Report

Divakar Kumar, ✉: d.kumar@warwick.ac.uk
Supervised by: Prof. Gareth Roberts

September 13, 2016

Contents

1 Introduction

The study behind generating samples from intractable distributions has attracted considerable interest (Andrieu *et al.* [2003]; Roberts & Rosenthal [2004]). Roberts & Tweedie [1996] proposed a *diffusion*-based method to approximate an intractable distribution. They considered *Langevin diffusion* whose invariant distribution is the distribution of the interest. The *transition density* of diffusions is generally unknown. Therefore, the Euler discretization method is customarily used for studying the underlying behavior. Using time-discretization methods to study intractable diffusion models often involves Monte Carlo simulations, in turn, giving rise to Monte Carlo and discretization errors. Controlling discretization error is computationally expensive which involves making time-discretization sufficiently negligible (Kloeden & Platen [1999]). Approximation methods for the simulation of diffusions are inexact. Thus, the need to exactly simulate trajectories of diffusions without an approximation error arises.

Exact algorithms (Beskos & Roberts [2005]; Beskos *et al.* [2006, 2008]) provides methods to *exactly* simulate the trajectories of diffusions. They are a class of retrospective Monte Carlo methods for simulating sample paths of diffusions over finite time interval at a finite collection of points. This method utilizes *rejection sampling* on the diffusion path space, which in turn relies on finding suitable proposal measures to draw the proposal sample paths exactly. However, exactly simulating the trajectories of *Langevin diffusion* corresponding to an intractable distribution is seen as a circular problem (Pollock *et al.* [2016]). The *Scalable Langevin Exact (ScaLE)* algorithm (Pollock *et al.* [2016]) is a diffusion-based approach to simulate from an intractable distribution. The ScaLE method approximates the intractable distribution of interest with the *quasi-stationary* distribution of a 'killed' *Brownian motion*. ScaLE is a recent alternative to gradient-based Langevin MCMC schemes such as *Metropolis Adjusted Langevin Algorithm* (MALA) (Roberts & Tweedie [1996]) which circumvents the need to use Metropolis-type correction (Pollock *et al.* [2016]). As a result, this methodology is highly applicable to 'Big Data' problems (Pollock *et al.* [2016]).

In this report we introduce the *Re-sampled Scalable Langevin Exact (ReScaLE)* method which is a novel method of sampling from an intractable distribution of interest. In contrast to ScaLE, ReScaLE uses an alternative approach to simulate from the quasi-stationary distribution of a killed Brownian motion whose invariant distribution is given by the intractable distribution of interest. This report introduces the current literature on exact simulation from diffusion trajectories. Furthermore, the report describes the ReScaLE method to approximate intractable distributions of interest using recent advancements in the study of quasi-stationary behaviour proposed by Blanchet *et al.* [2012]. We will then present some challenges with the objective to motivate future research in this area and within this, we shall outline the possible structure of the thesis.

This report is organized as follows:- Section 2 introduces some preliminaries and results involving diffusion processes. We present the notion of stochastic approximation which is essential for outlining the proof of an important result by Blanchet *et al.* [2012] which is central to this report. Blanchet *et al.* [2012] provides a regenerative mechanism to explore the quasi-stationary behaviour of an absorbing Markov chains which is an important element in the construction of the ReScaLE method. Section 3 presents an exact method of simulation from diffusion path space. We close section 3 with an application of the exact method to simulate from a toy diffusion process of interest. Section 4 introduces current literature on the ScaLE method. Further, we discuss a recent advance in exploring quasi-stationary behaviour of an absorbing Markov chain (Blanchet *et al.* [2012]) which is used to devise the ReScaLE method. We close this section with an application of the ReScaLE algorithm to a toy example. Section 5 presents some current challenges in this area

with the plan of future research and outlines the possible structure of the thesis.

2 Preliminaries

2.1 Diffusion processes

2.1.1 Overview and definition

In many real world problems we are interested in a process which is continuously changing, albeit change is not deterministic. Fluctuations in the stock market is a good example of such processes (Black & Scholes [1973]). The behaviour of these processes can be studied using their movements with respect to time. Such processes can often be studied as *diffusion processes*, which is a continuous-time stochastic process (see Karatzas & Shreve [1991]; Øksendal [2003] for detail) with almost sure continuous paths in $\mathbb{R}^n$. Further, such processes are often represented by decomposing their movements into two components; one which is deterministic, often called drift, while the other component being random noise. This leads to the notion of *Stochastic Differential Equations* (SDE) of which diffusion process $\{X_t : t \geq 0\}$ is a solution. SDE is a convenient way of accessing the sample paths of a diffusion which is of the form

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t, \tag{2.1}$$

$\mu(X_t, t)$ in (2.1) denotes the instantaneous drift coefficient, while $\sigma(X_t, t)$ is the instantaneous diffusion coefficient. $\mu(X_t, t)$ and $\sigma(X_t, t)$ are locally Lipschitz with a linear growth bound (see Karatzas & Shreve [1991]; Øksendal [2003] for detail). B_t is a *Brownian motion* (see Karatzas & Shreve [1991] for definition) which represents the random noise process.

2.1.2 Itô process and Itô calculus

Definition 2.1 *A Itô process $\{X_t : t \geq 0\}$ characterized by the SDE*

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t, \tag{2.2}$$

is a stochastic process defined on $(\mathbf{\Omega}, \mathcal{F}, \mathbb{P})$ such that

- *$\mu(X_t, t)$ is $\mathcal{F}_t$ measurable and $\mathbb{P}\left(\int_0^t |\mu(X_s, s)|ds < \infty, \forall\, t \geq 0\right) = 1$, where $\mathcal{F}_t$ is a filtration on $(\mathbf{\Omega}, \mathcal{F})$ - a family of increasing sigma algebra such that $\mathcal{F}_s \subset \mathcal{F}_t$ for $0 \leq s < t$, and*
- *$\mathbb{P}\left(\int_0^t |\sigma(X_s, s)|^2 ds < \infty, \forall\, t \geq 0\right) = 1$.*

For the purpose of this report we restrict our attention to an important subclass of Itô processes called Itô diffusion. Exact algorithm can only be applied to obtain sample trajectories of a Itô diffusion. Formally,

Definition 2.2 *A (time-homogeneous) $\mathbb{R}^d-$ valued Itô diffusion $\{X_t : t \geq 0\}$ characterized by the SDE*

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t, \tag{2.3}$$

is a stochastic process defined on $(\mathbf{\Omega}, \mathcal{F}, \mathbb{P})$ *such that* $X : \mathbf{\Omega} \times [0, \infty) \to \mathbb{R}^d$ *which satisfy following Lipschitz condition*

$$|\mu(x) - \mu(y)| + |\sigma(x) - \sigma(y)| \leq C|x - y|, \quad \forall\, x, y \in \mathbb{R}^d, \text{ for some constant } C > 0. \tag{2.4}$$

Now we present Itô's formula which is an important result in Itô's calculus. This gives a SDE representation of a transformation of a Itô process. Here we present 1-dimensional version of Itô's lemma.

Theorem 2.1 (Itô's formula) *Consider an Itô process* X_t *given by SDE*

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t. \tag{2.5}$$

Further, let g *be a twice continuously differentiable function on* $[0, \infty) \times \mathbb{R}$*. Then a transformation defined by* $Y_t := g(X_t, t)$ *is again an Itô process where*

$$dY_t = \frac{\partial g}{\partial t}(X_t, t)dt + \frac{\partial g}{\partial x}(X_t, t)dx + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(X_t, t) \cdot (dX_t)^2. \tag{2.6}$$

Here $(dX_t)^2 = (dX_t) \cdot (dX_t)$ *is evaluated using*

$$dt \cdot dt = dB_t \cdot dt = dt \cdot dB_t = 0, \quad dB_t \cdot dB_t = dt. \tag{2.7}$$

The result for $d-$dimensional Itô's formula with its formal proof can be found in Øksendal [2003].

Next we discuss an important result that best describes the change in the behaviour of a Itô process defined on $(\mathbf{\Omega}, \mathcal{F}, \mathbb{P})$ if there is change in probability measure from $\mathbb{P}$ to $\mathbb{Q}$. Informally, we apply transformation to a given Itô diffusion to obtain another Itô process to ease simulation studies. These transformations do not change the solutions but transformed SDE will have different drift and diffusion coefficient since the underlined probability measure is different.

Theorem 2.2 (Girsanov's Theorem) *Consider a Itô process* $dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t$ *defined on* $(\mathbf{\Omega}, \mathcal{F}, \mathbb{P})$*. Then the Radon - Nikodým derivative* [1] *for a drift-less Itô process* $Y_t = \sigma(Y_t, t)dB_t$ *defined on probability space* $(\mathbf{\Omega}, \mathcal{F}, \mathbb{Q})$ *satisfies,*

$$\frac{d\mathbb{P}}{d\mathbb{Q}}(X) = \exp\left(\int_0^t \frac{\mu(X_s, s)}{\sigma^2(X_s, s)}dX_s - \frac{1}{2}\int_0^t \frac{\mu^2(X_s, s)}{\sigma^2(X_s, s)}ds\right), \text{ for the path } X = \{X_s : 0 \leq s \leq t\}. \tag{2.8}$$

3 The Exact Method for Sampling from an Itô Diffusion

3.1 Path-space rejection sampling - an overview

The Exact algorithm (Beskos & Roberts [2005]; Beskos *et al.* [2006]) is a rejection mechanism on the diffusion path-space. Path-space rejection sampling is a rejection mechanism on a diffusion

[1] Let $(\mathbf{\Omega}, \mathcal{F})$ be a measurable space with two probability measures $\mathbb{Q}$ and $\mathbb{W}$ We say that $\mathbb{Q}$ has *Radon - Nikodým property* with respect to $\mathbb{W}$ if there exists a measurable function $f : \mathbf{\Omega} \to [0, \infty)$ such that $\mathbb{Q}(A) = \int_A f d\mathbb{W}$ for any measurable set $A \in \mathcal{F}$. Then $f := \frac{d\mathbb{Q}}{d\mathbb{W}}$ is called the *Radon - Nikodým derivative* of $\mathbb{Q}$ with respect to $\mathbb{W}$.

path-space over a fixed time interval. Formally, let $\{X_t : t \geq 0\}$ be a diffusion of interest defined on the probability space $(\mathbf{\Omega}, \mathcal{F}, \mathbb{Q})$. Suppose we are interested in simulating a sample path according to measure $\mathbb{Q}$ on the interval $[0, T]$. It is difficult to simulate from the measure $\mathbb{Q}$ and therefore we choose an equivalent[2] measure $\mathbb{W}$ on the space $(\mathbf{\Omega}, \mathcal{F})$. The measure $\mathbb{W}$ is chosen such that it is easier to obtain sample paths according to $\mathbb{W}$. In the rejection sampling setting we require that the *Radon - Nikodým derivative* is bounded. Formally, there exists $M > 0$ such that for any path $X = \{X_t : 0 \leq t \leq T\}$ we have,

$$\frac{d\mathbb{Q}}{d\mathbb{W}}(X) \leq M. \tag{3.1}$$

We propose a sample path X according to the measure $\mathbb{W}$ and accept it with probability

$$P_{\mathbb{W}}(X) := \frac{1}{M}\frac{d\mathbb{Q}}{d\mathbb{W}}(X) \in [0, 1]. \tag{3.2}$$

The average acceptance probability of the accepted sample path is given by $E_{\mathbb{W}}(P_{\mathbb{W}}(X)) = \frac{1}{M}$.

3.2 Model class

The exact algorithm only considers an Itô diffusion model of the form,

$$dX_t = \mu(X_t)dt + dB_t, \quad X_0 = x_0, \quad 0 \leq t \leq T. \tag{3.3}$$

However, for the one-dimensional Itô diffusion of the form $dX_t = \mu(X_t)dt + \sigma(X_t)dB_t$ with initial value $X_0 = x_0$, the *Lamperti transformation* can reduce to the Itô diffusion to the form (3.3). Such transformations might not exist in multi-dimensional settings. The Lamperti transformation is given by

$$Y_t = \alpha(X_t) = \int_{x_0}^{X_t} \frac{1}{\sigma(x)}dx. \tag{3.4}$$

Applying Itô's formula to (3.4) we have

$$\begin{aligned} dY_t &= \alpha'(X_t)dX_t + \frac{1}{2}\alpha''(X_t)(dX_t)^2 && (3.5)\\ &= \frac{1}{\sigma(X_t)}\left(\mu(X_t)dt + \sigma(X_t)dB_t\right) - \frac{1}{2}\frac{\sigma'(X_t)}{\sigma^2(X_t)}\sigma(X_t)dt && (3.6)\\ &= \left(\frac{\mu(X_t)}{\sigma(X_t)} - \frac{1}{2}\frac{\sigma'(X_t)}{\sigma^2(X_t)}\right)dt + dB_t. && (3.7) \end{aligned}$$

Thus, the transformed Itô diffusion Y_t is reduced to the form (3.3). Hence without loss of generality, we can consider the diffusion of the form (3.3).

[2]A probability measure $\mathbb{Q}$ is equivalent to another measure $\mathbb{W}$ defined on $(\mathbf{\Omega}, \mathcal{F})$ if $\mathbb{Q}(A) = 0$ iff $\mathbb{W}(A) = 0$, for any measurable set $A \in \mathcal{F}$.

3.3 The exact method

Let $\mathbb{Q}$ denote the measure induced by diffusion (3.3). To carry out a rejection sampling from $\mathbb{Q}$ we choose an equivalent measure $\mathbb{W}$ induced by the Brownian path $B = \{B_t : 0 \leq t \leq T\}$. Under Girsanov's transformation of equivalent probability measure for a path $X = \{X_t : 0 \leq t \leq T\}$, it implies that

$$\frac{d\mathbb{Q}}{d\mathbb{W}}(X) = \exp\left(\int_0^T \mu(X_t)dX_t - \frac{1}{2}\int_0^T \mu^2(X_t)dt\right). \tag{3.8}$$

Our aim is to apply the rejection mechanism using (3.8). The exact evaluation of (3.8) is impossible due to the presence of an Itô integral. However, (3.8) can be simplified using the Itô's Lemma on $A(x) := \int_0^x \mu(u)du$, which requires the following regularization condition:

The drift coefficient function $\mu(x)$ is differentiable. (R1)

Since $A(X_T) - A(x_0) = \int_0^T dA(X_s)$, by Ito's Lemma

$$\int_0^T dA(X_s) = \int_0^T \mu(X_s)dX_s + \int_0^T \frac{1}{2}\mu'(X_s)(dX_s)^2 \tag{3.9}$$

$$= \int_0^T \mu(X_s)dX_s + \int_0^T \frac{1}{2}\mu'(X_s)ds, \tag{3.10}$$

which implies that,

$$A(X_T) - A(x) - \frac{1}{2}\int_0^T \left(\mu^2(X_s) + \mu'(X_s)\right)ds = \int_0^T \mu(X_s)dX_s - \frac{1}{2}\int_0^T \mu^2(X_s)ds. \tag{3.11}$$

Equation (3.11) leads to the following simplification of (3.8):

$$\frac{d\mathbb{Q}}{d\mathbb{W}}(X) = \exp\left(A(X_T) - A(x) - \frac{1}{2}\int_0^T \left(\mu^2(X_t) + \mu'(X_t)\right)dt\right). \tag{3.12}$$

Rejection sampling using Brownian proposals requires (3.12) to be bounded, however $A(\cdot)$ can be unbounded, which in turn negates the existence of a bounding constant M in (3.2). To circumvent such an issue, we consider another probability measure, draws from which are identical to a Brownian motion started at the position x_0 except for the distribution of its end point. This process is called *Biased Brownian motion* as proposed by Beskos & Roberts [2005]. This is a new proposal for the target measure $\mathbb{Q}$. We need the following regularisation condition to facilitate the simulation of end point:

$$\textit{If } A(x) := \int_0^x \mu(u)du, \quad x \in R; \textit{ then } \int \exp(A(x) - (x - x_0)^2/2T))dx < \infty. \tag{R2}$$

Definition 3.1 *Biased Brownian motion is a process* $Z_t := \{B_t : B_0 = x, B_T = y \sim h\}$ *(with measure* $\mathbb{Z}$*) where* $x, y \in \mathbb{R}$, $0 \leq t \leq T$ *such that*

$$h(y; x, T) \propto \exp\left(A(y) - \frac{(y-x)^2}{2T}\right). \tag{3.13}$$

The choice of the proposal measure $\mathbb{Z}$ requires us to evaluate the *Radon - Nikodým derivative* $\frac{d\mathbb{Q}}{d\mathbb{Z}}$ in the rejection sampling setting. In order to get the form of $\frac{d\mathbb{Q}}{d\mathbb{Z}}$, we first evaluate $\frac{d\mathbb{W}}{d\mathbb{Z}}$. To achieve this, we first present a result which involves two stochastic processes, defined on the same measurable space. This result assumes the following: if the conditional laws of two stochastic processes conditioned on their end point are the same, then the *Radon - Nikodým derivative* of these conditional measures is given by the ratio of their densities at the end point. The following proposition due to Beskos *et al.* [2006] can be utilized to prove the equivalence between probability measures $\mathbb{Q}$ and $\mathbb{Z}$.

Proposition 3.1 (Beskos *et al.* [2006]) *Let* $M = \{M_t : 0 \leq t \leq T\}$ *and* $N = \{N_t : 0 \leq t \leq T\}$ *be two stochastic processes on* $(\mathbf{\Omega}, \mathcal{F})$ *with respective measures* $\mathbb{M}$ *and* $\mathbb{N}$*. Let* f_M *and* f_N *be the densities at the end point* T*. If* $(M \mid M_T = x) \stackrel{d}{=} (N \mid N_T = x)$ *for all* $x \in \mathbb{R}$ *then*

$$\frac{d\mathbb{M}}{d\mathbb{N}}(X) = \frac{f_M(X_T)}{f_N(X_T)}, \quad \textit{where}\, X = \{X_t : 0 \leq t \leq T\}. \tag{3.14}$$

To draw sample paths with respect to the law of a Biased Brownian motion, we first draw its end point according to the distribution h and then the intermediate points are drawn as per the law of a Brownian Bridge, conditioned on the starting point and end point. It is possible to draw from h due to existence of condition (R2), which makes the integral finite. It can be seen that measures $\mathbb{W}$ (corresponding to Brownian motion) and $\mathbb{Z}$ (corresponding to Biased Brownian motion) satisfy the conditions in proposition 3.1. Consequently,

$$\frac{d\mathbb{W}}{d\mathbb{Z}}(X) = \frac{\frac{1}{\sqrt{2\pi T}} \exp\left(-\frac{X_T^2}{2T}\right)}{h(X_T)}. \tag{3.15}$$

Using (3.12) and (3.15), we have

$$\frac{d\mathbb{Q}}{d\mathbb{Z}}(X) = \frac{d\mathbb{Q}}{d\mathbb{W}}(X)\frac{d\mathbb{W}}{d\mathbb{Z}}(X), \tag{3.16}$$

$$\frac{d\mathbb{Q}}{d\mathbb{Z}}(X) \propto \exp\left\{-\frac{1}{2}\int_0^T (\mu^2(X_t) + \mu'(X_t))dt\right\}. \tag{3.17}$$

Further, we assume that the function $(\mu^2 + \mu')/2$ is bounded i.e.

$$\textit{there exists}\ l, u\ \textit{such that}\ l \leq \frac{(\mu(x)^2 + \mu'(x))}{2} \leq u, \quad x \in \mathbb{R}. \quad . \tag{R3}$$

Using assumption (R3) it is easy to see that for the non-negative function $\phi := \frac{\mu^2+\mu'}{2} - l$,

$$\frac{d\mathbb{Q}}{d\mathbb{Z}}(X) \propto \exp\left\{ -\int_0^T \left(\frac{1}{2}(\mu^2(X_t) + \mu'(X_t)) - l \right) dt \right\}, \tag{3.18}$$

$$\frac{d\mathbb{Q}}{d\mathbb{Z}}(X) \propto \exp\left\{ -\int_0^T \phi(X_t)dt \right\} := P_{\mathbb{Z}}(X) \in (0,1]. \tag{3.19}$$

Since ϕ is strictly non-negative, (3.19) holds. Within a rejection sampling construction in which $X \sim \mathbb{Z}$, (3.19) gives us a $P_{\mathbb{Z}}(X)-$ coin to decide upon whether or not to retain X as a draw from $\mathbb{Q}$. However, the construction of a $P_{\mathbb{Z}}(X)-$ coin requires us to continuously sample X using $\mathbb{Z}$, which is not possible. Thus, there is a need to construct a $P_{\mathbb{Z}}(X)-$ coin based on finitely many sample points obtained according to $\mathbb{Z}$. Exact method circumvents this issue by choosing finitely many proposed points (often called a *skeleton*) over the time interval $[0, T]$. To understand this, we can observe that $P_{\mathbb{Z}}(X)$ is the probability of a Poisson random variable with intensity ϕ taken the value 0. Since by assumption (R3) we can conclude that

$$0 \leq \phi \leq u - l := M, \tag{3.20}$$

therefore, we can think of $P_{\mathbb{Z}}(X)$ as the probability that a Poisson process of unit intensity on the rectangle $\{(x, y) \in [0, T] \times [0, M]\}$ has no points in the epigraph $\{(x, y) \in [0, T] \times [0, M] : y \leq \phi(x)\}$. The above idea is made precise as follows:

Theorem 3.2 (Beskos *et al.* [2006]) *Let $X = \{X_t | t \geq 0,\ X_0 = x, X_T = y \sim h\}$ be the realizations of a Biased Brownian motion with $M = \sup_{X_t | t \geq 0} \phi(X_t)$. If Φ is a two-dimensional homogeneous Poisson process of unit intensity on $[0, T] \times [0, M]$ and N is the number of points of Φ found below the graph $\{(t, \phi(X_t)) | t \in [0, T]\}$ then,*

$$P(N = 0|X) = \exp\left\{ -\int_0^T \phi(X_t)dt \right\}.$$

Proof Using the definition of a two-dimensional Poisson process [3], the number of points below the graph $\{(t, \phi(X_t)) | t \in [0, T]\}$ is Poisson random variable with rate parameter $\left\{ \int_0^T \phi(X_t)dt \right\}$. We have,

$$P(N = 0 \,|\, X) = \exp\left\{ -\int_0^T \phi(X_t)dt \right\}.$$

[3] A two-dimensional Poisson process of rate λ is a continuous time stochastic process such that

1. the number of points in any given area A is Poisson distributed with mean λA;
2. the number of points in disjoint regions are independent.

A detailed description of simulating Poisson processes on a two-dimensional rectangle can be found in Ross [2010].

This result is profound in the sense that we can evaluate $P_{\mathbb{Z}}(X)-$coin while only partially unveiling the sample path. In particular, one can generate a realization of a homogeneous Poisson process with intensity M and then draw from the law of a Biased Brownian motion at the time instances where Poisson process events have occurred. Subsequently, we accept the realized skeleton from the Biased Brownian motion if all the points of Φ lie above the graph of ϕ. As a result, we only need a finite number of sample points from a path of Biased Brownian motion. Once the skeleton is accepted, the remaining path between our two successive selected points in the skeleton can be drawn from the law of a Brownian Bridge, conditioned on those two points. Algorithm 3.1 below summarizes the 'Exact algorithm'.

Algorithm 3.1: Exact Algorithm(μ, x_0, T)

$l \leftarrow \inf_{x\in\mathbb{R}} \left(\frac{\mu^2+\mu'}{2}\right)(x), \phi \leftarrow \frac{\mu^2+\mu'}{2} - l, M \leftarrow \sup_{x\in\mathbb{R}} \phi(x), A(x) \leftarrow \int_0^x \mu(s)ds$

$h(y; x_0, T) \propto \exp\left(A(y) - \frac{(y-x_0)^2}{2T}\right)$

while $N \neq 0$

do
- $I.$ $X_T \sim h(y; x_0, T)$
- $II.$ Generate $((t_1, x_1), ..., (t_k, x_k)) \sim \Phi -$ a two-dimensional Poisson process of unit intensity on $[0, T] \times [0, M]$
- $III.$ $(X_{t_1}, X_{t_2}, ..., X_{t_k}) \sim$ Biased Brownian Motion at times$(t_1, t_2, ..., t_k)$
- $IV.$ $N \leftarrow$ Number of times $\{x_i < \phi(X_{t_i}) : i \in \{1, 2, ...k\}\}$

return $(X_{t_1}, X_{t_2}, ..., X_{t_k}, X_T)$

3.4 Applying the Exact Algorithm - an example

We consider an application of the exact method to sample the trajectories of the following diffusion:

$$dX_t = \frac{1}{1+X_t^2}dt + dB_t \quad 0 \leq t \leq T,\ X_0 = 0. \tag{3.21}$$

Next, we check if (3.21) satisfies all the conditions laid down for the working of the Exact algorithm:

(R1): Drift function is differentiable everywhere with

$$\mu'(x) = \frac{-2x}{(1+x^2)^2}. \tag{3.22}$$

(R2): $A(x)$ has the following form:

$$A(x) = \int_0^x \mu(s)ds = \int_0^x \frac{1}{1+s^2}ds = \tan^{-1}(x), \tag{3.23}$$

and is integrable,

$$\int \exp\left\{A(x) - \frac{(x)^2}{2T}\right\} dx = \int \exp\left\{\tan^{-1}(x)\right\} \exp\left\{-\frac{(x)^2}{2T}\right\} dx \tag{3.24}$$

$$\leq \int \exp\left\{\frac{\pi}{2}\right\} \exp\left\{-\frac{(x)^2}{2T}\right\} dx \tag{3.25}$$

$$= \exp\left\{\frac{\pi}{2}\right\} \int \exp\left\{-\frac{(x)^2}{2T}\right\} dx \tag{3.26}$$

$$= \exp\left\{\frac{\pi}{2}\right\} \sqrt{2\pi T} < \infty. \tag{3.27}$$

(R3): Next we check that function ϕ is bounded. We note that

$$g(x) = \frac{\mu'(x) + \mu^2(x)}{2} = \frac{1}{2}\left[\frac{-2x}{(1+x^2)^2} + \left(\frac{1}{1+x^2}\right)^2\right] \tag{3.28}$$

$$= \frac{1}{2}\left[\frac{(1-2x)}{(1+x^2)^2}\right]. \tag{3.29}$$

To decide the lower and upper bound for the function $g(x)$, we first find out the stationary points of the function which is given by $g'(x) = 0$,

$$\frac{1}{2}\left[\frac{(1+x^2)^2(-2) - (1-2x)2(1+x^2)(2x)}{(1+x^2)^4}\right] = 0 \tag{3.30}$$

$$(1+x^2) + 2x(1-2x) = 0 \tag{3.31}$$

$$1 + 2x - 3x^2 = 0 \tag{3.32}$$

$$(1+3x)(1-x) = 0 \tag{3.33}$$

$$x = 1, -1/3. \tag{3.34}$$

It is easy to see using the second derivative that $x = 1$ is the point of minima while $x = -1/3$ is the point of maxima and so we have

$$-0.125 \leq \frac{\mu^2(x) + \mu'(x)}{2} \leq 0.675. \tag{3.35}$$

Further we define ϕ as follows:

$$\phi(x) = \frac{1}{2}\left[\frac{(1-2x)}{(1+x^2)^2}\right] + 0.125. \tag{3.36}$$

Next, we need to propose samples from the Biased Brownian motion, for which the end point is distributed according to

$$h(x) \propto \exp\left\{\tan^{-1}(x)\right\} \exp\left\{-\frac{(x)^2}{2T}\right\}. \tag{3.37}$$

The end point can be simulated using the rejection sampling (see for example [Ross, 2010, Chapter 11.2.2] for details) by proposing the samples from a $N(0, T)$ distribution.

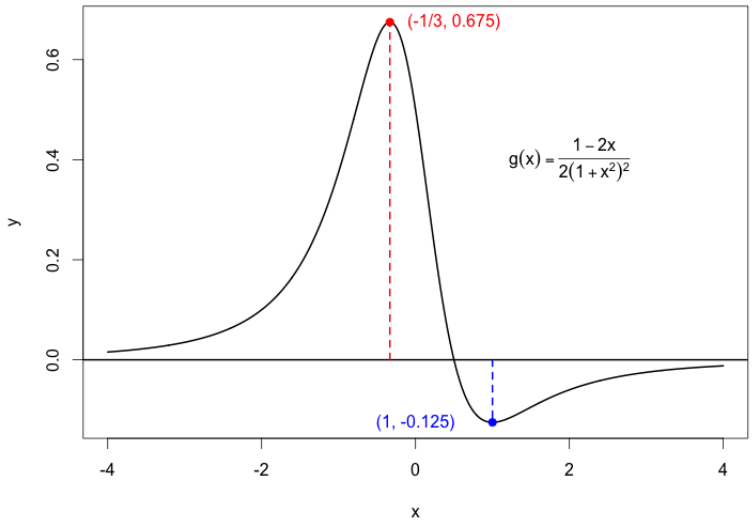

Figure 1: *Figure illustrates the function* $\frac{\mu^2+\mu'}{2}$ *with its global extrema.*

Figure 2 shows an accepted Exact skeleton of X, while figure 3 shows the transformation $X_t \to \phi(X_t)$ for the same skeleton. Green points in the figure 3 are the points from a two-dimensional Poisson process on the rectangle $[0, 4] \times [0, 0.8]$. Since all the intermediate green points lie above the graph $t \to \phi(X_t)$, this skeleton is accepted. On the contrary, figure 4 and figure 5 possess a similar idea for the case when the skeleton is rejected. Since it can be observed in figure 5 that at least one green point lies below the epigraph, this the skeleton is rejected.

This 6 shows the estimated kernel density function of X_1, based on a sample size of 100000 using the Exact algorithm and Euler's scheme with various discretization. It can be seen that the density of X_1, approximated using Exact algorithm is close to the Euler's approximation with really small discretization.

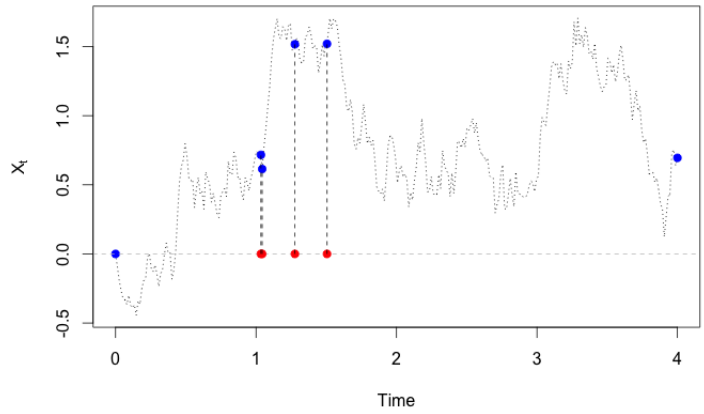

Figure 2: *This illustrates an accepted Exact skeleton of X. Red points show the instances when Poisson events are observed. The first and the last blue points are the starting and end points respectively for the Biased Brownian motion while the black dotted line is a Brownian Bridge.*

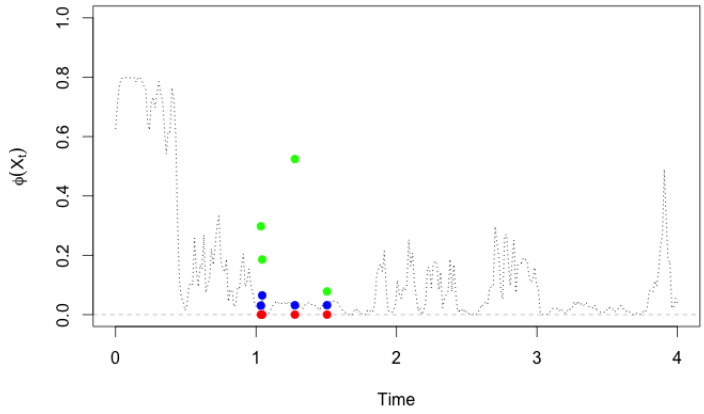

Figure 3: *This illustrates the transformation $X_t \to \phi(X_t)$ for an accepted Exact skeleton of X. The green points are the Poisson process points. Since no green points lie below the blue points, this skeleton is accepted.*

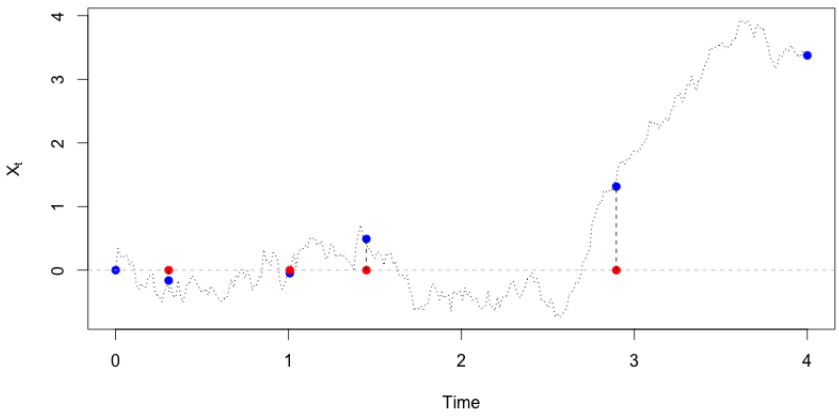

Figure 4: *This illustrates a rejected skeleton of X. Red points show the instances when Poisson events are observed. The first and the last blue points are the starting and end points respectively for the Biased Brownian motion while black dotted line is a Brownian Bridges.*

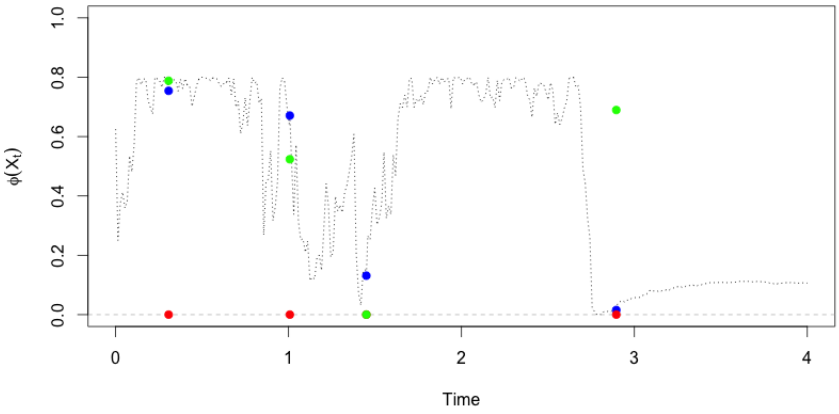

Figure 5: *This illustrates the transformation $X_t \to \phi(X_t)$ for a rejected skeleton of X. Green points are the Poisson Process points. Since more than one green points lie below the blue points i.e graph $t \to \phi(X_t)$, this skeleton is rejected.*

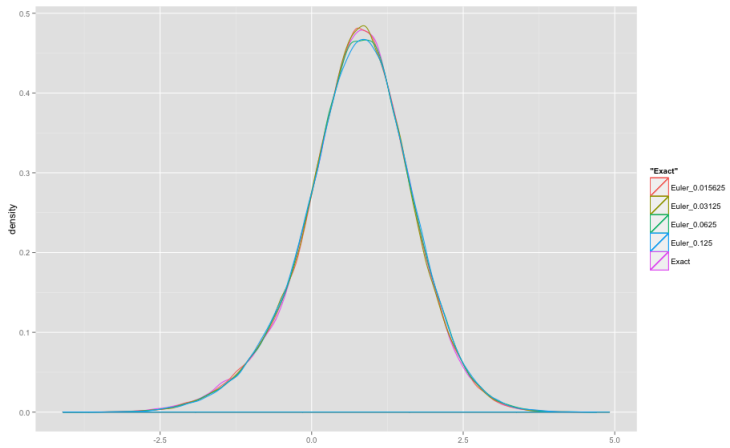

Figure 6: *This illustrates the kernel density approximation of X_1 based on a sample of size 100000 obtained via Exact algorithm and Euler discretization of $2^{-3}, 2^{-4}, 2^{-5}, 2^{-6}$, which has been colour coded accordingly.*

4 The ReScaLE (Re-sampled Scalable Langevin Exact) Algorithm

4.1 The Scalable Langevin Exact (ScaLE) Method

4.1.1 Motivation and overview

The Scalable Langevin Exact (ScaLE) (Pollock *et al.* [2016]) method uses the idea of exact simulation of diffusion trajectories. One motivation for the ScaLE algorithm derives from the following situation. Suppose we have been given N data points and we are interested in Bayesian inference for the parameter x belonging to the parameter space, which would require calculating

$$\pi(x) \propto p(x) \prod_{i=1}^{N} l_i(x), \tag{4.1}$$

at every iteration of an MCMC algorithm. Here $\pi(x)$ denotes the posterior density for the parameter x while $p(x)$ is the prior density and $l_i(x)$ is the likelihood function corresponding to $i-$th data. If N is large, calculating (4.1) at each iteration of the algorithm is costly from a computational point of view. There are various ways in which this problem can be addressed. One approach (using multi-core computing) is to break the data into various pieces, compute the posterior in parallel and then recombine the posterior (Scott *et al.* [2013]; Teh *et al.* [2015]). However, recombination of posteriors can be a problem from both the theoretical and computational point of view. An alternative approach which is based on a single *worker*, employs the gradient-based method (Pollock *et al.* [2016]). Gradient-based methods consider the gradient of the logarithm of $\pi(x)$ as the drift of the Langevin diffusion and is satisfied by the following stochastic differential equation,

$$dX_t = \frac{1}{2}\nabla \log \pi(X_t)dt + dB_t, \quad X_0 = x_0, t \in [0, T]. \tag{4.2}$$

The invariant distribution of (4.2) is given by π (Pollock *et al.* [2016]; Roberts & Tweedie [1996]). Solving (4.2) via time-discretization method is based on approximations and thus is not exact. Furthermore, it is computationally expensive and the size of error grows with the widening time-horizon. Exact simulation can be a way out of the issues. We note that (4.2) is a special case of an Itô diffusion with unit volatility seen in section 3 with the drift given by $\mu(X_t) = \nabla \log \pi(X_t)/2$. Simulating exactly from the above diffusion requires satisfying the conditions (R1) − (R3). Exact simulation of trajectories of (4.2) on a finite time interval would require us to calculate $\phi = \frac{\mu^2+\mu'}{2} - l$ to decide the acceptance of a simulated skeleton. However, it can be observed that ϕ takes the form

$$\phi(x) = \sum_{i=1}^{N}\sum_{j=1}^{N} g_i(x)g_j(x), \tag{4.3}$$

for function g defined as a function of π (a detailed description of this can be found in section (4.5.3)). When N is large, calculation of ϕ can be costly. Pollock *et al.* [2016] circumvents this issue by employing the sub-sampling idea that considers an unbiased estimator of ϕ, which are cheaper to evaluate for each point in the skeleton. For instance, $N^2 g_I(x)g_J(x)$ can be chosen as an unbiased estimator of ϕ, for I, J independent and randomly drawn from the set $\{1, ..., N\}$. Informally speaking, the ScaLE method scales well with the size of the data due to the fact that it does not calculate costly ϕ rather it evaluates its cheap unbiased estimators.

4.1.2 The methodology

Exact simulation of the trajectories of (4.2) over a finite time interval can be achieved by first simulating its terminal position at the final time point. Then it observes its trajectories at the intermediate time points using law of the Brownian bridges. Thus simulating the end point for the purpose of drawing from a Biased Brownian motion can be done via the density

$$h(x) \propto \exp\left\{A(x) - \frac{(x-x_0)^2}{2T}\right\} \tag{4.4}$$

$$\propto \exp\left\{\log(\{\pi(x)\}^{\frac{1}{2}}) - \frac{(x-x_0)^2}{2T}\right\} \tag{4.5}$$

$$\propto \{\pi(x)\}^{\frac{1}{2}} \exp\left\{-\frac{(x-x_0)^2}{2T}\right\}. \tag{4.6}$$

However, drawing from (4.6) is as difficult as drawing from π itself, due to the term $\{\pi(x)\}^{\frac{1}{2}}$ (Pollock *et al.* [2016]). Under the condition (R3) (which requires that $\frac{(\mu(x)^2+\mu'(x))}{2}$ is a bounded function) we have

$$l \leq \frac{(\mu(x)^2+\mu'(x))}{2} \leq u, \quad x \in \mathbb{R}, \tag{4.7}$$

and using the Dacunha-Castelle formula (Dacunha-Castelle & Florens-Zmirou [1986]; Pollock [2013]) we get the following form of the transition density (Pollock *et al.* [2016])

$$p_{0,t}(x_0,x) = \frac{1}{\sqrt{2\pi t}} \exp\left\{-\frac{(x-x_0)^2}{2t}\right\} \exp\{A(x)-A(x_0)\}\, \mathbb{E}_{x_0,x}\left(\exp\left\{-\int_0^t \frac{(\mu(X_s)^2+\mu'(X_s))}{2} ds\right\}\right). \tag{4.8}$$

Here the expectation is taken with respect to a Brownian bridge started at $X_0 = x_0$ and ending at $X_t = x$. With $\phi_\mu = \frac{\mu^2+\mu'}{2} - l$, (4.8) can be reduced to

$$p_{0,t}(x_0,x) = \frac{1}{\sqrt{2\pi t}} \exp\left\{-\frac{(x-x_0)^2}{2t}\right\} \exp\{A(x)-A(x_0)-lt\}\, \mathbb{E}_{x_0,x}\left(\exp\left\{-\int_0^t \phi_\mu(X_s) ds\right\}\right). \tag{4.9}$$

The transition density $p_{0,t}(x_0,x)$ converges to $\pi(x)$ as $t \to \infty$ (Pollock *et al.* [2016]). However, we observe that for a chain starting at $x_0 = 0$ the following holds:

$$p_{0,t}(0,x) \propto \exp\left\{-\frac{x^2}{2t}\right\} \exp\{A(x)\}\, \mathbb{E}_{x_0,x}\left(\exp\left\{-\int_0^t \phi_\mu(X_s) ds\right\}\right) \tag{4.10}$$

$$\propto \exp\left\{-\frac{x^2}{2t}\right\} \{\pi(x)\}^{\frac{1}{2}}\, \mathbb{E}_{x_0,x}\left(\exp\left\{-\int_0^t \phi_\mu(X_s) ds\right\}\right). \tag{4.11}$$

Now, if we drop the middle term $\{\pi(x)\}^{\frac{1}{2}}$ in (4.11), this biases the invariant distribution by a factor of $\{\pi(x)\}^{\frac{1}{2}}$ and the transition density (4.11) converges to the wrong distribution $\{\pi(x)\}^{\frac{1}{2}}$. However, if we double the drift (Pollock *et al.* [2016]) $2\mu(x) = \nabla \log(\pi(x))$, we define $\phi := \phi_{2\mu}$ then the transition density evaluates to

$$p_{0,t}(0,x) \propto \exp\left\{-\frac{x^2}{2t}\right\} \exp\{2A(x)\}\, \mathbb{E}_{x_0,x}\left(\exp\left\{-\int_0^t \phi(X_s)ds\right\}\right) \tag{4.12}$$

$$\propto \exp\left\{-\frac{x^2}{2t}\right\} \{\pi(x)\} \mathbb{E}_{x_0,x}\left(\exp\left\{-\int_0^t \phi(X_s)ds\right\}\right). \tag{4.13}$$

Now (4.13) converges to π^2. Similarly, ignoring the middle term $\{\pi(x)\}$ in (4.13) biases the invariant density by a factor of π and thus it converges to the true density π (Pollock *et al.* [2016]).

Expression (4.13) without the middle term is interpreted as the transition density of a killed Brownian motion with a state-dependent killing rate $\phi(X)$ ([Barber *et al.* , 2011, Chapter 4]; Pollock *et al.* [2016]; Øksendal [2003]). We present it formally as follows:

Theorem 4.1 *Consider a standard Brownian motion $\{X_t : t \geq 0\}$ which is killed at X_s with a state-dependent 'killing-rate' $\phi(X_s)$. Then the stationary density of this 'killed Brownian motion' conditional on its survival until time t is given by*

$$q_{0,t}(0,x) \propto \exp\left\{-\frac{x^2}{2t}\right\} \mathbb{E}_{x_0,x}\left(\exp\left\{-\int_0^t \phi(X_s)ds\right\}\right). \tag{4.14}$$

Proof See Appendix 6.3.

But the transition density (4.14) converges to π, which implies that π can be approximated using the long term behaviour of a Brownian motion before it is killed (such behaviour is also called *quasi-stationarity*) with a state-dependent killing rate $\phi(X)$. However, it is still impossible to draw continuously according to the transition density of a killed Brownian motion. We can use the *Poisson thinning* technique to circumvent this issue. Let us assume we use a *Poisson process* to describe the time of kill of a Brownian motion $\{X_t : t \geq 0\}$. If $\phi(X_s)$ denotes the instantaneous rate at which the process $\{X_t : t \geq 0\}$ is killed, then $\exp\left\{-\int_0^\tau \phi(X_s)ds\right\}$ will represent the probability that process survived until time τ. Suppose $M = \sup_{x \in R} \phi(x)$, then we have

$$\exp\left\{-\int_0^\tau \phi(X_s)ds\right\} = \exp\left\{-\int_0^\tau \frac{\phi(X_s)}{M} \times M ds\right\}. \tag{4.15}$$

The RHS of the above equality signifies that the process survives until time τ if it is instantaneously killed at time s with probability $\frac{\phi(X_s)}{M}$, where time s is drawn according to a *Poisson process* of rate M. We can summarize the above ideas as follows:

Theorem 4.2 *Let $\tau_1, \tau_2, ...,$ be the realizations from the homogeneous Poisson process of rate M. Let $X_{\tau_1}, X_{\tau_2}, ...$ be the realizations of Brownian motion $\{X_t : t \geq 0\}$ at $\tau_1, \tau_2, ...,$ with $M = \sup_{X_t | t \geq 0} \phi(X_t)$. If the Brownian motion started at 0 is killed at τ_i with probability $\frac{\phi(X_{\tau_i})}{M}$ then,*

$$P(\textit{Brownian motion survived until time } t) = \exp\left\{-\int_0^t \phi(X_s)ds\right\}. \tag{4.16}$$

Proof See Appendix 6.2.

Hence the expression (4.14) can be interpreted as follows: we continue drawing samples according to a Brownian motion at the event times of a Poisson process of rate M and decide to stop at a Poisson event time s with probability given by $\frac{\phi(X_s)}{M}$. Conditioned on the survival of the process, the transition density (4.14) converges to its *quasi-stationary* density π (Pollock *et al.* [2016]). So our problem of simulating according to π reduces to being able to simulate from the quasi-stationary density of a killed Brownian motion, where the path is instantaneously killed at rate ϕ. Pollock *et al.* [2016] uses a Sequential Monte Carlo (SMC) based approach to simulate from the quasi-stationary density given in (4.14), while we use a recent method suggested in Blanchet *et al.* [2012]. In the next subsection, we discuss the concept of quasi-stationarity and explore a novel approach to sample from a quasi-stationary distribution discussed in Blanchet *et al.* [2012].

4.2 Simulating from a quasi-stationary distribution

4.2.1 Quasi-stationarity - an overview

To understand the concept of the quasi-stationarity, we consider a continuous-time Markov process $\{X_t : t \geq 0\}$ on a discrete state-space which has a absorbing state: if the process enters it can never escape. The dynamics of such a process before it gets absorbed is termed as 'quasi-stationarity'. Quasi-stationarity was first coined by Darroch & Seneta [1965]. The quasi-stationarity is an important element in the study of birth-death processes and birth-catastrophe (Neill [2007]). For simplicity, we focus on a continuous-time Markov chain defined on a discrete state-space. Suppose d denotes the absorbing state and T is the set of all non-absorbing states. Thus the quasi-stationary distribution of X_t is a distribution on non-absorbing states T, conditioned on non-absorption of X_t until time t. Quasi-stationarity requires that such a distribution is independent of time.

Suppose R denotes the rate matrix of the continuous-time Markov process with absorbing state d. If we define the transition probabilities

$$p_{0,t}(i,j) = P(X_t = j | X_0 = i), \quad \text{and} \tag{4.17}$$

$$p_{0,t}(i,T) = 1 - p_{0,t}(i,d). \tag{4.18}$$

Here $p_{0,t}(i,j)$ denotes the probability that the process transits to state j at time t starting from i at time 0. $p_{0,t}(i,T)$ denotes the probability that the process does not visit absorbing state d starting from state i at time 0. Then the quasi-stationary distribution can be defined formally as: (Darroch & Seneta [1967]; Zheng [2014])

$$\pi_j = \lim_{t \to \infty} \frac{p_{0,t}(i,j)}{p_{0,t}(i,T)}. \tag{4.19}$$

4.2.2 A heuristic approach to sample from a quasi-stationary density

Now we present a heuristic analysis (de Oliveira & Dickamn [2005]) for interpreting the quasi-stationary behaviour of an absorbing Markov chain. Using Kolmogorov's forward equation, we have

$$\frac{d}{dt}(p_{0,t}(i,j)) = \sum_l p_{0,t}(i,l)R(l,j), \tag{4.20}$$

where $R(l,j)$ is the rate at which process moves from state l to state j. Using (4.18) and (4.20)

$$\frac{d}{dt}(p_{0,t}(i,T)) = \frac{d}{dt}(1-p_{0,t}(i,d)) = -\frac{d}{dt}(p_{0,t}(i,d)) = -\sum_l p_{0,t}(i,l)R(l,d). \tag{4.21}$$

The quasi-stationary distribution defined in (4.19) can be approximated by (de Oliveira & Dickamn [2005]; Blanchet *et al.* [2012])

$$\pi_j p_{0,t}(i,T) \approx p_{0,t}(i,j) \quad \text{for large } t. \tag{4.22}$$

Since π_j does not depend on t, differentiating (4.22) with respect to t yields,

$$\pi_j \frac{d}{dt}(p_{0,t}(i,T)) = \frac{d}{dt}(p_{0,t}(i,j)) \tag{4.23}$$

$$= \sum_l p_{0,t}(i,l)\, R(l,j). \quad \text{follows by (4.20)} \tag{4.24}$$

$$= \sum_l \pi_l\, p_{0,t}(i,T)R(l,j), \quad \text{follows by (4.22)}. \tag{4.25}$$

Similarly, using (4.21) and (4.22) we find

$$\frac{d}{dt}(p_{0,t}(i,T)) = -\sum_l p_{0,t}(i,l)R(l,d) \tag{4.26}$$

$$= -\sum_l \pi_l\, p_{0,t}(i,T)R(l,d). \tag{4.27}$$

Now, if we multiply (4.27) by π_j and subtract from (4.25) we have,

$$\sum_l \pi_l\, p_{0,t}(i,T)R(l,j) + \pi_j \left(\sum_l \pi_l\, p_{0,t}(i,T)R(l,d)\right) = \pi_j \frac{d}{dt}(p_{0,t}(i,T)) - \pi_j \frac{d}{dt}(p_{0,t}(i,T)) \tag{4.28}$$

$$\sum_l \pi_l\, p_{0,t}(i,T)R(l,j) + \pi_j \left(\sum_l \pi_l\, p_{0,t}(i,T)R(l,d)\right) = 0 \tag{4.29}$$

$$\left(\sum_l \pi_l\, R(l,j) + \pi_j \left(\sum_l \pi_l\, R(l,d)\right)\right) p_{0,t}(i,T) = 0 \tag{4.30}$$

$$\sum_l \pi_l\, R(l,j) + \pi_j \left(\sum_l \pi_l\, R(l,d)\right) = 0 \tag{4.31}$$

$$\sum_l \pi_l\, R(l,j) + \pi_j \left(\sum_l \pi_l\, R(l,d)\right) = \frac{d}{dt}(\pi_j). \tag{4.32}$$

(4.32) follows since π_j is independent of time. Therefore, π can be interpreted as the stationary point of the forward equation (4.32). The first part in the LHS of (4.32) is the term from general Kolmogorov forward equation. In the second part, $\pi\left(\sum_l \pi_l\, R(l,d)\right)$ is the probability of hitting the absorbing state starting from a non-absorbing state distributed according to π. We can use this idea in algorithm 4.1 (Blanchet *et al.* [2012]). We simulate the chain starting from a non-absorbing state until it hits the absorbing state. Once it hits the absorbing state, we simulate the starting position based on the empirical density of non-absorbing states traced by the chain. As $t \to \infty$, samples drawn are from quasi-stationary distribution (Blanchet *et al.* [2012]).

Algorithm 4.1: Estimating Quasi-Stationary distribution(π_0)

1. Initialize the probability vector $\pi = \pi_0$ on the non-absorbing states of Markov chain.
2. Select a non-absorbing state of the Markov chain x_0 and set $X_0 = x_0$
3. Simulate the Markov chain normally starting with X_0 until absorption. Update π by counting the number of visits to each state until absorption. That is, we count the total number to visits to a state until current iteration and then we re-normalise the probability vector π.
4. Choose an initial position according to normalized vector π and go to step 3.
5. Steps 3. and 4. are repeated many times to get an estimate of quasi-stationary dist.

return (π)

Subsection 4.3 discusses the outline of the proof of algorithm 4.1 for discrete-time absorbing Markov chain defined on a discrete state-space. The proof uses stochastic approximation method which converges to quasi-stationary distribution of the Markov chain.

4.3 An outline of the proof of algorithm 4.1

In this section we outline the proof of algorithm 4.1 for the discrete-time absorbing Markov chain defined on a discrete state-space. A formal extension of algorithm 4.1 to 'discrete-time general state-space Markov chain' and 'continuous-time discrete state-space Markov chain' can be found in Zheng [2014]. But no such extensions exists for a continuous-time absorbing Markov chain defined on a general state-space.

Let S be the state-space of the Markov chain with $T \subset S$ being the set of transient states. Further let Q be the sub-stochastic matrix over the set of transient states T with π_n as the sequence of normalized probability vector over T after the n^{th} iteration of algorithm 4.1. Thus $\pi_n(x)$ stores the cumulative empirical distribution upto and including the n-th iteration of algorithm 4.1 for the transient state x. We define $\{X_k^n\}$ as the Markov chain used in the n-th iteration of the algorithm 4.1. Next, we define the hitting time of the absorbing state in the n-th iteration of the Markov chain $\{X_k^n\}_{k\geq 1}$ as $\tau^{(n)} = \min\{k \geq 0 : X_k^n \notin T\}$, then the iterative updating of π in step 3 of algorithm 4.1 is as follows:

$$\pi_{n+1}(j) = \frac{\left(\sum_{k=0}^{n} \tau^{(k)}\right)\pi_n(j) + \sum_{k=0}^{\tau^{(n+1)}-1} \mathbb{I}(X_k^{(n+1)} = j | X_0^{(n+1)} \sim \pi_n)}{\sum_{k=0}^{n+1} \tau^{(k)}}. \tag{4.33}$$

Equation (4.33) can be further reduced to

$$\pi_{n+1}(j) = \pi_n(j) + \frac{\sum_{k=0}^{\tau^{(n+1)}-1} \mathbb{I}(X_k^{(n+1)} = j | X_0^{(n+1)} \sim \pi_n) - \tau^{(n+1)}\pi_n(j)}{\sum_{k=0}^{n+1} \tau^{(k)}}. \tag{4.34}$$

The iterative procedure (4.34) is expressed as a stochastic approximation algorithm similar to (6.1):

$$\pi_{n+1}(j) = \pi_n(j) + \frac{1}{n+1} \frac{\sum_{k=0}^{\tau^{(n+1)}-1} \Big(\mathbb{I}(X_k^{(n+1)} = j | X_0^{(n+1)} \sim \pi_n) - \pi_n(j)\Big)}{\frac{1}{n+1}\sum_{k=0}^{n+1} \tau^{(k)}}. \tag{4.35}$$

The sequence of probability vector π_n takes values in a probability simplex. The Assumption $A.2$ (6.14) in the theorem (6.2) requires that the second summand in (4.35) depends only on π_n. But the denominator $\sum_{k=0}^{n+1} \tau^{(k)}$ in (4.35) depends on the whole history of sequence π_n. To circumvent this issue, Blanchet *et al.* [2012] adds another iterative state T_n by defining

$$T_n = \frac{1}{n+1} \sum_{k=0}^{n} \tau^{(k)}, \tag{4.36}$$

which iterates as

$$T_{n+1} = T_n + \frac{1}{n+2}(\tau^{(n+1)} - T_n). \tag{4.37}$$

Now (4.35) can be rewritten as

$$\pi_{n+1}(j) = \pi_n(j) + \frac{1}{n+1} \frac{\sum_{k=0}^{\tau^{(n+1)}-1} \Big(\mathbb{I}(X_k^{(n+1)} = j | X_0^{(n+1)} \sim \pi_n) - \pi_n(j)\Big)}{T_n + \frac{\tau^{(n+1)}}{n+1}}. \tag{4.38}$$

The term $\frac{\tau^{(n+1)}}{n+1}$ in the denominator is asymptotically negligible. We define

$$\mathbf{Y}_n^1(\pi, T)(j) := \mathbf{Y}_n^1(\pi(j), T) := \frac{\sum_{k=0}^{\tau-1} (\mathbb{I}(X_k = j | X_0 \sim \pi) - \pi(j))}{T + \frac{\tau}{n+1}}, \tag{4.39}$$

$$\mathbf{Y}_n^2(\pi, T) := \frac{n+1}{n+2}(\tau - T). \tag{4.40}$$

Further, we define $\mathbf{Y}_n(\pi, T) = (\mathbf{Y}_n^1(\pi, T), \mathbf{Y}_n^2(\pi, T))'$ where $\mathbf{Y}_n^1(\pi, T) \in \mathbb{R}^r$ is the vector consisting of $\mathbf{Y}_n^1(\pi, T)(j_i)$ and $\{j_1, ..., j_r\}$ is the set of the transient states. Thus,

$$\begin{pmatrix} \mathbf{ß}_{n+1} \\ \mathbf{T}_{n+1} \end{pmatrix} = \begin{pmatrix} \mathbf{ß}_n \\ \mathbf{T}_n \end{pmatrix} + \epsilon_n \times \begin{pmatrix} \mathbf{Y}_n^1(\pi, T) \\ \mathbf{Y}_n^2(\pi, T) \end{pmatrix}, \quad \text{where} \quad \epsilon_n = \frac{1}{n+1}. \tag{4.41}$$

Hence we are able to express (4.41) in the form of (6.1) :

$$\theta_n = \begin{pmatrix} \mathbf{ß}_n \\ \mathbf{T}_n \end{pmatrix}, \quad \mathbf{Y}_n = \begin{pmatrix} \mathbf{Y}_n^1(\pi, T) \\ \mathbf{Y}_n^2(\pi, T) \end{pmatrix} \tag{4.42}$$

where $\epsilon_n = \frac{1}{n+1}$ clearly satisfies (6.2). Here θ_n lies in $H = H^1 \times [0, \infty)$ where $H^1 = \{\mathbf{j} = (j_1, ..., j_r) \in \mathbb{R}_+^r | \sum_{i=1}^{r} \pi_n(j_i) = 1\}$ for each n.

Building on the foundations of the stochastic approximation algorithm given in theorem (6.2) as in Kushner & Yin [2003], Blanchet *et al.* [2012] proved the convergence result. Here, we state the first part of the result in Blanchet *et al.* [2012] and skip the rate of convergence analysis.

Theorem 4.3 *Given an irreducible absorbing Markov chain over a finite state space S, with Q as the sub-stochastic matrix over the transient states, let π_0 be an initially chosen probability vector over the non-absorbing states and $T_0 \geq 1$ be the initial value of T_n (T_n as defined in* (4.36)*). Then there exists a unique quasi-stationary distribution π such that*

$$\pi' Q = \lambda \pi' \text{ with } \pi = (\pi(j_1), ..., \pi(j_r))' \geq 0, \sum_{i=1}^{r} \pi(j_i) = 1. \tag{4.43}$$

Therefore the algorithm 4.1 with iterates $\theta_n = (\pi_n, T_n)'$ and initial values (π_0, T_0) converges with probability one to the point $(\pi, 1/(1-\lambda))$.

A similar result holds for the continuous-time absorbing Markov chain on a discrete state-space and discrete-time absorbing Markov chain on a general state-space. Although the existence of such a convergence result has not been proved for continuous-time absorbing Markov chain on a general state-space, we assume its validity in our context to construct the ReScaLE algorithm. With this leap of faith, we construct the ReScaLE algorithm in the next subsection.

4.4 ReScaLE Algorithm - Pseudocode

As pointed out earlier in section 4.1, quasi-stationary behaviour of a killed Brownian motion, with killing rate ϕ approximates the intractable distribution of interest. The essence of the ReScaLE method lies in the use of re-sampling algorithm 4.1 of Blanchet *et al.* [2012] to simulate from the quasi-stationary distribution of a killed Brownian motion. We are using a natural extension of algorithm 4.1 although no such extension has been proved in Blanchet *et al.* [2012]. Since it is not possible to simulate a Brownian motion continuously, we use finitely many points from its trajectory. Theorem 4.2 helps us to simulate from a Poisson process of rate M, which in turn is used to obtain a finite-dimensional representation of a Brownian motion at each event times of this Poisson process. We decide to kill the Brownian motion at time s with probability $\phi(X_s)/M$, where X_s is the simulated Brownian motion at time s. Once the Brownian motion is killed at some point $\mathbf{t_{kill}}$, we re-sample its starting position from the states visited by the Brownian motion. This is achieved by first simulating a **starting time** uniformly between 0 and $\mathbf{t_{kill}}$ and then simulating the **starting value** according to the law of a Brownian Bridge conditioned on the two nearest neighbours of **starting time**, continuing the process similarly from time $\mathbf{t_{kill}}$. This idea has been

summarized in algorithm 4.2.

Algorithm 4.2: ReScaLE Algorithm(μ, x_0)

1. $l \leftarrow \inf_{x\in\mathbb{R}} \frac{\mu^2+\mu'}{2}, \phi \leftarrow \frac{\mu^2+\mu'}{2} - l, M \leftarrow \sup_{x\in\mathbb{R}} \phi(x)$
2. $t_0 \leftarrow 0; X_{t_0} \leftarrow x_0$
3. **do** $\begin{cases} (t_1, t_2, ...) \sim \text{Poisson Process of rate } M \text{ starting at } t_0 \\ (X_{t_1}, X_{t_2}, ...) \sim \text{Brownian Motion started at position } X_{t_0} \\ \text{Kill the process at } X_{t_i} \text{ with probability } \phi(X_{t_i})/M \\ \textbf{exit} \text{ once kill occurs} \end{cases}$
4. **starting time** $\sim U[0, \mathbf{t_{kill}}]$
5. **starting value** $\sim$ Brownian Bridge conditioned on neighbors of **starting time**
6. **GOTO** 2. with $t_0 \leftarrow \mathbf{t_{kill}}; X_{t_0} \leftarrow$ **starting value**

return $((X_{t_1}, X_{t_2}, ...))$

The above algorithm can be stopped after a sufficiently long time. Samples obtained come from quasi-stationary distribution of a killed Brownian motion. In the next subsection we apply the ReScaLE algorithm to a toy problem.

Figure 7 depicts the working of the algorithm 4.2. We initialize the algorithm with $X_0 = 0$ and continue sampling from a Brownian motion at the event times of a Poisson process of rate M until it gets killed. Killing in the first segment has been depicted through the termination of the black chain. The black dotted lines connecting the points are the indicative Brownian Bridge. Once the Brownian motion is killed, it regenerates by first simulating uniformly between 0 and $\mathbf{t_{kill}}$ and then simulating the Brownian motion at the observed time point. The regeneration point is denoted by a red cross that corresponds to the x-axis value of the rear part of the arrow connected to the first red dot. Once the starting time is observed it simulates the value of the Brownian motion by simulating from a Brownian Bridge, conditioned on the two nearest neighbours of the observed starting time. It has been denoted by the first black star in the figure. This point now becomes the starting position for the next segment and the process continues in a similar manner.

Figure 8 shows three different segments of a killed Brownian motion. The chain in this situation has a quasi-stationary distribution given by π.

4.5 Applying the ReScaLE algorithm - an example

We consider a toy example consisting of 5 data points y_i which are simulated from a density

$$\pi(y|x) \propto \frac{1}{1+(y-x)^2}. \tag{4.44}$$

The simulated data is given as the following:

```
> y
[1]  2.65226687  1.27648783  1.61011759  1.27433040  0.08721209.
```

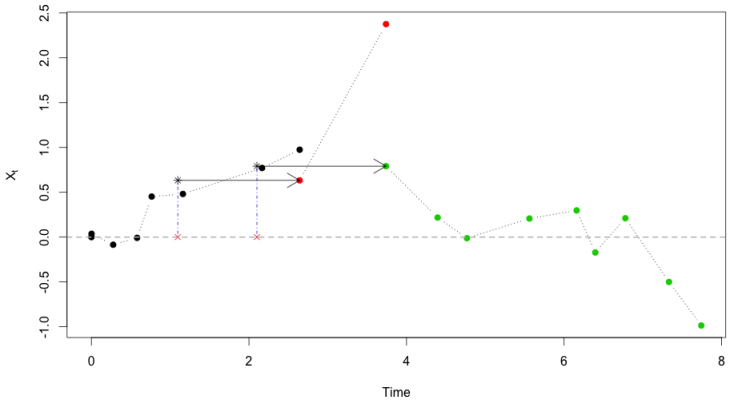

Figure 7: *This illustrates chain simulated using ReScaLE algorithm. It captures the situation when process is killed three times.*

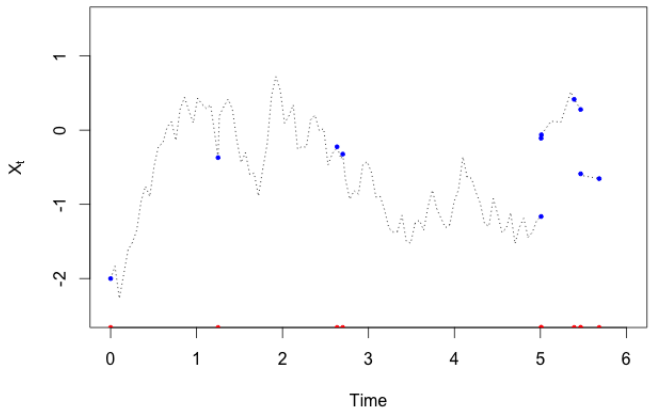

Figure 8: *Figure illustrates three different segments simulated using ReScaLE algorithm. Red points on time-axis are observed using Poisson process of rate M. Blue points are skeleton of chain while black dotted lines are Brownian bridges.*

We choose a standard Cauchy prior for the location parameter x. We are interested in simulating from the posterior density of the parameter x, which is given by

$$\pi(x|\mathbf{y} = (y_1, ..., y_5)) \propto \frac{1}{1+x^2} \prod_{i=1}^{5} \frac{1}{1+(y_i - x)^2}. \tag{4.45}$$

The drift function as in the ReScaLE algorithm is given by

$$\mu(x) = \frac{d}{dx} \log\left(\pi(x|\mathbf{y})\right) \tag{4.46}$$

$$= \sum_{i=1}^{5} \left(\frac{2(y_i - x)}{1+(y_i - x)^2} - \frac{1}{5} \frac{2x}{1+x^2} \right). \tag{4.47}$$

The derivative of the drift function is

$$\mu'(x) = \sum_{i=1}^{5} \left(\frac{-2(1-(y_i - x)^2)}{(1+(y_i - x)^2)^2} - \frac{1}{5} \frac{2(1-x^2)}{(1+x^2)^2} \right). \tag{4.48}$$

Based on numerical calculations, it can be observed that

$$\frac{\mu^2(x) + \mu'(x)}{2} \geq -2.379829.$$

Thus the rate of kill function ϕ can be defined as follows:

$$\phi(x) = \frac{\mu^2(x) + \mu'(x)}{2} + 2.379829.$$

The function ϕ is depicted in Figure 9. It is also clear from Figure 9 that $M = \max_{x \in \mathbb{R}} \phi(x) \leq 14$.

Figure 10 is a comparison of the quasi- stationary density obtained using the ReScaLE algorithm, to that of a numerically approximated density π. We approximate the normalising constant of the posterior density which is given by

$$\pi(x|\mathbf{y} = (y_1, ..., y_5)) = \frac{1}{0.06281079} \frac{1}{1+x^2} \prod_{i=1}^{5} \frac{1}{1+(y_i - x)^2}.$$

The kernel density is approximated based on samples obtained from the ReScaLE algorithm when it runs for a maximum time of 100000. The algorithm was initialised at the posterior mode, which is 1.191 in this case. It can be observed that the ReScaLE method approximates the true density really well.

4.5.1 Chain Diagnostics

In this section we examine how well the chain obtained using the ReScaLE method explores the support of the posterior distribution of interest. The mixing of the chain will be measured using the integrated autocorrelation time. The smaller value of the integrated autocorrelation time implies

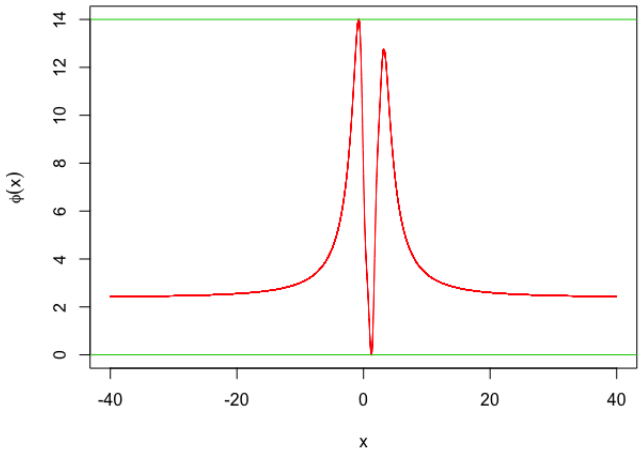

Figure 9: *This illustrates the 'rate of kill' function $\phi(x)$ together with its global extrema.*

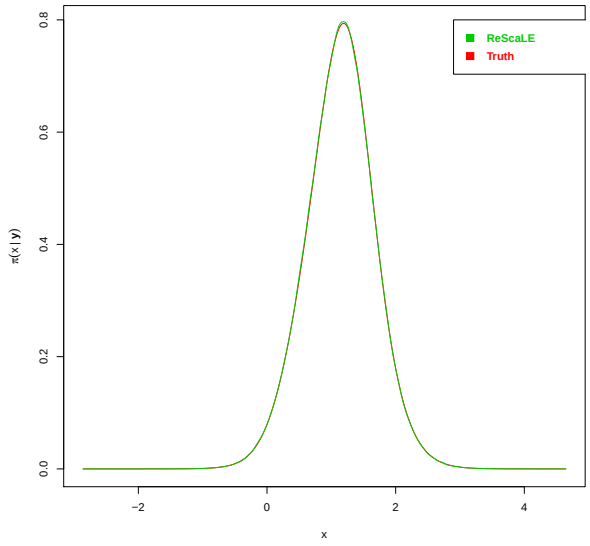

Figure 10: *This illustrates the kernel density approximation of samples obtained using the ReScaLE algorithm. Red curve is the true posterior density $\pi(x|\mathbf{y})$.*

that the chain explores the support quickly. For a scalar chain consisting of samples $X_0, X_1, ...$, the integrated autocorrelation time is defined as (see Straatsma *et al.* [1986] for details)

$$\tau_{int} = 1 + 2\sum_{i=1}^{\infty} \text{Corr}(X_0, X_i). \tag{4.49}$$

Informally, τ_{int} can be interpreted as the average number of samples required to obtain an independent sample. To estimate the integrated autocorrelation time we use the `IAT()` function in the R-package `LaplacesDemon`, which uses a random truncation of the infinite sum in (4.49) (Byron *et al.* [2016]). The integrated autocorrelation time for the sample obtained using the ReScaLE method approximates to 5.11892. The right plot in Figure 11 shows the convergence to τ_{int} by plotting the partial sum $\tau_{int}(l) = 1 + 2\sum_{i=1}^{l} \text{Corr}(X_0, X_i)$. Since $\tau_{int} = 5.11892$, which

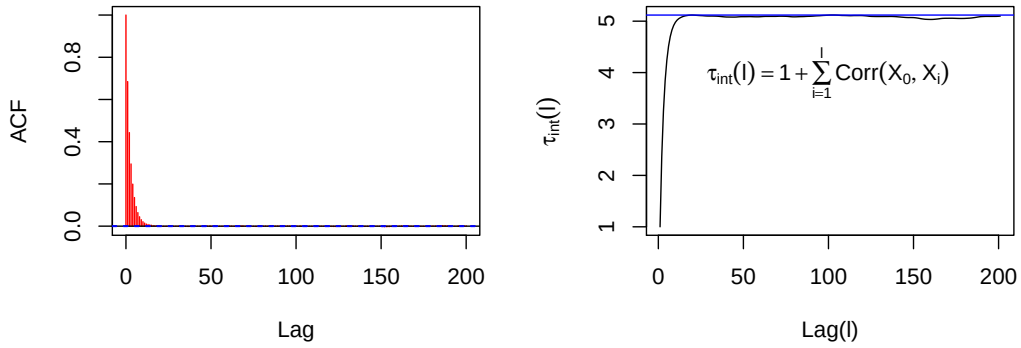

Figure 11: *The plot on the left illustrates the autocorrelation of the chain plotted against various lags. On the right* $\tau_{int}(l) = 1 + 2\sum_{i=1}^{l} Corr(X_0, X_i)$ *is plotted against lag*(l)*. The blue line indicates the integrated autocorrelation time* ($\tau_{int} = 5.11892$).

roughly means that every 6th sample in the chain will be independent of each other. For simplicity, we collect $\{X_j : j = 1 + 10i, j < N\}$ to diagnose the convergence of the empirical distribution $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{(-\infty,x]}(X_i)$ to the true posterior distribution $F(x)$. We use the Kolmogorov-Smirnov test statistic as the measure of discrepancy between the empirical distribution and the true posterior distribution. In this situation, the Kolmogorov-Smirnov test checks if the samples $\{X_j : j = 1 + 10i, j < N\}$ are drawn from the true posterior distribution. The p-value corresponding to this hypothesis testing is 0.491, which implies that the null hypothesis is accepted at a 5% level of significance. Thus, we can conclude that the samples obtained are from the true posterior distribution. This suggests that the kernel density approximation, based on samples obtained via the ReScaLE method approximates the true posterior distribution very well.

4.5.2 A case of 'non-uniform' rebirth strategy

In the previous subsections we assume that the 'regenerated time' is sampled uniformly between 0 and $\mathbf{t_{kill}}$. For a larger value of $\mathbf{t_{kill}}$, it is natural to assume that the samples obtained will be drawn from the quasi-stationary distribution. Considering this, we sample the regenerated time by sampling uniformly between $\lambda\mathbf{t_{kill}}$ and $\mathbf{t_{kill}}$ for some $0 \leq \lambda < 1$. Figure 13 shows a plot of the kernel density approximation of samples obtained using the ReScaLE method for different values of λ. We present the integrated autocorrelation time (IAT) and the p-value for the Kolmogorov-Smirnov test in Table 1. This empirical analysis suggests that a uniform distribution on $[\lambda\mathbf{t_{kill}}, \mathbf{t_{kill}}]$ can be chosen to sample the point of regeneration in the Glynn and Blanchet algorithm.

Case	λ	IAT	K-S p-value
1	0.0	5.119	0.4190
2	0.1	5.073	0.7419
3	0.2	5.095	0.5881
4	0.3	5.030	0.4435
5	0.4	5.072	0.2780
6	0.5	5.058	0.9691
7	0.6	5.029	0.6845
8	0.7	5.005	0.1104
9	0.8	5.005	0.4026
10	0.9	5.346	0.0002

Table 1: *IAT and K-S p-value table for varying values of λ for uniform density on the interval* $[\lambda\mathbf{t_{kill}}, \mathbf{t_{kill}}]$

Further, a natural question arises: can we regenerate time non-uniformly on the interval $[\lambda\mathbf{t_{kill}}, \mathbf{t_{kill}}]$? We test this assertion empirically by choosing a non-uniform distribution on the interval $[\lambda\mathbf{t_{kill}}, \mathbf{t_{kill}}]$. We choose a distribution such that it puts more weight towards $\mathbf{t_{kill}}$ so that more often quasi-stationary states are drawn. For simplicity, we choose a triangular and a quadratic distribution as presented in Figure 12. The empirical results involving the integrated autocorrelation time and the Kolmogorov-Smirnov p-values for different values of λ are given in Table 2. This table suggests that the integrated autocorrelation time is roughly close to 5, which was already observed for the case of uniform density in Table 1. We also noted in the previous subsection that the integrated autocorrelation time was close to 5 when no 'rebirth-strategies' were involved. This suggests that the chain mixes well even when non-uniform rebirth strategies are invoked. It should be noted that Kolmogorov-Smirnov p-value is mostly 'large' which implies that the convergence to the true posterior distribution is achieved in these situations as well.

4.5.3 Sub-sampling within ReScaLE

The strength of the ScaLE method lies in the fact that it scales very well with the size of the data (Pollock *et al.* [2016]). The scalability of the algorithm comes from the use of the sub-sampling ideas. In the context of the ReScaLE algorithm outlined earlier, we kill a Brownian motion with a killing rate ϕ. In a big-data set-up however, calculation of ϕ can be computationally expensive. So, instead of working with ϕ, we choose an unbiased estimator of ϕ as the rate of kill function (Pollock

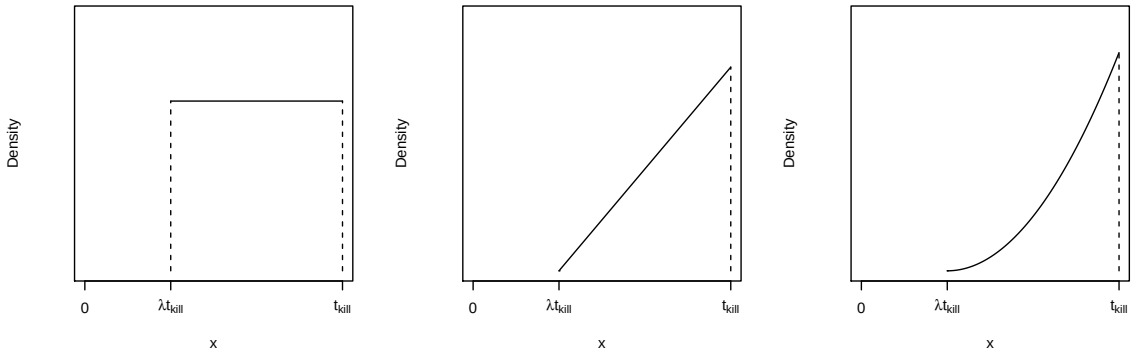

Figure 12: *This illustrates three types of densities considered for drawing regenerated times on* $[\lambda \mathbf{t_{kill}}, \mathbf{t_{kill}}]$. *On the left the uniform density* $f(x) = \frac{1}{(1-\lambda)t_{kill}}$, *in the middle the triangular density* $f(x) = \frac{2(x-\lambda t_{kill})}{(1-\lambda)^2 t_{kill}^2}$ *while on the right the quadratic density* $f(x) = \frac{3(x-\lambda t_{kill})^2}{(1-\lambda)^3 t_{kill}^3}$ *is plotted.*

Table 2: *IAT and K-S p-value table for varying values of* λ *for non-uniform densities on the interval* $[\lambda \mathbf{t_{kill}}, \mathbf{t_{kill}}]$

		Triangular distribution		**Quadratic distribution**	
Case	λ	IAT	K-S p-value	IAT	K-S p-value
1	0.0	5.034	0.0007	5.017	0.6384
2	0.1	5.026	0.0078	5.022	0.4560
3	0.2	5.043	0.2799	5.063	0.2450
4	0.3	5.055	0.4609	4.987	0.6823
5	0.4	5.031	0.6121	4.974	0.1766
6	0.5	5.016	0.0006	5.039	0.0444
7	0.6	5.019	0.7457	5.035	0.9115
8	0.7	5.038	0.6090	4.987	0.6047
9	0.8	5.028	0.4848	5.029	0.2785
10	0.9	4.984	0.9220	5.001	0.2481

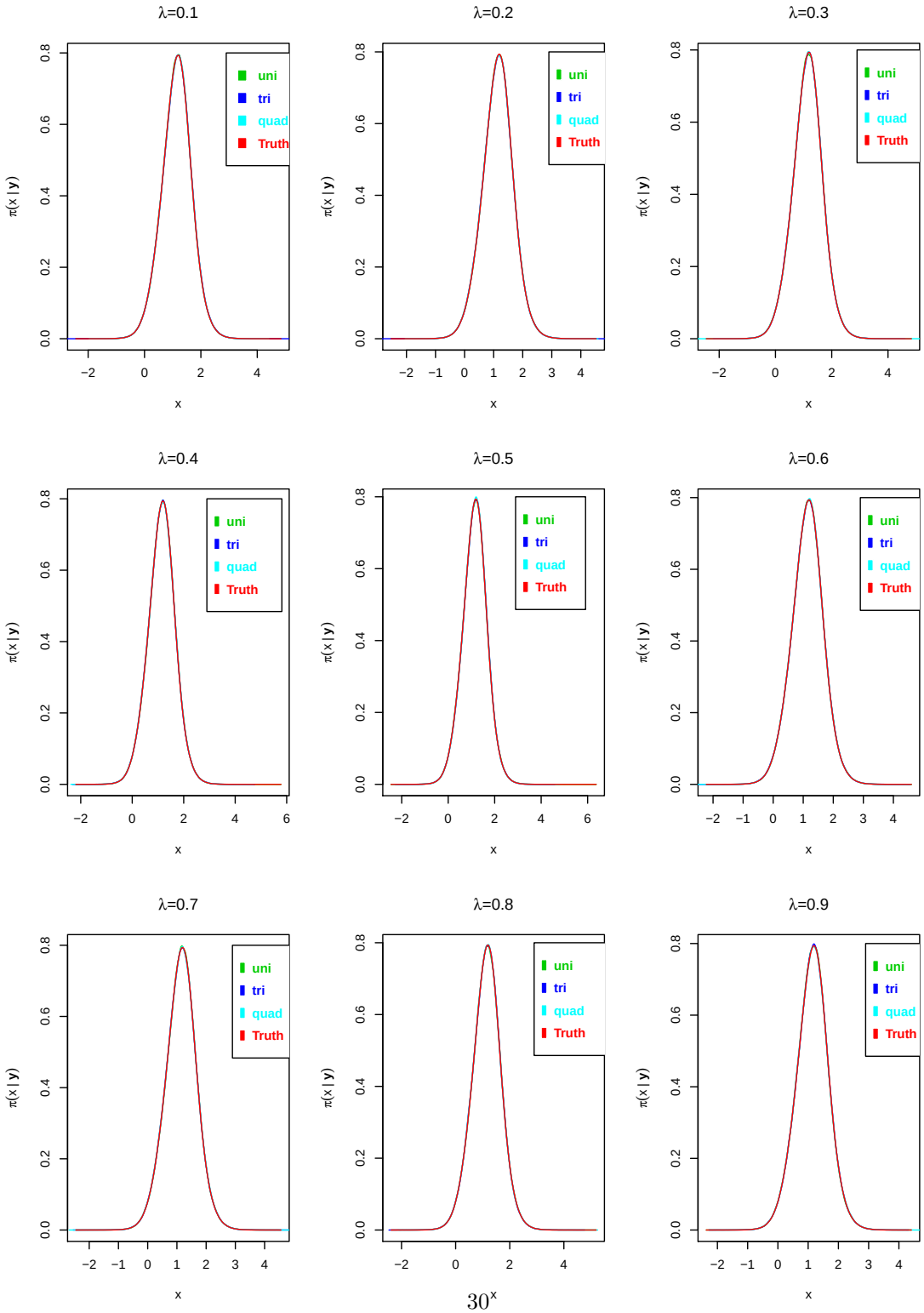

Figure 13: *This illustrates the kernel density approximation of samples obtained by regenrating according to different rebirth strategies ('uni' stands for uniform, 'tri' for triangular and 'quad' stands for quadratic density as in Figure 12) within the ReScaLE method. The time of regeneration is chosen according to these densitties on* $[\lambda \mathbf{t_{kill}}, \mathbf{t_{kill}}]$*, for different values of lambda within the ReScaLE algorithm . The red curve is the true posterior density* $\pi(x|\mathbf{y})$.

et al. [2016]). The form of the ϕ function in our case allows us to use the idea of sub-sampling. To illustrate this, we consider a posterior density of the form

$$\pi(x|\mathbf{y} = (y_1, ..., y_N)) \propto p(x) \prod_{i=1}^{N} \pi(y_i|x). \tag{4.50}$$

The logarithm of the posterior density is given by

$$\log(\pi(x|\mathbf{y})) = \log(p(x)) + \sum_{i=1}^{N} \log(\pi(y_i|x)) \tag{4.51}$$

$$= \sum_{i=1}^{N} \left(\log(\pi(y_i|x)) + \frac{1}{N} \log(p(x)) \right) \tag{4.52}$$

$$:= \sum_{i=1}^{N} f_i(x). \tag{4.53}$$

Now it can be observed that the drift function as in the ReScaLE algorithm is given by

$$\mu(x) = \sum_{i=1}^{N} f_i'(x), \tag{4.54}$$

while its derivative is

$$\mu'(x) = \sum_{i=1}^{N} f_i''(x) = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{N} f_i''(x). \tag{4.55}$$

Thus, the rate of kill function ϕ evaluates to

$$\phi(x) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \left(f_i'(x) f_j'(x) + \frac{1}{N} f_i''(x) \right)}{2} - l, \quad \text{where} \quad l = \inf_x \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \left(f_i'(x) f_j'(x) + \frac{1}{N} f_i''(x) \right)}{2}. \tag{4.56}$$

For independent discrete random variables I, J simulated uniformly on $\{1, ..., N\}$, we define the random variable

$$\phi_{I,J}(x) := \frac{N^2 f_I'(x) f_J'(x) + N f_K''(x)}{2} - l. \tag{4.57}$$

It is easy to observe that

$$\mathbb{E}(\phi_{I,J}(x)) = \phi(x). \tag{4.58}$$

We can choose a constant M' such that $\phi_{I,J}(x) \leq M'$ for all x, I, J. Therefore, we kill the process at x with probability $\phi_{I,J}(x)/M'$ (Pollock *et al.* [2016]). In the big data set -up, instead of calculating the large sum involved in the calculation of ϕ, we work with an unbiased estimator of it (Pollock *et al.* [2016]). Next, we consider the example from an earlier subsection incorporated

with the sub-sampling idea. Figure 14 shows a comparison of the kernel density approximation, based on samples obtained through the use of sub-sampling with the ReScaLE algorithm and the true posterior density. The samples were obtained based on a run of the ReScaLE method for a maximum time equal to 100000. It is evident from the *ordered* plots of samples of different sizes in Figure 15 that when subsampling is employed, the algorithm would take a long time to converge. The integrated autocorrelation time for the chain approximates to 9.100 as opposed to 5.119 when subsampling is not employed. This implies that under the subsampling method, chain explores the support of the posterior distribution slowly.

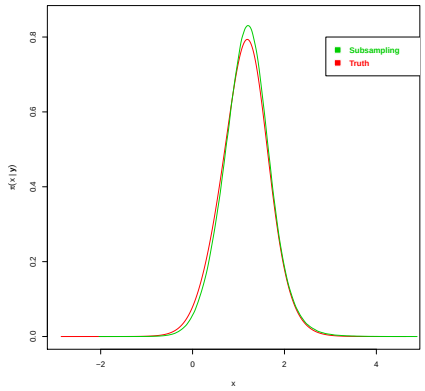

Figure 14: *This illustrates the kernel density approximation of samples obtained using subsampling within ReScaLE algorithm. The red curve is the true posterior density* $\pi(x|\mathbf{y})$.

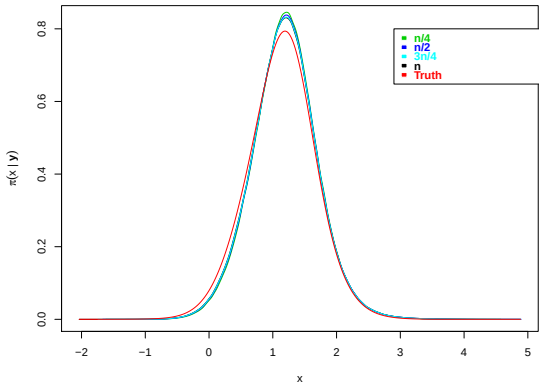

Figure 15: *This illustrates the kernel density approximation of samples of sizes* $\lfloor \frac{n}{4} \rfloor$, $\lfloor \frac{n}{2} \rfloor$, $\lfloor \frac{3n}{4} \rfloor$ *and* $n = 8945042$ *obtained using subsampling within the ReScaLE algorithm. The red curve is the true posterior density* $\pi(x|\mathbf{y})$.

5 Concluding remarks and possible further research

This report offers an introduction to the ReScaLE method, which uses a modern methodological advancement for exploring the quasi-stationary behaviour of an absorbing Markov chain. Compared to ScaLE, the ReScaLE method provides an alternative framework for simulating from an intractable distribution. The ScaLE method is a sequential Monte Carlo (SMC) based approach, which uses a large number of particles to converge to the quasi-stationary distribution of a killed Brownian motion. Comparatively, the ReScaLE method uses a single particle which evolves as a Brownian motion. Once killed, it regenerates again by sampling from the states visited by the particle. The unknown intractable distribution is then approximated by the quasi-stationary distribution of a killed Brownian motion. Through a toy example we noted that the chain exhibits a good mixing property and thus it explores the support of the intractable distribution well.

The current implementation of the ReScaLE method relies on a bounded rate of kill function ϕ. However, this idea can be extended to a situation where ϕ can possibly be unbounded. It is impossible to find a finite bound M for function ϕ over the entire path-space of a Brownian motion. Thus, we work over a finite interval $[l, u]$ (often called a *layer*) in the path-space of a Brownian motion, which guarantees that a bound $M_{[l,u]}$ exists for function ϕ over the interval $[l, u]$ (Beskos *et al.* [2008]; Pollock [2013]). This leads to the notion of a *layered Brownian motion*, which can be used as a proposal mechanism to design the ReScaLE method for an unbounded ϕ function (Pollock *et al.* [2016]). The proposal points at a Poisson event time (of rate $M_{[l,u]}$) can be simulated according to a layered Brownian motion, which is killed with probability $\phi/M_{[l,u]}$. For a Brownian motion starting at x_0 we construct a layer of width $2L$ ($L > 0$) until the first exit time for a Brownian motion on the layer $[x_0 - L, x_0 + L]$ (Burq & Jones [2008]; Chen & Huang [2013]; Pollock *et al.* [2016]). The proposal points can then be simulated according to the law of a three-dimensional Bessel bridge conditioned to remain inside the layer $[x_0 - L, x_0 + L]$ (Chen & Huang [2013]; Pötzelberger & Wang [2001]; Bertoin *et al.* [1999]).

As pointed in an earlier section, sub-sampling enables the ScaLE method to be employed into big data problems. ScaLE employs control-variate ideas with sub-sampling which results in sub-linear computational cost with respect to the size of data. This motivates us to question if the same can be achieved for the ReScaLE method. As compared to ReScaLE, the ScaLE method employs a large number of SMC particles; it would be interesting to note if the ReScaLE method can achieve an enhanced performance. We noted that under the subsampling method, chain gives rise to a higher integrated autocorrelation time. This implies that chain explores the support of the posterior distribution slowly as compared to a situation when the subsampling method is not employed.

We noted in an earlier section that the regenerative step in Glynn and Blanchet's algorithm can be tweaked slightly while convergence can be still achieved. We employed a non-uniform regeneration mechanism to a toy example. It was evident from the empirical analysis that the chain still exhibits good mixing properties and explores the support of the posterior distribution well. Further, the Kolmogorov-Smirnov test showed that the empirical distribution of the samples obtained converge to the true posterior distribution. This encourages us to explore the theoretical aspects of non-uniform rebirth strategies. The non-uniform rebirth strategy was employed with a view to sample more often from quasi-stationary states of a killed Brownian motion. Thus, it can be argued that this method could perform better in terms of its convergence to the quasi-stationary density. This leads to a question of whether or not a performance gain can be achieved for a big-data problem. We did not touch the algorithmic aspect of the ReScaLE method; can we overcome

the problem of a poor start? This is a problem of profound interest since Monte Carlo algorithms are often 'stuck' if they had a poor start. It would be interesting to note whether a non-uniform regenerative strategy can be employed to circumvent this issue. Secondly, we can explore other adaptive ways to recover from a poor start. If the ReScaLE method can be adapted to recover from a poor start, it would lead to a performance gain. So, we shall compare the efficiency of the ScaLE method with respect to an adaptive version of ReScaLE.

My possible thesis structure will be as follows:

Part - I : **Background and Motivation**

We would introduce the problem of sampling from an intractable distribution with a view of its application to big data context. Subsequently, we would lay down some current methodologies which address this issue, highlighting the challenges that these methodologies face when it comes to a big data set-up. Further, we would discuss how ScaLE resolves these issues and provide a rationale for the ReScaLE method as an alternative way to address similar issues.

Part - II : **The methodology**

Here, we would outline the theoretical aspects of the ReScaLE method. In particular, we will discuss the methodology for a general ϕ function, which would involve a layered Brownian motion proposal mechanism. Further, we would establish the results concerning non-uniform rebirth strategies together with their possible algorithmic advantages. Subsequently, this would include a discussion on an efficient algorithmic design of the ReScaLE method. We will explore ways in which the performance of the subsampling method can be enhanced for faster convergence to the unknown distribution.

Part - III : **Empirical results and conclusion**

Here we will consider an application of the ReScaLE method involving a big data problem. We will compare and contrast the efficiency of the method with respect to the existing ScaLE algorithm. Lastly, we will conclude by outlining potential research directions on this subject.

6 Appendix

6.1 Stochastic Approximation method

6.1.1 An overview

The basic stochastic approximation algorithm first introduced in Robbins & Monro [1951] was motivated by the problem of finding the root of a continuous function $g(\theta)$, where θ takes values in some Euclidean space $\mathbb{R}^r$. Later, Kushner & Clark [1978] showed that such a root can be linked to the solutions of an ODE. The function $g(\theta)$ is not known exactly rather we have access to 'noise-corrupted' observations Y of $g(\theta)$. The basic set-up in Robbins & Monro [1951] uses a recursive method to access the root $\bar{\theta}$ of $g(\theta)$ via a stochastic difference equation such as,

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n, \tag{6.1}$$

where the step size sequence $\epsilon_n > 0$ satisfies,

$$\sum_n \epsilon_n = \infty, \quad \sum_n \epsilon_n^2 < \infty, \quad \epsilon_n \to 0. \tag{6.2}$$

The random variable Y_n is a function of the noise-corrupted observations of $g(\theta)$ at θ_n. Under suitable stability conditions (see Robbins & Monro [1951]; Han-Fu [2002]; Kushner & Yin [2003] for details), $\theta_n \to \bar{\theta}$ as $n \to \infty$. In a more general setting, iterates θ_n may be constrained to a compact set H. If iterates ever leaves H, then algorithm (6.1) can be defined as:

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n + \epsilon_n Z_n. \tag{6.3}$$

Here $\epsilon_n Z_n$ is added (see [Kushner & Yin, 2003, Section 4.3] for details) to restrict $\theta_n + \epsilon_n Y_n$ back to the constraint set H if it ever leaves H otherwise $\epsilon_n Z_n$ is zero. Here we assume that 'noise' in the observation is a martingale difference i.e., there is a measurable function $g_n(\cdot)$ of θ and random variable β_n such that Y_n can be decomposed as

$$Y_n = g_n(\theta_n) + \delta M_n + \beta_n, \text{where} \tag{6.4}$$

$$\delta M_n = Y_n - E[Y_n|\theta_0, Y_i, i < n] \quad \text{and} \tag{6.5}$$

$$\beta_n = E[Y_n|\theta_0, Y_i, i < n] - g_n(\theta_n). \tag{6.6}$$

Here δM_n is the 'martingale difference' noise[4] while the sequence $\{\beta_n\}$ is asymptotically negligible. The function $g_n(\cdot)$ may or may not depend on n. We shall discuss the relationship between $g_n(\cdot)$ and $g(\cdot)$ when we lay down assumptions needed to get our main result. It can be shown (Kushner & Clark [1978]; Kushner & Yin [2003]) that under reasonable assumptions on β_n and $g_n()$, θ_n is related to the stationary points of ODE

$$\dot{\theta}(t) = g(\theta(t)). \tag{6.7}$$

Here we explain the idea briefly.

[4]Let X_t be a measurable sequence with respect to filtration $\mathcal{F}_t$ defined on the probability space $(\mathbf{\Omega}, \mathcal{F}, \mathbb{P})$. X_t is a martingale difference noise if it satisfies: (1) $E(X_t) < \infty$ and (2) $E(X_t \,|\, \mathcal{F}_s, s < t) = 0$

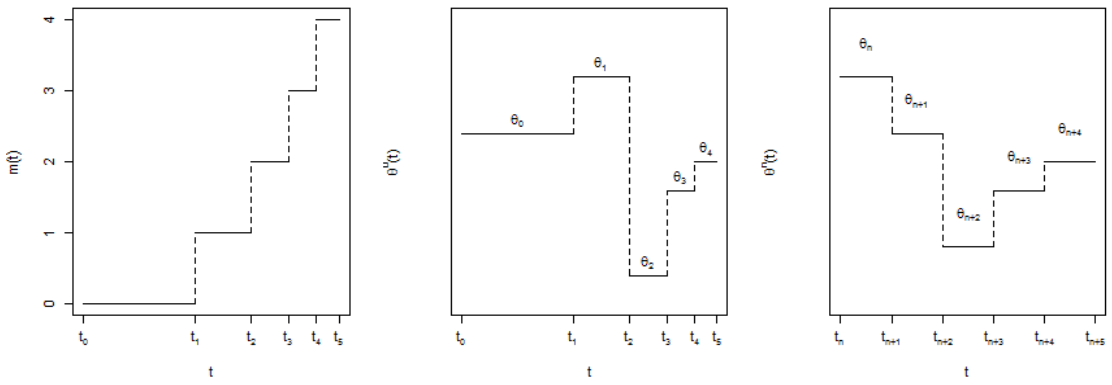

Figure 16: *This illustrates the function* $m(t), \theta^0(t)$ *and* $\theta^n(t)$ *respectively for* $\epsilon_n = 1/n$

6.1.2 The ODE method and Kushner's theorem

In this section we briefly outline the rationale behind the convergence of recursive method (6.3) and its relation to the stationary points of an ODE (6.7). This relation can be understood by considering a continuous-time interpolation of the estimates $\{\theta_n\}_{n\geq 1}$ generated by the recursive methods (6.3) with interpolating length equal to step sizes $\{\epsilon_n\}_{n\geq 1}$. Kushner & Yin [2003] showed that the tail part of the interpolating function satisfies (6.7). To understand the idea of interpolation we define,

$$t_0 = 0, \quad \text{and} \quad t_n = \sum_{i=0}^{n-1} \epsilon_i. \tag{6.8}$$

Further we define,

$$m(t) = \begin{cases} \max\{n : t_n \leq t < t_{n+1}\} & \text{if} \quad t \geq 0, \\ 0 & \text{if} \quad t < 0. \end{cases} \tag{6.9}$$

Next we define the continuous time interpolation $\theta^0(t)$ for $t \in \mathbb{R}$ by,

$$\theta^0(t) = \begin{cases} \theta_n & \text{if} \quad t_n \leq t < t_{n+1}, \\ \theta_0 & \text{if} \quad t \leq 0. \end{cases} \tag{6.10}$$

and the sequence of shifted functions defined by,

$$\theta^n(t) = \theta^0(t_n + t), \quad t \in \mathbb{R}. \tag{6.11}$$

Figures (16) illustrates the function $m(t)$, interpolations $\theta^0(t)$ and $\theta^n(t)$.

We are interested in the behaviour of the tail part of the iterates θ_n which is equivalent to the behaviour of $\theta^n(\cdot)$ for large n over a finite interval. Let $Z_i = 0$ and define $Z^n(\cdot)-$ a continuous-time interpolating functions of the terms $\epsilon_n Z_n$ in (6.3) by

$$Z^0(t) = \begin{cases} 0 & \text{if} \quad t \leq 0, \\ \sum_{i=0}^{m(t)-1} \epsilon_i Z_i & \text{if} \quad t > 0. \end{cases} \tag{6.12}$$

$$Z^n(t) = \begin{cases} \sum_{i=m(t_n)}^{m(t_n+t)-1} \epsilon_i Z_i & \text{if} \quad t \geq 0, \\ -\sum_{i=m(t_n+t)}^{m(t_n)-1} \epsilon_i Z_i & \text{if} \quad t < 0. \end{cases} \tag{6.13}$$

The continuous-time interpolations $\{\theta^n(\cdot), Z^n(\cdot)\}$ will be used to to present an important stochastic approximation result needed to prove algorithm (4.1). Next we lay down the assumptions for the main result of this subsection (due to Kushner & Yin [2003]).

A.1 $\sup_n E|Y_n|^2 < \infty$ where Y_n is as in (6.4).

A.2 There are functions $g_n(\cdot)$ which are continuous uniformly in n, a continuous function $g(\cdot)$ and a random variable $\beta_n \to 0$ w.p 1 such that

$$E[Y_n|\theta_0, Y_i, i < n] = g_n(\theta_n) + \beta_n, \tag{6.14}$$

and for each $\theta \in H$,

$$\lim_{n\to\infty} \left| \sum_{i=m(t_n)}^{m(t_n+t)} \epsilon_i (g_i(\theta) - g(\theta)) \right| \to 0, \quad \forall\, t > 0. \tag{6.15}$$

A.3 For the step size sequence $\epsilon_n > 0$, (6.2) holds .

Assumption $A.2$ ensures that the 'noise' in the tails of interpolated function $\theta^n(\cdot)$ converges to zero.

The stochastic approximation result relies on the extended version of Arzelà-Ascoli theorem for its completeness. We lay down down the important definition and result involving extended Arzelà-Ascoli theorem.

Definition 6.1 (Extended equicontinuity) *A sequence of function $f_n(\cdot)$ on $\mathbb{R}$ with bounded $f_n(0)$ is equicontinuous in the extended sense if for every $T > 0$ and $\epsilon > 0$, $\exists$ a $\delta > 0$ such that*

$$\limsup_n \sup_{|t-s|\leq\delta, |t|<T} |f_n(t) - f_n(s)| \leq \epsilon.$$

Lemma 6.1 (Extended Arzelà-Ascoli) *A sequence of extended-equicontinuous functions $\{f_n(\cdot)\}$ on $\mathbb{R}$ has a subsequence that converges to some continuous limit uniformly on each bounded interval.*

For the purpose of this report we present the stochastic approximation result without proof.

Theorem 6.2 *Let $(A.1) - (A.3)$ hold for the algorithm (6.3). Then there exists a null set N such that for $\omega \notin N$, the set of functions $\{\theta^n(\omega, \cdot), Z^n(\omega, \cdot)\}$ for $n < \infty$ is equicontinuous in the extended sense. Let $(\theta(\omega, \cdot), Z(\omega, \cdot))$ denote the limit of some convergent subsequence. Then this pair satisfies the projected ODE,*

$$\dot{\theta}(t) = g(\theta(t)) + z. \tag{6.16}$$

$\{\theta_n(\omega)\}$ converges to some limit set of the ODE in H, where z is added to restrict the solution in H. If $A \subset H$ is locally asymptotically stable in the sense of Liapunov[5] *for projected ODE* (6.16) *and θ_n is in some compact subset in the domain of attraction*[6] *of A infinitely often with probability at least υ, then $\theta_n \to A$ probability at least equal to υ.*

A formal proof of theorem (6.2) can be found in [Kushner & Yin, 2003, Chapter 5].

6.2 A proof of Theorem 4.2

We first establish the following lemma 6.3 to prove theorem 4.2. We consider a partition $P_n := \{0 = t_0 < t_1 < ... < t_n = t\}$ of interval $[0, t]$. We define the size of the partition by $|P_n| := \sup_{i=1,..,n} |t_i - t_{i-1}|$. The choice of the partition is made such that $\lim_{n\to\infty} |P_n| \longrightarrow 0$, which implies that for any $\delta > 0$ there exists $N_0 \in \mathbb{N}$

$$\sup_{i=1,..,n} |t_i - t_{i-1}| < \delta \quad \forall\, n \geq N_0 \tag{6.17}$$

$$(t_i - t_{i-1})(t_j - t_{j-1}) = o(\delta^2) \to 0 \quad \forall\, i \,\&\, j \tag{6.18}$$

Assumption (6.18) ensures that product of two or more intermediate lengths in this partition is negligible.

Lemma 6.3 *Let $\{N(t) : t \geq 0\}$ be the homogeneous Poisson Process of rate M. Let $X_{\tau_1}, X_{\tau_2}, ...$ are the realizations of Brownian motion $\{X_t : t \geq 0\}$ at $\tau_1, \tau_2, ...$, with $M = \sup_{X_t | t \geq 0} \phi(X_t)$. If the Brownian motion started at 0 is killed at τ_i with probability $\frac{\phi(X_{\tau_i})}{M}$. If ζ denotes the random variable associated with the instance when the Brownian motion is killed then the survival probability until time t is given by*

$$P(\zeta > t,\, N(t) = k \,|\, (X_s)_{0 \leq s \leq t}) = \exp\{-Mt\} \left(Mt - \int_0^t \phi(X_s) ds \right)^k \Big/ k! \tag{6.19}$$

Proof Consider a partition $0 = t_0 < t_1 < t_2 ... < t_n = t$ of the interval $[0, t]$. For $i_1 \neq i_2 \neq ... \neq i_k \in \{0, 1, ..., n\}$. Let $I = \{i_1, ..., i_k\}$. Survival of the process until time t is governed by the survival

[5] A set $A \subset H$ is said to be locally stable in the sense of Liapunov if for each $\delta > 0$, $\exists\, \delta_1 > 0$, $\delta > \delta_1$ such that all trajectories starting in δ_1 neighbourhood of A i.e $N_{\delta_1}(A)$ never leave $N_\delta(A)$.

[6] The domain of attraction of set $A \subset H$ is the set of all points in H for which sequences starting with these points converge to A.

of the process at the event times $t_{i_1}, ..., t_{i_k}$.

$$P(\zeta > t,\, N(t) = k \,|\, (X_s)_{0\le s\le t})$$

$$= \lim_{n\to\infty} \sum_{i_1<i_2<\ldots<i_k} \overbrace{\left(\prod_{j=1}^{k} M(t_{i_j} - t_{i_j-1})\right)}^{P(\text{Finding } k \text{ events})} \overbrace{\left(\prod_{j=1}^{k}\left(1 - \frac{\phi(X_{t_{i_j}})}{M}\right)\right)}^{P(\text{Not killing the process at event times})} \overbrace{\prod_{j\notin I}(1 - M(t_{j+1} - t_j))}^{P(\text{Rest } n-k \text{ intervals observes no events})} \tag{6.20}$$

$$= \lim_{n\to\infty} \sum_{i_1<i_2<\ldots<i_k} \left(\frac{M(t_{i_j} - t_{i_j-1})}{1 - M(t_{i_j} - t_{i_j-1})}\left(1 - \frac{\phi(X_{t_{i_j}})}{M}\right)\right) \prod_{j=1}^{n}(1 - M(t_j - t_{j-1})) \tag{6.21}$$

$$= \lim_{n\to\infty} \prod_{j=1}^{n}(1 - M(t_j - t_{j-1})) \lim_{n\to\infty} \sum_{i_1<i_2<\ldots<i_k} \prod_{j=1}^{k}\left(\frac{M(t_{i_j} - t_{i_j-1})}{1 - M(t_{i_j} - t_{i_j-1})}\left(1 - \frac{\phi(X_{t_{i_j}})}{M}\right)\right) \tag{6.22}$$

$$= \lim_{n\to\infty} \left(\prod_{j=1}^{n}(1 - M(t_j - t_{j-1}))\right) \lim_{n\to\infty} \sum_{i_1<i_2<\ldots<i_k} \prod_{j=1}^{k} M(t_{i_j} - t_{i_j-1})(1 + M(t_{i_j} - t_{i_j-1}))\left(1 - \frac{\phi(X_{t_{i_j}})}{M}\right) \tag{6.23}$$

Equation (6.21) follows due to multiplication and division of terms $(1 - M(t_j - t_{j-1}))$. (6.23) holds using (6.18) since $(1 - M(t_j - t_{j-1}))^{-1} = (1 + M(t_j - t_{j-1}))$. Assumption (6.18) ensures that the following approximation holds,

$$\lim_{n\to\infty}\left(\prod_{j=1}^{n}(1 - M(t_j - t_{j-1}))\right) = \exp\{-Mt\} \quad \text{and} \tag{6.24}$$

$$\lim_{n\to\infty} \sum_{i_1<i_2<\ldots<i_k} \prod_{j=1}^{k}\left(M(t_{i_j} - t_{i_j-1})(1 + M(t_{i_j} - t_{i_j-1}))\left(1 - \frac{\phi(X_{t_{i_j}})}{M}\right)\right) \tag{6.25}$$

$$= \lim_{n\to\infty} \sum_{i_1<i_2<\ldots<i_k} M(t_1 - t_0)(1 + M(t_1 - t_0))...M(t_k - t_{k-1})(1 + M(t_k - t_{k-1}))\left(1 - \frac{\phi(X_{t_0})}{M}\right)...\left(1 - \frac{\phi(X_{t_{k-1}})}{M}\right) \tag{6.26}$$

$$= \lim_{n\to\infty} \sum_{i_1<i_2<\ldots<i_k} M(t_1 - t_0)...M(t_k - t_{k-1}) + (-1)^k \phi(X_{t_0})(t_1 - t_0)...\phi(X_{t_{k-1}})(t_k - t_{k-1}) \tag{6.27}$$

$$= \lim_{n\to\infty}\left(\sum_{j=1}^{n} M(t_j - t_{j-1}) - \phi(X_{j-1})(t_j - t_{j-1})\right)^k \Big/ k! \tag{6.28}$$

$$= \left(Mt - \int_0^t \phi(X_s)ds\right)^k \Big/ k! \tag{6.29}$$

Consequently we have,

$$P(\zeta > t,\, N(t) = k \,|\, (X_s)_{0\le s\le t}) = \exp\{-Mt\}\, \frac{\left(Mt - \int_0^t \phi(X_s)ds\right)^k}{k!}. \tag{6.30}$$

Proof of theorem 4.2: Using the previous lemma we have,

$$P(\zeta > t,\, N(t) = k \,|\, (X_s)_{0\le s\le t}) = \exp\{-Mt\}\, \frac{\left(Mt - \int_0^t \phi(X_s)ds\right)^k}{k!} \tag{6.31}$$

$$P(\zeta > t \,|\, (X_s)_{0\le s\le t}) = \sum_{k=1}^{\infty} P(\zeta > t,\, N(t) = k \,|\, (X_s)_{0\le s\le t})) \tag{6.32}$$

$$= \sum_{k=1}^{\infty} \exp\{-Mt\}\, \frac{\left(Mt - \int_0^t \phi(X_s)ds\right)^k}{k!} \tag{6.33}$$

$$= \exp\{-Mt\} \sum_{k=1}^{\infty} \frac{\left(Mt - \int_0^t \phi(X_s)ds\right)^k}{k!} \tag{6.34}$$

$$= \exp\{-Mt\} \exp\left\{Mt - \int_0^t \phi(X_s)ds\right\} \tag{6.35}$$

$$= \exp\left\{-\int_0^t \phi(X_s)ds\right\}. \tag{6.36}$$

6.3 A proof of Theorem 4.1

For a Brownian motion $\{X_t : t \ge 0\}$, which is killed at rate $\phi(X_s)$ at time s; we have observed that

$$P(\zeta > t, \,|\, (X_s)_{0\le s\le t}) = \exp\left\{-\int_0^t \phi(X_s)ds\right\}, \tag{6.37}$$

where ζ denotes the random variable associated with the instance when the Brownian motion is killed. Thus the marginal survival probability is given by

$$P(\zeta > t \,|\, (X_0, X_t)) = \mathbb{E}_{X_0,X_t}\left[\exp\left\{-\int_0^t \phi(X_s)ds\right\}\right], \tag{6.38}$$

where $\mathbb{E}_{X_0,X_t}$ denotes the expectation with respect to Brownian bridge between X_0 and X_t. Let $X_0 = 0$ and $x \in \mathbb{R}$, the transition density of a killed Brownian motion is

$$\frac{P(X_t \in dx \,|\, \zeta > t)}{dx} \propto \frac{P(X_t \in dx)}{dx} P(\zeta > t \,|\, X_t \in dx) \tag{6.39}$$

$$\propto \exp\left(-\frac{x^2}{2t}\right) \mathbb{E}_{0,x}\left[\exp\left\{-\int_0^t \phi(X_s)ds\right\}\right]. \tag{6.40}$$

References

Andrieu, Christophe, De Freitas, Nando, Doucet, Arnaud, & Jordan, Michael I. 2003. An introduction to MCMC for machine learning. *Machine Learning*, **50**(1-2), 5–43.

Barber, David, Cemgil, A Taylan, Chiappa, Silvia, & Papaspiliopoulos, Omiros. 2011. *Bayesian Time Series Models*. 1 edn. Cambridge: Cambridge University Press.

Bertoin, Jean, Pitman, Jim, & De Chavez, Juan Ruiz. 1999. Constructions of a Brownian path with a given minimum. *Electronic Communications in Probability*, **4**, 31–37.

Beskos, A., Papaspiliopoulos, O., & Roberts, G. O. 2006. Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, **12**(6), 1077–1098.

Beskos, Alexandros, & Roberts, Gareth O. 2005. Exact simulation of diffusions. *Annals of Applied Probability*, **15**(4), 2422–2444.

Beskos, Alexandros, Papaspiliopoulos, Omiros, & Roberts, Gareth O. 2008. A factorisation of diffusion measure and finite sample path constructions. *Methodology and Computing in Applied Probability*, **10**(1), 85–104.

Black, Fischer, & Scholes, Myron. 1973. The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy*, **81**(3), 637–654.

Blanchet, Jose, Glynn, Peter, & Zheng, Shuheng. 2012. Empirical Analysis of a Stochastic Approximation Approach for Computing Quasi-stationary Distributions. *Evolve - A bridge between Probability*, 19–37.

Burq, Zaeem A., & Jones, Owen D. 2008. Simulation of Brownian motion at first-passage times. *Mathematics and Computers in Simulation*, **77**(1), 64–71.

Byron, Author, Hall, Martina, Brown, Eric, Hermanson, Richard, Charpentier, Emmanuel, Singmann, Henrik, & Henrik, Maintainer. 2016. Package ' LaplacesDemon '. *CRAN R Package*.

Chen, Nan, & Huang, Zhengyu. 2013. Localization and Exact Simulation of Brownian Motion-Driven Stochastic Differential Equations. *Mathematics of Operations Research*, **38**(3), 591–616.

Dacunha-Castelle, D, & Florens-Zmirou, D. 1986. Estimation of the Coefficients of a Diffusion from Discrete Observations. *Stochastics*, **19**(4), 263–284.

Darroch, J. N., & Seneta, E. 1965. On Quasi-Stationary Distribution in Absorbing Discrete-Time Finite Markov Chains. *Journal of Applied Probability*, **2**(1), 88–100.

Darroch, J. N., & Seneta, E. 1967. On Quasi-Stationary Distribution in Absorbing Continuous-Time Finite Markov Chains. *Journal of Applied Probability*, **4**(1), 192–196.

de Oliveira, M.M., & Dickamn, R. 2005. How to simulate the quasi-stationary state. *Phys. Rev. E*, **71**(1), 9.

Han-Fu, Chen. 2002. *Stochastic Approximation and Its Application*. 1 edn. Beijing, China: Kluwer Academic Publishers.

Karatzas, Ioannis, & Shreve, Steven E. 1991. *Brownian Motion and Stochastic Calculus*. 2 edn. Vol. 113. New York: Springer.

Kloeden, P. E., & Platen, E. 1999. *Numerical solution of stochastic differential equa- tions*. 3 edn. Berlin: Springer-Verlag.

Kushner, H. J., & Clark, D. S. 1978. *Stochastic Approximation for Constrained and Unconstrained System*. Berlin and New York: Springer-Verlag New York.

Kushner, H. J., & Yin, G. G. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*. 2 edn. New York: Springer-Verlag New York.

Neill, Philip D O. 2007. Constructing Population Processes With Specified Quasi-Stationary Distribution. *Stochastic Models*, **23**(3), 9.

Øksendal, Bernt. 2003. *Stochastic Differential Equations*. 6 edn. Berlin: Springer-Verlag Berlin Heidelberg.

Pollock, Murray. 2013. Some Monte Carlo Methods for Jump Diffusions. *PhD Thesis, University of Warwick*.

Pollock, Murray, Fearnhead, Paul, Johansen, Adam M, & Roberts, Gareth O. 2016. The Scalable Langevin Exact Algorithm : Bayesian Inference for Big Data. *Upcoming paper*, 1–45.

Pötzelberger, Klaus, & Wang, Liqun. 2001. Boundary Crossing Probability for Brownian Motion. *Journal of Applied Probability*, **38**(1), 152–164.

Robbins, Herbert, & Monro, Sutton. 1951. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, **22**(3), 400—-407.

Roberts, Gareth O, & Rosenthal, Jeffrey S. 2004. General state space Markov chains and MCMC algorithms. *Probability Surveys*, **1**, 20–71.

Roberts, Gareth O, & Tweedie, Richard L. 1996. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, **2**(4), pp. 341–363.

Ross, Sheldon. 2010. *Introduction to Probability Models*. 10 edn. Orlando, FL, USA: Academic Press, Inc.

Scott, Steven L, Blocker, Alexander W, Bonassi, Fernando V, Chipman, Hugh A, George, Edward I, & Mcculloch, Robert E. 2013. Bayes and Big Data : The Consensus Monte Carlo Algorithm. 1–22.

Straatsma, T.P., Berendsen, H.J.C., & a.J. Stam. 1986. Estimation of statistical errors in molecular simulation calculations. *Molecular Physics*, **57**(1), 89–95.

Teh, Yee Whye, Hasenclever, Leonard, Lienart, Thibaut, Vollmer, Sebastian, Webb, Stefan, Lakshminarayanan, Balaji, & Blundell, Charles. 2015. Distributed Bayesian Learning with Stochastic Natural-gradient Expectation Propagation and the Posterior Server. 1–30.

Zheng, Shuheng. 2014. Stochastic Approximation Algorithms in the Estimation of Quasi-Stationary Distribution of Finite and General State Space Markov Chains. *PhD Thesis, Columbia University*.