. A $d \times d$ symmetric real matrix M is said to be *non-negative definite* if $x^\intercal M x \geq 0$ for all $x \in \mathbb{R}^d$. The norm $\|M\|$ denotes the operator norm induced by the Euclidean norm.

$[N]$ – A set consisting of all non-negative integers up to N.

For $\Omega \subset \mathbb{R}^d$,
$(BC(\Omega), \|\cdot\|)$ – the Banach space of all bounded real-valued continuous functions defined on Ω, equipped with the sup-norm $\|f\| = \sup_{x \in \Omega} |f(x)|$.

$C_0(\Omega)$ – functions from $BC(\Omega)$ that vanish on $\partial\Omega$;

$C_c(\Omega)$ – the subset of $BC(\Omega)$ with compact supports;

$C^{j_1, \cdots, j_n}(\Omega)$, $j_i \in \mathbb{N}, 1 \leq i \leq n$ – the set of continuous functions from Ω to $\mathbb{R}$ possessing continuous partial derivatives up to the j_i-th order with respect to the i-th component.

$BC^{j_1, \cdots, j_n}(\Omega) = BC(\Omega) \cap C^{j_1, \cdots, j_n}(\Omega)$ – the set of continuous functions from Ω to $\mathbb{R}$ possessing bounded continuous derivatives up to the j_i-th order with respect to the i-th component.

$C^\infty(\Omega)$ – the set of smooth continuous functions from Ω to $\mathbb{R}$;

$C_c^\infty(\Omega) := C_c(\Omega) \bigcap C^\infty(\Omega)$;

$D([0,\infty), \mathbb{R}^d)$ – the Skorokhod space of càdlàg functions valued in $\mathbb{R}^d$.

$C(M, S)$ – the set of continuous mappings from M to S.

For functions $f(x), g(x)$:

$f(x) = \Theta(g(x))$ or $f(x) \asymp g(x)$ – There exists constant $K > 0$ s.t.

$$K^{-1}|g(x)| \leq |f(x)| \leq K|g(x)|;$$

$f(x) = o(g(x))$ or $f(x) \ll g(x)$ or $g(x) = \omega(f(x)) \Longleftrightarrow \lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$;

$f(x) \sim g(x) \Longleftrightarrow \lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$.

$(f \vee g)(x) = f(x) \vee g(x)$ and $(f \wedge g)(x) = f(x) \wedge g(x)$, where $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$.

$\Rightarrow$ – weak convergence;
$\xrightarrow{\mathbb{P}}$ – convergence in probability;
$\rho - \lim$ – convergence w.r.t. ρ-topology.

Chapter 1

Introduction

The *stochastic SIRS model* describes the spread of a disease with temporary immunity in a closed population of size N. Each susceptible individual is expected to contract the disease at rate $\lambda_o I/N$, where I denotes the current size of the infected population. Parameter $\lambda_o \in \mathbb{R}_+$ is known as the *transmission rate*. Once infected, each individual is immediately infectious and will recover at rate $\mu_o = 1$ independently of other individuals. Each recovered individual loses immunity at rate $\gamma_o \in \mathbb{R}_+$ and becomes susceptible independently. For future reference, we use interchangeably both the words infected and infectious, and the words recovered and immune when referring to population compartments in our epidemic models. The subscript 'o' stands for 'original' and is introduced to distinguish the original variables from the scaled variables.

Formally, the *stochastic SIRS model* is constructed as a two-dimensional continuous-time Markov chain $(I_t^N, R_t^N)_{t\geq 0}$, where I_t^N represents the size of infected population at time t, and R_t^N represents the size of immune population at time t. The model is associated with the transition rates:

$$\begin{aligned} (i,r) &\to (i+1,r), && \text{at rate } \lambda_o(N-i-r)i/N, \\ (i,r) &\to (i,r-1), && \text{at rate } \gamma_o r, \\ (i,r) &\to (i-1,r+1), && \text{at rate } i. \end{aligned} \tag{1.1}$$

We will state the definition of the stochastic SIRS model and the simpler SIS and

SIR models in full detail in Chapter 2.

After the basic parameters of a disease are understood, researchers naturally want to predict the course of development of the epidemic, and one of the variables of interest, known as the *extinction time*, represents how long it takes for the pathogen to die out within a population. In this thesis, the extinction time is modelled as the stopping time $T_o^N := \inf\{t : I_t^N = 0\}$.

We aim to study the scaling behaviour of the distribution of the extinction time of stochastic SIRS models, as the population size N tends to infinity. In particular, we are interested in how the scale of the parameters (λ_o, γ_o) and the initial states (I_0^N, R_0^N) affect the asymptotic distribution of the extinction time.

The study of this problem, associated with the simpler SIS and SIR models, dates back to as early as 1975 by Barbour [1]. Since then, most of the work has been on SIS models, and focused mainly on the expectation instead of the distribution [2–5]. For both SIS and SIR models, it has been proved that there exists a 'critical value' of the transmission rate, and that the extinction time is $O(\log N)$ when the transmission rate is strongly subcritical, and is $O(e^{cN})$ for some constant $c \in \mathbb{R}_+$ when the transmission rate is strongly supercritical. It has also been shown that the shape of the scaled distribution is affected by the scaling of I_0^N. For a while, only the cases with the initial size of infected population $I_0^N \asymp 1$ or $O(N)$ were studied.

Another observation made by existing literature is that the critical parameter regime is an $o(1)$-sized neighbourhood around the critical value. More specifically, if as $N \to \infty$, the transmission rate of the SIS (resp. SIR) model tends to the critical value faster than $N^{-1/2}$ (resp. $N^{-1/3}$), then the corresponding extinction time is $O(N^{1/2})$ (resp. $O(N^{1/3})$).

One important work on stochastic SIS models is by Foxall [6], who developed a framework that allows a comprehensive investigation over all possible combinations of transmission rates and initial sizes of infected population.

To our knowledge, this thesis is the first to study the scaling behaviour of SIRS extinction time in the subcritical ($\lambda_o \leq 1$) and near-critical ($\lambda_o \to 1$) regimes. We contribute in the following three directions.

Firstly, we extend Foxall's framework from the one-dimensional SIS model with a one-dimensional parameter space to the two-dimensional SIRS model with a two-dimensional parameter space. Despite both being two-dimensional, the SIRS model is significantly more complicated than the SIR model, since the latter has a monotonicity property. While the SIS and the SIR models are both driven by a single parameter (the transmission rate), the SIRS model incorporates a second parameter describing the average duration of immunity. In addition, there is no explicit solution to the ODE system describing deterministic SIRS models.

We succeed to identify the boundary of the critical regime, and obtain explicit expressions of the asymptotic distributions for a wide range of possibilities. We show how the scaling of the extinction time changes from $\log N$ to $N^{1/3}$ as the transmission rate approaches the criticality from below. We illustrate these cases with diagrams. Some cases with large initial sizes of the infected and immune populations are not covered by our results. However, we believe our techniques can be extended to all subcritical cases.

Secondly, we obtain an approximation for the scaled distribution in the critical regime. From a theoretical point of view, this is associated with the existence and uniqueness problem of a PDE associated with a degenerate operator. It can also be viewed as an application of Kühnemund's bi-continuous semigroup theory [7], and is especially interesting because the setting is non-Gaussian.

Thirdly, we run simulations to verify our theoretical results. In particular, we investigate the practicality of various methods for simulating an epidemic model with large N. We choose to implement the τ-leaping method alongside the standard SSA method. The simulation result suggests that the τ-leaping method can be an effective time-saver when simulating the long-term behaviour of near-critical epidemic models.

Now we will outline the structure of the rest of the thesis, and give a brief overview of the main results.

Summary of Chapters 2 and 3

We set up both mathematical and epidemiological background in these two chapters.

In Chapter 2, we introduce the basic notations and properties of continuous-time

Markov chains, the epidemiology background and motivation to our problem, and define the stochastic SIS, SIR, SIRS models and their deterministic counterparts.

In Chapter 3, we introduce the techniques and theories used in this thesis, including ODE approximation, diffusion approximation, the order-preserving coupling method, and the theory of bi-continuous semigroups.

Summary of Chapter 4

The main result of Chapter 4 is Theorem 4.6 below, in which we derive the asymptotic distribution of the stochastic SIRS model when the initial size of infection I_0^N is 'small'.

The process I_t^N is a birth-death chain with birth rate $\lambda_o(1-N^{-1}(R_t^N+I_t^N))$ and death rate 1. For models in the subcritical and near-critical parameter regimes, when I_0^N is small, I_t^N will remain small and thus its transition rates will be approximately linear. To our advantage, the extinction time of linear birth-death chains is explicitly known. Using the method of order-preserving coupling introduced in Section 3.2, we can sandwich each trajectory of the SIRS models between a pair of linear birth-death chains whose extinction times have the same asymptotic distribution, and in this way we can pinpoint the asymptotic distribution of the SIRS extinction time.

The technique described above has been applied by [1] to stochastic SIR models, and by [6, 8] to subcritical stochastic SIS models. The complexity of extending this technique to the SIRS models comes from the necessity to approximate R_t^N. This has not been an issue in the stochastic SIR model since R_t^N monotonically increases with respect to t.

Looking at the birth rate $\lambda_o(1-N^{-1}(R_t^N+I_t^N))$, it is natural to consider the cases $R_0^N \asymp N$ and $R_0^N = o(N)$ separately. This is the motivation for labelling the cases as Case 1.x and Case 2.x in Theorem 4.6 below.

Theorem [*4.6*]**.** *Consider a sequence of stochastic SIRS models defined in* (2.5)*, indexed by* $N \in \mathbb{N}$*, with parameters* $\lambda_o = \lambda_o(N) > 0$ *and* $\gamma_o = \gamma_o(N) > 0$*, and initial states* $(I_0^N, R_0^N) = (I_0(N), R_0(N))$*.*

Let $T_o^N := \inf\{t : I_t^N = 0\}$*. If one of the following conditions is satisfied, then we have the explicit expression of the asymptotic distribution of* T_o^N*:*

Cases 1.1-1.3 are cases where both the initial size of infection I_0 and immunity R_0 are small, whereas Cases 2.1 and 2.2 are cases where I_0 is small and R_0 is of order N.

- ***Case 1.1:*** $I_0|1-\lambda_o| \to 0$, $I_0 R_0 = o(N)$, $I_0 = o(N^{1/2}\gamma_o^{1/2})$.

  If $I_0 = O(1)$,
$$\lim_{N\to\infty} \mathbb{P}^N \left[T_o^N \le w\right] = \left(1+\frac{1}{w}\right)^{-I_0};$$
  and if $I_0 \to \infty$,
$$\lim_{N\to\infty} \mathbb{P}^N \left[\frac{T_o^N}{I_0} \le w\right] = e^{-\frac{1}{w}}.$$

- ***Case 1.2:*** $I_0(1-\lambda_o) \to a > 0$, $\lambda_o = \lambda_o(N) < 1$, *and* $I_0 = o\left(N^{1/2}\gamma_o^{1/2}\right)$, $I_0 R_0 = o\left(N\right)$.

  If $I_0 = O(1)$,
$$\lim_{N\to\infty} \mathbb{P}^N \left[T_o^N \le w\right] = \left(1+\frac{a}{e^{aw}-1}\right)^{-I_0};$$
  and if $I_0 \to \infty$,
$$\lim_{N\to\infty} \mathbb{P}^N \left[\frac{T_o^N}{I_0} \le w\right] = \exp\left\{-\frac{a}{e^{aw}-1}\right\}.$$

- ***Case 1.3:*** $I_0(1-\lambda_o) \to \infty$, $\lambda_o = \lambda_o(N) < 1$, $I_0 = o\left(\frac{N(1-\lambda_o)\gamma_o}{\log I_0(1-\lambda_o)}\right)$, *and* $R_0 \log I_0(1-\lambda_o) = o\left(N(1-\lambda_o)\right)$. *Then*
$$\lim_{N\to\infty} \mathbb{P}^N \left[(1-\lambda_o)T_o^N - \log(1-\lambda_o)I_0 \le w\right] = e^{-e^{-w}}.$$

- ***Case 2.1:*** $I_0 = O(1)$, $R_0 = r_0 N$, $r_0 > 0$, $\lambda_o = \lambda_o(N) \le 1$ *and* $\gamma_o = o(1)$. *Let* $a := \lim_{N\to\infty} 1-\lambda_o+\lambda_o r_0$, *then*
$$\lim_{N\to\infty} \mathbb{P}^N \left[T_o^N \le w\right] = \left(1+\frac{a}{e^{aw}-1}\right)^{-I_0}.$$

- ***Case 2.2:*** $I_0 \to \infty$, $R_0 = r_0 N$, $r_0 > 0$, *and there exists* $\epsilon_1, \epsilon_2 > 0$ *such that*

$I_0 = o\left(N^{1-\epsilon_1}\right)$ *and* $\gamma_o = o\left(N^{-\epsilon_2}\right)$. *Let* $a := \lim_{N\to\infty} 1 - \lambda_o + \lambda_o r_0$, *then*

$$\lim_{N\to\infty} \mathbb{P}^N \left[aT_o^N - \log(aI_0) \le w\right] = e^{-e^{-w}}.$$

Theorem 4.6 covers the entire domain $\{(\lambda_o, \gamma_o) \in \mathbb{R}_+^2 : \lambda_o \le 1\}$, and a subset of $\{(\lambda_o, \gamma_o) \in \mathbb{R}_+^2 : \lambda_o \ge 1\}$. The range of the latter is a function of the initial state (I_0, R_0).

Theorem 4.6 suggests that for 'small' I_0, $\{(\lambda_o, \gamma_o) \in \mathbb{R}_+^2 : \lambda_o \le 1\}$ can be divided into two regimes, the boundary of which is illustrated by the blue dotted line in Figure 1.1. For details on Figure 1.1 and the definition of $\langle\cdot\rangle$, see Section 2.3.2.

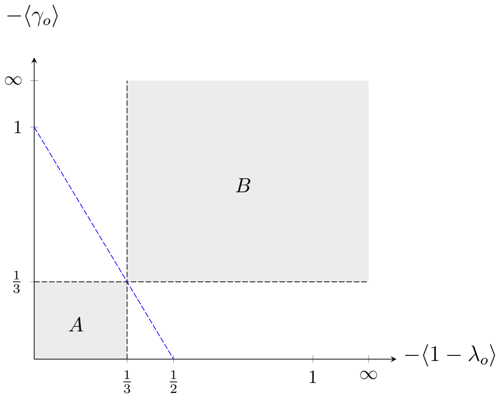


Figure 1.1: The division of the parameter space of SIRS models

In Figure 1.1, the regime below the blue line represents

$$\left\{(\lambda_o, \gamma_o) : N^{1/2}(1-\lambda_o)\gamma_o^{1/2} \to \infty\right\},$$

where given suitable (I_0, R_0), we can observe the behaviours of all five cases in Theorem 4.6. In the complement regime

$$\left\{(\lambda_o, \gamma_o) : N^{1/2}(1-\lambda_o)\gamma_o^{1/2} < \infty\right\},$$

only Cases 1.1, 2.1 and 2.2 can be observed.

The asymptotic distributions of the cases where $I_0 \to \infty$ are extreme value distributions. The intuition is that when the size of the infected population is small, infected individuals induce almost independent epidemics. The extinction time can be viewed as the maximum extinction time among all these local epidemics.

Summary of Chapter 5

In this chapter, we first identify the critical scaling of stochastic SIRS models in both time and space through a heuristic argument. Under the critical scaling, the scaled parameter space $(\hat{\lambda}, \gamma)$ is defined as

$$\hat{\lambda} := (1 - \lambda_o)N^{1/3}, \quad \gamma := \gamma_o N^{1/3}, \tag{1.2}$$

and the scaled stochastic SIRS model $(Y_t^N, Z_t^N)_{t\geq 0}$ is defined as

$$Y_t^N := \frac{I^N_{N^{1/3}t}}{N^{1/3}}, \quad Z_t^N := \frac{R^N_{N^{1/3}t}}{N^{2/3}}. \tag{1.3}$$

We also define the scaled extinction time $T^N = \inf\{t : Y_t^N = 0\}$.

The main results of this chapter concern the case where the scaled parameters and the scaled initial states are of $O(1)$. This is illustrated in Figure 1.1 as the shaded area B. In the first half of Chapter 5, we will show that (Y^N, Z^N) converges in distribution to a limit diffusion (Y, Z) and as do their extinction times, i.e., $T^N \Rightarrow T := \inf\{t : Y_t = 0\}$. The only analogous result available in the existing literature is by Foxall [6], who proved the same convergence for stochastic SIS models.

Compared to [6], we need to take one step further and make sure the limit diffusion is indeed well-defined, since the limit generator is not elliptic and has unbounded, non-Lipschitz coefficients. Fortunately, Brunick [9] has studied a type of degenerate martingale problems related to our limit diffusion, and our statement can be proved by a standard localisation argument [10].

Formally, the first half of the main results is stated as follows.

Theorem [*5.7*]**.** *Let* $(1 - \lambda_o(N))N^{1/3} \to \hat{\lambda} \in \mathbb{R}$, $\gamma := \lim_{N\to\infty} \gamma_o(N)N^{1/3} \geq 0$ *and* $Y_0^N \to y_0 > 0$, $Z_0^N \to z_0 > 0$. *Then the process* (Y^N, Z^N) *converges in distribution to* (Y, Z) *in* $D([0, \infty), \mathbb{R}^2)$, *where* (Y, Z) *is the unique weak solution to the stochastic* *differential equation system*

$$\begin{aligned} dY &= -(\hat{\lambda} + Z)Y\,ds + \sqrt{2Y}\,dW, \\ dZ &= (Y - \gamma Z)ds, \end{aligned} \tag{1.4}$$

with initial conditions $Y_0 = y_0$, $Z_0 = z_0$.

Theorem [*5.8*]**.** *Let* $(Y_0^N, Z_0^N) = (u_N, v_N)$ *and* $(Y_0, Z_0) = (u, v)$*. If* $(u_N, v_N) \to (u, v) \in \mathbb{R}_+^2$*, then* $T^N \Rightarrow T$*.*

In the second half of Chapter 5, we study the distribution of T. There is no known analogous result for stochastic SIS and SIR models.

It is a standard practice to express the distribution of the hitting time of a diffusion as the solution to a Cauchy-Dirichlet problem. However, the well-posedness of said problem does not directly follow from the well-posedness of the martingale problem on domain $C_c^\infty(\mathbb{R}_+^2)$, since the domain is not dense in $\widehat{BC}(\mathbb{R}_+^2)$ with respect to the uniform topology, where $(\widehat{BC}(\mathbb{R}_+^2), \|\cdot\|)$ is the Banach space of bounded continuous functions with continuous extensions to $[0, \infty)^2$. We construct the solution using a generalised Chernoff product formula proved by [11].

Theorem [*5.9*]**.** *Let* $V(t)$ *be a bounded linear operator on* $\widehat{BC}(\mathbb{R}_+^2)$ *for each* $t > 0$*, such that*

$$V(t)f(u, v) := \int_0^\infty g(t, ue^{-(\hat{\lambda}+v)t}; m) f(m, ve^{-\gamma t} + ut)\, dm,$$

for $t \geq 0$ *and* $f \in \widehat{BC}(\mathbb{R}_+^2)$*, where*

$$g(t, u; m) = \frac{1}{t} u^{1/2} m^{-1/2} e^{-(u+m)/t} I_1\left(\frac{2m^{1/2}u^{1/2}}{t}\right), m, u, t > 0,$$

and $I_1(\cdot)$ *is defined as* (5.7)*. Define*

$$U_n(u, v, t) := \left(V\left(\frac{t}{n}\right)\right)^n \mathbf{1}_{\mathbb{R}_+^2}(u, v).$$

The tail distribution of T*, i.e.* $\mathbb{P}\left[T > t \middle| (Y_0, Z_0) = (u, v)\right]$*, for each* $t > 0$*, is the limit of* $U_n(u, v, t)$ *as* $n \to \infty$*, for* $(u, v) \in \mathbb{R}_+^2$ *uniformly on compacts.*

Summary of Chapter 6

Chapter 6 focuses on the subcritical regime when the initial size of the infected population is too large to meet the assumptions in Theorem 4.6. We show that with suitable assumptions on (I_0, R_0), the trajectory of the stochastic SIRS epidemic can be well-approximated by the solution to some ODE systems, until a time when Theorem 4.6 is applicable. We can derive the asymptotics of the time taken for the corresponding ODE systems to travel between two given states (we will refer to this as '*elapsed time*'). The asymptotic distribution is then derived by shifting the asymptotic distribution in Theorem 4.6 according to the elapsed time.

The analogous results for stochastic SIR models can be found in [1] and the results for stochastic SIS models can be found in [6, 8]. In all these cases, the approximating ODEs are the corresponding deterministic epidemic models, where 'corresponding' means sharing the same parameters and initial states. The descriptions of the deterministic epidemic models are introduced in Section 2.2.2 to 2.2.4.

However, things are more complicated in stochastic SIRS models. Theorem 6.3 states that when γ_o is small and $I_0^N \asymp N$, the limit ODE is the corresponding deterministic SIR model. The variable $t_{\text{SIRS}}\,(a \to b)$ is the elapsed time of the deterministic SIRS model, and k_{SIR} and k_{SIRS} are the constants in the asymptotics of the elapsed time of the deterministic SIR and SIRS model respectively. The locations of their precise definitions are included in the statements of the main results of this chapter.

Theorem [*6.3*]**.** *Consider the stochastic SIRS model defined in* (1.1) *with parameters* $\lim_{N\to\infty} \lambda_o(N) =: \lambda_{lim} \leq 1$ *and* $\gamma_o = o\left(N^{-\epsilon_\gamma}\right)$ *for some* $\epsilon_\gamma > 0$*, and initial states*

$$\lim_{N\to\infty} I_0^N/N > 0, \ \lim_{N\to\infty} R_0^N/N \geq 0.$$

Then we have

$$\mathbb{P}\left[(1-\lambda_{lim}\theta^*_{lim})T_o^N - k_{\text{SIR}} - \log N - \log(1-\lambda_{lim}\theta^*_{lim}) \leq w\right] \to e^{-e^{-w}}.$$

where $\theta^*_{lim} = \lim_{N\to\infty} \theta^*(x_0^N, y_0^N; \lambda_o)$*, and* $k_{\text{SIR}} = k_{\text{SIR}}\left(\frac{N-I_0-R_0}{N}, \frac{I_0}{N}; \lambda_o\right)$ *are defined in Lemma 6.1.*

The second main result of this chapter, Theorem 6.7, concerns the parameter regime $\{(\lambda_o, \gamma_o) : 1 - \lambda_o \gg N^{-1/3}, \gamma_o \gg N^{-1/3}\}$. This regime is illustrated in Figure 1.1 as the shaded area A. The strongly subcritical case is a special case of Theorem 6.7 and is stated separately in Theorem 6.12. The reason why we state these two theorems separately is that we are able to obtain the exact asymptotics of $t_{\text{SIRS}}\left(I_0 N^{-1} \to a\right)$ only for the strongly subcritical case.

Theorem [*6.7*]**.** *Consider the stochastic SIRS model defined in* (2.5) *with parameters* $\lambda_o = \lambda_o(N) \uparrow 1$ *and* $\gamma_o = \gamma_o(N) \downarrow 0$*, satisfying* $(1-\lambda_o)N^{1/3} \to \infty$*,* $\gamma_o N^{1/3} \to \infty$*, and* $\lim_{N\to\infty} \frac{1-\lambda_o}{\gamma_o} \neq 1$*. If* $\gamma_o \ll 1 - \lambda_o$*, we in addition require that there exists some small* $\epsilon_p > 0$ *such that* $N^{\frac{\epsilon_p}{3}} \gamma_o^{1+\epsilon_p} \ll 1 - \lambda_o$*.*

Suppose the initial states of the model satisfy for some constants $c_y, d_y, c_z > 0$ *the conditions*

$$I_0^N = I_0(N) \leq \left[d_y \frac{N(1-\lambda_o)\gamma_o}{\log(N^{1/3}(1-\lambda_o))}, c_y(1-\lambda_o)\gamma_o N\right],$$

$$R_0^N = R_0(N) \leq \begin{cases} c_z N \gamma_o^{1+\epsilon_p}, \text{ if } \gamma_o \ll 1-\lambda_o, \\ c_z(1-\lambda_o)N, \text{ otherwise.} \end{cases}$$

Then

$$\mathbb{P}\left[(1-\lambda_o)T_o^N - (1-\lambda_o)t_{\text{SIRS}}\left(I_0 N^{-1} \to a\right) - \log a - \log N(1-\lambda_o) \leq w\right] \to e^{-e^{-w}},$$

where $a = a(N) \geq N^{-1}$ *can be chosen arbitrarily, as long as* $a = o((1-\lambda_o)\gamma_o)$*. The asymptotic distribution above is independent of the choice of* a*.*

Theorem [*6.12*]**.** *Suppose* $\lim_{N\to\infty} \lambda_o(N) = \lambda_{lim} < 1$ *and* $\lim_{N\to\infty} \gamma_o(N) = \gamma_{lim} > 0$ *are constants independent of* N*,* $\lambda_{lim} + \gamma_{lim} \neq 1$*, and* $\lambda_o(N) + \gamma_o(N) \neq 1$ *for* $N \in \mathbb{N}$*.*

Suppose further that the initial condition satisfies

$$\lim_{N\to\infty} \frac{I_0^N}{N} > 0, \lim_{N\to\infty} \frac{R_0^N}{N} > 0.$$

Then we have as $N \to \infty$*,*

$$\mathbb{P}\left[(1-\lambda_o)T_o^N - (k_{\text{SIRS}} + \log N + \log(1-\lambda_o))\right] \to e^{-e^{-w}},$$

where k_{SIRS} *is defined as in Lemma 6.11.*

Summary of Chapter 7

We conduct numerical experiments in MATLAB to verify our results in Chapter 4 and 6. We also review and compare the available methods of simulation in this chapter. The simulation shows that the asymptotic distributions we have derived are a fairly good approximation of the simulated data.

Chapter 2

Setting up the background

In this chapter, we start by introducing the mathematical and epidemiology background, before rigorously defining the stochastic SIS, SIR, SIRS models and their deterministic counterparts. Next, we will review the available results on near-critical behaviours and extinction times of epidemic models, and state precisely the objective of this thesis. Lastly, we introduce the diagrams we use throughout this thesis to illustrate the regimes of the parameter space and the initial state space.

2.1 Continuous-time Markov chains

Consider a continuous-time càdlàg Markov chain $X = (X_t)_{t\geq 0}$ on $(\Omega, \mathscr{F})$ with a finite state space $S \subset \mathbb{R}^d$, whose natural filtration is $(\mathscr{F}_t)_{t\geq 0}$, and let $\mathbb{P}_x, x \in S$ be the corresponding probability measure. Such a process is defined by its initial state $X_0 = x_0$ and the transition rates from state x to $x + j$, denoted as $q(x, j), j \in J$, where J is the set of possible increments/decrements X can have in the following sense:

$$\mathbb{P}_{x_0}\left[X_{t+\Delta t} = x + j | X_t = x\right] = q(x, j)\Delta t + o(\Delta t).$$

Denote a jump at time t as $\Delta X_t := X_t - X_{t-}$, then we can define random measures on $[0,\infty) \times J$ as

$$\mu := \sum_{t:\Delta X_t \neq 0} \delta_{(t,\Delta X_t)} \text{ and } \upsilon(dt,j) := q(X_{t-},j)dt.$$

Thus we can write

$$X_t = X_0 + \int_0^t \sum_{j\in J} j\mu(ds,j).$$

Let $(H(t,j))_{t\geq 0}$ be a left-continuous adapted process for each $j \in J$. It is known that if H satisfies for all $t \geq 0$,

$$\mathbb{E}\left[\int_0^t \sum_{j\in J} |H(s,j)|\upsilon(ds,j)\right] < \infty,$$

then it is known that

$$M_t = \int_0^t \sum_{j\in J} H(s,j)(\mu - \upsilon)(ds,j)$$

is a well-defined martingale (see e.g. Theorem 8.4, [12]).

It follows that we can decompose X as

$$X_t = x_0 + \int_0^t \sum_{j\in J} jq(X_{s-},j)ds + M_t, \tag{2.1}$$

where the martingale part $M_t := \int_0^t \sum_{j\in J} j(\mu - \upsilon)(ds,j)$ is called the *compensated martingale*, and $\int_0^t \sum_{j\in J} jq(X_{s-},j)ds$ is called the *compensator*.

Alternatively, we can view the process X as driven by an embedded discrete-time Markov chain $Y = (Y_n)_{n\in\mathbb{N},n\geq 0}$ and a sequence of holding times $\{S_n\}_{n\in\mathbb{N},n\geq 1}$.

Let $q(x) := \sum_{j\in J} q(x,j)$ and $Y_0 = x_0$. For $n \geq 0$, let Y follow the transition probability

$$\mathbb{P}_{x_0}\left[Y_{n+1} = x + j \middle| Y_n = x\right] = \pi(x,j) := \frac{q(x,j)}{q(x)}, \quad j \in J,$$

and $S_{n+1} \sim \text{Exp}\,(q(Y_n))$, the exponential distribution with parameter $q(Y_n)$. Then we can define $(X_t)_{t\geq 0}$ as $X_t = Y_n$ for $\sum_{i=1}^{n} S_i \leq t < \sum_{i=1}^{n+1} S_i$, $n \geq 1$, and $X_t = Y_0$ for $0 \leq t < S_1$.

If the finite state space S can be decomposed as a set of transient states $S \setminus \{0\}$ and an absorbing state $\{0\}$, then X will absorb at 0 almost surely. We can define the absorption time $T_X := \inf\{t : X_t = 0\}$.

Let $P_{mn}(t) := \mathbb{P}\left[X_{s+t} = n | X_s = m\right]$. The Q-matrix of X is defined as $Q = (Q_{mn})_{m,n\in S}$, where

- $Q_{mn} = q(m, n-m)$, for $m \neq n$, and
- $Q_{mm} = -q(m)$.

The *Kolmogorov forward equations* state that $P(t) = (P_{mn}(t))_{m,n\in S}$, $t \geq 0$, is the solution of

$$\frac{dP(t)}{dt} = P(t)Q, \quad P(0) = I,$$

where I represents the identity matrix.

The Kolmogorov forward equations give the exact expression of many quantities of interest, among which, the exact form of the distribution of T_X can be expressed as

$$\mathbb{P}\left[T_X \leq t \middle| X_0 = m\right] = P_{m0}(t).$$

The reader can find more details in [13].

2.2 Epidemic models

The SIRS epidemic models are mainly used to describe the behaviours of micro-parasite infections of humans. Throughout this thesis, we also need the definitions of two simpler models, SIS and SIR. In this section, we will introduce all three model structures, which are characterised by the different natural history of the infections (in other words, the journey a typical patient goes through). Of each model structure, we introduce the deterministic version and the stochastic version.

These models are known as *compartmental models*, which means that the individuals in the population are divided into broad subgroups and the model tracks individuals collectively.

For a comprehensive reading on epidemic modelling tailored for public health practitioners, one can refer to [14].

2.2.1 Background

In this subsection, we will introduce the basic concepts used in compartmental epidemic models. Many of these are widely used in epidemic and ecological modelling. Although usually considered oversimplified compared to the reality, the constructions below allow us to carry out more complicated analytical study.

Assumption 2.1 (Target population)**.** We assume that our target population is

- *closed / without demography*, i.e., there is no birth, death, migration into or out of the population; and
- *homogeneously mixing*, i.e., all individuals are considered to be identical. All individuals are assumed to be making *effective contact* with an arbitrary member of the rest of the population at equal rates.

In the stochastic version, we assume the population has a finite size $N \in \mathbb{N}$. This allows us to construct a sequence of stochastic models indexed by the parameter N.

In the deterministic version, we assume a continuum population, with compartmental variables interpreted as asymptotic proportions in a finite population as $N \to \infty$. Therefore, it makes sense to use the 'proportion', instead of the 'number' to measure the size of each subgroup. In the following, we construct the deterministic version as the average scenario when the population size tends to infinity.

The population is divided into some or all of the following compartments:

- Susceptible (S), which consists of individuals who are currently healthy but can get infected;
- Infectious/Infected (I), which consists of individuals who are infected by the disease and can transmit the infection to others;

- Recovered/Immune (R), which consists of individuals who are immune to the disease.

Another basic compartment studied in the literature is Pre-infectious (also called 'Exposed', E), consisting of individuals who are infected but not yet infectious.

Assumption 2.2 (Parameter space)**.** Assuming the target population has all three compartments S, I and R, the movement of a typical patient in the target population is determined by three parameters λ_o, μ_o and γ_o:

- The terminologies and verdicts below follow from [15].

  Each susceptible individual is expected to contract the disease at rate $\lambda_o I/N$, where I denotes the current size of infection. This rate is known as *force of infection*. Parameter $\lambda_o \in \mathbb{R}_+$ is a composite measure of contact rates and transmission probability, usually known as *transmission rate*. The assumption where the force of infection is assumed to depend on the proportion of infection, rather than the size of infection, is called *frequency dependent transmission*. We note, however, that many mathematical works (e.g., [16]) refer to the stochastic models under this assumption as 'density dependent process'. Frequency dependent transmission is usually considered a reasonable assumption for vector-borne diseases in human societies.

- Each infectious individual recovers at *recovery rate* μ_o independently. Its reciprocal $1/\mu_o$ is the average infectious period. Without loss of generality, from Section 2.2.3 onward, we assume $\mu_o = 1$. To illustrate how the deterministic and the stochastic models are defined, we keep μ_o arbitrary in Section 2.2.2.

- Each recovered individual loses immunity at the *rate of waning immunity* $\gamma_o \in [0, \infty)$. Its reciprocal $1/\gamma_o$ is the average period of immunity.

As we will see in the formulation of the SIS models below, the 'rates' above are understood in the context of continuous-time Markov chains when constructing the stochastic models. The deterministic version of each model can be heuristically interpreted as the limiting average trajectory of the stochastic version as $N \to \infty$.

The natural history of an infectious disease is reflected in different models by the transition route of a typical individual between different compartments. The idea of compartmental models is versatile in application. By introducing appropriate compartments and transition assumptions, we can address the heterogeneity in sociodemographic factors, contact structures, or the complexity in the transmission modes (e.g. vector-borne diseases) [17].

2.2.2 Without immunity: the SIS models

The SIS models assume that the state of a typical individual follows the pattern 'Susceptible - Infectious - Susceptible'. It is often used for curable sexually transmitted infections, for which the individual will gain negligible immunity following the infection.

Indexed by the population size N, the stochastic SIS model is defined as a continuous-time Markov chain I^N valued in $[N]$, representing the size of the Infectious compartment, with transition rates:

$$\begin{aligned} &i \to i+1, \text{ at rate } \lambda_o(1-i/N)i, \\ &i \to i-1, \text{ at rate } \mu_o i. \end{aligned}$$

The deterministic SIS model can be derived as follows:

Define the proportion of the Infectious compartment in the population at time t as $y(t)$:

$$y(t) := \lim_{N\to\infty} \frac{I_t^N}{N}.$$

In a small time interval $[t, t+\Delta t]$, the number of individuals moving from Susceptible to Infectious is,

$$S_t^N \times \text{per capita force of infection in } [t, t+\Delta t] \times \Delta t = \frac{\lambda_o S_t^N I_t^N}{N} \Delta t. \tag{2.2}$$

Similarly, the number of individuals moving from Infectious to Susceptible is $\mu_o I_t^N \Delta t$. Dividing by N on both sides of above and (2.2), and letting $N \to \infty$, we have

$$y(t+\Delta t) - y(t) = \lambda_o(1-y(t))y(t)\Delta t - \mu_o y(t)\Delta t.$$

We can relate the difference equation above to a one-dimensional ODE by taking the time step $\Delta t \to 0$:

$$\frac{dy}{dt} = \lambda_o(1-y)y - \mu_o y = (\lambda_o - \mu_o)y\left(1 - \frac{\lambda_o}{\lambda_o - \mu_o}y\right), \quad y(0) = y_0 \in (0, 1].$$

This equation is known as the *logistic equation* and is often used to model the density dependent population growth in ecology. It can be solved by separation of variables. The solution is: when $\lambda_o/\mu_o \neq 1$,

$$y(t) = \frac{\lambda_o - \mu_o}{\lambda_o}\left(1 + \frac{\lambda_o - \mu_o - \lambda_o y_0}{\lambda_o y_0} e^{(\mu_o - \lambda_o)t}\right)^{-1}, \quad t \geq 0,$$

and when $\lambda_o/\mu_o = 1$,

$$y(t) = \lambda_o^{-1}(t + y_0^{-1})^{-1}, \quad t \geq 0.$$

The long-term behaviour of $y(t)$ depends on the value of λ_o:

- when $\lambda_o/\mu_o \leq 1$, $\lim_{t\to\infty} y(t) = 0$, which indicates the extinction of the epidemic;
- when $\lambda_o/\mu_o > 1$, $\lim_{t\to\infty} y(t) = 1 - \frac{\mu_o}{\lambda_o}$, which indicates the prevalence of the epidemic. Such limit is often referred to as the *endemic equilibrium*.

It turns out that $\lambda_o/\mu_o = 1$ is the critical value dividing a quick extinction and an endemic for all three models in this chapter under Assumption 2.1. In the context of epidemiology, $\mathscr{R}_0 := \lambda_o/\mu_o$ is called *the basic reproduction number*. The basic reproduction number describes the ability of a disease to prevail and can be loosely interpreted as the average number of cases caused by an infectious individual during his/her entire infectious period in an entirely susceptible population [3].

2.2.3 Immunising infections: the SIR models

The SIR models assume that a typical individual follows the pattern Susceptible - Infectious - Recovered. The individuals in the Recovered compartment either have gained permanent immunity or have been removed from the population. Besides ‘immunising infections’ (i.e., those for which individuals gain permanent immunity), SIR models are also used to model infections with waning immunity (e.g., influenza, COVID-19) during a short time period after it is introduced to the population.

The stochastic SIR model is a continuous-time Markov chain (I^N, R^N) valued in $[N] \times [N]$, where I_t^N and R_t^N are the sizes of Infectious/Infected and Recovered/Immune compartments at time t respectively, with the transition rates:

$$\begin{aligned}
(i, r) &\to (i+1, r), && \text{at rate } \lambda_o (N - i - r) i / N, \\
(i, r) &\to (i-1, r+1), && \text{at rate } i.
\end{aligned}$$

Similar to the previous section, the deterministic SIR model is described by the ODE system

$$\begin{aligned}
&\frac{dx}{dt} = -\lambda_o y x, \\
&\frac{dy}{dt} = \lambda_o y x - y, \\
&\frac{dz}{dt} = y, \\
&(x(0), y(0)) = (x_0, y_0) \in [0,1) \times (0,1], \; x(0) + y(0) + z(0) = 1,
\end{aligned} \tag{2.3}$$

where x, y, z denote respectively the proportion of the size of S, I, R compartments in the target population.

These variables are dependent through the relation $x(t) + y(t) + z(t) \equiv 1$. Sometimes we find it more convenient to use (x, y)-coordinates while in other occasions we prefer (y, z)-coordinates.

From the ratio of the first and the third equations in (2.3), we have

$$\begin{aligned}
&\frac{dx}{dz} = -\lambda_o x, \\
&x(t) = x(0) e^{-\lambda_o (z(t) - z(0))} \geq x(0) e^{-\lambda_o},
\end{aligned}$$

from which we can express $x(t)$ in terms of $y(t)$, for all $t \geq 0$.

In particular, given a solution $(x(t), y(t))$ to (2.3), for all t such that $y(t) \leq y_0$, we can represent the value of $x(t)$ when $y(t) = a$ by the following mapping:

$$\theta : (0, y_0] \to [0,1], \; a \mapsto x \circ (y)^{-1}(a).$$

Notice that $\lambda_o \le 1$, we have $y(t) \le y_0$ for all $t \ge 0$.

Also it follows that for all $\lambda_o \in \mathbb{R}_+$, $\lim_{t\to\infty}(x(t), y(t)) = (\theta^*, 0)$ where $\theta^* := \lim_{a\downarrow 0}\theta(a)$ is a function of (x_0, y_0) and λ_o.

For numerical analysis, we need the following result:

Lemma 2.3 (A simple property of θ)**.** *The function θ is differentiable on $(0, y_0)$, and can be expressed in terms of the principal branch of Lambert W function W_0,*

$$\theta(y(t)) = -\frac{W_0(-x_0\lambda_o e^{-\lambda_o(x_0+y_0-y(t))})}{\lambda_o}. \tag{2.4}$$

In particular,

$$\theta^*(x_0, y_0; \lambda_o) := \lim_{a\downarrow 0}\theta(a) = -\frac{W_0(-x_0\lambda_o e^{-\lambda_o(x_0+y_0)})}{\lambda_o} \in [0, 1].$$

Proof. From $x(t) = x(0)e^{-\lambda_o(z(t)-z(0))}$, we have

$$-\lambda_o\theta(y)e^{-\lambda_o\theta(y)} = -\lambda_o x_0 e^{-\lambda_o(x_0+y_0)}e^{\lambda_o y}.$$

The rest of the statement follows from the definition and property of Lambert W function, which can be found in e.g. [18]. □

In the long term, all infectious individuals in this model will gain immunity. Therefore, unlike in the SIS models, we need to use the *final size of infection* to indicate whether the epidemic dies out quickly. The final size of infection $z(\infty)$ is defined as

$$z(\infty) := \lim_{t\to\infty} z(t) = 1 - \theta^*.$$

The criticality is at $\lambda_o = 1$ in the sense that when $x_0 \uparrow 1$, if $\lambda_o < 1$, then $z(\infty) \to 0$, whereas if $\lambda_o > 1$, then $z(\infty) > 0$. In other words, an epidemic outbreak from a single infectious individual will take place only when $\lambda_o > 1$.

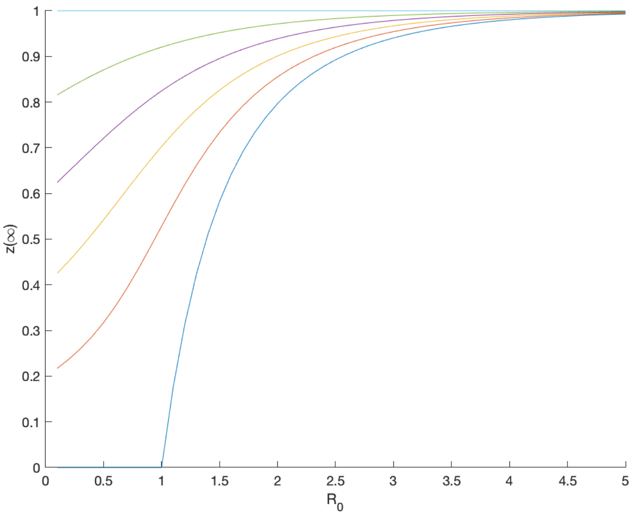


Figure 2.1: The value of $z(\infty)$ as a function of $\mathscr{R}_0$. Six lines from top to bottom: when $y_0 = 1, 0.8, 0.6, 0.4, 0.2, 0$.

2.2.4 Waning immunity: the SIRS models

The SIRS models assume that a typical individual moves according to the pattern 'Susceptible - Infectious - Recovered - Susceptible'. In contrast to the SIR model, the SIRS model assumes that immune individuals eventually lose their immunity and become susceptible again. A wide range of infections belong to this category, especially when being observed over a long time period.

With the same notations for the SIR model, the stochastic SIRS model is formulated as a two-dimensional continuous-time Markov chain (I^N, R^N), with the transition rates:

$$\begin{aligned} (i, r) &\to (i+1, r), && \text{at rate } \lambda_o(N-i-r)i/N, \\ (i, r) &\to (i, r-1), && \text{at rate } \gamma_o r, \\ (i, r) &\to (i-1, r+1), && \text{at rate } i. \end{aligned} \tag{2.5}$$

The deterministic SIRS model is defined as the ODE system

$$\begin{aligned}
\frac{dx}{dt} &= \gamma_o z - \lambda_o(1-y-z)y,\\
\frac{dy}{dt} &= \lambda_o(1-y-z)y - y,\\
\frac{dz}{dt} &= y - \gamma_o z,\\
(y(0), z(0)) &\in (0,1]\times[0,1), \quad x(0)+y(0)+z(0) = 1.
\end{aligned} \tag{2.6}$$

The criticality of (2.6) is at $\lambda_o = 1$ in the sense illustrated by the following theorem.

Theorem 2.4. *Consider the ODE system* (2.6).

For $\lambda_o \leq 1$*, the disease-free equilibrium* $(1,0,0)$ *is globally asymptotically stable.*

For $\lambda_o > 1$*, the endemic equilibrium*

$$(x^*, y^*, z^*) := \big(\lambda_o^{-1}, (1-\lambda_o^{-1})(\gamma_o+1)^{-1}\gamma_o,\ (1-\lambda_o^{-1})(\gamma_o+1)^{-1}\big)$$

is globally asymptotically stable.

Proof. We will prove the theorem using the (x,y)-coordinate of the system (2.6).

The construction of Lyapunov functions used below follows [19]. The proof follows from the global LaSalle's principle (Theorem 5.25, p.204, [20]).

The domain $\{(x,y) : x \geq 0, y \geq 0, x+y \leq 1\}$ is an invariant set. For the case $\lambda_o \leq 1$, we choose the Lyapunov function to be

$$V(x,y) = -(1-x-y) - \log x.$$

For the case $\lambda_o > 1$, we choose the Lyapunov function to be

$$\hat{V}(x,y) := x - x^* - \left(x^* + \frac{\gamma_o}{\lambda_o}\right)\log\frac{x+\frac{\gamma_o}{\lambda_o}}{x^*+\frac{\gamma_o}{\lambda_o}} + y - y^* - y^*\log\frac{y}{y^*}.$$

Both V and $\hat{V}$ are smooth and positive definite on their domain, and their global minimums satisfy

$$\inf_{(x,y)} V(x,y) = V(1,0) = \inf_{(x,y)} \hat{V}(x,y) = \hat{V}(x^*,y^*) = 0.$$

Next we check that

$$V'(x,y) := \frac{dV(x(t),y(t))}{dt} = -(1-\lambda_o)y - \gamma_o(1-x-y)(x^{-1}-1) \leq 0,$$

and

$$\Big\{(x,y) : V'(x,y) = 0\Big\} = (1,0).$$

Similarly,

$$\begin{aligned}\hat{V}'(x,y) :=& \frac{d\hat{V}(x(t),y(t))}{dt} = \frac{dx}{dt}\left(1 - \frac{x^* + \frac{\gamma_o}{\lambda_o}}{x + \frac{\gamma_o}{\lambda_o}}\right) + \frac{dy}{dt}\left(1 - \frac{y^*}{y}\right) \\ =& -\frac{\gamma_o(\lambda_o+\gamma_o)}{(x+\frac{\gamma_o}{\lambda_o})(1+\gamma_o)}(x-x^*)^2 \leq 0.\end{aligned}$$

The set

$$S_{\hat{V}} := \Big\{(x,y) : \hat{V}'(x,y) = 0\Big\} = \{(x,y) : x = x^*\},$$

contains no other trajectory except for the trivial trajectory $(x(t),y(t)) \equiv (x^*,y^*)$, since $dx/dt \neq 0$ for any point $(x,y) \neq (x^*,y^*)$ in $S_{\hat{V}}$.

Thus by the global LaSalle's principle (Theorem 5.25, p.204, [20]), the system is globally asymptotically stable. □

2.2.5 Deterministic models vs stochastic models

In the deterministic models, we assume that the randomness of the real world can be 'averaged out', in order to reveal the underlying disease dynamics. The stochastic models, however, aim to reflect this randomness.

In general, there are three ways to incorporate randomness into epidemic models [15]:

- adding random terms to population variables,
- defining parameters as random, and
- explicitly modelling individual-level events as random.

The third method is more popular and is the idea behind the stochastic epidemic models we introduced in the last three subsections.

The deterministic and stochastic models introduced above model the persistence behaviour of epidemics very differently. This is shown in two quantities that attract a lot of mathematical attention: the endemic state and the duration of the epidemic. An epidemic modelled as a deterministic model goes extinct for all SIR models and for SIS and SIRS models when $\mathscr{R}_0 \leq 1$. On the other hand, for SIS and SIRS models when $\mathscr{R}_0 > 1$, the epidemic prevails with a constant proportion of the population being infected. In other words, it reaches an endemic equilibrium. In this sense, it is difficult to define the duration of an epidemic in the deterministic model, and we shall turn to the stochastic version for help.

The stochastic models allow us to model the duration of the epidemics as hitting times of continuous-time Markov chains. This is shown to be effective in interpreting real-life data; for example, Broadfoot [21] studies the single-farm and inter-farm persistence of foot-and-mouth disease in livestocks, using both homogeneous mixing SIR models and SIR models on various graphs.

Since our models all have finite state spaces, the extinction will happen in finite time almost surely. To define a non-trivial endemic state for the stochastic models, the concept of *quasi-stationary distribution* is introduced. The quasi-stationary distribution of the stochastic SIS model is defined as the stationary distribution of I_t^N conditioned on that the extinction has not occurred, and was first investigated in 1960s by [22]. It can be obtained through an iterative scheme (See [4]).

The stochastic models and their deterministic counterparts are related in that: stochastic epidemic models can be well-approximated by their respective deterministic counterpart up to any constant time, in a sense that will be introduced in Section 3.1.1.

2.2.6 Near-critical behaviours

In previous sections, we identified the criticality of all three epidemic models and observed that the epidemic tends to die out when the transmission rate λ_o is below the criticality $\lambda_o = 1$, and tends to spread when λ_o is above the criticality. We refer to this phenomenon as '*phase transition*'.

As pointed out by [23], one of the arising challenges for stochastic epidemic models is the study of near-critical behaviours. They argued via examples that many epidemics of interest are neither strongly supercritical nor strongly subcritical, especially when the epidemic is still emerging, or is close to elimination under eradication effort. For example, a pathogen strain may switch from subcritical to supercritical, due to genetic or environmental change, and thus causes an outbreak. Conversely, a vaccination programme may push the transmissibility of a pathogen strain in the opposite direction.

By defining the parameters as functions of population size N, we are able to describe the near-critical behaviours with precision. We say the model is in *near-critical regime* if $\lambda_o(N) \to 1$ as $N \to \infty$. We say the model is *subcritical* if $\lambda_o(N) < 1$ for all $N \in \mathbb{N}$ and is *strongly subcritical* if $\lim_{N\to\infty} \lambda_o(N) < 1$.

It has been found in both stochastic SIS and SIR models that phase transition can be observed in a subset of the near-critical regime, which is often referred to as 'critical window', 'transition region' or 'critical regime'. In this thesis, we adopt the name *critical parameter regime*.

For the stochastic SIS model, Nåsell [4] observes a phase transition in the quasi-stationary distribution across the near-critical regime at $\lambda_o = 1 + cN^{-1/2}$, $c > 0$. From a diffusion approximation point of view, the same critical regime scaling $|\lambda_o - 1| \asymp N^{-1/2}$ is identified by Dolgoarshinnykh and Lalley [24] and Foxall [6]. Though sharing the same scaling with [4], the authors of [24] do not believe that there is a direct link between the two phenomena. From the perspective of extinction times, Doering, Sargsyan, and Sander [5] show that the Fokker-Planck equation provides an estimation with $O(1)$-error to the expected extinction time of the supercritical SIS model, applicable only when $\lambda_o = 1 + O(N^{-1/3-\epsilon})$ for some $\epsilon > 0$. Later, Foxall [6] proves that for $\lambda_o = 1 + O(N^{-1/2})$, the extinction time of an SIS model converges in distribution to the hitting time to 0 of its limit diffusion.

Dolgoarshinnykh and Lalley [24] also identify the critical parameter regime as $|\lambda_o - 1| \asymp N^{-1/3}$ for the SIR model, using the same diffusion approach as the one they used for SIS models. However, they didn't give a rigorous proof of their statement.

2.3 Objectives

There is no known result on the near-critical behaviours of the stochastic SIRS models. In this thesis, we study this problem from the perspective of extinction time of epidemics.

The extinction times of the three compartmental epidemic models above can all be defined as

$$T_o^N := \inf\{t : I_t^N = 0\}.$$

The extinction time provides valuable perspective to disease control and prevention policy and thus has received constant research interest. In mathematical literature, we have a lot of results regarding the stochastic SIS [2,3,6,8,25] and SIR models [1].

In particular, we would like to know how the distribution of T_o^N scales with $N, (\lambda_o, \gamma_o)$ and (I_0, R_0). That is, supposing that parameters (λ_o, γ_o) and initial states (I_0, R_0) are all given functions of N, we would like to find functions $f_1(N), f_2(N)$ such that the distribution of $\frac{T_o^N - f_1(N)}{f_2(N)}$ has a non-degenerate limit as $N \to \infty$.

Throughout the thesis, we assume λ_o, γ_o are finite as $N \to \infty$.

Estimating the duration T_o^N as a function of basic reproduction rate $\mathscr{R}_0 = \lambda_o$, immunity rate γ_o and population size N can help us understand the disease and suggest the control measures. For example, measles is observed to be prone to local extinction in small reasonably isolated communities, with a population of size N below some critical size. Empirical data suggests that the estimated mean period between measles outbreaks is of order $N^{-\frac{1}{2}}$ [26,27].

The exact expression of the distribution of the extinction time can be analysed using the corresponding Kolmogorov forward equations. However, the solution to the Kolmogorov forward equations can be algebraically cumbersome, and is not a straightforward indicator of how the extinction time is affected by various parameters. The available numerical simulation methods are also time-consuming when the target population is large. Therefore, many researchers have attempted to provide asymptotic results concerning the mean and distribution of the extinction time.

2.3.1 Literature review

There is few result on the extinction time of stochastic SIRS models. As far as we are aware, the only available result appears in [28], where it was shown that, given the strongly supercritcal case ($\mathscr{R}_0 > 1$) and $I_0^N, R_0^N \asymp N$, the extinction time is of order e^{cN} for some constant $c > 0$.

Nevertheless, the available results on the stochastic SIS and SIR models is a good indicator of what we shall expect.

The asymptotic distribution of the extinction time of the stochastic SIR model is obtained by [1] very early on. On the other hand, studies on SIS extinction time were mainly focusing on expectation (See [2–5]), until recent years when progress was made in the analogous results of the asymptotic distribution by [8].

Historically, for the strongly subcritical case ($\mathscr{R}_0 < 1$), Kryscio and Lefévre [2] attempt to derive the asymptotic distribution of the extinction time with 'large' I_0^N, but obtain an erroneous result as pointed out by [5]. We believe the error occurs because they directly quote the result of birth-death chain coupling of the SIR model in [1] when the method is not suitable for their assumptions. Andersson and Djehiche [29] show that with I_0^N being a constant, the SIS extinction time a.s. converges to the extinction time of a linear birth-death chain by suitable couplings. Nåsell [4] approaches this problem through the study of quasi-stationary distribution, and obtains the expected extinction time for both $I_0^N = 1$ and I_0^N at the quasi-stationary equilibrium.

For the strongly supercritical case $\mathscr{R}_0 > 1$, Andersson and Djehiche [29] prove that, if $I_0^N \asymp N$, then the extinction time weakly converges to an exponential distribution with an expectation of the order $N^{-1/2}e^{N(\log \mathscr{R}_0 - 1 + \mathscr{R}_0^{-1})}$. Later, Doering, Sargsyan and Sander [5] derive the same leading term for the expectation of the extinction time with the same initial state. They also show that the remaining term is of the order $N^{-3/2}e^{N(\log \mathscr{R}_0 - 1 + \mathscr{R}_0^{-1})}$.

The extinction times of models with non-classic assumptions have also been investigated. Here are a few examples: Nåsell [30] and Kamenev and Meerson [31] both attempt to remove the 'closed population' assumption by studying the SIS and SIR model with immigration and death rates respectively. Both models assume that the total populations are in steady states. Lopes and Luczak [32] extend the methodology of [8] to a two-dimensional classic SIS model. Ball, Britton and Neal [33] replace the Markovian assumption of the recovery of the classic SIS model to be any i.i.d. distributions, and study the expected extinction time. There are also various studies on heterogeneously-mixing models where spatial structure is introduced to the population, e.g. [33, 34].

2.3.2 Visualisation of assumptions

To better illustrate the various combinations of the scaling of the parameters and initial states, we introduce the notation $\langle\cdot\rangle$.

Definition 2.5. The mapping $\langle\cdot\rangle$ is defined as follows:

For any function $f = f(N)$, $\langle f\rangle = a \in \mathbb{R}$ if and only if $|f(N)| \asymp N^a$. If $|f(N)|$ tends to infinity faster then any polynomials, we say $\langle f\rangle = \infty$; and if $|f(N)|$ tends to infinity slower then any polynomials, we say $\langle f\rangle = 0+$ and $\langle 1/f\rangle = 0-$.

In this sense, we have the property that for any functions f, g,

$$|f(N)g(N)| \to 0 \iff \langle f\rangle + \langle g\rangle < 0 \text{ or } \langle f\rangle + \langle g\rangle = 0- .$$

Using this definition, we can describe the initial states by $(\langle I_0\rangle, \langle R_0\rangle) \in [0, 1]^2$ and the parameter regime by $(-\langle 1 - \lambda_o\rangle, -\langle\gamma_o\rangle) \in [0, \infty]^2$. One of the advantages of this set-up is that it can be effectively visualised through the diagrams.

For example, the sequence of stochastic SIRS models $\{(I^N, R^N)\}_{N\in\mathbb{N}}$ with constant parameters (λ_o, γ_o) and $I_0, R_0 \asymp N$ belongs to a family of model sequences. This family can be represented by a vector $(A_1, A_2) = (-\langle 1-\lambda_o\rangle, -\langle\gamma_o\rangle, \langle I_0\rangle, \langle R_0\rangle) \in \mathbb{R}^4$, where $A_1 = (0, 0)$ and $A_2 = (1, 1)$, and can be visualised as in Figure 2.2.

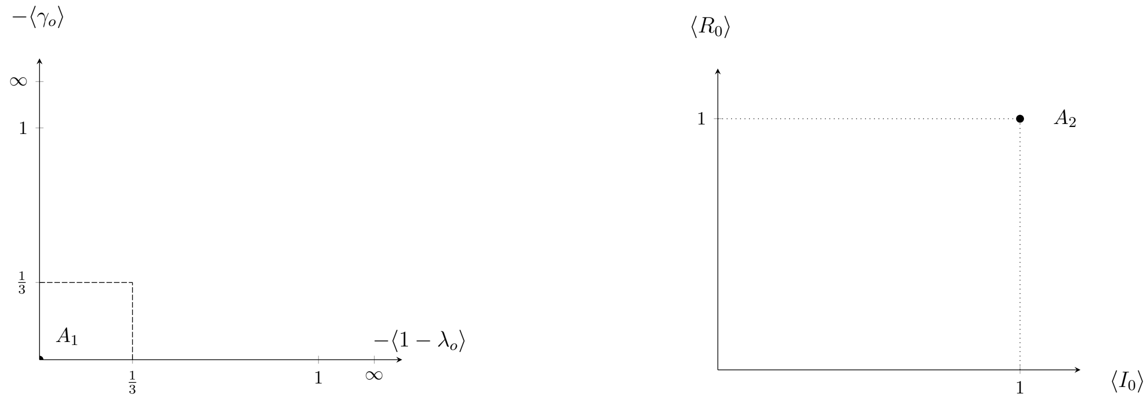


Figure 2.2: An example of our visualisation of assumptions

Chapter 3

Main techniques

In this chapter, we introduce the techniques used in this thesis. More specifically, the ODE approximation introduced in Section 3.1.1 is used in Chapter 4 and Chapter 6, the diffusion approximation and the theory of bi-continuous semigroups are used in Chapter 5, and the order-preserving coupling for birth-death chains is used in Chapter 4.

3.1 Sample path approximation

In the early days, a lot of works chose to study the extinction time by approximating the solution of Kolmogorov forward equations. Recently, more works have adopted the approach of sample path approximation. Sample path approximation seems to be the most fruitful approach so far, and has the advantage of providing an understanding of the entire trajectory of the epidemic.

In this section, we introduce two types of approximation: comparison to the solutions of ODEs and comparison to a diffusion limit.

Assumption 3.1 (Continuous-time Markov chains in finite population models). Consider a sequence of Markov chains indexed by N, valued in finite state space $S^N \subset \mathbb{R}^d$, and is denoted as $\{(X_t^N)_{t\geq 0}\}_{N\in\mathbb{N}}$. For each N, X^N is uniquely defined by its initial state $X_0^N = x_0^N \to x_0$ as $N \to \infty$, and transition rates $q^N(x, j)$, $j \in J^N$, where J^N is the set of possible jumps in column vectors. We assume the number of elements in J^N is finite and independent of N, and

$$\bar{j}^N := \max\left\{|j| : j \in J^N\right\} \to 0, \quad N \to \infty. \tag{3.1}$$

Assume that the transition rates $q^N(x,j)$ satisfy the following: for each $x \in S^N$,

$$\begin{aligned} b^N(x) &:= \sum_{j \in J^N} j q^N(x,j), \ \left|b^N(x)\right| < \infty, \\ a^N(x) &:= \sum_{j \in J^N} j j^\intercal q^N(x,j), \ \left\|a^N(x)\right\| < \infty, \\ &\text{and ucc} - \lim a^N(x) \to a(x), \text{ ucc} - \lim b^N(x) \to b, \end{aligned} \tag{3.2}$$

where $a \in C(\mathbb{R}^d, \mathbf{S}^d)$, and $b \in C(\mathbb{R}^d, \mathbb{R}^d)$ is globally Lipschitz with Lipschitz constant l_b, $|\cdot|$ denotes the Euclidean norm and $\|\cdot\|$ denotes the matrix norm induced by $|\cdot|$, and 'ucc' represents 'uniformly on compacts'.

3.1.1 Comparison to ODEs

As in (2.1), we can decompose X_t^N as the sum of the compensator

$$\int_0^t b^N(X_{s-}^N)ds,$$

and the compensated martingale M_t^N with zero mean.

The following proposition, which modifies Proposition 8.8, [12], shows that when the diffusivity of X^N, denoted as a^N in (3.2), is small in a suitable sense, M_t^N can be made arbitrarily small with high probability as $N \to \infty$.

Proposition 3.2. *Consider the sequence of Markov chains $\{X^N\}_{N\in\mathbb{N}}$ as defined in Assumption 3.1. Denote the i-th component of vector j as j_i. For each given N, $t_0 > 0$ and $\bar{a} = \bar{a}(N) > 0$, let*

$$\tau^{t_0}(i,\bar{a}) := \inf\left\{t : \int_0^{t\wedge t_0} \sum_{j\in J^N} q^N(X_{s-}^N, j) j_i^2 ds > \bar{a}(N)\right\}, \quad i = 1, 2, \cdots, d, \tag{3.3}$$

then for any $\delta = \delta(N) < \bar{a}/\max_{j\in J^N}|j_i|$,

$$\mathbb{P}\left[\sup_{s\le t_0\wedge\tau^{t_0}(i,\bar{a})}\left|\left\langle e_i, M_s^N\right\rangle\right| \ge \delta\right] \le 2\exp\left\{-\frac{\delta^2}{4\bar{a}}\right\},\ i = 1, 2, \cdots, d.$$

Proof. For any $\theta \in \mathbb{R}^d$, Define $h(x;\theta) := e^{\langle\theta,x\rangle} - \langle\theta, x\rangle - 1$ and

$$K_t^N(\theta) := \exp\left\{\langle\theta, M_t^N\rangle - \int_0^t \sum_{j\in J^N} q^N(X_{s-}^N, j)h(j;\theta)ds\right\}. \tag{3.4}$$

It is easy to see that $\left(K_t^N(\theta)\right)_{t\ge0}$ is a martingale with mean 1, since

$$\begin{aligned}
K_t^N(\theta) = \exp&\left\{\langle\theta, X_t^N\rangle - \int_0^t \sum_{j\in J^N} q^N\left(X_{s-}^N, j\right)\left(e^{\langle\theta,j\rangle} - 1\right) ds\right\},\\
=&K_0^N(\theta) - \int_0^t K_{s-}^N(\theta)\sum_{j\in J^N}\left(e^{\langle\theta,j\rangle} - 1\right)\upsilon(ds, j)\\
&+\int_0^t \exp\left\{-\int_0^s \sum_{j\in J^N} q^N\left(X_{u-}^N, j\right)\left(e^{\langle\theta,j\rangle} - 1\right) du\right\}\sum_{j\in J^N}\left(e^{\langle\theta,X_{s-}^N+j\rangle} - e^{\langle\theta,X_{s-}^N\rangle}\right)\mu(ds, j)\\
=&1 + \int_0^t K_{s-}^N(\theta)\sum_{j\in J^N}\left(e^{\langle\theta,j\rangle} - 1\right)(\mu - \upsilon)(ds, j),
\end{aligned}$$

where the definition of μ, υ and the martingale property follows from our discussion in Section 2.1, and $K^N(\theta)$ is bounded.

For any $x, \theta \in \mathbb{R}^d$,

$$h(x;\theta) \le e^{|\langle\theta,x\rangle|} - |\langle\theta, x\rangle| - 1 \le \frac{1}{2}|\langle\theta, x\rangle|^2 e^{|\langle\theta,x\rangle|}. \tag{3.5}$$

Conditioned on the event $\{t < \tau^{t_0}(i, \bar{a})\}$ and letting $\theta = ce_i$ for any $c = c(N) > 0$, we have

$$\int_0^{t\wedge t_0} \sum_{j\in J^N} q^N(X_{s-}^N, j)h(j;\pm ce_i)ds \le \frac{1}{2}c^2\exp\left\{c\max_{j\in J^N}|j_i|\right\}\bar{a}, \quad i = 1, 2, \cdots, d. \tag{3.6}$$

Now for $t > 0$, any $A = A(N) > 0$ and $B = B(N) > 0$, let

$$\tau_M(\theta, B) := \inf\{t : \langle \theta, M_t^N \rangle \geq B\}.$$

Since $h(j;\theta)$ is non-negative,

$$\begin{aligned}
&\mathbb{P}\left[\tau_M(\theta, B) \leq t_0 \wedge \tau^{t_0}(i,\bar{a}), \int_0^{\tau_M(\theta,B)\wedge t_0} \sum_{j\in J^N} q^N(X_{s-}^N, j)h(j;\theta)ds < A\right] \\
\leq &\mathbb{P}\left[\sup_{s\leq t_0\wedge\tau^{t_0}(i,\bar{a})} \langle \theta, M_s^N \rangle \geq B, \int_0^{\tau_M(\theta,B)\wedge t_0} \sum_{j\in J^N} q^N(X_{s-}^N, j)h(j;\theta)ds < A\right] \\
\leq &\mathbb{P}\left[K^N_{\tau_M(\theta,B)\wedge t_0}(\theta) \geq e^{B-A}\right] \\
\leq &e^{A-B}\mathbb{E}\left[K^N_{\tau_M(\theta,B)\wedge t_0}(\theta)\right] = e^{A-B},
\end{aligned}$$

where the last equality follows from *Doob's optional sampling theorem.*

Let

$$\theta = \pm\frac{\delta}{2\bar{a}}e_i.$$

It follows from (3.6) that on the event $\{t < \tau^{t_0}(i,\bar{a})\}$,

$$\int_0^{t\wedge t_0} \sum_{j\in J^N} q^N(X_{s-}^N, j)h(j;\pm\frac{\delta}{2\bar{a}}e_i)ds \leq \frac{\delta^2}{4\bar{a}}, \quad i = 1, 2, \cdots, d.$$

We then have

$$\begin{aligned}
&\mathbb{P}\left[\sup_{s\leq t_0\wedge\tau^{t_0}(i,\bar{a})} \left|\langle e_i, M_s^N \rangle\right| \geq \delta\right] \\
\leq &\mathbb{P}\left[\tau_M\left(\frac{\delta}{2\bar{a}}e_i, \frac{\delta^2}{2\bar{a}}\right) \leq t_0 \wedge \tau^{t_0}(i,\bar{a})\right] + \mathbb{P}\left[\tau_M\left(-\frac{\delta}{2\bar{a}}e_i, \frac{\delta^2}{2\bar{a}}\right) \leq t_0 \wedge \tau^{t_0}(i,\bar{a})\right] \\
\leq &2\exp\left\{\frac{\delta^2}{4\bar{a}} - \frac{\delta^2}{2\bar{a}}\right\} = 2\exp\left\{-\frac{\delta^2}{4\bar{a}}\right\},
\end{aligned}$$

and the statement follows. □

Proposition 3.2 is helpful when we approximate continuous-time Markov chains with the solutions of ODEs. In particular, when the ODE is the mean-field differential equation of the continuous-time Markov chain, the approximation is known as the *law of large numbers*, as found in extensive literature, e.g., [16], [12].

The idea is the following:

Recall the drift of X^N and its limit b defined in (3.2). Let $x(t)$ be the unique solution to the ODE

$$dx = b(x)dt, \quad x(0) = x_0. \tag{3.7}$$

The existence and uniqueness of $x(t)$ defined for all $t \in [0, \infty)$ is guaranteed by b being continuous and globally Lipschitz. The discrepancy between X^N and the limit function x can be measured by the largest deviation between the two on a compact time interval, i.e., $\sup_{s\in[0,t]} \left|X_s^N - x(s)\right|$.

To bound the deviation stated above, we need the following Gronwall's inequality.

Theorem 3.3 (Gronwall's inequality, p.498, [16])**.** *Let $\epsilon \geq 0$, and f be a Borel measurable function that is bounded on compact intervals, and satisfies for some $M > 0$,*

$$0 \leq f(t) \leq \epsilon + M \int_0^t f(s)ds,\, t \geq 0,$$

then

$$f(t) \leq \epsilon e^{Mt},\, t \geq 0.$$

From (2.1) and (3.7), we have that the following holds pathwise,

$$\begin{aligned}\left|X_t^N - x(t)\right| \leq & \left|X_0^N - x(0)\right| + \int_0^t \left|b(X_{s-}^N) - b(x(s))\right|ds \\ & + \int_0^t \left|\sum_{j\in J^N} jq^N(X_{s-}^N, j) - b(X_{s-}^N)\right|ds + |M_t^N|,\end{aligned}$$

and after taking supremum over a compact time interval,

$$\sup_{s\in[0,t]} \left|X_s^N - x(s)\right| \leq \left|X_0^N - x(0)\right| + \int_0^t l_b \sup_{u\in[0,s]} \left|X_u^N - x(u)\right| ds$$
$$+ \int_0^t \left|b^N(X_{s-}^N) - b(X_{s-}^N)\right| ds + \sup_{s\in[0,t]} \left|M_s^N\right|.$$

Then we apply Gronwall's inequality pathwise and obtain

$$\sup_{s\in[0,t]} \left|X_s^N - x(s)\right| \leq \left(\left|X_0^N - x(0)\right| + \int_0^t \left|b^N(X_{s-}^N) - b(X_{s-}^N)\right| ds + \sup_{s\in[0,t]} \left|M_s^N\right|\right) e^{l_b t}$$

The first two terms in the parentheses on the RHS above can be made arbitrarily small due to Assumption 3.1 and the third term can be bounded using Proposition 3.2.

In Chapter 5, we apply a modified version of this approximation.

3.1.2 Comparison to diffusions

Definition 3.4 (Martingale problem)**.** Consider a linear operator $\mathcal{A} : \mathrm{Dom}\,(\mathcal{A}) \mapsto C(\mathbb{R}^d)$ defined as

$$\mathcal{A}f(x) := \frac{1}{2} \sum_{1\leq m,n\leq d} a_{mn}(x) \frac{\partial^2 f}{\partial x_m \partial x_n} + \sum_{1\leq m\leq d} b_m(x) \frac{\partial f}{\partial x_m}, \tag{3.8}$$

where the covariance matrix $a : \mathbb{R}^d \to \mathbf{S}^d$ and the drift vector $b : \mathbb{R}^d \to \mathbb{R}^d$ are locally bounded measurable functions. For $D \subset \mathrm{Dom}\,(\mathcal{A})$, a $\mathbb{R}^d$-valued process X with càdlàg paths (resp. the corresponding probability measure on the Skorokhod space) solves *the* $(\mathcal{A}, D)$*-martingale problem with the initial state* x if $X_0 = x$ -a.s. and $f(X_t) - f(X_0) - \int_0^t \mathcal{A}f(X_s)ds$ is a martingale for all $f \in D$.

Definition 3.5 (Well-posed)**.** For $D \subset \mathrm{Dom}\,(\mathcal{A})$, a $(\mathcal{A}, D)$-martingale problem is said to be *well-posed* if for any initial state x, the problem has a unique solution X.

Definition 3.6 (Stopped martingale problem)**.** Let X_t be a càdlàg process with

$X_0 = x \in U$. For an open subset $U \subset \mathbb{R}^d$, define the exit time of X_t from U as

$$\tau_U^x := \inf\{t \geq 0 : X_t \notin U\}.$$

We say that X solves *the stopped* $(\mathcal{A}, D)$*-martingale problem with initial state* x if $X_t = X_{t\wedge\tau_U^x}$ -a.s. and

$$f(X_t) - \int_0^{t\wedge\tau_U^x} \mathcal{A}f(X_s)ds$$

is a martingale for all $f \in D$.

Assuming we have a diffusion limit candidate that is the solution of a well-posed martingale problem, then we can proceed to discuss the weak convergence of the sequence $\{X^N\}_{N\in\mathbb{N}}$.

The following theorem provides a sufficient condition of the existence and uniqueness of a $(\mathcal{A}, C_c^\infty(\mathbb{R}^d))$-martingale problem.

Theorem 3.7 (Theorem 10.2.2, [35]). *Let* $a : \mathbb{R}^d \to \mathbf{S}^d$ *and* $b : \mathbb{R}^d \to \mathbb{R}^d$ *be locally bounded measurable functions in* (3.8)*. If* a *is positive definite, i.e.,*

$$\inf_{|\theta|=1} \theta^\intercal a(x)\theta > 0, \, \forall x \in \mathbb{R}^d,$$

and if there exists $C < \infty$ *such that*

$$\max\{\|a(x)\|, \langle x, b(x)\rangle\} \leq C(1 + |x|^2), \, x \in \mathbb{R}^d,$$

then the $(\mathcal{A}, C_c^\infty(\mathbb{R}^d))$*-martingale problem is well-posed.*

The stochastic differential equation

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = x_0,$$

satisfying $a(x) = \sigma(x)\sigma^\intercal(x)$ has a solution unique in law if and only if the corresponding $(\mathcal{A}, C_c^\infty(\mathbb{R}^d))$-martingale problem is well-posed.

If a is not strictly positive definite, then the martingale problem is called *degenerate*, whose well-posedness needs to be investigated on a case-by-case basis.

In the following, we will present a theorem tailored to the type of continuous-time Markov chains defined in Section 3.1, which is a direct corollary of Theorem 4.1, p.354, [16].

Theorem 3.8. *Consider operator* $\mathcal{A}$ *defined in* (3.8). *Let* $a \in C(\mathbb{R}^d, \mathbf{S}^d)$, $b \in C(\mathbb{R}^d, \mathbb{R}^d)$. *Suppose the* $(\mathcal{A}, C_c^\infty(\mathbb{R}^d))$*-martingale problem is well-posed. For* $N \geq 1$, *let* X^N *and* B_N *be processes with sample paths in* $D([0,\infty), \mathbb{R}^d)$, *and let* A_N *be a symmetric* $d \times d$ *matrix-valued process such that* $(A_N)_{mn}$ *has sample paths in* $D([0,\infty), \mathbb{R})$ *and* $A_N(t) - A_N(s) \in \mathbf{S}^d$ *for* $t > s \geq 0$. *Set* $\mathscr{F}_t^N = \sigma(X_s^N, B_N(s), A_N(s) : s \leq t)$.

Let $\tau_r^N := \inf\{t : |X_t^N| \geq r$ *or* $|X_{t-}^N| \geq r\}$, *and suppose that* $M_N(t) := X_t^N - B_N(t)$, *and* $(M_N)_m (M_N)_n - (A_N)_{mn}$, $1 \leq m, n \leq d$, *are* $\mathscr{F}_t^N$*-local martingales, and that for each* $r > 0$, $T > 0$,

$$
\begin{aligned}
&\lim_{N\to\infty} \mathbb{E}\left[\sup_{t \leq T \wedge \tau_r^N} \left|X_t^N - X_{t-}^N\right|^2\right] = 0, \\
&\lim_{N\to\infty} \mathbb{E}\left[\sup_{t \leq T \wedge \tau_r^N} \left|B_N(t) - B_N(t-)\right|^2\right] = 0, \\
&\lim_{N\to\infty} \mathbb{E}\left[\sup_{t \leq T \wedge \tau_r^N} \left|(A_N(t))_{mn} - (A_N(t-))_{mn}\right|\right] = 0, \\
&\sup_{t \leq T \wedge \tau_r^N} \left|(B_N(t))_m - \int_0^t b_m(X_s^N) ds\right| \xrightarrow{\mathbb{P}} 0, \\
&\sup_{t \leq T \wedge \tau_r^N} \left|(A_N(t))_{mn} - \int_0^t a_{mn}(X_s^N) ds\right| \xrightarrow{\mathbb{P}} 0,
\end{aligned}
\tag{3.9}
$$

and $\lim_{N\to\infty} X_0^N = x_0$.

Then $\{X^N\}_{N\in\mathbb{N}}$ *converges in distribution to the solution of the* $(\mathcal{A}, C_c^\infty(\mathbb{R}^d))$*-martingale problem with initial state* x_0.

The following is a corollary of Theorem 3.8 when applied to the sequence of Markov chains $\{X^N\}_{N\in\mathbb{N}}$ defined in Assumption 3.1.

Theorem 3.9 (Weak convergence to the diffusion limit)**.** *Consider a sequence of continuous-time Markov chains* $\{X^N\}_{N\in\mathbb{N}}$ *defined as in Assumption 3.1. Let operator* $\mathcal{A}$ *be defined as* (3.8) *and let the* $(\mathcal{A}, C_c^\infty(\mathbb{R}^d))$*-martingale problem be well-posed.*

Then $\{X^N\}_{N\in\mathbb{N}}$ weakly converges to the solution of the $(\mathcal{A}, C_c^\infty(\mathbb{R}^d))$-martingale problem with initial state x_0.

Proof. Let

$$B_N(t) = \int_0^t \sum_{j\in J^N} jq^N(X_{s-}^N, j)ds = \int_0^t b^N(X_{s-}^N)ds,$$

and

$$A_N(t) = \int_0^t \sum_{j\in J^N} jj^\intercal q^N(X_{s-}^N, j)ds.$$

Recall the decomposition (2.1) of X^N, we have $M_N(t) := X_t^N - B_N(t) = M_t^N$ is a martingale, and

$$\begin{aligned}
&\mathbb{E}\left[(M_t^N)_m(M_t^N)_n - (M_s^N)_m(M_s^N)_n \Big| \mathscr{F}_s\right] \\
&= \mathbb{E}\left[\int_s^t \sum_{j\in J^N} j_m(\mu-\upsilon)(dr,j) \int_s^t \sum_{k\in J^N} k_n(\mu-\upsilon)(dr,k)\Big|\mathscr{F}_s\right] \\
&= \mathbb{E}\left[\int_s^t \sum_{j,k\in J^N} j_m k_n \, d\Big[(\mu-\upsilon)([s,t],j),(\mu-\upsilon)([s,t],k)\Big]\Big|\mathscr{F}_s\right] \\
&= \mathbb{E}\left[\int_s^t \sum_{j\in J^N} j_m j_n \mu(dr,j)\Big|\mathscr{F}_s\right] = \mathbb{E}\left[\int_s^t \sum_{j\in J^N} j_m j_n \upsilon(dr,j)\Big|\mathscr{F}_s\right] \\
&= \mathbb{E}\left[(A_N(t) - A_N(s))_{mn}\Big|\mathscr{F}_s\right],
\end{aligned}$$

which shows that $M_t^N(M_t^N)^\intercal - A_N(t)$ is also a martingale.

Assumptions (3.9) can be easily verified: the first line follows from (3.1), the second and third lines follow from the definition of A_N, B_N, and the fourth and fifth lines follow from our assumption (3.2). □

3.2 Stochastic dominance and coupling

In this section, we introduce the basic definitions and facts that allow us to compare two Markov processes.

Definition 3.10 (Partially ordered set and increasing set)**.** Let $(S, \preceq)$ be a partially ordered set, that is, the binary relation $\preceq$ satisfies for any $a, b, c \in S$,

1. $a \preceq a$;
2. If $a \preceq b$ and $b \preceq a$ then $a = b$;
3. If $a \preceq b$ and $b \preceq c$ then $a \preceq c$.

A subset $F \subset S$ is called an increasing set if $x \preceq y$, $x \in F$ implies $y \in F$.

Definition 3.11 (Stochastic dominance)**.** Let $(S, \preceq)$ be a partially ordered set. Probability measures $\mathbb{P}$ and $\mathbb{P}'$ are defined on S. Then $\mathbb{P}$ is said to be *stochastically dominated* by $\mathbb{P}'$ if for all increasing sets $F \subset S$, $\mathbb{P}[F] \leq \mathbb{P}'[F]$.

Consider Markov processes X, Y valued in S with transition probabilities

$$p_t^X(x, A) := \mathbb{P}\left[X_{s+t} \in A \middle| X_s = x\right], \quad p_t^Y(y, A) := \mathbb{P}\left[Y_{s+t} \in A \middle| Y_s = y\right].$$

Process X is said to be stochastically dominated by Y if probability measure $p_t^X(x, \cdot)$ is stochastically dominated by $p_t^Y(y, \cdot)$ for all $x \preceq y$, $t \geq 0$.

Definition 3.12 (Order-preserving coupling)**.** Let stochastic processes X, Y each take values in a countable partially ordered set $(S, \preceq)$. An *order-preserving* coupling is a stochastic process (X', Y') valued in $S \times S$, whose marginals are distributed the same as the original processes X, Y, and which satisfies that, for given constant initial states $X_0' \preceq Y_0'$, the following condition holds:

$$\mathbb{P}\left[X_t' \preceq Y_t', \forall t \geq 0\right] = 1.$$

The following theorem establishes the relationship between stochastic dominance and order-preserving coupling in Markov chains.

Theorem 3.13 (Existence of Markov order-preserving coupling, Theorem 1, [36])**.** *Let X, Y be non-explosive continuous-time Markov chains where X is stochastically dominated by Y. Then there exists a non-explosive order-preserving coupling (X', Y') which is a Markov chain.*

We can explicitly construct an order-preseving coupling between simple Markov chains, e.g. birth-death chains.

Example 3.14 (Order-preserving coupling for birth-death chains)**.** Consider two birth-death chains Z^1, Z^2, both defined on a state space $\mathbb{N}$. For $i = 1, 2$, Z^i has transition rates

$$\begin{cases} z \to z+1, \text{ at rate } b_i(z), \\ z \to z-1, \text{ at rate } d_i(z), \end{cases}$$

where $b_1(z) \geq b_2(z)$, $d_1(z) \leq d_2(z)$ for all $z \in \mathbb{N}$, and their initial states satisfy $Z_0^1 \geq Z_0^2$. The order-preserving coupling $(\hat{Z}^1, \hat{Z}^2)$ is defined on $\mathbb{N} \times \mathbb{N}$, satisfying:

At state (z, z), $(\hat{Z}^1, \hat{Z}^2)$ has transition rates

$$\begin{cases} (z, z) \to (z+1, z), \text{ at rate } b_1(z) - b_2(z), \\ (z, z) \to (z+1, z+1), \text{ at rate } b_2(z), \\ (z, z) \to (z-1, z-1), \text{ at rate } d_1(z), \\ (z, z) \to (z, z-1), \text{ at rate } d_2(z) - d_1(z), \end{cases}$$

and at state (z_1, z_2), $z_1 \neq z_2$, $\hat{Z}^1$ and $\hat{Z}^2$ jump independently. Since $\hat{Z}^1$ and $\hat{Z}^2$ will a.s. not jump at the same time, their paths will a.s. not cross each other when $|z_1 - z_2| = 1$.

The intuition as to why this coupling is order-preserving is that, $\hat{Z}^1$, with higher birth rates and lower death rates, will stay above $\hat{Z}^2$ until they meet, in which case, they will jump together until either $\hat{Z}^1$ moves upward, or $\hat{Z}^2$ moves downward.

Lastly, we state the following theorem which is useful for comparing diffusions.

Theorem 3.15 (Comparison theorem, Theorem 3.7, p.394, [37])**.** *For $i = 1, 2$, let $(X_t^i)_{t \geq 0}$ be a diffusion valued in $\mathbb{R}$, with drift coefficient $b_i(t, x)$ and diffusion coefficient $\sigma(t, x)$. Let X^1 and X^2 be defined with respect to the same Brownian motion. If*

- *b_1, b_2 are bounded Borel functions such that $b_1 \geq b_2$ everywhere and at least one of them satisfies a Lipschitz condition,*
- *$(\sigma(t, x) - \sigma(t, y))^2 \leq C|x - y|$ for some positive constant C, and*
- *$X_0^1 \geq X_0^2$ -a.s.,*

then

$$\mathbb{P}\left[X_t^1 \geq X_t^2, \forall t \geq 0\right] = 1.$$

3.3 Bi-continuous semigroups

In this section, we outline the theory of bi-continuous semigroups following [7].

Firstly, we need to introduce a topology coarser than the topology induced by the uniform norm $\|\cdot\|$.

Definition 3.16 (Seminorm)**.** A *seminorm* on Φ is a map $p : \Phi \to \mathbb{R}$ such that for all $f, g \in \Phi$: (1) $p(f) \geq 0$; (2) $p(\alpha f) = |\alpha| p(f)$ for every scalar α; and (3) $p(f + g) \leq p(f) + p(g)$.

Definition 3.17 (Locally convex topology)**.** Let Φ be a vector space and $P := \{p_q\}_{q \in Q}$ be a family of seminorms on Φ. The locally convex topology generated by P is the coarsest topology ρ on Φ s.t. each p_q is continuous.

Assumption 3.18 (Assumptions 1.1, [7])**.** Let $(\Phi, \|\cdot\|)$ be a Banach space with topological dual Φ', and let ρ be a locally convex topology on Φ with the following properties:

1. The space (Φ, ρ) is sequentially complete on $\|\cdot\|$-bounded sets, i.e., every $\|\cdot\|$-bounded ρ-Cauchy sequence converges in (Φ, ρ).

2. The topology ρ is Hausdorff and coarser than the $\|\cdot\|$-topology.

3. The space $(\Phi, \rho)'$ is norming for $(\Phi, \|\cdot\|)$, i.e., for all $x \in \Phi$,

$$\|x\| = \sup\{|f(x)| : f \in (\Phi, \rho)', \|f\|_{(\Phi, \|\cdot\|)'} \leq 1\},$$

   where $\|\cdot\|_{(\Phi, \|\cdot\|)'}$ denotes the operator norm.

On the Banach space $(\Phi, \|\cdot\|)$, with additional topology ρ induced by a family of seminorms $\{p_q\}_{q \in Q}$, we can define the bi-continuous semigroup and related concepts. Most of the definitions in the theory of bi-continuous semigroups mirror the ones in the theory of strong continuous semigroups. Since we aim to present an application of the theory, we will only list below the definitions and theorems that are directly relevant to our proofs.

Definition 3.19 (Bi-equicontinuity)**.** A semigroup $(\mathcal{R}_t)_{t\geq 0}$ is called (globally) *bi-equicontinuous* if

$$\rho - \lim_{n\to\infty} \mathcal{R}_t f_n \to 0$$

holds, for any uniformly bounded sequence $\{f_n\}_{n\in\mathbb{N}}$ ρ-converging to 0 uniformly w.r.t. $t \in [0,\infty)$.

It is called *locally bi-equicontinuous* if the convergence is uniform w.r.t. t in compact intervals.

Remark 3.20. The following condition, known as *locally equicontinuous*, implies local bi-equicontinuity:

For each $q \in Q$, we can find $\tilde{q} \in Q$ independent of t, such that for all $f \in \Phi$,

$$p_q(\mathcal{R}_t f) \leq p_{\tilde{q}}(f).$$

The converse, however, is not true. Consider the family of operators defined in (5.20), we can prove by contradiction that $(\mathcal{S}_t)_{t\geq 0}$ is not locally equicontinuous w.r.t. $(\widehat{BC}(\mathbb{R}^2_+), \text{ucc})$.

Assuming for every compact set $K \subset \mathbb{R}^2_+$, we can find compact set $K_0 \subset \mathbb{R}^2_+$ such that for all $f \in \widehat{BC}(\mathbb{R}^2_+)$,

$$\|\mathcal{S}_t f\|_K \leq \|f\|_{K_0},$$

then it must be true that for all $h \in \widehat{BC}(\mathbb{R}^2_+)$ satisfying $h = 0$ on K_0, and strictly positive elsewhere,

$$\|\mathcal{S}_t h\|_K \leq \|h\|_{K_0} = 0.$$

The LHS of the inequality above is strictly positive, so there is a contradiction.

In order to use the local equicontinuity condition, one may want to work with a finer topology.

Definition 3.21 (Bi-continuous semigroup)**.** A semigroup $(\mathcal{R}_t)_{t\geq 0}$ is said to be *bi-continuous* w.r.t. ρ-topology if

- $\mathcal{R}_0 = \mathcal{I}$, $\mathcal{R}_s\mathcal{R}_t = \mathcal{R}_{s+t}$.
- $\mathcal{R}_t$ is exponentially bounded w.r.t. to the classic operator norm on $(\Phi, \|\cdot\|)$.
- $(\mathcal{R}_t)_{t\geq 0}$ is strongly bi-continuous, i.e.,

$$\rho - \lim_{t\downarrow 0}(\mathcal{R}_t - \mathcal{I})f = 0, \quad f \in \Phi.$$

- $(\mathcal{R}_t)_{t\geq 0}$ is locally bi-equicontinuous.

Definition 3.22 (Bi-dense)**.** A subset of $S \subset \Phi$ is called *bi-dense* if for every $f \in \Phi$ there exists a $\|\cdot\|$-bounded sequence $\{f_n\}_{n\in\mathbb{N}} \subset S$ which ρ-converges to f.

Definition 3.23 (Generator and domain)**.** Let $(\mathcal{R}_t)_{t\geq 0}$ be a bi-continuous semigroup on Φ. The generator $(\mathcal{A}, \mathrm{Dom}\,(\mathcal{A}))$ is defined as

$$\mathcal{A}f := \rho - \lim_{t\downarrow 0}\frac{\mathcal{R}_t f - f}{t},$$

for $f \in \mathrm{Dom}\,(\mathcal{A})$, where $\mathrm{Dom}\,(\mathcal{A})$ is the collection of $f \in \Phi$ such that

$$\sup_{t\in(0,1]}\left\|\frac{\mathcal{R}_t f - f}{t}\right\| < \infty, \quad \rho - \lim_{t\downarrow 0}\frac{\mathcal{R}_t f - f}{t} \in \Phi.$$

Definition 3.24 (Bi-closure)**.** Consider the operator $(\mathcal{A}, \mathrm{Dom}\,(\mathcal{A}))$. For any $\{f_n\}_{n\in\mathbb{N}} \subset \mathrm{Dom}\,(\mathcal{A})$ such that $\{f_n\}_{n\in\mathbb{N}}$ and $\{\mathcal{A}f_n\}_{n\in\mathbb{N}}$ are $\|\cdot\|$-bounded, and have respective limits $f = \rho - \lim_{n\to\infty} f_n$ and $y = \rho - \lim_{n\to\infty}\mathcal{A}f_n$. If $f \in \mathrm{Dom}\,(\mathcal{A})$ and $\mathcal{A}f = y$, then we say $(\mathcal{A}, \mathrm{Dom}\,(\mathcal{A}))$ is *bi-closed*.

Proposition 3.25 (Proposition 1.18 (d), [7])**.** *Let $(\mathcal{A}, Dom\,(\mathcal{A}))$ be the generator of a bi–continuous semigroup $(\mathcal{R}_t)_{t\geq 0}$ on Φ. Then the subspace $\Phi_0 := \overline{Dom\,(\mathcal{A})}^{\|\cdot\|} \subset \Phi$ is $(\mathcal{R}_t)_{t\geq 0}$–invariant and $\mathcal{R}_t|_{\Phi_0}$ is the strongly continuous semigroup on Φ_0 generated by $\mathcal{A}|_{\Phi_0}$ ($\mathcal{A}|_{\Phi_0}$ is known as the part of $\mathcal{A}$ in Φ_0), where*

$$\mathcal{A}|_{\Phi_0} f := \mathcal{A}f \text{ for all } f \in Dom\,(\mathcal{A}|_{\Phi_0})\,,$$

and

$$Dom\,(\mathcal{A}|_{\Phi_0}) := \{f \in Dom\,(\mathcal{A}) \cap \Phi_0 : \mathcal{A}f \in \Phi_0\}.$$

The generalised version of the Chernoff product formula is particularly relevant to our problem.

In the following, we denote the space of linear bounded operators on the space Φ as $\mathscr{L}(\Phi)$.

Theorem 3.26 (Theorem 4.1, [11]). *Let $V : [0, \infty) \longrightarrow \mathscr{L}(\Phi)$ satisfy the following conditions:*

1. *$V(0) = \mathcal{I}$.*
2. *$\|V(t)^m\| \leq Me^{mwt}$ for all $t \geq 0$, $m \in \mathbb{N}$, and for some constants $M \geq 1$ and $w \in \mathbb{R}$.*
3. *The operator family $\{(e^{-wt}V(t))^k : t \geq 0\}$ is locally bi-equicontinuous uniformly for $k \in \mathbb{N}$.*
4. *The family $\left(\frac{V(s)f-f}{s}\right)_{s\in[0,t]}$ is $\|\cdot\|$-bounded for any $t > 0$ and*

$$\mathcal{A}f := \rho - \lim_{s\downarrow 0} \frac{V(s)f - f}{s}$$

exists for all $f \in D \subset \Phi$, where D and $(\alpha - \mathcal{A})D$ are bi-dense subsets in Φ for some $\alpha > w$.

Then the bi-closure of $(\mathcal{A}, D)$ generates a bi-continuous semigroup $(\mathcal{R}_t)_{t\geq 0}$ which is given by the Chernoff Product Formula, i.e.,

$$\mathcal{R}_t f = \rho - \lim_{n\to\infty} \left(V\left(\frac{t}{n}\right)\right)^n f,$$

for all $f \in \Phi$ and uniformly for t in compact intervals of $[0, \infty)$.

Remark 3.27. Theorem 4.1, [11] improves Proposition 2.9, [7] in the sense that the former only requires $\{\frac{V(s)f-f}{s}\}_{s\in[0,t]}$ to be $\|\cdot\|$-bounded and ρ-convergent as $s \downarrow 0$, while the latter requires the stronger $\|\cdot\|$-convergence.

Chapter 4

Small initial infections

In this chapter, we study the cases where the size of the infected population is so small that the randomness becomes the dominating effect. In subcritical regimes with medium or large initial size of infections, the final phases of all epidemics fall into this category, and the behaviours we study below determine the shape of the asymptotic distribution of the extinction time T_o^N.

One way to look at this is that when the size of infection is small, the jump rates of I^N is approximately linear, and thus we can compare I^N to linear birth-death chains, whose properties are well-understood. The other way to look at this is to compare I^N to a branching process, where each infected individual induces its own epidemic. Since the size of infection is small, these epidemics can be viewed as almost independent and the extinction time is the maximum extinction time among all the small epidemics. This explains intuitively why the asymptotic distributions in this chapters appear to be extreme value distributions.

The structure of this chapter is as follows: after a brief introduction in Section 4.1, we prove some preliminary properties of linear birth-death chains in Section 4.2. The main result of this chapter is stated and proved in Section 4.3 and illustrated in diagrams in Section 4.4.

4.1 Introduction

It was noticed from very early on, in the study of the extinction time of stochastic epidemic models, that the trajectory of the size of infected population I^N can be well-approximated by linear birth-death chains when I^N is small. By *linear birth-death chains*, we mean the continuous-time Markov chains $(L_t)_{t\geq 0}$ valued in $\mathbb{N}$, with transition rates

$$\begin{aligned} x &\to x+1, \text{ at rate } \lambda x, \\ x &\to x-1, \text{ at rate } \mu x, \end{aligned}$$

where parameters λ, μ are known as *birth rate* and *death rate* respectively.

The approximation is rigorously justified by coupling I^N between two linear birth-death chains whose extinction times have asymptotically identical distributions. The theory of coupling is introduced in Section 3.2.

The existing works using this technique include [1] on stochastic SIR models, and chronologically [2, 8, 29] on subcritical stochastic SIS models. Among these, the result of [29] is deduced from a remark in [2]. The authors of [2] try to quote the result of [1] but fail to pose the correct conditions for the coupling, i.e., I_0^N needs to be sufficiently small. This error is pointed out in [8]. The authors of [8] present a rigorous discussion of coupling subcritical SIS models with linear birth-death chains.

Finally, Foxall [6] shows that we can use this technique to obtain the asymptotic distribution of extinction time of small initial infections for subcritical, critical and supercritical SIS models.

When we extend this technique to the SIRS models, the situation is considerably more complicated. In this chapter, we are only able to cover the parameter regime when $\lambda_o(N) \leq 1$ and a subset of cases when $\lambda_o(N)$ tends to 1 from above sufficiently quickly.

4.2 Properties of linear birth-death chains

Theorem 4.1 (Asymptotic of linear birth-death chains, Theorem 1, [6]). *Let $\{L^N\}_{N\in\mathbb{N}}$ be a sequence of linear birth-death chains with birth rate $\lambda_N > 0$ and death rate $\mu_N > 0$.*

Let $T^N_{bdp} := \inf\left\{t \geq 0 : L^N_t = 0\right\}$. The distribution of T^N_{bdp} converges to the following limits as $N \to \infty$:
Suppose that $L^N_0 = L_0$ is a constant independent of N:

1. *If $\mu_N - \lambda_N \to 0$, then*

$$\lim_{N\to\infty} \mathbb{P}^N_{L_0}\left[T^N_{bdp} \leq w\right] = \left(1 + \frac{1}{w}\right)^{-L_0}, \quad w > 0;$$

2. *If $\mu_N - \lambda_N \to a > 0$, then*

$$\lim_{N\to\infty} \mathbb{P}^N_{L_0}\left[T^N_{bdp} \leq w\right] = \left(1 + \frac{a}{e^{aw} - 1}\right)^{-L_0}, \quad w > 0.$$

Suppose that $L^N_0 = L_0(N) \to \infty$:

3. *If $L_0(\mu_N - \lambda_N) \to 0$, then*

$$\lim_{N\to\infty} \mathbb{P}^N_{L_0}\left[\frac{T^N_{bdp}}{L_0} \leq w\right] = e^{-\frac{1}{w}}, \quad w > 0;$$

4. *If $L_0(\mu_N - \lambda_N) \to a > 0$, then*

$$\lim_{N\to\infty} \mathbb{P}^N_{L_0}\left[\frac{T^N_{bdp}}{L_0} \leq w\right] = \exp\left\{-\frac{a}{e^{aw} - 1}\right\}, \quad w > 0;$$

5. *If $L_0(\mu_N - \lambda_N) \to \infty$, then*

$$\lim_{N\to\infty} \mathbb{P}^N_{L_0}\left[(\mu_N - \lambda_N)T^N_{bdp} - \log L_0(1 - \lambda_N/\mu_N) \leq w\right] = e^{-e^{-w}}, \quad w \in \mathbb{R}.$$

In particular, if for some $a(N) \sim L_0(\mu_N - \lambda_N)$, we have $L_0(\mu_N - \lambda_N) - a(N) =$

$o\left(\frac{a}{\log a}\right)$, *then we can also write the limit distribution as*

$$\lim_{N\to\infty} \mathbb{P}^N_{L_0}\left[a(N)\frac{T^N_{bdp}}{L_0} - \log\frac{a(N)}{\mu_N} \le w\right] = e^{-e^{-w}}.$$

Proof. We shall drop the subscription and the superscription of the probability measure below, since there is no confusion.

It is possible to obtain the closed form of the distribution of T^N_{bdp}.

Fix arbitrary N, and let $P_n(t) := \mathbb{P}\left[L^N_t = n\right]$. By the Kolmogorov forward equation,

$$\frac{dP_n(t)}{dt} = -(\lambda_N + \mu_N)nP_n(t) + \lambda_N(n-1)P_{n-1}(t) + \mu_N(n+1)P_{n+1}(t). \tag{4.1}$$

Denote the probability generating function $G(z;t) = \sum_{n=0}^{\infty} z^n P_n(t), |z| < 1$, which has the properties

$$\frac{\partial G}{\partial t}(z;t) = \sum_{n=0}^{\infty} z^n \frac{dP_n(t)}{dt},$$

$$\frac{\partial G}{\partial z}(z;t) = \sum_{n=0}^{\infty} nz^{n-1}P_n(t).$$

Multiplying z^n to both sides of (4.1) and adding up from $n = 0$ to infinity, we have

$$\frac{\partial G}{\partial t} - (z-1)(\lambda_N z - \mu_N)\frac{\partial G}{\partial z} = 0.$$

Consider characteristic curves parametrised by r. Let $t = r$, $z = z(r)$, and

$$\frac{dG}{dr} = \frac{\partial G}{\partial t} + \frac{\partial G}{\partial z}\frac{dz}{dr} = 0,$$
$$\frac{dz}{dt} = \frac{dz}{dr} = -(z-1)(\lambda_N z - \mu_N).$$

When $\lambda_N \neq \mu_N$, the solution has the form

$$\frac{z - \mu_N/\lambda_N}{z-1}\exp\{-(\lambda_N - \mu_N)t\} = \text{constant}.$$

Thus we have a solution of the form $G(z;t) = f\left(\frac{z-\mu_N/\lambda_N}{z-1}e^{-(\lambda_N-\mu_N)t}\right)$. Insert the initial condition $G(z;0) = z^{L_0^N}$, we have that f has the form

$$f(x) = \left(\frac{x-\mu_N/\lambda_N}{x-1}\right)^{L_0}.$$

Therefore

$$G(z;t) = \left(\frac{z(\lambda_N e^{-(\lambda_N-\mu_N)t}-\mu_N)-(\mu_N e^{-(\lambda_N-\mu_N)t}-\mu_N)}{z(\lambda_N e^{-(\lambda_N-\mu_N)t}-\lambda_N)-(\mu_N e^{-(\lambda_N-\mu_N)t}-\lambda_N)}\right)^{L_0}. \tag{4.2}$$

We have

$$\begin{aligned}\mathbb{P}\left[T_{bdp}^N \le t\right] = \mathbb{P}\left[L_t^N = 0\right] = G(0;t) &= \left(\frac{\exp\{-(\lambda_N-\mu_N)t\}-1}{\exp\{-(\lambda_N-\mu_N)t\}-\lambda_N\mu_N^{-1}}\right)^{L_0} \\ &= \left(1+\frac{\hat{\lambda}_N}{e^{\hat{\lambda}_N t}-1}\right)^{-L_0},\end{aligned} \tag{4.3}$$

where $\hat{\lambda}_N := \mu_N - \lambda_N$.

Notice that this expression does not require $\mu_N > \lambda_N$.

Case 1 and 2 can be derived by taking $\hat{\lambda}_N \to a \ge 0$ in (4.3).

For Case 3, let $t = wL_0$, since $L_0\hat{\lambda}_N = o(1)$, we have

$$\frac{\hat{\lambda}_N}{\exp\left\{\hat{\lambda}_N t\right\}-1} = \frac{\hat{\lambda}_N}{\exp\left\{w\hat{\lambda}_N L_0\right\}-1} \sim (wL_0)^{-1} \to 0.$$

It follows that

$$\mathbb{P}\left[\frac{T_{bdp}^N}{L_0} \le w\right] = \left(1+(wL_0)^{-1}+o\left((wL_0)^{-1}\right)\right)^{-L_0} \to e^{-\frac{1}{w}}.$$

For Case 4, let $t = wL_0$, then $\hat{\lambda}_N t \sim aw$, and $\frac{\hat{\lambda}_N}{e^{\hat{\lambda}_N t}-1} \sim \frac{\hat{\lambda}_N}{e^{aw}-1} \to 0$.

$$\mathbb{P}\left[\frac{T^N_{bdp}}{L_0} \leq w\right] = \mathbb{P}\left[T^N_{bdp} \leq t\right] = \left(1 + \frac{\hat{\lambda}_N}{\exp\left\{\hat{\lambda}_N t\right\} - 1}\right)^{-L_0}$$
$$= \left(1 + \frac{\hat{\lambda}_N}{e^{aw} - 1} + o\left(\frac{\hat{\lambda}_N}{e^{aw}-1}\right)\right)^{-L_0} \to \exp\left\{-\frac{a}{e^{aw-1}}\right\}.$$

For Case 5, let the leading asymptotic order of $\hat{\lambda}_N L_0$ be $a(N)$, and $b(N) := \hat{\lambda}_N L_0 - a(N) = o(a(N))$.

Let $t = \left(w + \log \hat{\lambda}_N L_0\right) \hat{\lambda}_N^{-1}$, we have

$$\mathbb{P}\left[T^N_{bdp} \leq t\right] = \mathbb{P}\left[\hat{\lambda}_N T^N_{bdp} - \log \hat{\lambda}_N L_0 \leq w\right] = \left(1 + \frac{\hat{\lambda}_N}{\exp\left\{\hat{\lambda}_N t\right\} - 1}\right)^{-L_0}$$
$$= \left(1 + \frac{\hat{\lambda}_N}{\hat{\lambda}_N L_0 e^w - 1}\right)^{-L_0} \to \exp\left\{-e^{-w}\right\}.$$

Now let $t = (w + \log a(N))\, L_0/a(N)$ instead.

$$\mathbb{P}\left[T^N_{bdp} \leq t\right] = \mathbb{P}\left[a\frac{T^N_{bdp}}{L_0} - \log a \leq w\right] = \left(1 + \frac{\hat{\lambda}_N}{\exp\left\{\hat{\lambda}_N t\right\} - 1}\right)^{-L_0}$$
$$= \left(1 + \frac{\hat{\lambda}_N}{(ae^w)^{1+b/a} - 1}\right)^{-L_0} = \exp\left\{-\frac{a+b}{e^{(1+b/a)w}a^{1+b/a} - 1}\right\}.$$

When $a^{b/a} \to 1$, the last expression tends to $\exp\{-e^{-w}\}$. And $a^{b/a} \to 1$ if and only if $\frac{b}{a}\log a \to 0$, hence the second part of Case 5 is proved. □

The following three lemmas estimate the probability of a birth-death chain hitting some given larger state starting from a given state. Although the approach is routine, since we are interested in the case when all parameters are of various scaling of N, we will state the full proof.

Lemma 4.2 (Hitting probability of linear birth-death chains)**.** *Let L be a linear birth-death chain with birth rate $\lambda(N) > 0$ and death rate 1, and $L_0 = l(N) \in \mathbb{N}$.*

As $N \to \infty$, the probability that L_t ever reaches $k(N) > l(N)$ tends to 0, if either

1. *$(1-\lambda)(k-l) \to \infty$, $\lambda(N) < 1$ for all $N \in \mathbb{N}$ and $\lim_{N\to\infty} \lambda(N) \leq 1$, or*

2. *$l(N) = o(k(N))$, $(1-\lambda)k \to 0$ and $\lambda \to 1$.*

Proof. Let h_i be the probability of L ever hitting k from $L_0 = i$, $i \in \mathbb{N}$. Then $\{h_i\}_{i\leq k}$ is the minimal non-negative solution (See Theorem 3.3.1, p.112, [13]) of

$$\begin{aligned} 0 &= \lambda(h_{i+1} - h_i) + (h_{i-1} - h_i), \quad 1 < i < k, \\ h_k &= 1, \quad h_0 = 0. \end{aligned}$$

It has a solution $\left\{\frac{\lambda^{-i}-1}{\lambda^{-k}-1}\right\}_{i\leq k}$. Since $\{h_i\}_{i\in\mathbb{N}}$ is the minimal solution,

$$\frac{\lambda^{-i}-1}{\lambda^{-k}-1} \geq h_i, \quad i \leq k.$$

If $(1-\lambda)(k-l) \to \infty$, $\lambda(N) < 1$ for all $N \in \mathbb{N}$, and $\lim_{N\to\infty} \lambda(N) \leq 1$, we have

$$h_l \leq \lambda^{k-l} = \left((1-(1-\lambda))^{\frac{1}{1-\lambda}}\right)^{(1-\lambda)(k-l)} \to 0.$$

If $(1-\lambda)k \to 0$ and $\lambda \to 1$, we have

$$\lambda^{-i} = \left((1-(1-\lambda))^{\frac{1}{1-\lambda}}\right)^{(1-\lambda)i} \to 1, \quad i \leq k.$$

Notice that this is true even if $\lambda(N) \geq 1$ for some $N \in \mathbb{N}$.

It follows that

$$\lim_{N\to\infty} h_l = \lim_{N\to\infty} \frac{\sum_{i=0}^{l-1} \lambda^{-i}}{\sum_{i=0}^{k-1} \lambda^{-i}} = \lim_{N\to\infty} \frac{l-1}{k-1} = 0,$$

when $l = o(k)$. □

Lemma 4.3 (Hitting probability of immigration-death chains absorbing at 0)**.** *For any given $N \in \mathbb{N}$, let L be an immigration-death chain with immigration rate $\alpha(N) >$*

0 and death rate $\mu(N) > 0$, absorbing at 0. That is, L has the transition rates for $x \geq 1$:

$$
\begin{aligned}
x &\to x+1, \text{ at rate } \alpha, \\
x &\to x-1, \text{ at rate } \mu x;
\end{aligned}
$$

and L remains at 0 once it hits 0.

Let $L_0 = l(N) \to \infty$. Then the probability that L_t ever reaches $2l(N)$ tends to 0 as $N \to \infty$, if

$$\lim_{N\to\infty} \frac{l(N)\mu(N)}{\alpha(N)} > e.$$

Proof. Let h_i be the probability of L ever hitting k from $L_0 = i$, $i \leq 2l$. Then $\{h_i\}_{i\leq 2l}$ is the minimal non-negative solution of

$$
\begin{aligned}
0 &= \alpha(h_{i+1} - h_i) + \mu i(h_{i-1} - h_i), \quad 1 \leq i < 2l, \\
h_{2l} &= 1.
\end{aligned}
$$

It has a general solution $\{x_i\}_{i\leq 2l}$, where

$$
\begin{aligned}
x_i &= \frac{\sum_{k=0}^{i-1}(\mu/\alpha)^k k!}{\sum_{k=0}^{2l-1}(\mu/\alpha)^k k!} + h_0\left(1 - \frac{\sum_{k=0}^{i-1}(\mu/\alpha)^k k!}{\sum_{k=0}^{2l-1}(\mu/\alpha)^k k!}\right), \quad 1 \leq i \leq 2l, \\
x_0 &= h_0,
\end{aligned}
$$

for any $h_0 \in [0,1]$. The minimal non-negative solution is reached when $h_0 = 0$.

$$\frac{\sum_{k=0}^{l-1}(\mu/\alpha)^k k!}{\sum_{k=0}^{2l-1}(\mu/\alpha)^k k!} = h_l$$

When $\lim_{N\to\infty} \frac{l\mu}{e\alpha} > 1$, we have

$$\sum_{k=l}^{2l-1}(\mu/\alpha)^k k! \geq (\mu/\alpha)^l l! \sum_{k=0}^{l-1}(\mu/\alpha)^k k!,$$

and for sufficiently large N,

$$h_l \le \frac{\sum_{k=0}^{l-1}(\mu/\alpha)^k k!}{\sum_{k=0}^{l-1}(\mu/\alpha)^k k! + (\mu/\alpha)^l l! \sum_{k=0}^{l-1}(\mu/\alpha)^k k!} = \frac{1}{1+(\mu/\alpha)^l l!} \le \frac{1}{1+l^{1/2}\left(\frac{l\mu}{e\alpha}\right)^l \sqrt{2\pi}} \to 0.$$

□

Lemma 4.4 (Hitting probability of immigration-death chains)**.** *For given $N \in \mathbb{N}$, let L be an immigration-death chain with immigration rate $\alpha(N) > 0$ and death rate $\mu(N) > 0$. That is, L has the transition rates for $x \ge 0$:*

$$x \to x+1, \text{ at rate } \alpha,$$
$$x \to x-1, \text{ at rate } \mu x.$$

If $L_0 = l(N) \to \infty$, $\mu = O(1)$ and $\alpha = o(l\mu)$, then for $t_0 = t_0(N) \to \infty$ satisfying $t_0 = o\left(\left(\frac{l\mu}{\alpha e}\right)^l\right)$, the probability of the event 'L_t reaches $2l$ before $t = t_0$' tends to 0 as $N \to \infty$.

Proof. Notice that $L_0 = l(N) \to \infty$, $\mu = O(1)$ and $\alpha = o(l\mu)$ imply

$$\lim_{N\to\infty} \frac{l(N)\mu(N)}{\alpha(N)} > e.$$

Under this condition, in Lemma 4.3, we have estimated the probability for L_t starting from l to ever reach $2l$ before reaching 0, denoted as h_l, and have

$$h_l = o\left(\left(\frac{l\mu}{\alpha e}\right)^{-l}\right).$$

To prove the statement in Lemma 4.4, we argue that with probability tending to 1, L_t can reach l from 0 at most $\lceil t_0 \rceil$ times within time interval $[0, t_0]$. If this is indeed the case, then

$$\mathbb{P}\left[\sup_{t\in[0,t_0]} L_t \ge 2l\right] \le (\lceil t_0 \rceil + 1)h_l \to 0.$$

Since L_t is stochastically dominated by Poisson process C_t with rate α, by order-preserving coupling, we have

$$\mathbb{P}\left[L_t \text{ travels from 0 to } l \text{ at least} \lceil t_0 \rceil \text{ times within } [0,t_0]\right] \leq \mathbb{P}\left[C_{t_0} \geq l\lceil t_0 \rceil\right],$$

where $C_{t_0} \sim \text{Poisson}(\alpha t_0)$. Since $l\lceil t_0 \rceil > \alpha t_0$ for all sufficiently large N, we have the following bound of the tail probability of Poisson distributions (Theorem 5.4, p.97, [38])

$$\mathbb{P}\left[C_{t_0} \geq l\lceil t_0 \rceil\right] \leq \left(\frac{e\alpha t_0}{l\lceil t_0 \rceil}\right)^{l\lceil t_0 \rceil} e^{-\alpha t_0} \to 0.$$

□

There will be several times when we need to bound the value of $\int_0^t I_s^N ds$ for $t > 0$. In the following lemma, we provide an upper bound for this quantity.

Lemma 4.5. *Let $L(l) = (L_t)_{t\geq 0}$ be a linear birth-death chain with birth rate $\lambda = \lambda(N) > 0$, death rate $\mu = \mu(N) > \lambda(N)$, and $L_0 = l(N)$ for $N \in \mathbb{N}$.*

Let

$$T_L := \inf\left\{t : L_t = 0\right\},$$

and

$$H(l) := \int_0^{T_L} L_s ds.$$

Then for each $N \in \mathbb{N}$, $L_0 = l(N)$ and $\delta = \delta(N) > 0$, we have

$$\mathbb{P}\left[H(l) > \delta\right] \leq \frac{l}{(\mu - \lambda)\delta}. \tag{4.4}$$

Proof. We fix N throughout the proof.

Denote the Laplace transform of H with $L_0 = l(N)$ as

$$H^*(a;l) := \mathbb{E}\left[e^{-aH(l)}\right], \quad a \geq 0.$$

Let S denote the sojourn time of $L(l)$ before its first jump. The explicit expression of $H^*(a;l)$ can be obtained following a first-step analysis (e.g. p.482, [39]):

$$\begin{aligned}\mathbb{E}\left[e^{-aH(l)}\right] &= \frac{\lambda}{\lambda+\mu}\mathbb{E}\left[e^{-a(H(l+1)+Sl)}\right] + \frac{\mu}{\lambda+\mu}\mathbb{E}\left[e^{-a(H(l-1)+Sl)}\right],\\ H^*(a;l) &= \frac{\lambda}{\lambda+\mu}H^*(a;l+1)\int_0^\infty e^{-las}(\lambda+\mu)le^{-(\lambda+\mu)ls}ds\\ &\quad + \frac{\mu}{\lambda+\mu}H^*(a;l-1)\int_0^\infty e^{-las}(\lambda+\mu)le^{-(\lambda+\mu)ls}ds,\end{aligned}$$

Then

$$H^*(a;l) = \left(\lambda H^*(a;l+1) + \mu H^*(a;l-1)\right)(\lambda+\mu+a)^{-1}.$$

The solution of the above is

$$H^*(a;l) = \left(\frac{\lambda+\mu+a-\sqrt{(\lambda+\mu+a)^2-4\lambda\mu}}{2\lambda}\right)^l, \quad l \geq 1.$$

$$\mathbb{E}\left[H(l)\right] = -\frac{dH^*(a;l)}{da}\Big|_{a=0} = \frac{l}{\mu-\lambda}.$$

By the Markov inequality, we have for each $N \in \mathbb{N}$ and any $\delta = \delta(N) > 0$,

$$\mathbb{P}\left[H(l) > \delta\right] \leq \frac{\mathbb{E}\left[H(l)\right]}{\delta} = \frac{l}{(\mu-\lambda)\delta}.$$

□

4.3 Main result

In this section, we state and prove our main result regarding the asymptotic distribution of the extinction time of SIRS epidemics with small initial size of infection.

Theorem 4.6 (Small initial infections)**.** *Consider a sequence of stochastic SIRS models defined in* (2.5)*, indexed by* $N \in \mathbb{N}$*, with parameters* $\lambda_o = \lambda_o(N) > 0$ *and* $\gamma_o = \gamma_o(N) > 0$*, and initial states* $(I_0^N, R_0^N) = (I_0(N), R_0(N))$*.*

Let $T_o^N := \inf\{t : I_t^N = 0\}$*. If one of the following conditions is satisfied, then we have the explicit expression of the asymptotic distribution of* T_o^N*:*

Cases 1.1-1.3 are cases where both the initial size of infection I_0 and immunity R_0 are small, whereas Cases 2.1 and 2.2 are cases where I_0 is small and R_0 is of order N.

- ***Case 1.1:*** $I_0|1-\lambda_o| \to 0$, $I_0 R_0 = o(N)$, $I_0 = o(N^{1/2}\gamma_o^{1/2})$.

  If $I_0 = O(1)$,
$$\lim_{N\to\infty} \mathbb{P}^N \left[T_o^N \leq w\right] = \left(1+\frac{1}{w}\right)^{-I_0};$$
  and if $I_0 \to \infty$,
$$\lim_{N\to\infty} \mathbb{P}^N \left[\frac{T_o^N}{I_0} \leq w\right] = e^{-\frac{1}{w}}.$$

- ***Case 1.2:*** $I_0(1-\lambda_o) \to a > 0$, $\lambda_o = \lambda_o(N) < 1$, *and* $I_0 = o\left(N^{1/2}\gamma_o^{1/2}\right)$, $I_0 R_0 = o\left(N\right)$.

  If $I_0 = O(1)$,
$$\lim_{N\to\infty} \mathbb{P}^N \left[T_o^N \leq w\right] = \left(1+\frac{a}{e^{aw}-1}\right)^{-I_0};$$
  and if $I_0 \to \infty$,
$$\lim_{N\to\infty} \mathbb{P}^N \left[\frac{T_o^N}{I_0} \leq w\right] = \exp\left\{-\frac{a}{e^{aw}-1}\right\}.$$

- ***Case 1.3:*** $I_0(1-\lambda_o) \to \infty$, $\lambda_o = \lambda_o(N) < 1$, $I_0 = o\left(\frac{N(1-\lambda_o)\gamma_o}{\log I_0(1-\lambda_o)}\right)$, *and* $R_0 \log I_0(1-\lambda_o) = o\left(N(1-\lambda_o)\right)$. *Then*
$$\lim_{N\to\infty} \mathbb{P}^N \left[(1-\lambda_o)T_o^N - \log(1-\lambda_o)I_0 \leq w\right] = e^{-e^{-w}}.$$

- ***Case 2.1:*** $I_0 = O(1)$, $R_0 = r_0 N$, $r_0 > 0$, $\lambda_o = \lambda_o(N) \leq 1$ *and* $\gamma_o = o(1)$. *Let* $a := \lim_{N\to\infty} 1 - \lambda_o + \lambda_o r_0$, *then*
$$\lim_{N\to\infty} \mathbb{P}^N \left[T_o^N \leq w\right] = \left(1+\frac{a}{e^{aw}-1}\right)^{-I_0}.$$

- ***Case 2.2:*** $I_0 \to \infty$, $R_0 = r_0 N$, $r_0 > 0$, *and there exists* $\epsilon_1, \epsilon_2 > 0$ *such that*

$I_0 = o\left(N^{1-\epsilon_1}\right)$ and $\gamma_o = o\left(N^{-\epsilon_2}\right)$. Let $a := \lim_{N\to\infty} 1 - \lambda_o + \lambda_o r_0$, then

$$\lim_{N\to\infty} \mathbb{P}^N \left[aT_o^N - \log(aI_0) \le w\right] = e^{-e^{-w}}.$$

The cases above cover all of the parameter regime $\{(\lambda_o, \gamma_o) : \lambda_o \le 1\}$, and Case 1.1 also covers a subset of the parameter regime $\{(\lambda_o, \gamma_o) : \lambda_o \ge 1\}$. At the end of this section, we will use diagrams to illustrate the different combinations of parameters and initial states covered in the theorem above.

The rest of this section is dedicated to proving Theorem 4.6.

The process I^N has the following transition rates at time t when $I_t^N = x$:

$$\begin{aligned} &x \to x+1, \text{ at rate } \lambda_o \left(1 - N^{-1}(x + R_t^N)\right) x, \\ &x \to x-1, \text{ at rate } x. \end{aligned}$$

The general idea of the proof is that we will sandwich I^N between two linear birth-death chains whose extinction times have the same asymptotic distributions, according to Theorem 4.1. The construction of such coupling follows from Example 3.14. To make sure the birth rates and death rates are in the correct order, we will need to find upper-bounds held with high probability for I^N and R^N.

The intuition behind discussing two broad scenarios depending on the order of R_0 is as follows:

If $R_0(N)/N \to 0$, then I^N, with small initial value and additional assumptions, will have a birth rate close to λ_o. Looking at Theorem 4.1, it makes sense to discuss three different cases within this scenario based on the limit of $I_0(1 - \lambda_o)$.

If $R_0(N)/N \to r_0 \in (0, 1]$, then I^N, with small initial value and additional assumptions, will have a birth rate close to $\lambda_o(1 - r_0)$. Depending on whether $I_0 = O(1)$, we can divide this scenario into two cases corresponding to the last two cases in Theorem 4.1.

The proof of Case 1.1 to 1.3 follows the same idea: we choose an appropriate

$k(N)$ and $m(N)$ such that $R_0 \le m(N)$ for sufficiently large N, and as $N \to \infty$,

$$\mathbb{P}\left[R_t^N \le 2m(N), I_t^N \le k(N), \forall t \ge 0\right] \to 1.$$

Define two linear birth-death chains $\underline{L}$ and $\overline{L}$, such that $\overline{L}$ has birth rate λ_o and death rate 1, and $\underline{L}$ has birth rate $\lambda_o(1 - \frac{k(N)+2m(N)}{N})$ and death rate 1. Let $\underline{L}_0 = \overline{L}_0 = I_0$. Then we only need to check that the extinction times $T_{\underline{L}}$ and $T_{\overline{L}}$ have the same asymptotic distributions.

We will state the proof of Case 1.1 in full detail, and omit the repeated content in Case 1.2 and 1.3.

Case 1.1: $I_0|1-\lambda_o| \to 0$, $I_0R_0 = o(N)$, $I_0 = o(N^{1/2}\gamma_o^{1/2})$.
Notice in this case it is necessary that $|1-\lambda_o| \to 0$.
Since $I_0 = o(N^{1/2}\gamma_o^{1/2})$, we can find $\kappa(N) \to \infty$ such that

$$\kappa(N) \ll \left(N^{1/2}\gamma_o^{1/2} \wedge |1-\lambda_o|^{-1}\right) I_0^{-1}.$$

Let $k(N) := I_0\kappa(N)$. Define linear birth-death chain $\overline{L}$ with birth rate λ_o and death rate 1, and linear birth-death chain $\underline{L}$ with birth rate $\lambda_o(1 - \frac{k(N)+2m(N)}{N})$ and death rate 1. Let $\overline{L}_0 = \underline{L}_0 = I_0$. From Example 3.14, there is an order-preserving coupling between I^N and $\overline{L}$ such that $I_t^N \le \overline{L}_t$, for all $t \ge 0$. Since $I_0 \ll k$, $(1-\lambda_o)k(N) \to 0$ and $\lambda_o \to 1$, we can apply the second case in Lemma 4.2 to $\overline{L}$, and obtain that with probability tending to 1, $I_t^N \le \overline{L}_t \le k(N)$.

For $N \in \mathbb{N}$, conditioned on $\{I_t^N \le k(N)\}$, each R^N is stochastically dominated by an immigration-death chain $M = (M_t)_{t\ge 0}$ with immigration rate $k(N)$ and death rate γ_o and $M_0 \ge R_0$.

Let $M_0 = m(N) := N^{1/2}\gamma_o^{-1/2} \vee R_0$, and $t_0 = M_0\gamma_o$. It is obvious that $M_0 \to \infty$ and $k = o(M_0\gamma_o)$.
Since $\frac{M_0\gamma_o}{ke} \to \infty$ and $M_0\gamma_o = O(M_0)$, we have

$$t_0 = o\left(\left(\frac{M_0\gamma_o}{ke}\right)^{M_0}\right).$$

Thus all the conditions of Lemma 4.4 are met, and we have

$$\mathbb{P}\left[R_t^N \geq 2m(N),\, \forall t \leq t_0 \,\middle|\, I_t^N \leq k(N)\right] \leq \mathbb{P}\left[M_t \geq 2M_0,\, \forall t \leq t_0\right] \to 0.$$

It follows that with probability tending to 1,

$$\lambda_o \geq \lambda_o\left(1 - \frac{I_t^N + R_t^N}{N}\right) \geq \lambda_o\left(1 - \frac{k(N) + 2m(N)}{N}\right).$$

Denote $T_{\underline{L}} := \inf\{t : \underline{L}_t = 0\}$ and $T_{\overline{L}} := \inf\{t : \overline{L}_t = 0\}$. From Case 1 and 3 of Theorem 4.1, we have $T_{\overline{L}}$ is of order $I_0 = o(t_0)$. It follows that as $N \to \infty$,

$$\mathbb{P}\left[T_{\overline{L}} < t_0\right] \to 1.$$

For each $N \in \mathbb{N}$, conditioned on

$$\left\{I_t^N \leq k(N),\, R_t^N \leq 2m(N),\, \forall t \leq t_0\right\},$$

there is an order-preserving coupling between $\underline{L}$ and I^N and between I^N and $\overline{L}$ such that $\underline{L}_t \leq I_t^N \leq \overline{L}_t$ for all $t \geq 0$. For sufficiently large N, we have

$$\mathbb{P}\left[T_{\underline{L}} \leq T_o^N \leq T_{\overline{L}} < t_0\right] \geq \mathbb{P}\left[I_t^N \leq k(N), R_t^N \leq 2m(N),\, \forall t \leq t_0\right].$$

Notice that $I_0 M_0 \leq I_0 N^{1/2}\gamma_o^{-1/2} + I_0 R_0 = o(N)$. Since

$$\begin{aligned}
\lim_{N\to\infty}\left(1 - \lambda_o\left(1 - \frac{k(N) + 2m(N)}{N}\right)\right)\underline{L}_0 &= \lim_{N\to\infty}(1-\lambda_o)I_0 + \lim_{N\to\infty}\lambda_o\frac{I_0(k(N)+2m(N))}{N} \\
&= \lim_{N\to\infty}(1-\lambda_o)\overline{L}_0 = 0,
\end{aligned}$$

the asymptotic distribution of T_o^N follows from Case 1 in Theorem 4.1 if $I_0 = O(1)$, and Case 3 if $I_0 \to \infty$.

Case 1.2: $I_0(1-\lambda_o) \to a > 0$ and $I_0 = o\left(N^{1/2}\gamma_o^{1/2}\right)$, $I_0 R_0 = o(N)$.

Notice that this is only possible if $(1-\lambda_o)N^{1/2}\gamma_o^{1/2} \to \infty$. This case covers the scenarios where λ_o is independent of N and $I_0 = O(1)$.

Let $m(N) := N^{1/2}\gamma_o^{-1/2} \vee R_0 \to \infty$.

Since $I_0 = o\left(m\gamma_o\right)$, by letting $k(N) := \sqrt{I_0 m\gamma_o} \to \infty$, we have

$$(1-\lambda_o)\sqrt{I_0 m\gamma_o} \gg (1-\lambda_o)I_0 \asymp 1.$$

By the first case in Lemma 4.2, with probability tending to 1, $I_t^N \leq \sqrt{I_0 m\gamma_o}$ for all $t \geq 0$.

Again, let $M = (M_t)_{t\geq 0}$ be the immigration-death chain dominating R^N. Let $M_0 = m(N)$, and we have $k = o(M_0\gamma_o)$. For

$$t_0 = m\gamma_o = o\left(\left(\frac{M_0\gamma_o}{ke}\right)^{M_0}\right),$$

by Lemma 4.4 and the argument similar to the previous case, we have

$$\mathbb{P}\left[R_t^N \geq 2m(N),\, \forall t \leq t_0 \,\middle|\, I_t^N \leq \sqrt{I_0 m\gamma_o},\, \forall t \geq 0\right] \to 0.$$

The extinction times $T_{\underline{L}}$ and $T_{\overline{L}}$ have the same asymptotic distribution as specified in Theorem 4.1 (Case 2 when $I_0 = O(1)$ and Case 4 when $I_0 \to \infty$). As in the previous case, as $N \to \infty$,

$$\mathbb{P}\left[T_{\overline{L}} < t_0\right] \to 1.$$

Since $I_0\sqrt{I_0 m\gamma_o} = o(I_0 m\gamma_o) = O(I_0 M_0)$, and $I_0 M_0 \leq I_0 N^{1/2}\gamma_o^{-1/2} + I_0 R_0 = o(N)$, we have

$$\begin{aligned}\lim_{N\to\infty}\left(1-\lambda_o\left(1-\frac{\sqrt{I_0 m\gamma_o}+2m}{N}\right)\right)\underline{L}_0 &= \lim_{N\to\infty}(1-\lambda_o)I_0 + \lim_{N\to\infty}\lambda_o\frac{I_0(\sqrt{I_0 m\gamma_o}+2m)}{N}\\ &= \lim_{N\to\infty}(1-\lambda_o)\overline{L}_0 = a.\end{aligned}$$

Case 1.3: $I_0(1-\lambda_o) \to \infty$, $I_0 = o\left(\frac{N(1-\lambda_o)\gamma_o}{\log I_0(1-\lambda_o)}\right)$, and $R_0 = o\left(\frac{N(1-\lambda_o)}{\log I_0(1-\lambda_o)}\right)$.

This case is possible only if $(1-\lambda_o)N^{1/2}\gamma_o^{1/2} \to \infty$. It covers the scenarios where λ_o is independent of N, and $I_0 \to \infty$.

Let $k(N) := 2I_0$. Since $(1-\lambda_o)I_0 \to \infty$ and $\lambda_o(N) < 1$, by the first case in Lemma 4.2, with probability tending to 1, $I_t^N \leq 2I_0$.

Since $I_0 = o\left(\frac{N(1-\lambda_o)\gamma_o}{\log I_0(1-\lambda_o)}\right)$, and $R_0 = o\left(\frac{N(1-\lambda_o)}{\log I_0(1-\lambda_o)}\right)$, we can find $\widetilde{m}(N)$ such that

$$I_0 \ll \widetilde{m}\gamma_o \ll \frac{N(1-\lambda_o)\gamma_o}{\log I_0(1-\lambda_o)}.$$

Let

$$m(N) = \frac{N(1-\lambda_o)}{\log^2 N(1-\lambda_o)} \vee R_0 \vee \widetilde{m}.$$

We have the properties: $I_0 = o(m\gamma_o)$ and

$$m = o\left(\frac{N(1-\lambda_o)}{\log I_0(1-\lambda_o)}\right).$$

At the beginning of this proof, we state that $N^{1/2}(1-\lambda_o)\gamma_o^{1/2} \to \infty$, from which we also have $(1-\lambda_o)^{-1} \ll N(1-\lambda_o)$.

Define linear birth-death chains $\overline{L}$ and $\underline{L}$ the same way as in Case 1.2.

The extinction time of $T_{\overline{L}}$, according to Case 5, Theorem 4.1, is of order $(1-\lambda_o)^{-1}\log I_0(1-\lambda_o)$. Notice that

$$(1-\lambda_o)^{-1}\log I_0(1-\lambda_o) \ll N^2(1-\lambda_o)^2.$$

Let $t_0 = N^2(1-\lambda_o)^2$, then similarly we have

$$\mathbb{P}\left[T_{\overline{L}} < t_0\right] \to 1.$$

Since

$$\log t_0 = 2\log N(1-\lambda_o) \ll N^{1/2}(1-\lambda_o)^{1/2}\log\frac{m\gamma_o}{2eI_0} \ll m\log\frac{m\gamma_o}{2eI_0},$$

it follows from Lemma 4.4 that,

$$\mathbb{P}\left[R_t^N \geq 2m(N),\, \forall t \leq t_0 \,|\, I_t^N \leq 2I_0,\, \forall t \geq 0\right] \to 0.$$

Since $I_0 = o\left(m(N)\right)$, we have

$$\lim_{N\to\infty}\left(1-\lambda_o\left(1-\frac{2(I_0+m)}{N}\right)\right)\underline{L}_0 - \lim_{N\to\infty}(1-\lambda_o)I_0 = \lim_{N\to\infty}\lambda_o\frac{2I_0(I_0+m)}{N}$$
$$= o\left(\frac{I_0(1-\lambda_o)}{\log I_0(1-\lambda_o)}\right),$$

and the rest follows from the order-preserve coupling.

For Case 2.1 and 2.2, we require that γ_o is sufficiently small, so that R^N does not move far away from $R_0(N) \sim r_0 N$, $r_0 \in (0,1)$ before extinction. According to Theorem 4.1, when $I_0 = O(1)$, we expect the extinction time to be of order $O(1)$; whereas when $I_0 \to \infty$, we expect the extinction time to be of order $\log I_0 + O(1)$.

Firstly, we estimate the probability that R^N will remain close to $r_0 N$ for duration of order $\log N$.

Lemma 4.7. *Let*

$$X_t^{N,1} := I_t^N/N, \quad X_t^{N,2} := R_t^N/N,$$

with initial states $X_0^{N,1} = I_0(N)/N$ *and* $X_0^{N,2} \to r_0 > 0$. *Let* $\delta = \delta(N) > 0$. *For sufficiently large* N, *if* $t_1 = t_1(N)$ *satisfies* $0 < t_1 < \delta\gamma_o^{-1}$, *then we have*

$$\mathbb{P}\left[\sup_{t\le t_1}\left|X_t^{N,2} - X_0^{N,2}\right| > 4\delta\right] \le 2\exp\left\{-\frac{\delta^2 N}{4(\gamma_o+1)t_1}\right\} + \frac{I_0}{(1-\lambda_o+\lambda_o r_0/2)\delta N}. \tag{4.5}$$

Proof. We consider $N \in \mathbb{N}$ to be sufficiently large and fixed throughout the proof.

Let

$$T_R(\epsilon) := \inf\left\{t \ge 0 : \sup_{s\le t}\left|X_s^{N,2} - X_0^{N,2}\right| > \epsilon\right\}.$$

The process X^N has transition rates:

$$q^N((x_1,x_2),j) = \begin{cases} N\gamma_o x_2, \quad j = (0, -\frac{1}{N}), \\ N x_1, \quad j = (-\frac{1}{N}, \frac{1}{N}), \\ N\lambda_o(1-x_1-x_2)x_1, \quad j = (\frac{1}{N}, 0). \end{cases} \tag{4.6}$$

It is also easy to see that the state space of X^N is a subset of $[0,1]^2$.

By the argument introduced in Section 3.1.1, we can write

$$\begin{aligned} X_t^{N,2} &= X_0^{N,2} + \int_0^t \sum_j j_2 q^N \left((X_s^{N,1}, X_s^{N,2}), j\right) ds + M_t^N \\ &= X_0^{N,2} + \int_0^t \left(-\gamma_o X_s^{N,2} + X_s^{N,1}\right) ds + M_t^N, \end{aligned} \tag{4.7}$$

where M^N is a zero-mean martingale. We also have for any $x_1, x_2 \in [0,1]$,

$$\sum_{j \in J^N} j_2^2 q^N((x_1,x_2),j) = N^{-1}\gamma_o x_2 + N^{-1} x_1 < (\gamma_o + 1)N^{-1}.$$

For any given N and $\epsilon = \epsilon(N) > 0$, let

$$T_M(\epsilon) := \inf\left\{t : \sup_{s\le t} \left|M_s^N\right| > \epsilon\right\}.$$

By Proposition 3.2, we have for any $t_1 = t_1(N)$,

$$\mathbb{P}\left[T_M(\epsilon) \le t_1\right] \le 2\exp\left\{-\frac{\epsilon^2 N}{4(\gamma_o+1)t_1}\right\}. \tag{4.8}$$

Taking the supremum and applying Gronwall's inequality to (4.7), we have

$$\begin{aligned} &\sup_{s\le t} X_s^{N,2} \le X_0^{N,2} + \int_0^t \gamma_o \sup_{u\le s} X_u^{N,2} ds + \int_0^t X_s^{N,1} ds + \sup_{s\le t}\left|M_s^N\right|, \\ &\sup_{s\le t} X_s^{N,2} \le \left(X_0^{N,2} + \sup_{s\le t}\left|M_s^N\right| + \int_0^t X_s^{N,1} ds\right) e^{\gamma_o t}, \\ &\sup_{s\le t}\left(X_s^{N,2} - X_0^{N,2}\right) \le X_0^{N,2}(e^{\gamma_o t} - 1) + \left(\sup_{s\le t}\left|M_s^N\right| + \int_0^t X_s^{N,1} ds\right) e^{\gamma_o t}. \end{aligned} \tag{4.9}$$

On the other hand, from (4.7) we have, for all $t \ge 0$, $X_t^{N,2} \ge X_0^{N,2} - \gamma_o t - \sup_{s\le t}\left|M_s^N\right|$. It follows that for all $t > 0$,

$$\inf_{s\le t}\left(X_s^{N,2} - X_0^{N,2}\right) \ge -\gamma_o t - \sup_{s\le t}\left|M_s^N\right|. \tag{4.10}$$

Combining (4.9) and (4.10), we have

$$\sup_{s\le t}\left|X_s^{N,2}-X_0^{N,2}\right| \le (e^{\gamma_o t}-1)+\left(\sup_{s\le t}\left|M_s^N\right|+\int_0^t X_s^{N,1}ds\right)e^{\gamma_o t}.$$

Define $T_{int}(\epsilon):=\inf\left\{t:\int_0^t X_s^{N,1}ds>\epsilon\right\}$.

For $t_1(N)$ and $\delta=\delta(N)\to 0$ satisfying $\gamma_o(N)t_1(N)<\delta(N)$ for sufficiently large N, on the event

$$\{t<T_M(\delta)\wedge T_{int}(\delta)\wedge t_1\},$$

we have

$$\sup_{s\le t}\left|X_s^{N,2}-X_0^{N,2}\right| = \left(\gamma_o t_1+O(\gamma_o^2 t_1^2)\right)+2\delta(1+\gamma_o t_1+O(\gamma_o^2t_1^2))<4\delta.$$

In other words, $\mathbb{P}\left[T_R(4\delta)>t\middle|t<T_M(\delta)\wedge T_{int}(\delta)\wedge t_1\right]=1$.

It follows that

$$\mathbb{P}\left[T_M(\delta)\wedge T_{int}(\delta)\wedge t_1\le T_R(4\delta)\right]=1. \tag{4.11}$$

Let $T_{sum}(x):=\inf\left\{t:X_t^{N,1}+X_t^{N,2}\le x\right\}$.

It follows from (4.11) that for sufficiently large N, on the event $\{t<T_M(\delta)\wedge T_{int}(\delta)\wedge t_1\}$,

$$X_t^{N,1}+X_t^{N,2}\ge X_0^{N,2}-4\delta>\frac{3r_0}{4},$$

which suggests that

$$\mathbb{P}\left[T_{sum}\left(\frac{r_0}{2}\right)>T_M(\delta)\wedge T_{int}(\delta)\wedge t_1\right]=1. \tag{4.12}$$

The inequality is strict because $(X^{N,1}+X^{N,2})$ has jump sizes of order N^{-1} and cannot reach $r_0/2$ from above $3r_0/4$ in one jump.

Also from (4.11),

$$\mathbb{P}\left[T_R(4\delta)<t_1\right]\le\mathbb{P}\left[T_M(\delta)<t_1\right]+\mathbb{P}\left[T_{int}(\delta)=T_{int}(\delta)\wedge T_M(\delta)<t_1\right].$$

The upper bound of $\mathbb{P}\left[T_M(\delta)<t_1\right]$ is obtained in (4.8). For the second term on the

RHS above, conditioned on the event $\{T_{int}(\delta) = T_{int}(\delta) \wedge T_M(\delta) < t_1\}$, the equality (4.12) is equivalent to

$$\mathbb{P}\left[T_{sum}\left(\frac{r_0}{2}\right) > T_{int}(\delta)\right] = 1.$$

Then

$$\mathbb{P}\left[T_{int}(\delta) = T_{int}(\delta) \wedge T_M(\delta) < t_1\right] \le \mathbb{P}\left[T_{int}(\delta) < T_{sum}\left(\frac{r_0}{2}\right) \wedge t_1\right].$$

On the event $\left\{t < T_{sum}\left(\frac{r_0}{2}\right) \wedge t_1\right\}$, the process I^N is dominated by a linear birth-death chain $L = (L_t)_{t\ge 0}$ with birth rate $\lambda_o\left(1 - \frac{r_0}{2}\right)$ and death rate 1. Therefore, $\int_0^t X_s^{N,1} ds$ is stochastically bounded by

$$N^{-1}\int_0^{T_L} L_s ds = N^{-1}H(I_0),$$

where T_L and $H(I_0)$ are defined as Lemma 4.5.

For any $t > 0$, the probability $\mathbb{P}\left[T_{int}(\delta) \le t\right]$ is then bounded by the probability $\mathbb{P}\left[H(I_0) > N\delta\right]$. By (4.4),

$$\begin{aligned}
&\mathbb{P}\left[T_{int}(\delta) = T_{int}(\delta) \wedge T_M(\delta) < t_1\right] \le \mathbb{P}\left[T_{int}(\delta) < t_1\right] \le \mathbb{P}\left[H(I_0) > N\delta\right] \\
\le &\frac{I_0}{(1 - \lambda_o + \lambda_o r_0/2)N\delta}.
\end{aligned}$$

Together with (4.8), we have

$$\mathbb{P}\left[T_R(4\delta) < t_1\right] \le 2\exp\left\{-\frac{\delta^2 N}{4(\gamma_o + 1)t_1}\right\} + \frac{I_0}{(1 - \lambda_o + \lambda_o r_0/2)N\delta}.$$

□

Now we are ready to discuss different cases under the second scenario, depending on the size of I_0.

Case 2.1: $I_0 = O(1)$ and $\gamma_o = o(1)$.

Let $\delta(N) = \gamma_o^{1-\epsilon} \vee N^{-1/3}$ for a small $\epsilon > 0$. Then we have any $t_1 = O(1)$ satisfies

$t_1 < \gamma_o^{-\epsilon} \leq \delta\gamma_o^{-1}$. By Lemma 4.7,

$$\mathbb{P}\left[\sup_{s\leq t_1}\left|R_s^N - R_0\right| > 4\delta N\right] \to 0.$$

Define $\delta_0(N) := |R_0/N - r_0|$.

For each $N \in \mathbb{N}$, define two linear birth-death chains $\underline{L}$ and $\overline{L}$ such that $\overline{L}$ has birth rate $\lambda_o(1-r_0+4\delta+\delta_0)$ and death rate 1, and $\underline{L}$ has birth rate $\lambda_o(1-r_0-4\delta-\delta_0)$ and death rate 1. Let $\underline{L}_0 = \overline{L}_0 = I_0$.

Denote

$$T_R(4\delta) := \inf\left\{t \geq 0 : \sup_{s\leq t}\left|R_s^N - R_0\right| > 4\delta N\right\}.$$

Then for $t < T_R(4\delta)$,

$$I_0\lambda_o(1 - r_0 - 4\delta - \delta_0) \leq I_0\lambda_o\left(1 - \frac{I_t^N + R_t^N}{N}\right) \leq I_0\lambda_o(1 - r_0 + 4\delta + \delta_0),$$

and all three terms tend to $\lim_{N\to\infty} I_0\lambda_o(1 - r_0)$.

Recall $a := \lim_{N\to\infty} 1 - \lambda_o + \lambda_o r_0 \in (0, \infty)$.

The conclusion then follows from Case 2 of Theorem 4.1:

$$\begin{aligned}\mathbb{P}\left[T_o^N \leq w\right] &= \mathbb{P}\left[T_o^N \leq w,\, T_R(4\delta) \leq w\right] + \mathbb{P}\left[T_o^N \leq w,\, T_R(4\delta) > w\right]\\ &\to \left(1 + \frac{a}{e^{aw} - 1}\right)^{-I_0},\end{aligned}$$

where we use the fact from above that for all $w > 0$

$$\mathbb{P}\left[T_o^N \leq w,\, T_R(4\delta) \leq w\right] \leq \mathbb{P}\left[T_R(4\delta) \leq w\right] \to 0.$$

Case 2.2 $I_0 \to \infty$ and there exists $\epsilon_1, \epsilon_2 > 0$ such that $I_0 = o\left(N^{1-\epsilon_1}\right)$ and $\gamma_o = o\left(N^{-\epsilon_2}\right)$.

Let $\delta = N^{-b\epsilon_2}$, where positive constant b is chosen to satisfy $b < \frac{1}{2\epsilon_2} \wedge \frac{\epsilon_1}{\epsilon_2} \wedge 1$. Then we have that any $t_1 = O(\log N)$ satisfies $t_1 < \delta\gamma_o^{-1} = N^{(1-b)\epsilon_2}$.

Since $\delta^2 N = N^{1-2b\epsilon_2}$ and $\frac{I_0}{\delta N} \ll N^{b\epsilon_2-\epsilon_1}$ are both of negative polynomial orders of N, by Lemma 4.7,

$$\mathbb{P}\left[\sup_{s\le t_1}\left|R_s^N - R_0\right| > 4\delta N\right] \to 0.$$

The constructions of $\underline{L}$ and $\overline{L}$, and the stopping time $T_R(4\delta)$ remain the same as Case 2.1. We still have that, with probability tending to 1, for $t \le T_R(4\delta)$,

$$\lambda_o(1 - r_0 - 4\delta - \delta_0) \le \lambda_o\left(1 - \frac{I_t^N + R_t^N}{N}\right) \le \lambda_o(1 - r_0 + 4\delta + \delta_0).$$

Let $a := \lim_{N\to\infty} 1 - \lambda_o + \lambda_o r_0 \in (0, \infty)$.

Following from Case 5 of Theorem 4.1, the extinction times of $\underline{L}$ and $\overline{L}$ tend to the same limit if $I_0(4\delta + \delta_0) = o\left(\frac{aI_0}{\log(aI_0)}\right)$, which is equivalent to $N^{-b\epsilon_2}\log N \to 0$.

The conclusion then follows from Case 5 of Theorem 4.1:

$$\begin{aligned}
&\mathbb{P}\left[aT_o^N - \log(aI_0) \le w\right] = \mathbb{P}\left[T_o^N \le a^{-1}\log(aI_0) + a^{-1}w < T_R(4\delta)\right] \\
&\quad + \mathbb{P}\left[T_o^N \le a^{-1}\log(aI_0) + a^{-1}w,\, T_R(4\delta) \le a^{-1}\log(aI_0) + a^{-1}w\right] \\
&\to e^{-e^{-w}},
\end{aligned}$$

where we use the fact that for any constants $c_1, c_2 > 0$,

$$\mathbb{P}\left[T_R(4\delta) \le c_1 \log N + c_2\right] \to 0.$$

4.4 Summary

We illustrate the conditions of different cases in the previous section using the diagram introduced in Definition 2.5. Notice that except for Case 1.1, it is assumed that $\lambda_o(N) \le 1$.

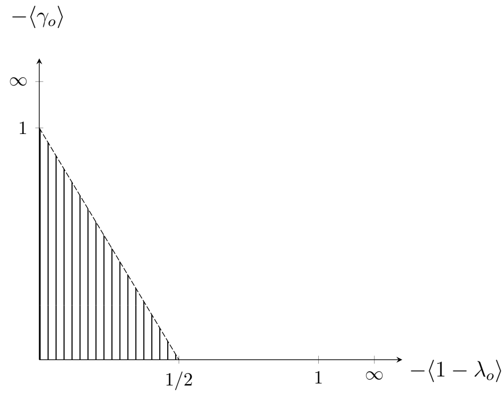


Figure 4.1: Parameter regime: divisions with small initial size of the infected population

In Figure 4.1, the area shaded with vertical lines represents

$$\left\{(\lambda_o,\gamma_o) : -\langle 1-\lambda_o\rangle < \frac{1+\langle\gamma_o\rangle}{2}\right\} \text{ or equivalently } \left\{(\lambda_o,\gamma_o) : N^{1/2}(1-\lambda_o)\gamma_o^{1/2} \to \infty\right\}.$$

In this regime, based on the scaling of (I_0, R_0), as illustrated in the first diagram in Figure 4.2, we can find behaviours of all five cases in Theorem 4.6.

Similarly, the complement area

$$\left\{(\lambda_o,\gamma_o) : -\langle 1-\lambda_o\rangle > \frac{1+\langle\gamma_o\rangle}{2}\right\} \text{ or equivalently } \left\{(\lambda_o,\gamma_o) : N^{1/2}(1-\lambda_o)\gamma_o^{1/2} \to 0\right\}$$

is illustrated in the second diagram in Figure 4.2.

The boundary scenario

$$\left\{(\lambda_o,\gamma_o) : -\langle 1-\lambda_o\rangle = \frac{1+\langle\gamma_o\rangle}{2}\right\}$$

is similar to the second diagram in Figure 4.2. In Chapter 4, we will give the illustration when $-\langle 1-\lambda_o\rangle = -\langle\gamma_o\rangle = 1/3$.

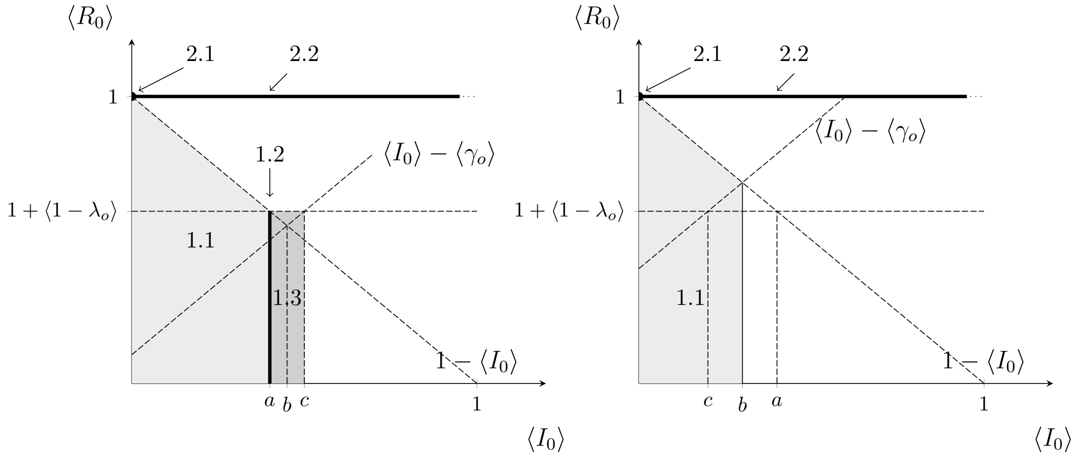


Figure 4.2: Diagram of small initial infection cases, where $a = -\langle 1 - \lambda_o \rangle$, $b = \frac{1+\langle \gamma_o \rangle}{2}$, $c = 1 + \langle 1 - \lambda_o \rangle + \langle \gamma_o \rangle$, and the numbers denote the cases in Theorem 4.6.

We can see for all the parameter combinations we are interested in, as long as the initial size of infection I_0 is sufficiently small, and the initial size of immunity R_0 satisfies certain conditions, we have the explicit expression of the asymptotic distribution of the extinction time T_o^N.

Chapter 5

The critical parameter regime

In this chapter, we identify the critical parameter regime of the stochastic SIRS model, and investigate the behaviour of its extinction time. We show that as $N \to \infty$, with suitable initial states, the stochastic SIRS model in the critical regime weakly converges to a degenerate diffusion, and the asymptotic distribution of the SIRS extinction time is equal to the distribution of the hitting time of the limit diffusion.

5.1 Introduction

The critical regime, also known as 'transition region' or 'critical window' in literature, is a subset of the parameter space of the stochastic epidemic models, in which we can observe phase transitions.

The existence of critical regimes in stochastic epidemic models was first discovered by Nåsell [4] in stochastic SIS models. In the last decade, a more detailed picture was established for the stochastic SIS and SIR models, both of which have one-dimensional parameter space $\lambda_o \in \mathbb{R}_+$. Our theoretical motivation is to establish the analogous result in the stochastic SIRS model, which has a two-dimensional parameter space $(\lambda_o, \gamma_o) \in \mathbb{R}_+^2$.

The stochastic SIS model has a critical regime of width $N^{-1/2}$. That is $\{\lambda_o : |\lambda_o(N) - 1| = O(N^{-1/2})\}$. This was observed by various authors through two different approaches: in [4], it is observed through the asymptotic approximation of the exact expression of quasi-stationary distributions and expected extinction times; whereas in [24] and [6], it is observed through performing a suitable scaling to a sequence of Markov chains indexed by the population size N in such way that the limit of the scaled processes is a diffusion. Such scaling is called the *critical scaling*. The authors of [24] comment that they cannot find a direct link between their observations and the one by [4].

The stochastic SIR model has a critical regime of width $N^{-1/3}$. This is observed by [40] and [24] through the scaling approach.

The behaviour of the extinction time within the critical regime is well studied only in the case of the stochastic SIS model. The earlier works include [25] and [5]. Nåsell [25] studied the quasi-stationary distribution, and the expected extinction time of SIS models initiated at quasi-stationary equilibrium, as well as at $I_0^N = 1$. Doering, Sargsyan and Sander [5] derived the expected extinction time of SIS models at criticality $\lambda_o = 1$ with an $O(1)$ error term. The most comprehensive study so far is by Foxall [6], who showed that the asymptotic distribution of the extinction time is equal to the distribution of the hitting time of the limit diffusion.

Less is known about the distribution of the hitting time of the diffusion obtained as the limit of stochastic epidemic models. Foxall [6] did not attempt to make any description of the hitting time itself. While studying the total size of the infection in the stochastic SIS model in the critical regime, Dolgoarshinnykh and Lalley [24] proposed the idea of random time change. Extending this idea to the stochastic SIR model, we can associate the extinction time with 'the hitting time of a Wiener process to a parabola' through a random time change, where the probability density of the latter is derived explicitly in Theorem 2, [40]. It seems that there is no straightforward way to extend the same idea to the stochastic SIRS model. The random time change approach is also not helpful in providing analytical approximations.

The limit of multidimensional discrete stochastic models in epidemiology and population genetics often turns out to be degenerate diffusions. There is no universal result for the well-posedness of the martingale problem/parabolic problems associated with degenerate generators, and different types of degeneracy are usually investigated on a case-by-case basis.

Our approach in the first half of this chapter is an extension of [6] to the stochastic SIRS model. In Section 5.2, we find the critical scaling of the stochastic SIRS model through a heuristic argument. For completeness, in Section 5.3, we state the asymptotic distribution of the extinction time within the critical regime, when the initial sizes of the infection is small. This is a straightforward corollary of Theorem 4.6. In Section 5.4, we prove that the extinction time of the stochastic SIRS model converges in distribution to the hitting time of the limit diffusion as $N \to \infty$. Special attention is required here as the limit diffusion is degenerate on the entire $\mathbb{R}_+^2$ with unbounded coefficients.

It is a standard practice to express the distribution of hitting time of diffusions as the solution of PDEs. As a result, we obtain a time-homogeneous PDE with the end condition in $\widehat{BC}(\mathbb{R}_+^2)$, the Banach space of bounded continuous functions which have continuous extensions to $[0,\infty)^2$. The well-posedness of the PDE problem does not directly follow from the well-posedness of the martingale problem on domain $C_c^\infty(\mathbb{R}_+^2)$, since $C_c^\infty(\mathbb{R}_+^2)$ is not dense in $(\widehat{BC}(\mathbb{R}_+^2), \|\cdot\|)$, where $\|\cdot\|$ denotes the uniform norm. To deal with this scenario, new types of semigroups have been defined in literature, usually with either a weaker continuity property, or a definition of strong continuity for a weaker topology. Analogous generation theorems and approximation theorems have been developed (See [7] for a comprehensive review). Kühnemund [7] develops a general framework of the bi-continuous semigroup to unite many of these individual results, and proves a generalised version of the Chernoff product formula.

The Chernoff product formula motivates a set of numerical approximations, which is often referred to in the area of computation as *splitting method* [41]. The much-better-known *Trotter product formula* can be viewed as a corollary of the classic Chernoff approximation. In Section 5.5, we introduce and apply the generalised Chernoff approximation based on the framework developed by [7], and prove the well-posedness of the PDE associated with the distribution of our extinction time.

5.2 Critical scaling of Markov chains

We would like to scale the SIRS process $(I_t^N, R_t^N)_{t\geq 0}$ in both time and space, and will denote the scaled process as $(Y_s^N, Z_s^N)_{s\geq 0}$. We use a heuristic argument to explore what the critical scaling should be.

Define scaled parameters $\hat{\lambda}$ and γ such that

$$\begin{aligned}
I_t^N =& N^\beta Y_{N^{-\alpha}t}^N, \quad 0<\beta\le 1,\\
R_t^N =& N^\xi Z_{N^{-\alpha}t}^N, \quad 0<\xi\le 1,\\
\lambda_o =& 1-\frac{\hat{\lambda}}{N^\delta}, \quad \delta>0,\\
\gamma_o =& \gamma N^{-\kappa}, \quad \kappa\ge 0.
\end{aligned}$$

Clearly, a stochastic SIRS process is subcritical when $\hat{\lambda}>0$, and supercritical when $\hat{\lambda}<0$.

Let $\left(Y_0^N, Z_0^N\right)=(y,z)$. Define $\Delta Y^N := Y_s^N - y$, $\Delta Z^N := Z_s^N - z$. Then we have the following expected linear and quadratic increments in time s:

$$\begin{aligned}
\mathbb{E}_{y,z}\left[\Delta Y^N\right] &= y(-\hat{\lambda}N^{-\delta} - N^{\beta-1}y - N^{\xi-1}z + \hat{\lambda}yN^{\beta-1-\delta} + \hat{\lambda}zN^{\xi-1-\delta})N^\alpha s + o(s)\\
&\sim y(-\hat{\lambda}N^{-\delta} - N^{\xi-1}z)N^\alpha s,\\
\mathbb{E}_{y,z}\left[\Delta Z^N\right] &\sim \left(N^{\beta-\xi}y - \gamma N^{-\kappa}z\right)N^\alpha s,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{y,z}\left[(\Delta Z^N)^2\right] \quad &\sim N^{-\xi}\left(N^{\beta-\xi}y + \gamma N^{-\kappa}z\right)N^\alpha s,\\
\mathbb{E}_{y,z}\left[(\Delta Y^N)^2\right] \quad &= N^{-\beta}y\left(2 - N^{\beta-1}y - N^{\xi-1}z + \lambda yN^{\beta-1-\delta} + \hat{\lambda}zN^{\xi-1-\delta}\right)N^\alpha s + o(s)\\
&\sim 2yN^{\alpha-\beta}s,\\
\mathbb{E}_{y,z}\left[\Delta Z^N \Delta Y^N\right] &= -N^{-\beta-\xi}N^\beta yN^\alpha s + o(s) \sim -N^{\alpha-\xi}ys.
\end{aligned}$$

The first observation is that a non-trivial scaling should satisfy $\beta<\xi$; otherwise, $\mathbb{E}_{y,z}\left[\Delta Z^N\right]$ would be asymptotically dominating the other four expectations above.

The second observation is that $\mathbb{E}_{y,z}\left[(\Delta Z^N)^2\right]$ and $\mathbb{E}_{y,z}\left[\Delta Z^N \Delta Y^N\right]$ are both asymptotically dominated by the $O\left(N^{\alpha+\beta-\xi}\right)$ term in $\mathbb{E}_{y,z}\left[\Delta Z^N\right]$. Therefore, out of the five expectations above, only the following three can be of the leading order of magnitude:

$$\begin{aligned}\mathbb{E}_{y,z}\left[\Delta Y^N\right] &\sim y(-\hat{\lambda}N^{-\delta} - N^{\xi-1}z)N^{\alpha}s,\\ \mathbb{E}_{y,z}\left[\Delta Z^N\right] &= \left(N^{\beta-\xi}y - \gamma N^{-\kappa}z\right)N^{\alpha}s,\\ \mathbb{E}_{y,z}\left[(\Delta Y^N)^2\right] &\sim 2yN^{-\beta}N^{\alpha}s.\end{aligned}$$

The only possibility for all three terms above to be of the same order of N is when $\alpha = \beta = \delta = \kappa = 1/3$, $\xi = 2/3$ and $t = N^{1/3}s$. In other words, it is natural to use the scaled parameters

$$\hat{\lambda}(N) := (1-\lambda_o)N^{1/3}, \quad \gamma(N) := \gamma_o N^{1/3}, \tag{5.1}$$

and consider the following scaled SIRS process:

$$\begin{aligned}Y_t^N &:= \frac{I^N_{N^{1/3}t}}{N^{1/3}},\\ Z_t^N &:= \frac{R^N_{N^{1/3}t}}{N^{2/3}},\end{aligned} \tag{5.2}$$

given the deterministic initial state $(Y_0^N, Z_0^N) = (y_0^N, z_0^N)$. We may drop the subscript or superscript of $\mathbb{P}$ when there is no confusion about which initial state we are discussing.

It is not unexpected that the scaling we applied to (I^N, R^N) and λ_o is consistent with what is found for stochastic SIR model in [24], as the latter is a special case of stochastic SIRS model with $\gamma_o = 0$.

The transition rates of (Y^N, Z^N) at (y, z) are

$$\begin{aligned}&(y,z) \to (y + N^{-1/3}, z) \text{ at rate } N^{2/3}(1-\hat{\lambda}N^{-1/3})(1 - yN^{-2/3} - zN^{-1/3})y,\\ &(y,z) \to (y, z - N^{-2/3}) \text{ at rate } N^{2/3}\gamma z,\\ &(y,z) \to (y - N^{-1/3}, z + N^{-2/3}) \text{ at rate } N^{2/3}y.\end{aligned} \tag{5.3}$$

The critical scaling divides the parameter space into four regimes depending on whether $-\langle 1-\lambda_o\rangle$ and $-\langle \gamma_o\rangle$ are smaller than 1/3. See Section 2.3.2 for the definition of $\langle\cdot\rangle$.

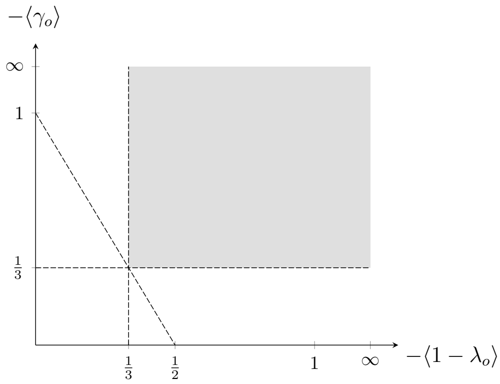


Figure 5.1: Parameter regime: the shaded area represents the critical regime.

In this chapter, we look at the parameter regime $(-\langle 1-\lambda_o\rangle, -\langle\gamma_o\rangle) \in [1/3,\infty)^2$, which is called the *critical parameter regime*.

5.3 Small infection

For completeness, we state the asymptotic distribution of the extinction time in the critical regime with small initial size of the infection.

Theorem 5.1. *Let the parameters* $(\lambda_o(N), \gamma_o(N))$ *satisfy* $|1-\lambda_o|N^{1/3} \to \hat{\lambda} \in \mathbb{R}$ *and* $\gamma_o N^{1/3} \to \gamma \geq 0$*, and let the initial state of the stochastic SIRS model be* $(I_0^N, R_0^N) = (I_0, R_0)$*. Then we have the following results:*

1. *When* $I_0 = o\left(N^{1/3}\right)$ *and* $R_0 = o\left(NI_0^{-1}\right)$*:*

   if in addition $I_0 \to \infty$*,*

$$\lim_{N\to\infty} \mathbb{P}^N\left[\frac{T_o^N}{I_0} \leq w\right] = e^{-\frac{1}{w}},\ w > 0;$$

   if in addition $I_0 = O(1)$*,*

$$\lim_{N\to\infty} \mathbb{P}^N\left[\frac{T_o^N}{I_0} \leq w\right] = \left(1+\frac{1}{w}\right)^{-I_0},\ w > 0.$$

2. When $I_0 = O(1)$ *and* $R_0 \sim r_0 N$ *for* $r_0 \in (0,1]$*,*

$$\lim_{N\to\infty} \mathbb{P}^N \left[T_o^N \le w\right] = \left(1 + \frac{r_0}{e^{r_0 w} - 1}\right)^{-I_0}, \; w > 0.$$

3. When $I_0 = o\left(N^{1-\epsilon}\right)$ *for some* $\epsilon > 0$*, and* $R_0 \sim r_0 N$ *for* $r_0 \in (0,1]$*,*

$$\lim_{N\to\infty} \mathbb{P}^N \left[r_0 T^N - \log(r_0 I_0) \le w\right] = e^{-e^{-w}}, \; w \in \mathbb{R}.$$

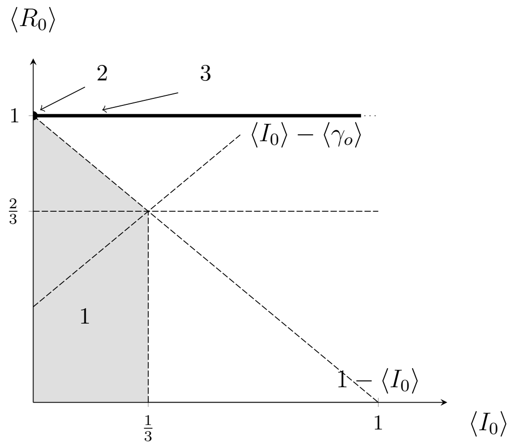


Figure 5.2: Diagram of small initial infection cases, where $-\langle 1 - \lambda_o \rangle = \frac{1+\langle \gamma_o \rangle}{2}$

Proof. This is a corollary of Theorem 4.6. $\square$

For the rest of this chapter, we focus on the case where $I_0 \asymp N^{1/3}$ and $R_0 \asymp N^{2/3}$.

5.4 Diffusion limit

Given a fixed initial state of infection $(I_0^N, R_0^N) = (I_0(N), R_0(N))$ in our original model, the initial state of the scaled process (Y^N, Z^N) is also a function of N: $(y_0^N, z_0^N) := (N^{-1/3} I_0(N), N^{-2/3} R_0(N))$. Under this scaling, we consider the cases where $\lim_{N\to\infty} (y_0^N, z_0^N) \to (y_0, z_0) \in \mathbb{R}_+^2$.

We shall denote the extinction time of Y^N as

$$T^N := \inf \left\{ t : Y_t^N = 0 \right\}.$$

The stopping time T^N is related to the extinction time T_o^N of the original process in the following sense: given $(Y_0^N, Z_0^N) = (N^{-1/3} I_0(N), N^{-2/3} R_0(N))$,

$$T^N = T_o^N N^{-1/3}.$$

From the heuristic analysis above, we expect the diffusion limit of (Y^N, Z^N) to be a Markov process generated by some extension of the operator

$$\mathcal{G} f(y,z) = -(\hat{\lambda} + z) y \partial_y f + (y - \gamma z) \partial_z f + y \partial_{yy} f, \quad f \in C_c^2(\mathbb{R}_+^2). \tag{5.4}$$

The covariance matrix $\left(\begin{smallmatrix} 2y & 0 \\ 0 & 0 \end{smallmatrix}\right)$ is degenerate on the entire domain $\mathbb{R}_+^2$. Since the operator is not uniformly elliptic, we need to make sure that such limit process is well-defined, which is equivalent to proving the $\mathcal{G}$-martingale problem with initial state $(y_0, z_0) \in \mathbb{R}_+^2$ is well-posed.

This particular type of degeneracy is studied in [9].

Theorem 5.2 (Theorem 5.14, [9])**.** *Let $d = d_0 + d_1$ with $d_1 \le d_0$. Let $a : [0,\infty) \times \mathbb{R}^d \to \mathbf{S}^{d_0}$, $b^{(0)} : [0,\infty) \times \mathbb{R}^d \to \mathbb{R}^{d_0}$ be measurable functions, and $b^{(1)} \in C^{1,2,1}(\mathbb{R}_+ \times \mathbb{R}^{d_0} \times \mathbb{R}^{d_1}, \mathbb{R}^{d_1})$. Let $b : [0,\infty) \times \mathbb{R}^d \to \mathbb{R}^d$ denote $b(s,x) = (b^{(0)}(s,x), b^{(1)}(s,x))^\intercal$. Suppose that for all $T > 0$, $x \in \mathbb{R}^d$,*

$$\inf_{s \in [0,T]} \inf_{\substack{\theta \in \mathbb{R}^{d_0}, \\ |\theta| = 1}} \langle \theta, a(s,x) \theta \rangle > 0,$$

$$\lim_{y \to x} \sup_{s \in [0,T]} \| a(s,y) - a(s,x) \| = 0.$$

Further suppose that there exists constant C such that

$$\sup_{s \in [0,\infty)} \| a(s,x) \| + |b(s,x)|^2 \le C(1 + |x|^2),$$

and that the Jacobian matrix of $b^{(1)}(s,x)$ restricted to the first d_0 components is of *rank* d_1 *for all* $(s,x) \in [0,\infty) \times \mathbb{R}^d$*, i.e.,*

$$\operatorname{rank} \begin{pmatrix} \frac{\partial b_{d_0+1}(s,x)}{\partial x_1} & \cdots & \frac{\partial b_{d_0+1}(s,x)}{\partial x_{d_0}} \\ \vdots & \ddots & \vdots \\ \frac{\partial b_{d_0+d_1}(s,x)}{\partial x_1} & \cdots & \frac{\partial b_{d_0+d_1}(s,x)}{\partial x_{d_0}} \end{pmatrix} = d_1.$$

Denote

$$\mathcal{A} := \sum_{1 \le m \le d} b_m(s,x) \frac{\partial}{\partial x_m} + \frac{1}{2} \sum_{1 \le m,n \le d_0} a_{mn}(s,x) \frac{\partial^2}{\partial x_m \partial x_n}.$$

Then the $(\mathcal{A}, C_c^\infty(\mathbb{R}^2))$*-martingale problem is well-posed.*

For convenience, in the rest of this subsection, we also refer to the degenerate martingale problem in Theorem 5.2 as the (a,b)-martingale problem. This should not be confused with the term '$(\mathcal{A}, D)$-martingale problem' defined in Definition 3.4.

Notice that a is valued in S^{d_0} rather than S^d. The (a,b)-martingale problem stopped by τ is then defined as finding (Y_t, Z_t) with initial state (y,z) such that

$$f(Y_{t\wedge\tau}, Z_{t\wedge\tau}) - \int_0^{t\wedge\tau} \mathcal{A} f(Y_s, Z_s) ds$$

is a martingale for all $f \in C_c^\infty(\mathbb{R}_+^2)$.

We apply the localisation argument of a generalised martingale problem on subsets of $\mathbb{R}^2$ from [10] to Theorem 5.2 in order to study the $\mathcal{G}$-martingale problem defined at the beginning of this section.

We will frequently be using the following sequence of functions in the rest of this chapter.

Definition 5.3 (Exhaustion of $\mathbb{R}_+^2$ and $\{\iota_n\}_{n\in\mathbb{N}}$)**.** Let $\{D_n\}_{n\in\mathbb{N}}$ be a sequence of open subsets of $\mathbb{R}_+^2$ such that $\bar{D}_n \subset D_{n+1}$ and $\bigcup_{n=1}^\infty D_n = \mathbb{R}_+^2$. We refer to said sequence $\{D_n\}_{n\in\mathbb{N}}$ as an exhaustion of $\mathbb{R}_+^2$.

We define a sequence of functions $\{\iota_n\}_{n\in\mathbb{N}} \subset C_c^\infty(\mathbb{R}_+^2)$ on $\{D_n\}_{n\in\mathbb{N}}$ satisfying:

$$\iota_n(u,v)\begin{cases} = 1 & (u,v) \in \bar{D}_n, \\ \in [0,1] & (u,v) \in D_{n+1}\backslash \bar{D}_n, \\ = 0 & (u,v) \in \mathbb{R}_+^2\backslash \bar{D}_{n+1}. \end{cases} \tag{5.5}$$

Proposition 5.4. *For $\mathcal{G}$ defined in* (5.4)*, the $(\mathcal{G}, C_c^\infty(\mathbb{R}_+^2))$-martingale problem is well-posed.*

Proof. Denote the one-point compactification of $D = \mathbb{R}_+^2$ as $\widehat{\mathbb{R}_+^2} = \mathbb{R}_+^2 \cup \{\Delta\}$. Take the exhaustion of $\mathbb{R}_+^2$ as $\{D_n := (1/n, n)^2\}_{n\in\mathbb{N}, n\geq 2}$. Define stopping times $\tau_n := \inf\{t : Y_t \notin D_n\}$. Then $\tau_\Delta \equiv \lim_{n\to\infty} \tau_n$ is also a stopping time.

Let $\widehat{\Omega}_D$ be the space of continuous trajectories valued in $\widehat{D} = \widehat{\mathbb{R}_+^2}$, satisfying the following: either $\tau_\Delta = \infty$, or $\tau_\Delta < \infty$ and $(Y_{\tau_\Delta+t}, Z_{\tau_\Delta+t}) = \Delta$ for all $t > 0$. Let $\widehat{\mathscr{F}^D}$ denote the Borel σ-algebra on $\widehat{\Omega}_D$ and define the filtration $\widehat{\mathscr{F}_t^D} = \sigma((Y_s, Z_s), s \in [0,t])$.

Define

$$\begin{aligned} a_n(y,z) &:= |2y|\iota_n(y,z) + (1 - \iota_n(y,z)), \\ b_n(y,z) &:= \begin{pmatrix} -(\hat{\lambda} + z)y \\ y - \gamma z \end{pmatrix} \iota_n(y,z) + \begin{pmatrix} 0 \\ y \end{pmatrix} (1 - \iota_n(y,z)). \end{aligned}$$

Let $\Omega = C([0,\infty), \mathbb{R}^2)$ and $\mathscr{F}$ be the Borel σ-algebra on Ω. Define $\mathscr{F}_t = \sigma((Y_s, Z_s), s \in [0,t])$.

By Theorem 5.2, for each n, the (a_n, b_n)-martingale problem with initial state $(y,z) \in \mathbb{R}^2$ has a unique solution $\{\mathbb{P}_{y,z}^{(n)}, (y,z) \in \mathbb{R}^2\}$ on $(\Omega, \mathscr{F}_{\tau_n})$.

Our localisation argument relies on the statement of Theorem 13.1, [10], the proof of which proceeds exactly as in Theorem 10.4, pp. 34-35, [10], applying Theorem 10.5 of the same reference. The proof only requires the well-posedness of the original martingale problem and therefore is applicable to our degenerate generator. It follows that there exists a unique solution $\{\widehat{\mathbb{P}}_{y,z}, (y,z) \in \widehat{D}\}$ on $(\widehat{\Omega}_D, \widehat{\mathscr{F}^D})$ to the generalised (a,b)-martingale problem with initial state $(y,z) \in D$ which satisfies

$$\widehat{\mathbb{P}}_{y,z}\Big|_{\widehat{\mathscr{F}}^D_{\tau_n}} = \mathbb{P}^{(n)}_{y,z}\Big|_{\mathscr{F}_{\tau_n}},$$

for all $n \in \mathbb{N}$, where

$$a(y,z) = |2y|, \quad b(y,z) = \begin{pmatrix} -(\hat{\lambda}+z)y \\ y - \gamma z \end{pmatrix}.$$

We now argue that $\{\widehat{\mathbb{P}}_{y,z}, (y,z) \in D\}$ corresponds to a unique solution $\{\mathbb{P}_{y,z}, (y,z) \in D\}$ to the (a,b)-martingale problem on $(\Omega, \mathscr{F})$.

Let (Y_t, Z_t) be the solution corresponding to the measure $\{\widehat{\mathbb{P}}_{y,z}, (y,z) \in D\}$, then it satisfies the following SDE system for $t \in [0, \tau_\Delta]$:

$$\begin{aligned} dY_t &= -(\hat{\lambda} + Z_t)Y_t dt + \sqrt{2Y_t} dW_t, \\ dZ_t &= (Y_t - \gamma Z_t)dt. \end{aligned}$$

Recall that the process approaches Δ when one or both of Y_t and Z_t approaches 0 or ∞.

Firstly, notice the solution $(Y_{t\wedge\tau_\Delta}, Z_{t\wedge\tau_\Delta})$ with $(y,z) \in D$ satisfies

$$Z_{t\wedge\tau_\Delta} = \int_0^{t\wedge\tau_\Delta} e^{-\gamma(t\wedge\tau_\Delta - s)} Y_s ds + e^{-\gamma(t\wedge\tau_\Delta)} z.$$

Hence $\inf\{t : Y_t = 0\} < \inf\{t : Z_t = 0\}$ and $\inf\{t : Y_t = \infty\} \leq \inf\{t : Z_t = \infty\}$. It follows that the only possible scenarios are either $\tau_\Delta = \inf\{t : Y_t = 0\}$ or $\tau_\Delta = \inf\{t : Y_t = \infty\}$.

Secondly, we show that in fact $\inf\{t : Y_t = \infty\} = \infty$.

Since $Z_t - e^{-\gamma t} z$ is a strictly increasing function of t given $\{t \leq \tau_\Delta\}$, all $\omega \in \widehat{\Omega}_D$ belong to exactly one of the following two events:

1. $Z_t - ze^{-\gamma t} < |\hat{\lambda}|$ for all $t \in [0, \tau_\Delta)$, in which case Y must have reached 0 before reaching infinity;

2. There exists $\tau(\omega) < \tau_\Delta(\omega)$ such that $\tau = \inf\{t : Z_t - e^{-\gamma t} z \geq |\hat{\lambda}|\}$, in which case $\hat{\lambda} + Z_{\tau+t} \geq 0$ for all $t > 0$ and $Y_{\tau+t}$ will have a non-positive drift and

the same diffusion coefficient as a squared Bessel process. By Comparison Theorem 3.15, $Y_{\tau+t}$ is stochastically dominated by $\frac{1}{2}BESQ^1(2y)$, as defined in Definition 5.5, while the latter absorbs at 0 almost surely (p.314, [42]). Hence Y_t reaches 0 almost surely.

In either event, (Y, Z) is non-explosive, and we can conclude $\tau_\Delta = \inf\{t : Y_t = 0\}$.

Finally, notice that the (a, b)-martingale problem with initial state $(0, z)$, $z > 0$ has a unique trivial solution

$$Y_t = 0,\ Z_t = e^{-\gamma t} z > 0.$$

For any $\widehat{B} \subset \widehat{\Omega}_D$ that is not a null event of $\widehat{\mathbb{P}}_{y,z}$, we can find a set $B \subset \Omega$ such that $\mathbb{P}_{y,z}(B) = \widehat{\mathbb{P}}_{y,z}(\widehat{B})$ by replacing the trajectory $(Y_{\tau_\Delta+t}, Z_{\tau_\Delta+t}) = \Delta$, $t \geq 0$ in $\widehat{B}$ to be $(Y_{\tau_\Delta+t}, Z_{\tau_\Delta+t}) = (0, e^{-\gamma t} Z_{\tau_\Delta})$. □

The infinitesimal generator $\mathcal{G}$ has the same diffusion part as a well-studied family of processes, namely the squared Bessel process. It is helpful to introduce the basic properties of the squared Bessel process. See [42] for a nice survey of this topic.

Definition 5.5 (Squared Bessel process)**.** For every $\delta \geq 0$ and $x_0 \geq 0$, the unique strong solution to the equation

$$X_t = x_0 + \delta t + 2\int_0^t \sqrt{|X_s|}dW_s$$

is called the δ-dimensional squared Bessel process starting at x_0 and is denoted by $BESQ^\delta(x_0)$.

Denote for $u > 0$ the transition density function of $X_t \sim BESQ^0$ from u to $m > 0$ as $q_t^{BESQ}(u, m)$. We have (see Appendix A.2, [42])

$$q_t^{BESQ}(u, m) = \frac{1}{2t} u^{1/2} m^{-1/2} e^{-(u+m)/2t} I_1\left(\frac{m^{1/2}u^{1/2}}{t}\right),$$

for $m > 0$, where $I_1(\cdot)$ is the modified Bessel function of the first kind of index 1, and

$$\mathbb{P}_u\left[X_t = 0\right] = \exp\left\{-\frac{u}{2t}\right\}. \tag{5.6}$$

It is also useful to note that X_t will reach absorption at 0 almost surely.

Definition 5.6 (Modified Bessel function of the first kind, I_α, pp.375-377, [43])**.** The modified Bessel function of the first kind of index α is defined as

$$I_\alpha(x) := \sum_{m=0}^{\infty} \frac{1}{m!\,\Gamma(m+\alpha+1)} \left(\frac{x}{2}\right)^{2m+\alpha}. \tag{5.7}$$

We have the properties

$$I_\alpha(x) \sim \frac{1}{\Gamma(\alpha+1)} \left(\frac{x}{2}\right)^{\alpha},\ x \downarrow 0, \tag{5.8}$$

and

$$I_\alpha(x) \sim \frac{e^x}{\sqrt{2\pi x}},\ x \to \infty. \tag{5.9}$$

Theorem 5.7. *Let* $(1-\lambda_o(N))N^{1/3} \to \hat{\lambda} \in \mathbb{R}$, $\gamma := \lim_{N\to\infty} \gamma_o(N)N^{1/3} \geq 0$ *and* $Y_0^N \to y_0 > 0$, $Z_0^N \to z_0 > 0$. *Then the process* (Y^N, Z^N) *converges in distribution to* (Y, Z) *in* $D([0,\infty), \mathbb{R}^2)$, *where* (Y, Z) *is the unique weak solution to the stochastic differential equation system*

$$\begin{aligned} dY &= -(\hat{\lambda} + Z)Y\,ds + \sqrt{2Y}\,dW, \\ dZ &= (Y - \gamma Z)ds, \end{aligned} \tag{5.10}$$

with initial conditions $Y_0 = y_0$, $Z_0 = z_0$.

Proof. The system (5.10) corresponds to the infinitesimal generator

$$\mathcal{G}f(y,z) = -(\hat{\lambda}+z)y\partial_y f + (y-\gamma z)\partial_z f + y\partial_{yy} f,\ f \in C_c^2(\mathbb{R}^2).$$

The well-posedness of the $\mathcal{G}$-martingale problem implies the existence and uniqueness of a weak solution to (5.10). To prove the weak convergence to the diffusion limit, it remains to check that all the assumptions in Theorem 3.9 are satisfied.

Firstly, for each y,

$$\begin{aligned}
b_1^N(y) &= y(-\hat{\lambda}N^{-1/3} - N^{1/3-1}y - N^{2/3-1}z + \hat{\lambda}yN^{1/3-1-1/3} + \hat{\lambda}zN^{2/3-1-1/3})N^{1/3} \\
&\to y(-\hat{\lambda} - z), \\
b_2^N(y) &= \left(N^{1/3-2/3}y - \gamma N^{-1/3}z\right) N^{1/3} \to y - \gamma z,
\end{aligned}$$

and

$$\begin{aligned}
a^N(y) = &\begin{pmatrix} N^{-2/3} & -N^{-1} \\ -N^{-1} & N^{-4/3} \end{pmatrix} N^{2/3}y + \begin{pmatrix} 0 & 0 \\ 0 & N^{-4/3} \end{pmatrix} N^{2/3}\gamma z \\
&+ \begin{pmatrix} N^{-2/3} & 0 \\ 0 & 0 \end{pmatrix} N^{2/3}y\left(1 - N^{1/3-1}y - N^{2/3-1}z\right)\left(1 - \hat{\lambda}N^{-1/3}\right) \\
&\to \begin{pmatrix} 2y & 0 \\ 0 & 0 \end{pmatrix}.
\end{aligned}$$

Secondly, the maximum jump size (under Euclidean norm) is bounded by $(N^{-2/3} + N^{-4/3})^{1/2} \to 0$, and lastly, the convergence of the initial state is assumed. The weak convergence then follows. □

Denote the extinction time of the diffusion limit as

$$T := \inf\{t : Y_t = 0\}.$$

Next, we want to prove that T^N weakly converges to T, given the initial states converge.

The weak convergence to the diffusion limit is not sufficient for concluding the weak convergence of the extinction times. Essentially, we would also need the family $\{X^N\}_{N\in\mathbb{N}}$ to have a tendency of moving downward.

Theorem 5.8. *Let* $(Y_0^N, Z_0^N) = (u_N, v_N)$ *and* $(Y_0, Z_0) = (u, v)$*. If* $(u_N, v_N) \to (u, v) \in \mathbb{R}_+^2$*, then* $T^N \Rightarrow T$.

Our proof of Theorem 5.8 is inspired by [6], in which an analogous statement was proved in one dimension.

Proof. Let the probability measure $\mathbb{P}_{u,v}$ be the solution to the $(\mathcal{G}, C_c^\infty(\mathbb{R}_+^2))$-martingale problem with initial state (u, v).

For $y \geq 0$, define $\tau_y(f) := \inf\{t : (f(t))_1 \leq y\}$ for càdlàg $f \in D([0,\infty), \mathbb{R}^2)$ with the topology of uniform convergence on compacts (ucc), where $(f(t))_1$ denotes the first component of $f(t)$.

We first show that for small $y > 0$, τ_y is a.s. continuous at (Y, Z).

For each sample path ω, taking any sequence $\{f_n\}_{n\in\mathbb{N}} \subset D([0,\infty), \mathbb{R}^2)$ such that $\text{ucc} - \lim_{n\to\infty} f_n = f := (Y(\omega), Z(\omega))$, we have that the following two statements are true:

1. For small $y > 0$, $\liminf_{n\to\infty} \tau_y(f_n) \geq \tau_y(f)$.

2. For small $y > 0$, $\limsup_{n\to\infty} \tau_y(f_n) \leq \tau_y(f)$.

Item 1 is in fact true for all $f \in D([0,\infty), \mathbb{R}^2)$. Otherwise, i.e., if $\liminf_{n\to\infty} \tau_y(f_n) < \tau_y(f)$, then there exist s such that

$$\liminf_{n\to\infty} \tau_y(f_n) < s < \tau_y(f),$$

and a subsequence $\{f_{n_i}\}_{i\in\mathbb{N}}$ satisfying $\lim_{i\to\infty} \tau_y(f_{n_i}) < s$. This contradicts to $\text{ucc} - \lim_{i\to\infty} f_{n_i} = f$, since $\inf\{(f(t))_1 : t \leq s\} > y$.

Item 2 requires a little more work.

It is sufficient to prove that for all

$$y \in \begin{cases} (0,\infty), & \hat{\lambda} \geq 0, \\ (0, |4\hat{\lambda}|^{-1}), & \hat{\lambda} < 0, \end{cases} \tag{5.11}$$

we have

$$\limsup_{n\to\infty} \tau_y(f_n) \leq \tau_y(f).$$

Consider $(Y(\omega), Z(\omega))$ after time $\tau_y(f)$. We claim that for any $\epsilon > 0$, the path of Y a.s. intersects $[0, y)$ by time $\tau_y(f) + \epsilon$.

Let $\tau^\epsilon := \inf\{t : Y_t \geq 2y\} \wedge \epsilon$. It is sufficient to show that for any $(Y_0, Z_0) = (y, z)$,

$z \geq 0$, the following holds almost surely:

$$\inf_{t\in(0,\tau^\epsilon)} Y_t < y.$$

From the second equation of (5.10), we have

$$\begin{aligned} Z_{t\wedge\tau^\epsilon} &= \int_0^{t\wedge\tau^\epsilon} e^{-\gamma(t\wedge\tau^\epsilon - s)} Y_s ds + e^{-\gamma(t\wedge\tau^\epsilon)} z \\ &\leq 2y\gamma^{-1} + (z - 2y\gamma^{-1})e^{-\gamma(t\wedge\tau^\epsilon)} =: \bar{z}(t), \end{aligned}$$

and $\bar{z}(t)$ is bounded on $t \in (0, \tau^\epsilon]$.

Our choice of y guarantees that the drift term of $Y_{t\wedge\tau^\epsilon}$ is bounded for $t \in (0, \tau^\epsilon]$ regardless of the sign of $\hat{\lambda}$:

- if $\hat{\lambda} \geq 0$, then $0 \geq -(\hat{\lambda} + Z_t)Y_t \geq -2y(\hat{\lambda} + \bar{z}(t))$; and
- if $\hat{\lambda} < 0$, then $1/2 \geq -(\hat{\lambda} + Z_t)Y_t \geq -2y\bar{z}(t)$.

For $t \leq \tau^\epsilon$, by comparison theorem 3.15, $2Y_t$ is stochastically dominated by $BESQ^1(2y)$, which can also be seen as the square of Wiener process $\widetilde{W}_t$, i.e., $2Y_t$ is stochastically dominated by $(\widetilde{W}_t + \sqrt{2y})^2$. For any $\epsilon' > 0$ and $\delta > 0$,

$$\begin{aligned} \mathbb{P}_{y,z}\left[\inf_{s\leq\epsilon'} Y_s \leq y - \delta \middle| \epsilon' < \tau^\epsilon\right] &\geq \mathbb{P}\left[\inf_{s\leq\epsilon'} (\widetilde{W}_s + \sqrt{2y})^2 \leq 2(y-\delta)\right] \\ = \mathbb{P}\left[\sup_{s\leq\epsilon'} \widetilde{W}_s \geq \sqrt{2y} - \sqrt{2(y-\delta)}\right] &= 2\mathbb{P}\left[\widetilde{W}_{\epsilon'} \geq \sqrt{2y} - \sqrt{2y - 2\delta}\right], \end{aligned}$$

where the first equality above uses the fact that $\widetilde{W}$ and $-\widetilde{W}$ have the same distribution, and the second equality follows from the reflection principle of Wiener processes. It follows that

$$\mathbb{P}_{y,z}\left[\inf_{s\leq\epsilon'} Y_s < y \middle| \epsilon' < \tau^\epsilon\right] = \lim_{\delta\downarrow 0} \mathbb{P}_{y,z}\left[\inf_{s\leq\epsilon} Y_s \leq y - \delta \middle| \epsilon' < \tau^\epsilon\right] = 1,$$

Since $\mathbb{P}_{y,z}[\tau^\epsilon \in (0, \epsilon]] = 1$, we can take any positive sequence $\{\epsilon_n\}_{n\in\mathbb{N}}$ converging to

0 and have

$$\mathbb{P}_{y,z}\left[\inf_{s\leq\tau^\epsilon} Y_s < y\right] = \mathbb{P}_{y,z}\left[\bigcup_{n\in\mathbb{N}}\left\{\inf_{s\leq\epsilon_n} Y_s < y \text{ and } \epsilon_n < \tau^\epsilon\right\}\right]$$
$$= \lim_{\epsilon'\downarrow 0}\mathbb{P}_{y,z}\left[\inf_{s\leq\epsilon'} Y_s < y \middle| \tau^\epsilon > \epsilon'\right]\mathbb{P}_{y,z}\left[\tau^\epsilon > \epsilon'\right] = 1.$$

For each sample path ω, take any sequence $\{f_n\}_{n\in\mathbb{N}}$ converging to $(Y(\omega), Z(\omega))$ uniformly on compacts, sufficiently small $y > 0$ in the sense of (5.11), and any $\epsilon > 0$, we have $\tau_y(f_n) < \tau_y(f) + \epsilon$ for all sufficiently large n.

Combining (1) and (2) above, for $y \in (0, |4\hat{\lambda}|^{-1})$, we have τ_y a.s. continuous at (Y, Z).

Given the weak convergence $(Y^N, Z^N) \Rightarrow (Y, Z)$, the Skorokhod Representation Theorem suggests that it is possible to choose a sample space in which $(Y^N, Z^N) \to (Y, Z)$ almost surely in the Skorokhod topology. Moreover, since (Y, Z) is continuous, it follows that $\text{ucc} - \lim_{N\to\infty}(Y^N, Z^N) = (Y, Z)$ almost surely. By the *continuous mapping theorem*, for $y \in (0, |4\hat{\lambda}|^{-1})$,

$$\tau_y((Y^N, Z^N)) \Rightarrow \tau_y((Y, Z)). \tag{5.12}$$

Our goal is to show that

$$\lim_{N\to\infty}\mathbb{P}^N_{u_N,v_N}\left[T^N \leq t\right] = \mathbb{P}_{u,v}\left[T \leq t\right],$$

at all continuous points t of the RHS. This breaks down to showing the following

$$\limsup_{N\to\infty}\mathbb{P}^N_{u_N,v_N}\left[T^N < t\right] \leq \mathbb{P}_{u,v}\left[T < t\right], \tag{5.13}$$
$$\liminf_{N\to\infty}\mathbb{P}^N_{u_N,v_N}\left[T^N < t\right] \geq \mathbb{P}_{u,v}\left[T < t\right]. \tag{5.14}$$

Since (Y, Z) is continuous, $\lim_{y\downarrow 0}\tau_y((Y, Z)) = T$ almost surely, which implies

$$\lim_{y\downarrow 0}\mathbb{P}_{u,v}\left[\tau_y((Y, Z)) < t\right] = \mathbb{P}_{u,v}\left[T < t\right]. \tag{5.15}$$

We also have $\lim_{y\downarrow 0}\tau_y((Y^N,Z^N))\le T^N$, which implies

$$\limsup_{N\to\infty}\mathbb{P}^N_{u_N,v_N}\left[T^N<t\right]\le\limsup_{y\downarrow 0}\lim_{N\to\infty}\mathbb{P}_{u_N,v_N}\left[\tau_y((Y^N,Z^N))<t\right]. \tag{5.16}$$

The inequality (5.13) follows from (5.15) and(5.16).

For the opposite direction, consider any large N and denote $S_Y:=\{nN^{-1/3}: n\in[N]\}$ and $S_Z:=\{nN^{-2/3}: n\in[N]\}$, which are the state space of Y^N and Z^N respectively. For any $\epsilon\in(0,t)$ and $a\in(0,u_N)\cap S_Y$, we have

$$\mathbb{P}^N_{u_N,v_N}\left[T^N<t\right]\ge\mathbb{P}^N_{u_N,v_N}\left[\tau_a((Y^N,Z^N))<t-\epsilon\right]\inf_{b\in S_Z}\mathbb{P}^N_{a,b}\left[T^N<\epsilon\right].$$

For each given initial state, $\tau_y((Y^N,Z^N))(\omega)$ is a non-increasing function of sufficiently small y.

If for each $\epsilon>0$,we can find $a=a_\epsilon>0$ such that $\mathbb{P}^N_{a,b}\left[T^N<\epsilon\right]>1-\epsilon$ for all $b>0$, then

$$\begin{aligned}\liminf_{N\to\infty}\mathbb{P}^N_{u_N,v_N}\left[T^N<t\right]&\ge\liminf_{N\to\infty}\mathbb{P}^N_{u_N,v_N}\left[\tau_a((Y^N,Z^N))<t-\epsilon\right](1-\epsilon)\\&=\mathbb{P}_{u,v}\left[\tau_a((Y,Z))<t-\epsilon\right](1-\epsilon)\ge\mathbb{P}_{u,v}\left[T<t-\epsilon\right](1-\epsilon),\end{aligned}$$

where the equality follows from (5.12).

Since $\mathbb{P}_{u,v}\left[T<t-\epsilon\right](1-\epsilon)$ monotonically increases as $\epsilon\downarrow 0$, we have

$$\liminf_{N\to\infty}\mathbb{P}^N_{u_N,v_N}\left[T^N<t\right]\ge\mathbb{P}_{u,v}\left[T<t\right]. \tag{5.17}$$

In fact, it suffices to take any $a=a_\epsilon$ satisfying

$$a_\epsilon<\begin{cases}-\log(1-\epsilon)\frac{|e^{\hat\lambda\epsilon}-1|}{|\hat\lambda|}\wedge\left|4\hat\lambda\right|^{-1}, & \hat\lambda\ne 0,\\ -\epsilon\log(1-\epsilon), & \hat\lambda=0.\end{cases}$$

To see this, assuming we take such a_ϵ, we first look at the case when $\hat\lambda\ne 0$. Recall that

$$\mathbb{P}^N_{a,b}\left[T^N<\epsilon\right]=\mathbb{P}^N_{a,b}\left[T^N_o<N^{1/3}\epsilon\right],$$

where T^N_o denotes the extinction time of I^N with initial state $(I^N_0,R^N_0)=(N^{1/3}a,N^{2/3}b)$.

It is obvious that I^N is stochastically dominated by a linear birth-death chain L^N with birth rate $\lambda_o = \lambda_o(N)$, death rate 1 and initial state $L_0^N = N^{1/3}a$, satisfying $L_0^N(1-\lambda_o) \to a\hat{\lambda}$. Stochastic dominance between the processes implies that T_o^N is stochastically dominated by the extinction time of L^N. We have the exact expression (4.3) of the distribution of the latter when $\hat{\lambda} \neq 0$. Notice that this bound is independent of the status of R^N.

For any

$$0 < a_\epsilon \leq -\frac{\left|e^{\hat{\lambda}\epsilon} - 1\right|}{|\hat{\lambda}|} \log(1-\epsilon),$$

the following statement is true for all sufficiently large N:

$$a_\epsilon \leq -\frac{\left|e^{\hat{\lambda}\epsilon} - 1\right|}{|\hat{\lambda}|} \log(1-\epsilon) \leq -\log(1-\epsilon) N^{-1/3} \log^{-1}\left(1 + \frac{|\hat{\lambda}|N^{-1/3}}{\left|e^{\hat{\lambda}\epsilon} - 1\right|}\right),$$

since the last term above is a monotonically decreasing function of N, tending to the limit $-\log(1-\epsilon)\frac{|e^{\hat{\lambda}\epsilon}-1|}{|\hat{\lambda}|}$. It follows that

$$\mathbb{P}_{a,b}^N\left[T^N < \epsilon\right] \geq \mathbb{P}^N\left[T_{bdp}^N < N^{1/3}\epsilon \middle| L_0^N = N^{1/3}a\right] = \left(1 + \frac{\hat{\lambda}N^{-1/3}}{e^{\epsilon\hat{\lambda}} - 1}\right)^{-aN^{1/3}} \geq 1 - \epsilon,$$

The case of $\hat{\lambda} = 0$ can be treated in the same way. For any

$$0 < a_\epsilon < -\epsilon \log(1-\epsilon),$$

the following statement is true for all sufficiently large N:

$$\begin{aligned} a_\epsilon &\leq -\log(1-\epsilon) N^{-1/3} \log^{-1}\left(1 + \frac{\hat{\lambda}N^{-1/3}}{\exp\left\{\hat{\lambda}\epsilon\right\} - 1}\right) \\ &= -\log(1-\epsilon) N^{-1/3} \log^{-1}\left(1 + N^{-1/3}\epsilon^{-1}\right) \to -\epsilon\log(1-\epsilon). \end{aligned}$$

The exact expression of the extinction time of L^N with equal birth and death rates can be derived, e.g., taking the limit in (4.3) as $\lambda_N - \mu_N \to 0$. Similar to the above,

$$\mathbb{P}^N_{a,b}\left[T^N < \epsilon\right] \geq \mathbb{P}^N\left[T^N_{bdp} < N^{1/3}\epsilon \middle| L^N_0 = N^{1/3}a_\epsilon\right] = \left(\frac{N^{1/3}\epsilon}{N^{1/3}\epsilon + 1}\right)^{a_\epsilon N^{1/3}} \geq 1 - \epsilon.$$

With both (5.13) and (5.14), we conclude $T^N \Rightarrow T$ when initial state $(u_N, v_N) \to (u, v)$ and thus prove the theorem. □

5.5 Distribution of the extinction time T of the limit diffusion

First, we transform the problem of obtaining the distribution of the hitting time to solving a second-order PDE.

Define $U(u,v,t) := \mathbb{P}\left[Y_{t_0} > 0 \middle| Y_t = u, Z_t = v\right] = \mathbb{P}\left[T > t_0 - t \middle| (Y_0, Z_0) = (u,v)\right]$ on the domain $\mathbb{R}^2_+ \times [0, t_0]$.

By the Feynman-Kac formula, $U(u,v,s)$ is the solution to the following PDE:

$$\frac{\partial U}{\partial t} = -u\frac{\partial^2 U}{\partial u^2} + (\hat{\lambda} + v)u\frac{\partial U}{\partial u} - (u - \gamma v)\frac{\partial U}{\partial v}, \tag{5.18}$$

with the end condition $U(u,v,t_0) = \mathbf{1}_{\mathbb{R}^2_+}(u,v)$, and the boundary condition

$$\lim_{u \downarrow 0} U(u,v,t) = 0,\ t \in [0, t_0),$$

where $\mathbf{1}_A$ denotes the indicator function of set A.

Consider the Banach space of bounded continuous functions which have continuous extensions to $[0,\infty)^2$, equipped with the uniform norm, denoted as $(\widehat{BC}(\mathbb{R}^2_+), \|\cdot\|)$.

Theorem 5.9. *Let $V(t)$ be a bounded linear operator on $\widehat{BC}(\mathbb{R}^2_+)$ for each $t > 0$, such that*

$$V(t)f(u,v) := \int_0^\infty g(t, ue^{-(\hat{\lambda}+v)t}; m) f(m, ve^{-\gamma t} + ut)\, dm,$$

for $t \geq 0$ *and* $f \in \widehat{BC}(\mathbb{R}_+^2)$*, where*

$$g(t,u;m) = \frac{1}{t} u^{1/2} m^{-1/2} e^{-(u+m)/t} I_1\left(\frac{2m^{1/2}u^{1/2}}{t}\right), m, u, t > 0,$$

and $I_1(\cdot)$ *is defined as* (5.7)*. Define*

$$U_n(u,v,t) := \left(V\left(\frac{t}{n}\right)\right)^n \mathbf{1}_{\mathbb{R}_+^2}(u,v).$$

The tail distribution of T*, i.e.* $\mathbb{P}\left[T > t \middle| (Y_0, Z_0) = (u,v)\right]$*, for each* $t > 0$*, is the limit of* $U_n(u,v,t)$ *as* $n \to \infty$*, for* $(u,v) \in \mathbb{R}_+^2$ *uniformly on compacts.*

Define linear operators $\mathcal{L}, \mathcal{H}$ on $f \in C^2(\mathbb{R}_+^2)$, the space of twice differentiable functions on $\mathbb{R}_+^2$, as

$$\mathcal{L} := u\frac{\partial^2}{\partial u^2} \text{ and } \mathcal{H} := -(\hat{\lambda} + v)u\frac{\partial}{\partial u} + (u - \gamma v)\frac{\partial}{\partial v}.$$

Recalling the definition of the squared Bessel process in Definition 5.5, we see that $2\mathcal{L}$ is the infinitesimal generator of $BESQ^0(u)$. This motivates us to construct a solution analogous to the Lie-Trotter product.

The Lie-Trotter product formula [44] is an extension to generators of strongly continuous semigroups on Banach spaces of the following result for $n \times n$ matrices A and B:

$$e^{t(A+B)} = \lim_{n\to\infty} \left(e^{tA/n} e^{tB/n}\right)^n.$$

The Lie-Trotter product formula can be seen as a consequence of the Chernoff product formula [45, 46] below.

Theorem 5.10. *Let* $(F(t))_{t\geq 0}$ *be a family of bounded linear operators on a Banach space* X*. Assume that*

1. $F(0) = \mathcal{I}$*,*
2. $\left\|F^k(t)\right\| \leq Me^{wkt}$ *for some* $M \geq 1$*, some* $w > 0$*, all* $k \in \mathbb{N}$ *and all* $t \geq 0$*,*
3. *the limit* $\mathcal{A}f := \lim_{t\downarrow 0} \frac{F(t)-\mathcal{I}}{t} f$ *for all* $f \in D$*, where* D *and* $(\alpha - \mathcal{A})D$ *are dense subspaces in* X *for some* $\alpha > w$*.*

Then the closure $(\mathcal{A}, Dom(\mathcal{A}))$ of $(\mathcal{A}, D)$ w.r.t. graph norm generates a strongly continuous semigroup $(\mathcal{R}_t)_{t\geq 0}$ given by

$$\mathcal{R}_t f = \lim_{n\to\infty} (F(t/n))^n f.$$

The problem is, the generator we are interested in does not generate a strongly continuous semigroup on $\widehat{BC}(\mathbb{R}_+^2)$. In order to prove Theorem 5.9, we adopt the theory of bi-continuous semigroups established in [7]. It allows us to obtain the Chernoff product formula w.r.t. an appropriately chosen topology. The basic definitions and important theorems are introduced in Section 3.3.

5.5.1 Chernoff approximation of extinction times in the critical regime

In this section, we apply Theorem 3.26 to our problem.

The underlying topological space

Let $\Phi = \widehat{BC}(\mathbb{R}_+^2)$ and endow $\widehat{BC}(\mathbb{R}_+^2)$ with the uniform on compacts topology (ucc), which is induced by the set of seminorms $\{\|\cdot\|_K\}_{K\subset\mathbb{R}_+^2}$, defined as

$$\|f\|_K := \sup_{(u,v)\in K} |f(u,v)|,$$

for every compact set $K \subset \mathbb{R}_+^2$.

Following standard theorems, we can verify that $(\widehat{BC}(\mathbb{R}_+^2), \text{ucc})$ satisfies Assumption 3.18. The details of verification is included below for completeness.

Item 2 of Assumption 3.18 follows from the fact that the Banach space $(\widehat{BC}(\mathbb{R}_+^2), \|\cdot\|)$ is continuously embedded in $(\widehat{BC}(\mathbb{R}_+^2), \text{ucc})$ (denoted in symbol as$(\widehat{BC}(\mathbb{R}_+^2), \|\cdot\|) \hookrightarrow (\widehat{BC}(\mathbb{R}_+^2), \text{ucc})$). We can see this because for each $K \subset \mathbb{R}_+^2$,

$$\|f\|_K \leq \|f\|.$$

Item 1 of Assumption 3.18 is satisfied due to the following lemma.

Lemma 5.11. *The space* $(\widehat{BC}(\mathbb{R}^2_+), \mathrm{ucc})$ *is sequentially complete on* $\|\cdot\|$*-bounded sets.*

Proof. Any $\|\cdot\|$-bounded ucc-Cauchy sequence $\{f_n\}_{n\in\mathbb{N}} \subset \widehat{BC}(\mathbb{R}^2_+)$ has a limit coinciding with the pointwise limit

$$f(u,v) := \lim_{n\to\infty} f_n(u,v), \quad (u,v) \in \mathbb{R}^2_+.$$

It is easy to check that such f belongs to $\widehat{BC}(\mathbb{R}^2_+)$. □

Item 3 is true since the topological dual of $(\widehat{BC}(\mathbb{R}^2_+), \mathrm{ucc})$ is the set of Radon measures with compact support, which contains the Dirac measures. Hence $(\widehat{BC}(\mathbb{R}^2_+), \mathrm{ucc})'$ is norming for $(\widehat{BC}(\mathbb{R}^2_+), \|\cdot\|)$.

Lemma 5.12. *The space* $C_c(\mathbb{R}^2_+)$ *is bi-dense in* $\widehat{BC}(\mathbb{R}^2_+)$.

Proof. For each $f \in \widehat{BC}(\mathbb{R}^2_+)$, we can take an arbitrary exhaustion of $\mathbb{R}^2_+$ denoted as $\{D_n\}_{n\in\mathbb{N}}$ and find a sequence of functions $f_n(u,v) := f(u,v)\iota_n(u,v) \in C_c(\mathbb{R}^2_+), n \in \mathbb{N}$, where ι_n is defined as (5.5). The sequence $\{f_n\}_{n\in\mathbb{N}}$ is $\|\cdot\|$-bounded by $\|f\|$ and converges uniformly on compacts to f. □

The second-order problem associated with $\mathcal{L} = u\frac{\partial^2}{\partial u^2}$

We will now examine the operator $\mathcal{L} = u\frac{\partial^2}{\partial u^2}$. Luckily, we are able to describe the semigroup generated by $\mathcal{L}$ precisely.

Definition 5.13 (Green's function, p.82, [47])**.** The solution $y(x,t) = G(t,x;x_0)$ of the following PDE

$$\frac{\partial y}{\partial t} = \mathcal{A}y$$

on domain $D \times [0,T]$, with initial condition $\lim_{t\downarrow 0} y(x,t) = \delta(x-x_0)$, $x \in D$ and boundary condition $y(x,t) = 0$, $x \in \partial D \times (0,T]$ is called the *Green's function* of the operator $\partial_t - \mathcal{A}$.

Consider the process $(X_t)_{t\geq 0}$ such that $2X_t$ is a squared Bessel process with $2X_0 = 2u > 0$ below.

Lemma 5.14. *The Green's function of the parabolic operator $\partial_t + \mathcal{L}$ can be written as*

$$g(t_0-t,u;u_0) = \frac{1}{t_0-t}u^{1/2}u_0^{-1/2}e^{-(u+u_0)/(t_0-t)}I_1\left(\frac{2u_0^{1/2}u^{1/2}}{t_0-t}\right), (u,u_0)\in\mathbb{R}_+^2, t\in[0,t_0),$$

where $I_1(\cdot)$ is the modified Bessel function of the first kind of index 1.

Proof. The Green's function $g(t_0-t,u;u_0)$ of $\partial_t+\mathcal{L}$ should satisfy

$$\begin{aligned} &(\partial_t+\mathcal{L})g(t_0-t,u;u_0) = 0,\ t\in[0,t_0), \\ &\lim_{t\uparrow t_0} g(t_0-t,u;u_0) = \delta(u-u_0), \\ &g(t_0-t,0;u_0) = 0,\ t\in[0,t_0). \end{aligned} \tag{5.19}$$

To solve (5.19), we notice that it is the Kolmogorov backward equation of a process X_t where $2X_t \sim BESQ^0(2u)$, and has the solution

$$g(t_0-t,u;u_0) = 2q_t^{BESQ}(2u,2u_0) = \frac{1}{t_0-t}u^{1/2}u_0^{-1/2}e^{-(u+u_0)/(t_0-t)}I_1\left(\frac{2u_0^{1/2}u^{1/2}}{t_0-t}\right).$$

It is easy to verify that $g(t_0-t,u;u_0)$ satisfies the initial and boundary conditions. □

Define for each $t\geq 0$, a family of operators $\mathcal{S}_t \in \mathscr{L}(\widehat{BC}(\mathbb{R}_+^2))$ as

$$\mathcal{S}_t f(u,v) := \begin{cases} \int_0^\infty g(t,u;m)f(m,v)dm, & t>0, \\ f(u,v), & t=0, \end{cases} \tag{5.20}$$

where g is defined in Lemma 5.14. We should also note that the definition of $\mathcal{S}_t$ can be extended to bounded measurable functions on $\mathbb{R}_+^2$.

From the probability interpretation used in the proof of Lemma 5.14, it is easy to check that $(\mathcal{S}_t)_{t\geq 0}$ relates to the squared Bessel process X_t defined above in the following sense: for $f\in\widehat{BC}(\mathbb{R}_+^2)$, denote $\tau^X := \inf\{t : X_t = 0\}$, then

$$\mathcal{S}_t f(u,v) = \mathbb{E}_u\left[f(X_t,v)\right] - f(0,v)\mathbb{P}_u\left[\tau^X \leq t\right] = \mathbb{E}_u\left[f(X_t,v), t<\tau^X\right].$$

Proposition 5.15. *The family of operators $(\mathcal{S}_t)_{t\geq 0}$ is a bi-continuous contraction semigroup on $(\widehat{BC}(\mathbb{R}_+^2), \text{ucc})$. The generator of $(\mathcal{S}_t)_{t\geq 0}$ restricted to $C_c^\infty(\mathbb{R}_+^2)$ coincides with $\mathcal{L}$.*

Proof. It is straightforward to see that $\mathcal{S}_0 = I$, $\mathcal{S}_s\mathcal{S}_t = \mathcal{S}_{s+t}$, and $\|\mathcal{S}_t f\| \leq \|f\|$.

Next we check that $(\mathcal{S}_t)_{t\geq 0}$ is locally bi-equicontinuous. Let $K \subset \mathbb{R}_+^2$ be compact, $t_0 > 0$ and $\epsilon > 0$. By (A.23)–(A.26), we can find a compact interval $K_\epsilon^1 \subset \mathbb{R}_+$ such that, uniformly for $t \in (0, t_0]$,

$$\int_{\mathbb{R}_+\setminus K_\epsilon^1} g(t, u; m)dm < \frac{\epsilon}{2\sup_{n\in\mathbb{N}} \|f_n\|}$$

for all $(u, v) \in K$.

Denote $K_\epsilon := (K_\epsilon^1 \times \mathbb{R}_+) \cap K \subset \mathbb{R}_+^2$.

Given ϵ and K_ϵ as above, for any $\|\cdot\|$-bounded sequence $\{f_n\}_{n\in\mathbb{N}}$ ucc-converging to 0, we can find some $n_0 \in \mathbb{N}$ such that $\|f_n\|_{K_\epsilon} < \epsilon/2$ for all $n > n_0$.

$$\begin{aligned}\|\mathcal{S}_t f_n\|_K &\leq \sup_{(u,v)\in K} \int_{K_\epsilon^1} g(t, u; m)|f_n(m, v)|dm + \int_{\mathbb{R}_+\setminus K_\epsilon^1} g(t, u; m)|f_n(m, v)|dm \\ &< \epsilon.\end{aligned}$$

Hence $(\mathcal{S}_t)_{t\geq 0}$ is locally bi-equicontinuous w.r.t. ucc-topology.

Thirdly we check that $(\mathcal{S}_t)_{t\geq 0}$ is bi-continuous. We will prove this using the property of the Green's function g.

Let $f \in \widehat{BC}(\mathbb{R}_+^2)$, $K \subset \mathbb{R}_+^2$ be compact, and $\epsilon > 0$. By local bi-equicontinuity, there exists $f_0 \in C_0(\mathbb{R}_+^2)$ satisfying $\|f_0 - f\|_K < \epsilon$ and $\|\mathcal{S}_t(f_0 - f)\|_K < \epsilon$.

Consider $\hat{f} : [0.1]^2 \to \mathbb{R}$ such that $\hat{f}(e^{-u}, e^{-v}) = f_0(u, v)$ for all $(u, v) \in \mathbb{R}_+^2$. It is known that as a continuous function on compact domain, $\hat{f}$ can be uniformly approximated by a sequence of Berstein Polynomials (see e.g. [48]). In other words, there exists positive integer n_0 and a sequence of polynomials of degree n w.r.t. both x and y, denoted as $\hat{h}_n(x, y), n \in \mathbb{N}$, such that for all $n > n_0$,

$$\sup_{(u,v)\in\mathbb{R}_+^2} \left|\hat{f}(e^{-u}, e^{-v}) - \hat{h}_n(e^{-u}, e^{-v})\right| = \sup_{(x,y)\in[0,1]^2} \left|\hat{f}(x, y) - \hat{h}_n(x, y)\right| < \epsilon.$$

We have

$$\|\mathcal{S}_t f - f\|_K \le \|\mathcal{S}_t(f_0 - f)\|_K + \|\mathcal{S}_t f_0 - f_0\|_K + \|f - f_0\|_K.$$

For all $(u, v) \in K$,

$$\begin{aligned}
&|\mathcal{S}_t f_0(u,v) - f_0(u,v)| \\
\le &\int_{\mathbb{R}_+} g(t,u;m)\Big|f_0(m,v) - \hat{h}_n(e^{-m}, e^{-v})\Big|dm \\
&+ \Big|f_0(u,v) - \hat{h}_n(e^{-u}, e^{-v})\Big| + \Big|\int_{\mathbb{R}_+} g(t,u;m)\hat{h}_n(e^{-m}, e^{-v})dm - \hat{h}_n(e^{-u}, e^{-v})\Big| \\
\le &2\epsilon + \Big|\int_{\mathbb{R}_+} g(t,u;m)\hat{h}_n(e^{-m}, e^{-v})dm - \hat{h}_n(e^{-u}, e^{-v})\Big|. \qquad (5.21)
\end{aligned}$$

Each $\hat{h}_n$ is a linear combination of $\{e^{-ju-kv},\ j,k = 1,2,\cdots,n\}$. Notice that by using Berstein Polynomials, we have made sure that $j,k \neq 0$. This is because f_0 vanishes at infinity, and hence $\hat{f}$ vanishes when one or both components are 0.

By Lemma A.6, for each $(u, v) \in K$,

$$\begin{aligned}
&\Big|\int_{\mathbb{R}_+} g(t,u;m)\hat{h}_n(e^{-m}, e^{-v})dm - \hat{h}_n(e^{-u}, e^{-v})\Big| \\
\le &\sum_{1\le j,k\le n} |a_{jk}|e^{-kv}\Big|\int_{\mathbb{R}_+} g(t,u;m)e^{-jm}\,dm - e^{-ju}\Big| \\
= &\sum_{1\le j,k\le n} |a_{jk}|e^{-kv}\Big|\exp\Big\{-\frac{u}{t+j^{-1}}\Big\} - \exp\Big\{-\frac{u}{t}\Big\} - \exp\{-uj\}\Big| \to 0 \text{ as } t \downarrow 0.
\end{aligned}$$

Together with (5.21) and our definition of f_0, we can conclude that there exists $t_\epsilon > 0$ such that

$$\|\mathcal{S}_t f - f\|_K \le 5\epsilon.$$

For each $v \in \mathbb{R}_+$, let $f_v(u) := f(u,v) \in C_0(\mathbb{R}_+)$, then $\mathcal{S}_t f_v(u) = \mathbb{E}_u\left[f(X_t, v)\right]$ where $2X_t$ is a 0-dimensional squared Bessel process. The generator of the semigroup associated with a one-dimensional diffusion can be fully characterised (see Chapter 8.1, [16]), from which we have $(\mathcal{S}_t)_{t\ge0}$ is strong continuous on $C_0(\mathbb{R}_+)$ with the domain $\{f \in C_0(\mathbb{R}_+) : \mathcal{L}f \in C_0(\mathbb{R}_+)\} \supset C_c^\infty(\mathbb{R}_+)$. This can also be proved by approximation by linear combinations of $\{e^{-ju-kv}, j,k \in \mathbb{N}\}$ similar to our argument above. By the dominated convergence theorem, $\mathcal{S}_t$ preserves the limit at $v \downarrow 0$ and $v \to \infty$. Hence $(\mathcal{S}_t)_{t\geq 0}$ is strongly continuous on $C_0(\mathbb{R}_+^2)$. Since $\mathcal{L}$ does not act on component v, we also have the corresponding domain contains $C_c^\infty(\mathbb{R}_+^2)$.

This concludes the proof. □

We denote by $C^{i,j}(\mathbb{R}_+^2)$ the subspace of $\widehat{BC}(\mathbb{R}_+^2)$ with continuous partial derivatives up to order i w.r.t. the first component and up to order j w.r.t. the second component. We denote by $\widehat{BC}^{i,j}(\mathbb{R}_+^2)$ the subspace of $C^{i,j}(\mathbb{R}_+^2)$ whose partial derivatives above are also bounded and has a continuous extension to $[0,\infty)^2$.

The first-order problem associated with $\mathcal{H} = -(\hat{\lambda}+v)u\frac{\partial}{\partial u} + (u-\gamma v)\frac{\partial}{\partial v}$

We now shift our attention to the first order operator $\mathcal{H}$ and the semigroup $(\mathcal{T}_t)_{t\geq 0}$ it generates. Consider the solution $F = F(u,v,t)$ to the first-order PDE

$$\frac{\partial F}{\partial t} = -\mathcal{H}F, \tag{5.22}$$

$$F(u,v,0) = f(u,v) \in C^{1,1}(\mathbb{R}_+^2), \tag{5.23}$$

$$\lim_{u\downarrow 0} F(u,v,t) = 0,\ t>0.$$

Denote characteristics parametrised by $s \geq 0$ as $(t(s), u(c_1,c_2,s), v(c_1,c_2,s))$, where

$$\begin{aligned} \frac{dt}{ds} &= 1, \\ \frac{du}{ds} &= -(\hat{\lambda}+v)u, \\ \frac{dv}{ds} &= u - \gamma v, \end{aligned} \tag{5.24}$$

and $u(c_1,c_2,0) = c_1$, $v(c_1,c_2,0) = c_2$, $t(0) = 0$. By the Inverse function theorem, calculating the Jacobian determinant at any $(c_1, c_2, 0)$, we find that

$$\det\begin{pmatrix} \frac{\partial u}{\partial c_1}, & \frac{\partial u}{\partial c_2}, & \frac{\partial u}{\partial s} \\ \frac{\partial v}{\partial c_1}, & \frac{\partial v}{\partial c_2}, & \frac{\partial v}{\partial s} \\ 0, & 0, & \frac{\partial t}{\partial s} \end{pmatrix} \equiv 1,$$

and therefore we can represent c_1, c_2, s as functions of u, v, t for all u, v and small t. We can construct a solution to (5.22) as

$$F(u, v, t) = f(c_1(u, v, t), c_2(u, v, t)).$$

Alternatively, we can denote $x_1(u_0, v_0, t), x_2(u_0, v_0, t)$ as the solution to the characteristic equations (5.24) given initial state (u_0, v_0). Then we can define a shift operator $\mathcal{T}_t \in \mathscr{L}(\widehat{BC}(\mathbb{R}_+^2))$ such that

$$\mathcal{T}_t f(u, v) = f(x_1(u, v, t), x_2(u, v, t)),\ f \in \widehat{BC}(\mathbb{R}_+^2).$$

The fixed points of the system (5.24) are $(0, 0)$ and $(-\gamma\hat{\lambda}, -\hat{\lambda})$. The Jacobian matrix at (u, v) is

$$J(u, v) = \begin{pmatrix} -\left(\hat{\lambda} + v\right) & -u \\ 1 & -\gamma \end{pmatrix}.$$

The eigenvalues of $J(0, 0)$ are $-\hat{\lambda}$ and $-\gamma$, whereas the eigenvalues of $J(-\gamma\hat{\lambda}, -\hat{\lambda})$ are $\frac{1}{2}\left(-\gamma \pm \sqrt{\gamma^2 + 4\hat{\lambda}\gamma}\right)$. It follows that the region $[0, \infty)^2$ is an invariant set, and regardless of the value of $(\hat{\lambda}, \gamma)$, the characteristics will not hit $u = 0$ in finite time. This ensures that the solutions with our boundary conditions exists.

With simple manipulation of the characteristic equations, we have

$$u(t) = u(0)e^{-\hat{\lambda}t}\exp\left\{-\int_0^t v(s)ds\right\} \leq u(0)e^{|\hat{\lambda}|t} \tag{5.25}$$

and

$$v(t) = e^{-\gamma t}\int_0^t e^{\gamma s}u(s)ds + e^{-\gamma t}v(0).$$

Lemma 5.16. *The family of operators $(\mathcal{T}_t)_{t\geq 0}$ is a bi-continuous semigroup w.r.t.* ucc-*topology.*

Proof. Firstly, it is easy to see that $\mathcal{T}_{t+s} = \mathcal{T}_t\mathcal{T}_s$, $\mathcal{T}_0 = \mathcal{I}$ and $\|\mathcal{T}_t\| = 1$.

Next, we prove the local bi-equicontinuity.

Taking any constant $t_0 > 0$, any compact set $K \subset \mathbb{R}_+^2$, we can always find $M > 0$ such that $K \subset [0, \gamma M] \times [0, M]$. It is easy to check that for any $x > 0$, the characteristic curve does not cross the boundary $[0, \gamma x] \times \{x\}$ in the direction toward $+\infty$. Recall that by (5.25), the solution to the characteristic equations has the following property:

Since $u(t) \le u(0)e^{|\hat{\lambda}|t_0}$, for $t \in [0, t_0]$, we can define compact set

$$K_0 = [0, \gamma M e^{|\hat{\lambda}|t_0}] \times [0, M e^{|\hat{\lambda}|t_0}],$$

and the characteristics initiated in K will remain in K_0 up to time t_0. It follows that

$$\|\mathcal{T}_t f\|_K \le \|f\|_{K_0}, \quad t \in [0, t_0],\ f \in \widehat{BC}(\mathbb{R}_+^2). \tag{5.26}$$

Thirdly, we prove the bi-continuity property of the semigroup $(\mathcal{T}_t)_{t\ge 0}$.
For every $f \in C_c^\infty(\mathbb{R}_+^2)$, $t \in [0, t_0]$,

$$\begin{aligned}
&\|(\mathcal{T}_t - \mathcal{I})f\| = \sup_{(u,v)\in\mathbb{R}_+^2} \Big| f(x_1(u,v,t), x_2(u,v,t)) - f(u,v) \Big| \\
&= \sup_{(u,v)\in\mathbb{R}_+^2} \Big| f^{(1,0)}(u,v)(x_1 - u) + f^{(0,1)}(u,v)(x_2 - v) \\
&\quad + \frac{1}{2}\big(f^{(2,0)}(u_1,v_1)(x_1-u)^2 + f^{(0,2)}(u_1,v_1)(x_2-v)^2 + 2f^{(1,1)}(u_1,v_1)(x_1-u)(x_2-v)\big) \Big| \\
&= \sup_{(u,v)\in\mathbb{R}_+^2} t\Big| -f^{(1,0)}(u,v)(\hat{\lambda}+v_2)u_2 + f^{(0,1)}(u,v)(u_3 - \gamma v_3) \\
&\quad + \frac{1}{2}t\left(f^{(2,0)}(u_1,v_1)\left(-(\hat{\lambda}+v_2)u_2\right)^2 + f^{(0,2)}(u_1,v_1)\left(u_3-\gamma v_3\right)^2 \right. \\
&\quad \left. -2f^{(1,1)}(u_1,v_1)(\hat{\lambda}+v_2)u_2(u_3-\gamma v_3)\right)\Big|,
\end{aligned} \tag{5.27}$$

where $(u_i, v_i) = a_i(u,v) + (1-a_i)(x_1(u,v,h), x_2(u,v,h))$ for some $a_i \in (0,1), i = 1,2,3$, are taken according to the Lagrange's form of Taylor expansion: the second order remainder of the expansion of f, and the first order remainders of the expansion of x_1 and x_2 respectively.

Since $f \in C_c^\infty(\mathbb{R}_+^2)$, the supremum in the last line of (5.27) is finite, and $((\mathcal{T}_t - \mathcal{I})f)_{t\geq 0}$ is uniformly $\|\cdot\|$-bounded by $2\|f\|$. Hence

$$\text{ucc} - \lim_{t\downarrow 0}(\mathcal{T}_t - \mathcal{I})f = 0. \tag{5.28}$$

Recall that $C_c^\infty(\mathbb{R}_+^2)$ is $\|\cdot\|$-dense in $C_c(\mathbb{R}_+^2)$ and thus is bi-dense in $\widehat{BC}(\mathbb{R}_+^2)$ by Lemma 5.12. For any $f \in \widehat{BC}(\mathbb{R}_+^2)$ and any $K \subset \mathbb{R}_+^2$, the local bi-equicontinuity ensures that we can find for every $\epsilon > 0$ an $f_c \in C_c^\infty(\mathbb{R}_+^2)$ such that $\|f - f_c\|_K \vee \|\mathcal{T}_t(f - f_c)\|_K < \epsilon$. By (5.28), we can also find $t_\epsilon > 0$ such that for all $t \in [0, t_\epsilon]$,

$$\|(\mathcal{T}_t - \mathcal{I})f_c\|_K < \epsilon.$$

It follows that for $t \leq t_\epsilon$

$$\|(\mathcal{T}_t - \mathcal{I})f\|_K \leq \|\mathcal{T}_t(f - f_c)\|_K + \|(\mathcal{T}_t - \mathcal{I})f_c\|_K + \|f - f_c\|_K < 3\epsilon,$$

and hence the bi-continuity of $(\mathcal{T}_t)_{t\geq 0}$ on $\widehat{BC}(\mathbb{R}_+^2)$ is proved. □

It is not possible to solve the characteristic equations and derive an explicit expression of $\mathcal{T}_t$. Since the Chernoff product formula only uses the property of $\mathcal{T}_t$ when t is small, we will work with an approximation of $(\mathcal{T}_t)_{t\geq 0}$ instead.

Define the mapping $\xi : (t, u, v) \mapsto (\xi_t^1(u, v), \xi_t^2(u, v))$ from $[0, \infty) \times \mathbb{R}_+^2 \to \mathbb{R}_+^2$, where

$$\xi_t^1(u, v) := ue^{-(\hat{\lambda}+v)t}, \quad \xi_t^2(u, v) := ve^{-\gamma t} + ut.$$

Define operator $W : [0, \infty) \to \mathscr{L}(\widehat{BC}(\mathbb{R}_+^2))$ such that

$$W(t)f(u, v) := f(\xi_t^1(u, v), \xi_t^2(u, v)) = f(ue^{-(\hat{\lambda}+v)t}, ve^{-\gamma t} + ut). \tag{5.29}$$

Lemma 5.17. *The family of operators $(W(t))_{t\geq 0}$ is locally bi-equicontinuous and bi-continuous w.r.t.* ucc-*topology.*

Proof. First we prove the local bi-equicontinuity. Let $t_0 > 0$ and $K \subset \mathbb{R}_+^2$ be any compact set. Define $\overline{u} = \max\{u : (u, v) \in K\}$ and $\underline{u} = \min\{u : (u, v) \in K\}$.

Similarly, we can define $\overline{v}, \underline{v}$. Then we have for all $t \in [0, t_0]$, $(u, v) \in K$,

$$\xi_t^2(u,v) \in \left[\underline{v}e^{-\gamma t_0},\, \overline{u}t_0 + \overline{v}\right],$$

$$\xi_t^1(u,v) \in \left[\underline{u}e^{-(|\hat{\lambda}|+\overline{u}t_0+\overline{v})t_0},\, \overline{u}e^{|\hat{\lambda}|t_0}\right].$$

We can always find compact set $K_0 \subset \mathbb{R}_+^2$ such that

$$K_0 \supset \left[\underline{u}e^{-(|\hat{\lambda}|+\overline{u}t_0+\overline{v})t_0},\, \overline{u}e^{|\hat{\lambda}|t_0}\right] \times \left[\underline{v}e^{-\gamma t_0},\, \overline{u}t_0 + \overline{v}\right].$$

Then for all $t \in [0, t_0]$,

$$\|W(t)f\|_K \le \|f\|_{K_0},$$

which implies local bi-equicontinuity.

For bi-continuity, we notice that for any compact set $K \subset \mathbb{R}_+^2$, each $f \in \widehat{BC}(\mathbb{R}_+^2)$ is uniformly continuous on K, i.e., for each $\epsilon > 0$, there exists $\delta > 0$ depending only on ϵ, such that $\sup_{|u-u'|\vee|v-v'|<\delta} |f(u,v) - f(u',v')| < \epsilon$. For each K, there also exists $t_\delta > 0$ such that

$$\sup_{t\in[0,t_\delta]} \sup_{(u,v)\in K} |ue^{-(\hat{\lambda}+v)t} - u| \vee |ve^{-\gamma t} + ut - v| < \delta,$$

from which we conclude $\mathrm{ucc} - \lim_{t\downarrow 0}(W(t) - \mathcal{I})f = 0$. The proof of bi-continuity is then concluded since $\sup_{t\ge 0} \|(W(t) - \mathcal{I})f\| \le 2\|f\|$. □

Lemma 5.18. *For all $f \in C_c^\infty(\mathbb{R}_+^2)$,*

$$\mathrm{ucc} - \lim_{t\downarrow 0} \frac{W(t) - \mathcal{I}}{t} f = \mathcal{H}f.$$

Proof. For $f \in C_c^\infty(\mathbb{R}_+^2)$, $t > 0$, and $(u,v) \in \mathbb{R}_+^2$,

$$\begin{aligned}
&\frac{W(t)-\mathcal{I}}{t} f(u,v) - \mathcal{H}f(u,v) \\
=& f^{(1,0)}(u,v)\frac{ue^{-(\hat{\lambda}+v)t}-u}{t} + f^{(0,1)}(u,v)\frac{ve^{-\gamma t}+ut-v}{t} \\
&+ \frac{1}{2}f^{(2,0)}(u_1,v_1)\frac{(ue^{-(\hat{\lambda}+v)t}-u)^2}{t} + f^{(1,1)}(u_1,v_1)\frac{(ue^{-(\hat{\lambda}+v)t}-u)(ve^{-\gamma t}+ut-v)}{t} \\
&+ \frac{1}{2}f^{(0,2)}(u_1,v_1)\frac{(ve^{-\gamma t}+ut-v)^2}{t} - \mathcal{H}f \\
=& f^{(1,0)}(u,v)\frac{1}{2}u(\hat{\lambda}+v_2)^2 t + f^{(0,1)}(u,v)\frac{1}{2}v\gamma_2^2 t + \frac{1}{2}f^{(2,0)}(u_1,v_1)u^2(\hat{\lambda}+v_3)^2 t \\
&+ f^{(1,1)}(u_1,v_1)u(\hat{\lambda}+v_3)(-\gamma_3+u)t + \frac{1}{2}f^{(0,2)}(u_1,v_1)(-\gamma_3+u)^2 t,
\end{aligned}$$

where we use Lagrange's form of the remainder to choose all the variables (u_1, v_1) and $(v_i, \gamma_i), i = 2, 3$, such that:

$$\begin{aligned}
&f(\xi_t^1(u,v), \xi_t^2(u,v)) = f(u,v) + f^{(1,0)}(u,v)(\xi_t^1(u,v)-u) + f^{(0,1)}(u,v)(\xi_t^2(u,v)-v) \\
&+ \frac{1}{2}f^{(2,0)}(u_1,v_1)(\xi_t^1(u,v)-u)^2 + \frac{1}{2}f^{(1,1)}(u_1,v_1)(\xi_t^1(u,v)-u)(\xi_t^2(u,v)-v) \\
&+ \frac{1}{2}f^{(0,2)}(u_1,v_1)(\xi_t^2(u,v)-v)^2,
\end{aligned}$$

and

$$\begin{aligned}
e^{-(\hat{\lambda}+v)t} &= 1-(\hat{\lambda}+v_3)t = 1-(\hat{\lambda}+v)t + \frac{1}{2}(\hat{\lambda}+v_2)^2 t^2, \\
e^{-\gamma t} &= 1-\gamma_3 t = 1-\gamma t + \frac{1}{2}\gamma_2^2 t^2.
\end{aligned}$$

Since $f \in C_c^\infty(\mathbb{R}_+^2)$, we have $\sup_{t\in(0,1]} \left\| \frac{W(t)-\mathcal{I}}{t} f \right\| < \infty$, and for any compact set $K \subset \mathbb{R}_+^2$, there exists constant $C_{f,K} > 0$ such that

$$\left\| \frac{W(t)-\mathcal{I}}{t} f - \mathcal{H}f \right\|_K \leq tC_{f,K} \to 0, \quad t \downarrow 0.$$

Hence the proof is concluded. $\square$

5.5.2 Main result

Theorem 5.19. *Let* $(\mathcal{S}_t)_{t\geq 0}$ *be the bi-continuous semigroup on* $\widehat{BC}(\mathbb{R}^2_+)$ *defined in* (5.20)*, and* $(W(t))_{t\geq 0}$ *as defined in* (5.29)*.*

The bi-closure of $(\mathcal{L}+\mathcal{H}, C_c^\infty(\mathbb{R}^2_+))$ *generates a bi-continuous semigroup* $\mathcal{U}_t$ *given by the Chernoff product formula, i.e.,*

$$\mathcal{U}_t f = \rho - \lim_{n\to\infty}\left(W\left(\frac{t}{n}\right)\mathcal{S}_{\frac{t}{n}}\right)^n f,\ f \in \widehat{BC}(\mathbb{R}^2_+),$$

uniformly for t *in compact intervals in* $[0,\infty)$*.*

Proof. Let

$$V(t) := W\,(t)\,\mathcal{S}_t, \quad t \geq 0.$$

We shall check that V satisfies the conditions in Theorem 3.26.

Firstly, by definition, $V(0) = \mathcal{I}$.

$$\left\|(V(t))^k\right\| \leq \|W(t)\|^k\|\mathcal{S}_t\|^k \leq 1.$$

Secondly, we check that for each $f \in C_c^\infty(\mathbb{R}^2_+)$,

$$\text{ucc} - \lim_{t\downarrow 0}\frac{V(t)f-f}{t} - \mathcal{L}f - \mathcal{H}f = 0.$$

We have for each $t > 0$,

$$\begin{aligned}\frac{V(t)f-f}{t} - \mathcal{L}f - \mathcal{H}f &= \frac{W(t)\mathcal{S}_t f - f}{t} - \mathcal{L}f - \mathcal{H}f\\ = \left(\frac{W(t)-\mathcal{I}}{t}f - \mathcal{H}f\right) &+ W(t)\left(\frac{\mathcal{S}_t-\mathcal{I}}{t}f - \mathcal{L}f\right) + (W(t)-\mathcal{I})\mathcal{L}f.\end{aligned}$$

Notice that $\mathcal{L}f \in \widehat{BC}(\mathbb{R}^2_+)$, and hence the first and the third term above ucc-converge to 0 following Lemma 5.18, and the second term ucc-converges to 0 following Proposition 5.15.

Thirdly, for each $(u,v) \in \mathbb{R}_+^2$, let $m_n = u$. Then

$$
\begin{aligned}
&\left|(V(t))^n f(m_n, v)\right| \\
&\leq \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_n, v); m_{n-1}\right) \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_{n-1}, \xi_t^2(m_n, v)); m_{n-2}\right) \\
&\quad \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_{n-2}, \xi_t^2(m_{n-1}, \xi_t^2(m_n, v))); m_{n-3}\right) \\
&\quad \cdots \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_1, \Xi^{2\to n}(v)); m_0\right) \left|f(m_0, \Xi^{1\to n}(v))\right| dm_0 \cdots dm_{n-1},
\end{aligned}
$$

where we use the shorthand notation $\Xi^{k\to l}(v)$ to represent $\xi_t^2(m_k, \xi_t^2(m_{k+1}, \cdots \xi_t^2(m_l, v))$ for each $k, l \in \mathbb{N}, k < l$.

To prove the local bi-equicontinuity, we take any $K \subset \mathbb{R}_+^2$, $t_0 > 0$ and $\epsilon > 0$. Recall that $\xi_t^1(u,v) \leq u e^{|\hat{\lambda}|t}$ for any v, and then

$$
\begin{aligned}
&\left|(V(t))^n f(u, v)\right| \leq \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_n, v); m_{n-1}\right) \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_{n-1}, \xi_t^2(m_n, v)); m_{n-2}\right) \\
&\quad \cdots \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_1, \Xi^{2\to n}(v)); m_0\right) \left|f(m_0, \Xi^{1\to n}(v))\right| dm_0 \cdots dm_{n-1}, \\
&= \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_n, v); m_{n-1}\right) \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_{n-1}, \xi_t^2(m_n, v)); m_{n-2}\right) \\
&\quad \cdots \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_1, \Xi^{2\to n}(v)); m_0\right) (1 + m_0 + v) \frac{\left|f(m_0, \Xi^{1\to n}(v))\right|}{1 + m_0 + v} dm_0 \cdots dm_{n-1}, \\
&\leq \sup_{(u,v)\in\mathbb{R}_+^2} \frac{|f(u,v)|}{1+u+v} \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_n, v); m_{n-1}\right) \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_{n-1}, \xi_t^2(m_n, v)); m_{n-2}\right) \\
&\quad \cdots \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_2, \Xi^{3\to n}(v)); m_1\right) \left(1 + \xi_t^1(m_1, \Xi^{2\to n}(v)) + v\right) dm_1 \cdots dm_{n-1} \\
&\leq \sup_{(u,v)\in\mathbb{R}_+^2} \frac{|f(u,v)|}{1+u+v} e^{|\hat{\lambda}|t} \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_n, v); m_{n-1}\right) \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_{n-1}, \xi_t^2(m_n, v)); m_{n-2}\right) \\
&\quad \cdots \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_2, \Xi^{3\to n}(v)); m_1\right) m_1 \, dm_1 \cdots dm_{n-1} + (1+v) \sup_{(u,v)\in\mathbb{R}_+^2} \frac{|f(u,v)|}{1+u+v}.
\end{aligned}
$$

Iterate this procedure and we have

$$\left|\left(e^{-|\hat{\lambda}|t}V(t)\right)^n f(u,v)\right| \leq u \sup_{(u',v')\in\mathbb{R}_+^2} \frac{|f(u',v')|}{1+u'+v'} + (1+v) \sup_{(u',v')\in\mathbb{R}_+^2} \frac{|f(u',v')|}{1+u'+v'}.$$

For any $\|\cdot\|$-bounded sequence $\{f_j\}_{j\in\mathbb{N}} \subset \widehat{BC}(\mathbb{R}_+^2)$ ucc-converging to 0, let $M > 2\sup_{j\in\mathbb{N}} \|f_j\|$. We can decompose each f_j as $f_j = h_j + \hat{f}_j$ where $\hat{f}_j(u,v)$ vanishes when one or both components tend to 0, and

$$h_j(u,v) \leq M(1 - x_j^{-1}u)\mathbf{1}_{u\in(0,x_j^{-1}]} + M(1 - x_j^{-1}v)\mathbf{1}_{v\in(0,x_j^{-1}]},$$

with $\{x_j\}_{j\in\mathbb{N}}$ being a sequence of positive constants tending to 0.

Similar to the previous argument,

$$\begin{aligned}
&M^{-1}|(V(t))^n h_j(u,v)| \leq (1 - x_j^{-1}v)\mathbf{1}_{v\in(0,x_j^{-1}]} \\
&+ \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_n, v); m_{n-1}\right) \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_{n-1}, \xi_t^2(m_n, v)); m_{n-2}\right) \\
&\cdots \int_{\mathbb{R}_+} g\left(t, \xi_t^1(m_1, \Xi^{2\to n}(v)); m_0\right)(1 - x_j^{-1}m_0)\mathbf{1}_{m_0\in(0,x_j^{-1}]}\, dm_0 \cdots dm_{n-1}.
\end{aligned}$$

By the estimation (A.27), for $t > 0$, we can find j_0 such that for all $j > j_0$,

$$(\mathbb{R}_+ \times (0, x_j^{-1}]) \cap K = ((0, x_j^{-1}] \times \mathbb{R}_+) \cap K = \emptyset,$$

and for $(u,v) \in K$,

$$\begin{aligned}
M^{-1}|(V(t))^n h_j(u,v)| \leq& (1 - x_j^{-1}v)\mathbf{1}_{v\in(0,x_j^{-1}]} + \sup_{(u,v)\in\mathbb{R}_+^2} \int_0^{x_j} g\left(t, u; m\right)(1 - x_j^{-1}m)\, dm \\
<& \sup_{u\in\mathbb{R}_+} I_2\left(\frac{2\sqrt{ux_j}}{t}\right) e^{-u/t} < \epsilon/M, \\
\|(V(t))^n h_j(u,v)\|_K <& \epsilon.
\end{aligned}$$

In addition, by bi-continuity of both $W(t)$ and $\mathcal{S}_t$, and the $\|\cdot\|$-boundedness of

$(V(t))^n$, we have

$$\lim_{t\downarrow 0} \|(V(t))^n h_j\|_K = \|h_j\|_K = 0,$$

for all sufficiently small x_j. Together, we can choose such j_0 uniformly for t on compact intervals.

Now, since $\hat{f}_j$ vanishes as one or both components tend to 0, and ucc$-\lim_{j\to\infty} \hat{f}_j = 0$, we can find j_1 such that for all $j > j_1$,

$$\sup_{(u,v)\in\mathbb{R}_+^2} \frac{\left|\hat{f}_j(u,v)\right|}{1+u+v} < \epsilon.$$

It follows that for all $j > j_0 \vee j_1$, and each $(u,v) \in K$,

$$\left|\left(e^{-|\hat{\lambda}|t}V(t)\right)^n f_j(u,v)\right| \le |(V(t))^n h_j(u,v)| + \left|\left(e^{-|\hat{\lambda}|t}V(t)\right)^n \hat{f}_j(u,v)\right| < \epsilon + (1+u+v)\epsilon,$$
$$\left\|\left(e^{-|\hat{\lambda}|t}V(t)\right)^n f_j\right\|_K < (2 + \|u+v\|_K)\epsilon.$$

This concludes the local bi-equicontinuity.

In the next part, we show that both $C_c^\infty(\mathbb{R}_+^2)$ and $(\alpha - \mathcal{L} - \mathcal{H})C_c^\infty(\mathbb{R}_+^2)$ are bi-dense subsets of $\widehat{BC}(\mathbb{R}_+^2)$.

The first half of the statement is true because, by Lemma 5.12, $C_c(\mathbb{R}_+^2)$ is bi-dense in $\widehat{BC}(\mathbb{R}_+^2)$, and it is well-known that $C_c^\infty(\mathbb{R}_+^2)$ is $\|\cdot\|$-dense in $C_c(\mathbb{R}_+^2)$.

By Proposition 3.25, $(\mathcal{S}_t)_{t\ge 0}$ and $(\mathcal{T}_t)_{t\ge 0}$ restricted to $C_0(\mathbb{R}_+^2)$ are strongly continuous semigroups, and their generators are extension of operators $(\mathcal{L}, C_c^\infty(\mathbb{R}_+^2))$ and $(\mathcal{H}, C_c^\infty(\mathbb{R}_+^2))$. By Theorem 2.12, [16], $(\mathcal{L}, C_c^\infty(\mathbb{R}_+^2))$ and $(\mathcal{H}, C_c^\infty(\mathbb{R}_+^2))$ are dissipative. By Theorem 3.14 (iii), [49], $(\mathcal{L}+\mathcal{H}, C_c^\infty(\mathbb{R}_+^2))$ is also dissipative.

In Proposition 5.4, we showed that the $(\mathcal{L}+\mathcal{H}, C_c^\infty(\mathbb{R}_+^2))$-martingale problem is well-posed. By Theorem 8.1.1, [50] and Proposition 3.4, [16], the closure of $(\mathcal{L} + \mathcal{H}, C_c^\infty(\mathbb{R}_+^2))$, denoted as $\left(\overline{\mathcal{L}+\mathcal{H}}, \mathrm{Dom}\left(\overline{\mathcal{L}+\mathcal{H}}\right)\right)$, generates a strongly continuous contraction semigroup on $C_0(\mathbb{R}_+^2)$.

By Proposition 2.1, [16], for all $\alpha > 0$, $(\alpha - \overline{\mathcal{L}+\mathcal{H}})$ is a one-to-one mapping from $\mathrm{Dom}\left(\overline{\mathcal{L}+\mathcal{H}}\right)$ to $C_0(\mathbb{R}_+^2)$. From the property of the domain of strongly continuous semigroups, we also know that

$$\text{Dom}\left(\overline{\mathcal{L}+\mathcal{H}}\right) \subset \{f \in C^{2,1}(\mathbb{R}_+^2) : \|(\mathcal{L}+\mathcal{H})f\| < \infty\}.$$

Let $K \subset \mathbb{R}_+^2$ be compact, $\alpha > 0$ and $\epsilon > 0$. Since $(\alpha - \overline{\mathcal{L}+\mathcal{H}})\text{Dom}\left(\overline{\mathcal{L}+\mathcal{H}}\right) = C_0(\mathbb{R}_+^2)$ is bi-dense in $\widehat{BC}(\mathbb{R}_+^2)$. For each $h \in \widehat{BC}(\mathbb{R}_+^2)$, we can find a $\|\cdot\|$-bounded sequence $\{h_n\}_{n\in\mathbb{N}} \subset C_0(\mathbb{R}_+^2)$ and some n_0 such that for all $n > n_0$, $\|h_n - h\|_K < \frac{1}{3}\epsilon$.

Let $f_n = (\alpha - \overline{\mathcal{L}+\mathcal{H}})^{-1}h_n \in \text{Dom}\left(\overline{\mathcal{L}+\mathcal{H}}\right)$. Let $\{M_m\}_{m\in\mathbb{N}}$ be an exhaustion of $\mathbb{R}_+^2$. For each $n \in \mathbb{N}$, we can construct a $\|\cdot\|$-bounded sequence $\{f_{nm}\}_{m\in\mathbb{N}} \subset C_c^\infty(\mathbb{R}_+^2) \subset D$ satisfying $f_{nm} = f_n$ on M_m . Since $f_n \in C_0(\mathbb{R}_+^2)$, such $\{f_{nm}\}_{m\in\mathbb{N}}$ has the property $\|\cdot\| - \lim_{m\to\infty} f_{nm} = f_n$, and $\|\cdot\| - \lim_{m\to\infty}(\mathcal{L}+\mathcal{H})f_{nm} = \overline{\mathcal{L}+\mathcal{H}}f_n$ for each n.

For any $\epsilon > 0$, there exist $n_0 \in \mathbb{N}$ such that for all $n > n_0$,

$$\begin{aligned}
&\|(\alpha - \mathcal{L} - \mathcal{H})f_{nn} - h\|_K \leq \left\|(\alpha - \overline{\mathcal{L}+\mathcal{H}})f_n - h\right\|_K + \alpha\|f_n - f_{nn}\|_K \\
&+ \left\|\overline{\mathcal{L}+\mathcal{H}}(f_n - f_{nn})\right\|_K \\
<&\frac{1}{3}\epsilon + \frac{1}{3}\epsilon + \left\|\overline{\mathcal{L}+\mathcal{H}}(f_n - f_{nn})\right\|_K < \epsilon.
\end{aligned}$$

The uniform boundedness of $\{(\alpha - \mathcal{L} - \mathcal{H})f_{nn}\}_{n\in\mathbb{N}}$ follows from the uniform boundedness of $\{h_n\}_{n\in\mathbb{N}}$. Hence $\text{ucc} - \lim_{n\to\infty}(\alpha - \mathcal{L} - \mathcal{H})f_{nn} = h$, which suggests that $(\alpha - \mathcal{L} - \mathcal{H})C_c^\infty(\mathbb{R}_+^2)$ is bi-dense in $\widehat{BC}(\mathbb{R}_+^2)$.

It is easy to check the uniform boundedness on $C_c^\infty(\mathbb{R}_+^2)$:

$$\sup_{t\in(0,1]}\left\|\frac{V(t)f - f}{t}\right\| \leq \sup_{t\in(0,1]}\|W(t)\|\left\|\frac{\mathcal{S}_t - \mathcal{I}}{t}f\right\| + \sup_{t\in(0,1]}\left\|\frac{W(t) - \mathcal{I}}{t}f\right\| < \infty.$$

As all four conditions in Theorem 3.26 are met, we can conclude that the bi-closure of $(\mathcal{L}+\mathcal{H}, C_c^\infty(\mathbb{R}_+^2))$ generates a bi-continuous semigroup $(\mathcal{U}_t)_{t\geq 0}$ given by

$$\mathcal{U}_t f = \rho - \lim_{n\to\infty}\left(V\left(\frac{t}{n}\right)\right)^n f = \lim_{n\to\infty}\left(W\left(\frac{t}{n}\right)\mathcal{S}_{\frac{t}{n}}\right)^n f,$$

for all $f \in \widehat{BC}(\mathbb{R}_+^2)$ and uniformly for t in compact intervals in $[0, \infty)$. □

Proof of Theorem 5.9. The main theorem 5.9 follows from Theorem 5.19. Since $\mathbf{1}_{\mathbb{R}_+^2} \in \widehat{BC}(\mathbb{R}_+^2)$, $U(u,v,t_0-t) = \mathcal{U}_t \mathbf{1}_{\mathbb{R}_+^2}(u,v)$ is the unique mild solution of (5.18) (Proposition 6.4, [51]). From the standard probability interpretation of the solution, we have

$$U(u,v,t_0-t) = \mathbb{E}_{u,v}\left[\mathbf{1}_{\mathbb{R}_+^2}(Y_{t\wedge\tau_\Delta}, Z_{t\wedge\tau_\Delta})\right],$$

where τ_Δ is the stopping time of one of Y_t and Z_t first hits $\{0,\infty\}$. The boundary condition at $u \downarrow 0$ alone suffices to define a unique solution because as we have shown in the proof of Proposition 5.4, $\tau_\Delta = \inf\{t : Y_t = 0\}$ almost surely. □

5.6 A closer look at the approximation

When the first-order operator $\mathcal{H}$ is simple enough, $(\mathcal{T}_t)_{t\geq 0}$ can be written explicitly and we can take $W(t) = \mathcal{T}_t$. Notice in our proof, apart from having smooth coefficients, we only require the following:

Condition 1: For each $t_0 > 0$, we can find compact sets in $[0,\infty)^2$ such that the characteristics (analogous to the solution of (5.24)) initiated in K will remain in K_0 up to time t_0.

Under Condition 1, the exact solution of the PDE associated with $\mathcal{L}+\mathcal{H}$ and the same domain and boundary conditions, can be expressed in terms of the Chernoff product formula.

In the following example, we are able to find the closed-form expression of the Chernoff product formula through deduction. For straightforward comparison, we use the same notations as our main problem, and only change the definition of $\mathcal{H}$.

Example 5.20.

$$\frac{\partial U}{\partial t} = -u\frac{\partial^2 U}{\partial u^2} + auv\frac{\partial U}{\partial u} + bv\frac{\partial U}{\partial v},$$
$$U(u,v,t_0) = \mathbf{1}_{\mathbb{R}_+^2}(u,v), \quad \lim_{u\downarrow 0} U(u,v,t) = 0 \text{ for } t \in [0,t_0),$$

where parameters $a, b \in \mathbb{R}$.

The characteristic equation of the first-order problem is

$$\frac{du}{dt} = -auv,\ \frac{dv}{dt} = -bv,$$

whose behaviour at $t \to \infty$ is $(u, v) \to (u_0 \exp\{-av_0 b^{-1}\}, 0)$. The solution of the characteristic equation is

$$u(t) = u_0 \exp\big\{-ab^{-1}v_0(1 - e^{-bt})\big\},\ v(t) = v_0 e^{-bt},$$

and therefore we can write $\mathcal{T}_t f(u, v) = f(u \exp\big\{avb^{-1}(e^{-bt} - 1)\big\}, ve^{-bt})$.

Condition 1 is satisfied, since for each compact set K, we can find $K \subset [0, x]^2$ for some $x > 0$, and let $K_0 = [0, x] \times [0, xe^{bt_0}]$. Hence, the solution to the PDE can be expressed in the form of the Lie-Trotter product.

We calculate the first few terms in the Lie-Trotter product sequence, and deduce that

$$\begin{aligned}(\mathcal{S}_h\mathcal{T}_h)^n \mathbf{1}_{\mathbb{R}_+^2}(u, v) &= 1 - \exp\left\{-u\left(\sum_{j=0}^{n-1} h \exp\big\{rv(1 - e^{-jbh})\big\}\right)^{-1}\right\} \\ &\to 1 - \exp\left\{-\frac{u}{\int_0^t \exp\{ab^{-1}v(1 - e^{-bx})\}dx}\right\},\end{aligned}$$

as $h = t/n$ and $n \to \infty$.

We can verify that

$$U(u, v, t) = \mathcal{U}_{t_0 - t}\mathbf{1}_{\mathbb{R}_+^2}(u, v) = 1 - \exp\left\{-\frac{u}{\int_0^{t_0 - t} \exp\{ab^{-1}v(1 - e^{-bx})\}dx}\right\}$$

is indeed a solution.

Proof. We denote $\exp\left\{-\frac{u}{\int_0^{t_0-t}\exp\left\{ab^{-1}v(1-e^{-bx})\right\}dx}\right\}$ as $\exp\{\cdot\}$ in the following proof.

$$
\begin{aligned}
\frac{\partial U}{\partial t} &= \exp\{\cdot\}u\exp\left\{ab^{-1}v(1-e^{-b(t_0-t)})\right\}\left(\int_0^{t_0-t}\exp\left\{ab^{-1}v(1-e^{-bx})\right\}dx\right)^{-2},\\
\frac{\partial U}{\partial u} &= \exp\{\cdot\}\left(\int_0^{t_0-t}\exp\left\{ab^{-1}v(1-e^{-bx})\right\}dx\right)^{-1},\\
\frac{\partial^2 U}{\partial u^2} &= -\exp\{\cdot\}\left(\int_0^{t_0-t}\exp\left\{ab^{-1}v(1-e^{-bx})\right\}dx\right)^{-2},\\
\frac{\partial U}{\partial v} &= -\exp\{\cdot\}u\left(\int_0^{t_0-t}\exp\left\{ab^{-1}v(1-e^{-bx})\right\}dx\right)^{-2}\\
&\quad\cdot\int_0^{t_0-t}\exp\left\{ab^{-1}v(1-e^{-bx})\right\}ab^{-1}(1-e^{-bx})dx.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\frac{\partial U}{\partial t}-\left(-u\frac{\partial^2 U}{\partial u^2}+auv\frac{\partial U}{\partial u}+bv\frac{\partial U}{\partial v}\right)\\
=&u\exp\{\cdot\}\left(\int_0^{t_0-t}\exp\left\{ab^{-1}v(1-e^{-bx})\right\}dx\right)^{-2}\\
&\left(\exp\left\{ab^{-1}v(1-e^{-b(t_0-t)})\right\}-1+av\int_0^{t_0-t}\exp\left\{ab^{-1}v(1-e^{-bx})\right\}(1-e^{-bx})dx\right.\\
&\left.-av\int_0^{t_0-t}\exp\left\{ab^{-1}v(1-e^{-bx})\right\}dx\right)\\
&=0,
\end{aligned}
$$

where in the last step, we use the following results derived through integration by parts:

$$
\int_0^{t_0-t}\exp\left\{ab^{-1}v(1-e^{-bx})\right\}(-e^{-bx})dx=-(av)^{-1}(\exp\left\{ab^{-1}v(1-e^{-b(t_0-t)})\right\}-1).
$$

□

Chapter 6

The subcritical parameter regime

In this chapter, we discuss the cases where the initial size of infection is not small enough for us to apply Theorem 4.6 in Chapter 4.

The stochastic SIRS model is significantly more complicated comparing to SIS and SIR models. This is because the SIS model is one-dimensional, and although the SIR model is two-dimensional, one of its components monotonically increases almost surely.

In this chapter, we focus on two types of initial states:

- Large initial size of the infected populations, i.e., $I_0 \asymp N$: we find that in this case, if we also have that $\lambda_o(N) \leq 1$ and $\gamma_o(N)$ tend to 0 faster than some negative power of N, then the behaviour of the extinction time is similar to the stochastic SIR model. This is discussed in Section 6.2.

  Notice that this case covers the half of the critical parameter regime satisfying $\lambda_o \leq 1$.

- Medium initial size of the infected and immune populations: In this case, we can approximate the stochastic SIRS model under critical scaling by the corresponding deterministic model, until I_t^N reaches a state sufficiently small for Theorem 4.6 to be applicable. This is discussed in Section 6.3.

As we will show in Section 6.4, when the parameters are both bounded away from the criticality, i.e., $\lim_{N\to\infty} \lambda_o(N) = \lambda_{lim} < 1$ and $\lim_{N\to\infty} \gamma_o(N) = \gamma_{lim} > 0$, the conditions for the medium and the large initial size of the infected and immune populations overlap.

6.1 Introduction

The paradigm of this chapter is to compare the long-term behaviour of a Markov chain with deterministic equations. The important result in this area is Kurtz's *law of large numbers* and can be found in [16]. It says that density dependent Markov chains can be well approximated up to a fixed constant time by the solution of the corresponding ODE, which describes their average drift. We briefly mentioned the concept of density dependent processes in Chapter 2. By Kurtz's definition, a sequence of Markov chains $\{X^N\}_{N\in\mathbb{N}}$ is density dependent if Markov chains $\widetilde{X}^N := N^{-1}X^N$, $N \in \mathbb{N}$, has jump rates depending only on the state of $\widetilde{X}^N$.

There is a large volume of literature proving Kurtz's *law of large numbers*, with various extensions with respect to different types of convergence and estimates of the rate of convergence. Among those, we pay particular attention to the exponential martingale estimate approach of [12], which is introduced in details in Section 3.1.1.

Mathematically, the extinction time problem of subcritical and near-critical epidemic models is related to the long-term behaviour of Markov chains near the stable fixed point of their corresponding ODEs. More specifically, it requires that we extend the classical result to over a time interval that grows with N. The available approaches are fundamentally related, and the key is to use the negative linear drift of Markov chains near the stable fixed point. Barbour et al. [52] discuss a fairly general set of models using integration by parts and exponential martingale estimate of ODEs. However, as pointed out by [32], the results obtained in [52] contain non-explicit constants, and their estimations are too weak to investigate near-critical phenomena. Following a similar idea, the authors of [8] and [32] study specific one-dimensional and two-dimensional epidemic models respectively. In their problem, they can explicitly bound the non-linear part of the drift near the stable fixed point, and thus obtain a much refined estimate. Alternatively, Foxall [6] solves the problem by applying the ODE approximation to Markov chains under critical scaling (space and time, e.g., as in Chapter 4 for SIRS models), instead of average scaling (N^{-1}

in space). His argument uses a drift barrier estimate derived to solve problems in a more general setting [53], and uses L_2-estimate instead of exponential estimate, which is a weaker bound according to [12].

Our approach combines the advantages of both. By applying the integration by parts to the critically-scaled Markov chains and then applying the exponential martingale estimate of ODEs, we avoid the argument of martingale transform in [32] and significantly reduce the algebraic work.

6.2 Large initial sizes of infection and immunity

In this section, we consider the case when $\lim_{N\to\infty} \lambda_o \leq 1$, and there exists $\epsilon_\gamma > 0$ such that $\gamma_o = o\left(N^{-\epsilon_\gamma}\right)$, with initial states

$$\lim_{N\to\infty} I_0^N/N > 0, \quad \lim_{N\to\infty} R_0^N/N \geq 0.$$

Before we state the main result of this section, we first introduce the following lemma, which obtains the time it takes for the deterministic SIR model to travel between two states. This will allow us to express our main result in a more concise form.

Recall the deterministic SIR model (2.3) with parameter $\lambda_o > 0$. In this subsection, (x_1, x_2, x_3) represents the solution of the following deterministic SIR model:

$$\begin{aligned} \frac{dx_1}{dt} &= -\lambda_o x_1 x_2, \\ \frac{dx_2}{dt} &= \lambda_o x_1 x_2 - x_2 = \lambda_o x_2(1 - x_2 - x_3) - x_2, \\ \frac{dx_3}{dt} &= x_2. \end{aligned} \tag{6.1}$$

Here, as well as for the deterministic SIRS model in the next section, we state all three components, since we will use both representations (x_1, x_2) and (x_2, x_3) of the solution. We use variables x_i, $i = 1, 2, 3$, since we have used (y, z) for the deterministic model under critical scaling in Chapter 4.

For each N, let $(x_1^N(t), x_2^N(t))$ be the solution of (6.1) with

$$(x_1^N(0), x_2^N(0)) = (x_0^N, y_0^N).$$

Let $t_{\text{SIR}}^N(a \to b)$ be the time taken for x_1^N to travel from a to b.

Recall our discussion in Lemma 2.3, where for each $y \in [0, y_0^N)$, $\theta = \theta^N(y)$ is the solution of

$$x_0^N + y_0^N + \frac{\log(\theta/x_0^N)}{\lambda_o} - \theta = y. \tag{6.2}$$

Let $\theta^{*,N} := \theta^*(x_0^N, y_0^N; \lambda_o) = \lim_{a \downarrow 0} \theta^N(a)$.

Lemma 6.1 (Elapsed time of the deterministic SIR model)**.** *For each N, $0 < \theta^{*,N} < b(N) < a(N) \leq x_0^N$, we have*

$$t_{\text{SIR}}^N(a \to b) = \int_b^a u^{-1} \left(x_0^N + y_0^N + \frac{\log(u/x_0^N)}{\lambda_o} - u \right)^{-1} du. \tag{6.3}$$

Also, there exists constant k_{SIR} such that for any $a = a(N) \downarrow 0$ as $N \to \infty$,

$$k_{\text{SIR}} = k_{\text{SIR}}(x_0^N, y_0^N; \lambda_o) = \lim_{N\to\infty} -\log a^{-1} + (1 - \lambda_o \theta^{*,N}) t_{\text{SIR}}^N \left(x_0^N \to \theta^N(a) \right).$$

Remark 6.2. When λ_o is independent of N, the statement on $t_{\text{SIR}}^N(a \to b)$ and θ above is consistent with $J(b, a)$ and function θ defined in [1].

Proof. See Appendix A.1. □

Our main result below suggests that when γ_o tends to 0 sufficiently fast, and the initial size of infection is of order N, the behaviour of the stochastic SIRS model resembles the stochastic SIR model, as studied by [1].

Theorem 6.3. *Consider the stochastic SIRS model defined in* (2.5) *with parameters* $\lim_{N\to\infty} \lambda_o(N) =: \lambda_{lim} \leq 1$ *and* $\gamma_o = o(N^{-\epsilon_\gamma})$ *for some* $\epsilon_\gamma > 0$*, and initial states*

$$\lim_{N\to\infty} I_0^N/N > 0, \ \lim_{N\to\infty} R_0^N/N \geq 0.$$

Then we have

$$\mathbb{P}\left[(1-\lambda_{lim}\theta^*_{lim})T^N_o - k_{\text{SIR}} - \log N - \log(1-\lambda_{lim}\theta^*_{lim}) \le w\right] \to e^{-e^{-w}}.$$

where $\theta^*_{lim} = \lim_{N\to\infty}\theta^*(x^N_0, y^N_0; \lambda_o)$, *and* $k_{\text{SIR}} = k_{\text{SIR}}\left(\frac{N-I_0-R_0}{N}, \frac{I_0}{N}; \lambda_o\right)$ *are defined in Lemma 6.1.*

The idea behind the proof of Theorem 6.3 is that we can approximate the stochastic SIRS model by the deterministic SIR model until I^N_t is sufficiently small, and then we can apply the extinction time result for small initial infections in Theorem 4.6.

We refer to the part of the stochastic SIRS model that closely resembles a deterministic SIR model as the 'initial phase' of the epidemic.

Lemma 6.4 (Initial phase)**.** *Under the same assumptions as in Theorem 6.3, define*

$$X^{N,1}_t = I^N_t/N,\ X^{N,2}_t = R^N_t/N,$$

then we have for any $t_1(N) \le \frac{1}{8}\epsilon_\gamma \log N + c_1$, $c_1 \in \mathbb{R}$,

$$\mathbb{P}\left[\sup_{s\le t_1}\left|X^{N,1}_s - x^N_2(s)\right| + \left|X^{N,2}_s - x^N_3(s)\right| \ge N^{-2\epsilon_\gamma}\right] \to 0, \quad N\to\infty,$$

where (x^N_2, x^N_3) *is the solution of the corresponding ODE system* (6.1) *with initial condition*

$$(x^N_2(0), x^N_3(0)) = (I^N_0/N, R^N_0/N).$$

Proof. We use the ODE approximation argument introduced in Section 3.1.

As stated in Section 3.1, X^N has the following decomposition

$$X^{N,1}_t = X^{N,1}_0 + \int_0^t \lambda_o\left(1 - X^{N,1}_s - X^{N,2}_s\right)X^{N,1}_s - X^{N,1}_s ds + M^{N,1}_t,$$
$$X^{N,2}_t = X^{N,2}_0 + \int_0^t X^{N,1}_s - \gamma_o X^{N,2}_s ds + M^{N,2}_t,$$

where $M^{N,i}, i = 1, 2$, are zero-mean martingales.

It follows that

$$
\begin{aligned}
\Big|X_t^{N,1} - x_2^N(t)\Big| \leq & |X_0^{N,1} - x_2^N(0)| + \int_0^t (1+\lambda_o)\big|X_s^{N,1} - x_2^N(s)\big| + \lambda_o\big|X_s^{N,2} - x_3^N(s)\big| ds \\
& + \Big|M_t^{N,1}\Big|, \\
|X_t^{N,2} - x_3^N(t)| \leq & |X_0^{N,2} - x_3^N(0)| + \int_0^t \big|X_s^{N,1} - x_2^N(s)\big| + \gamma_o X_s^{N,2} ds + \Big|M_t^{N,2}\Big|.
\end{aligned}
$$

We combine the two and obtain for each N,

$$
\begin{aligned}
& \sup_{s\leq t_1} \big|X_s^{N,1} - x_2^N(s)\big| + \big|X_s^{N,2} - x_3^N(s)\big| + \frac{\gamma_o}{3} \leq \Big|X_0^{N,1} - x_2^N(0)\Big| + \Big|X_0^{N,2} - x_3^N(0)\Big| \\
& + 3\int_0^{t_1} \left(\sup_{u\leq s} \big|X_u^{N,1} - x_2^N(u)\big| + \big|X_u^{N,2} - x_3^N(u)\big| + \frac{\gamma_o}{3}\right) ds \\
& + \sup_{s\leq t_1} \left(\big|M_s^{N,1}\big| + \big|M_s^{N,2}\big| + \frac{\gamma_o}{3}\right).
\end{aligned}
$$

By Gronwall's inequality,

$$
\begin{aligned}
& \sup_{s\leq t_1} \big|X_s^{N,1} - x_2^N(s)\big| + \big|X_s^{N,2} - x_3^N(s)\big| + \frac{\gamma_o}{3} \qquad (6.4) \\
\leq & \left(\Big|X_0^{N,1} - x_2^N(0)\Big| + \Big|X_0^{N,2} - x_3^N(0)\Big| + \sup_{s\leq t_1} \big|M_s^{N,1}\big| + \big|M_s^{N,2}\big| + \frac{\gamma_o}{3}\right) e^{3t_1}.
\end{aligned}
$$

Applying Proposition 3.2 where we choose $\bar{a}(N) = t_1(\lambda_o(N)+1)/N$ when $i=1$ and $\bar{a}(N) = t_1(\gamma_o(N)+1)/N$ when $i=2$, we have

$$
\begin{aligned}
& \mathbb{P}\left[\sup_{s\leq t_1} \big|M_s^{N,1}\big| + \big|M_s^{N,2}\big| > 2N^{-\epsilon_\gamma}\right] \leq \mathbb{P}\left[\sup_{s\leq t_1} \big|M_s^{N,1}\big| > N^{-\epsilon_\gamma}\right] + \mathbb{P}\left[\sup_{s\leq t_1} \big|M_s^{N,2}\big| > N^{-\epsilon_\gamma}\right] \\
\leq & 2\exp\left\{-\frac{N^{1-2\epsilon_\gamma}}{4(1+\lambda_o)t_1}\right\} + 2\exp\left\{-\frac{N^{1-2\epsilon_\gamma}}{4(1+\gamma_o)t_1}\right\} = o(1).
\end{aligned}
$$

By (6.4),

$$\begin{aligned}
&\mathbb{P}\left[\sup_{s\le t_1}\left|X_s^{N,1}-x_2^N(s)\right|+\left|X_s^{N,2}-x_3^N(s)\right|>N^{-\frac{1}{2}\epsilon_\gamma}\right]\\
\le&\mathbb{P}\left[\sup_{s\le t_1}\left|X_s^{N,1}-x_2^N(s)\right|+\left|X_s^{N,2}-x_3^N(s)\right|>\left(2N^{-\epsilon_\gamma}+\frac{\gamma_o}{3}\right)e^{3t_1}\right]\\
\le&\mathbb{P}\left[\sup_{s\le t_1}\left|M_s^{N,1}\right|+\left|M_s^{N,2}\right|>2N^{-\epsilon_\gamma}\right]=o(1).
\end{aligned}$$

where we use the fact that

$$\left(2N^{-\epsilon_\gamma}+\frac{\gamma_o}{3}\right)e^{3t_1}\le\frac{7}{3}N^{-\epsilon_\gamma}e^{c_1}N^{\frac{3}{8}\epsilon_\gamma}=o\left(N^{-\frac{1}{2}\epsilon_\gamma}\right).$$

□

Now we are ready to prove the main result of this section.

Proof of Theorem 6.3. We write for shorthand the following quantities in Lemma 6.1:

$$\theta^*(N):=\theta^*\left(\frac{N-I_0-R_0}{N},\frac{I_0}{N};\lambda_o\right),\quad k_{\mathrm{SIR}}:=k_{\mathrm{SIR}}\left(\frac{N-I_0-R_0}{N},\frac{I_0}{N};\lambda_o\right).$$

First, we note that

$$\lim_{N\to\infty}\left(\lambda_o^{-1}-\theta^*\left(\frac{N-I_0-R_0}{N},\frac{I_0}{N};\lambda_o\right)\right)=(\lambda_{lim}^{-1}-\theta_{lim}^*)>0.$$

For each N, let

$$t_1(N)=(1-\lambda_o\theta^*)^{-1}k_{\mathrm{SIR}}+\frac{\lambda_o}{8}\epsilon_\gamma\log N.$$

For $\epsilon=\epsilon_\gamma(1-\lambda_{lim}\theta_{lim}^*)$, by the definition of k_{SIR} in Lemma 6.1, we have $x_2^N(t_1)=N^{-\frac{1}{8}\epsilon}$ and $x_3^N(t_1)\sim 1-\theta^*>0$.

By Lemma 6.4, with probability tending to 1, $\left|X_{t_1}^{N,1}-x_2^N(t_1)\right|$ and $\left|X_{t_1}^{N,2}-x_3^N(t_1)\right|$ are smaller or equal to $N^{-\frac{1}{2}\epsilon_\gamma}$.

Since $I_{t_1}^N\le N^{1-\epsilon/8}+2N^{1-\epsilon_\gamma/2}$ with probability tending to 1, it suggests that, starting from time t_1, we can apply Case 2.2 of Theorem 4.6 and have

$$\mathbb{P}\left[(1-\lambda_{lim}\theta^*_{lim})(T^N_o - t_1) - \log(1-\lambda_{lim}\theta^*_{lim})I^N_{t_1} \le w\right] \to e^{-e^{-w}},$$

where $\theta^*_{lim} := \lim_{N\to\infty}\theta^*(N)$.

$$\begin{aligned}&(1-\lambda_{lim}\theta^*_{lim})T^N_o - \frac{1-\lambda_{lim}\theta^*_{lim}}{1-\lambda_o\theta^*}k_{\text{SIR}} - \frac{1}{8}\epsilon_\gamma(1-\lambda_{lim}\theta^*_{lim})\log N \\ &- \log N^{1-\frac{1}{8}\epsilon} - \log\left(X^{N,1}_{t_1}N^{\frac{1}{8}\epsilon}\right) - \log(1-\lambda_{lim}\theta^*_{lim}) \\ =&(1-\lambda_{lim}\theta^*_{lim})T^N_o - k_{\text{SIR}} - \log N - \log(1-\lambda_{lim}\theta^*_{lim}) + o(1),\end{aligned}$$

It follows from Lemma A.4 that

$$\mathbb{P}\left[(1-\lambda_{lim}\theta^*_{lim})T^N_o - k_{\text{SIR}} - \log N - \log(1-\lambda_{lim}\theta^*_{lim}) \le w\right] \to e^{-e^{-w}}.$$

□

6.3 Medium initial sizes of infection and immunity

In this section, we assume that the parameters satisfy $(1-\lambda_o)N^{1/3} = \hat{\lambda} \to \infty$ and $\gamma_o N^{1/3} = \gamma \to \infty$.

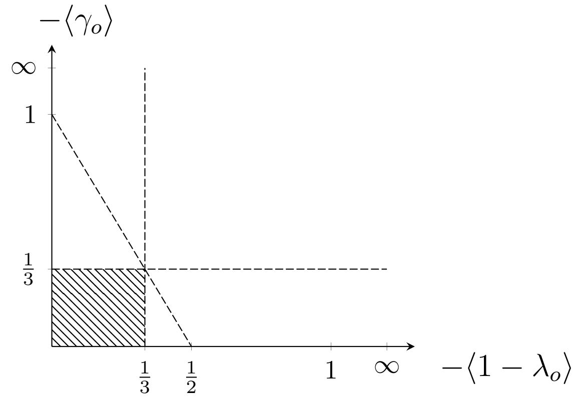


Figure 6.1: Parameter regime: the shaded area represents the regime of interest.

Similar to the previous section, this section starts with the necessary variables to represent the elapsed time of the deterministic model, before we proceed to the statement of the main theorem.

Recall the deterministic SIRS model defined in (2.6). To avoid ambiguity in notation, we will denote the solution of the deterministic SIRS model as $(x_1(t), x_2(t), x_3(t))$:

$$\begin{aligned}\frac{dx_1}{dt} &= \gamma_o(1 - x_1 - x_2) - \lambda_o x_1 x_2 \\ \frac{dx_2}{dt} &= \lambda_o x_1 x_2 - x_2 = \lambda_o(1 - x_2 - x_3)x_2 - x_2, \\ \frac{dx_3}{dt} &= x_2 - \gamma_o x_3,\end{aligned} \tag{6.5}$$

Lemma 6.5 (Elapsed time of the deterministic SIRS model)**.** *Let $(x_1(t), x_2(t))$ be the solution of* (6.5) *with $(x_1(0), x_2(0)) = (s_0, i_0)$. Let $t_{\text{SIRS}}(m \to n)$ be the time it takes for x_2 to travel from m to n, $0 < n < m \leq i_0$. Then*

$$t_{\text{SIRS}}(m \to n) = \int_n^m \frac{v(y)}{y} dy, \tag{6.6}$$

where $v(y)$ is the solution of differential equation

$$\frac{dv}{dy} = -v^2\left(\lambda_o + \frac{\gamma_o}{y}\right) + v^3\left(\lambda_o + \lambda_o\gamma_o + \frac{\gamma_o - \lambda_o\gamma_o}{y}\right), y \in (0, i_0],$$

with the end condition $v(i_0) = \frac{1}{1-\lambda_o s_0}$.

Proof. The proof is a simple manipulation of (6.5), see Appendix A.2. □

Remark 6.6. When the initial states and parameters depend on N, $t_{\text{SIRS}}(m \to n)$ is defined for all sufficiently large N.

When $\gamma_o = 0$, for each N,

$$t_{\text{SIRS}}(m \to n) = t^N_{\text{SIR}}(\theta(m) \to \theta(n)).$$

Now we are ready to state the main result of this section.

Theorem 6.7. *Consider the stochastic SIRS model defined in* (2.5) *with parameters $\lambda_o = \lambda_o(N) \uparrow 1$ and $\gamma_o = \gamma_o(N) \downarrow 0$, satisfying $(1 - \lambda_o)N^{1/3} \to \infty$, $\gamma_o N^{1/3} \to \infty$,* *and* $\lim_{N\to\infty}\frac{1-\lambda_o}{\gamma_o}\neq 1$*. If* $\gamma_o \ll 1-\lambda_o$*, we in addition require that there exists some small* $\epsilon_p > 0$ *such that* $N^{\frac{\epsilon_p}{3}}\gamma_o^{1+\epsilon_p} \ll 1-\lambda_o$*.*

Suppose the initial states of the model satisfy for some constants $c_y, d_y, c_z > 0$ *the conditions*

$$I_0^N = I_0(N) \leq \left[d_y \frac{N(1-\lambda_o)\gamma_o}{\log(N^{1/3}(1-\lambda_o))}, c_y(1-\lambda_o)\gamma_o N\right],$$

$$R_0^N = R_0(N) \leq \begin{cases} c_z N \gamma_o^{1+\epsilon_p}, \text{ if } \gamma_o \ll 1-\lambda_o, \\ c_z(1-\lambda_o)N, \text{ otherwise.} \end{cases}$$

Then

$$\mathbb{P}\left[(1-\lambda_o)T_o^N - (1-\lambda_o)t_{\text{SIRS}}\left(I_0 N^{-1} \to a\right) - \log a - \log N(1-\lambda_o) \leq w\right] \to e^{-e^{-w}},$$

where $a = a(N) \geq N^{-1}$ *can be chosen arbitrarily, as long as* $a = o((1-\lambda_o)\gamma_o)$*. The asymptotic distribution above is independent of the choice of* a*.*

Remark 6.8. We do not have a result for the exact asymptotics of

$$(1-\lambda_o)t_{\text{SIRS}}\left(I_0 N^{-1} \to a\right) + \log a$$

under the assumptions of Theorem 6.7, besides that it is $O(\log N)$. We know this because $I_0 N^{-1} \leq 1$ and $a(N) \geq N^{-1}$, and we know that $x_2(t)$ decays as fast as $e^{-(1-\lambda_o)t}$.

However, we do have the exact asymptotics of this quantity when $\lambda_o < 1$ and $\gamma_o > 0$ have limits bounded away from the criticality, which allows us to give a more accurate description of the asymptotic distribution for some special cases. This is discussed in Section 6.4.

For our purpose, it is easier to work with a change of variables. Therefore, in the first subsection below, we introduce the change of variables and the integration by parts transformation to the scaled SIRS model (Y^N, Z^N) defined as in (5.2). In the second subsection, we will state the precise form of the main result and complete the proof.

6.3.1 Transformation to (Y^N, Z^N)

Recall the scaled parameters (5.1) and the scaled process (5.2):

$$\hat{\lambda}(N) := (1-\lambda_o)N^{1/3}, \quad \gamma(N) := \gamma_o N^{1/3},$$
$$Y_t^N := \frac{I^N_{N^{1/3}t}}{N^{1/3}}, \quad Z_t^N := \frac{R^N_{N^{1/3}t}}{N^{2/3}}.$$

In the case when $\hat{\lambda} \neq \gamma$ for all $N \in \mathbb{N}$, we can perform a helpful change of variables so that the stable manifolds of the corresponding ODE are tangent to the axes at the origin. For the case where $\hat{\lambda} = \gamma$ for some $N \in \mathbb{N}$, we expect that we can perform a different change of variables analogous to [8], and the treatment will be the same in principle. Since our focus is on the scaling of the parameters, we will only discuss the case when $\hat{\lambda} \neq \gamma$ for all $N \in \mathbb{N}$.

Similar to above, we introduce the notations

$$(\hat{\lambda} \vee \gamma)(N) := \hat{\lambda}(N) \vee \gamma(N), \quad (\hat{\lambda} \wedge \gamma)(N) := \hat{\lambda}(N) \wedge \gamma(N).$$

To avoid too many variables, we will recycle the definition of the stochastic processes X^N, M^N and V^N from previous chapters.

Introduce the change of variables to the scaled stochastic SIRS model (Y^N, Z^N) defined in (5.2) as the follows:

$$\begin{pmatrix} \tilde{Y}^N \\ \tilde{Z}^N \end{pmatrix} := \begin{pmatrix} 1 & 0 \\ (\hat{\lambda}-\gamma)^{-1} & 1 \end{pmatrix} \begin{pmatrix} Y^N \\ Z^N \end{pmatrix}. \tag{6.7}$$

The process $(\tilde{Y}^N, \tilde{Z}^N)$ has the following transition rates at state (y, z):

$$\begin{aligned}
&(y,z) \to \left(y + N^{-1/3}, z + \frac{N^{-1/3}}{\hat{\lambda}-\gamma}\right), \text{ at rate} \\
&\qquad N^{2/3}(1-\hat{\lambda}N^{-1/3})\left(1 + y\left(\frac{N^{-1/3}}{\hat{\lambda}-\gamma} - N^{-2/3}\right) - zN^{-1/3}\right)y, \\
&(y,z) \to \left(y, z - N^{-2/3}\right), \text{ at rate } N^{2/3}\gamma(-(\hat{\lambda}-\gamma)^{-1}y + z), \\
&(y,z) \to \left(y - N^{-1/3}, z - \frac{N^{-1/3}}{\hat{\lambda}-\gamma} + N^{-2/3}\right), \text{ at rate } N^{2/3}y.
\end{aligned} \tag{6.8}$$

It has the drift coefficients

$$\mu(\tilde{Y}^N) = -\hat{\lambda}\tilde{Y}^N + \lambda_o \left((\tilde{Y}^N)^2((\hat{\lambda}-\gamma)^{-1} - N^{-1/3}) - \tilde{Y}^N\tilde{Z}^N \right),$$

$$\begin{aligned}\mu(\tilde{Z}^N) &= -\gamma\tilde{Z}^N + \frac{1-\hat{\lambda}N^{-1/3}}{\hat{\lambda}-\gamma}(\tilde{Y}^N)^2((\hat{\lambda}-\gamma)^{-1} - N^{-1/3}) - \tilde{Y}^N\tilde{Z}^N \\ &= -\gamma\tilde{Z}^N + \frac{\lambda_o}{\hat{\lambda}-\gamma}\left((\tilde{Y}^N)^2((\hat{\lambda}-\gamma)^{-1} - N^{-1/3}) - \tilde{Y}^N\tilde{Z}^N \right),\end{aligned}$$

and the diffusion coefficients

$$\begin{aligned}\sigma^2(\tilde{Y}^N) =& \tilde{Y}^N \left(2 - \hat{\lambda}N^{-1/3} + \lambda_o N^{-1/3}\left(\tilde{Y}^N((\hat{\lambda}-\gamma)^{-1} - N^{-1/3}) - \tilde{Z}^N \right) \right), \\ \sigma^2(\tilde{Z}^N) =& \tilde{Y}^N(\hat{\lambda}-\gamma)^{-2}\left(\lambda_o + N^{-1/3}\lambda_o \left(\tilde{Y}^N((\hat{\lambda}-\gamma)^{-1} - N^{-1/3}) - \tilde{Z}^N \right) \right. \\ & \left. + ((\hat{\lambda}-\gamma)N^{-1/3} - 1)^2 \right) + N^{-2/3}\gamma\left(-\tilde{Y}^N(\hat{\lambda}-\gamma)^{-1} + \tilde{Z}^N \right).\end{aligned}$$

By the properties of the continuous Markov chains, we have the standard decomposition

$$\tilde{Y}^N_t = \tilde{Y}^N_0 - \int_0^t \hat{\lambda}\tilde{Y}^N_s ds - \lambda_o\tilde{Y}^N_s \left((N^{-1/3} - (\hat{\lambda}-\gamma)^{-1})\tilde{Y}^N_s + \tilde{Z}^N_s \right) ds + M^{N,1}_t, \quad (6.9)$$

$$\tilde{Z}^N_t = \tilde{Z}^N_0 - \int_0^t \gamma\tilde{Z}^N_s ds - \frac{\lambda_o}{\hat{\lambda}-\gamma}\tilde{Y}^N_s \left((N^{-1/3} - (\hat{\lambda}-\gamma)^{-1})\tilde{Y}^N_s + \tilde{Z}^N_s \right) ds + M^{N,2}_t,$$

where $M^{N,i}_t$, $i = 1, 2$, are zero-mean martingales.

The mean value ODE for each N is

$$\begin{pmatrix} d\tilde{y}/dt \\ d\tilde{z}/dt \end{pmatrix} := \begin{pmatrix} -\hat{\lambda} & 0 \\ 0 & -\gamma \end{pmatrix}\begin{pmatrix} \tilde{y} \\ \tilde{z} \end{pmatrix} - \lambda_o\tilde{y}\left((N^{-1/3} - (\hat{\lambda}-\gamma)^{-1})\tilde{y} + \tilde{z} \right)\begin{pmatrix} 1 \\ (\hat{\lambda}-\gamma)^{-1} \end{pmatrix},$$

$$(\tilde{y}(0), \tilde{z}(0)) = (\tilde{Y}^N_0, \tilde{Z}^N_0), \quad (6.10)$$

whose solution is denoted as $(\tilde{y}^N, \tilde{z}^N)$.

We introduce the function $f^N(y, z) := y(N^{-1/3} - (\hat{\lambda}-\gamma)^{-1}) + z$. Performing integration by parts on (6.10), we have

$$\tilde{y}^N(t) = e^{-\hat{\lambda}t}\tilde{y}^N(0) - \lambda_o \int_0^t e^{-\hat{\lambda}(t-s)}\tilde{y}^N(s) f^N(\tilde{y}^N(s), \tilde{z}^N(s))ds, \tag{6.11}$$

$$\tilde{z}^N(t) = e^{-\gamma t}\tilde{z}^N(0) - \lambda_o(\hat{\lambda} - \gamma)^{-1} \int_0^t e^{-\gamma(t-s)}\tilde{y}^N(s) f^N(\tilde{y}^N(s), \tilde{z}^N(s))ds. \tag{6.12}$$

There is a corresponding representation of continuous-time Markov chain $(\tilde{Y}^N, \tilde{Z}^N)$, according to Lemma 4.1, [54].

$$\tilde{Y}_t^N = e^{-\hat{\lambda}t}\tilde{Y}_0^N - \lambda_o \int_0^t e^{-\hat{\lambda}(t-s)}\tilde{Y}_s^N f^N(\tilde{Y}_s^N, \tilde{Z}_s^N)ds + e^{-\hat{\lambda}t}V_t^{N,1}, \tag{6.13}$$

$$\tilde{Z}_t^N = e^{-\gamma t}\tilde{Z}_0^N - \frac{\lambda_o}{\hat{\lambda} - \gamma} \int_0^t e^{-\gamma(t-s)}\tilde{Y}_s^N f^N(\tilde{Y}_s^N, \tilde{Z}_s^N)ds + e^{-\gamma t}V_t^{N,2}, \tag{6.14}$$

where

$$V_t^{N,1} := \int_0^t e^{\hat{\lambda}s} dM_s^{N,1}, \tag{6.15}$$

$$V_t^{N,2} := \int_0^t e^{\gamma s} dM_s^{N,2}. \tag{6.16}$$

To find fluctuation bounds for $V^{N,i}$, we define

$$X_t^{N,1} = \int_0^t e^{b_1 s} d\tilde{Y}_s^N, \tag{6.17}$$

$$X_t^{N,2} = \int_0^t e^{b_2 s} d\tilde{Z}_s^N,$$

where $b_1 = \hat{\lambda}$ and $b_2 = \gamma$.

Let J^N and $q^N((\tilde{Y}^N, \tilde{Z}^N), j)$, $j \in J^N$, be the set of possible jumps and the corresponding transition rates of $(\tilde{Y}^N, \tilde{Z}^N)$. Then for $i = 1, 2$,

$$V_t^{N,i} = X_t^{N,i} - X_0^{N,i} - \int_0^t \sum_{j \in J^N} q^N((\tilde{Y}_s^N, \tilde{Z}_s^N), j) e^{b_i s} j_i ds,$$

where j_i is the i-th component of j.

We will discuss the fluctuation under different parameter regimes below, and prove the ODE approximation respectively.

6.3.2 Proof of Theorem 6.7

The idea behind the proof is that under the assumptions above, the scaled, transformed stochastic SIRS model $(\tilde{Y}^N, \tilde{Z}^N)$ can be well-approximated by the corresponding ODE (6.10), up until $I_t^N = N^{1/3} Y_t^N$ and $R_t^N = N^{2/3} Z_t^N$ are small enough for us to apply Theorem 4.6.

Proposition 6.9 (Initial phase)**.** *Consider for each $N \in \mathbb{N}$ the Markov chain $(\tilde{Y}^N, \tilde{Z}^N)$ defined in* (6.8)*, with parameters $\hat{\lambda} = (1-\lambda_o)N^{1/3} \to \infty$ and $\gamma = \gamma_o N^{1/3} \to \infty$, and initial states satisfying $\tilde{Y}_0^N \in [y_*, y^*]$ and $\tilde{Z}_0^N = O\left(\hat{\lambda} \wedge \gamma\right)$ where $y^* = c_y \hat{\lambda}\gamma$ for some $c_y > 0$ and $y_* = \hat{\lambda}^{1-\epsilon}\gamma$ for some sufficiently small constant $\epsilon > 0$.*

Let $(\tilde{y}^N, \tilde{z}^N)$ be the solution to (6.10) *with $(\tilde{y}^N(0), \tilde{z}^N(0)) = (\tilde{Y}_0^N, \tilde{Z}_0^N)$. For any constant $t > 0$ independent of N, we have as $N \to \infty$,*

$$\mathbb{P}\left[\sup_{s\le t}\left|\tilde{Y}_s^N - \tilde{y}^N(s)\right| > (\hat{\lambda}\vee\gamma)^{1-2\epsilon} \text{ and } \sup_{s\le t}\left|\tilde{Z}_s^N - \tilde{z}^N(s)\right| > \frac{3}{2}(\hat{\lambda}\vee\gamma)^{-2\epsilon}\right] = o(1).$$

To prove this, we first need to introduce some new variables.

Define $t_* := \inf\{t : \tilde{y}^N(t) \le y_*\}$, $\delta\tilde{Y}_t^N := \tilde{Y}_t^N - \tilde{y}^N(t)$ and $\delta\tilde{Z}_t^N := \tilde{Z}_t^N - \tilde{z}^N(t)$.

Define the stopping times

$$\tau_X^N := \inf\left\{t : \left|\delta\tilde{Y}_t^N\right| \ge (\hat{\lambda}\vee\gamma)^{1-2\epsilon} \text{ or } \left|\delta\tilde{Z}_t^N\right| \ge \frac{3}{2}(\hat{\lambda}\vee\gamma)^{-2\epsilon}\right\},$$

$$\tau_{2,X}^N := \inf\left\{t : \left|\delta\tilde{Y}_t^N\right| \ge 2(\hat{\lambda}\vee\gamma)^{1-2\epsilon} \text{ or } \left|\delta\tilde{Z}_t^N\right| \ge 2(\hat{\lambda}\vee\gamma)^{-2\epsilon}\right\},$$

and

$$\tau_V^N := \inf\left\{t : \left|V_t^{N,1}\right| \vee \hat{\lambda}\left|V_t^{N,2}\right| \ge (\hat{\lambda}\vee\gamma)^{1-3\epsilon}\right\}.$$

Clearly $\tau_X^N \le \tau_{2,X}^N$ -a.s.

We can derive estimations for $\tilde{y}^N(t)$ and $\tilde{z}^N(t)$ from (6.10).

Firstly, it is easy to see that $\tilde{y}^N$ monotonically decreases. Multiplying (6.11) by $(\hat{\lambda}-\gamma)^{-1}$ and subtracting it from (6.12), we have

$$\begin{aligned}\left|\tilde{z}^N(t)\right| =&\left|e^{-\gamma t}\tilde{z}^N(0)+\tilde{y}^N(t)(\hat{\lambda}-\gamma)^{-1}-e^{-\hat{\lambda}t}\tilde{y}^N(0)(\hat{\lambda}-\gamma)^{-1}\right| \\ \leq& 2\tilde{y}^N(0)\left|\hat{\lambda}-\gamma\right|^{-1}+\left|\tilde{z}^N(0)\right| = O(\hat{\lambda}\wedge\gamma). \end{aligned} \tag{6.18}$$

Returning to the original variables $(y^N(t), z^N(t))$, we have

$$dy^N/dt = -\hat{\lambda}y^N - \lambda_o y(N^{-1/3}y^N + z^N) \leq -\hat{\lambda}y^N,$$

and therefore

$$\tilde{y}^N(t) = y^N(t) \leq y^N(0)e^{-\hat{\lambda}t}. \tag{6.19}$$

Furthermore,

$$\begin{aligned}&\int_0^t \left|f^N\left(\tilde{y}^N(s),\tilde{z}^N(s)\right)\right|ds = \int_0^t \left|\left(N^{-1/3}-(\hat{\lambda}-\gamma)^{-1}\right)\tilde{y}^N(s)+\tilde{z}^N(s)\right|ds \\ =&\int_0^t\left|N^{-1/3}\tilde{y}^N(s)+e^{-\gamma s}\tilde{z}^N(0)-e^{-\hat{\lambda}s}\tilde{y}^N(0)(\hat{\lambda}-\gamma)^{-1}\right|ds \\ \leq&\int_0^t\left(N^{-1/3}\tilde{y}^N(0)e^{-\hat{\lambda}s}+e^{-\gamma s}\left|\tilde{z}^N(0)\right|+\left|\hat{\lambda}-\gamma\right|^{-1}\tilde{y}^N(0)e^{-\hat{\lambda}s}\right)ds \\ \leq&\left(N^{-1/3}+\left|\hat{\lambda}-\gamma\right|^{-1}\right)\frac{\tilde{y}^N(0)}{\hat{\lambda}}+\frac{\left|\tilde{z}^N(0)\right|}{\gamma}=O(1).\end{aligned}$$

The next lemma gives us the bound for V^N.

Lemma 6.10. *For any $t_0 > 0$, $\epsilon < \frac{1}{7}$,*

$$\mathbb{P}\left[\sup_{s\leq t_0\wedge\tau_{2,X}^N}\left|V_s^{N,1}\right| > e^{2\hat{\lambda}t_0}(\hat{\lambda}\vee\gamma)^{1-3\epsilon}\right]\to 0,$$

$$\mathbb{P}\left[\sup_{s\leq t_0\wedge\tau_{2,X}^N}\left|V_s^{N,2}\right| > e^{2\gamma t_0}(\hat{\lambda}\vee\gamma)^{-3\epsilon}\right]\to 0,$$

as $N\to\infty$.

Proof. The proof is an application of Proposition 3.2 to the process X^N defined in (6.17).

In the following, we use the fact that

$$\tilde{Z}^N \geq \tilde{Y}^N(\hat{\lambda}-\gamma)^{-1},$$

and therefore almost surely

$$1+\tilde{Y}^N\left(\frac{N^{-1/3}}{\hat{\lambda}-\gamma}-N^{-2/3}\right)-\tilde{Z}^N N^{-1/3}\in(0,1). \tag{6.20}$$

For all sufficiently large N, conditioning on the event $\{t \leq \tau^N_{2,X}\}$, we can apply (6.20) to the diffusion coefficient $\sigma^2\left(\tilde{Y}^N_s\right)$ and have for $i=1$,

$$\begin{aligned}
&\int_0^t \sum_{j\in J^N} q^N(X^N_{s-},j)(e^{\hat{\lambda}s}j_i)^2 ds \leq 2\int_0^t e^{2\hat{\lambda}s}\tilde{Y}^N_s\, ds\\
\leq& 2\frac{e^{2\hat{\lambda}t}-1}{2\hat{\lambda}}\left(c_y\hat{\lambda}\gamma+2(\hat{\lambda}\vee\gamma)^{1-2\epsilon}\right)\\
\leq& e^{2\hat{\lambda}t}(\hat{\lambda}\vee\gamma)^{1+\epsilon}.
\end{aligned}$$

We can choose $\delta = e^{2\hat{\lambda}t_0}(\hat{\lambda}\vee\gamma)^{1-3\epsilon}$, and then by Proposition 3.2,

$$\mathbb{P}\left[\sup_{s\leq t_0\wedge\tau^N_{2,X}}\left|V^{N,1}_s\right|>e^{2\hat{\lambda}t_0}(\hat{\lambda}\vee\gamma)^{1-3\epsilon}\right]\leq 2\exp\left\{-\frac{e^{4\hat{\lambda}t_0}(\hat{\lambda}\vee\gamma)^{2-6\epsilon}}{e^{2\hat{\lambda}t_0}(\hat{\lambda}\vee\gamma)^{1+\epsilon}}\right\}\to 0,\quad N\to\infty.$$

Similarly, for $i=2$,

$$\begin{aligned}
&\int_0^t \sum_{j\in J^N} q^N(X^N_{s-},j)(e^{\gamma s}j_i)^2 ds\\
\leq&\frac{e^{2\gamma t}-1}{2\gamma}\Big[(\hat{\lambda}-\gamma)^{-2}\left(\hat{\lambda}\gamma+2(\hat{\lambda}\vee\gamma)^{1-2\epsilon}\right)+N^{-2/3}\gamma(2(\hat{\lambda}\vee\gamma)^{-2\epsilon}+\tilde{z}^N(0))\\
&+\left(N^{-2/3}+(\hat{\lambda}-\gamma)^{-2}+2N^{-1/3}\left|\hat{\lambda}-\gamma\right|^{-1}\right)\left(\hat{\lambda}\gamma+2(\hat{\lambda}\vee\gamma)^{1-2\epsilon}\right)\Big]\\
\leq& e^{2\gamma t}(\hat{\lambda}\vee\gamma)^{-1+\epsilon}.
\end{aligned}$$

For any $t_0 > 0$, we can choose $\delta = e^{2\gamma t_0}(\hat{\lambda} \vee \gamma)^{-3\epsilon}$, and then by Proposition 3.2, when constant $\epsilon > 0$ is chosen to be sufficiently small,

$$\mathbb{P}\left[\sup_{s \le t_0 \wedge \tau_{2,X}^N} \left|V_s^{N,2}\right| > e^{2\gamma t_0}(\hat{\lambda} \vee \gamma)^{-3\epsilon}\right] \le 2\exp\left\{-\frac{e^{4\gamma t_0}(\hat{\lambda} \vee \gamma)^{-6\epsilon}}{e^{2\gamma t_0}(\hat{\lambda} \vee \gamma)^{-1+\epsilon}}\right\} \to 0, \quad N \to \infty.$$

□

Proof of Proposition 6.9. Subtract (6.11) from (6.13), we have

$$e^{\hat{\lambda} t}\delta \tilde{Y}_t^N = \delta \tilde{Y}_0^N - \lambda_o \int_0^t e^{\hat{\lambda} s}\left(\tilde{Y}_s^N f^N(\tilde{Y}_s^N, \tilde{Z}_s^N) - \tilde{y}^N(s) f^N(\tilde{y}^N(s), \tilde{z}^N(s))\right) ds + e^{\hat{\lambda} t} V_t^{N,1}. \tag{6.21}$$

Taking absolute value on both sides and then the supremum over time interval $[0, t]$, we have

$$\begin{aligned}
e^{\hat{\lambda} t}\sup_{s\in[0,t]}\left|\delta\tilde{Y}_s^N\right| \le& \left|\delta\tilde{Y}_0^N\right| + \int_0^t e^{\hat{\lambda} s}\sup_{u\in[0,s]}\left|\delta\tilde{Y}_u^N\right|\left|f^N(\tilde{y}^N(s), \tilde{z}^N(s))\right| ds \\
&+ \int_0^t e^{\hat{\lambda} s}\sup_{u\in[0,s]}\left|\delta\tilde{Y}_u^N\right|\left|f^N(\tilde{Y}_s^N, \tilde{Z}_s^N) - f^N(\tilde{y}^N(s), \tilde{z}^N(s))\right| ds \\
&+ \int_0^t e^{\hat{\lambda} s}\tilde{y}^N(s)\left|f^N(\tilde{Y}_s^N, \tilde{Z}_s^N) - f^N(\tilde{y}^N(s), \tilde{z}^N(s))\right| ds + e^{\hat{\lambda} t}\sup_{s\in[0,t]}\left|V_s^{N,1}\right|, \\
e^{\hat{\lambda} t}\sup_{s\in[0,t]}\left|\delta\tilde{Y}_s^N\right| \le& \left|\delta\tilde{Y}_0^N\right| + \int_0^t e^{\hat{\lambda} s}\sup_{u\in[0,s]}\left|\delta\tilde{Y}_u^N\right|\left|f^N(\tilde{y}^N(s), \tilde{z}^N(s))\right| ds \\
&+ \int_0^t e^{\hat{\lambda} s}\sup_{u\in[0,s]}\left|\delta\tilde{Y}_u^N\right|\left(\left(N^{-1/3} + \left|\hat{\lambda} - \gamma\right|^{-1}\right)\sup_{u\in[0,s]}\left|\delta\tilde{Y}_u^N\right| + \sup_{u\in[0,s]}\left|\delta\tilde{Z}_u^N\right|\right) ds \\
&+ \int_0^t e^{\hat{\lambda} s}\tilde{y}^N(0)e^{-\hat{\lambda} s}\left(\left(N^{-1/3} + \left|\hat{\lambda} - \gamma\right|^{-1}\right)\sup_{u\in[0,s]}\left|\delta\tilde{Y}_u^N\right| + \sup_{u\in[0,s]}\left|\delta\tilde{Z}_u^N\right|\right) ds \\
&+ e^{\hat{\lambda} t}\sup_{s\in[0,t]}\left|V_s^{N,1}\right|,
\end{aligned}$$

where we use

$$f^N(\tilde{Y}_s^N, \tilde{Z}_s^N) - f^N(\tilde{y}^N(s), \tilde{z}^N(s)) = (N^{-1/3} - (\hat{\lambda} - \gamma)^{-1})\delta\tilde{Y}_s^N + \delta\tilde{Z}_s^N.$$

By Gronwall's inequality (Theorem 3.3), we have for $t \leq \tau_{2,X}^N \wedge \tau_V^N$,

$$
\begin{aligned}
&e^{\hat{\lambda}t} \sup_{s\in[0,t]} \left|\delta\tilde{Y}_s^N\right| \leq \left(\left|\delta\tilde{Y}_0^N\right| + e^{\hat{\lambda}t}\left|V_t^{N,1}\right| + \sup_{s\in[0,t]}\left|\delta\tilde{Z}_s^N\right|\tilde{y}^N(0)t \right) \\
&\exp\left\{ \int_0^t \left| f^N(\tilde{y}^N(s),\tilde{z}^N(s))\right| + \left(N^{-1/3} + \left|\hat{\lambda}-\gamma\right|^{-1}\right)\tilde{y}^N(0)e^{-\hat{\lambda}s}ds \right\} \\
&\exp\left\{ 2\left(N^{-1/3} + \left|\hat{\lambda}-\gamma\right|^{-1}\right)(\hat{\lambda}\vee\gamma)^{1-2\epsilon}t + 2(\hat{\lambda}\vee\gamma)^{-2\epsilon}t \right\} \\
\leq &\left(\left|\delta\tilde{Y}_0^N\right| + e^{\hat{\lambda}t}\left|V_t^{N,1}\right| + \sup_{u\in[0,s]}\left|\delta\tilde{Z}_u^N\right|\tilde{y}^N(0)t \right) \exp\left\{ 2\left(N^{-1/3} + \left|\hat{\lambda}-\gamma\right|^{-1}\right)\frac{\tilde{y}^N(0)}{\hat{\lambda}} + \frac{\tilde{z}^N(0)}{\gamma} \right\} \\
&\exp\left\{ 2(\hat{\lambda}\vee\gamma)^{-2\epsilon}t\left(\left(N^{-1/3} + \left|\hat{\lambda}-\gamma\right|^{-1}\right)(\hat{\lambda}\vee\gamma) + 1 \right) \right\}.
\end{aligned}
$$

Hence, for constant $t \leq \tau_{2,X}^N \wedge \tau_V^N$, given our assumptions to $\tilde{y}^N(0)$ and $\tilde{z}^N(0)$, there exists a constant $c_0 > 0$ such that for sufficiently large N,

$$
\begin{aligned}
e^{\hat{\lambda}t} \sup_{s\in[0,t]} \left|\delta\tilde{Y}_s^N\right| &\leq \left(\left|\delta\tilde{Y}_0^N\right| + e^{\hat{\lambda}t}\left|V_t^{N,1}\right| + \sup_{u\in[0,t]}\left|\delta\tilde{Z}_u^N\right|\tilde{y}^N(0)t \right) e^{c_0}, \\
\sup_{s\in[0,t]} \left|\delta\tilde{Y}_s^N\right| &\leq \left(\left|V_t^{N,1}\right| + \sup_{u\in[0,t]}\left|\delta\tilde{Z}_u^N\right|\tilde{y}^N(0)e^{-\hat{\lambda}t}t \right) e^{c_0} \leq (\hat{\lambda}\vee\gamma)^{1-2\epsilon}. \qquad (6.22)
\end{aligned}
$$

Similarly, from (6.12) and (6.14) we have

$$
\begin{aligned}
\delta\tilde{Z}_t^N =& \delta\tilde{Z}_0^N - \frac{\lambda_o}{\hat{\lambda}-\gamma}\int_0^t e^{-\gamma(t-s)}\left(\tilde{Y}_s^N f^N(\tilde{Y}_s^N,\tilde{Z}_s^N) - \tilde{y}^N(s)f^N(\tilde{y}^N(s),\tilde{z}^N(s))\right)ds + V_t^{N,2} \\
=& \delta\tilde{Z}_0^N - (\hat{\lambda}-\gamma)^{-1}\delta\tilde{Y}_0^N + (\hat{\lambda}-\gamma)^{-1}\delta\tilde{Y}_t^N - (\hat{\lambda}-\gamma)^{-1}V_t^{N,1} + V_t^{N,2}.
\end{aligned}
$$

The fluctuation of $\tilde{Z}^N$ can be bounded by a combination of the fluctuation of $\tilde{Y}^N$

and martingales $V^{N,i}$, $i = 1, 2$:

$$\begin{aligned}\sup_{s\in[0,t]}\left|\delta\tilde{Z}_s^N\right| \leq& \left|\delta\tilde{Z}_0^N\right| + \left|\hat{\lambda}-\gamma\right|^{-1}\left|\delta\tilde{Y}_0^N\right| + \left|\hat{\lambda}-\gamma\right|^{-1}\sup_{s\in[0,t]}\left|\delta\tilde{Y}_s^N\right| \\ &+ \left|\hat{\lambda}-\gamma\right|^{-1}\sup_{s\in[0,t]}\left|V_s^{N,1}\right| + \sup_{s\in[0,t]}\left|V_s^{N,2}\right|, \\ \sup_{s\in[0,t]}\left|\delta\tilde{Z}_s^N\right| \leq& \left|\hat{\lambda}-\gamma\right|^{-1}(\hat{\lambda}\vee\gamma)^{1-2\epsilon} + \left|\hat{\lambda}-\gamma\right|^{-1}(\hat{\lambda}\vee\gamma)^{1-3\epsilon} + (\hat{\lambda}\vee\gamma)^{-3\epsilon} \\ \leq& \frac{3}{2}(\hat{\lambda}\vee\gamma)^{-2\epsilon}.\end{aligned} \tag{6.23}$$

Together with (6.22) and (6.23), we can see that $\tau_{2,X}^N \wedge \tau_V^N \leq \tau_X^N$ almost surely. Combining with $\tau_X^N \leq \tau_{2,X}^N$ almost surely, we have for any constant $t > 0$,

$$\mathbb{P}\left[\tau_X^N \leq t\right] \leq \mathbb{P}\left[\tau_V^N \leq t\right],$$

and the statement follows from Lemma 6.10. □

Proof of Theorem 6.7. Under the assumptions in Theorem 6.7,

$$\tilde{Y}_0^N = I_0^N N^{-1/3} \in \left[d_y\frac{\hat{\lambda}\gamma}{\log\hat{\lambda}}, c_y\hat{\lambda}\gamma\right], \quad \tilde{Z}_0^N \asymp \hat{\lambda}\gamma\left|\hat{\lambda}-\gamma\right|^{-1} + \hat{\lambda}\wedge\gamma = O(\hat{\lambda}\wedge\gamma),$$

which satisfies the conditions of the initial states in Proposition 6.9.

Take a sufficiently small constant $\epsilon > 0$. The time it takes for $\tilde{y}^N$ to reach $y_* = \hat{\lambda}^{1-\epsilon}\gamma$, denoted as t_{ini}, equals to $N^{-1/3}t_{\mathrm{SIRS}}\left(I_0/N \to (1-\lambda_o)^{1-\epsilon}\gamma_o\right)$.

By (6.19), $t_{ini} \asymp \frac{\log\hat{\lambda}}{\hat{\lambda}} = o(1)$.

It is sufficient to take $t = 1$ in Proposition 6.9, and we have that, as $N \to \infty$,

$$\mathbb{P}\left[\sup_{s\leq 1}\left|\tilde{Y}_s^N - \tilde{y}^N(s)\right| > (\hat{\lambda}\vee\gamma)^{1-2\epsilon} \text{ and } \sup_{s\leq 1}\left|\tilde{Z}_s^N - \tilde{z}^N(s)\right| > \frac{3}{2}(\hat{\lambda}\vee\gamma)^{-2\epsilon}\right] = o(1).$$

With probability tending to 1, the event

$$\left\{\left|\tilde{Y}_{t_{ini}}^N - \tilde{y}^N(t_{ini})\right| \leq (\hat{\lambda}\vee\gamma)^{1-2\epsilon} \text{ and } \left|\tilde{Z}_{t_{ini}}^N - \tilde{z}^N(t_{ini})\right| \leq \frac{3}{2}(\hat{\lambda}\vee\gamma)^{-2\epsilon}\right\}$$

happens. Conditioned on this event,

$$I^N_{N^{1/3}t_{ini}} = N^{1/3}\tilde{Y}^N_{t_{ini}} \leq N^{1/3}\hat{\lambda}^{1-\epsilon}\gamma\left(1+\hat{\lambda}^{-\epsilon}+\gamma^{-2\epsilon}\right) = N(1-\lambda_o)^{1-\epsilon}\gamma_o\left(1+\hat{\lambda}^{-\epsilon}+\gamma^{-2\epsilon}\right),$$
$$R^N_{N^{1/3}t_{ini}} = N^{2/3}Z^N_{t_{ini}} \asymp N^{2/3}\left(z^N(t_{ini})+(\hat{\lambda}\vee\gamma)^{-2\epsilon}\right),$$

where $z^N(t_{ini})$ is the solution of the ODE system before transformation:

$$\begin{aligned} z^N(t_{ini}) &= \tilde{z}^N(t_{ini}) - (\hat{\lambda}-\gamma)^{-1}\tilde{y}^N(t_{ini}) = e^{-\gamma t_{ini}}z^N(0) + (\hat{\lambda}-\gamma)^{-1}y^N(0)(e^{-\gamma t_{ini}} - e^{-\hat{\lambda}t_{ini}}) \\ &= \begin{cases} O(\gamma), & \gamma \ll \hat{\lambda}, \\ O\left(\hat{\lambda}^{1-\frac{\gamma}{\hat{\lambda}}}\right), & \text{otherwise.} \end{cases} \end{aligned}$$

It is easy to check that this satisfies the conditions of Case 1.3, Theorem 4.6, and therefore

$$\begin{aligned} &\lim_{N\to\infty}\mathbb{P}\left[(1-\lambda_o)(T^N_o - N^{1/3}t_{ini}) - \log(1-\lambda_o)aN \leq w\right] \\ =&\lim_{N\to\infty}\mathbb{P}\left[(1-\lambda_o)T^N_o - (1-\lambda_o)t_{\text{SIRS}}\left(I_0/N \to a\right) - \log a - \log(1-\lambda_o)N \leq w\right] \\ =&e^{-e^{-w}}, \end{aligned}$$

where $a = (1-\lambda_o)^{1-\epsilon}\gamma_o$.

The asymptotic distribution should be independent of the choice of a. This is indeed the case when a is sufficiently small. Consider $0 < a_1(N) < a_2(N) = o((1-\lambda_o)\gamma_o)$. By (A.12),

$$\phi^N(a) = x^N_3(t_{ini}) = O\left(x^N_3(0) + x^N_2(0)|1-\lambda_o-\gamma_o|^{-1}\right) = o(\gamma_o).$$

Then we can apply Lemma A.3 and have

$$\begin{aligned} &(1-\lambda_o)T^N_o - (1-\lambda_o)t_{\text{SIRS}}\left(I_0/N \to a_2\right) - (1-\lambda_o)t_{\text{SIRS}}\left(a_2 \to a_1\right) - \log a_1 \\ &- \log(1-\lambda_o)N \\ =&(1-\lambda_o)T^N_o - (1-\lambda_o)t_{\text{SIRS}}\left(I_0/N \to a_2\right) - \log a_2 - \log(1-\lambda_o)N + o(1). \end{aligned}$$

The statement of Theorem 6.7 follows from Lemma A.4. □

6.4 A special case of Theorem 6.7: strongly subcritical

In general, we do not know the asymptotics of $t_{\text{SIRS}}\,(m \to n)$ when m, n are functions of N and $\lambda_o = \lambda_o(N) \uparrow 1$. However, we do know that when λ_o is bounded away from 1, $t_{\text{SIRS}}\,(m \to n)$ is the sum of a constant and a term of order $\log N$. More precisely:

Lemma 6.11. *Consider $t_{\text{SIRS}}\,(m \to n)$ defined in Lemma 6.5. When*

$$\lim_{N\to\infty} \lambda_o = \lambda_{lim} < 1, \quad \lim_{N\to\infty} \gamma_o = \gamma_{lim} > 0, \quad \lambda_o(N) + \gamma_o(N) \neq 1,$$

and

$$\lim_{N\to\infty} \frac{I_0^N}{N} > 0, \quad \lim_{N\to\infty} \frac{R_0^N}{N} > 0.$$

Then there exists a constant k_{SIRS} such that for any $a = a(N) \to 0$,

$$k_{\text{SIRS}} = \lim_{N\to\infty} \log a + (1-\lambda_o) t_{\text{SIRS}}\left(I_0^N/N \to a\right).$$

Proof. See Appendix A.2. □

Theorem 6.12. *Suppose $\lim_{N\to\infty} \lambda_o(N) = \lambda_{lim} < 1$ and $\lim_{N\to\infty} \gamma_o(N) = \gamma_{lim} > 0$ are constants independent of N, $\lambda_{lim} + \gamma_{lim} \neq 1$, and $\lambda_o(N) + \gamma_o(N) \neq 1$ for $N \in \mathbb{N}$.*

Suppose further that the initial condition satisfies

$$\lim_{N\to\infty} \frac{I_0^N}{N} > 0, \ \lim_{N\to\infty} \frac{R_0^N}{N} > 0.$$

Then we have as $N \to \infty$,

$$\mathbb{P}\left[(1-\lambda_o)T_o^N - (k_{\text{SIRS}} + \log N + \log(1-\lambda_o))\right] \to e^{-e^{-w}},$$

where k_{SIRS} is defined as in Lemma 6.11.

Proof. Notice that the proof of Theorem 6.7 up to the derivation of the asymptotic distribution containing a covers the case where $1-\lambda_o$ and γ_o are bounded away from 0. Therefore, we can apply the same proof with $c_y = i_0$ and $c_z = r_0$ and sufficiently small $a = N^{-\epsilon}, \epsilon > 0$, and obtain the asymptotic distribution

$$\lim_{N\to\infty} \mathbb{P}\left[(1-\lambda_o)T_o^N - (1-\lambda_o)t_{\text{SIRS}}\left(I_0^N \to a\right) - \log a - \log(1-\lambda_o)N \leq w\right] = e^{-e^{-w}}.$$

What is different here is that instead of Lemma A.3, we will use Lemma 6.11 and have

$$\begin{aligned} &\lim_{N\to\infty} \mathbb{P}\left[(1-\lambda_o)T_o^N - k_{\text{SIRS}} - \log(1-\lambda_o)N \leq w\right] \\ =&\lim_{N\to\infty} \mathbb{P}\left[(1-\lambda_o)T_o^N - (1-\lambda_o)t_{\text{SIRS}}\left(I_0^N \to a\right) - \log a - \log(1-\lambda_o)N \leq w\right] = e^{-e^{-w}}. \end{aligned}$$

□

Chapter 7

Numerical experiments

7.1 Method options

In general, there are two groups of methods to simulate the extinction time of the stochastic epidemic models.

The first group of methods are designed to approximate the solution to the Kolmogorov equation. In particular, to simulate the distribution of the extinction time, we need to obtain the probability vector of the model at multiple time points. One of the advantages of this type of method is that the error is easily controlled by choosing appropriate step size.

Among the methods we are aware of in the first group, we find that the implicit Euler method [55] is the most efficient. Even so, the time cost to approximate the extinction time of a stochastic SIRS model at $N = 10^8$ is too much (roughly 25 days by estimation). Other well-known methods include the Krylov subspace approximation (KSA) method (MATLAB Package `expokit.m`) [56], the finite state projection (FSP) approach and its variations [57], etc. We find through testing that KSA methods are not sufficiently efficient for our purposes, and FSP-based methods are not suitable for large population cases.

It is worth mentioning that, according to [55], the implicit Euler method is especially efficient for the models where the population process and its embedded counting process has a one-to-one mapping (e.g., SIR model). Such a model allows us to solve the Kolmogorov equation of the embedded counting process instead, and benefit from the fact that the transition matrix is lower triangular after appropriate reordering. However, neither the SIS nor the SIRS model has this property, which makes it difficult to further simplify the algorithm.

The other group of methods aims to simulate the paths of Markov chains. The naive method, known as the stochastic simulation algorithm (SSA or Gillespie), simulates the sojourn time between consecutive jumps, and then simulates the type of jump that takes place (See [58] for the description of the algorithm). It is often regarded as 'mathematically exact', in the sense that the sample paths it generates follow the same distribution as the Kolmogorov equation. The SSA method is computationally expensive when we simulate a large population until its extinction. Therefore, various methods are developed to approximate the SSA method, among which we choose the modified Poisson τ-leaping method proposed by [58].

The idea behind Poisson τ-leaping is that, instead of generating a random sojourn time as in the SSA, we preselect a sequence of time increments during which the transition rates are not expected to change significantly, and simulate the number of jumps in each time increment by Poisson random variables. The criteria of 'significant change' is controlled by the parameter ϵ.

However, since a Poisson random variable can take arbitrarily large values, there is positive probability that this method will generate samples with negative states. To avoid this, Cao, Gillespie and Petzold [58] propose the modified Poisson τ-leaping method. The intuition behind this modification is that we set a number n_c, and whenever a component of the model reaches a value below n_c, we simulate this component by the SSA method. When $n_c = \infty$, the modified Poisson τ-leaping algorithm is reduced to the SSA method; and when $n_c = 0$, the modified Poisson τ-leaping algorithm is reduced to the Poisson τ-leaping method.

We find that, in terms of computation time, it is infeasible to use the SSA method to simulate stochastic SIRS model with $N > 10^7$, especially for near-critical cases, and introducing the τ-leaping method significantly reduces the computation time.

7.2 Our implementation

We implement the simulation of stochastic SIRS models in MATLAB (R2019b) using both the SSA method and the modified τ-leaping. The modified τ-leaping method is used in near-critical cases where the convergence is slower and the computation time for large N is very long. In general, we find that it is feasible to simulate up to $N = 10^5$ for very near-critical cases. For comparison purposes, for each result using the SSA method, we always present the counterpart result from the modified τ-leaping simulation.

The modified τ-leaping method is an approximation to the SSA method. As previously explained, the discrepancy between the SSA and the τ-leaping can be tuned through n_c. We set the parameters of the modified τ-leaping to be $n_c = 200, \epsilon = 0.02$.

The value of $t_{\text{SIRS}}\,(m \to n)$ defined in Lemma 6.5 is obtained entirely numerically. Firstly, we obtain $v(y)$ using MATLAB ODE solver `ode45`, which is based on an explicit Runge-Kutta (4,5) formula named 'the Dormand-Prince pair'. Secondly, we use MATLAB `trapz` to perform trapezoidal numerical integration following (6.6) to obtain $t_{\text{SIRS}}\,(m \to n)$. The step size in our numerical integration is determined by the `ode45` solver. We find that in MATLAB, $t_{\text{SIRS}}\,(m \to n)$ with $\gamma_o = 0$ converges faster than $t^N_{\text{SIR}}\,(\theta(m) \to \theta(n))$ as defined in (6.3). This is particularly obvious when $\theta(m)$ is small. This is why we use the former when we plot the conclusion of Theorem 6.3.

For each case, determined by $(\lambda_o(N), \gamma_o(N))$ and $(I_0(N), R_0(N))$, we run 700 simulations for a set of N of different orders. For each case, the results are presented in three sub-figures. The time axes presented are always scaled according to the scaling of the asymptotic distribution.

The lines representing different N are colour-coded by a gradient from dark red to yellow, where the closer to yellow, the larger the corresponding value of N. In all of our figures, the dashed lines represent the simulation done by the modified τ-leaping method and the solid lines represent the simulation done by the SSA method.

For each N, we randomly choose a simulation and present its sample paths as follows: in each figure, we present in sub-figure (a) the log-scaled sample paths

$\log_N(I_t^N)$ (in the thicker lines) and $\log_N(R_t^N)$ (in the thinner lines) over the scaled time.

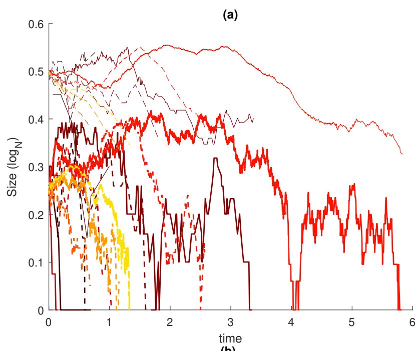
In sub-figure (b), we present the histogram of extinction times for different N, normalised so that the sum of the bar areas is less than or equal to 1 (i.e., sub-figure (b) is a simulation for the probability density function of the extinction time). Sub-figure (c) presents the histogram of extinction times for different N, normalised so that the height of the last bar is less than or equal to 1 (i.e., sub-figure (c) is a simulation for the cumulative distribution function of the extinction time). In (c), the blue line represents the asymptotic distribution we have derived through analysis, and in (b) the blue line represents the first-order derivative of the asymptotic distribution function.

7.3 Results

The figures below each present an example of one of the cases we analysed. We can see that all our asymptotic results provide fairly good approximations.

Consistent with intuition, the strongly subcritical cases converge faster than near-critical cases and the theoretical result reflects the simulation data well for N as small as 10^5.

For near-critical cases, the use of τ-leaping method significantly reduced the running time. The discrepancy between the exact method (SSA) and τ-leaping also appears to be small enough to justify using the latter.



Case 1.1 ($\lambda_o(N) < 1$)

$$N = 10^2, 10^3, \cdots, 10^6, 10^8, 10^{10}, 10^{12},$$
$$\lambda_o = 1 - N^{-1/2},\ \gamma_o = N^{-1/6},$$
$$I_0 = N^{1/4},\ R_0 = N^{1/2}.$$

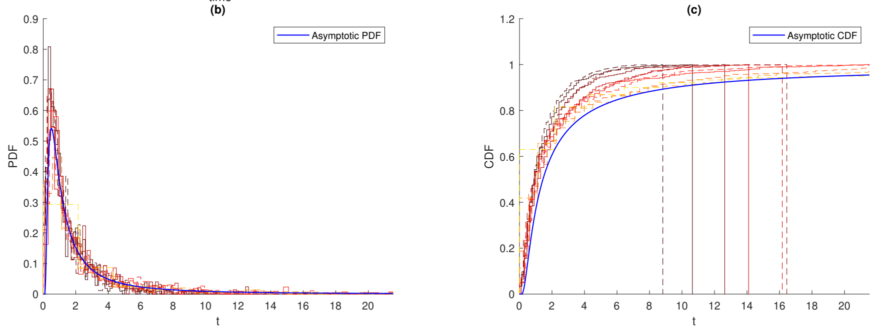
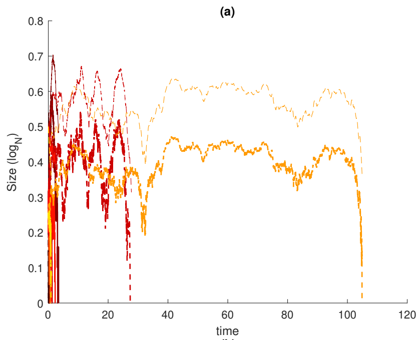


Figure 7.1: Verification of Case 1.1 ($\lambda_o(N) < 1$), Theorem 4.6



Case 1.1 ($\lambda_o(N) > 1$)

$$N = 10^2, 10^3, \cdots, 10^6, 10^8, 10^{10}, 10^{12},$$
$$\lambda_o = 1 + N^{-1/2},\ \gamma_o = N^{-1/6},$$
$$I_0 = N^{1/4},\ R_0 = N^{1/2}.$$

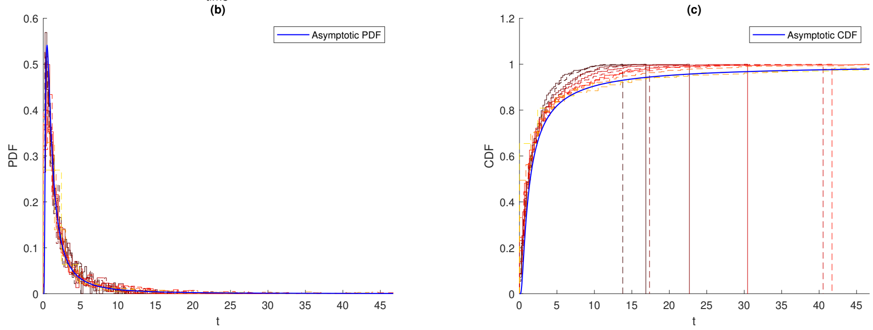
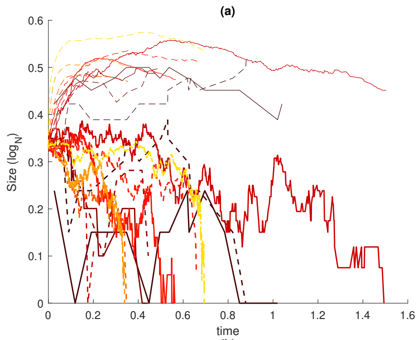


Figure 7.2: Verification of Case 1.1 ($\lambda_o(N) > 1$), Theorem 4.6



Case 1.2

$$N = 10^2, 10^3, \cdots, 10^6, 10^8, 10^{10}, 10^{12},$$
$$\lambda_o = 1 - N^{-1/3},\ \gamma_o = N^{-1/4},$$
$$I_0 = N^{1/3},\ R_0 = N^{1/3}.$$

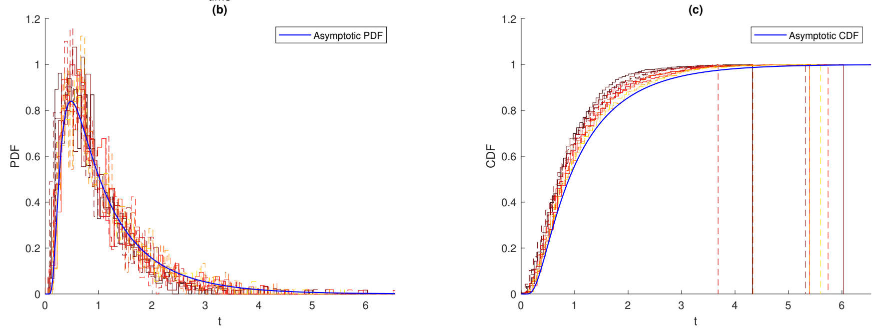
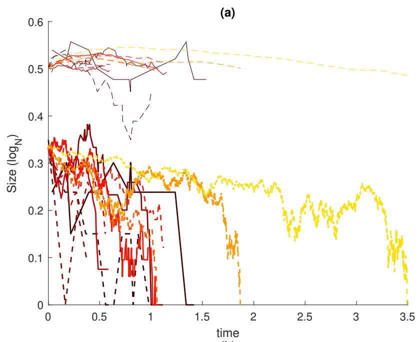


Figure 7.3: Verification of Case 1.2, Theorem 4.6



Case 1.3

$$N = 10^2, 10^3, \cdots, 10^6, 10^8, 10^{10}, 10^{12},$$
$$1 - \lambda_o = N^{-1/4},\ \gamma_o = N^{-1/4},$$
$$I_0 = N^{1/3},\ R_0 = N^{1/2}.$$

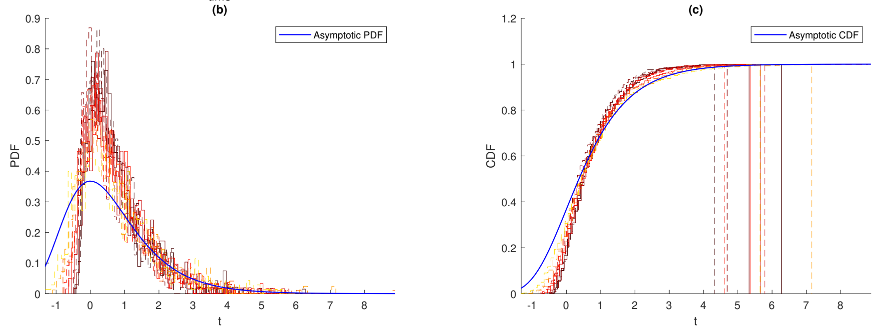
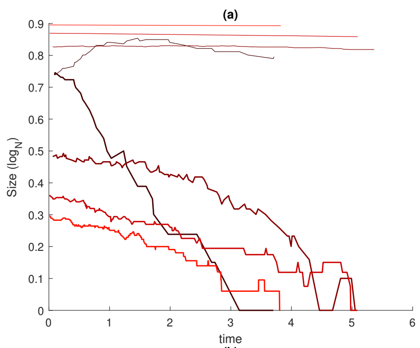


Figure 7.4: Verification of Case 1.3, Theorem 4.6



Case 2.1

$$N = 10^2, 10^3, \cdots, 10^5,$$
$$\lambda_o = 0.8 - N^{-1/2},\ \gamma_o = N^{-5/12},$$
$$I_0 = 30,\ R_0 = 0.3N.$$

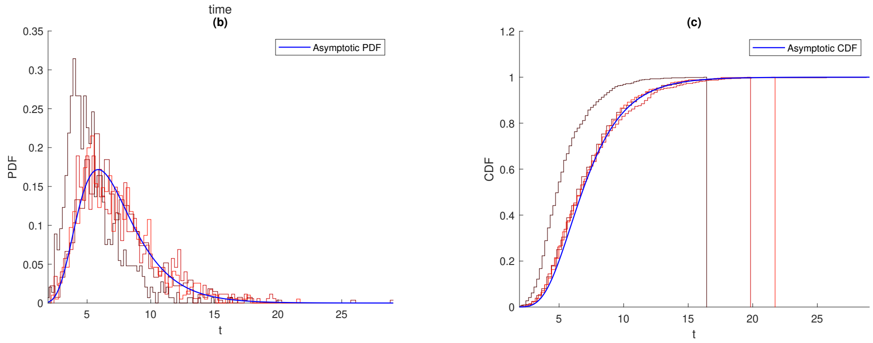
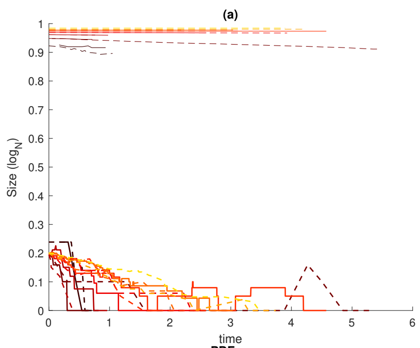


Figure 7.5: Verification of Case 2.1, Theorem 4.6



Case 2.2

$$N = 10^2, 10^3, \cdots, 10^8, 10^{10},$$
$$\lambda_o = 1 - N^{-1/4},\ \gamma_o = N^{-1/2},$$
$$I_0 = N^{1/5},\ R_0 = 0.7N.$$

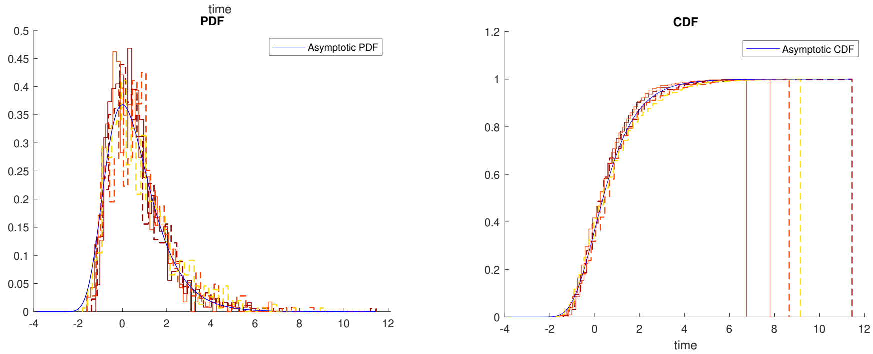
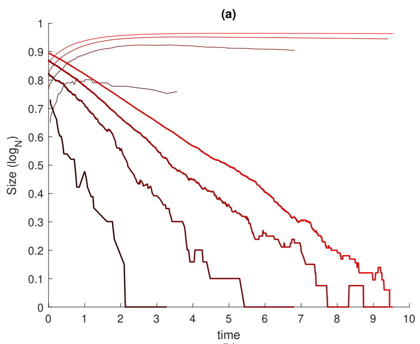


Figure 7.6: Verification of Case 2.2, Theorem 4.6



$$N = 10^2, 10^3, \cdots, 10^5,$$
$$\lambda_o = 1 - N^{-1/4},\ \gamma_o = N^{-1/2},$$
$$I_0 = 0.3N,\ R_0 = 0.2N.$$

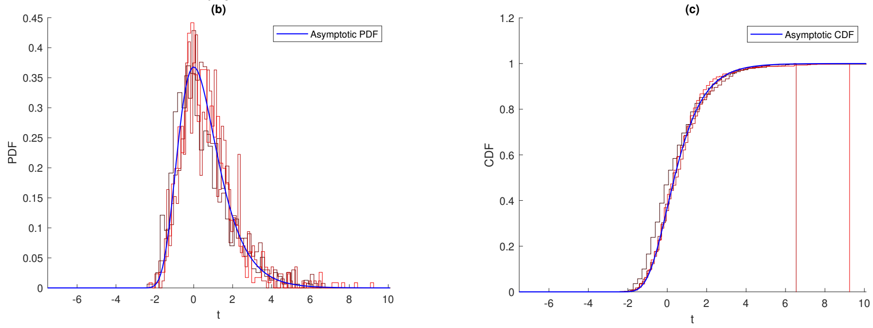
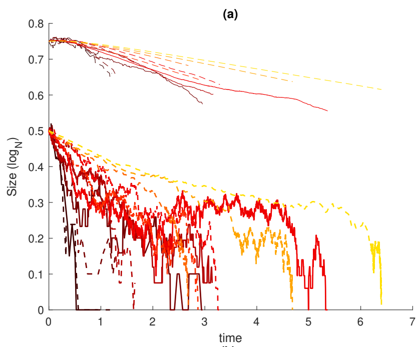


Figure 7.7: Verification of Theorem 6.3



$$N = 10^2, 10^3, \cdots, 10^6, 10^8, 10^{10}, 10^{12},$$
$$\lambda_o = 1 - N^{-1/4},\ \gamma_o = 0.7N^{-1/4},$$
$$I_0 = N^{1/2},\ R_0 = N^{3/4}.$$

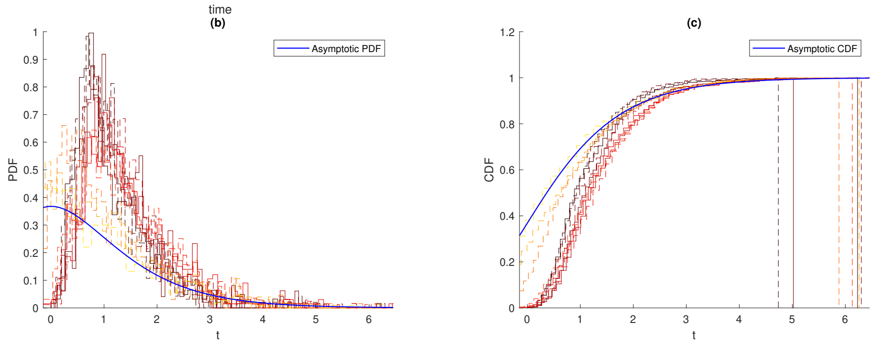
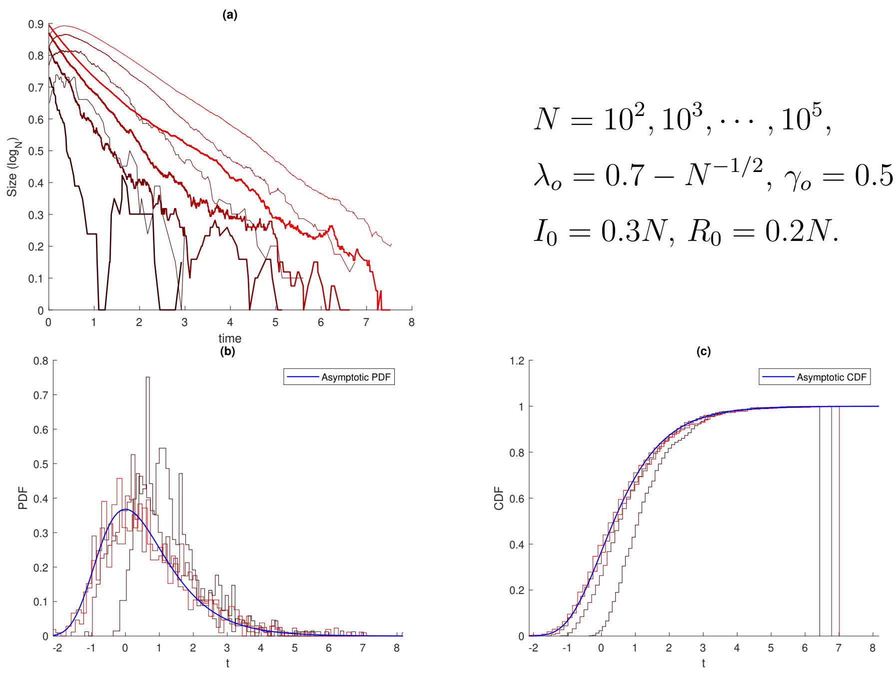


Figure 7.8: Verification of Theorem 6.7



$$N = 10^2, 10^3, \cdots, 10^5,$$
$$\lambda_o = 0.7 - N^{-1/2},\ \gamma_o = 0.5 + N^{-1/2},$$
$$I_0 = 0.3N,\ R_0 = 0.2N.$$

Figure 7.9: Verification of Theorem 6.12

Appendix A

Appendix

A.1 The elapsed time of the deterministic SIR model

Lemma A.1. *For $b > 0$,*

$$\int_0^b \log y^{-1} dy = b \log b^{-1} + b.$$

Proof. For any $b > 0$,

$$\begin{aligned} 0 \le \lim_{a \downarrow 0} \int_a^b \log y^{-1} dy &= \lim_{a \downarrow 0} -y \log y + y \Big|_a^b = \lim_{a \downarrow 0} (b-a) \log b^{-1} + a \log(a/b) + b - a \\ &= b \log b^{-1} + b. \end{aligned} \tag{A.1}$$

□

Proof of Lemma 6.1. For each $N \in \mathbb{N}$, let $(x_1^N(t), x_2^N(t))$ be the solution to:

$$\begin{aligned} &\frac{dx_1}{dt} = -\lambda_o x_1 x_2, \\ &\frac{dx_2}{dt} = \lambda_o x_1 x_2 - x_2, \\ &(x_1(0), x_2(0)) = (x_0^N, y_0^N) \in [0,1) \times (0,1], \end{aligned}$$

then for $a = a(N)$, $\theta^N(a) = x_1^N \circ (x_2^N)^{-1}(a)$ is the solution of the following:

$$x_0^N + y_0^N + \frac{\log\left(\theta^N/x_0^N\right)}{\lambda_o} - \theta^N = a. \tag{A.2}$$

From the first equation of the ODE system above,

$$t_{\text{SIR}}^N\left(a \to b\right)(N) = \int_b^a \lambda_o^{-1} u^{-1} \left(x_0^N + y_0^N + \frac{\log\left(u/x_0^N\right)}{\lambda_o} - u\right)^{-1} du. \tag{A.3}$$

Take the derivative with respect to a on both sides of (A.2) and we have

$$\frac{d\theta^N}{da} = \frac{\theta^N(a)}{\lambda_o^{-1} - \theta^N(a)} > 0.$$

For sufficiently large N, $a(N) \in [0, y_0^N]$, define $D\left(\theta^N(a)\right) := \lambda_o t_{\text{SIR}}^N\left(x_0^N \to \theta^N(a)\right)$, and substitute u in (A.3) for $\theta^N(a)$:

$$\begin{aligned}
D(\theta^N(a)) &= \int_{\theta^N(a)}^{x_0^N} u^{-1}\left(x_0^N + y_0^N + \frac{\log\left(u/x_0^N\right)}{\lambda_O} - u\right)^{-1} du = \int_a^{y_0^N} \theta^N(y)^{-1} y^{-1} d\theta^N(y) \\
&= \int_a^{y_0^N} \theta^N(y)^{-1} y^{-1} \frac{d\theta^N}{dy} dy = \int_a^{y_0^N} (\theta^N)^{-1} \frac{\theta^N}{\lambda_o^{-1} - \theta} d\log y \\
&= (\lambda_o^{-1} - \theta^N(y))^{-1} \log y \Big|_a^{y_0^N} - \int_a^{y_0^N} \log y \, d(\lambda_o^{-1} - \theta^N)^{-1} \\
&= (\lambda_o^{-1} - x_0^N)^{-1} \log y_0^N + (\lambda_o^{-1} - \theta^N(a))^{-1} \log a^{-1} \\
&+ \int_a^{y_0^N} (\lambda_o^{-1} - \theta^N(y))^{-3} \theta^N(y) \log y^{-1} dy.
\end{aligned}$$

The last term $\int_a^{y_0^N} (\lambda_o^{-1} - \theta(y))^{-3} \theta(y) \log y^{-1} dy$ is non-negative and converges to a finite constant, since

$$\theta^{*,N}(\lambda_o^{-1} - \theta^*)^{-3} \leq \theta^N(y)(\lambda_o^{-1} - \theta^N(y))^{-3} \leq x_0^N(\lambda_o^{-1} - x_0^N)^{-3}, \text{for all } y \in [0, y_0^N],$$

where $\theta^{*,N} = \theta^*(x_0^N, y_0^N; \lambda_o)$, and both ends of the inequality above have a positive limit as $N \to \infty$.

By Lemma A.1, as $N \to \infty$, $y_0^N \to i_0$,

$$\int_a^{i_0} (\lambda_o^{-1} - \theta^N(y))^{-3} \theta^N(y) \log y^{-1} dy$$

increases and is bounded above, and therefore converges. In other words, let $\lambda_{lim} := \lim_{N\to\infty} \lambda_o$ and $\theta^*_{lim} := \lim_{N\to\infty} \theta^{*,N}$, then

$$(\lambda_{lim}^{-1} - \theta^*_{lim}) \lim_{N\to\infty} \left(D(\theta(a)) - (\lambda_o^{-1} - \theta^N(a))^{-1} \log a^{-1} \right)$$

exists, which we denote as k_{SIR}.

Hence, as $a \downarrow 0$,

$$\begin{aligned}
& - \log a^{-1} + (1 - \lambda_o \theta^{*,N}) t^N_{\text{SIR}} \left(x_0^N \to \theta^N(a) \right) \\
=& (\lambda_o^{-1} - \theta^{*,N}) \left[D\left(\theta^N(a)\right) - \left(\lambda_o^{-1} - \theta^N(a) \right)^{-1} \log a^{-1} \right] + \frac{\theta^N(a) - \theta^{*,N}}{\lambda_o^{-1} - \theta^N(a)} \log a^{-1} \\
=& \frac{\lambda_o^{-1} - \theta^{*,N}}{\lambda_{lim}^{-1} - \theta^*_{lim}} (\lambda_{lim}^{-1} - \theta^*_{lim}) \left[D\left(\theta^N(a)\right) - \left(\lambda_o^{-1} - \theta^N(a) \right)^{-1} \log a^{-1} \right] \\
& + \frac{\lambda_o^{-1} \log\left(\theta^N(a)/\theta^{*,N}\right) - a}{\lambda_o^{-1} - \theta^N(a)} \log a^{-1} \\
\to & k_{\text{SIR}}.
\end{aligned}$$

where the second equality is due to (A.2). □

A.2 The elapsed time of the deterministic SIRS model

Proof of Lemma 6.5. In this proof we use the notations $x_i' = dx_i/dt$, $i = 1, 2$.

Rearranging the second equation in (6.5) and differentiating by t, we have

$$\lambda_o x_1 = \frac{x_2' + x_2}{x_2}, \tag{A.4}$$

$$\lambda_o x_1' = \frac{x_2'' x_2 - (x_2')^2}{x_2^2}. \tag{A.5}$$

Insert (A.4) into the first equation in the deterministic SIRS model,

$$x_1' = -(\lambda_o x_2 + \gamma_o)x_1 + \gamma_o(1 - x_2) = -\frac{\lambda_o x_2 + \gamma_o}{\lambda_o}\frac{x_2' + x_2}{x_2} + \gamma_o(1 - x_2).$$

Together with (A.5),

$$-(\lambda_o x_2 + \gamma_o)\frac{x_2' + x_2}{x_2} + \lambda_o\gamma_o(1 - x_2) = \frac{x_2'' x_2 - (x_2')^2}{x_2^2},$$

$$x_2'' x_2 - (x_2')^2 = -x_2'(\lambda_o x_2^2 + \gamma_o x_2) - x_2^3(\lambda_o + \lambda_o\gamma_o) + x_2^2(\lambda_o\gamma_o - \gamma_o). \tag{A.6}$$

Since $x_2 : [0, \infty) \to (0, i_0]$ is an injection, we can define for $y \in (0, i_0]$,

$$v : y \mapsto \frac{y}{-x_2' \circ x_2^{-1}(y)},$$

then

$$\frac{dv}{dx_2} = -\frac{dv(x_2(t))}{dt}/\frac{dx_2}{dt} = -\frac{(x_2')^2 - x_2 x_2''}{(x_2')^3}. \tag{A.7}$$

By (A.6), we have

$$\frac{dv}{dx_2} = \frac{-x_2'(\lambda_o x_2^2 + \gamma_o x_2) - x_2^3(\lambda_o + \lambda_o\gamma_o) + x_2^2(\lambda_o\gamma_o - \gamma_o)}{(x_2')^3} \tag{A.8}$$

$$= v^3(\lambda_o + \lambda_o\gamma_o - \frac{\lambda_o\gamma_o - \gamma_o}{x_2}) - v^2(\lambda_o + \frac{\gamma_o}{x_2}). \tag{A.9}$$

Recall by definition

$$v(x_2(t)) = -\frac{x_2(t)}{dx_2}dt,$$

$$dt = \frac{v(x_2(t))}{x_2(t)}d(-x_2).$$

Then for $0 < n < m \leq i_0$,

$$t_{\text{SIRS}}(m \to n) = \int_n^m \frac{v(x_2)}{x_2}dx_2.$$

□

Lemma A.2. *Let (x_2, x_3) be the second and third components of the deterministic SIRS model (6.5) with parameters $\lambda_o \in (0,1), \gamma_o > 0$. Then the following inequalities hold:*

$$c_2 e^{-(1-\lambda_o+\lambda_o x_3(0))t} \le x_2(t) \le x_2(0)e^{-(1-\lambda_o)t}, \tag{A.10}$$

$$x_3(t) \le x_3(0) + x_2(0)\frac{e^{-(1-\lambda_o)t} - e^{-\gamma_o t}}{\lambda_o + \gamma_o - 1}, \tag{A.11}$$

$$x_3(t) \le \left(x_3(0) + x_2(0)|1 - \lambda_o - \gamma_o|^{-1}\right) e^{-\gamma_o t}. \tag{A.12}$$

for some constant c_2 depending only on $x_2(0)$, λ_o and γ_o.

Proof. It is easy to see

$$x_2(t) \le x_2(0)e^{-(1-\lambda_o)t},$$

and

$$\begin{aligned} x_3(t) = e^{-\gamma_o t}x_3(0) + \int_0^t e^{-\gamma_o(t-s)}x_2(s)ds &\le x_3(0) + e^{-\gamma_o t}x_2(0)\int_0^t e^{(\gamma_o-1+\lambda_o)s}ds \\ &= x_3(0) + x_2(0)\frac{e^{-(1-\lambda_o)t} - e^{-\gamma_o t}}{\lambda_o + \gamma_o - 1}. \end{aligned}$$

We also have

$$\begin{aligned} e^{\gamma_o t}x_3(t) = x_3(0) + \int_0^t e^{\gamma_o s}x_2(s)ds &\le x_3(0) + \int_0^t e^{-(1-\lambda_o-\gamma_o)s}x_2(0)ds \\ &\le x_3(0) + x_2(0)|1 - \lambda_o - \gamma_o|^{-1}, \end{aligned}$$

and (A.12) follows.

Applying the bound of x_3 to the second equation of the deterministic SIRS model, we have

$$\begin{aligned}
x_2(t) =& x_2(0)\exp\left\{-(1-\lambda_o)t - \lambda_o\int_0^t x_2(s)+x_3(s)ds\right\}\\
\geq& x_2(0)\exp\left\{-(1-\lambda_o)t - \lambda_o\int_0^t x_2(0)e^{-(1-\lambda_o)s} + x_3(0) + x_2(0)\frac{e^{-(1-\lambda_o)s}-e^{-\gamma_o s}}{\lambda_o+\gamma_o-1}ds\right\}\\
=& x_2(0)\exp\left\{x_2(0)\frac{\lambda_o+\gamma_o}{\lambda_o+\gamma_o-1}\left(\frac{1-e^{-(1-\lambda_o)t}}{1-\lambda_o} - \frac{1-e^{-\gamma t}}{\gamma_o(\lambda_o+\gamma_o)}\right)\right\}\\
&\cdot\exp\{-(1-\lambda_o+\lambda_o x_3(0))t\}\\
\geq& c_2\exp\{-(1-\lambda_o+\lambda_o x_3(0))t\},
\end{aligned}$$

for positive constant $c_2 = x_2(0)\exp\left\{\frac{-1}{\gamma_o(\lambda_o+\gamma_o-1)}\right\}$. □

Lemma A.3. *Let (x_2^N, x_3^N) be the second and third component of the deterministic SIRS model (6.5), and let $\phi^N : a \mapsto x_3^N \circ (x_2^N)^{-1}(a)$ map the value $x_2^N(t) = a$ to the value of $x_3^N(t)$ for all $t \geq 0$.*

For two states $n(N) < m(N) = o\left((1-\lambda_o)\gamma_o\right)$ for all $N \in \mathbb{N}$, if $\phi^N(m) = o(\gamma_o)$, then

$$(1-\lambda_o)t_{\text{SIRS}}\left(m \to n\right) - \log\frac{m}{n} = o(1), \quad N \to \infty.$$

Proof. For the ODE expression, we have

$$\frac{\mathrm{d}e^{(1-\lambda_o)t}x_2^N(t)}{\mathrm{d}t} = -\lambda_o e^{(1-\lambda_o)t}x_2^N(t)\left(x_2^N(t)+x_3^N(t)\right).$$

Consider $x_2^N(0) = m(N)$ and $x_3^N(0) = o(\gamma_o)$ and apply the bounds in Lemma A.2,

$$\begin{aligned}
&\left|e^{(1-\lambda_o)t}x_2^N(t) - x_2^N(0)\right|\\
=&\lambda_o\int_0^t e^{(1-\lambda_o)s}x_2^N(s)\left(x_2^N(s)+x_3^N(s)\right)ds\\
\leq&\lambda_o\int_0^t x_2^N(0)\left(x_2^N(0)e^{-(1-\lambda_o)s} + x_3^N(0)e^{-\gamma_o s} + x_2^N(0)|1-\lambda_o-\gamma_o|^{-1}e^{-\gamma_o s}\right)ds\\
\leq&\lambda_o x_2^N(0)\left(x_2^N(0)(1-\lambda_o)^{-1} + x_2^N(0)|1-\lambda_o-\gamma_o|^{-1}\gamma_o^{-1} + x_3^N(0)\gamma_o^{-1}\right)\\
=&x_2^N(0)o(1).
\end{aligned}$$

It follows that

$$(1-\lambda_o)t_{\text{SIRS}}\,(m\to n)=\log\frac{m(1+o(1))}{n}=\log\frac{m}{n}+o(1).$$

□

Proof of Lemma 6.11. To simplify the notation, we rename the variables λ_{lim} and γ_{lim} as λ and γ respectively.

Firstly, given $(x_2^N(0),x_3^N(0))\to(x_2(0),x_3(0))\in(0,1]\times[0,1]$, according to (A.10) and (A.12),

$$\begin{aligned}\frac{x_3^N(t)}{(x_2^N(t))^{\gamma_o/(2+2\gamma_{lim})}}&\le\frac{\left(x_3^N(0)+x_2^N(0)|1-\lambda_o-\gamma_o|^{-1}\right)e^{-\gamma_o t}}{c_2^{\gamma_o/(2+2\gamma_{lim})}\exp\{-2(1+\gamma_{lim})t\cdot\gamma_o/(2+2\gamma_{lim})\}}\\ =&\frac{x_3^N(0)+x_2^N(0)|1-\lambda_o-\gamma_o|^{-1}}{c_2^{\gamma_o/(2+2\gamma_{lim})}}=O(1),\end{aligned}\tag{A.13}$$

Next, we prove that for any $a=a(N)\downarrow 0$,

$$\lim_{N\to\infty}\left(a+\phi^N(a)\right)\log a^{-1}=0,\tag{A.14}$$

where ϕ^N is defined in Lemma A.3. By (A.13), there exists constant $b_1>0$ independent of N such that

$$\begin{aligned}\left(a+\phi^N(a)\right)\log a^{-1}&=\frac{\phi^N(a)}{a^{\gamma_o/(2+2\gamma_{lim})}}\frac{\log a^{-1}}{a^{-\gamma_o/(2+2\gamma_{lim})}}+\frac{\log a^{-1}}{a^{-1}}\\&\le b_1\frac{\log a^{-1}}{a^{-\gamma_o/(2+2\gamma_{lim})}}+\frac{\log a^{-1}}{a^{-1}}.\end{aligned}$$

Now we are ready to deal with the main problem.

For each sufficiently large N, applying integration by parts, we have

$$\begin{aligned}&(1-\lambda_o)t_{\text{SIRS}}\left(x_2^N(0)\to a\right)=(1-\lambda_o)\int_a^{x_2^N(0)}\frac{v(y)}{y}dy\\=&(1-\lambda_o)v(x_2^N(0))\log x_2^N(0)-(1-\lambda_o)v(a)\log a-(1-\lambda_o)\int_a^{x_2^N(0)}\log y\frac{dv}{dy}dy.\end{aligned}$$

The first term above is independent of a, and the second term is

$$
\begin{aligned}
&(1-\lambda_o)v(a)\log a && \text{(A.15)}\\
=&-\frac{(1-\lambda_o)}{1-\lambda_o+\lambda_o(a+\phi^N(a))}\log a^{-1} && \text{(A.16)}\\
=&-\log a^{-1}+\lambda_o\frac{a+\phi^N(a)}{1-\lambda_o+\lambda_o(a+\phi^N(a))}\log a^{-1}=-\log a^{-1}+o(1),
\end{aligned}
$$

since by (A.14) as $N\to\infty$,

$$
\lambda_o\frac{a+\phi^N(a)}{1-\lambda_o+\lambda_o(a+\phi^N(a))}\log a^{-1}\le\frac{\lambda_o}{1-\lambda_o}(a+\phi^N(a))\log a^{-1}=o(1).
$$

What remains to prove is that the third term above is $O(1)$.

$$
\begin{aligned}
&-\int_a^{x_2^N(0)}\log y\frac{dv}{dy}dy && \text{(A.17)}\\
=&\lambda_o\int_a^{x_2^N(0)}(1-\lambda_o+\lambda_o(y+\phi^N(y)))^{-3}\log y^{-1}\frac{\lambda_o(1-y-\phi^N(y))y-\gamma_o\phi^N(y)}{-y}dy\\
\le&\lambda_o^2(1-\lambda_o)^{-3}\int_a^{x_2^N(0)}\log y^{-1}dy+(1-\lambda_o)^{-3}\lambda_o\gamma_o\int_a^{x_2^N(0)}\log y^{-1}\frac{\phi^N(y)}{y}dy. && \text{(A.18)}
\end{aligned}
$$

By Lemma A.1, the first term of (A.18) is of order $O(1)$.

For the second term above,

$$
\begin{aligned}
\int_a^{x_2^N(0)}\frac{\phi^N(y)}{y}\log y^{-1}dy&=\int_a^{x_2^N(0)}\frac{\phi^N(y)}{y^{\gamma_o/(2+2\gamma_{lim})}}\frac{\log y^{-1}}{y^{1-\gamma_o/(2+2\gamma_{lim})}}dy\\
&\le b_1\int_a^{x_2^N(0)}\frac{\log y^{-1}}{y^{1-\gamma_o/(2+2\gamma_{lim})}}dy.
\end{aligned}
$$

The last line above is of order $O(1)$ because for any $0<k<1$, $n>0$

$$
\lim_{m\downarrow 0}\int_m^n\frac{\log(y^{-1})}{y^k}dy=\lim_{m\downarrow 0}-\frac{y^{1-k}\log y}{1-k}+\frac{y^{1-k}}{(1-k)^2}\Big|_{y=m}^{n}=-\frac{n^{1-k}\log n}{1-k}+\frac{n^{1-k}}{(1-k)^2},
$$

and is bounded.

Hence we can see that indeed $\int_a^{x_2^N(0)} \log y \frac{dv}{dy} dy = O(1)$ and thus we can define

$$k_{\text{SIRS}} := \lim_{N\to\infty} (1-\lambda_o) \left(t_{\text{SIRS}} \left(x_2^N(0) \to a\right) - v(a) \log a^{-1}\right).$$

From (A.16), we have

$$k_{\text{SIRS}} = \lim_{N\to\infty} (1-\lambda_o) t_{\text{SIRS}} \left(x_2^N(0) \to a\right) - \log a^{-1},$$

which is useful when we state our main result. □

Lemma A.4 (Taking the limit of the scaling)**.** *Let $a(N)$, $b(N)$ be functions of N tending to either a positive constant or infinity and $\epsilon_a(N), \epsilon_b(N) \to 0$ as $N \to \infty$. A sequence of continuous random variables $\{T^N\}_{N\in\mathbb{N}}$ has the asymptotic distribution*

$$\mathbb{P}\left[(a(N)+\epsilon_a(N))T^N \leq w + b(N) + \epsilon_b(N)\right] \to F(w), \quad N \to \infty.$$

Then

$$\mathbb{P}\left[a(N)T^N \leq w + b(N)\right] \to F(w)$$

if and only if $b\epsilon_a/a \to 0$.

Proof.

$$\mathbb{P}\left[aT^N \leq w + b\right] = \mathbb{P}\left[(a+\epsilon_a)T^N - b - \epsilon_b \leq w + \frac{b\epsilon_a}{a} + \frac{w\epsilon_a}{a} - \epsilon_b\right].$$

□

A.3 Estimation of $\mathcal{S}_t f$ for special functions f

Definition A.5 (Dirac δ-function)**.** The Dirac δ-function has the following heuristic characteristics:

$$\delta_a(x) = \begin{cases} \infty & x \neq a, \\ 0 & x = a, \end{cases}$$

and

$$\int_{-\infty}^{\infty} \delta_a(x)dx = 1.$$

It can be defined as the pointwise limit of a sequence of normal distributions $\{\delta_0^\epsilon(x)\}$ as $\epsilon \downarrow 0$, defined as

$$\delta_0^\epsilon(x) = \frac{1}{\sqrt{2\pi}\epsilon} e^{-\frac{x^2}{2\epsilon^2}}.$$

The Dirac δ-function has the property

$$\int_{-\infty}^{\infty} f(x)\delta_0^{(n)}(x)dx = (-1)^n f^{(n)}(0), \quad f \in C_0^\infty(\mathbb{R}).$$

We are able to obtain the exact value of $\mathcal{S}_t f$ for a set of special functions f.

Lemma A.6 (Special integrals with g as kernel)**.** *Let*

$$R_k(t,s,u) := \mathcal{S}_t \left(u^k \exp\{-u/s\}\right) = \int_0^\infty g(t,u;m)m^k \exp\{-m/s\}\, dm.$$

We have

$$R_0(t,s,u) = \exp\left\{-\frac{u}{t+s}\right\} - \exp\left\{-\frac{u}{t}\right\}, \tag{A.19}$$

and

$$R_k(t,s,u) = \sum_{n=1}^{k} u^n e^{-u/(s+t)} \binom{k}{n} \frac{(k-1)!}{(n-1)!}(1+t/s)^{-k-n} t^{k-n}. \tag{A.20}$$

Proof of Lemma A.6 . Let $X_t \sim BESQ^0(u)$ and recall the notations in Definition 5.5, we have that the Laplace transform of $g(t,u;m)$ is

$$\begin{aligned}
(Lg)(\rho) &= \int_0^\infty g(t,u;m)e^{-\rho m}\, dm \\
&= \exp\left\{-\frac{\rho}{1+\rho t}u\right\} \int_0^\infty g\left(t(1+\rho t)^{-1}, u(1+\rho t)^{-2}; m\right)\, dm \\
&= \exp\left\{-\frac{\rho u}{1+\rho t}\right\} - \exp\left\{-\frac{u}{t}\right\},
\end{aligned}$$

and

$$R_0(s_1, s_2, u) = Lg(1/s_2).$$

For the second result, we consider the inverse Laplace transform of $f_k(t^{-1}) = m^k e^{-m/t}$, which is $\left(L^{-1} f_k\right)(\rho) = \delta^{(k)}(\rho - t^{-1})$. Let

$$\begin{aligned} R_k(t, s, u) &= \int_0^\infty g(t, u; m) f_k(s^{-1})\, dm \\ =&\int_0^\infty \left(\exp\left\{-\frac{\rho}{1+\rho s_1} u\right\} - \exp\left\{-\frac{u}{t}\right\}\right) \delta^{(k)}(\rho - s^{-1})\, d\lambda \\ =&(-1)^k \frac{d^k (Lg)}{d\rho^k}(s^{-1}). \end{aligned} \tag{A.21}$$

To evaluate $\frac{d^k(Lg)}{d\rho^k}$, we use Faà di Bruno's formula.

Let $h(x) = e^{-xu}$, $l(\rho) = \frac{\rho}{1+\rho t}$, then

$$h^{(k)}(x) = (-u)^k e^{-ux}, \quad l^{(k)}(\rho) = (-1)^{k-1} k! (1+\rho t)^{-(k+1)} t^{k-1}.$$

By Faà di Bruno's formula [59],

$$\begin{aligned} (-1)^k \frac{d^k (Rg)}{d\rho^k}(\rho) &= (-1)^k \frac{d^k h(l(\rho))}{d\rho^k} \\ &= (-1)^k \sum_{n=1}^{k} h^{(n)}(l(\rho)) B_{k,n}\left(l^{(1)}(\rho), \cdots, l^{(k-n+1)}(\rho)\right), \end{aligned} \tag{A.22}$$

where $B_{k,n}$ denotes the partial or incomplete exponential Bell polynomials. It is known that

$$\begin{aligned} &B_{k,n}(x_1, x_2, \ldots, x_{k-n+1}) \\ =&\sum \frac{k!}{m_1! m_2! \cdots m_{k-n+1}!} \left(\frac{x_1}{1!}\right)^{m_1} \left(\frac{x_2}{2!}\right)^{m_2} \cdots \left(\frac{x_{k-n+1}}{(k-n+1)!}\right)^{m_{k-n+1}}, \end{aligned}$$

where the sum is taken over all sequences $\mathbf{m}_{k-n+1} := \{m_1, \cdots, m_{k-n+1}\}$ of non-negative integers such that the following two conditions are satisfied:

$$m_1 + m_2 + \cdots + m_{k-n+1} = n,$$
$$m_1 + 2m_2 + 3m_3 + \cdots + (k-n+1)m_{k-n+1} = k.$$

Continue with (A.22),

$$(-1)^k \frac{d^k(Lg)}{d\rho^k}(\rho) = (-1)^k \sum_{n=1}^{k} (-u)^n e^{-ul(\rho)} \sum_{\mathbf{m}_{k-n+1}} \left[\frac{k!}{m_1! \cdots m_{k-n+1}!} \left(\frac{1!(1+\rho t)^{-2}}{1!} \right)^{m_1} \right.$$
$$\left. \cdots \left(\frac{(-1)^{k-n}(k-n+1)!(1+\rho t)^{-(k-n+2)} t^{k-n}}{(k-n+1)!} \right)^{m_{k-n+1}} \right]$$
$$= (-1)^k \sum_{n=1}^{k} (-u)^n e^{-ul(\rho)} \sum_{\mathbf{m}_{k-n+1}} \frac{k!}{m_1! ... m_{k-n+1}!} (-1)^{k-n} (1+\rho t)^{-k-n} t^{k-n}$$
$$= \sum_{n=1}^{k} u^n e^{-ul(\rho)} \sum_{\mathbf{m}_{k-n+1}} \frac{k!}{m_1! ... m_{k-n+1}!} (1+\rho t)^{-k-n} t^{k-n}$$

By p.135, [60], we see that

$$\sum_{\mathbf{m}_{k-n+1}} \frac{k!}{m_1! ... m_{k-n+1}!} = B_{k,n}(1!, 2!, \cdots, (k-n+1)!) = \binom{k-1}{n-1} \frac{k!}{n!} = \binom{k}{n} \frac{(k-1)!}{(n-1)!}.$$

It follows that

$$R_k(t,s,u) = \sum_{n=1}^{k} u^n e^{-u/(s+t)} \binom{k}{n} \frac{(k-1)!}{(n-1)!} (1+t/s)^{-k-n} t^{k-n}.$$

□

The following tail estimate based on the first and second moments of X turns out to be useful.

Recall our notation for the squared Bessel process X_t generated by the semigroup $\mathcal{S}_t$, with $X_0 = u$. We have from Lemma A.6,

$$\mathbb{E}[X_t] = u, \ \mathbb{E}\left[X_t^2\right] = 2ut + u^2.$$

By Chebyshev's inequality, for $u < M < K$,

$$\mathbb{P}\left[X_t > K\right] \leq \mathbb{P}\left[|X_t - \mathbb{E}\left[X_t\right]| > K - u\right] \leq \frac{\mathrm{Var}\left(X_t\right)}{(K-u)^2} = \frac{2u}{(K-u)^2}t$$
$$< \frac{2M}{(K-M)^2}t. \tag{A.23}$$

It follows that

$$\begin{aligned}\mathbb{E}_u\left[X_t\mathbf{1}_{X_t\in[A,B]}\right] &= \int_0^\infty \mathbb{P}\left[X_t\mathbf{1}_{X_t\in[A,B]} > x\right]dx \\ &= \int_0^A \mathbb{P}\left[X_t \in [A,B]\right]dx + \int_A^B \mathbb{P}\left[X_t \in [x,B]\right]dx \\ &\leq A\frac{2ut}{(A-u)^2} + \int_A^B \frac{2ut}{(x-u)^2}dx \\ &= 2ut\left(\frac{A}{(A-u)^2} + \frac{1}{A-u} - \frac{1}{B-u}\right). \end{aligned} \tag{A.24}$$

Now we prove the second statement.

In Definition 5.6, we state the properties that the power expansion of the modified Bessel functions of the first kind $I_\alpha(x)$, $\alpha \in \mathbb{N}$, is

$$I_\alpha(x) = \sum_{j=0}^{\infty} \frac{(x/2)^{2j+\alpha}}{j!(j+\alpha)!}.$$

The asymptotic expansion of $I_\alpha(x)$ at infinity is

$$I_\alpha(x) = \frac{e^x}{\sqrt{2\pi x}}(1 + O(x^{-1})), \quad \alpha \in \mathbb{N},\ x \to \infty.$$

By the Monotone convergence theorem, for each $u \in \mathbb{R}_+$ and $t > 0$,

$$
\begin{aligned}
\mathbb{P}\left[X_t < x\right] = \int_0^x g(t,u;m)dm &= \lim_{n\to\infty} \int_0^x \frac{1}{t} u^{1/2} m^{-1/2} e^{-(u+m)/t} \frac{\sqrt{um}}{t} \sum_{j=0}^{n} \frac{(um/t^2)^j}{j!(j+1)!} dm \\
&= \lim_{n\to\infty} t^{-2} u e^{-u/t} \sum_{j=0}^{n} \frac{(u/t^2)^j}{j!(j+1)!} \int_0^x e^{-m/t} m^j \, dm \\
&\leq \lim_{n\to\infty} e^{-u/t} \sum_{j=0}^{n} \frac{(ux/t^2)^{j+1}}{(j+1)!(j+1)!} \\
&= e^{-u/t} \left(I_0\left(\frac{2\sqrt{ux}}{t}\right) - 1 \right).
\end{aligned}
$$

For given $u > 0$ and $x < u/4$, as $t \downarrow 0$,

$$
e^{-u/t} \left(I_0\left(\frac{2\sqrt{ux}}{t}\right) - 1 \right) \sim \frac{1}{\sqrt{2\pi}} \exp\left\{ -u/t + \frac{2\sqrt{ux}}{t} \right\} \left(\frac{2\sqrt{ux}}{t} \right)^{-1/2} \to 0. \quad \text{(A.25)}
$$

For given $u, t > 0$,

$$
e^{-u/t} \left(I_0\left(\frac{2\sqrt{ux}}{t}\right) - 1 \right) \sim x \frac{u}{t^2} e^{-u/t}. \quad \text{(A.26)}
$$

Using a similar method, we also have

$$
\begin{aligned}
&\int_0^x g(t,u;m)(1-m/x)dm \\
&= \lim_{n\to\infty} \int_0^x \frac{1}{t} u^{1/2} m^{-1/2} e^{-(u+m)/t} \frac{\sqrt{um}}{t} \sum_{j=0}^{n} \frac{(um/t^2)^j}{j!(j+1)!} (1-m/x) dm \\
&= \lim_{n\to\infty} t^{-2} u e^{-u/t} \sum_{j=0}^{n} \frac{(u/t^2)^j}{j!(j+1)!} \int_0^x e^{-m/t} m^j (1-m/x) \, dm \\
&\leq \lim_{n\to\infty} e^{-u/t} \sum_{j=0}^{n} \frac{(ux/t^2)^{j+1}}{(j+1)!(j+1)!} - \lim_{n\to\infty} e^{-u/t} \sum_{j=0}^{n} \frac{(ux/t^2)^{j+1}}{j!(j+2)!} \\
&= e^{-u/t} \left(I_0\left(\frac{2\sqrt{ux}}{t}\right) - 1 - I_2\left(\frac{2\sqrt{ux}}{t}\right) \right) = e^{-u/t} \left(I_1\left(\frac{2\sqrt{ux}}{t}\right) / \left(\frac{\sqrt{ux}}{t}\right) - 1 \right) \\
&< e^{-u/t} I_2\left(\frac{2\sqrt{ux}}{t}\right). \quad \text{(A.27)}
\end{aligned}
$$

References

[1] A. D. Barbour, "The duration of the closed stochastic epidemic," *Biometrika*, vol. 62, no. 2, pp. 477–482, 1975. 2, 4, 9, 26, 27, 46, 113

[2] R. J. Kryscio and C. Lefévre, "On the extinction of the S-I-S stochastic logistic epidemic," *Journal of Applied Probability*, vol. 26, no. 4, pp. 685–694, 1989. 2, 26, 27, 46

[3] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*. Oxford university press, 1992. 2, 18, 26, 27

[4] I. Nåsell, "The quasi-stationary distribution of the closed endemic sis model," *Advances in Applied Probability*, vol. 28, no. 3, pp. 895–932, 1996. 2, 24, 25, 27, 70, 71

[5] C. R. Doering, K. V. Sargsyan, and L. M. Sander, "Extinction times for birth-death processes: Exact results, continuum asymptotics, and the failure of the Fokker–Planck approximation," *Multiscale Modeling & Simulation*, vol. 3, no. 2, pp. 283–299, 2005. 2, 25, 27, 71

[6] E. Foxall, "Extinction time of the logistic process," *ArXiv e-prints*, 2018, 1805.08339. 2, 4, 7, 9, 25, 26, 46, 47, 71, 72, 83, 111

[7] F. Kühnemund, *Bi-continuous semigroups on spaces with two topologies: theory and applications*. PhD thesis, Eberhard-Karls-Universität Tübingen, 2001. 3, 41, 43, 44, 72, 91

[8] G. Brightwell, T. House, and M. Luczak, "Extinction times in the subcritical stochastic SIS logistic epidemic," *Journal of Mathematical Biology*, Jan 2018. 4, 9, 26, 27, 28, 46, 111, 120

[9] G. Brunick, "Uniqueness in law for a class of degenerate diffusions with continuous covariance," *Probability Theory and Related Fields*, vol. 155, pp. 265–302, Feb 2013. 7, 77

[10] R. G. Pinsky, *Positive Harmonic Functions and Diffusion.* Cambridge Studies in Advanced Mathematics, Cambridge University Press, 1995. 7, 78, 79

[11] A. A. Albanese and E. Mangino, "Trotter–kato theorems for bi-continuous semigroups and applications to feller semigroups," *Journal of Mathematical Analysis and Applications*, vol. 289, no. 2, pp. 477 – 492, 2004. 8, 44

[12] R. W. R. Darling and J. R. Norris, "Differential equation approximations for markov chains," *Probab. Surveys*, vol. 5, pp. 37–79, 2008. 13, 31, 34, 111, 112

[13] J. R. Norris, *Markov chains.* Cambridge university press, 1998. 14, 51

[14] E. Vynnycky and R. G. White, *An introduction to infectious disease modelling.* Oxford university press, 2010. 15

[15] M. J. Keeling and P. Rohani, *Modeling infectious diseases in humans and animals.* Princeton University Press, 2011. 16, 23

[16] S. N. Ethier and T. G. Kurtz, *Markov processes: Characterization and Convergence*, vol. 282. John Wiley & Sons, 2009. 16, 34, 37, 95, 105, 111

[17] K. E. Nelson and C. M. Williams, *Infectious Disease Epidemiology: Theory and Practice.* Jones & Bartlett Learning, 2014. 17

[18] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Advances in Computational Mathematics*, vol. 5, no. 1, pp. 329–359, 1996. 20

[19] A. Korobeinikov and G. C. Wake, "Lyapunov functions and global stability for SIR, SIRS, and SIS epidemiological models," *Applied Mathematics Letters*, vol. 15, no. 8, pp. 955–960, 2002. 22

[20] S. Sastry, *Nonlinear Systems: Analysis, Stability, and Control*, vol. 10. Springer Science & Business Media, 2013. 22, 23

[21] K. Broadfoot, H. Charlesworth, N. Foster, J. Hilton, M. Keeling, and S. Gubbins, "The critical community size and extinctions for livestock diseases." 24

[22] J. N. Darroch and E. Seneta, "On quasi-stationary distributions in absorbing continuous-time finite markov chains," *Journal of Applied Probability*, vol. 4, no. 1, pp. 192–196, 1967. 24

[23] T. Britton, T. House, A. L. Lloyd, D. Mollison, S. Riley, and P. Trapman, "Five challenges for stochastic epidemic models involving global transmission," *Epidemics*, vol. 10, pp. 54–57, 2015. 25

[24] R. G. Dolgoarshinnykh and S. P. Lalley, "Critical scaling for the SIS stochastic epidemic," *Journal of applied probability*, vol. 43, no. 3, pp. 892–898, 2006. 25, 71, 74

[25] I. Nåsell, *Extinction and quasi-stationarity in the stochastic logistic SIS model*. Springer, 2011. 26, 71

[26] M. S. Bartlett, "Measles periodicity and community size," *Journal of the Royal Statistical Society. Series A (General)*, vol. 120, no. 1, pp. 48–70, 1957. 26

[27] C. E. Gunning and H. J. Wearing, "Probabilistic measures of persistence and extinction in measles (meta) populations," *Ecology letters*, vol. 16, no. 8, pp. 985–994, 2013. 26

[28] R. G. Dolgoarshinnykh, *Epidemic Modelling: SIRS Models*. PhD thesis, University of Chicago, Department of Statistics, 2003. 27

[29] H. Andersson and B. Djehiche, "A threshold limit theorem for the stochastic logistic epidemic," *Journal of Applied Probability*, vol. 35, no. 03, pp. 662–670, 1998. 27, 46

[30] I. Nåsell, "On the time to extinction in recurrent epidemics," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 2, pp. 309–330, 1999. 27

[31] A. Kamenev and B. Meerson, "Extinction of an infectious disease: a large fluctuation in a nonequilibrium system," *Physical Review E*, vol. 77, no. 6, p. 061107, 2008. 27

[32] F. Lopes and M. Luczak, "Extinction time for the weaker of two competing SIS epidemics," *ArXiv e-prints*, Feb. 2018, 1802.04037. 27, 111, 112

[33] F. Ball, T. Britton, and P. Neal, "On expected durations of birth-death processes, with applications to branching processes and sis epidemics," *Journal of Applied Probability*, vol. 53, no. 1, pp. 203–215, 2016. 28

[34] A. W. Park, S. Gubbins, and C. A. Gilligan, "Extinction times for closed epidemics: the effects of host spatial structure," *Ecology Letters*, vol. 5, no. 6, pp. 747–755, 2002. 28

[35] D. W. Stroock and S. R. S. Varadhan, *Multidimensional diffusion processes*. Springer-Verlag, 2006. 36

[36] F. J. López, S. Martínez, and G. Sanz, "Stochastic domination and Markovian couplings," *Advances in Applied Probability*, vol. 32, no. 4, pp. 1064–1076, 2000. 39

[37] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*. Grundlehren der mathematischen Wissenschaften, Springer-Verlag, 2013. 40

[38] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005. 54

[39] D. R. McNeil, "Integral functionals of birth and death processes and related limiting distributions," *The Annals of Mathematical Statistics*, vol. 41, no. 2, pp. 480–485, 1970. 55

[40] A. Martin-Löf, "The final size of a nearly critical epidemic, and the first passage time of a Wiener process to a parabolic barrier," *Journal of applied probability*, vol. 35, no. 3, pp. 671–682, 1998. 71

[41] H. Holden, K. Karlsen, K.-A. Lie, and H. Risebro, *Splitting methods for partial differential equations with rough solutions. Analysis and Matlab programs.* European Mathematical Society, 2010. 72

[42] A. Göing-Jaeschke and M. Yor, "A survey and some generalizations of Bessel processes," *Bernoulli*, vol. 9, no. 2, pp. 313–349, 2003. 81

[43] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* National Bureau of Standards, 1972. 82

[44] H. F. Trotter, "On the product of semi-groups of operators," *Proceedings of the American Mathematical Society*, vol. 10, no. 4, pp. 545–551, 1959. 90

[45] P. R. Chernoff, "Note on product formulas for operator semigroups," *Journal of Functional analysis*, vol. 2, no. 2, pp. 238–242, 1968. 90

[46] P. R. Chernoff, *Product formulas, nonlinear semigroups, and addition of unbounded operators*, vol. 140. American Mathematical Soc., 1974. 90

[47] A. Friedman, *Partial differential equations of parabolic type.* R.E. Krieger Pub. Co., 1983. 92

[48] E. H. Kingsley, "Bernstein polynomials for functions of two variables of class c(k)," *Proceedings of the American Mathematical Society*, vol. 2, no. 1, pp. 64–71, 1951. 94

[49] P. P. Clément, H. J. Heijmans, S. Angenent, C. J. van Duijn, and B. de Pagter, "One-parameter semigroups," *CWI monographs*, vol. 5, 1987. 105

[50] B. Oksendal, *Stochastic differential equations: an introduction with applications.* Springer Science & Business Media, 2013. 105

[51] K.-J. Engel and R. Nagel, *One-parameter semigroups for linear evolution equations*, vol. 194. Springer Science & Business Media, 1999. 107

[52] A. D. Barbour, K. Hamza, H. Kaspi, and F. C. Klebaner, "Escape from the boundary in Markov population processes," *Advances in Applied Probability*, vol. 47, no. 4, pp. 1190–1211, 2015. 111

[53] E. Foxall, "The naming game on the complete graph," *Electron. J. Probab.*, vol. 23, p. 42 pp., 2018. 112

[54] A. D. Barbour and M. J. Luczak, "A law of large numbers approximation for Markov population processes with countably many types," *Probability Theory and Related Fields*, vol. 153, no. 3-4, pp. 727–757, 2012. 122

[55] G. Jenkinson and J. Goutsias, "Numerical integration of the master equation in some models of stochastic epidemiology," *PloS one*, vol. 7, no. 5, p. e36160, 2012. 132

[56] R. B. Sidje, "Expokit: a software package for computing matrix exponentials," *ACM Transactions on Mathematical Software (TOMS)*, vol. 24, no. 1, pp. 130–156, 1998. 132

[57] B. Munsky and M. Khammash, "The finite state projection algorithm for the solution of the chemical master equation," *The Journal of chemical physics*, vol. 124, no. 4, p. 044104, 2006. 132

[58] Y. Cao, D. T. Gillespie, and L. R. Petzold, "Avoiding negative populations in explicit poisson tau-leaping," *The Journal of chemical physics*, vol. 123, no. 5, p. 054104, 2005. 133

[59] E. W. Weisstein, "Faà di bruno's formula. From MathWorld–A Wolfram Web Resource." Last visited on 19/11/2019. 151

[60] L. Comtet, *Advanced Combinatorics: The art of finite and infinite expansions.* Springer Science & Business Media, 2012. 152