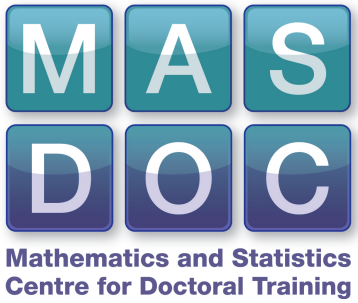


On Zig-Zag Extensions and Related Ergodicity Properties

by

Georgios Vasdekis

Thesis

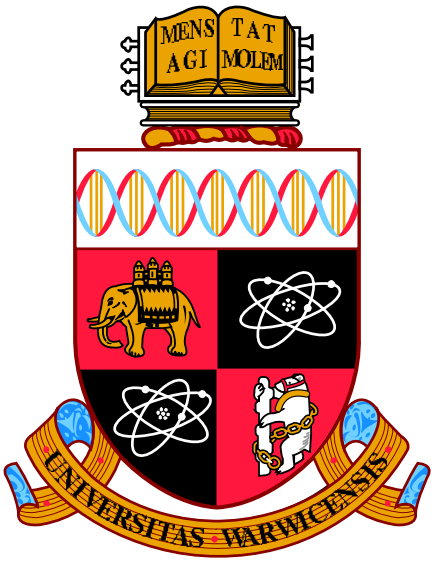
Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Mathematics Institute, Department of Statistics
The University of Warwick

February 2021



Contents

List of Tables

List of Figures

Acknowledgments

First of all, I would like to thank my supervisor, Gareth Roberts for his guidance, patience and utmost understanding throughout the PhD period. Our discussions have been inspirational and forged my enthusiasm towards the area.

Secondly, I would like to thank my family and especially my parents Vasilis and Gianna, along with my sister Konstantina. Without their constant encouragement and support I wouldn't had completed this work.

There have been many people in Warwick close to me these last years. I would like to thank all of my house mates and friends, but especially Quirin and Shannon, who made my time in Warwick wonderful.

I would, also, like to thank the guys from the office and the rest of the MASDOC DTC. I am grateful for meeting all of them and being part of this program.

My time in Warwick wouldn't had been the same without Leo and his amazing travel plans.

I am grateful to Lionel and Melanie for their company and for hosting me for the last few days before thesis submission.

I am grateful to all my friends in Greece, especially Kostas, Alkis, Xenofon and Zannis.

I would like to extend my thanks to Dr. Jon Warren and Dr. Krys Latuszynski for sitting through my Personal Advisory Committee, their helpful comments on how to improve the thesis and for their kind recommendation letters.

I would like to thank Professor Apostolos Giannopoulos, Professor Dimitris Cheliotis, Professor Leoni Dalla, Dr Perla Sousi and Professor Nikos Zygouras for all their help throughout my studies.

I would like to thank Professor Anthony Lee for pointing out a simpler way to prove Theorem 4.1.1 in this work.

I would like to thank my PhD examiners, Dr. Krzysztof Latuszynski and Professor George Deligiannidis for reading my work and for their comments that greatly helped the improvement of this work.

This work wouldn't had been completed without the funding from EPSRC and the Department of Statistics, grant reference number EP/N509796/1. I would, also like to thank both the Departments of Mathematics and Statistics for giving me the opportunity to pursue a PhD in Warwick.

Declarations

I hereby declare that this thesis is a product of my own work and the ideas discussed with my supervisor, Professor Gareth Roberts, unless specifically referenced or cited otherwise and it is in accordance with the university's guidelines on plagiarism. I also declare that part of Section 3.1 of this thesis contains material presented in the author's MASDOC M.Sc dissertation at the university of Warwick. This is clearly noted in the text. I also declare that this thesis has not been submitted for any other degree or qualification at this or any other university.

Abstract

In this thesis, we study the Zig-Zag process, which was recently proposed as an MCMC algorithm. We propose two extensions for this process and we prove geometric ergodicity results for both of them. The first extension, allows the process to move in more directions than just parallel to $\{-1, +1\}^d$, which can help explore the target distribution more efficiently. The second extension is motivated by a convergence problem we identify for the Zig-Zag process in heavy tailed targets. More precisely, we prove that the Zig-Zag process fails to converge geometrically fast towards the target distribution and we identify the rates of convergence for some specific targets. As a solution to this problem, the second extension allows the process to move with non constant speed, depending on its current location. This allows us to recover geometric ergodicity even for heavy tailed targets.

Abbreviations

1. MCMC: Markov Chain Monte Carlo
2. PDMP: Piecewise Deterministic Markov Process
3. ZZ: Zig-Zag
4. BPS: Bouncy Particle Sampler
5. MDZZ: Multi-Directional Zig-Zag
6. MDCNZZ: Multi-Directional Closest Neighbour Zig-Zag
7. SUZZ: Speed Up Zig-Zag

Chapter 1

Introduction

In Bayesian parametric statistics the parameter space which governs the behaviour of the experiment is assigned a prior distribution which is updated to a posterior distribution after the experiment has been conducted. One can use this posterior distribution as the current belief of the true value of the parameter. The focus, then, moves to calculate quantities of interest in the form $\mathbb{E}_\pi[f(X)]$, where $X \sim \pi$, π is the posterior distribution on the parameter space (which for the purpose of this work will be assumed to be a subset of $\mathbb{R}^d$) and $f : \mathbb{R}^d \to \mathbb{R}$ an observable of the parameter space. One of the main problems in the area is that the posterior distribution π can be very complex, making the integration with respect to π impossible to do by hand, especially when the dimension of the parameter space d is large, as is typically the case in applications. One of the main tools to overcome this problem is Markov Chain Monte Carlo (MCMC) [GRS96]. The main idea of this technique is to create a Markov chain that has the posterior distribution π as unique invariant measure and the law of the chain converges to π as time goes to infinity. Then, if we run a simulation with the chain for long time, eventually we will simulate approximately from the law of the posterior. More precisely, if $(X_n)_{n\geq 0}$ is a realisation of the Markov Chain, we create the ergodic sums

$$\hat{f}_N = \frac{1}{N}\sum_{i=1}^{N} f(X_i) \tag{1.1}$$

and under some standard assumptions a Law of Large Numbers (see for example [MT09] Theorem 17.1.7) ensures that as $N \to \infty$ these sums converge to $\mathbb{E}_\pi[f(X)]$. However, implementing such a Markov Chain until time N, for a large N, can be very costly, which forces us to use Markov chains converging to the invariant measure fast. At the same time, we would like the estimator $\hat{f}_N$ to have a small variance so

that we can convince ourselves that the estimation provided is close to the truth.

The most well known class of MCMC algorithms is the Metropolis-Hastings [MRR$^+$53,Has70]. According to this class of algorithms, when the process is at some point x, it chooses a new state y according to a proposal distribution with density $q(x,y)$ and it jumps to that state with probability $a(x,y) = \min\left\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\}$ or stays at the same position x with probability $1 - a(x,y)$. The main cost of running this algorithm comes from having to evaluate the quantity $a(x,y)$, which in turn depends on evaluating the quantities $\pi(x)$ and $\pi(y)$ in each step. In practice, these two quantities are quite costly to evaluate as they depend on all the observations of the experiment that created the posterior distribution π. On the other hand, we know that the algorithm will converge to the posterior of interest for a large family of proposal distributions and until recently the research in the area of MCMC focused on choosing a proposal distribution that can estimate the posterior fast and without a lot of variance. Some well studied algorithms, induced by a different choice of proposal, include the Independent Sampler, the Random Walk Metropolis [MRR$^+$53], the Gibbs Sampler [GG84] and the MALA algorithm [Bes94, RT96].

A main feature of the Metropolis-Hastings algorithm and possibly the main reason it is so widely studied and used, is the fact that it is always time-reversible for any choice of proposal. Roughly speaking, this means that if we start the chain from the invariant measure, the law of the process ran forwards in time is the same as the law of the process ran backwards in time. The main advantage of a reversible chain is that its invariant measure needs to satisfy the detailed balance equations. These are equations that involve point-wise evaluation of probability densities and are easier to verify than the more general global conditions, which involve integration. Therefore, it is easier to verify that a specific probability density of interest is the invariant measure of a time-reversible Markov chain. At the same time, a more operator theoretic approach identifies the time-reversible Markov chains with those chains whose transition operator is self adjoint in $L^2(\pi)$, which allows one to use the spectral decomposition theorem to study the convergence properties of these chains (see for example [LPW09]), making time reversible chains even more attractive. However, it was proven (see [CLP99, DHN00] that in some cases, the addition of a drift in the chain, makes it converge faster to equilibrium. More specifically, they introduced the concept of "Lifting" a Markov chain. This technique introduces a copy of the state of the Markov chain and adds a drift in each of the copies, allowing the Markov chain to jump from one copy to another at the same time. In this way, the time-reversibility of the chain was destroyed, but interestingly, in many cases, the rate of convergence of the chain to equilibrium increased. It is now better understood (see for example

[HHMS93, Gus98, CLP99, DHN00, SSG10, CH13, LNP13, RBS15, Bie14, DLP15]) that the addition of a momentum guiding the Markov chain could make the process non-reversible, but at the same time can help the process not to get stuck in some area of the space and can improve its convergence properties.

In this direction Hamiltonian Monte Carlo (HMC) [DKPR87, Nea11] was proposed as an algorithm that introduces a notion of momentum, which can help the algorithm to avoid getting stuck in some area. The idea comes from Physics and more specifically the Hamiltonian Dynamics, where the particle has a position, a velocity and total energy which is the sum of the potential and the kinetic energies. The former depends on the position of the particle and the latter on the velocity. The particle moves in such a way that the total energy remains the same. In an MCMC setting, the idea is to see the parameter space as the position space of the particle and then augment it by assigning a velocity space, which consists of the velocities the particle is allowed to take. In this situation, the role of potential energy is played by the target distribution, while the kinetic energy is free to be chosen to make the algorithm work better. Then the Markov chain follows the Hamiltonian dynamics, preserving the total energy of the system and it marginally targets the right distribution in the position space. However, in order to simulate this Markov chain one needs to find a closed form formula for the solution of the Hamiltonian dynamics, which is not always possible. Instead, one tends to discretise the Hamiltonian dynamics and approximate them using some numerical scheme, most notably the leapfrog integration. However, the resulting process no longer has the distribution of interest as invariant and typically one compensates for that by introducing a Metropolis-Hastings accept or reject step. This results in a Metropolis-Hastings algorithm with a new choice of a proposal measure, which is still time-reversible.

The last couple of years, there is a tendency in MCMC to use path dynamics that have a closed form formula as functions of time. This is achieved using a family of continuous time processes, called Piecewise Deterministic Markov processes (PDMPs). Roughly speaking, PDMPs move in deterministic dynamics, for example straight lines, for a random period of time. Then, they randomly switch to a different type of deterministic dynamics and they follow them for another random period of time, etc. These random periods occur as the first arrival times of Poisson processes and as will be seen in Chapter 2, these are objects one can typically simulate without any numerical error. At the same time, if one picks the rates of these Poisson processes appropriately, the resulting PDMP has the distribution of interest as invariant, which makes them useful tools to construct MCMC algorithms with.

The main two PDMPs used as building blocks in MCMC are the Bouncy Particle Sampler [BCVD18] and the Zig-Zag sampler [BFR19]. Incidentally, they were both seen as possible MCMC algorithms around the same time period. Except for both of them being non-reversible, a very interesting feature is that there are variants of these algorithms that can adapt well in a Bayesian statistics scenario where the target distribution is a posterior constructed by a large number of observations. More specifically, in order to implement these algorithms in practice, one needs to accept or reject some switches of deterministic dynamics with probabilities depending point-wise on the gradient of the minus log-likelihood of the target posterior. This quantity is computationally hard to evaluate as it depends on all the observations that induce the posterior. Instead of using the actual probability to accept or reject though, one can use an unbiased estimator of this probability, in a similar fashion to [AR09], or more recently [PFJR20]. The unbiased estimator does not need to depend on all the observations. This version of the algorithm is, therefore, computationally easier to implement, making the use of PDMPs a prominent tool in the era of big data.

The goal of this work is to motivate and introduce some extensions of the Zig-Zag sampler and study their convergence behaviour. As an example, the d-dimensional Zig-Zag process is only allowed to move in straight lines, parallel to a vector from the set $\{-1, +1\}^d$. We extend the algorithm so that it can use more directions to move. At the same time, a very interesting problem, which is common between MCMC algorithms, is that the performance of the Zig-Zag sampler is not very efficient when targeting heavy tails distributions and we propose a variant of the algorithm that tackles this problem.

The rest of this work is structured as follows. In Chapter 2 we provide a background with the mathematical notions we will use in the rest of the thesis. We also introduce the PDMPs and do a literature review on work that has been presented in the area. In Chapter 3 we provide an extension of the Zig-Zag process, allowing it to move in more directions. We prove that the process is Ergodic and Geometrically ergodic under assumptions very similar to the ones of [BRZ19]. In Chapter 4 we provide a proof that the Zig-Zag process (and any other unit speed PDMP algorithm) is not Geometrically ergodic when targeting heavy tailed distributions, which is a common problem between MCMC algorithms. We also extend this proof to a variant of the one-dimensional Zig-Zag, where the speed is allowed to be chosen from some probability distribution in $\mathbb{R}$, instead of just $\{\pm 1\}$. We focus on the convergence behaviour of one-dimensional Zig-Zag process on heavy tailed distributions and we prove exact rates of polynomial convergence when the

target distribution has tails similar to that of a student distribution. Furthermore, on Chapter 5 we propose a solution to the problem of slow convergence of the Zig-Zag process in heavy tails. We propose a variant of the algorithm, called Speed Up Zig-Zag (SUZZ), which allows the process, instead of moving with unit speed, to move with speed $s(x)$, depending on its current position $x \in \mathbb{R}^d$. We establish conditions on the speed function s for the process to be non-explosive, to have the right invariant measure and to be Geometrically ergodic. In both Chapters 3 and 5 we accompany our theoretical results with simulations. Finally, on Chapter 6 we discuss some further directions that this work could take.

Chapter 2

Preliminaries

In this chapter we will state some basic definitions we will use throughout the thesis and present some literature review on the use of Piecewise Deterministic Markov Processes (PDMP) in the field of MCMC. First, we recall some facts on Markov processes, their convergence to stationary (invariant) measure and their generators. Then, we briefly introduce the PDMP, a family of processes that evolve deterministically in space for some random period of time at the end of which they jump randomly to a new position random and start over. The two processes in this family that are mainly used in MCMC are the Zig-Zag (ZZ) sampler and the Bouncy Particle Sampler (BPS), which are introduced in the next two sections. Finally some complementary algorithms to ZZ or BPS are introduced.

2.1 Some results on Markov Processes

2.1.1 Discrete Time Markov Chains

In this section we will recall some basic facts about Markov Processes that we will encounter throughout the thesis.

Definition 2.1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space. A stochastic process of discrete time $(X_n)_{n\geq 0}$ is a (time homogeneous) Markov Chain (MC) if for any $n, m \in \mathbb{N}$ and $A \in \mathcal{F}$*

$$\mathbb{P}(X_{m+n} \in A | X_m = x, X_i = x_i, i < m) = \mathbb{P}(X_{m+n} \in A | X_m = x) = P^n(x, A)$$

where $\{P^n, n \in \mathbb{N}\}$ are the transition probabilities of the Chain after n steps.

We call the set E where X_n takes values the state space and we can think of the family of $\{P^n, n \geq 0\}$ as a family of operators, acting on bounded measurable

functions $f : E \to \mathbb{R}$ by setting $P^n f(x) = \mathbb{E}_x[f(X_n)]$ for all $x \in E$, where we write $\mathbb{E}_x$ to denote the expectation when the process starts from $x \in E$. We can, also think that P^n acts on measures on E by setting, for a measure π, $\pi P^n(A) = \int_E P^n(x, A) d\pi(x)$.

An important role in the theory of Markov Chains is played by invariant measures, i.e. by measures μ that satisfy the property $\pi P^n = \pi$ for all n. This is due to the fact that the law of the process can only get asymptotically close to an invariant measure. In order to measure the distance of the law of the process from an invariant measure we introduce the total variation distance $\|\cdot\|_{TV}$ between two probability measures μ, ν as

$$\|\mu - \nu\|_{TV} = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)| = \sup_{|f| \leq 1} |\mu(f) - \nu(f)| \tag{2.1}$$

where the supremum in the second equation is taken over all the functions $f : E \to \mathbb{R}$ such that $|f(x)| \leq 1$ for all $x \in E$ and we write $\mu(f)$ to denote the expectation of f under the measure μ. Similarly, for a positive function $V : E \to \mathbb{R}_+$, we define the V-norm as

$$\|\mu - \nu\|_V = \sup_{|f| \leq V} |\mu(f) - \nu(f)| . \tag{2.2}$$

Definition 2.1.2. *A Markov Chain $(X_n)_{n \geq 0}$ on state space E, with invariant measure π is called Ergodic if for all starting points $x \in E$*

$$||P^n(x, \cdot) - \pi(\cdot)||_{TV} \xrightarrow{n \to \infty} 0. \tag{2.3}$$

Definition 2.1.3. *A Markov Chain $(X_n)_{n \geq 0}$ on state space E, with invariant measure π is Geometrically Ergodic if there exists a function $V : E \to [1, +\infty)$ and constants $R > 0$, $\rho < 1$ such that for all $x \in E$*

$$||P^n(x, \cdot) - \pi(\cdot)||_V \leq RV(x)\rho^n. \tag{2.4}$$

The main part of this work is the study of convergence properties a some PDMPs. For this reason, we now introduce the main tools use to prove ergodicity and geometric ergodicity.

A Markov Chain X is ϕ-irreducible if there exists a non-trivial measure ϕ such that for any point z and any set $Aı\mathcal{B}(E)$ such that $\phi(A) > 0$, there exists an n with $P^n(z, A) > 0$. We call ϕ irreducibility measure.

Given $m \in \mathbb{N}$ and a probability measure ν_m, a set $C \in \mathcal{F}$ is ν_m-small if there

exists $c_m > 0$ such that for all $x \in C$

$$\mathbb{P}_x\left(X_m \in \cdot\right) \geq c_m \; \nu_m(\cdot), \tag{2.5}$$

where we write $\mathbb{P}_x$ to denote the law of the process starting from $x \in E$. A set $C \in \mathcal{F}$ is ν-petite for some probability measure ν if there exists a distribution a on $\mathbb{N}$ and $c > 0$ so that for all $x \in C$

$$K_a(x, \cdot) = \int_0^\infty \mathbb{P}_x\left(X_n \in \cdot\right) a(dn) \geq c \; \nu(\cdot). \tag{2.6}$$

If the dependence on ν_m or ν is not of interest we just call the set small or petite respectively. Note that a small set is trivially petite by setting the distribution a to be the Dirac distribution with full mass on $m \in \mathbb{N}$, where m as in equation (2.5). The idea behind the definition of a small set is that we can write equation (2.5) as

$$\mathbb{P}_x\left(X_m \in \cdot\right) = c \; \nu(\cdot) + q(x, \cdot)$$

therefore if the process starts from a small set C, it has a positive probability $c = \nu(E)$, to forget the point it starts from and decide where to jump m steps in the future according to the law of $\frac{1}{c}\nu$. The property of aperiodicity is then defined using the notion of small sets. More precisely, if C is $\nu = \nu_M$ small such that $\nu(C) > 0$ we define $E_C = \{n \geq 1 : C \text{ is } \nu_n - \text{small for } \nu_n = \delta_n \nu \text{ for some } \delta_n > 0\}$. The process is called aperiodic if the greatest common divisor of E_C is one. We have the following.

Proposition 2.1.4 (Theorem 5.5.7 Meyn and Tweedie 2009)**.** *If the chain is ϕ-irreducible and aperiodic then every petite set is small.*

Furthermore, the process is called strongly aperiodic if there exists a ν_1-small set C (i.e. a set satisfying (2.5) with $m = 1$) with $\nu_1(C) > 0$. This is sometimes easier to verify than aperiodicity and one can see that it implies aperiodicity. This is a generalisation of the countable state space definition of strong aperiodicity which needs $P(x, \{x\}) > 0$ for some $x \in E$. The small set can play the role of a single state x as the process has a positive probability to move in a uniform way starting from C. This potential memory loss inside a petite set means that one should expect the process to converge quickly to the stationary measure if it visits small sets quite often. Maybe even more interestingly, the other direction is true as well and any geometrically convergent process should visit small sets quite often.

A petite set C is said to have the self-geometric recurrence property (SGR)

if there exists a $b > 1$ so that for the stopping time τ_C of the first return time to C

$$\sup_{x \in C} \mathbb{E}_x[b^{\tau_C}] < \infty. \tag{2.7}$$

Indeed, we have the next significant result. It is a combination of Theorems 15.0.1 and 16.0.1 in [MT09].

Theorem 2.1.5 (Meyn-Tweedie 2009)**.** *Assume that A Markov Chain is ϕ-irreducible and aperiodic.*

- *Assume that there exists a petite set C and constant $b < \infty$ and $c > 0$ and a function $V \geq 1$ finite everywhere such that for all $x \in E$*

$$P^1 V(x) \leq (1 - c)V(x) + b1_C(x) \tag{2.8}$$

  where we write 1_C for the indicator function of C, that is 1 inside C and 0 outside. Then the process is Geometrically ergodic for all starting points $x \in E$ and V is the function appearing on (2.4).

- *Assume that the Chain is Geometrically Ergodic, then there exists a petite set C that has the self geometric recurrence property for some $b > 1$.*

The first part of the theorem can be used to prove that a chain is Geometrically ergodic and the function satisfying (2.8) is called drift or Lyapunov function. The second part of the theorem can be used in order to prove that a process is not geometrically ergodic, by proving that no petite set can have the SGR property. This is used for example in the study of Random Walk Metropolis algorithm targeting heavy tailed distributions in [JH00].

If the chain is ϕ irreducible and for any $A \in \mathcal{F}$ with $\phi(A) > 0$ and $x \in A$ we have
$\mathbb{P}_x \left(\sum_{n=1}^{\infty} 1_{X_n \in A} = +\infty \right) = 1$ the chain is called Harris recurrent.

Finally we say that the chain is a $T - chain$ if there exists a sub-stochastic transition kernel T such that $T(x, E) > 0$ for all $x \in E$, $T(\cdot, A)$ is lower semi-continuous function on E for every $A \in \mathcal{F}$ and there exists a distribution a on $\mathbb{N}$ such that for all $x \in E$ and all $A \in \mathcal{F}$ we have

$$K_a(x, A) \geq T(x, A) \tag{2.9}$$

where K_a is defined in (2.6).

2.1.2 Continuous Time Markov Processes

Let $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, a stochastic process of continuous time $(X_t)_{t\geq 0}$ taking values on a state space E is a (time homogeneous) Markov process if for all t, s we have

$$\mathbb{P}(X_{t+s} \in A | X_s = x, X_u, u < s) = \mathbb{P}(X_{t+s} \in A | X_s = x) = P^t(x, A)$$

where $(P^s)_{s\geq 0}$ the transition probabilities after s steps. We see $(P^t)_{t\geq 0}$ as a semigroup on continuous functions by setting for any bounded measurable f and $x \in E$ $P^t f(x) = \mathbb{E}_x[f(X_t)]$. The definitions of the previous section are analogously transferred to continuous time so we only give some examples. For example a set C is petite if there exists a non-trivial probability measure ν, $c > 0$ and a distribution a on $\mathbb{R}_+$ such that for any $x \in C$ and $A \in \mathcal{F}$

$$K_a(x, A) = \int_0^{+\infty} \mathbb{P}_x \left(X_t \in A\right) da(t) \geq c\nu(A). \tag{2.10}$$

Furthermore, the process is called aperiodic if there exists a petite set C and a $T > 0$ such that for any $x \in C$ and $t \geq T$, $\mathbb{P}\left(X_t \in C\right) > 0$.
Additionally, a ϕ-irreducible process is Harris recurrent if for all $x \in E$ and A such that $\phi(A) > 0$ we have $\mathbb{P}_x \left(\int_0^{+\infty} 1_A(X_t) = +\infty\right) = 1$. It can be shown (see [MT93a]) that this implies existence and uniqueness up to constant multiples of an invariant measure. If this measure is finite the process is called positive Harris recurrent.
Also, the process is T-process if equation (2.10) holds for some a a distribution on $[0, +\infty)$.
We define the strong generator $\mathcal{L}$ as a linear operator with domain $\mathcal{D}(\mathcal{L})$ all the functions $f : E \to \mathbb{R}$ for which there exists a g such that

$$\lim_{s\to 0} \left\| \frac{P^t f - f}{s} - g \right\|_\infty = 0 \tag{2.11}$$

and then we define

$$\mathcal{L}f(x) = g(x). \tag{2.12}$$

We defined the limit in the uniform norm, but it is also common to define it in any L^p norm. Heuristically, the generator can be though as the derivative of a function f along the average path that the Markov process will take if it starts from $x \in E$. With this interpretation, the following theorem (see [Dyn65] or Section 7.4 of [Oks03]) plays the role of the fundamental theorem of calculus.

Theorem 2.1.6 (Dynkin 1965)**.** *Let $(X_t)_{t\geq 0}$ a Strong Markov process that is strongly measurable (see [Dyn65] page 98, 80 and 81) with $\mathcal{L}$ the strong generator, then*

$$M_t = f(X_t) - f(x) - \int_0^t \mathcal{L}f(X_s)ds$$

is a $\mathbb{P}_x$ Martingale for any starting position $x \in E$.

This motivates the following definition, which extends the strong generator. We say that$\mathcal{D}(\mathcal{L}) \subset \mathcal{B}(E)$ (where $\mathcal{B}(E)$ are the Borel functions in E) is the domain of the extended generator of the process is for all $f \in \mathcal{D}(\mathcal{L})$ there exists $g \in \mathcal{B}(E)$ such that $s \to g(X_s)$ is a.s. locally integrable and

$$M_t = f(X_t) - f(x) - \int_0^t g(X_s)ds \tag{2.13}$$

is a $\mathcal{P}_x$ Martingale for any starting position $x \in E$. We define the extended generator of the process to be all such pairs (f, g). Later in this work, if $f \in \mathcal{D}(\mathcal{L})$ we may write $\mathcal{L}f$ to denote a specific g such that (f, g) is in the extended generator.

Remark 2.1.7. *In [Dav84] in the definition of extended generator, it is only assumed that M_t in (2.13) is a Local Martingale. We will stick with the original definition of the extended generator, but in Chapter 5 we will deal with the case where M_t is only a Local Martingale.*

An important use of the generator is that it can help identify an invariant measure for the process. Again the idea is that since the generator is the derivative along the dynamics of the process we expect for a measure to be invariant to have an average zero derivative along the dynamics. The following proposition is Proposition 4.9.2 in [EK86]

Definition 2.1.8. *Let $\mathcal{A} : \mathcal{D}(\mathcal{A}) \to L$ be an operator on a Banach space L. A set $\mathcal{D}$ is a* **core** *for the operator $\mathcal{A}$ if for every $f \in \mathcal{D}(\mathcal{A})$, there exists a sequence $f_n \in \mathcal{D}$ with $f_n \xrightarrow{n\to\infty} f$ and $\mathcal{A}f_n \xrightarrow{n\to\infty} \mathcal{A}f$.*

Proposition 2.1.9 (Ethier and Kurtz 1986)**.** *Assume that $\mathcal{A}$ generates a strongly continuous contraction semi-group on L, a subset of the borel functions of E. Assume that L is separating the state space E and the martingale problem is well posed, so that there exists a stochastic process X_t and for any $f \in L$, $f(X_t) - f(x) - \int_0^t \mathcal{A}f(X_s)ds$ is a $\mathbb{P}_x$ martingale. Assume further that $\mathcal{D}$ is a core for $\mathcal{A}$. Let μ be a measure on E and such that for all $f \in \mathcal{D}$*

$$\int_E f(x)\mu(dx) = 0. \tag{2.14}$$

Then μ is invariant for X_t.

A second use of the generator is to formulate a drift condition, similar to (2.8), that is sometimes easier to verify and which establishes geometric ergodicity for the process. The following is Theorem 5.2 from [DMT95]

Theorem 2.1.10 (Down-Meyn-Tweedie 1995)**.** *Assume that a Markov process is ϕ-irreducible and aperiodic and let $\mathcal{L}$ be the extended generator. Assume that there exist $c > 0$ $b < \infty$, a petite set C and a function $V \geq 1$ such that the following drift condition holds*

$$\mathcal{L}V(x) \leq -cV(x) + b1_C(x). \tag{2.15}$$

in which case we call V a Lyapunov function. Then there exist a unique invariant measure π and the process is Geometrically ergodic, i.e. there exists $R > 0$ and $\rho < 1$ such that for any $x \in E$,

$$||P^t(x, \cdot) - \pi(\cdot)||_V \leq RV(x)\rho^t. \tag{2.16}$$

Finally, we note that except for proving fast convergence the drift condition (2.15) can be used to prove Central Limit Theorems (CLT) for the Markov process. This is Theorem 4.4 in [GM96].

Theorem 2.1.11 (Glynn-Meyn 1996)**.** *Assume that the drift condition (2.15) holds for some $V \geq 1$ and that a function $g : E \to \mathbb{R}$ satisfies $g^2(x) \leq V(x)$ for all $x \in E$. Then there exists a unique (up to an additive constant) ϕ such that $\mathcal{L}\phi = -g$. Furthermore, for any $n \in \mathbb{N}$ define the function $Z_n : [0, 1] \to \mathbb{R}$ by*

$$Z_n(t) = \frac{1}{\sqrt{n}} \int_0^{nt} g(X_s, \Theta_s) - \pi(g)ds.$$

where π is the unique invariant measure. Then there exists a constant $\gamma_g^2 = 2\int_E \phi(x)\left(g(x) - \pi(g)\right)dx \in [0, +\infty)$ and Z_n converges weakly in the Skorokhod topology (see [Bil99]) in the space of continuous function in $[0, 1]$, $\mathcal{D}[0, 1]$ to $\gamma_g^2 B$, where B is a standard Brownian motion on $[0, 1]$.

2.2 Poisson Point Processes

In this section we will summarise the main properties of Poisson Processes as they are the main tool we use to represent the randomness of the Piecewise Deterministic process.

Definition 2.2.1. *Let $m : [0, +\infty) \to [0, +\infty)$. A process $N(t), t \geq 0$ is called Poisson Process (PP) with intensity m if*

1. $N(0) = 0$
2. $N(t+s) - N(s) \sim Poi\left(\int_s^{t+s} m(u)du\right)$
3. *For all $r \leq s$, $N(t+s) - N(s)$ is independent of $N(r)$.*

One can think of the Poisson Process as a mechanism that randomly generates points on the real line $[0, +\infty)$ and $N(t)$ counts the number of points generated up to time t. The higher the intensity rate, the more points are generated. Now, let us consider T_1 to be the time at which the first point is generated by the Poisson Process with rate m. Then, by the definition of the process we have

$$\mathbb{P}(T_1 \geq t) = \mathbb{P}(N(t) = 0) = \exp\left\{-\int_0^t m(s)ds\right\}.$$

On the other hand, consider the increasing function $f(t) = \int_0^t m(s)ds$ and the inverse function f^{-1}. If $E \sim \exp(1)$ then

$$\mathbb{P}(f^{-1}(E) \geq t) = \mathbb{P}(E \geq \int_0^t m(s)ds) = \exp\{-\int_0^t m(s)ds\}.$$

Therefore, $f^{-1}(E)$ and T_1 have the same law and this gives the following.

Proposition 2.2.2 (Exponential Representation of Poisson Process)**.** *Suppose m a positive function and $E \sim \exp(1)$. Then $T = \inf\{t \geq 0 : \int_0^t m(s)ds = E\}$ has the law of the first arrival time of a Poisson Process T.*

A consequence of this proposition is that in order to simulate directly a Poisson process with rate λ one needs to be able to integrate m, construct the function $f(t) = \int_0^t m(s)ds$ and then invert it. When this is not possible, one can use a useful property of the Poisson Process called Poisson Thinning. This allows one to couple any two Poisson processes together and generate their points at the same time using the following algorithm. The reader is referred to [Jac06, Kin92, LS79] for more details.

Algorithm 2.2.3 (Poisson Thinning)**.** *Let $\{N(t), t \geq 0\}$ a PP with intensity $m(t)$ and $\{N'(t), t \geq 0\}$ a PP with intensity $M(t)$. Assume further that $m(t) \leq M(t)$ for all t. Use the following method to generate random points:*

1. *Simulate a PP with intensity M.*

2. *Assuming that the points generated in Step 1 were* $\{t_i, i = 1, 2, ...\}$*, for each* i *accept the point* t_i *with probability* $\dfrac{m(t_i)}{M(t_i)}$*, else reject it.*

Then the accepted points are generated according to the law of a PP with intensity m*.*

This thesis will study stochastic processes whose random behaviour is encaptured by Poisson processes, having rate functions from which one cannot sample directly in practice. If one can bound above the rate m, which is the rate of interest, from a bounding rate M, from which one can simulate from, then one can simulate a process with rate M and then accept or reject the generated points with probability equal to the ratio between m and M evaluated at these points. The accepted points are distributed according to a Poisson process with the rate of interest and this is an implementable algorithm as one only needs to evaluate the rate m point-wise at the points generated by M, instead of having to perform an integration and invert a complicated function. This naturally raises the question of what type of bounding rates M one can use to perform Poisson Thinning. There are two types of rate functions currently used in the context of PDMPs for MCMC, the one where M is either a constant or a linear (up to a constant) function. We now present a way to simulate these processes, found in [BFR19].

Example 2.2.4 (Constant Bounding Process)**.** *Consider the case* $m(t) = M$ *for all* $t \geq 0$*. This Poisson process can be simulated until a given time horizon* T *as follows. First pick the number of points generated until time* T *which is distributed according to a Poisson distribution* $Poi(MT)$*. Given a number of* N *points generated before time* T*, these are distributed uniformly in the interval* $[0, T]$*.*

Example 2.2.5 (Linear Bounding Process)**.** *Consider the case* $m(t) = a + bt$ *,* $b > 0$*. This Poisson process can be simulated until a given time horizon* T *as follows. Construct the function* $f(t) = \int_0^t m(s)ds = \int_0^t a + bsds = at + \frac{b}{2}t^2$*. Simulate an* $E_1, E_2, ... \sim \exp(1)$ *i.i.d. and for any* $n \in \mathbb{N}$ *solve the quadratic equation* $f(T_n) = E_1 + E_2 + ... + E_n$*.* T_n *is the* n*'th arrival time of the process, due to Proposition 2.2.2. The number of* n *such that* $T_n \leq T$ *is a.s. finite so in order to simulate the process until time* T *this procedure will terminate a.s.*

It is also worth noting that one can simulate from rates that are piecewise linear and piecewise constant on different intervals. This can be done by combining the techniques used in a constant rate on the regions where the rate is constant and the linear techniques on the linear regions. This is justified by the lack of memory property of the Poisson process. Note as well that one could simulate from quadratic

intensities of the form $m(t) = a + bt + cb^2$ using the same procedure as in the linear case. This is because the function $f(t) = \int_0^t \mu(s)ds$ will be a cubic polynomial and there are known formulas to solve cubic equations of the form $f(T) = E$.

2.3 Piecewise Deterministic Markov Processes

Piecewise deterministic Markov Processes were first introduced in [Dav84] (see also [Dav18]). Their main feature is that they move in a deterministic way according to some dynamics and after a random period of time they randomly switch the dynamics in which they move. They then start over moving deterministically according to the new dynamics. The set of different dynamics which the process may deterministically follow are parametrized by a parameter $v \in K$ for some parameter space K. More precisely, for any $v \in K$ the process evolves on the space $M_v \times \{v\}$ where $M_v \subset \mathbb{R}^{d(v)}$ for some $d : K \to \mathbb{N}$. Overall, the state space of the process is $E = \{(x, v), v \in K, x \in M_v\}$ and the state of the process at time t is denoted by $\zeta_t = (x_t, v_t)$ where x_t represents the position in the space and v_t the dynamics under which the process tends to move at the current state. Let $\mathcal{M}_v$ be the Borel subsets of M_v and let $\mathcal{E} = \sigma\left(\{\cup_{v\in K} A_v \times \{v\}, A_v \in \mathcal{M}_v\}\right)$. The measurable space is $(E, \mathcal{E})$. The law of the process is characterized by a triplet $((g_v)_{v\in K}, \lambda, Q)$. For every $v \in K, g_v : M_v \to M_v$ is a function that characterises the dynamics parametrised by v. More precisely, we assume that g_v is such that for every $x \in M_v$ there exists a unique curve $(\Phi_v(x, t))_{t\geq 0}$ such that

$$\begin{cases} \dfrac{d\Phi_v(x,t)}{dt} = g_v(\Phi_v(x,t)) \\ \Phi_v(x,0) = x \end{cases} \tag{2.17}$$

Davis also requires this curve to be defined on $\mathbb{R}^{d(v)}$ for all $t \geq 0$. We will not make this assumption on Chapter 5, but we stick with it for now. Let ∂M_v be the boundary of M_v and let $\partial^* M_v$ be the points of the boundary where integral curves of the form of (2.17) can exit, i.e.

$$\partial^* M_v = \{y \in \partial^* M_v : \Phi_v(x, t) = y \text{ for some } (t, x) \in \mathbb{R}_+ \times M_v\}$$

and let $\Gamma^* = \cup_{v\in K} (\partial^* M_v \times \{v\})$. Then for $z = (x, v) \in E$ we define the exit time

$$t^*(x, v) = \inf\{t \geq 0 : \Phi_v(x, t) \in \partial^* M_v\}$$

If the process is at (x, v) then it tends to follow the deterministic flow $\Phi_v(x, t)$ for $t \in [0, t^*(x, v))$. Furthermore, $\lambda : E \to \mathbb{R}_+$ is a function such that for all $(x, v) \in E$ the function $s \to \lambda(\Phi_v(x, s), v)$ is locally integrable . $Q : \Gamma^* \times \mathcal{E} \to [0, 1]$ is such that for every $A \in \mathcal{E}$ the function $Q(\cdot, A)$ is measurable and for every $z \in E \cup \Gamma^*, Q(z, \cdot)$ is a probability measure. The process is then defined inductively as follows. Suppose that the process starts from $z = (x, v)$, let T_1 be the first minimum between $t^*(x, v)$ and the first arrival time of a Poisson process with rate function $t \to \lambda(\Phi_v(x, t), v)$. T_1 is the first switching event and we define the process as $Z_t = \Phi_v(x, t)$ for all $t \in [0, T_1)$. Then we select a random variable (X, V) in E with law $Q((\Phi_v(x, T_1), v), \cdot)$ and we set $Z_{T_1} = (X, V)$. Given that we have defined the process until time T_n for some $n \in \mathbb{N}$, $T_{n+1} - T_n$ is the minimum between $t^*(Z_{T_n})$ and the first arrival time of a Poisson process with rate $t \to \lambda(\Phi_{T_n}(X_{T_n}, t), V_{T_n})$. The process is defined as $Z_t = (\Phi_{V_{T_n}}(X_{T_n}, t - T_n), V_{T_n})$ for $t \in [T_n, T_{n+1})$ and we sample (X, V) in E according to the law $Q(\Phi_{V_{T_n}}(X_{T_n}, T_{n+1} - T_n), v, \cdot)$ and set $Z_{T_{n+1}} = (X, V)$.

The authors make the following assumption as a way to ensure that the process will not explode by having an infinite amount of switches in finite time.

Assumption 2.3.1. *Let $N_t = \sup\{T_n : n \in \mathbb{N}, T_n \leq t\}$ be the number of switches until time t, then $\mathbb{E}[N_t] < \infty$ for any t.*

Given the filtration $(\mathcal{F}_t)_{t \geq 0}$ in which the process is adapted and a stopping time T we define the σ-algebra of T to be $\mathcal{F}_T = \{A \in \mathcal{F} : A \cap \{T = t\} \in \mathcal{F}_t \text{ for all } t\}$. We say that the process X has the **strong Markov property** if on the event $\{T < \infty\}$ we have

$$\mathbb{E}[f(X_t + T)|\mathcal{F}_T] = \mathbb{E}_{X_T}[f(X_t)] \tag{2.18}$$

for all bounded measurable functions f. Note that the random switching time of a PDMP is the first arrival time of a Poisson process, which is convenient as it allows the process to maintain the Markov property. Indeed the process even possesses the Strong Markov property. This is a very convenient property when we want to simulate a PDMP as we can condition on the values of the process during the switching event and then we can start the process afresh without any need to know the past path in order to simulate the future path. This is not trivial without the Strong Markov property as the switching events happen during random stopping times.

Proposition 2.3.2 (Davis 1984)**.** *Assume that assumption 2.3.1 holds. Then the PDMP satisfying assumption has the Strong Markov property.*

The most interesting result of [Dav84] is probably that the extended generator of a PDMP is completely characterised. Before we present the result we need some

definitions. First of all, out of the points in the boundary Γ^*, some will a.s. never be reached by the process. We define Γ to be the set of points $(y,v) \in \Gamma^*$ such that $\lim_{x\to y} \mathbb{P}_{x,v}(T_1 = t^*(x,v)) = 1$. We also let $\mathcal{F}_t = \sigma(Z_s, s \leq t)$ be the filtration of the PDMP.

Theorem 2.3.3 (Davis 1984)**.** *The domain $\mathcal{D}(\mathcal{A})$ of the extended generator of the PDMP, i.e. the set of measurable functions $f : E \cup \Gamma \to \mathbb{R}$ such that there exists a function $\mathcal{A}f$ so that $s \to \mathcal{A}f(z_s)$ is locally integrable with*

$$M_t = f(Z_t) - f(z) - \int_0^t \mathcal{A}f(Z_s)ds$$

being a $\mathbb{P}_z$ Local Martingale on the $(\mathcal{F}_t)_{t\geq 0}$ filtration for all $z \in E \cup \Gamma$, is exactly the functions that satisfy the following three properties.

1. *For every $(x,v) \in E$ the function $t \to f(\Phi_v(x,t),v)$, $t \in (0, t^*(x,v)]$ is absolutely continuous., i.e. it has a weak derivative.*

2. *For all $z \in \Gamma$ we have*
$$f(z) = \int_E f(z')Q(z,dz') \tag{2.19}$$

3. *If Ω is the probability space and $\mathcal{B}f : E \times \mathbb{R}_+ \times \Omega$ as $\mathcal{B}f(z,t,\omega) = f(z) - \lim_{s\to t^-} f(Z_s)$, then $\mathcal{B}f$ is jointly measurable and for any t*
$$\mathbb{E}\left[\sum_{T_n \leq t} f(Z_{T_n}) - f(Z_{T_n^-})\right] < \infty. \tag{2.20}$$

Furthermore, if $f \in \mathcal{D}(\mathcal{A})$ then

$$\mathcal{A}f(x,v) = \langle g_v(x), \nabla f(x,v)\rangle + \lambda(x,v)\int_E \big(f(z') - f(x,v)\big)\, Q((x,v),dz). \tag{2.21}$$

where we write
$$\nabla f(x,v) = \frac{d}{dt} f(\Phi_v(x,t),v)\,|_{t=0} \tag{2.22}$$

In the formula for the generator (2.21), $\langle\cdot,\cdot\rangle$ is the inner product on $\mathbb{R}^{d(v)}$ and one should interpret $g_v(x)$ and $\nabla f(x,v)$ for fixed x, v as vectors in $\mathbb{R}^{d(v)}$. If $f \in C^1$ with respect to x, as is usually the case, then $\nabla f(x,v) = (\partial_1 f(x,v), ..., \partial_{d(v)} f(x,v))$ and the first part of the sum of (2.21) is equal to $\sum_{k=1}^{d(v)} g_v^k(x)\partial_k f(x,v)$ where $g_v(x) = (g_v^1(x), ..., g_v^{d(v)}(x))$. Here we write ∂_i for the operator of the partial derivative in the i coordinate.

Corollary 2.3.4. *Assume that $\mathcal{D}(\mathcal{A})$ is the domain of the extended generator of a PDMP and $f \in \mathcal{D}(\mathcal{A})$. Then ∇f, defined in (2.22) is the weak derivative of f, under the flow Φ, i.e. for all $(x, v) \in E$ and $t \geq 0$*

$$f(\Phi_v(x,t),v) - f(x,v) = \int_0^t \nabla f(\Phi_v(x,s),v)ds.$$

Proof of Corollary 2.3.4. We know from Theorem 2.3.3 that for all (x, v), the function $t \to f(\Phi_v(x,t),v)$ has a weak derivative $h_{x,v} : [0,+\infty) \to \mathbb{R}$. This means that there exists an $h : E \times [0,+\infty) \to \mathbb{R}$ such that for all $(x, v) \in E$ and all $t \geq 0$

$$f(\Phi_v(x,t),v) - f(x,v) = \int_0^t h(x,v,s)ds. \tag{2.23}$$

We write for all $s, t \geq 0$

$$\begin{aligned}
&\int_0^{t+s} h(x,v,u)du = f(\Phi_v(x,t+s),v) - f(x,v) = \\
&= f(\Phi_v(x,t+s),v) - f(\Phi_v(x,t),v) + f(\Phi_v(x,t),v) - f(x,v) = \\
&= \int_0^s h(\Phi_v(x,t),v,u)du + \int_0^t h(x,v,u)du
\end{aligned}$$

therefore for all $(x, v) \in E$, $ts \geq 0$

$$\int_t^{t+s} h(x,v,u)du = \int_0^s h(\Phi_v(x,t),v,u)du$$

so

$$\int_0^s h(x,v,t+u)du = \int_0^s h(\Phi_v(x,t),v,u)du$$

So for all $t \geq 0$,

$$h(x,v,t+\cdot) = h(\Phi_v(x,t),v,\cdot)$$

in the L^1 sense and then

$$h(x,v,t) = h(\Phi_v(x,t),v,0) = \nabla f(\Phi_v(x,t),v)$$

which proves the result when combined with (2.23). □

2.4 The Zig-Zag Sampler

In this section we will introduce the Zig-Zag process, which is the main focus of this work. Trying to construct a non-reversible algorithm to sample from the Curie-Weiss

model (see for example [LLP07]) the authors of [BR17] created a Lifted version (see [TCV11, DHN00]) of a Metropolis algorithm on the space of magnetisation of the Curie-Weiss model. The one dimensional Zig-Zag algorithm appeared as a scaling limit of that Lifted Random Walk Metropolis-Hastings, although a simpler version of the process was introduced in [Gol51] as the telegraph process (see also [Kac74, FGM16, FGM12]). The process was later extended in higher dimensions in [BFR19] and has been proposed as a PDMP which can be used as an MCMC algorithm to target posterior distributions (see also [FBPR18, VBCDD17]).

2.4.1 Definition and Invariant Measure

The d dimensional Zig-Zag process $(Z_t)_{t\geq 0} = ((X_t, \Theta_t))_{t\geq 0}$ is a Piecewise Deterministic Markov Process with state space $E = \mathbb{R}^d \times \{\pm 1\}^d$. One can think of the process as a particle moving in $\mathbb{R}^d$ in straight lines. When the process is at point $(x, \theta) \in E$ the particle is at point $x \in \mathbb{R}^d$ and moves towards the direction $\theta \in \{\pm 1\}^d$, which we will also call the velocity. After some random time, which depends on the path that the particle follows it stops and changes the direction it moves towards. More formally, writing $\Phi_{(x,\theta)}(t) = \Phi_\theta(x, t)$, for the deterministic flow starting of the process starting from (x, θ) the dynamics are governed by

$$\begin{cases} \dfrac{dX_t}{dt} = \dfrac{d\Phi_{(x,\theta)}(t)}{dt} = \theta, & X_0 = \Phi_{(x,\theta)}(0) = x \\ \dfrac{d\Theta}{dt} = 0, & \Theta_0 = \theta \end{cases} \tag{2.24}$$

In each of the d coordinates we assign a Poisson Process with rate $m_i(t) = \lambda(x + \theta t, \theta), i = 1, ..., d$, for some function $\lambda : E \to [0, +\infty)$ and we assume that m_i is locally integrable. For each of the processes we generate its first arrival time T_i, we pick the smallest of them $T = \min\{T_i, i = 1, ..., d\}$ and the process $j = argmin\{T_j, j = 1, ..., d\}$ from which T came from. The process follows the deterministic flow until time T and then it will switch the direction from $\theta = (\theta_1, ..., \theta_d)$ to $F_j[\theta] \in \{\pm 1\}^d$, where

$$\begin{cases} F_j[\theta]_j = -\theta_j \\ F_j[\theta]_i = \theta_i \text{ for } i \neq j. \end{cases}$$

Then the process is at point $(x', \theta') = (x + T\theta, F_j[\theta])$ and it starts moving again according to the deterministic flow of $\{\Phi_{(x',\theta')}(t), t \geq 0\}$. The process is described in the following algorithm.

Algorithm 2.4.1 (The Zig-Zag Algorithm)**.**

1. *Set* $t_{current} = 0$

2. *Start from point* $(X_{t_{current}}, \Theta_{t_{current}}) = (x, \theta)$.

3. *The* x*-component moves along the line* $\{x + t\theta, t \geq 0\}$.

4. *For every coordinate* $i \in \{1, .., d\}$ *construct a Poisson Process with intensity* $\{m_i(t) = \lambda_i(x + t\theta, \theta), t \geq 0\}$.

5. *Let* T^i *be the first arrival time of the* i*'th Poisson Process, i.e. for all* $t \geq 0$, $\mathbb{P}(T^i \geq t) = \exp\{-\int_0^t m_i(s)ds\}$. *Let* $j = argmin\{T^i, i = 1, ..., d\}$ *and* $T = T^j$ *the first arrival time of all the processes.*

6. *For* $t \in [t_{current}, t_{current} + T)$ *set* $X_t = x + (t - t_{current})\theta$ *and* $\Theta_t = \theta$.

7. *Set* $x = x + T\theta$, $\theta = F_j[\theta]$ *and* $t_{current} = t_{current} + T$.

8. *Repeat from the Step 2.*

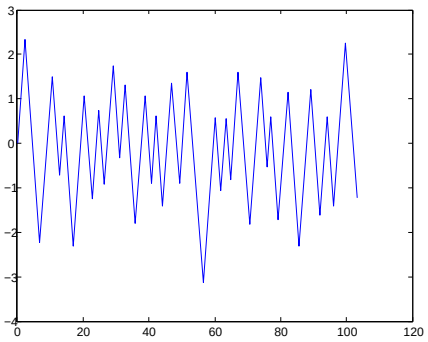
We can view this process in the context of [Dav84] by writing $E = \cup_{\theta \in \{\pm 1\}^d}\{\mathbb{R}^d \times \{\theta\}\}$ and have a deterministic flow given by (3.3). The rate function is $\lambda(x, \theta) = \sum_{i=1}^d \lambda_i(x, \theta)$ and the jump measure is $Q(x, \theta, \cdot) = \sum_{i=1}^d \frac{\lambda_i(x,\theta)}{\lambda(x,\theta)} \delta_{(x,F_i[\theta])}(\cdot)$.
In Figure 2.1 we present the trace plots of three one dimensional Zig-Zag process. The first targets a $\mathcal{N}(0, 1)$, the second an exponential distribution with parameter 1, symmetrically extended to the negative real line and the third a Cauchy distribution. In all the plots the process starts from 0 facing the upwards direction therefore the process starts from $(0, +1)$. Likewise in figure 2.2 we present the trace plots of two two dimensional Zig-Zag processes. The first one targets a positive correlated Gaussian and the second an uncorrelated Cauchy. Since the particle moves in straight lines between switching direction events, all the randomness of the process is hidden in the rates λ. By tuning these rates appropriately one can have the measure of interest invariant for the process as shown in the following proposition.

Proposition 2.4.2 (Bierkens-Fearnhead-Roberts 2019)**.** *Suppose that there exists a function* $U \in C^1$ *with* $\int_{\mathbb{R}^d} \exp\{-U(x)\}dx < \infty$ *such that the rates of the Zig-Zag process satisfy*

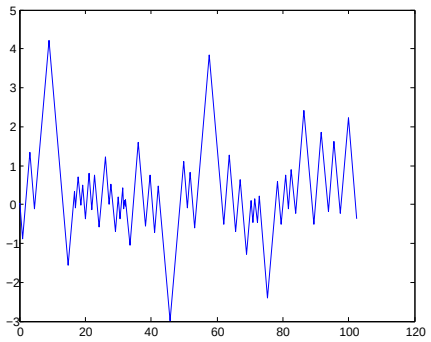
$$\lambda_i(x, \theta) = [\theta_i \partial_i U(x)]^+ + \gamma_i(x, \theta_{-i}) \tag{2.25}$$

where $a^+ = \max\{a, 0\}$ *and* γ_i *is a non-negative function that does not depend on the* i *component of* θ. *Then the process admits the measure*

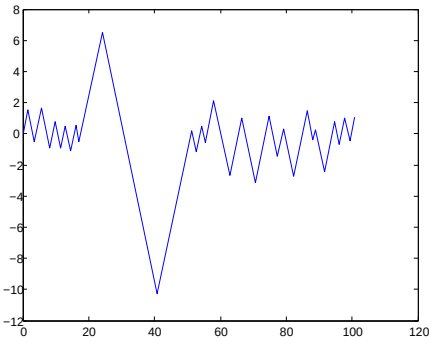
$$\mu(dx, d\theta) = \frac{1}{2^d Z} \exp\{-U(x)\}dxd\theta \tag{2.26}$$



(a) Normal distribution



(b) Exponential distribution

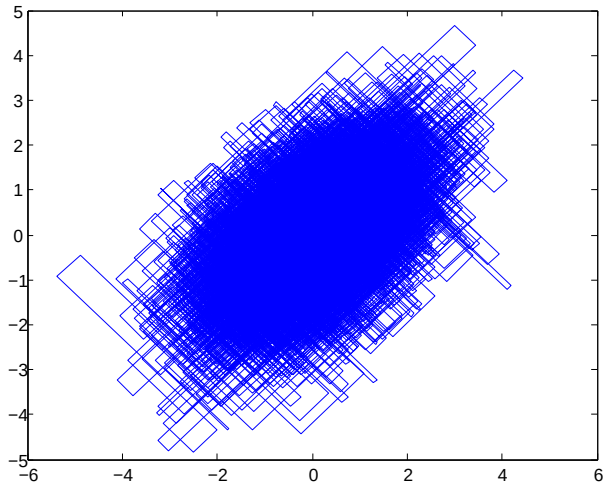


(c) Cauchy distribution

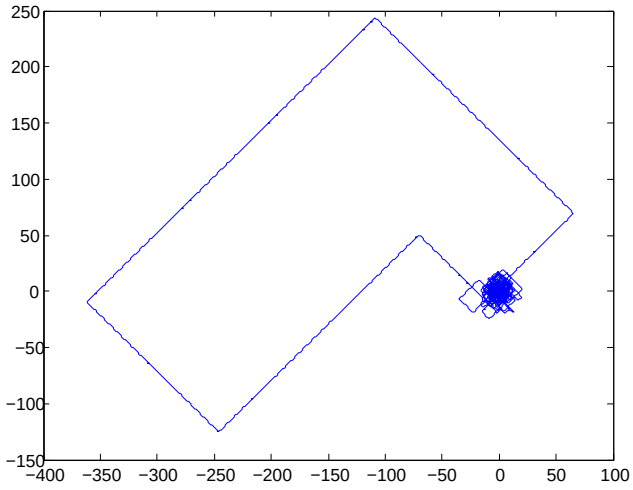
Figure 2.1: *Trace plot Zig-Zag for three different one dimensional distributions.*

on E as invariant.

The quantity $\theta_i \partial_i U(x)$ tends to become large when the process tends to move towards regions of lower density along the ith coordinate. It signifies the tendency of the process to switch the direction it moves in the ith coordinate when moving towards areas of lower density. The function γ_i describes an extra refreshment rate that we can add in the algorithm, which tends to flip the ith component of the velocity, independently of where the particle tends to move towards. We call this function γ_i the refresh rate. Because this rate tends to switch the ith coordinate from $+1$ to -1 as much as from -1 to $+1$ it does not change the invariant distribution of the algorithm, but it tends to make the algorithm more reversible with respect to time. [AL19] extends results of Peskun ordering (see for example [Pes73, Tie98, Mir01]) to some non-reversible processes, with the Zig-Zag being one of them. Their results hint that in one dimension the choice of refresh



(a) Positive Correlated Normal distribution



(b) Uncorrelated Cauchy distribution

Figure 2.2: *Plot of Zig-Zag for two different two dimensional distributions run until time* $T = 10^4$

$\gamma(x) \equiv 0$ minimizes the asymptotic variance of the sampler and is in this sense the best. This is in accordance with the folklore that the non-reversible algorithms can outperform reversible ones. In general though, being flexible enough to pick any γ_i non-negative is the key in order to implement Zig-Zag with sub-sampling [BFR19], which seems to have applications when targeting a posterior that is created from a large number of observed data.

2.4.2 Sub-sampling

As with most of the state of the art MCMC procedures, the main cost of implementing the algorithm is the evaluation of the likelihood of the posterior distribution. For example, in Metropolis-Hastings, in each iteration one needs to accept or reject the proposed move and the probability of acceptance depends on the likelihood (or the gradient of the likelihood) of the target. In the case of Zig-Zag these likelihood evaluations appear in a more subtle way. In order to implement an one dimensional Zig-Zag one needs to simulate the times where the direction switches. Since these times are distributed according to the first arrival time of a Poisson process with rate described by $\lambda(x, \theta) = [\theta U'(x)]^+$, if one cannot integrate and then invert these rates directly, one needs to use Poisson thinning, as in Algorithm 2.2.3, in order to simulate these arrival times. This would include bounding the rates from a rate $\Lambda(x)$ from which we can simulate from and then accept any point x_0 generated from the Λ-process according to a probability $\frac{\lambda(x_0,\theta)}{\Lambda(x_0)} = \frac{[\theta U'(x_0)]^+}{\Lambda(x_0)}$, or otherwise reject that point. Since this probability of accepting the point involves the quantity U', this procedure of deciding whether to accept that point or no involves evaluating the gradient of the log-likelihood. In a Bayesian setting the posterior density is typically of the form

$$\pi(x) = \frac{1}{Z} \prod_{k=1}^{n} f(y_k|x) pr(x)$$

where $f(\cdot|x)$ is the likelihood of the experiment under x, pr the prior distribution, $y_1, ..., y_n$ the observations and Z a normalising constant. This means that if $\pi(x) = \frac{1}{Z} \exp\{-U(x)\}$ then

$$\partial_i U(x) = \frac{1}{n} \sum_{k=1}^{n} E_k^i(x) \tag{2.27}$$

where

$$E_k^i(x) = \partial_i U_k(x) = \frac{\partial}{\partial x_i} \left[-\log(pr(x)) - n \log(f(y_k|x)) \right] \tag{2.28}$$

One can notice that evaluating $\partial_i U$ can be very costly when the number of observations n is large. On the other hand, each the E_k^i needs a much cheaper evaluation

since it depends only on one observation. In [BFR19] the authors use (2.27) to introduce a variant of the Zig-Zag algorithm that can lead to cheaper implementations. Instead of using $\partial_i U$ to construct the probability of accepting or rejecting a switch from a bounding process, the idea is to pick an index $k \in \{1, ..., n\}$ at uniformly at random and use E_k^i. Since $\partial_i U = \frac{1}{n}\sum_{k=1}^n E_k^i$, the quantity E_K^i where $K \sim unif\{1, ..., d\}$ is and unbiased estimator of $\partial_i U$ and evaluating E_K^i is much cheaper than $\partial_i U$. The following algorithm is called Zig-Zag with sub-sampling, due to the fact that one uses only part of the observations each time one needs to calculate a value related to the gradient of the log-likelihood.

Algorithm 2.4.3 (Zig-Zag with Sub-sampling)**.**

1. *Set $t_{current} = 0$, specify a value $T_{run} > 0$.*
2. *Start from point $(X_{t_{current}}, \Theta_{t_{current}}) = (x, \theta)$.*
3. *The x-component moves along the line $\{x + t\theta, t \geq 0\}$.*
4. *For each $i \in \{1, ..., d\}$, find a rate function $\{M^i(t), t \geq 0\}$ such that $[\theta E_k^i(x + \theta t)]^+ \leq M^i(t)$ for all $t \leq T_{run}$, for all $k \in \{1, ..., n\}$ and such that one can simulate the first arrival times of all the intensities M^i.*
5. *Let T^i be the first arrival time of the Poisson Process, with rate $\{M^i(t), t \geq 0\}$, i.e. for all $t \geq 0$, $\mathbb{P}(T \geq t) = \exp\{-\int_0^t M^i(s)ds\}$. Let $j = argmin\{T^i, i = 1, ..., d\}$ and $T = T^j$ the first arrival time of all the processes.*
6. *If $T > t_{current} + T_{run}$*
   (a) *For $t \in [t_{current}, t_{current} + T_{run})$ set $X_t = x + (t - t_{current})\theta$ and $\Theta_t = \theta$. Set $x = x + T_{run}\theta$, $t_{current} = t_{current} + T_{run}$.*

   Else

   (a) *For $t \in [t_{current}, t_{current} + T)$ set $X_t = x + (t - t_{current})\theta$ and $\Theta_t = \theta$. Set $x = x + T\theta$, $t_{current} = t_{current} + T$.*
   (b) *Pick $K \sim unif\{1, ..., n\}$.*
   (c) *With probability $\frac{[\theta E_K^j(x)]^+}{M^j(T)}$ set $\theta = F_j[\theta]$, else leave θ unchanged.*
7. *Repeat from the Step 2.*

Proposition 2.4.4 (Bierkens-Fearnhead-Roberts 2019)**.** *The process introduced in (2.4.3) has μ introduced in (2.26) invariant.*

The proof boils down to the fact that the algorithm described is a Zig-Zag process with switching rates for any $i \in \{1, ..., d\}$, $\lambda_i(x, \theta) = \frac{1}{n}\sum_{k=1}^{n}[\theta E_k^i(x)]^+ = [\theta_i \partial_i U(x)]^+ + \gamma(x, \theta)$ where

$$\gamma_i(x, \theta) = \sum_{k=1}^{n}[\theta_i E_k^i(x)]^+ - [\theta_i \frac{1}{n}\sum_{k=1}^{n} E_k^i(x)]^+ \geq 0$$

is non-negative due to the fact that the operation $^+$ is a convex function. Furthermore, one can check that $\gamma_i(x, F_i[\theta]) = \gamma_i(x, \theta)$. Then we can conclude using Proposition 2.4.2. It is important to note that the reason the process still targets the same distribution is that we have some flexibility in picking the switching rate, by appropriately tuning the functions γ_i. On the other hand, increasing γ_i seems to lead to weaker convergence behaviour (as established in [AL19]). At the same time, using sub-sampling means that one has to increase the bounding function M, as it needs to bound all the functions $[\theta E_k]^+$ for all k. This leads to more proposed switching times and more evaluations of the gradient log-likelihood needed.

We note here that on step 6*b* of the sub-sampling algorithm 2.4.3, we use the information provided by some observation $K \in \{1, ..., n\}$, where K is sampled uniformly from $\{1, ..., n\}$. However, as analysed in [SSLD20] one can put different weights on each observation and favour the choice of some specific observations over others. This could have applications in some settings and the authors argue that it can be useful in an example of Bayesian logistic regression with sparse predictors. Let us, also note that ideas on implementing sub-sampling for the Bouncy Particle Sampler (see Section 2.5) have also been presented in the literature (see for example [PGCP17])

A more efficient way of sub-sampling from Zig-Zag, was also proposed in [BFR19], under the crucial assumption that the gradient of all components of the log-likelihood are uniformly Lipschitz, meaning that there exists a C and $p \in [1, +\infty]$ such that for all $x_1, x_2 \in \mathbb{R}^d$, for all $k \in \{1, ..., n\}$ and all $i \in \{1, ..., d\}$,

$$|\partial_i U_k(x_1) - \partial_i U_k(x_2)| \leq C_i |x_1 - x_2|_p. \tag{2.29}$$

This property is satisfied, for example, when the posterior is formed by i.i.d. Gaussians. They call this variant Zig-Zag with control variates as it uses the technique of control variates to minimize the variance of the information given by the different components U_k of the posterior. The idea is then to choose and fix a reference point x^* at some distance $O(n^{-1/2})$ from the mode, which can happen using some numerical method with an $O(n)$ cost. Then, for all $i \in \{1, ..., d\}$, they calculate the

quantities $\partial_i U(x^*)$ and $\partial_i U_k(x^*)$ which also has an $O(n)$ cost. After these initial calculations they can implement the Zig-Zag with sub-sampling by choosing slightly different rates, using

$$E_k^i(x) = \partial_i U(x^*) + \partial_i U_k(x) - \partial_i U_k(x^*) \tag{2.30}$$

Choosing these rates allows them to use the Lipschitz condition and reduce the variance between different values of E_k^i over k. Indeed, using the Lipschitz property, one can bound for any $x \in \mathbb{R}$ and $\theta \in \{\pm 1\}$

$$|E_k^i(x + t\theta)| \leq |\partial_i U(x^*)| + C_i|x + t\theta - x^*|_p \leq |\partial_i U(x^*)| + C_i|x - x^*|_p + C_i\sqrt{d}t$$

Therefore one can simulate the first switching time of the i coordinate of the process starting from (x, θ) using Poisson thinning with linear bounding rate $M_i(t) = a_i + b_i t$ where $a_i = [\theta \partial_i U(x^*)]^+ + |x - x^*|_p$ and $b_i = C_i\sqrt{d}$. As shown in Section 2.2.5, one can simulated directly from these bounding rates. Since x^* is chosen to be close to the mode, $\partial_i U(x^*)$ is not very large (more specifically the authors argue that it is $O(n^{1/2})$) and as long as x is close to the mode, $|x - x^*|_p$ is small so the bounding process has a not very large rate. The authors effectively use that in the limit as $n \to \infty$, the posterior will look like a Gaussian (for example from Berstein Von Mises theorem [Cam86]) and the process will spend most of the time close to the mode, so most of the times the process switches direction, the new bounding process will not have a very large rate and simulating the next switching point will not be computationally heavy. Even if we assume that the Lipschitz constant C_i increases with n in an $O(n)$ rate, the authors argue that after choosing the point x^* and calculate the quantities $\partial_i U(x^*)$ and $\partial_i U_k(x^*)$ for all k and i, provided the process starts from the stationary distribution, it only needs $O(1)$ computational cost to generate an essentially independent from the starting position sample. They call this property super-efficiency and it was tested in simulations on models like logistic regression with positive results. This efficiency of the algorithm is one of the main motivations of this work that studies some properties and variants of the Zig-Zag process.

2.4.3 Convergence Theorems

Early results on the convergence of the one dimensional Zig-Zag sampler can be found in [BR17]. In [BRZ19] the authors prove the following convergence results for the Zig-Zag sampler in an arbitrary dimension d.

Theorem 2.4.5 (Bierkens-Roberts-Zitt 2019 Theorem 1)**.** *Assume that $U \in C^3(\mathbb{R}^d)$, U has a non-degenerate local minimum, in the sense that the Hessian matrix is strictly positively defined at the local minimum and assume that for some $c > d$ and $c' \in \mathbb{R}$, for all $x \in \mathbb{R}^d$*

$$U(x) \geq c \log(|x|) - c'. \tag{2.31}$$

Then, the Zig-Zag process is ergodic, in the sense that for any starting point $(x, \theta) \in E$

$$\|\mathbb{P}_{x,\theta}(Z_t \in \cdot) - \mu(\cdot)\|_{TV} \xrightarrow{t\to\infty} 0. \tag{2.32}$$

The authors prove this theorem by first proving ϕ-irreducibility and then using some standard tools from [MT93a]. In order to prove irreducibility, they introduce the notion of reachability and prove that the process has this property. They first prove that the reachability holds when the target distribution is a Gaussian and then they argue that around the non-degenerate local minimum the distribution looks like a Gaussian to prove the reachability property holds locally around the local minimum. They then extend the result for all points away from the minimum by analysing the structure of the sets of points that can reach each other. In terms of the assumptions made in the theorem, the C^3 property and the existence of a non-degenerate local minimum, are very often satisfied in an MCMC setting. The growth condition (2.31) is equivalent to ask that the density of the target measure satisfies

$$\frac{1}{2^d Z} \exp\{-U(x)\} \leq C|x|^{-c}$$

for some $c > d$. Since the function $|x|^{-c}$ is integrable over $\mathbb{R}$ if and only if $c > d$, this growth assumption is not unreasonable if we are targeting a probability measure, where the density should be integrable. An immediate consequence of Theorem 2.32 is a Law of Large Numbers.

Corollary 2.4.6 (Bierkens-Roberts-Zitt 2019)**.** *For all starting points $(x, \theta) \in E$ and all $g \in L^1(E)$ such that $s \to g(X_s, \Theta_s)ds$ is a.s. locally integrable, we have*

$$\int_0^T g(X_t, \Theta_t)dt \xrightarrow{T\to\infty} \mu(g). \tag{2.33}$$

If one further assumes some heavier growth assumptions for U, we have the following.

Theorem 2.4.7 (Bierkens-Roberts-Zitt 2019 Theorem 2)**.** *Assume that $U \in C^3$, U has a non-degenerate local minimum, in the sense that the Hessian matrix is strictly positively defined at the local minimum and assume that the following growth*

conditions hold

$$\lim_{\|x\|\to\infty} \|\nabla U(x)\| = +\infty \ , \ \lim_{\|x\|\to\infty} \frac{\|\nabla U(x)\|}{U(x)} = 0 \ , \ \lim_{\|x\|\to\infty} \frac{\|Hess(U(x))\|}{\|\nabla U(x)\|} = 0. \tag{2.34}$$

Assume, further that all the refresh rates γ_i are bounded. Then, the process is geometrically ergodic, i.e. there exist $M > 0$, $V : E \to [1, +\infty)$ and $\rho < 1$ such that for any $(x, \theta) \in E$

$$\|\mathbb{P}_{x,\theta}(Z_t \in \cdot) - \mu(\cdot)\|_{TV} \leq M(V(x, \theta) + 1)\rho^t \tag{2.35}$$

Remark 2.4.8. *Assumption $\lim_{\|x\|\to\infty} \|\nabla U(x)\| = +\infty$, adopted by the theorem, corresponds to target distributions with light tails. The behaviour of the Zig-Zag process in the case where the target distributions has heavy tails will be studied in Chapter 4.*

Using results from [GM96] the following CLT was proved.

Theorem 2.4.9 (Bierkens-Roberts-Zitt 2019 Theorem 3)**.** *Assume that all assumptions of Theorem 2.4.7 hold. Assume further that for some $\eta \in (0, 1)$, $\int_{\mathbb{R}^d} \exp\{-\eta U(x)\}dx < \infty$. Let $g : E \to \mathbb{R}$ such that there exists $0 \leq b < \frac{1-\eta}{2}$ and $k > 0$ such that $|g(x, \theta)| \leq k \exp\{bU(x)\}$ for all $(x, \theta) \in E$. Define*

$$Z_n(t) = \frac{1}{\sqrt{n}} \int_0^{nt} g(X_s, \Theta_s) - \mu(g)ds \ , \ t \geq 0.$$

Then there exists $\gamma_g \in [0, +\infty)$ such that for any starting distribution Z_n converges as $n \to \infty$ in distribution in $D[0, 1]$ to $\gamma_g B$, where B is a standard Brownian motion on time $[0, 1]$.

More CLT results for the one dimensional Zig-Zag process, which also include some heavy tails proposals can be found in [BD17]. An interesting result on that article, which we will also use in this work in Chapter 5 is that in some cases the asymptotic variance can be explicitly written as an integral of the observable function g. This is due to the representation of the asymptotic variance in terms of the solution to the Poisson equation [GM96].

Some large deviations approaches have already been used in MCMC literature to study convergence properties of the sampler. For example in [DLPD12] a Large Deviation principle for the empirical measures was established and the rate function was proposed as a tool for analysing parallel tempering. As a result a new algorithm was proposed, called *infinite swapping*. Also, on [RBS15] the authors use

the same approach to study the performance of some non-reversible Langevin samplers. This large deviation method, which allows one to compare the performance of different algorithms by comparing their rate functions, is a new criterion, different to the widely used in the literature criteria of having Geometric ergodicity or having small asymptotic variance. In [BNC19] the authors take a similar approach to study the Zig-Zag sampler. They prove a Large deviation principle for the one dimensional Zig-Zag sampler and they identify some conditions for the log-likelihood that allow them to write down the rate function explicitly. They also prove that the rate function is decreasing as a function of a constant rate γ which is in accordance with the results of [AL19].

A way to obtain a quantitative rate of converge to the stationary distribution is to study the spectral properties of the transition semi-group. In [BVL19] the authors study the L^2 spectrum of the one dimensional process with zero refresh rate and explicitly characterise it under some assumptions on the target distribution, including uni-modality. Under some further assumptions they also compare the spectral gap between a zero refresh process and processes with a small constant refresh rate. Further work on the spectrum of the generator of the Zig-Zag and the BPS has been conducted by [GN20] in the case of low temperature, i.e. when the target density is proportional to $\exp\{-\frac{1}{h}U(x)\}$ for h very small.

Even though these results provide some guarantees of good performance for the sampler, they do not explain how the sampler scales as the dimension of the target d increases to infinity, as is in practice the case. The limiting behaviour of the Zig-Zag sampler and the Bouncy Particle Sampler (see section 2.5 of this work) is studied in [BKR18] in the case where the the target density is formed of i.i.d. Gaussian components. They consider three different observables and obtain scaling limits for both processes for all three observables, assuming that the processes start from stationarity. This is in very similar spirit to work that has been done for other state of the art MCMC algorithms, for example in [RGG97, RR98, RR01, MPS12, BPR$^+$13] This allows the authors to approximate the amount of time the algorithms need to generate an essentially independent sample as a function of the dimension d. More importantly it allows to approximate the complexity of generating essentially independent samples as a function of d and compare this complexity between the two algorithms. In the case of Bouncy Particle Sample, further analysis of the limiting process of one of the three observables can be performed and this allows the authors to discuss and argue on what is the optimal refresh rate for the Bouncy Particle Sampler Algorithm.

Another interesting feature of the Zig-Zag process can be observed when

the target density is of a product form $\pi(x) = \prod_{i=1}^{d} f(x_i)$. In that case $U(x) = -\sum_{k=1}^{d} \log f(x_k)$ so if we pick refresh rate $\gamma(x) = 0$ we get from (2.25) that $\lambda_k(x,\theta) = \left[-\theta_i \frac{d}{dx_k} \log f(x_k)\right]^+$ which depends on x and θ only through their k coordinates. This means that every coordinate of the resulting Zig-Zag sampler behaves independently of the other coordinates and the d-dimensional process is a product of d independent one dimensional Zig-Zags, which favours fast convergence. This gives some intuition that the Zig-Zag sampler should perform well even on high dimension densities assuming that they have a weak correlation structure. In [BGvdMS20] the authors take advantage of the good performance of the Zig-Zag in weakly correlated structures and use Zig-Zag to sample from diffusions and diffusion bridges.

Finally, there is a different theoretical perspective for acquiring convergence rates for PDMPs that goes under the name of hypocoercivity. This technique (see [Vil06]) is based on work of [DMS09,DMS15] and was used in the context of PDMPs by [ADNR21]. The main result is a geometric decay of the $L^2(\mu)$ norm of the transition semi-group of the process as time increases. More specifically, if P^t the transition semi-group of a large family of processes that includes the Zig-Zag and the Bouncy Particle Sampler (as will be introduced in Section 2.5) and if the target measure is μ, the authors, under assumptions, prove a result of the form

$$\|P^t f - \mu(f)\|_{L^2(\mu)} \leq C \exp\{-\nu t\} \|f - \mu(f)\|_{L^2(\mu)} \tag{2.36}$$

for some $C > 1$, $\nu > 0$ and all $t \geq 0$. The main idea of the proof is to consider a norm different , but equivalent to $\|\|_{L^2(\mu)}$, for which the authors can prove that the semi-group $P^t f$ is contracting. This proves the geometric decay of P^t under the new norm and due to equivalence of norms, under $L^2(\mu)$ as well. The most interesting part of this work is that the constants C, ν are given explicitly as functions of other parameters of the process. Equation (2.36) can be considered as a measure of convergence to equilibrium and the larger the ν the better the performance of the algorithm. The authors of [ADNR21] proceed to compare different values of ν for different PDMP algorithms. Recently, more results in this direction have been provided in [LW20]. Finally, in [BRB19] the authors consider algorithms for which the hypocoercivity results hold and they use these results to provide non-asymptotic confidence intervals for the ergodic averages, estimating the expectation of some observable.

2.5 The Bouncy Particle Sampler

The second Piecewise Deterministic Markov Process that is used for MCMC in the current literature is the Bouncy Particle Sampler. This process was introduced in [BCVD18] around the same period with the Zig-Zag sampler, although seeds of this idea where presented already in [PdW12]. The algorithm tries to sample from the measure $\pi(dx) = \frac{1}{Z}\exp\{-U(x)\}dx$ for $x \in \mathbb{R}^d$ and sees the space $\mathbb{R}^d$ as a field incorporated by contours, which are sets of the form $C_a = \{x \in \mathbb{R}^d : U(x) = a\}$. A particle moves in $\mathbb{R}^d$ in straight lines with some velocity $v \in \mathbb{R}^d$. As the particle moves in direction v, a Poisson process, having rate $\lambda_0(x, v)$ that depends on the position and the speed of the particle, generates a first arrival time τ in a manner similar to the Zig-Zag. However, if the particle is at point x_τ after time τ, instead of switching the sign of one of the direction's coordinates, the particle bounces and reflects against the contour line that crosses point x_τ. The new velocity v' is therefore chosen deterministically according to the formula

$$v' = R_x v = v - 2\frac{\langle v, \nabla U(x)\rangle}{\|\nabla U(x)\|^2}\nabla U(x)$$

and the particle keeps moving in the direction of v'. This bouncing behaviour against the contours gives the process its name and we call λ_0 the bouncing rate. For technical reasons it is sometimes more efficient to add a second Poisson process with some rate $\gamma(x)$ which plays the role of randomly refreshing the velocity. More precisely, if the first arrival time of the process with rate $\gamma(x)$ is less than the one with rate λ_0, instead of bouncing against the contours the process stops and a new velocity is picked according to some rotationally invariant distribution μ_V, for example $N(0, I_d)$, or the uniform distribution on the unit sphere $\mathcal{S}^{d-1}$ if we put the restriction that $\|v\|_2 = 1$. We call this $\gamma(x)$ rate the refresh rate. This process can be seen to be PDMP with state space $E = \mathbb{R}^d \times \mathcal{V}$, where the velocity space $\mathcal{V}$ is either $\mathbb{R}^d$ or the unit sphere $\mathcal{S}^{d-1}$, depending on the definition of the algorithm. A point $(x, v) \in E$ represents the particle being at position x and moving with velocity v. The characteristics of the PDMP are of the following form. Write $E = \cup_{v\in V}\{\mathbb{R}^d \times \{v\}\}$ For the deterministic flow induced on $\mathbb{R}^d \times \{v\}$ we write $\Phi_v(t, x) = \Phi_{(x,v)}(t)$ and have

$$\begin{cases} \dfrac{dX_t}{dt} = \dfrac{d\Phi_{(x,v)}(t)}{dt} = v \\ \dfrac{dV_t}{dt} = 0 \end{cases} \tag{2.37}$$

The rate is $\lambda(x,v) = \lambda_0(x,v) + \gamma(x)$ and the jumping measure is such that $Q(x,v,dy,dv') = \frac{\lambda_0(x,v)}{\lambda(x,v)}\delta_{(x,R_x v)}(dx,dv) + \frac{\gamma(x)}{\lambda(x,v)}1_{y=x}\mu_V(dv')$.

In algorithmic terms the process is given by the following.

Algorithm 2.5.1 (Bouncy Particle Sampler)**.**

1. *Set* $t_{current} = 0$
2. *Start from point* $(X_{t_{current}}, V_{t_{current}}) = (x,v)$.
3. *The* x*-component moves along the line* $\{x + tv, t \geq 0\}$.
4. *Construct a Poisson Process with intensity* $\{m(t) = \lambda(x + tv, v), t \geq 0\}$.
5. *Let* T *be the first arrival time of the Poisson Process, i.e. for all* $t \geq 0$, $\mathbb{P}(T \geq t) = \exp\{-\int_0^t m(s)ds\}$.
6. *For* $t \in [t_{current}, t_{current} + T)$ *set* $X_t = x + (t - t_{current})\theta$ *and* $V_t = v$.
7. *Set* $x = x + Tv$, $t_{current} = t_{current} + T$
8. *With probability* $\frac{\lambda_0(x,v)}{\lambda(x,v)}$ *set*

$$v = v - 2\frac{\langle v, \nabla U(x)\rangle}{\|\nabla U(x)\|^2}\nabla U(x).$$

9. *Otherwise randomly pick* v *according to the measure* μ_V
10. *Repeat from the Step 2.*

Remark 2.5.2. *Note that this algorithm can only be implemented in target densities for which* U *is differentiable. However, some ways to generalise the algorithm to targets where* U *is only piecewise continuous can be found in [Pak17]. Furthermore, a way to define the process on bounded domains, subsets of* $\mathbb{R}^d$ *can be found in [BBCD*$^+$*18]. Finally, we note that the Bouncy Particle Sampler, as a continuous time algorithm, has inspired the construction of discrete time algorithms that try to mimic its dynamics, for example [VBCDD17, ST17, LS19].*

In a similar fashion to Zig-Zag one can identify the bouncing rate in terms of the gradient log-likelihood of the posterior so that the the process targets the posterior.

Proposition 2.5.3 (Proposition 1 of Bouchard-Cote, Volmer, Doucet 2018)**.** *Assume that there exists a function* $U \in C^1$ *such that* $\int_{\mathbb{R}^d} \exp\{-U(x)\}dx < \infty$ *and such that the overall rate of the BPS is of the form*

$$\lambda(x,v) = [\langle \nabla U(x), v\rangle]^+ + \gamma(x). \tag{2.38}$$

Then the process admits

$$\mu(dx, dv) = \frac{1}{Z} \exp\{-U(x)\} dx \mu_V(dv)$$

as invariant.

2.5.1 Local Bouncy Particle Sampler

In [BCVD18] the authors propose a different version of the algorithm that takes into consideration the structure of the target distribution. This is in similar flavour to the sub-sampling approach of the Zig-Zag. In many applications the target density is of the form $\pi(x) = \prod_{f \in F} \pi_f(x_f)$ where F is an index set called set of factors. For any $f \in F$ we assign $N_f \subset \{1, ..., d\}$ a subset of coordinates and we define x_f to be the restriction of x to N_f. In that case we can write

$$U(x) = \sum_{f \in F} U_f(x) \tag{2.39}$$

where

$$\frac{\partial}{\partial x_k} U_f(x) = 0 \text{ for all } x_k \notin N_f \tag{2.40}$$

This allows the following modification of the BPS. Instead of one Poisson process with rate $\lambda(x, v) = [\langle \nabla U(x), v \rangle]^+$ we consider one Poisson process for each factor $f \in F$ with rate $\lambda_f(x, v) = [\langle \nabla U_f(x_f), v \rangle]^+$. If an event is generated from λ_f the velocity v bounces against the contours of U_f, i.e.

$$v' = R_x^f v = v - 2 \frac{\langle v, \nabla U_f(x) \rangle}{\|\nabla U_f(x)\|^2} \nabla U_f(x) \tag{2.41}$$

The interesting part of this modification is that due to (2.40), the bounce on (2.41) does not change the components k of v such that $k \notin N_f$. The particle keeps moving with the same speed in the coordinate that is not in the factor f that generated the bouncing event. All these Poisson events can be generated via Poisson Thinning and the algorithm has the following form.

Algorithm 2.5.4.

1. *Set $t_{current} = 0$, specify a value $T_{run} > 0$*
2. *Start from point $(X_{t_{current}}, V_{t_{current}}) = (x, v)$.*
3. *The x-component moves along the line $\{x + tv, t \geq 0\}$.*

4. *For every $f \in F$ construct a Poisson Process with intensity $\{m_f(t) = \lambda_f(x + tv, v), t \geq 0\}$.*

5. *Find a rate $M(t), t \geq 0$ such that $m_f(t) \leq M(t)$ for all $t \leq T_{run}$ and $f \in F$ and such that one can simulate from rate M*

6. *Let T be the first arrival time of the Poisson Process with rate $M(t)$, i.e. for all $t \geq 0$, $\mathbb{P}(T \geq t) = \exp\{-\int_0^t M(s)ds\}$.*

7. *If $T > T_{run}$*

   (a) *For $t \in [t_{current}, t_{current} + T)$ set $X_t = x + (t - t_{current})\theta$ and $V_t = v$. Set $x = x + Tv$, $t_{current} = t_{current} + T$*

8. *Else*

   (a) *For $t \in [t_{current}, t_{current} + T)$ set $X_t = x + (t - t_{current})\theta$ and $V_t = v$.*

   (b) *Set $x = x + Tv$, $t_{current} = t_{current} + T$*

   (c) *Pick f uniformly at random from F*

   (d) *With probability $\frac{m_f(x,v)}{M(t)}$ set $v = R_v^f v$.*

9. *Repeat from the Step 2.*

2.5.2 Convergence Results

In [BCVD18] the authors prove ergodicity and Law of Large Numbers for the BPS, assuming that the refresh rate $\gamma(x) = \gamma > 0$. They also prove that this assumption of having positive refresh rate is essential, as they present a counterexample where a BPS with zero refresh rate targeting a product i.i.d. gaussians and starting from $x = (1, 0)$ and $v = (0, 1)$ will never be able to enter the unit sphere and is therefore reducible. This problem can be overcome with the addition of the refresh rate which allows the process to change direction more often. We have the following.

Proposition 2.5.5 (Bouchard-Cote, Vollmer, Doucet 2018)**.** *Let $(Z_t)_{t\geq 0} = ((X_t, V_t))_{t\geq 0}$ be a BPS on $\mathbb{R}^d$, assume assumptions of Proposition 2.5.3 hold, $\gamma(x) = \gamma > 0$ and μ_V is the $N(0, I_d)$. Then for μ-almost every starting point*

$$\frac{1}{T}\int_0^T f(Z_s)\ ds = \mathbb{E}_\mu[f]\ \ a.s. \tag{2.42}$$

In [DBCD19] the authors establish Geometric ergodicity for the process. Results in the same direction with different assumptions were later proven in [DGM20].

Proposition 2.5.6 (Deligiannidis, Bouchard-Cote, Doucet 2019)**.** *Assume that* $U \in C^2(\mathbb{R}^2)$*, with* $\frac{\partial^2}{\partial x_i \partial x_j} U(x)$ *is locally Lipschitz for all* i, j*, that*

$$\int_{\mathbb{R}^d} \exp\{-U(x)\} \|\nabla U(x)\| dx < \infty$$

and that

$$\liminf_{\|x\| \to \infty} \frac{\exp\{U(x)/2\}}{\sqrt{\|\nabla U(x)\|}} > 0.$$

Suppose further that there exists $c > 0$ *such that for all* $(x, v) \in E$

$$V(x, v) = \frac{\exp\{U(x)/2\}}{\lambda^{1/2}(x, -v)} \geq c.$$

Also, assume that the refresh rate satisfies $\gamma(x) = \gamma$ *and that on a refresh event the new velocity is distributed according to the uniform measure on the unit sphere of* $\mathbb{R}^d$*. Assume further that one of the following two conditions hold*

1. $\lim_{\|x\| \to \infty} \|\nabla U(x)\| = \infty$ *,* $\limsup_{\|x\| \to \infty} \|\Delta U(x)\| \leq a$ *and* $\gamma > (2a+1)^2$*.*
2. $\liminf_{\|x\| \to \infty} \|\nabla U(x)\| = 2b > 0$ *,* $\limsup_{\|x\| \to \infty} \|\Delta U(x)\| \leq C < \infty$ *and* $\gamma \leq b/c_d$ *where* $c_d = 16\sqrt{d}$*.*

Then the process is Geometrically ergodic, with Lyapunov function V*.*

Note, also, that the authors prove geometric ergodicity under different assumptions in the case of very thin tails and in the case of heavier tails they argue that one can transform the space to lighter tails, run the BPS there and then rescale back to the original state and gain Geometric ergodicity. For more information one can look at Theorem 3.3 on [DBCD19].

Addressing the issue of the need of extra refreshment for the Bouncy Particle Sampler, a natural question is how large this refreshment should be. Although we understand it should be positive, one could argue that it should also not be very large as the Bouncy Particle Sampler would tend to lose its non-reversible property. As we mentioned in section 2.4.3, in [BKR18] the authors prove a scaling limit for some observables of the Bouncy Particle Sampler and try to infer on the optimal behaviour of the refresh rate via studying the limiting process. Similarly, in [DPBCD21] the authors prove a scaling limit for the Bouncy Particle Sampler as the number of dimensions increases to infinity in a slightly different setting. They consider the Bouncy Particle Sampler where the velocity is refreshed with a constant rate γ and the new velocity if sampled from an isotropic Gaussian distribution in $\mathbb{R}^d$. They consider both the position and the velocity of the first coordinate

of the bouncy particle sampler and they prove that as the number of dimensions increases to infinity, this pair converges to a process called Randomized Hamiltonian Monte Carlo. This is a process, introduced in [BRSS17], which follows deterministic Hamiltonian dynamics and these dynamics refresh after a random time period T that follows exponential random variable with some parameter γ. In the context of the next theorem, this γ is the same as the refresh rate for the BPS.

Theorem 2.5.7 (Deligiannidis, Paulin, Bouchard-Cote, Doucet 2021)**.** *Assume that for the n dimensional BPS the invariant measure is*

$$\pi_n(dx) = \frac{1}{Z}\exp\{-U(x)\}\exp\{-1/2v^2\}dxdv.$$

Assume that $U_n \in C^2(\mathbb{R}^d)$ and that there exist $M > m > 0$ such that for all $x \in \mathbb{R}^d$,

$$mI_d \leq Hess(U_n)(x) \leq MI_d.$$

Assume as well that U_n achieves its minimum at $0 \in \mathbb{R}^d$. Assume further that if $f : \mathbb{R} \to [0, +\infty)$ is the marginal of π_n on the first coordinate of position, i.e.

$$f(x) = \int_{\mathbb{R}^{d-1}\times\mathbb{R}^d} \pi_n(x, x_2..., x_n; v_1, ..., v_n)dx_2...dx_ndv_1...dv_n$$

then $f(x) = \exp\{-W(x)\}$, where $W \in C^\infty(\mathbb{R})$, $\lim_{|x|\to+\infty} W(x) = 0$ and

$$\int_{\mathbb{R}} \exp\{W(x)\}\left(|W''(x)| + |W'(x)|^2\right)dx < +\infty.$$

Let $a \in [0, 1)$ and assume that the velocity is refreshed with rate γ_{ref} and when the velocity v is refreshed, the new velocity is picked as $av+\sqrt{1-a^2}\xi$ where $\xi \sim N(0, I_n)$. Assume that the BPS process $(Z_n(t))_{t\geq 0} = (Z_n^1(t), ..., Z_n^n(t))_{t\geq 0}$ starts from π_n. Then the process $(Z_n^1(t))_{t\geq 0}$ corresponding to the first coordinate and velocity components of the BPS converges weakly to the one dimensional RHMC process $(Z_0(t))_{t\geq 0}$ as $n \to \infty$ with refreshment rate γ_{ref}.

Furthermore, the authors observe that the above theorem would hold if the potential U_n is defined in $\mathbb{R}^{n\cdot d}$ and can be decomposed into d-dimensional blocks as $U_n(x_1, ..., x_d) = \sum_{k=1}^n U(x_k)$ where U is defined in $\mathbb{R}^d$. When n is much larger than d this is a weakly correlated target. The authors observe that for d fixed, as $n \to \infty$ the first d coordinates of the BPS converge to a d-dimensional RHMC.

The RHMC cannot be used in practice in an MCMC context as one would need to simulate directly from the Hamiltonian dynamics to implement the pro-

cess, which is a non realistic scenario. It can, however, be used as a proxy for the behaviour of the high dimensional BPS. The authors study the behaviour of a d dimensional RHMC and prove the following dimension free convergence rates under some constraints in the Hessian of the targeted potential.

Theorem 2.5.8 (Deligiannidis, Paulin, Bouchard-Cote, Doucet 2021)**.** *Suppose that $U \in C^2(\mathbb{R}^d)$ and there exist $M > m > 0$ such that for all $x, v \in \mathbb{R}^d$,*

$$m\|v\|_2^2 \leq \langle v, \nabla^2 U(x) v\rangle \leq M\|v\|_2^2.$$

Let $a \in [0, 1)$. Consider a d-dimensional RHMC $(Z_0(t))_{t\geq 0}$ targeting the product between $\pi(dx) = \frac{1}{Z}\exp\{-U(x)\}dx$ and $N(0, I_d)$. Let the refreshment be

$$\gamma = \frac{1}{1-a^2}(2\sqrt{M+m} - \frac{m(1-a)}{\sqrt{m+M}}) \tag{2.43}$$

and let

$$c = \frac{(1+a)m}{\sqrt{m+M}} - \frac{am^{3/2}}{2\sqrt{m+M}}.$$

Then there exists an explicit $C > 0$ such that for all $f \in L^2(\pi)$ with $\pi(f) = 0$ and all $t \geq 0$, we have

$$\|P^t f\|_2^2 \leq \min\{1, C\exp\{-ct\}\}\|f\|_2^2. \tag{2.44}$$

The authors argue that the value (2.43) seems to be optimal for the convergence of RHMC in the sense that for any other refresh rate the values of μ that satisfy (2.44) are smaller than the one given in the theorem. Equation (2.43) could also serve as the rate one should pick for the BPS.

2.6 Alternatives or Complements to Zig-Zag or BPS

In this section we will present some other algorithms arising in the context of Piecewise Deterministic Processes in MCMC.

2.6.1 NUZZ Sampler

In [CHP20] a different numerical implementation of the Zig-Zag sampler is proposed. Instead of using Poisson thinning, which would in some cases need many useless evaluations of the gradient log-likelihood of the posterior, the authors use the exponential representation of Poisson processes (Proposition 2.2.2) in the following way. If the process starts from (x, θ) the first switching time τ satisfies

$\int_0^\tau \lambda(x+s\theta,\theta)ds = E$ where $E \sim \exp(1)$, so they propose to sample E, approximate λ via some polynomial $\tilde{\lambda}$ and then numerically find the root of the equation $\int_0^\tau \tilde{\lambda}(x+s\theta,\theta)ds - E = 0$. This creates a biased stochastic process, but they argue that the distance between the new invariant measure and the one they initially try to target is not large. They also provide some numerical examples.

2.6.2 The Coordinate Sampler

In [WR20] a PDMP alternative to Zig-Zag or Bouncy Particle Sampler was introduced. The algorithm generalised the Gibbs sampler in continuous time. The state space is $E = \mathbb{R}^d \times V$ where $V = \{\pm e_k, k = 1, ..., d\}$ and $\{e_k, k = 1, ..., d\}$ is the standard basis of $\mathbb{R}^d$. The process moves in straight lines with velocity $v \in V$, which means that one of the coordinates moves with speed ± 1 and all other coordinates stay frozen. The deterministic dynamics are therefore given by

$$\begin{cases} \dfrac{dX_t}{dt} = \dfrac{d\Phi_{(x,v)}(t)}{dt} = v \\ \dfrac{dV_t}{dt} = 0 \end{cases} \tag{2.45}$$

Denoting $\lambda(x) = \sum_{v \in V} \lambda(x, v)$, the event rate is $\lambda(x, v) = [\langle v, \nabla U(x)\rangle]^+ + \gamma$ where γ denotes a refresh rate and the jumping measure is
$Q((x, v), dy, dv_0) = \sum_{v' \in V} \frac{\lambda(x,-v')}{\lambda(x)} \delta_{(x,v')}(dy, dv_0)$. This means that when the process switches, it picks a new direction $\pm e_k$ with weight proportional to $\lambda(x, \mp e_k)$ and starts moving in that direction. They authors prove the following.

Proposition 2.6.1 (Wu-Robert 2020)**.** *If $\gamma > 0$ the coordinate sampler has $\mu(dx, dv) = \frac{1}{2dZ} \exp\{-U(x)\}dxdv$ as invariant distribution.*

The authors also prove Geometric Ergodicity for the process under similar conditions to Proposition 2.5.6 in section 2.5.2 of this work, which proves geometric ergodicity for the BPS.

An interesting feature of the algorithm is that the switching rates tend to be smaller than the corresponding of Zig-Zag since the overall rate of switching for the coordinate sampler for $v = \pm e_k$ is $\lambda(x, v) = [v_k \partial_k U(x)]^+$ whereas the overall rate for the Zig-Zag while at state (x, θ) is $\lambda(x, \theta) = \sum_{k=1}^d [\theta_k \partial_k U(x)]^+$. This means that the coordinate sampler has smaller computational cost to implement per unit of time, which compensates the drawback that the process only moves in one coordinate at the time and might explore the space slower.They also present several numerical examples and they claim that the coordinate sampler tends to outperform the Zig-Zag sampler in some high dimensional examples.

2.6.3 Generalised Bouncy Particle Sampler

The Generalised Bouncy Particle Sampler was introduced in [WR19] as a way to overcome the reducibility problems that may appear in Bouncy Particle Sampler if there is no extra refreshment. The authors observe that when the Bouncy Particle Sampler bounces against the contours while at point (x, v), the new velocity has the form

$$v' = R_x v = v - 2\frac{\langle v, \nabla U(x)\rangle}{\|\nabla U(x)\|^2}\nabla U(x) = v_2 - \frac{\langle v, \nabla U(x)\rangle}{\|\nabla U(x)\|^2}\nabla U(x)$$

for $v_2 = v - \frac{\langle v, \nabla U(x)\rangle}{\|\nabla U(x)\|^2}\nabla U(x)$ and note that $\langle v_2, \nabla U(x)\rangle = 0$. This is the way to decompose the new velocity v' to the projection on the one dimensional subspace generated by $\nabla U(x)$ and its perpendicular component v_2. In the generalised BPS, the component v_2 is instead randomized and sampled from the normal distribution from the $d-1$ dimensional subspace $N_x^{\perp} = \{v_2 \in \mathbb{R}^d : \langle v_2, \nabla U(x)\rangle = 0\}$. They allow, however, the projection on $\nabla U(x)$ to be the same as it would had been in the classic Bouncy Particle Sampler. The new velocity is then chosen to be $v' = v_2 - \frac{\langle v, \nabla U(x)\rangle}{\|\nabla U(x)\|^2}\nabla U(x)$.

The deterministic dynamics are chosen to be straight lines and when at point $(x, v) \in E$ the process is at $x \in \mathbb{R}^d$ and tends to move with velocity v. The bouncing rate is set as in the case of the Bouncy Particle Sampler to be $\lambda(x, v) = [\langle v, \nabla U(x)\rangle]^+$. The authors prove that an invariant measure for the process is $\mu(dx, dv) = \frac{1}{Z}\exp\{-U(x)\}\psi_d(dv)dx$ where ψ_d is the density of the d-dimensional standard normal distribution. They also find conditions under which the Generalised BPS $(X_t, V_t)_{t\geq 0}$, with velocities appropriately normalised such that $\|V_t\|_2 = 1$, has as unique invariant measure the product between $\pi(dx) = \frac{1}{Z}\exp\{-U(x)\}dx$ and the uniform distribution in the unit sphere. This overcomes the reducibility problem of the standard BPS.

Finally the authors present some numerical examples where they compare the Generalised BPS without any refreshment against the standard BPS with different refreshment rates and they argue that the generalised process sometimes works better and has the advantage that since the refresh rate is set equal to zero there is no parameter to tune, in contrast to the classic BPS.

2.6.4 Boomerang Sampler

In [BGKR20] the authors propose a new PDMP that, unlike the Zig-Zag or the BPS does not move in straight lines. Instead they move along the contours of a Gaussian distribution. More precisely, if the process starts from $(x_0, v_0) \in \mathbb{R}^d \times \mathbb{R}^d$, the deter-

ministic dynamics are of the form $X_t = x_0 \cos(t) + v_0 \sin(t)$ and $V_t = -x_0 \sin(t) + v_0 \cos(t)$. This gives the name to the process as, when $x, v \in \mathbb{R}$, the pair (x, v) is moving in circles. The idea of using this type of dynamics in a Piecewise Deterministic Process setting was already introduced in [VBCDD17] (see also [MS18]). The variable v is assumed to be auxiliary and plays the role of the speed. The target measure is assumed to have a density of the form $\exp\{-U(x) - 1/2x'\Sigma^{-1}x - 1/2v'\Sigma^{-1}v\}$ with respect to the Lebesgue measure on $\mathbb{R}^d \times \mathbb{R}^d$, for some positive definite matrix Σ. Equivalently, the target distribution has density $\exp\{-U(x)\}$ with respect the the product of two Gaussian measures having correlation Matrix Σ. The rate of switching is set to be $\lambda(x, v) = [\langle \nabla U(x), v \rangle]^+$. When a switching event happens at position (x, v), the new velocity is set to be $R(x)v = v - 2\frac{\nabla U(x), v}{\|\Sigma^{1/2}\nabla U(x)\|^2}\Sigma\nabla U(x)$. The authors also allow a positive constant refresh rate $\lambda^{ref} > 0$ in order to avoid irreducibility problems similar to BPS. When a refresh event occurs, the new velocity is picked independently from a $N(0, \Sigma)$.

The authors, also, introduce a version of the Boomerang Sampler, called Factorised Boomerang Sampler that is closer in nature to the Zig-Zag process. More specifically, the process still follows the same cyclic dynamics but it has one switching rate in every coordinate i, $\lambda_i(x, v) = [v_i \partial_i U(x)]^+ + \lambda_i^{ref}$. When an event from the i'th rate occurs only the i'th component, v_i, of the velocity is updated. If the event was due to the refreshment part λ_i^{ref} then a new v_i is picked according to some normal distribution, while if the event was due to the $[v_i \partial_i U(x)]^+$ part, the v_i switches to $-v_i$.

Furthermore, the authors introduce sub-sampling variants of the Boomerang Sampler, using control variates techniques, similar to [BFR19] to make the algorithm suitable for a Big Date setting, when the number of observations n that induce the posterior is large. As observed in simulations on logistic regression model and diffusion bridges the Boomerang sampler can outperform other PDMP algorithms like Zig-Zag and Bouncy Particle Sampler, when either the dimension of the space d or the number of observations n increases to infinity.

Chapter 3

Multi-Directional Zig-Zag

In its original formulation the Zig-Zag process in $\mathbb{R}^d$ moves only in directions parallel to a vector from the set $\{-1,+1\}^d$. However, one could, instead, allow more directions, hoping that it will be easier for the process to explore the space and, therefore, have faster convergence. In this chapter we will introduce the Zig-Zag process where more directions allowed. We will, then, present equations the switching rates need to satisfy for the probability measure of interest to be invariant for the process. Furthermore, we will prove Ergodicity results for the process which extends to Geometric ergodicity when the target distribution satisfies some growth condition at the tails, including being light tailed. Finally we will support our results by presenting some simulations.

3.1 Introduction of the Process and Invariant Measure

In this section we will introduce the Zig-Zag process with more directions and we will establish conditions on the rate function that ensure that the process has the distribution of interest invariant. The process was already introduced in the (unpublished) author's Master thesis [Vas17] and this section summarises results from that work.
We begin by assuming that our space is $E = \mathbb{R}^d \times \{\theta^1, \theta^2, ..., \theta^n\}^d$, where $\theta^1 < ... < \theta^n$ are real numbers and we want to target a measure μ on E , with

$$\mu(dx, d\theta) = \frac{1}{Z} \exp\{-U(x)\}\mu_0(dx, d\theta), \tag{3.1}$$

where μ_0 is the product measure between the Lebesgue measure in $\mathbb{R}^d$ and the uniform measure on $\{\theta^1, ..., \theta^n\}^d$ and $Z = \int_{\mathbb{R}^d} \exp\{-U(x)\}dx$. Write $\Theta_0 = \{\theta^1, \theta^2, ..., \theta^n\}$ for the set of available values for the coordinates of the velocity and write $\Theta = \Theta_0^d$

for the set of all available velocities.

Remark 3.1.1. *For the rest of this chapter we will be using the following notation. We write θ^j to symbolise an element of $\Theta_0 = \{\theta^1, ..., \theta^n\}$, while we write θ_i to symbolise the i coordinate of the velocity $\theta = (\theta_1, ..., \theta_d) \in \Theta_0^d$.*

The use of this extension of the original Zig-Zag is more apparent in more than one dimensions, where the new process will be allowed to explore the space through more angles. In dimension one, there are still only two directions where the process can be directed towards, but by allowing more vectors θ, we allow the process to move in a different speed than just 1.
This extension of the Zig-Zag process will be called **Multi-Directional Zig-Zag (MDZZ)**. The main difference between this extension and the original Zig-Zag process is that, in the original Zig-Zag if the process was to change the i coordinate of its velocity, the change would happen deterministically from $+1$ to -1 and vice versa, whereas in the multi-directional Zig-Zag, there are more than one possible new choices to jump to. More precisely, the new coordinate of the velocity has to be chosen from the elements of Θ_0. If the process has velocity $\theta = (\theta_1, ..., \theta_d)$, for any coordinate $i \in \{1, ..., d\}$ and for any $\theta^j \in \Theta_0$, we need to introduce a Poisson Process that gives the rate in which the process changes the i coordinate of the velocity from θ_i to θ^j. We introduce the notation $F_i^j[\theta] \in \Theta$ for the velocity $\theta = (\theta_1, ..., \theta_d)$ whose i coordinate has switched to θ^j, i.e.

$$\begin{cases} F_i^j[\theta]_k = \theta^j \text{ for } k = i \\ F_i^j[\theta]_k = \theta_k \text{ for } k \neq i, \end{cases} \tag{3.2}$$

and we write $\lambda_i(x, \theta, \theta^j), (x, \theta) \in E$ for the hazard rate of the Poisson Process that controls the switching from θ to $F_i^j[\theta]$. We assume, as usual, that this λ is a non-negative function, locally integrable. The evolution of the process is described in the following algorithm.

Algorithm 3.1.2.

1. *Start from point $(x, \theta) = (x_1, .., x_d; \theta_1, .., \theta_d) \in E$.*
2. *The x-component moves along the line $\{x + t\theta, t \geq 0\}$.*
3. *For every coordinate $i \in \{1, .., d\}$ and every $\theta^j \neq \theta_i$ we construct a Poisson Process with intensity $\{m_{i,j}(t) = \lambda_i(x + t\theta, \theta, \theta^j), t \geq 0\}$. Suppose that the first process to generate a point is the process $m_{i,j}$ and the time the point was generated is T.*

4. *Set $x = x + T\theta$ and switch the i component of θ to θ^j, i.e. switch θ to $F_i^j[\theta]$.*

5. *Repeat from the first step.*

Note that the multi-directional Zig-Zag as described in Algorithm 3.1.2 can be simulated using Poisson thinning in order to simulate the first arrival times of the Poisson processes, similarly to the standard Zig-Zag. The process can be seen in the setting of Davis (in Section 2.3) for PDMPs. When the process is at $(x, \theta) \in E$ it follows the deterministic dynamics given by

$$\begin{cases} \dfrac{dX_t}{dt} = \dfrac{d\Phi_{(x,\theta)}(t)}{dt} = \theta \\ \dfrac{d\Theta}{dt} = 0 \end{cases} \tag{3.3}$$

the jump intensity is $\lambda(x,\theta) = \sum_{i=1}^{d} \sum_{\substack{\theta^j \in \Theta_0 \\ \theta^j \neq \theta_i}} \lambda_i(x, \theta, \theta^j)$ and the jumping measure is $Q((x,\theta), dy, d\eta) = \delta_x(dy) \sum_{i=1}^{d} \sum_{\substack{\theta^j \in \Theta_0 \\ \theta^j \neq \theta_i}} \frac{\lambda_i(x,\theta,\theta^j)}{\lambda(x,\theta)} \delta_{F_i^j[\theta]}(d\eta)$. Since in each coordinate the process has a speed at most $\theta_{max} = \max\{|\theta_1|, |\theta_n|\}$, after any given time T and starting from any point (x, θ) the process will not have exit a ball of radius $\sqrt{d}\theta_{max}T$ around the starting point x. Therefore, assuming that the switching rates are bounded on bounded sets, the main Assumption 3.1 of [Dav84] (presented as Assumption 2.3.1 in Chapter 2 of this work) is satisfied.

Proposition 3.1.3. *Assume that for all i and all θ^j the switching rates $\lambda(x, \theta, F_i^j[\theta])$ are locally bounded functions of x. For any starting point (x, θ) and any time $T > 0$, if N_T is the number of switches of the multi-directional Zig-Zag before time T, we have $\mathbb{E}_{x,\theta}[N_T] < \infty$.*

An immediate consequence of Proposition 3.1.3 is that we can describe its generator.

Proposition 3.1.4 (Strong Generator)**.** *For any $f \in C^1(E)$ the generator of the multi-directional Zig-Zag on $\mathbb{R}^d$ is*

$$\mathcal{L}f(x,\theta) = \sum_{i=1}^{d} \left(\theta_i \partial_i f(x,\theta) + \sum_{\substack{m=1 \\ \theta^m \neq \theta_i}}^{n} \lambda_i(x, \theta, \theta^m)[f(x, F_i^m[\theta]) - f(x,\theta)] \right). \tag{3.4}$$

Proof of Proposition 3.1.4. This is a consequence of Theorem 2.3.3 (see also Theorem 5.5 of [Dav84]). □

This in turn allows us to describe the equations that the switching rates must satisfy for the process to have the probability measure of interest, invariant.

Proposition 3.1.5 (Invariant Measure)**.** *Assume that $U \in C^1$, for all $i = 1, .., d$, $\lambda_i \in C^1$ and for every $(x, \theta) \in E$*

$$\sum_{\substack{m=1 \\ \theta^m \neq \theta_i}}^{n} \lambda_i(x, \theta, F_i^m[\theta]) - \lambda_i(x, F_i^m[\theta], \theta) = \theta_i \partial_i U(x, \theta). \tag{3.5}$$

If $U \in C^1$ then the multi-directional Zig-Zag process has the measure μ defined in (3.1) invariant.

Proof of Proposition 3.1.5. For any $k \in \{1, ..., d\}$ and $f \in C_c^1(E)$ let

$$\mathcal{L}_k f(x, \theta) = \theta_k \partial_k f(x, \theta) + \sum_{\substack{m=1 \\ \theta^m \neq \theta_k}}^{n} \lambda_k(\theta, F_k^m \theta, x)[f(x, F_k^m \theta) - f(x, \theta)]$$

so that

$$\mathcal{L} f(x, \theta) = \sum_{k=1}^{d} \mathcal{L}_k f(x, \theta)$$

Consider the first term of the sum and the quantity $\mathbb{E}_\mu[\mathcal{L}_1 f]$. Using, integration by parts and rearranging the sums on the third equality, we get for any $f \in C_c^1(E)$

$$\begin{aligned}
&Z|\Theta| \int_E \mathcal{L}_1 f(x, \theta) \mu(dx, d\theta) = \\
&= \sum_{\theta \in \Theta^d} \int_{\mathbb{R}^d} \exp\{-U(x)\} \left(\theta_1 \partial_1 f(x, \theta) + \sum_{\substack{m=1 \\ \theta^m \neq \theta_1}}^{n} \lambda_1(x, \theta, F_1^m[\theta])(f(x, F_1^m[\theta]) - f(x, \theta)) \right) dx = \\
&= \sum_{\theta \in \Theta^d} \int_{\mathbb{R}^d} \theta_1 \partial_1 U(x) f(x, \theta) \exp\{-U(x)\} dx + \\
&+ \sum_{\theta \in \Theta^d} \int_{\mathbb{R}^d} \exp\{-U(x)\} \sum_{\substack{k=1 \\ \theta^m \neq \theta_1}}^{n} \lambda_1(x, \theta, F_1^m[\theta]) f(x, F_1^m[\theta]) dx - \\
&- \sum_{\theta \in \Theta^d} \int_{\mathbb{R}^d} \exp\{-U(x)\} \sum_{\substack{m=1 \\ \theta^m \neq \theta_1}}^{n} \lambda_1(x, \theta, F_1^m[\theta]) f(x, \theta) dx = \\
&= \sum_{\theta \in \Theta^d} \int_{\mathbb{R}^d} \exp\{-U(x)\} f(x, \theta) \left\{ \theta_1 \partial_1 U(x) - \sum_{\substack{m=1 \\ \theta^m \neq \theta_1}}^{n} \lambda_1(x, \theta, F_1^m[\theta]) + \sum_{\substack{m=1 \\ \theta^m \neq \theta_1}}^{n} \lambda_1(x, F_1^m[\theta], \theta) \right\} dx \\
&= 0
\end{aligned}$$

where the last equality is due to (3.5). The same argument can be applied to any

$\mathcal{L}_k f$ for any k and therefore for any $f \in C_c^1(E)$

$$\mathbb{E}_\mu[\mathcal{L}f] = 0.$$

All the assumptions of Corollary 19 of [DGM18] are satisfied so $C_c^1(E)$ is core for the generator seen as a semigroup on $C_0(E)$ and therefore μ is invariant for the process from Proposition 2.1.9. □

If we consider $\lambda_i(x, \theta, F_i^m[\theta]) - \lambda_i(x, F_i^m[\theta], \theta)$ as the unknowns of the system (3.5) then there are $O(n^2)$ unknown quantities and only $O(n)$ equations that need to be satisfied. For large n, this system has a huge number of degrees of freedom and this number becomes even bigger as n increases. In the original case, there was only one unknown and one equation and therefore a unique solution for the quantity $\lambda(x, 1) - \lambda(x, -1)$. This means that one can use many different processes to target the interested measure and compare them to find the most efficient one.
However, no matter how one picks the rates of switching, the system of equations induced by (3.5) forces a restriction on the type of directions we are allowed to choose.

Proposition 3.1.6. *Suppose that $U \in C^1(E)$. If the process has μ (introduced in 3.1) as invariant distribution and assuming that μ is a probability measure, then*

$$\sum_{m=1}^{n} \theta^m = 0. \tag{3.6}$$

Proof of Proposition 3.1.6. Fix $i \in \{1, ..., d\}$. Fix an $x \in \mathbb{R}^d$ and consider a $\theta = (\theta_1, ..., \theta_d) \in \{-1, +1\}^d$. Fix the values of all coordinates of θ except for θ_i. Let us consider the sum of all the equations, involving λ_i on (3.5), over all the possible values of θ_i (taken from the set $\{\theta^1, ..., \theta^n\}$). We get

$$\begin{aligned}
\left(\sum_{m=1}^{n} \theta^m\right) \partial_i U(x) &= \sum_{\theta_i \in \{\theta^1, ..., \theta^n\}} \sum_{\substack{m=1 \\ \theta^m \neq \theta_i}}^{n} (\lambda_i(x, \theta, F_i^m[\theta]) - \lambda_i(x, F_i^m[\theta], \theta)) = \\
&= \sum_{\theta_i \in \{\theta^1, ..., \theta^n\}} \sum_{\substack{m=1 \\ \theta^m \neq \theta_i}}^{n} \lambda_i(x, \theta, F_i^m[\theta]) - \sum_{\theta_i \in \{\theta^1, ..., \theta^n\}} \sum_{\substack{m=1 \\ \theta^m \neq \theta_i}}^{n} \lambda_i(x, F_i^m[\theta], \theta) = \\
&= \sum_{\theta_i \in \{\theta^1, ..., \theta^n\}} \sum_{\substack{m=1 \\ \theta^m \neq \theta_i}}^{n} \lambda_i(x, \theta, F_i^m[\theta]) - \sum_{\theta_i \in \{\theta^1, ..., \theta^n\}} \sum_{\substack{m=1 \\ \theta^m \neq \theta_i}}^{n} \lambda_i(x, \theta, F_i^m[\theta]) = 0
\end{aligned}$$

Overall, we get that for all $i = 1, .., d$ and all x,

$$\left(\sum_{m=1}^{n} \theta^m \right) \partial_i U(x) = 0.$$

This is true for all $i = 1, ..., d$. Given that U cannot be constant everywhere, since we are targeting a probability measure, we get the result. □

The reason we need the directions to sum to zero is that we force the process to have invariant measure whose marginal with respect to the directions is the uniform distribution. If we relax this assumption we get a different system of equations for the rates than (3.5). Having an invariant measure with uniform marginal on the directions means that the process tends to spend the same amount of time travelling in every direction $\theta \in \Theta$. This is not necessarily a useful property. In practice, one would like to spend more time travelling parallel to the contour lines of the invariant distribution and it would be very interesting to create an algorithm that could do exactly that. Most likely this would involve using an adaptive algorithm. Even though we still haven't developed an adapted algorithm in this multi-directional setting, we can, however, identify the rates that one needs to choose to target a measure with non-uniform marginals. More specifically, suppose that we are now trying to target a measure

$$\mu(dx, d\theta) = \frac{1}{Z} \exp\{-U(x)\} \Phi(\theta) \mu_0(dx, d\theta) \tag{3.7}$$

for some probability mass function Φ on Θ, which assigns different **weights** to each direction.

Proposition 3.1.7 (Invariant Weighted Measure)**.** *Consider a multi-directional Zig-Zag process, with velocity space $\Theta = \Theta_0^d$ and the rates picked such that for all i, $\lambda_i \in C^1(E)$ and for all $i \in \{1, ..., d\}$*

$$\sum_{\substack{m=1 \\ \theta^m \neq \theta_i}}^{n} \Phi(\theta) \lambda_i(x, \theta, F_i^m[\theta]) - \Phi(F_i^m[\theta]) \lambda_i(x, F_i^m[\theta], \theta) = \theta_i \Phi(\theta) \partial_i U(x) \tag{3.8}$$

If $U \in C^1(\mathbb{R})$ then the process has μ introduced in 3.7 invariant.

Proof of Proposition 3.1.7. For a function $f \in C_c^1(E)$ we write $\mathcal{L}f = \sum_{i=1}^{d} \mathcal{L}_i f$,

where

$$\mathcal{L}_i f(x,\theta) = \theta_i \partial_i f(x,\theta) + \sum_{\substack{m=1 \\ \theta^m \neq \theta_i}}^{n} \lambda_i(x,\theta,F_i^m\theta)[f(x,F_i^m\theta) - f(x,\theta)]$$

as in the proof of Proposition 3.1.5 and for $i \in \{1,...,d\}$ we write

$$\begin{aligned}
&Z\int_E \mathcal{L}_i f(x,\theta)\mu(dx,d\theta) = \sum_{\theta\in\Theta}\int_{\mathbb{R}^d} \mathcal{L}_i f(x,\theta)\Phi(\theta)\exp\{-U(x)\}dx = \\
&= \sum_{\theta\in\Theta}\Phi(\theta)\int_{\mathbb{R}^d} \theta_i\partial_i U(x) f(x,\theta)\exp\{-U(x)\}dx + \\
&+ \sum_{\theta\in\Theta}\Phi(\theta)\int_{\mathbb{R}^d} \sum_{\substack{m=1 \\ \theta^m\neq\theta_i}}^{n} \exp\{-U(x)\}\lambda_i(x,\theta,F_i^m[\theta])\left[f(x,F_i^m[\theta]) - f(x,\theta)\right]dx \\
&= \sum_{\theta\in\Theta}\int_{\mathbb{R}^d} \Phi(\theta)\theta_i\partial_i U(x) f(x,\theta)\exp\{-U(x)\}dx + \\
&+ \sum_{\theta\in\Theta}\int_{\mathbb{R}^d} \left(\sum_{\substack{m=1 \\ \theta^m\neq\theta_i}}^{n} \Phi(F_i^m[\theta])\lambda_i(x,F_i^m[\theta],\theta) - \Phi(\theta)\lambda_i(x,\theta,F_i^m[\theta])\right)\exp\{-U(x)\}f(x,\theta) = \\
&= 0
\end{aligned}$$

and we conclude as in the proof of Proposition 3.1.5. □

Finally, let us consider the more general case where we want to target a more general invariant measure of the form

$$\mu(dx,d\theta) = \frac{1}{Z}\exp\{-U(x,\theta)\}\mu_0(dx,d\theta) \tag{3.9}$$

for some function $U \in C^1(E)$.

Proposition 3.1.8. *Consider a multi-directional Zig-Zag process, with velocity space $\Theta = \Theta_0^d$ and the rates picked such that for all i, $\lambda_i \in C^1(E)$ and*

$$\begin{aligned}
&\sum_{\substack{m=1 \\ \theta^m\neq\theta_i}}^{n} \lambda_i(x,\theta,F_i^m[\theta])\exp\{-U(x,\theta)\} - \lambda_i(x,\theta,F_i^m[\theta])\exp\{-U(x,F_i^m[\theta])\} = \\
&= \theta_i\partial_i U(x,\theta_i)\exp\{-U(x,\theta)\}
\end{aligned} \tag{3.10}$$

for some function $U \in C^1(E)$. Then the process has μ introduced in (3.9) invariant.

Proof of Proposition 3.1.8. For some $f \in C_c^1(E)$ let as write $\mathcal{L}f = \sum_{i=1}^d \mathcal{L}_i f$ as

usual. Then,

$$
\begin{aligned}
&Z\int_E \mathcal{L}_i f(x,\theta)d\mu = \sum_{\theta\in\Theta}\int_{\mathbb{R}^d} \mathcal{L}_i f(x,\theta)\exp\{-U(x,\theta)\}dx = \\
&= \sum_{\theta\in\Theta}\int_{\mathbb{R}^d} \theta_i \partial_i f(x,\theta)\exp\{-U(x,\theta)\}dx + \\
&+ \sum_{\theta\in\Theta}\int_{\mathbb{R}^d} \sum_{\substack{m=1\\ \theta^m\neq\theta_i}}^{n} \lambda_i(x,\theta,F_i^m[\theta])(f(x,F_i^m[\theta]) - f(x,\theta))\exp\{-U(x,\theta)\}dx = \\
&= \sum_{\theta\in\Theta}\int_{\mathbb{R}^d} f(x,\theta)\theta_i(\partial_i U(x,\theta))\exp\{-U(x,\theta)\}dx - \\
&- \sum_{\theta\in\Theta}\int_{\mathbb{R}^d} f(x,\theta)\exp\{-U(x,\theta)\}\left(\sum_{\substack{m=1\\ \theta^m\neq\theta_i}}^{n} \lambda_i(x,\theta,F_i^m[\theta])\right)dx + \\
&+ \sum_{\theta\in\Theta}\int_{\mathbb{R}^d} \exp\{-U(x,\theta)\}\left(\sum_{\substack{m=1\\ \theta^m\neq\theta_i}}^{n} \lambda_i(x,\theta,F_i^m[\theta]\right) f(x,F_i^m[\theta]))dx = \\
&= \sum_{\theta\in\Theta}\int_{\mathbb{R}^d} f(x,\theta)\theta_i\partial_i U(x,\theta)\exp\{-U(x,\theta)\} - \\
&- \sum_{\theta\in\Theta}\int_{\mathbb{R}^d} f(x,\theta)\sum_{\substack{m=1\\ \theta^m\neq\theta_i}}^{n} (\lambda_i(x,\theta,F_i^m[\theta])\exp\{-U(x,\theta)\} - \lambda_i(x,F_i^m[\theta],\theta)\exp\{-U(x,F_i^m[\theta])\})\,dx \\
&= 0.
\end{aligned}
$$

We conclude as in the proof of Proposition 3.1.5. □

3.2 Multi-Directional Closest Neighbour Zig-Zag

As we have already seen in the previous section, there are many different rate functions we can choose in order to satisfy the system of equations (3.5) so that the MDZZ targets the distribution of interest. In this chapter we will focus on the following choice of algorithm.
Suppose that the allowed directions are

$$\Theta_0^d = \{\theta^1, \theta^1, ..., \theta^n\}^d$$

on $\mathbb{R}^d$, with $\theta^1 < ... < \theta^n$ such that

$$\sum_{m=1}^{n} \theta^m = 0.$$

We impose the following condition on the switches. Assuming that some coordinate of the direction is on θ^j and it has to switch, then it can only switch either to θ^{j+1} or to θ^{j-1}. In other words we only allow the directions to change in the smallest possible way, to a neighbouring direction. We call this **Multi-Directional Closest Neighbour Zig-Zag** (MDCNZZ). It turns out that we can find an explicit formula for the rates in this case.

Consider first the case where a velocity $\theta = (\theta_1, ..., \theta_d)$ satisfies $\theta_i = \theta^1$. The only neighbour of θ^1 is θ^2 and therefore for any $m \neq 2$ the process is not allowed to jump from θ to $F_i^m[\theta]$ or the other way around, i.e. $\lambda_i(x, \theta, F_i^m[\theta]) = \lambda_i(x, F_i^m[\theta], \theta) = 0$. Equation (3.5) then gives

$$\lambda_i(x, \theta, F_i^2[\theta]) - \lambda_i(x, F_i^2[\theta], \theta) = \theta^1 \partial_i U(x). \tag{3.11}$$

Now, we consider the case where the velocity θ satisfies $\theta_i = \theta^2$. This time the process is allowed to jump either to θ^1 or θ^3. Note here that in this case $F_i^1[\theta]$ can play the role of a velocity with first coordinate equal to θ^1 and from equation (3.11) we have

$$\lambda(x, F_i^1[\theta], \theta) - \lambda(x, \theta, F_i^1[\theta]) = \theta^1 \partial_i U(x). \tag{3.12}$$

Using equation (3.5) again, along with (3.12), we get

$$\begin{aligned}
&\big(\lambda_i(x, \theta, F_i^3[\theta]) - \lambda_i(x, F_i^3[\theta], \theta)\big) + \big(\lambda_i(x, \theta, F_i^1[\theta]) - \lambda_i(x, F_i^1[\theta], \theta)\big) = \theta^2 \partial_i U(x) \iff \\
&\iff \lambda_i(x, \theta, F_i^3[\theta]) - \lambda_i(x, F_i^3[\theta], \theta) - \theta^1 \partial_i U(x) = \theta^2 \partial_i U(x) \iff \\
&\iff \lambda_i(x, \theta, F_i^3[\theta]) - \lambda_i(x, F_i^3[\theta], \theta) = (\theta^1 + \theta^2) \partial_i U(x)
\end{aligned}$$

If we continue in the same way and using induction on the value of m when $\theta_i = \theta^m$, we get that if $\theta_i = \theta^m$ for some $m < n$ we have

$$\lambda_i(x, \theta, F_i^{m+1}[\theta]) - \lambda_i(x, F_i^{m+1}[\theta], \theta) = (\theta^1 + \theta^2 + ... + \theta^m) \partial_i U(x). \tag{3.13}$$

We conclude that if $\theta_i = \theta^m$ for $m < n$, then the two rates that appear in the last equation must satisfy

$$\lambda_i(x, \theta, F_i^{m+1}[\theta]) = \big[(\theta^1 + \theta^2 + ... + \theta^m) \partial_i U(x)\big]^+ + \gamma_i(x, \theta) \tag{3.14}$$

and

$$\lambda_i(x, F_i^{m+1}[\theta], \theta) = \left[(\theta^1 + \theta^2 + ... + \theta^m)\partial_i U(x)\right]^- + \gamma_i(x, \theta) \tag{3.15}$$

for some function γ. Furthermore note that, for any (x, θ), at least one of the quantities
$\left[(\theta^1 + \theta^2 + ... + \theta^m)\partial_i U(x)\right]^+$ or $\left[(\theta^1 + \theta^2 + ... + \theta^m)\partial_i U(x)\right]^-$ must be zero and since λ_i are hazard rates and therefore non-negative, the function γ has to be non-negative. For some $\theta_0 \in \{\theta^i, i = 1, .., n\}$ we introduce the notation

$$\theta_0^{\leq} = \sum_{\theta' \leq \theta_0} \theta', \ \theta_0^{<} = \sum_{\theta' < \theta_0} \theta'. \tag{3.16}$$

Also, for some $\theta \in \Theta$ with $\theta_i = \theta^m$, $m < n$ we write $F_i^+[\theta] \in \Theta$ to denote the velocity that assigns the i coordinate of θ the smallest possible, larger than the current one, value, i.e.

$$\begin{cases} (F_i^+[\theta])_i = \theta^{m+1} \\ (F_i^+[\theta])_j = \theta_j \text{ for } j \neq i \end{cases} \tag{3.17}$$

and similarly, for $\theta \in \Theta$ with $\theta_i = \theta^m$, $m > 1$ we write $F_i^-[\theta] \in \Theta$ to denote the velocity with

$$\begin{cases} (F_i^-[\theta])_i = \theta^{m-1} \\ (F_i^-[\theta])_j = \theta_j \text{ for } j \neq i \end{cases} \tag{3.18}$$

Then, we get from (3.14) and (3.15) the following.

Proposition 3.2.1 (Invariant Measure for MDCNZZ)**.** *Consider a d-dimensional MDCNZZ with velocity space $\{\theta^1, ...\theta^n\}^d$ and assume that the rates must satisfy*

$$\lambda_i^+(x, \theta) = \lambda_i(x, \theta, F_i^+[\theta]) = \left[\theta_i^{\leq} \ U'(x)\right]^+ + \gamma_i(x, \theta) \tag{3.19}$$

and

$$\lambda_i^-(x, \theta) = \lambda_i(x, \theta, F_i^-[\theta]) = \left[\theta_i^{<} \ U'(x)\right]^- + \gamma_i(x, F_i^-[\theta]). \tag{3.20}$$

for non-negative functions γ_i. Then the process has measure μ in (3.1) as invariant.

3.3 Ergodicity of Multi-Directional Closest Neighbour Zig-Zag

In this section we will identify conditions that ensure that the Multi-Directional Closest Neighbour Zig-Zag, as introduced in Section 3.2, is ergodic, meaning that

the law of the process converges to the law of the invariant distribution as time goes to infinity. For this section, we will restrict ourselves to the case where the process can move in directions from the space

$$\Theta = \Theta_0^d = \{-\theta^n < ... < -\theta^1 < \theta^1 < ... < \theta^n\}^d,$$

where we will assume that $\theta_1 \neq 0$. The state space is $E = \mathbb{R}^d \times \Theta_0^d$ and the process targets the probability measure $\mu(dx) = \frac{1}{Z}\exp\{-U(x)\}\mu_0(dx)$ with uniform marginal distribution in the space of directions.

We emphasise that in order to prove ergodicity of the process, we will assume that the process cannot have a velocity with 0 in any coordinate. We note, however, that in general the process can be defined while allowing $0 \in \Theta_0$.

In order to prove ergodicity of the process, we will have to make the two following assumptions. Both of them are typically satisfied in applications.

Assumption 3.3.1. *$U \in C^3$ and U has a non-degenerate local minimum, i.e. there exists an x_0 which is a local minimum for U and such that the Hessian matrix $Hess(U)(x_0)$ is strictly positive definite.*

Assumption 3.3.2 (Growth Condition)**.** *$U \in C^3$ and there exists $c > d$, $c' \in \mathbb{R}^d$ so that for all $x \in \mathbb{R}^d$*

$$U(x) \geq c \cdot log(1 + \|x\|) - c'. \tag{3.21}$$

The main goal of this section is to prove the following result.

Theorem 3.3.3 (Ergodicity of the MDCNZZ process)**.** *Assume that Assumptions 3.3.1 and 3.3.2 hold. Then the Multi-Directional Closer Neighbour Zig-Zag is ergodic, i.e. for any $(x, \theta) \in E$,*

$$\lim_{t\to\infty} ||P_{x,\theta}\left((X_t, \Theta_t) \in \cdot\right) - \mu(\cdot)||_{TV} = 0. \tag{3.22}$$

The proof closely follows the steps of the proof of the original Zig-Zag, as shown in [BRZ19] and assumes the validity of the result in the case of the original Zig-Zag to generalise it to the case where the velocities can take more values than $\{\pm 1\}^d$. The main observation is that since $\sum_{k=-n}^{n} \theta^k = 0$ (due to equation (3.6), where we will write $\theta^{-k} = -\theta^k$ for convenience of notation), we have that the quantities $\theta_0^{\leq}, \theta_0^{<}$ introduced in (3.16) of the previous section are all non-positive. Therefore, if we pick refresh rate functions $\gamma_i \equiv 0$ in (3.19) and (3.20), we see that whether the rates $\lambda_i^+(x, \theta)$ and $\lambda^-(x, \theta)$ will be strictly positive or zero depends completely on the sign of $\partial_i U(x)$. This sign is independent of the direction the process moves

towards and only depends on the position. We will use this property in the next section to prove a notion of irreducibility for the process in Section 3.3.1. We will then prove that the process will almost surely not drift away to infinity in Section 3.3.2. Finally, we will use these results to prove ergodicity of the process in section 3.3.3.

3.3.1 Reachability

The main challenge in the proof of ergodicity is to prove that the process is irreducible, meaning that it can explore the whole state. To get that, we first introduce a deterministic version of irreducibility which we call reachability.

More specifically, we define as **control sequence** an object $u = (t, \iota, s)$, where $t = (t_0, ..., t_m) \in (0, +\infty)^{m+1}$, $\iota = (i_1, ..., i_m) \in \{1, ..., d\}^m$, $s = (s_1, ..., s_m) \in \{-n, ..., n\}^m$ for some $m \in \mathbb{N}$. Starting from $(x, \theta) \in E$ a control sequence u gives rise to a Multi-Directional Zig-Zag trajectory (X_t, Θ_t) starting from any $(x, \theta) \in E$ as follows:

Start from (x, θ) and follow direction θ for t_0 time, i.e. set $X_t = x + t\theta, \Theta_t = \theta$ for $t \in [0, t_0)$. Then, switch the i_1'th component of θ to θ^{s_1}, i.e. switch θ to $F_{i_1}^{s_1}[\theta]$ and set $\Theta_{t_0} = F_{i_1}^{s_1}[\theta]$. Then, follow that direction for t_1 time, i.e. set $X_t = x + t_0\theta + (t - t_0)F_{i_1}^{s_1}[\theta], \Theta_t = F_{i_1}[\theta]$ for $t \in [t_0, t_0 + t_1)$. Then switch the i_2'th component of $F_{i_1}^{s_1}[\theta]$ to θ^{s_2}, i.e. set $\Theta_{t_0+t_1} = F_{i_1,i_2}^{s_1,s_2}[\theta] = F_{i_2}^{s_2}[F_{i_1}^{s_1}[\theta]]$. Continue similarly until time $t_0 + ... + t_m$. Write $\tau_k = \sum_{i=0}^{k-1} t_i$ for the time of the k'th switch, write $\tau_0 = 0$ and denote the final position $(X_{\tau_m}, \Theta_{\tau_m})$ of the path by $\Psi_u(x, \theta)$. The control sequence u (and the induced path) is called admissible from (x, θ) if path starting from (x, θ) satisfies $\lambda_{i_k}(X_{\tau_k}, \Theta_{\tau_{k-1}}, \Theta_{\tau_k}) > 0$ for all $k \in \{1, ..., m\}$, i.e. when the rates of switching at the switching points of the path are strictly positive. One can think of an admissible path as a path that has a "positive probability" of appearing as the process has probability zero to switch in neighbourhoods with zero intensity. To ease notation sometimes we will identify the control sequence u with the induced path.

Definition 3.3.4.

1. *We say that (y, η) is reachable from (x, θ) and write $(x, \theta) \to (y, \eta)$ if there exists an admissible path u from (x, θ) to (y, η), i.e. with $\Psi_u(x, \theta) = (y, \eta)$.*

2. *We say that a velocity η is reachable from θ and write $\theta \to \eta$ if for any $x \in \mathbb{R}^d$ there exists a $y \in \mathbb{R}^d$ such that $(x, \theta) \to (y, \eta)$.*

3. *We write* $(x, \theta) \not\rightarrow (y, \eta)$ *if there exists an admissible path from* (x, θ) *to* (y, η) *which changes the sign of every coordinate of* θ *along the way.*

Throughout this section we will impose Assumption 3.3.1, i.e. that $U \in C^3$ and U has a non-degenerate local minimum. We will, further, impose the following assumption, which is a relaxed version of Assumption 3.3.2.

Assumption 3.3.5 (Growth Condition 1)**.** $\lim_{|x|\to\infty} U(x) = \infty$.

The main goal of this section is to prove the following theorem.

Theorem 3.3.6. *Assume that Assumptions 3.3.1 and 3.3.5 hold. Then, for all* $x, y \in \mathbb{R}^d$ *,* $\theta, \eta \in \Theta$*,* $(x, \theta) \not\rightarrow (y, \eta)$*.*

The rest of this section is devoted to the proof of Theorem 3.3.6, which is similar to the proof of the original Zig-Zag in [BRZ19]. We begin with some preliminary results.

First of all, we note that we can assume without loss of generality that all the refresh functions γ_i introduced in (3.19) and (3.20) are equal to zero. This is because if the rates of a path at the switching points are strictly positive when $\gamma_i \equiv 0$, the same will hold if we add a non-negative quantity γ_i. Therefore, for the rest of this section we can assume that all $\gamma_i \equiv 0$, therefore the upwards and downwards rates of the process satisfy

$$\lambda_i^+(x, \theta) = \left[\theta_i^{\leq} \partial_i U(x)\right]^+ \tag{3.23}$$

and

$$\lambda_i^-(x, \theta) = \left[-\theta_i^{<} \partial_i U(x)\right]^+ = \left[\theta_i^{<} \partial_i U(x)\right]^- . \tag{3.24}$$

Note once again that if $\theta_i < \theta^n$ then $\theta_i^{\leq} < 0$ and therefore $\lambda_i^+(x, \theta) > 0$ if and only if $\partial_i U(x) < 0$. This means that assuming there is a larger than the current value for the i coordinate, the process has a positive rate of jumping upwards as long as $\partial_i U(x) < 0$. The same holds for downwards jumps and if $\theta_i > -\theta_n$ we have $\theta_i^{<} < 0$ and therefore $\lambda_i^-(x, \theta) > 0$ if and only if $\partial_i U(x) > 0$. See Figure 3.1 for an illustration of the process' admissible moves. Using this fact, we can observe a property of admissible paths that if one reverses time and traverses them backwards, they remain admissible. We emphasise that this is only true when the refresh rates are $\gamma_i \equiv 0$, which we have already assumed for this section, without loss of generality.

Proposition 3.3.7 (Time Reversal of Admissible paths)**.** *Let* $u = (t, \iota, s)$ *an admissible control sequence from* (x, θ) *to* (y, η)*. Then the same path, traversed with time reversed, i.e. from* $(y, -\eta)$ *to* $(x, -\theta)$ *is also admissible.*

Proof of Proposition 3.3.7. Let τ be a turning time and $x \in \mathbb{R}^d$ the turning position of the path traversed forward or backwards in time. Suppose that the switch in velocity happened forward in time from $\theta = (\theta_1, ..., \theta_d)$ to $F_i^j[\theta]$ in which case the change in velocity backwards in time happened from $-F_i^j[\theta]$ to $-\theta$. Assume that $\theta_i < \theta^j$, i.e. the i coordinate of the velocity increased. Since the path is admissible, the rate of this upwards jump in the turning point is positive and from equation (3.23) we get that $\partial_i U(x) < 0$. Since $-\theta^j < -\theta_i$ the jump backwards in time is also upwards and since $\partial_i U(x) < 0$ we get from (3.23) that the switching rate is positive. A similar argument for downwards jump can be applied in the case where $\theta^j < \theta_i$, using equation (3.24) for the downwards rates and we conclude that the path backwards in time is also admissible. □

The rest of the section is organised as follows. We first prove the result in the case where the invariant measure is a **Gaussian**. Then, under the assumption that U admits a non-degenerate local minimum, we prove that all the points **locally around this minimum** can reach each other. Furthermore, we prove that no matter where the process starts from, it can **switch all the coordinates** of the velocity through an admissible path. After that, we prove the **Continuous Component Lemma**, Lemma 3.3.21, which connects the notion of reachability with the notion of irreducibility and essentially proves that if the process can reach a point, then it has a positive probability to reach a neighbourhood of that point. Finally, we combine all these results to study the structure of sets of points that reach each other and prove that we can **extend the local reachability to a global statement**. All these statements are dealt with in the next five subsections.

Gaussian Case

We first consider the case where our target distribution is Gaussian, and more specifically we assume that for all $x \in \mathbb{R}^d$, $U(x) = \langle x\,, Ax \rangle$ for some positive definite, symmetric matrix A. The main goal of this section is to prove the reachability result in this specific case.

Proposition 3.3.8 (Reachability in the Gaussian case)**.** *Assume that $U(x) = \langle x\,, Ax \rangle$ for some positive definite matrix A. For any $x, y \in \mathbb{R}^d$, $\theta, \eta \in \Theta$ we have $(x, \theta) \leadsto (y, \eta)$.*

Our first step in order to prove this proposition will be to ensure that from any starting velocity we can reach any other velocity in the space. We begin with a definition to simplify notation.

Definition 3.3.9. *We will say that for a component $\theta_0 \in \{\theta^1, .., \theta^n\}$ we say that η_0 is* **opposite** *of θ_0 if $\eta_0 < \theta_0$. For a $\theta_0 \in \{-\theta^n, .., -\theta^1\}$ say η_0 is* **opposite** *of θ_0 if $\theta_0 < \eta_0$.*
A velocity η is opposite of θ in coordinate k if the k component of η is opposite of the k component of θ. A velocity η is opposite of θ if it is opposite in every coordinate $k \in \{1, ..., d\}$.

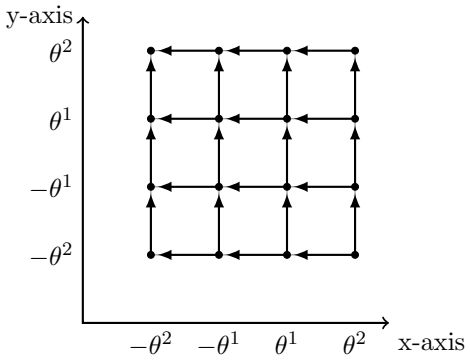
Heuristically η_0 is opposite of θ_0 if in order to move from θ_0 to η_0 via a series of jumps to the closer neighbour, we need to first move to values that are closer to zero.
The proof will rely on the notion of Asymptotic Flippability. Roughly speaking, a velocity is Asymptotically Flippable if the process that starts from this direction can eventually jump to any opposite velocity.

Definition 3.3.10. *For a velocity θ, the coordinate k is Asymptotically Flippable (AF) if $\theta_k(A\theta)_k > 0$. A velocity is called Asymptotically Flippable if all d coordinates are Asymptotically Flippable for this velocity.*

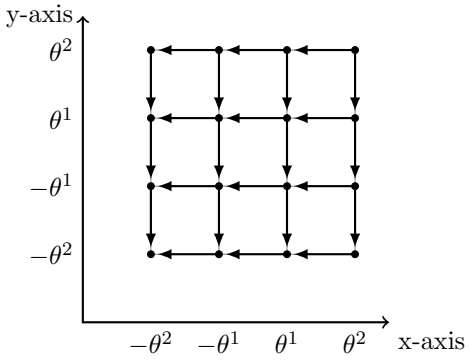
Note that in this Gaussian setting we have $\nabla U(x) = Ax$ and therefore for any (x, θ) such that $\theta_k < \theta^n$ we have $\lambda_k^+(x, \theta) > 0$ iff $(Ax)_k < 0$. Also, for any (x, θ) such that $-\theta^n < \theta_k$, $\lambda_k^-(x, \theta) > 0$ iff $(Ax)_k > 0$. Directly from the definition of AF we get the following.

Lemma 3.3.11. *Suppose k is AF for θ and η agrees with θ in all coordinates except for the k'th one and suppose further that η_k is opposite of θ_k. Then starting from θ and spending a large time t on it, the process can have an admissible path that ends in velocity η.*

Proof. Assume that $\theta = (\theta_1, ..., \theta_d)$ and that $\theta_k > 0$, so $\eta_k < \theta_k$. In this case, θ being AF in k coordinate means that $(A\theta)_k > 0$.
Now, assume that the process starts at (x, θ), for some $x \in \mathbb{R}^d$ does not switch direction for some large time t. After time t it will be at point $(x + t\theta, \theta)$ and $\lambda_k^-(x + t\theta, \theta) > 0 \iff [A(x + t\theta)]_k > 0$. Note the sign of $A(x + t\theta) = Ax + tA\theta$ is the sign of $A\theta$ for large t. Therefore the process is allowed to jump down by one in the k coordinate. Now, since the derivative of U is continuous, $\partial_k U$ will remain negative in a small neighbourhood of $x + t\theta$. Therefore, we can continue to decrease the k component of the velocity of the process until we reach η_k, while the space position remains in this small neighbourhood, thus making those changes admissible.
A similar argument treats the case when $\theta_k < 0$. □



(a) $\partial_1 U(x) > 0, \partial_2 U(x) < 0$



(b) $\partial_1 U(x) > 0, \partial_2 U(x) > 0$

Figure 3.1: *Two pictures that illustrate all the available jumps from one velocity to the other for two different values of $x \in \mathbb{R}^d$ for a MDCNZZ process. Every node in this 4×4 grid denotes a velocity. The switching rates are given by (3.23) and (3.23). An arrow denotes an admissible jump from one velocity to the other for a specific value of x. When $\partial_1 U(x) > 0$ only down-steps are admissible and when $\partial_1 U(x) < 0$ only up-steps. The same goes for the second coordinate. Since $\partial_1 U$ is continuous, in a small neighbourhood of x its sign does not change so the same type of jumps are allowed in this neighbourhood of x. By successively jumping from one velocity to some neighbouring one with a very small time period between the jumps the process can admissibly go from any starting node to any node that is connected via a sequence of arrows.*

By repeatedly using the above lemma in every coordinate we get the following.

Lemma 3.3.12. *If θ is AF for any coordinate in the index set I and η is opposite of θ in the k-coordinate for every $k \in I$ and agrees with θ on any coordinate outside I, then $\theta \to \eta$. That is, for any $x \in \mathbb{R}^d$ there exists a $y \in \mathbb{R}^d$ so that $(x, \theta) \to (y, \eta)$.*

Proof. Suppose the process starts from (x, θ). If we pick t large enough and let the process move in direction θ for time t, there will be a neighbourhood of $x + t\theta$ in which the rates of successive jumps from θ to η will be allowed in every coordinate $k \in I$. By applying all these successive jumps during a small time period we can ensure that all the jump points will be inside this neighbourhood and therefore the path will be admissible. □

The importance of this lemma is that it guarantees that once the process has reached an AF velocity θ, it can then reach any other velocity that is opposite of θ. The following Lemma guarantees that the process can use AF velocities.

Lemma 3.3.13. *For any direction θ there exists an AF direction η so that $\theta \to \eta$.*

The definition of the AF in the multi-directional Zig-Zag is tailored so that the proof of the equivalent lemma used in the original Zig-Zag, carries over here. We include it here for the sake of completeness. Before we give the proof of this lemma, we introduce the following notation. Given a velocity $\theta \in \Theta$, and an index set $I \subset\in \{1, .., d\}$, we write $G_I(\theta) \in \Theta$ for the velocity that satisfies, for all $i \in I$

$$\begin{cases} G_I(\theta)_i = -\theta_i \text{ if } i \in I \\ G_I(\theta)_i = \theta_i \text{ if } i \notin I. \end{cases}$$

Proof of Lemma 3.3.13. We first use the LL^T representation of A to find vectors $u_1, ..., u_d$ so that $A_{ij} = \langle u_i, u_j \rangle$ for all i, j, where $\langle \cdot, \cdot \rangle$ is the usual dot product in $\mathbb{R}^d$. for all i, j. Note that since A is positive definite, these vectors are linearly independent.
The idea is to create a process that, given a starting velocity, it targets a new velocity, each time reachable from the previous one and the only way for that process to stop is for an Asymptotic Flippable velocity to have been reached. We then ensure that this process has to stop after visiting a finite number of velocities and the proof will be complete.
Assume that we start from a velocity $\theta = (\theta_i)_{i=1}^d \in \Theta$. Consider the vector

$$u(\theta) = \sum_{i=1}^{d} \theta_i u_i.$$

We observe, that for θ to be AF in the i coordinate we must have $\theta_i(A\theta)_i > 0 \iff \langle \theta_i u_i, u(\theta)\rangle > 0$.
Assume that θ is not Asymptotic Flippable, so if I is the index set of coordinates on which θ is Asymptotic Flippable then $I \neq \{1, ..., d\}$. Consider the flipped velocity $G_I(\theta)$. Since θ is AF in every coordinate $i \in I$, we get from Lemma 3.3.12 that $\theta \to G_I(\theta)$.
Let $u_+ = \sum_{i\in I} u_i\theta_i$ and $u_- = \sum_{\notin I} u_i\theta_i$. Then, $u(\theta) = u_+ + u_-$ and $u(G_I(\theta)) = u_- - u_+$. Then, from the polarization equality, we get

$$\|u(G_I(\theta))\|^2 = \|u(\theta)\|^2 - 4\langle u_+, u_-\rangle.$$

Note that $\langle u_+, u_-\rangle = \langle u(\theta), u_-\rangle - ||u_-||^2 = \sum_{i\notin I}\langle u(\theta), \theta_i u_i\rangle - ||u_-||^2$. From the definition of I, every part of the sum is non-positive. Since θ is not AF, $I^c \neq \emptyset$ and since all $\theta_i \neq 0$ and u_i linearly independent, we get $u_- \neq 0$. Therefore, $\langle u_+, u_-\rangle < 0$. This means that if θ is not AF then

$$\|u(G_I(\theta))\| > \|u(\theta)\|.$$

We then consider a sequence of velocities (by applying the operator G_I each time) that strictly increases the norm of u and since the velocities are finitely many, this sequence of velocities will have to stop. However, since for every non AF velocity we can continue the sequence by applying the G operator and increase the norm, we conclude that the last velocity of the sequence is AF. Since each velocity in the sequence can be reached by the previous velocity, we get that the the first velocity will reach the AF velocity. □

In the proof of the original Zig-Zag, where the direction space is $\{-1, +1\}^d$, θ being AF means that after the process spends a long time moving in direction θ, the process can ultimately jump to any other direction, since any two directions are opposite to each other. This type of irreducibility in the space of directions is used in the proof of reachability, but it is not immediate here. However, we can prove it using the reachability of the original case, combined with Lemma 3.3.13.

Proposition 3.3.14. *For any directions θ and η, we have $\theta \to \eta$.*

Before we prove this proposition we will need the following which concentrates on the behaviour of points that can be seen as in the setting of the original Zig-Zag.

Lemma 3.3.15. *Let $\theta^k \in \{-\theta^n, .., -\theta^1, \theta^1, .., \theta^n\}$ and consider the hypercube $H_k = \{-\theta^k, \theta^k\}^d$.*

Then, if the process starts from a velocity $\theta \in H_k$, it can reach an AF velocity $\eta \in H_k$.

Proof of Lemma 3.3.15. Suppose we start from point (x, θ). Let $Z_t = (X_t, \Theta_t)$ be our multi-direction Zig-Zag process with velocity space $\{-\theta^n, .., -\theta^1, \theta^1, .., \theta^n\}^d$ and let $Z'_t = (X'_t, \Theta'_t)$ be different multi-directional Zig-Zag targeting the same Gaussian distribution, but with velocity space H_k. We can use Lemma 3.3.13 in the case where the velocity space is $\Theta = H_k$, which means that there exists a velocity $\eta \in H_k$ that is AF, i.e. $\eta_i(A\eta)_i > 0$ for all i and, furthermore, there exists an admissible path of the Z' process, starting from (x, θ) and ending to (y, η) for some $y \in \mathbb{R}^d$. The fact that $\eta_i(A\eta)_i > 0$ for all i means that η is AF also in the setting of the multi-dimensional Zig-Zag process Z_t with velocity space $\{-\theta^n, ..., \theta^n\}^d$. We will now prove that the existence of such admissible path for the process (Z'_t) implies the existence of an admissible path $(x, \theta) \to (z, \eta)$ for some $z \in \mathbb{R}^d$ for the process (Z_t) and this will conclude the proof.

Assume for simplicity and without loss of generality that $x = 0$, $\theta = (\theta^k, ..., \theta^k)$ and $\eta = -\theta$. There exist times $t_1, .., t_d$ so that at time t_i , (Z'_t) can admissibly switch its direction on the i coordinate from θ^k to $-\theta^k$. For simplicity assume without loss of generality that $t_1 < ... < t_d$. Suppose that the switch happens for the i coordinate at point $y \in \mathbb{R}^d$. Then, $\theta^k(Ay)_i > 0 \iff (Ay)_i > 0$. Then, by continuity, there exists an $\epsilon > 0$ so that for all i and for all $y \in B(y_i, \epsilon)$, we have $(Ay_i)_i > 0$. In the setting of the multi-dimensional Zig-Zag $(Z_t)_{t\geq 0}$, this means that at point y_1, the process is allowed to switch the first coordinate from θ^k to θ^{k-1}. Then, for a very small δ_1 (chosen appropriately after all y_i and ϵ are fixed) Z_t can admissibly change its first coordinate every δ_1 time-steps to the next smaller allowed number (i.e. from θ^{k-1} to θ^{k-2} etc.), until it reaches $-\theta^k$. This is because if δ_1 is small enough, all these switches happen inside the ball $B(y_1, \epsilon)$. Then, the process travels in direction $(-\theta^k, \theta^k, ..., \theta^k)$ until it enters the ball $B(y_2, \epsilon)$. This will happen, since $y_2 - y_1$ and $(-\theta^k, \theta^k, .., \theta^k)$ are collinear and the process starts moving in direction $(-\theta^k, \theta^k, .., \theta^k)$ from inside the ball $B(y_1, \epsilon)$. While inside $B(y_2, \epsilon)$, (Z'_t) is allowed to admissibly switch its second coordinate to $-\theta^k$, so the Multi-Directional Zig-Zag is allowed to do the same by using the same argument as in the first coordinate. This is done by having the successful switches happen every δ_2 time periods for some δ_2 chosen appropriately small so that the process remains inside the ball $B(y_2, \epsilon)$ throughout these switches. Then, it travels in direction $(-\theta^k, -\theta^k, \theta^k, ..., \theta^k)$ until it enters the ball $B(y_3, \epsilon)$, which will happen for the same reason as in the second coordinate. We can, then, inductively construct an admissible path $(x, \theta) \to (z, \eta)$ for some $z \in B(y, \epsilon)$. □

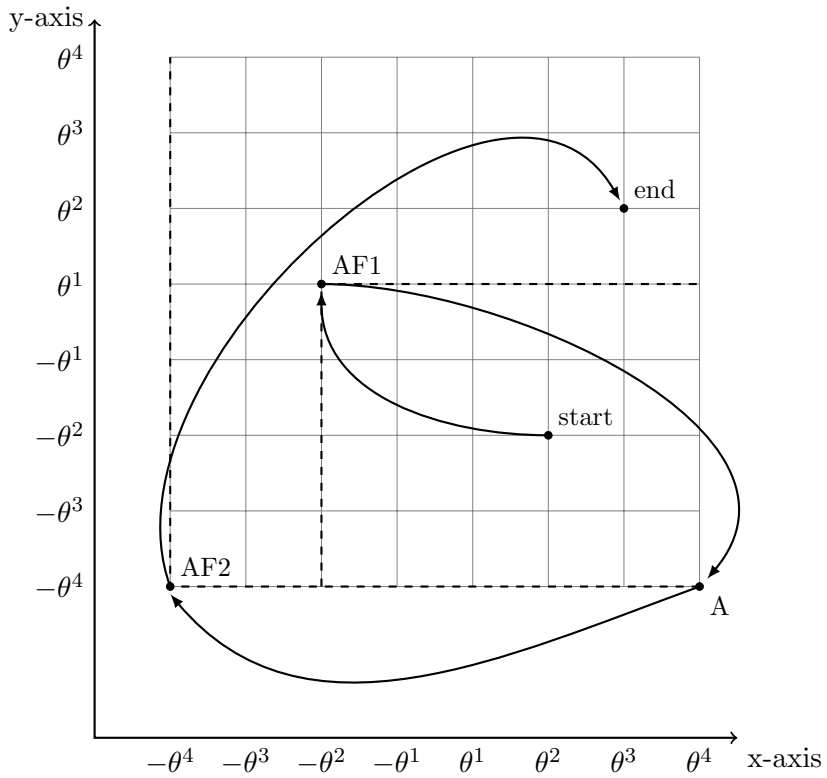


Figure 3.2: *A description of the proof of Proposition 3.3.14. The grid contains all the* 8×8 *velocities of the state space and the arrows signify admissible moves from one velocity to the other via a sequence of local neighbours. AF1 and AF2 are Asymptotic Flippable velocities and the spaces enclosed by the doted lines signify all the velocities that are reachable from the Asymptotic Flippable ones after the process moves with the Asymptotic Flippable velocity for a large period of time.*

We can now prove that any velocity can be reached. For an illustration of the proof, see Figure 3.2.

Proof of Proposition 3.3.14. From Lemma 3.3.13 we can find an AF velocity $AF1$ so that $\theta \to AF1$. Consider the velocity $A \in H_n$ with signs exactly the opposite of the signs of the coordinates of $AF1$. Then, since $AF1$ is AF and A is opposite of $AF1$, we get $AF1 \to A$. Now, we can use Lemma 3.3.15 to find an AF velocity $AF2 \in H_n$ and furthermore, $A \to AF2$. Finally, since $AF2 \in H_n$, η is opposite of $AF2$ and since $AF2$ is AF we have $AF2 \to \eta$. □

The result that will allow us to extend the reachability of velocities result to a reachability of both position and velocities is the following.

Lemma 3.3.16. *If $\theta \in \{-\theta_1, \theta_1\}^d$ is AF, then for any $x, y \in \mathbb{R}^d$ we have $(x, \theta) \to (y, -\theta)$.*

Proof of Lemma 3.3.16. Let $x, y \in \mathbb{R}^d$ and $\theta \in \{-\theta_1, \theta_1\}^d$. It is known from [BRZ19] that in the context of the original Zig-Zag targeting the same Gaussian distribution, if $\eta = \frac{1}{\theta^1}\theta \in \{-1, 1\}^d$ satisfies $\eta_k(A\eta)_k > 0$ for all $k \in \{1, ..., d\}$ then there exists an admissible path $(X_t, H_t)_{t\geq 0}$ from (x, η) to $(y, -\eta)$. The rates of this Zig-Zag process are $\lambda_k^{ZZ}(z, \eta) = [\eta_k(Az)_k]^+$ so if $\tau_1, \tau_2, ...$ the switching points of the path and $i_1, i_2, ...$ the coordinates switched during these switches, we have $\lambda_{i_m}^{ZZ}((X_t)_{\tau_m}, H_{\tau_m}) = [(H_{\tau_m})_{i_k}(AX_{\tau_m})_{i_m}]^+ > 0$ for all $m = 1, 2, ...$ Now, consider the path $(Y_t, \Theta_t) = (X_{\theta^1 t}, \theta^1 H_{\theta^1 t})_{t\geq 0}$. This is the same path as before but we traverse it with speed θ^1 so it is a MDCN Zig-Zag path with velocity space $\{-\theta^n, ..., \theta^n\}^d$, but for this specific path the velocities are taken from the subset $\{-\theta^1, \theta^1\}^d$. The switching times of the path are $\frac{1}{\theta^1}\tau_1, \frac{1}{\theta^1}\tau_2, ...,$ while the coordinates switched are still $i_1, i_2, ...$ Suppose that during the first switch, the i_1 coordinate of H_t changes from $+1$ to -1 so that the i_1 coordinate of Θ_t changes from θ^1 to $-\theta^1$. Then $(AX_{\tau_1})_{i_1}]^+ > 0$ so $(AY_{\tau_1})_{i_1}]^+ > 0$ which means that $\lambda^-(Y_{\tau_1}, \theta) > 0$. So the first switch which changed the i_1 coordinate from θ^1 to $-\theta^1$ is admissible and using the symmetric argument we see that the same holds if the change was from $-\theta^1$ to θ^1. The same holds for all the other switches and therefore the entire path is admissible in the context of MDCN Zig-Zag. □

We can now, finish the proof of Proposition 3.3.8, using a similar argument as in [BRZ19].

Proof of 3.3.8. Let $(x, \theta), (y, \eta)$ two points of the space. From Lemma 3.3.15 there exists
$\theta' \in \{-\theta_1, \theta_1\}^d$ AF and from Proposition 3.3.14 we can find a $z_1 \in \mathbb{R}^d$ with $(x, \theta) \to (z, \theta')$ and a $z_2 \in \mathbb{R}^d$ so that $(y, -\eta) \to (z_2, \theta')$. Using the time reversal property of admissible paths (Proposition 3.3.7) we get that $(z_2, -\theta') \to (y, \eta)$. Finally, from Lemma 3.3.16, $(z_1, \theta') \to (z_2, -\theta')$. Therefore,

$$(x, \theta) \to (z_1, \theta') \to (z_2, -\theta') \to (y, \eta).$$

and since this means that $(x, \theta) \to (y, \eta) \to (x, \theta) \to (y, \eta)$ which means that we can find an admissible path from (x, θ) to (y, η) by switching the sign of all coordinates of velocity, i.e.

$$(x, \theta) \rightsquigarrow (y, \eta)$$

which proves the result. □

Local Reachability

We will now consider more general target measures and extend the reachability result locally around a local minimum of their minus log-likelihood function U. The idea is that around the local minimum, U looks like the minus log-likelihood of a Gaussian and we can then conclude using the result of the previous section. The main goal of this paragraph is to prove the following.

Proposition 3.3.17 (Local Reachability)**.** *Suppose $U \in C^3$ and has a non-degenerate local minimum, meaning that at some point x_0, $\nabla U(x_0) = 0$ and $V = Hess_U(x_0)$ is positive definite. Then, there exists a $\gamma > 0$ so that for every x, y with $\|x - x_0\| < \gamma, \|y - x_0\| < \gamma$ and any velocities $\theta, \eta \in \Theta$, $(x, \theta) \rightsquigarrow (y, \eta)$.*

We begin with the following lemma which will allow us to target a neighbourhood of points, given that we target a point via a fully flipped path. This will also be used a lot in subsequent sections.

Lemma 3.3.18. *Assume that $U \in C^1$. If $(x, \theta) \rightsquigarrow (y, \eta)$ then there is an open neighbourhood O of (y, η) such that, if $(y', \eta') \in O$, $(x, \theta) \rightsquigarrow (y', \eta')$.*

Proof. Suppose that $y = x + t_0\theta + t_1 F_{i_1}^{s_1}\theta + ... + t_m F_{i_1,..,i_m}^{s_1,..,s_m}\theta$ and that the path A induced by the control sequence $u = (t, \iota, v)$, $t = (t_0, ..., t_m)$, $\iota = (i_1, ..., i_m)$, $v = (v_1, ..., v_m)$ is admissible and changes the sign of all coordinates of θ. Consider the linear function $\Phi : (s_0, .., s_m) \to s_0\theta + s_1 F_{i_1}^{u_1}\theta + ... + s_m F_{i_1,..,i_m}^{v_1,..,v_m}\theta$.
The matrix of the function is written in columns as $(\theta, F_{i_1}^{v_1}\theta, .., F_{i_1,..,i_m}^{v_1,..v_m}\theta)$. For the function to be onto, we need the columns to span $\mathbb{R}^d$. Note that the difference between two succeeding columns is of the form $\pm(\theta^k - \theta^{k-1})e_i$ for some $k \in \{-n + 1, ..., n\}$, $i \in \{1, .., d\}$. Since all the coordinates have changed at least once, all the e_k's appear in this set and therefore the differences of columns span the space. So the columns span the space as well, therefore Φ is onto. Consequently, the images of open balls through Φ are open in $\mathbb{R}^d$. This means that there exists a small neighbourhood of $\Phi(t_0, .., t_m)$ that is the image of small perturbations of $(t_0, .., t_m)$ through Φ. Therefore, by translation, we can find a small neighbourhood around $y = x + \Phi(t_0, ..., t_m)$ for which all the points can be targeted by a path from x that is induced by the control sequence $u' = (t', \iota, v)$ for some small perturbation t' of the times $t = (t_0, .., t_m)$. Since the rates are continuous functions and the path A is admissible, these small perturbations won't change the admissibility of the new path. □

We can now prove the main result of this section.

Proof of Proposition 3.3.17. Let us assume for simplicity that the local minimum is at $x_0 = 0$. Recall that μ is the invariant measure of the process with minus log-likelihood U. Consider the Gaussian measure μ^V, whose density is proportional to $\exp\{-1/2x^T Vx\}$, where $V = Hess_U(x_0)$. Let $(Z_t^V)_t$ be the Multi-directions Zig-Zag process with minimal rates λ^V that targets μ^V. We write $(x,\theta) \looparrowright^V (y,\eta)$ to denote that the path connecting (x,θ) and (y,η) is admissible under the rates λ^V.

Consider a control sequence $(t,i,s) = (t_0,..,t_p;i_1,..,i_p;s_1,..,s_p)$ starting from $(0,\theta)$, and ending at $(x(\tau_{p+1}),\theta(\tau_{p+1}))$, where $\tau_k = \sum_{i=0}^{k-1} t_i$. Assume that the control sequence's switching points are $(x(\tau_i),\theta(\tau_k))_{k=1}^p$. We define

$$\lambda^V_{min}(t,i,s) = \min_{j=1,..,p} \lambda^V_{i_j}(x(\tau_j),\theta(\tau_j)) \ , \ r_{max}(t,i,s) = \max_{j=0,..,p+1} |x(\tau_j)|$$

to be the minimum switching rate and the maximum distance from the origin of the path.
For $m \in \mathbb{N}$, $\theta,\eta \in \Theta$, let's define

$$\begin{aligned} U_{m,\theta,\eta} =& \{y \in \mathbb{R}^d, |y| < 2, (0,\theta) \looparrowright^V (y,\eta) \text{ via a control sequence } (t,i,s) \text{ with} \\ & \lambda^V_{min}(t,i,s) > 1/m, r_{max}(t,i,s) < m\}. \end{aligned}$$

If $y \in U_{m,\theta,\eta}$, from Lemma 3.3.18 we can find a neighbourhood around y so that for any y' in the neighbourhood, $(0,\theta) \looparrowright^V (y',\eta)$ via a control sequence (t',ι,s) for t' a small perturbation of t. By making the neighbourhood small enough and given the continuity of the rates, we can have $\lambda^V_{min}(t',i,s) > 1/m$ and $r_{max}(t',i,s) < m$. This means that $(y',\eta) \in U_{m,\theta,\eta}$ and therefore $U_{m,\theta,\eta}$ is open for all $m \in \mathbb{N}$. Now, from the reachability of Gaussian case, for any $\theta,\eta \in \Theta$, we have $(U_{m,\theta,\eta})_{m\in\mathbb{N}}$ is an open cover of the unit disc $D = \{y \in \mathbb{R}^d, |y| \leq 1\}$. Since this is a compact set and since $U_{m,\theta,\eta}$ are increasing in m, we can find $N_{\theta,\eta} \in \mathbb{N}$ so that $D \subset U_{N_{\theta,\eta},\theta,\eta}$. By fixing an $N > \max_{\theta,\eta\in\Theta} N_{\theta,\eta}$ we have $D \subset U_{N,\theta,\eta}$ for all $\theta,\eta \in \Theta$. Note as well that we can take N to be as large as we want and the last inclusion will still be true.

Overall, for any $\theta,\eta \in \Theta$, $|x| < 1$, we have $(0,\theta) \looparrowright^V (x,\eta)$ through a trajectory with minimum rate $\lambda^V_{min} > 1/N$ and maximum distance from 0 $r_{max} < N$. Note as well, that using a Taylor expansion on the unit ball, we can find $c > 1$ so that

$$\|\nabla U(x) - Vx\| \leq c|x|^2, |x| \leq 1 \tag{3.25}$$

Now, fix $\theta,\eta \in \Theta$, set $\gamma = \dfrac{1}{cN^4}$ and assume that $|y| < \gamma$. Set $z = y/\gamma$ so that $|z| < 1$. Then, we can find an admissible control (t,i,s) so that $(0,\theta) \looparrowright^V (z,\eta)$, with $\lambda^V_{min} > 1/N$ and $r_{max} < N$. If we rescale the times of this control sequence

by γ and consider the control (t', i, s), $t' = \gamma t$, this will induce a path from $(0, \theta)$ that targets (y, η). Due to the fact that the rates λ^V target a Gaussian structure, they treat such a rescale in a linear way. Therefore $\lambda^V_{min}(t', i, s) > \frac{\gamma}{N} = \frac{1}{cN^5}$, $(0, \theta) \overset{V}{\rightarrowtail} (y, \eta)$ and the trajectory lies in a ball of radius $\gamma N = \frac{1}{2cN^3} < 1$, as long as we pick N large enough. This means that we can apply (3.25) in all switching points of the control (t', i, s), i.e. if (x_0, θ_0) is a switching point of the trajectory, then

$$\|\nabla U(x_0) - Vx_0\| \leq c|x_0|^2 < c\gamma^2 N^2 = \frac{1}{4cN^6}.$$

To finish the proof, we will prove that the trajectory (t', i, s) is also admissible under the rates of U. Write $b = \sum_{j=1}^n \theta^j$ so

$$\begin{aligned}
\lambda_i^+(x_0, \theta_0) = \left[\theta_i^{\leq} \partial_i U(x_0)\right]^+ &= \left[\theta_i^{\leq}(Vx_0)_i - \theta_i^{\leq}((Vx_0)_i - \partial_i U(x_0))\right]^+ \geq \\
&\geq (\theta_i^{\leq}(Vx_0)_i)^+ - (\theta_i^{\leq}((Vx_0)_i - \partial_i U(x_0)))^+ \geq \\
&\geq \lambda_i^{+,V}(x_0, \theta_0) - |\theta_i^{\leq}||(Vx_0)_i - \partial_i U(x_0)| \geq \lambda_i^{+,V}(x_0, \theta_0) - b\|Vx_0 - \nabla U(x_0)\| \geq \\
&\geq \frac{\gamma}{N} - bc|x_0|^2 = \frac{1}{cN^5} - b\frac{1}{4cN^6}
\end{aligned}$$

and the last quantity is positive for N appropriately large. The same calculations are true for any other turning point of the path. This proves that $(0, \theta) \overset{}{\rightarrowtail} (y, \eta)$ for all $|y| < \gamma$ and all $\theta, \eta \in \Theta$. By time reversibility of admissible paths, we also have $(y, \eta) \overset{}{\rightarrowtail} (0, \theta)$ and this proves Lemma 3.3.17. □

Flippability

So far we have managed to prove that any two points close to the local minimum can reach each other. In order to extend this result to a non-local statement, we will later study the structure of sets of points that can reach each other. In this direction, a useful property for the process is the full flippability.

Definition 3.3.19. *The process is called* **fully flippable** *if for any (x, θ) there exists a (y, η) so that $(x, \theta) \overset{}{\rightarrowtail} (y, \eta)$, i.e. there exists an admissible path starting from (x, θ) that switches the sign of all coordinates of the velocity.*

The importance of the full flippability property can be understood by reviewing Lemma 3.3.18, which guarantees that a fully flippable process, independently of the starting position, can always reach an open ball of points. In this paragraph we prove the following.

Proposition 3.3.20 (Full Flippability)**.** *Assume that $U \in C^1$ and $\lim_{x\to\infty} U(x) = +\infty$. Then the MDCNZZ process is fully flippable.*

Proof of Proposition 3.3.20. Suppose otherwise, so there exists a point (x, θ) so that any admissible path from that point leaves the sign of at least one coordinate unchanged. Let's assume for now the following statement, which we prove later in the proof.

Statement: *For all $\epsilon > 0$, for all $T > 0$ and for all $\delta > 0$, there exists an admissible trajectory $(x_t, \theta_t)_{0\le t\le T}$ of length T, starting from (x, θ) so that for all $i \in \{1, ..., d\}$ and all $t \in [\delta, T]$*

$$(\theta_t)_i \ \partial_i U(x_t) < \epsilon. \tag{3.26}$$

If we have that, then let $\epsilon, \delta > 0$ and $T > 0$ and let us construct an admissible trajectory of length T from (x, θ) such that (3.26) holds. This means that for all $t \in [\delta, T]$

$$\sum_{i=1}^{d} (\theta_t)_i \ \partial_i U(x_t) < d\epsilon. \tag{3.27}$$

If we integrate (3.27) over time in $[\delta, T]$, we get

$$U(x_T) - U(x_\delta) \le \epsilon d(T - \delta) \le \epsilon dT$$

so

$$U(x_T) \le U(x_\delta) + \epsilon T d.$$

Note that since the process is not fully flippable, this path has to leave the sign of one coordinate unchanged (assume without loss of generality that this is coordinate $i = 1$) and therefore $\|x_T - x\|_\infty \ge \|(x_T)_1 - (x)_1\|_\infty \ge T\theta_1$. Therefore,

$$\inf\{U(y) : \|y - x\|_\infty \ge \theta_1 T\} \le U(x_\delta) + \epsilon T d$$

for all $\epsilon, \delta, T > 0$. If we let $\delta, \epsilon \to 0$ we get

$$\inf\{U(y) : ||y - x|| \ge \theta_1 T\} \le U(x)$$

for all $T > 0$. If we let $T \to \infty$, we can find a sequence y_n with $\|y_n\| \to +\infty$ and $U(y_n) \le U(x)$. This contradicts the growth condition on the energy, where we assume that $\lim_{\|x\|\to\infty} U(x) = +\infty$. Therefore, assuming the statement of equation (3.26) we have proven that the process must be fully flippable.

Now we will prove the statement of (3.26), which will conclude the proof. Fix $\epsilon > 0$. Let $i \in \{1, ..., d\}$. Suppose that $\theta_i \partial_i U(x) < \epsilon$. Since $U \in C^1$, there

is a neighbourhood V_i around x so that $\theta_i \partial_i U(y) < \epsilon$ for all $y \in V_i$. If instead, $\theta_i \partial_i U(x) \geq \epsilon$, then there exists a V_i such that the sign of $\partial_i U$ is unchanged inside V_i. Define $V = \cap_{i=1}^d V_i$. Take $\delta > 0$ small enough so that any trajectory from (x, θ) will not have left V until time δ. We define the set $Nice \subset [\delta, \infty)$ as follows

$$Nice = \{T \in [\delta, +\infty] : \text{ there exists an admissible trajectory } (x_t)_{0 \leq t \leq T} \text{ of length } T, \\ \text{starting from } (x, \theta), \text{ so that (3.26) holds for all } t \in [\delta, T]\}.$$

Our goal is to prove that $Nice = [\delta, \infty)$. From the connectedness of real numbers it suffices to prove that $Nice$ is non-empty, open and closed.
For the non-empty, we use the following construction. Suppose that for some $i \in \{1, ..., d\}$, we have $\theta_i \partial_i U(x) > 0$. Then in a very small period of time, and before time δ, we can, admissibly, flip θ in coordinate i, step by step until it changes sign, from θ_i to $-\theta_i$. After that, we leave the coordinate i unchanged. For all other coordinates j that satisfy $\theta_j \partial_j U(x) > 0$, we can admissibly make the same switches from θ_j to $-\theta_j$ and these switches can happen before time δ. After we have flipped all these coordinates, we extend the path, without any more switches until time δ. Since the path has remained inside V all this time, the sign of $\partial_i U$ is left unchanged for all $i \in \{1, ..., d\}$ and we have $(\theta_\delta)_i \, \partial_i U(x_\delta) < 0 < \epsilon$ for all i. Therefore, $\delta \in Nice$.

Since $U \in C^1$, we can immediately get that $Nice$ is open, since we can always extend an admissible path, without switching any part of the velocity, for a very small period of time, so that it still satisfies (3.26).

It remains to show that $Nice$ is closed. The idea to do this will be to switch the velocity in a similar way as in when we prove $\delta \in Nice$. Suppose that $T_m \in Nice$ is an increasing sequence with $T_m \xrightarrow{m \to \infty} T$. We will prove that we can extend an admissible path, satisfying (3.26) until time T, so $T \in Nice$. Until time T, any path from (x, θ) must lie in a compact set, since the set of available velocities is finite and therefore bounded above by θ^n. Since $U \in C^2$, we can find an upper bound C_T of all the entries of the Hessian matrix of U in this compact set. Take m large enough so that $T - T_m < \epsilon / (2(\theta^n)^2 d C_T)$ and consider an admissible path satisfying (3.26) until time T_m. If for some i, $(\theta_{T_m})_i \partial_i U(x_{T_m}) > 0$ we can admissibly extend the path and quickly flip the sign of the i coordinate, in the same manner as when we proved that $\delta \in Nice$. We can do that for the rest of the coordinates j such that $\theta_j \partial_j U(x_{T_m}) > 0$ and extend the admissible path until some time $S \in [T_m, T]$ so that for all $i \in \{1, ..., d\}$, $(\theta_S)_i \partial_i U(x_S) < 0$. We then extend the path until time T without any switches. Then, the path until time T is admissible, (3.26) holds until

time S and for any $i \in \{1, ..., d\}$ and for any $t \in [S, T]$,

$$|(\theta_t)_i \ \partial_i U(x_t) - (\theta_S)_i \ \partial_i U(x_S)| = |(\theta_S)_i||\partial_i U(x_t) - \partial_i U(x_S)| \leq$$
$$(\theta^m)^2 (T - S) dC_T \leq (\theta^n)^2 (T - T_m) dC_T \leq \frac{\epsilon}{2}$$

Therefore, $(\theta_t)_i \partial_i U(X_t) < \epsilon$ for all $t \in [S, T]$ and $i \in \{1, ..., d\}$, so (3.26) holds until time T. This means that $T \in Nice$ so $Nice$ is closed. This completes the proof. $\square$

Continuity of Reachability

In this paragraph, we will prove the main lemma that extends the deterministic idea of reachability to the probabilistic idea of irreducibility.
The following Lemma will be used extensively in the rest of this section. It confirms that if we can reach one point from another, we can also reach a neighbourhood of that point starting from anywhere close to the other point. Furthermore, in a sense this reachability property is uniform in the starting position. From now on we write Leb for the Lebesgue measure in $\mathbb{R}^d$ (for the appropriate d).

Lemma 3.3.21 (Continuous Component Lemma)**.** *Assume that $U \in C^1$. If $(x, \theta) \looparrowright (y, \eta)$, then there exist opens neighbourhoods U', V of x, y respectively, and constants $\epsilon, t_0, c > 0$, so that if $t \in [t_o, t_o + \epsilon]$, $x' \in U'$, then*

$$\mathbb{P}_{x',\theta} \left(X_t \in \cdot, \Theta_t = \eta \right) \geq \ c \ Leb(\cdot \cap V). \tag{3.28}$$

Proof of 3.3.21. Since $(x, \theta) \looparrowright (y, \eta)$, there exists an admissible sequence
$u = (t, \iota, s) = (t_0, t_1, .., t_m; i_1, ..., i_m; s_1, .., s_m)$ such that the path induced by this control sequence ends at $(x + \theta t_0 + F_{i_1}^{s_1}[\theta] t_1 + ... + t_m F_{i_1,..,i_m}^{s_1,...,s_m}[\theta], F_{i_1,..,i_m}^{s_1,...,s_m}[\theta], \eta) = (y, \eta)$. Let $\tau_k = \sum_{j=0}^{k-1} t_j$ the time of the k'th jump and let $t = \tau_{m+1}$ the ending time of the path.
Since the process has velocity θ, where $\|\theta\| \leq \sqrt{d}\theta^n$, until time $T_0 = 1 + \tau_{m+1}$ the process will not have exit a ball of radius $\sqrt{d}\theta^n(T_0 + 1)$ around x, if the starting point of the process is close to x. Let $\bar{\lambda}$ be an upper bound for all the switching rates in this ball $B = B(x, \sqrt{d}\theta^n(T_0 + 1))$. Using Poisson thinning, we can construct the switches of the process using a bounding Poisson clock of intensity $2d\bar{\lambda}$ in the following way. Once an event from this bounding process occurs, let's say after time $\tilde{t}$, if the process is at $(X_{\tilde{t}}, \Theta_{\tilde{t}})$, we pick $\tilde{\iota} \sim unif\{1, , , .d\}$, $\tilde{s} \sim unif\{+, -\}$ and $\tilde{u} \sim unif[0, 1]$ independent of each other and then we accept a switch of direction

from $\Theta_{\tilde{t}}$ to $F_{\tilde{\iota}}^{\tilde{s}}[\Theta_{\tilde{t}}]$ on the event that

$$\left\{\tilde{u} \leq \frac{\lambda_{\tilde{\iota}}^{\tilde{s}}(X_{\tilde{t}}, \Theta_{\tilde{t}})}{\bar{\lambda}}\right\}. \tag{3.29}$$

For any control sequence (t', ι, s), consider $\lambda_{min}(x', \theta, t', \iota, s)$ to be the minimum of the rates at the switching points of the path induced by the control sequence (t', ι, s), starting from (x', θ). By continuity of the rates, we can find $\underline{\lambda} > 0$, U'' an open neighbourhood of x and for all $i = 1, ..., m$, we can find O_i open neighbourhood of τ_i, so that for all $x' \in U''$ and for any time sequence $t' = (t'_0, .., t'_m)$ with $\sum_{l=0}^{j-1} t'_l \in O_j$ for all j, we have

$$\lambda_{min}(x', \theta, t', \iota, s) > \underline{\lambda}.$$

We can also assume, without loss of generality, that the sets O_i's do not intersect each other, by making them even smaller if needed.

Let A be the event that: *m Poisson events $T_1, ..., T_m$ (according to the bounding process) occur before time t, $T_j \in O_j$ for all $j = 1, ..., m$, all the proposed switches were done according to (i, s) and all the uniform in $[0, 1]$ random variables $v_1, ..., v_m$, used to decide whether we accept of reject a switch according to (3.29), satisfied*

$$v_i \leq \underline{\lambda}/\bar{\lambda} \text{ for all } i \in \{1, ..., m\}. \tag{3.30}$$

Note that since all the rates at the switching points are bounded below by $\underline{\lambda}$,(3.30) implies that all the proposed switches were accepted. At the same time, since the uniform distributions $v_1, ..., v_m$ are independent of the rest of the randomness of the process, the event (3.30) also occurs independently of the rest of the process.
Then, if we write $\Omega(x', t, T_1, ..., T_m) = x' + T_1\theta + (T_2 - T_1)F_{i_1}^{s_1}\theta + ... + (t - T_m)F_{i_1,..,i_m}^{s_1,..,s_m}\theta$ we get, for all $x' \in U''$,

$$\begin{aligned}
&\mathbb{P}_{x',\theta}\left(X_t \in \cdot, \Theta_t = \eta\right) \geq \mathbb{P}_{x',\theta}\left((X_t \in \cdot, \Theta_t = \eta) \cap A\right) \geq \mathbb{P}_{x',\theta}\left(\left(\Omega(x', \theta, T_1, .., T_m) \in \cdot\right) \cap A\right) \geq \\
&\geq \left(\frac{\underline{\lambda}}{2d\bar{\lambda}}\right)^m \mathbb{P}_{x',\theta}\left(\left(\Omega(x', \theta, T_1, .., T_m) \in \cdot\right) \cap \{\text{m events occur}\} \cap \{T_j \in O_j \text{ for all } j\}\right) \geq \\
&\geq c_0 \mathbb{P}_{x',\theta}\left(\Omega(x', t, u_1, .., u_m) \in \cdot\right)
\end{aligned} \tag{3.31}$$

where $u_i \sim unif(O_i)$ and c_0 does not depend on (x', θ). Here we used that the events of a Poisson process with constant rate, conditioned on lying in a set, they are uniformly distributed on that set. Now, since $(x, \theta) \looparrowright (y, \eta)$, we can assume that $\{1, ..., d\} \subset \{i_1, ..., i_m\}$. Therefore, for all $t \in [\tau_m, \tau_m + \epsilon]$, the (up to translation) linear map $(u_1, ..., u_m) \to \Omega(x, \theta, t, u_1, ..., u_m)$ is of full rank as its matrix has column

vectors

$$\left\{\theta - F_{i_1}^{s_1}[\theta], ..., F_{i_1,...,i_{m-1}}^{s_1,...,s_{m-1}}[\theta] - F_{i_1,...,i_m}^{s_1,...,s_m}[\theta]\right\},$$

which span $\mathbb{R}^d$. From Lemma 6.3 of [BLBMZ15] we get that there exists a neighbourhood U_t of x and $c' > 0$ and a neighbourhood V_t of y such that for all $x' \in U_t$,

$$\mathbb{P}_{x',\theta}\left(\Omega(x,\theta,t,u_1,...,u_m) \in \cdot\right) \geq c'\lambda(\cdot \cap V_t).$$

Now, since $\Omega(x,\theta,t,u_1,...,u_m) = x + u_1\theta + (u_2-u_1)F_{i_1}^{s_1}[\theta] + ... + (t-u_m)F_{i_1,...,i_m}^{s_1,...,s_m}[\theta]$, a change in t effects on Ω as a translation in direction $F_{i_1,...,i_m}^{s_1,...,s_m}[\theta]$. This means that, if ϵ is small enough, for every $t \in [\tau_{m+1}, \tau_{m+1}+\epsilon]$ if we pick a starting point (x',θ) with $x' \in U_{\tau_m}$, we get for all $A \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}_{x',\theta}\left(\Omega(x',\theta,t,T_1,...,T_m) \in A\right) = \mathbb{P}_{x',\theta}\left(\Omega(x',\theta,\tau_m,T_1,...,T_m) \in [A - (t-\tau_m)F_{i_1,...,i_m}[\theta]]\right)$$
$$\geq c'\lambda\left([A - (t-\tau_m)F_{i_1,...,i_m}[\theta]] \cap V_{\tau_m}\right) = c'\lambda(A \cap (V_{\tau_m} + (t-\tau_m)F_{i_1,...,i_m}[\theta])).$$

If $\epsilon > 0$ is picked small enough then $\cap_{t\in[\tau_m,\tau_m+\epsilon]}\left(V_{\tau_m} + (t-\tau_m)F_{i_1,...,i_m}[\theta]\right)$ is not empty and contains an open set V. Then, for all $x' \in U_{\tau_m}$, and all $A \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}_{x',\theta}\left(\Omega(x,\theta,t,u_1,...,u_m) \in A\right) \geq c'\lambda(A \cap V).$$

Overall, and setting $U' = U'' \cap U_{\tau_m}$ (where recall that U'' is the open neighbourhood of x such that (3.31) holds for all $x' \in U'$) , for all $x' \in U$ and all $A \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}_{x',\theta}\left(X_t \in A, \Theta_t = \eta\right) \geq c_0\mathbb{P}_{x',\theta}\left(\Omega(x,\theta,t,u_1,...,u_m) \in A\right) \geq c_0c'\lambda(A \cap V).$$

which proves the result. □

Reachability in the General Case

The goal of this section is to extend the reachability result, which we have already proved for the Gaussian case, in a more general class of distributions and finally prove Theorem 3.3.6.

Definition 3.3.22. *For $(x,\theta),(y,\eta) \in E$, we define*

$$(x,\theta) \sim (y,\eta) \iff (x,\theta) = (y,\eta) \text{ or } (x,\theta) \nrightarrow (y,\eta) \nrightarrow (x,\theta).$$

We further let

$$Cl(x,\theta) = \{(y,\eta) : (x,\theta) \sim (y,\eta)\}$$

be the equivalence class of $(x, \theta) \in E$ *under* $\sim$.

Our goal is to prove that there is only one equivalence class. The idea is to study the structure of these equivalence classes and then use the result of Section 3.3.1 to conclude that there can be only one of them. We follow the proof of [BRZ19].

We recall that a set $A \subset E$ is open if for every $(x, \theta) \in A$ there exists a ball $O \subset \mathbb{R}^d$ containing x such that $O \times \{\theta\} \subset A$. In other words we equip Θ with the discrete topology and E with the product topology between $\mathbb{R}^d$ and Θ.

Lemma 3.3.23. *Assume that* $U \in C^1$. *For any* (x, θ), *the equivalence class* $Cl(x, \theta)$ *is either a single point or an open set of* E. *Moreover, if* R *is the velocity reversal operator (applied on points of sets), we have* $R(Cl(x, \theta)) = Cl(R(x, \theta))$. *In particular, the classes of* (x, θ) *and* $(x, -\theta)$ *are of the same type (open or singleton).*

Proof. Suppose that $(x, \theta), (x', \theta')$ are two different points in the same equivalence class. Then $(x, \theta) \not\mapsto (x', \theta')$ so from Lemma 3.3.18, $(x, \theta) \not\mapsto (y, \eta)$ for all (y, η) in a neighbourhood A_1 of (x', θ'). Also, $(x', \theta') \not\mapsto (x, \theta)$ so by time reversal $(x, -\theta) \not\mapsto (x', -\theta')$, so $(x, -\theta)$ leads to all points in a neighbourhood $B_2 \times \{-\theta'\}$ of $(x', -\theta')$. Again by time reversal, the points in the neighbourhood $A_2 = B_2 \times \{\theta'\}$ of (x', θ') will lead to (x, θ). If a point (y, η) lies in $A_1 \cap A_2$ then $(x, \theta) \not\mapsto (y, \eta) \not\mapsto (x, \theta)$ and therefore $(y, \theta) \in Cl(x, \theta)$ so $A_1 \cap A_2 \subset Cl(x, \theta)$. Since $A_1 \cap A_2$ forms a neighbourhood of (x', θ') we conclude that $Cl(x, \theta)$ is open. This proves the first part of the Lemma.

If $(y, \eta) \in R(Cl(x, \theta))$ then $(y, -\eta) \not\mapsto (x, \theta) \not\mapsto (y, -\eta)$ and by time reversal $(x, -\theta) \not\mapsto (y, \eta) \not\mapsto (x, -\theta)$, so $(y, \eta) \in Cl(R(x, \theta))$. Therefore $R(Cl(x, \theta)) \subset Cl(R(x, \theta))$. Since this is true for all (x, θ) we can apply it for $(x, -\theta)$ and get $R(Cl(x, -\theta)) \subset Cl(R(x, -\theta)) = Cl(x, \theta)$. Applying R to both sides we get $Cl(x, -\theta) \subset R(Cl(x, \theta))$ and therefore $Cl(R(x, \theta)) \subset R(Cl(x, \theta))$. Overall we get $R(Cl(x, \theta)) = Cl(R(x, \theta))$ and the result follows. □

Now we can characterise the structure of equivalence classes.

Proposition 3.3.24. *Assume that that the process is fully flippable and* $U \in C^1$. *Let* $(x, \theta) \in E$. *If* $Cl(x, \theta)$ *is open, then for every* $(y, \eta) \in E$ *we have* $(x, \theta) \not\mapsto (y, \eta) \iff (y, \eta) \not\mapsto (x, \theta) \iff (x, \theta) \sim (y, \eta)$. *Furthermore, the non-singleton classes are of the form* $\mathbb{R}^d \times V$ *for some* $V \subset \Theta$.

Proof of Proposition 3.3.24. Suppose that $O = Cl(x, \theta)$ is a non-empty and therefore open class. Recall that μ is the invariant measure of the process and let O^+ be

the future of O, that is

$$O^+ = \{(y,\eta) : \text{ there exists } (z,\xi) \in O \text{ with } (z,\iota) \not\hookrightarrow (y,\eta)\} = \{(y,\eta) : (x,\theta) \not\hookrightarrow (y,\eta)\}.$$

Let $(y,\eta) \in O^+$. Then $(x,\theta) \not\hookrightarrow (y,\eta)$, so from Lemma 3.3.18, there exists a neighbourhood A_1 of y such that $(x,\theta) \not\hookrightarrow (z,\eta)$ for all $z \in A_1$. Therefore, O^+ is open and consequently measurable. Now, if $(P^t)_{t\geq 0}$ is the transition semi-group of MDCNZZ, since μ is invariant, we have $\mu = \mu P^t$ for all t, therefore

$$\begin{aligned}\mu(O^+) &= \int_0^\infty \exp\{-t\}\mu P^t(O^+)dt = \int_0^\infty \int_E \exp\{-t\}\mathbb{P}_{(x',\theta')}\left((X_t,\Theta_t) \in O^+\right) d\mu(x',\theta')dt = \\ &= \int_0^\infty \int_E \exp\{-t\}1_{(x',\theta')\in O^+}\mathbb{P}_{(x',\theta')}\left((X_t,\Theta_t) \in O^+\right) d\mu(x',\theta')dt + \\ &+ \int_0^\infty \int_E \exp\{-t\}1_{(x',\theta')\notin O^+}\mathbb{P}_{(x',\theta')}\left((X_t,\Theta_t) \in O^+\right) d\mu(x',\theta')dt.\end{aligned}$$

Now, note that any point in the future of O^+ is still in O^+ so all the probabilities in the first integral are 1 and therefore the first integral is equal to $\mu(O^+)$. Therefore,

$$\int_0^\infty \int_E \exp\{-t\}1_{(x',\theta')\notin O^+}\mathbb{P}_{(x',\theta')}\left((X_t,\Theta_t) \in O^+\right) d\mu(x',\theta')dt = 0.$$

Therefore, there exists a set A of full μ-measure so that if $(x',\theta') \in A$,

$$1_{(x',\theta)'\notin O^+}\int_0^\infty \exp\{-t\}\mathbb{P}_{(x',\theta')}\left((X_t,\Theta_t) \in O^+\right) dt = 0. \tag{3.32}$$

Now, assume that $(y,\eta) \in A$ and that $(y,\eta) \not\hookrightarrow (x,\theta)$. Then, for any $(z,\xi) \in O^+$, $(y,\eta) \not\hookrightarrow (z,\iota)$. From Continuous Component Lemma, Lemma 3.3.21, there exists $c > 0$, an interval $[t_0, t_0+\epsilon]$ and for all $t \in [t_0, t_0+\epsilon]$, $\mathbb{P}_{(y,\eta)}\left((X_t,\Theta_t) \in O^+\right) \geq c$. Therefore

$$\int_0^\infty \exp\{-t\}\mathbb{P}_{(x,\theta)}[(X_t,\Theta_t) \in O^+]dt > 0,$$

and due to (3.32), we have that $(y,\eta) \in O^+$ meaning that $(x,\theta) \not\hookrightarrow (y,\eta)$. Since by assumption, we also have $(y,\eta) \not\hookrightarrow (x,\theta)$ we get $(x,\theta) \sim (y,\eta)$.
Assume now, that $(y,\eta) \in A$ and that $(x,\theta) \not\hookrightarrow (y,\eta)$. Since A has full μ measure and μ is uniform in the set Θ, we can assume without loss of generality (if needed, by discarding some subsets of A that have μ-measure zero and they break the symmetry from A) that A is stable under reversal of velocities. Then, by time reversal $(y,-\eta) \not\hookrightarrow (x,-\theta) \in R(O)$ and since $(y,-\eta) \in A$ we get from the previous part of the proof that $(y,-\eta) \sim (x,-\theta)$ so $(x,\theta) \sim (y,\eta)$ from time reversal.

We have, so far, proven the first result of the proposition in a set A of full μ measure. Now, consider any (y,η) such that $(x,\theta) \looparrowright (y,\eta)$. Since the process is fully flippable, we get that there exists a (z,ξ) so that $(y,\eta) \looparrowright (z,\xi)$. Therefore, (y,η) targets a neighbourhood of (z,ξ), due to Lemma 3.3.18. Since A is of full μ measure and therefore of full Lebesgue measure, there exists a point $(z',\xi) \in A$ which is also in that neighbourhood of (z,ξ). Then, $(x,\theta) \looparrowright (y,\eta) \looparrowright (z',\xi) \in A$. Using the previous part we get that $(z',\xi) \in Cl(x,\theta)$. Then $(z',\xi') \looparrowright (x,\theta)$ and overall $(x,\theta) \looparrowright (y,\eta) \looparrowright (z',\xi) \looparrowright (x,\theta)$. So $(y,\eta) \in Cl(x,\theta)$. We have proved that if $Cl(x,\theta)$ contains more than one elements and $(x,\theta) \looparrowright (y,\eta)$ then $(x,\theta) \sim (y,\eta)$.

Assume, now, that $(y,\eta) \looparrowright (x,\theta)$ therefore $(x,-\theta) \looparrowright (y,-\eta)$. From Lemma 3.3.23 we have thar $Cl(x,-\theta)$, also contains more than one elements and from the result we just proved we get $(x,-\theta) \sim (y,-\eta)$, therefore $(x,\theta) \sim (y,\eta)$ from time reversal.

To sum up, we have proven the first part of the proposition, that if $Cl(x,\theta)$ contains more than one elements, then for any $(y,\eta) \in E$, $(x,\theta) \looparrowright (y,\eta) \iff (y,\eta) \looparrowright (x,\theta) \iff (x,\theta) \sim (y,\eta)$.

Finally, we will prove that, under full flippability, any non-singular equivalence class O is of the form $O = \mathbb{R}^d \times V$ for some $V \subset \Theta$, which will conclude the proof of the proposition. In order to do this, we will prove that any non-singular equivalence class O is a closed set. Since, from Lemma 3.3.23, O must also be open, from the connectedness of $\mathbb{R}^d$, we get that O contains copies of $\mathbb{R}^d$, i.e. $O = \mathbb{R}^d \times V$ for some $V \subset \Theta$.

Suppose that O is a non-singular equivalence class. Let $(x,\theta) \in \bar{O}$, the closure of O. Since the process is fully flippable, from Lemma 3.3.18 we can find a point (y,η) with $(x,\theta) \looparrowright (y,\eta)$. From Lemma 3.3.18, there exists a neighbourhood V around (y,η) such that for any $(z,\xi) \in V$, $(x,\theta) \looparrowright (z,\xi)$. Now, (y,η) is reached by (x,θ) via an admissible path. Since the rates are continuous, for all x' sufficiently close to x, the MDCNZZ path induced by the same control sequence but starting from (x',θ) (a "parallel path") will be admissible and will reach the point $(y+x'-x,\eta)$, i.e. $(x',\theta) \looparrowright (y+x'-x,\eta)$. Also, for all x' sufficiently close to x we have $(y+x'-x,\eta) \in V$ and therefore $(x,\theta) \looparrowright (y+x'-x,\eta)$ for all x' sufficiently close to x. Since $(x,\theta) \in \bar{O}$ we can find x' sufficiently close to x with $(x',\theta) \in O$. Then, $(x',\theta) \looparrowright (y+x'-x,\eta)$ so $(y+x'-x,\eta) \in O$ from the previous part of the theorem. But recall that we also have $(x,\theta) \looparrowright (y+x'-x,\eta)$, therefore $(x,\theta) \in O$. Therefore, O is closed. □

We can now prove the reachability result which was the original goal of the section.

Proof of Theorem 3.3.6. Since $U \in C^3$ and has a non-degenerate local minimum,

from Proposition 3.3.17 we get that there exists one equivalence class $\mathcal{K}$ that has a subset of the form $B \times \Theta$ for some ball $B \subset \mathbb{R}^d$ around the local minimum. Since $\lim_{x\to\infty} U(x) = +\infty$, from Proposition 3.3.20 the process is fully flippable and, therefore, from Proposition 3.3.24, $\mathcal{K}$ must be of the form $\mathbb{R}^d \times H$, for some $H \subset \Theta$. We conclude that $\mathcal{K} = \mathbb{R}^d \times \Theta$ which proves the result. □

3.3.2 Non-Evanescence

In this section we prove that the MDCNZZ will almost surely not diverge to infinity. Under some extra properties that we will establish in the next section, this will be equivalent to having positive Harris recurrence which is essential for ergodicity.

Definition 3.3.25. *A point (x, θ) is non-evanescent if $\mathbb{P}_{x,\theta}(\lim_{t\to\infty} |X_t| = \infty) = 0$. The process is non-evanescent if every point is non-evanescent.*

Throughout this section we will impose Assumption 3.3.2. Recall that this means that $U \in C^3$ and there exists $c > d$, $c' \in \mathbb{R}^d$ so that for all $x \in \mathbb{R}^d$

$$U(x) \geq c \cdot log(1 + \|x\|) - c'. \tag{3.33}$$

As noted in the notes after Theorem 2.4.7 in Chapter 2 this is an assumption that guarantees that the invariant measure is probability and in practice is a condition that holds. The main goal of the section is to prove the following.

Proposition 3.3.26 (Non-Evanescence of MDCNZZ)**.** *Assume that growth condition 3.3.2 holds and that $U \in C^3$. Then the MDCNZZ process is non-evanescent.*

The rest of this section is devoted to the proof of Proposition 3.3.26. The method of proof closely follows [BRZ19].
We begin by introducing some notation which is needed for our first Lemma.

Definition 3.3.27. *Let $(T_i)_{i=1}^{\infty}$ be the random times where the components of the velocity switch. Let N the random integer so that T_N is the first time when $d-1$ components of the velocity have switched (and $T_N = \infty$ if this does not occur). Let $\tau = T_{N+1}$ if T_N is finite and $\tau = \infty$ else , i.e. τ is the next switching time after T_N.*

Lemma 3.3.28 (Continuity of Jump Times)**.** *Assume that $U \in C^1$ and $(X_t, \Theta_t)_{t\geq 0}$ a MDCNZZ process that targets measure μ introduced in (3.1). Let τ be as in the Definition 3.3.27. Then the distribution of X_τ is absolutely continuous with respect to the Lebesgue measure, i.e. if $B \subset \mathbb{R}^d$ with zero Lebesgue measure, then for all $(x, \theta) \in E$*

$$\mathbb{P}_{x,\theta}(\tau < \infty, X_\tau \in B) = 0.$$

Proof of Lemma 3.3.28. Let B a set of zero Lebesgue measure in $\mathbb{R}^d$. We will prove that for any T,

$$\mathbb{P}_{x,\theta}\left(\tau \leq T, X_\tau \in B\right) = 0$$

and the result will follow, on letting $T \to \infty$, by monotone convergence.
Fix $T \geq 0$. Since the process has finite speed, the path until time T is contained in a ball of radius $\theta^n T$ and centre x. Find an upper bound $\bar{\lambda}$ for the rates inside this ball. Then using a Poisson thinning argument, in the same manner as in the proof of Proposition 3.3.21, the path in $[0, T]$ can be constructed by using a Poisson process with intensity $2d\bar{\lambda}$ and simulating arrival times according to that process. Whenever the process is at (x, θ), the x-component moves with constant velocity θ until the next Poisson event. If the event occurs at time t and the process is at X_t, then we pick an element (i, s) of $\{1, .., d\} \times \{-, +\}$ uniformly at random and accept the change in the i coordinate towards s (upwards or downwards) with probability

$$\frac{\lambda_i^s(X_t, \Theta_t)}{\bar{\lambda}}.$$

According to this construction, after the first time the bounding process generates an event, the new velocity might have changed in one coordinate (say coordinate i) either by increasing, decreasing or staying exactly the same (if the switch is rejected). Write $F_i^s[\theta]$ for this new direction, if the previous one was θ. Here i symbolises the coordinate that might change and s is allowed the values $-1, +1$ or 0 to signify whether the change effects a decrease, an increase or does not do anything. After k Poisson events we will have velocity $F_{i_1,..,i_k}^{s_1,..,s_k}[\theta]$. At the same time, the time between any two events generated by the bounding process is an exponential with parameter $2d\bar{\lambda}$. Also, any two times between events are independent. Therefore, at time τ we have

$$X_\tau = x + E_1\theta + E_2 F_{I_1}^{S_1}[\theta] + .. + E_{M+1} F_{I_1,..,I_M}^{S_1,..,S_M}[\theta],$$

where M, I_k, S_k, E_k are all random variables, possibly correlated to each other and M is the number of bounding Poisson events that occur until the process has switched $d-1$ components of the velocity, i.e.

$$M = \inf\{n \in \mathbb{N} : |A_n| = d - 1\}$$

where

$$A_n = \{i \in \{1, ..., d\} : \text{ there exists } k \leq n : I_k = i \text{ and } S_k \neq 0\}.$$

Note that E_k are i.i.d. exponential random variables with parameter $2d\bar{\lambda}$ and I_k are

i.i.d. uniform in $\{1, ..., d\}$ and independent of all E_k's. Then,

$$\mathbb{P}(\tau \leq T, X_\tau \in B) =$$
$$= \sum_{m=1}^{+\infty} \sum_{(i_1,..,i_m)\in\{1,...,d\}^m} \sum_{(s_1,..,s_m)\in\{-1,0,1\}^m}$$
$$\mathbb{P}(\tau \leq T, M = m, I_k = i_k, S_k = s_k \text{ for all } k \leq m, X_\tau \in B)$$
$$= \sum_{m=1}^{+\infty} \sum_{(i_1,..,i_m)\in\{1,...,d\}^m} \sum_{(s_1,..,s_m)\in\{-1,0,1\}^m} \tag{3.34}$$
$$\mathbb{P}\left(\tau \leq T, M = m, I_k = i_k, S_k = s_k \text{ for all } k \leq m, x + E_1\theta + E_2 F_{i_1}^{s_1}[\theta] + .. + E_{m+1} F_{i_1,..,i_m}^{s_1,..,s_m}[\theta] \in B\right)$$

Now, let us fix a specific configuration of $m, i_1, ..., i_m, s_1, ..., s_m$ and consider the probability in the above sum. Because of the definition of M and since we consider an event where $M = m$, if there exist at least two coordinates in which all the velocities $\theta, F_{i_1}^{s_1}\theta, ..., F_{i_1,..,i_m}^{s_1,..,s_m}\theta$ have the same value, the above event has zero probability. If there exists at most one coordinate in which all the velocities $\theta, F_{i_1}^{s_1}\theta, ..., F_{i_1,..,i_m}^{s_1,..,s_m}\theta$ have the same value, then the differences between any two of these velocities, along with θ, span $\mathbb{R}^d$. Therefore $\{\theta, F_{i_1}^{s_1}\theta, ..., F_{i_1,..,i_m}^{s_1,..,s_m}\theta\}$ span $\mathbb{R}^d$. Therefore for such configuration of $m, i_1, ..., i_m, s_1, ..., s_m$,

$$\mathbb{P}\left(\tau \leq T, M = m, I_k = i_k, S_k = s_k \text{ for all } k \leq m, x + E_1\theta + E_2 F_{i_1}^{s_1}[\theta] + .. + E_{m+1} F_{i_1,..,i_m}^{s_1,..,s_m}[\theta] \in B\right)$$
$$\leq \mathbb{P}\left(x + E_1\theta + E_2 F_{i_1}^{s_1}[\theta] + .. + E_{m+1} F_{i_1,..,i_m}^{s_1,..,s_m}[\theta] \in B\right) = 0$$

because $\{\theta, F_{i_1}^{s_1}\theta, ..., F_{i_1,..,i_m}^{s_1,..,s_m}\theta\}$ span $\mathbb{R}^d$, $E_1, ..., E_m$ are i.i.d. are exponential with parameter $2d\bar{\lambda}$ and B has Lebesgue measure zero.
Overall, every part of the sum in (3.34) is zero and this proves the result. □

We now use Lemma 3.3.28 to present a useful lower bound of the probability of not escaping to infinity.

Lemma 3.3.29. *Assume that the invariant measure μ is a probability and $U \in C^1$. Let τ be the first switching time after $d-1$ coordinates of the velocity have switched signs, as introduced before Lemma 3.3.28. If the process starts from (x, θ), then*

$$\mathbb{P}_{x,\theta}\left(\liminf_{t\to\infty} |X_t| < \infty\right) \geq \mathbb{P}_{x,\theta}(\tau < \infty). \tag{3.35}$$

Proof of 3.3.29. Let π be the marginal probability measure of μ on $\mathbb{R}^d$. Assume first that we start the process from the stationary measure μ. Take $K \subset \mathbb{R}^d$ compact.

Then $\liminf_{t\to\infty} 1_{X_t\notin K} = 1_{\{X_t \text{ eventually leaves } K\}}$ and by Fatou,

$$\mathbb{P}_\mu\left(X_t \text{ eventually leaves } K\right) \leq \liminf_{t\to\infty} \mathbb{P}_\mu\left(X_t \notin K\right) = 1 - \mu(K\times\Theta)$$

Now since $\{\lim_{t\to+\infty}\|X_t\| = +\infty\} = \cap_K\{X_t \text{ eventually leaves } K\}$, we get $\mathbb{P}_\mu\left(\lim_{t\to+\infty}\|X_t\| = +\infty\right) \leq \lim_{K\nearrow\mathbb{R}^d} 1-\mu(K\times\Theta) = 0$. Therefore, if A is defined as

$$A = \{(x,\theta)\in E : (x,\theta) \text{ is non-evanescent }\},$$

then $\mu(A) = 1$. Now A can be partitioned as $A = \cup_{\theta\in\Theta} A_\theta\times\{\theta\}$ where for any $\theta\in\Theta$

$$A_\theta = \{x\in\mathbb{R}^d : (x,\theta) \text{ is non-evanescent. }\}$$

therefore for any $\theta\in\Theta$,

$$\mu(A_\theta^c\times\{\theta\}) = 0.$$

Now, since μ is the product of π and the uniform measure on Θ, we have that for all $\theta,\eta\in\Theta$,

$$\mu(A_\theta\times\{\eta\}) = \mu(A_\theta^c\times\{\theta\}) = 0$$

and therefore for all $\theta\in\Theta$

$$\pi(A_\theta) = \mu(A_\theta\times\Theta) = 0$$

Therefore if N is defined as

$$N = \{x\in\mathbb{R}^d : \text{ for all } \theta\in\Theta, (x,\theta) \text{ is non-evanescent }\},$$

then

$$\pi(N^c) = \pi(\cup_{\theta\in\Theta} A_\theta^c) = 0$$

so $\pi(N) = 1$.
Now, let us start the process from any (x,θ) and let τ be the first switching time after $d-1$ coordinates of the velocity have switched, as in the Definition 3.3.27. Then

$$\mathbb{P}_{x,\theta}\left(\liminf_{t\to\infty}|X_t| < \infty\right) \geq \mathbb{P}_{x,\theta}\left(\tau<\infty, \liminf_{t\to\infty}|X_t|<\infty\right) =$$
$$= \mathbb{E}_{x,\theta}\left[1_{\tau<\infty}\mathbb{P}_{X_\tau,\Theta_\tau}\left(\liminf_{t\to\infty}|X_t|<\infty\right)\right] \geq \mathbb{E}_{x,\theta}\left[1_{\tau<\infty}1_{(X_\tau,\Theta_\tau)\in N}\right] = \mathbb{P}_{x,\theta}\left(\tau<\infty, X_\tau\in N\right)$$

Now, from equivalence of measures, since $\pi(N^c) = 0$ we have $\lambda(N^c) = 0$ and from

absolute continuity of jumps as in Lemma 3.3.28 we get

$$\mathbb{P}_{x,\theta}\left(\tau < \infty, X_\tau \notin N\right) = 0$$

and therefore

$$\mathbb{P}_{x,\theta}\left(\lim_{t\to\infty} |X_t| \neq \infty\right) \geq \mathbb{P}_{x,\theta}\left(\tau < \infty\right).$$

□

Lemma 3.3.29 allows us to prove the next result, which will be our main stepping stone in order to prove Proposition 3.3.26.

Lemma 3.3.30. *Assume that $U \in C^1$. Assume further that $d \geq 2$ and the MDCNZZ process has invariant probability measure μ. Assume that $\mathbb{P}_{x,\theta}\left(|X_t| \to \infty\right) > 0$ for some (x, θ). Then, there exist coordinates i, j so that*

$$\mathbb{P}_{x,\theta}\left(i, j \text{ coordinates of velocity never switch}\right) > 0. \tag{3.36}$$

Proof of 3.3.30. Assume otherwise, i.e. for starting point (x, θ), a.s. at most one coordinate of the velocity never switches. Then, for the stopping time T_N, the first time all but one coordinates have switched, we have $T_N < \infty$ a.s. From Lemma 3.3.31 we know that the process will have infinitely many switches a.s., therefore $\tau = T_{N+1} < \infty$ a.s. But then, from Lemma 3.35 $\mathbb{P}_{x,\theta}\left(\lim_{t\to\infty} |X_t| \neq \infty\right) \geq \mathbb{P}_{x,\theta}\left(\tau < \infty\right) = 1$. □

Finally, before we prove Proposition 3.3.26 we prove the following Lemma.

Lemma 3.3.31. *Assuming Growth Condition 3.3.5 and $U \in C^1$ the MDCNZZ a.s. has infinitely many switches.*

Proof of 3.3.31. We can assume without loss of generality that we have minimal switching rates. Suppose we start from (x, θ), then

$$\begin{aligned}
&\mathbb{P}_{x,\theta}\left(\text{ no switches untill time } T\right) = \\
&= \exp\{-\int_0^T \sum_{i=1}^d [\theta_i^{\leq}\ \partial_i U(x+\theta s)]^+ + [\theta_i^{<}\ \partial_i U(x+\theta s)]^- ds\} \leq \\
&\leq \exp\{-\int_0^T \sum_{i=1}^d \theta_i^{\leq}\ \partial_i U(x+s\theta) - \theta_i^{<}\ \partial_i U(x+s\theta) ds\} = \\
&= \exp\{-\int_0^T \sum_{i=1}^d \theta_i \partial_i U(x+s\theta) ds\} = \\
&= \exp\{-U(x+\theta T) + U(x)\} \xrightarrow{T\to+\infty} 0.
\end{aligned}$$

Therefore if $T_1, T_2, ..$ the sequence of jumps $\mathbb{P}_{x,\theta}[T_1 < \infty] = 1$.
By the Strong Markov Property, $\mathbb{P}[T_{k+1} < \infty] = \mathbb{E}_{x,\theta}[1_{T_k<\infty}\mathbb{P}_{X_{T_k},\Theta_{T_k}}[T_1 < \infty]] = \mathbb{P}_{x,\theta}[T_k < \infty]$ and therefore all the jump times are a.s. finite by induction. □

We can now prove the main result of this section, Proposition 3.3.26. We will use Lemmas 3.3.30 and 3.3.31. The proof uses induction in the dimension of the space. If a $d+1$-dimensional process fails to be non-evanescent, starting from some point (x, θ), then we construct a d-dimensional process that, also, fails to be non-evanescent. Although we follow very closely the proof of Proposition 4 in [BRZ19], we need to be a bit careful and distinguish between different possibilities for the value of θ.

Proof of Proposition 3.3.26. The proof will proceed with induction on the dimension of state space d. Consider the proposition $\mathcal{P}_d$: For every C^3 function $U : \mathbb{R}^d \to \mathbb{R}$ satisfying Growth Condition 3.3.2, the MDCNZZ process induced by U is non-evanescent.

Recall here that Growth Condition 3.3.2 for the d-dimensional process states that there exists a $c > d$ and $c' \in \mathbb{R}$ such that for all $x \in \mathbb{R}^d$, $U(x) \geq c\log(1+\|x\|)-c'$.

For $d = 1$, if τ is the first switching time after the process has switched all but one coordinates, then in this case τ is also the first switching time. Since we know that a.s. there are infinitely many switches, from Lemma 3.3.29 we get for all $(x, \theta) \in \mathbb{R} \times \{\pm 1\}$

$$\mathbb{P}_{x,\theta}\left(|X_t| \to \infty\right) \geq \mathbb{P}_{x,\theta}\left(\tau < \infty\right) = 1.$$

which proves non-evanescence.

For $d = 2$ assume that for some (x, θ), $\mathbb{P}_{x,\theta}\left(|X_t| \to \infty\right) > 0$. From Lemma 3.3.26, with positive probability the process will never switch and this contradicts Lemma 3.3.31.

Now, assume that $\mathcal{P}_d$ is true for some $d \geq 2$. Suppose that $\mathcal{P}_{d+1}$ is not true. That means that there exists a C^3 function $U : \mathbb{R}^{d+1} \to \mathbb{R}$ and constants $c > d + 1 > d$, $c' \in \mathbb{R}$ such that for all $x \in \mathbb{R}^{d+1}$

$$U(x) \geq c\log(1 + |x|) - c',$$

but such that there exists a point $(x, \theta) = ((x_1, ..., x_{d+1}); (\theta_1, ..., \theta_{d+1}))$ with
$\mathbb{P}_{x,\theta}\left(|X_t| \to \infty\right) > 0$. From Lemma 3.3.26, there exist two coordinates, and suppose without loss of generality that these are d and $d + 1$, so that
$\mathbb{P}_{x,\theta}\left(d, d+1 \text{ coordinate of velocity never change}\right) > 0$.

Exponential Representation of the Process: *Let us formulate the MDCNZZ in the following way, inspired by the exponential representation of the Poisson process, Proposition 2.2.2. Take* $(U_j^k)_{j=1,..,d+1}^{k\in\mathbb{N}}$ *and* $(D_j^k)_{j=1,..,d+1}^{k\in\mathbb{N}}$ *i.i.d. exponential random variables with parameter 1 and say that the k'th "upwards" jump of the j'th component of the velocity happens at time* $T_j^{k,+}$*, which is defined as,*

$$T_j^{k,+} = \inf\left\{ t \geq 0 : \int_{T_j^{k-1,+}}^{t} \lambda_j^+(X_s, \Theta_s)ds \geq U_j^k \right\}.$$

Same idea applies to the "downwards" jumps $T_j^{k,-}$ *by using* D_j^k *instead of* U_j^k *and* λ^- *instead of* λ^+*.*

Recall here that λ^+, λ^- are the rates of up-wards/down-wards steps for the process and are defined by (3.23) and (3.24).

Using this construction we will couple our $d+1$-dimensional MDCNZZ process, with a d-dimensional MDCN Zig-Zag that violates $\mathcal{P}_d$.

We will distinguish between the following five cases for θ (where $(x, \theta) = (x_1, ..., x_{d+1}; \theta_1, ..., \theta_{d+1})$ is the starting point of the $d+1$-dimensional process).

The first case is when $\theta_{d+1}^{\leq} \neq 0$ and $\theta_{d+1}^{<} \neq 0$.

In that case, consider the d-dimensional potential

$$V(y_1, .., y_d) = U(y_1, ..., y_d, y_d) \tag{3.37}$$

which induces a d-dimensional MDCN Zig-Zag process $(Y, H) = (Y_1, .., Y_d; H_1, .., H_d)$, when V is the minus log-likelihood of the process' invariant measure. We can observe that

$$V(y_1, .., y_d) \geq c\log(1 + \|(y_1, .., y_d, y_d)\|) - c' \geq c\log(1 + \|y_1, .., y_d\|) - c'$$

and $V \in C^3$ so (Y, H) must be non-evanescent by induction hypothesis.

Coupling Construction: *We will couple the two processes us follows. As* (Y, H) *is a MDCNZZ process, it can be constructed using a sequence of i.i.d. exponential variables with parameter* 1*, using the exponential representation of the process, described above. Let us use* $(U_j^k)_{j=1,..,d-1}^{k\in\mathbb{N}}$ *and* $(D_j^k)_{j=1,..,d-1}^{k\in\mathbb{N}}$ *for the first* $d-1$ *coordinates of* (Y, H)*. These exponential random variables are the same as the ones used to define the process* (X, Θ)*. For the d coordinate of* (Y, H)*, use i.i.d.* $(\tilde{U}_d^k)_{k\in\mathbb{N}}$ *and* $(\tilde{D}_d^k)_{k\in\mathbb{N}}$ *exponentials with parameter* 1*, independent of the others.*

Let us start the d-dimensional process from $((x_1, ..., x_d); (\theta_1, ..., \theta_d))$. The processes Θ_d and H_d both start from θ_d, whereas Θ_{d+1} starts from θ_{d+1}. Let τ be

the first time one of Θ_d, Θ_{d+1} or H_d changes. Suppose that $t \leq \tau$, so until time t, $\Theta_d = H_d$. Furthermore, due to the way we coupled the processes, the first d coordinates of the process X are equal to the d coordinates of Y until time t.
Using the bound $(a+b)^+ \leq a^+ + b^+$ we get the following estimates for the upward rates of the d-coordinate of the (Y, H)-process,

$$\begin{aligned}
&\int_0^t \lambda^+_{Y,H,d}(Y_s, H_s)ds = \int_0^t [H_d^{\leq}(s)\partial_d V(Y_s)]^+ ds = \int_0^t [\theta_d^{\leq}\partial_d V(Y_s)]^+ ds = \\
&= \int_0^t [\theta_d^{\leq}\partial_d U(X_s) + \theta_d^{\leq}\partial_{d+1} U(X_s)]^+ ds \leq \\
&\leq \int_0^t [\theta_d^{\leq}\ \partial_d U(X_s)]^+ + [\theta_d^{\leq}\ \partial_{d+1} U(X_s)]^+ ds \leq \\
&\leq \int_0^t [\theta_d^{\leq}\ \partial_d U(X_s)]^+ + \frac{\theta_d^{\leq}}{\theta_{d+1}^{\leq}} [\theta_{d+1}^{\leq}\ \partial_{d+1} U(X_s)]^+ ds. \qquad (3.38)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\int_0^t \lambda^-_{Y,H,d}\ (Y_s, H_s)ds = \int_0^t [-H_d(s)^{<}\ \partial_d V(Y_s)]^+ ds \leq \\
&\leq \int_0^t [-\theta_d^{<}\ \partial_d U(X_s))]^+ ds + \frac{\theta_d^{<}}{\theta_{d+1}^{<}} \int_0^t [-\theta_{d+1}^{<}\ \partial_{d+1} U(X_s)]^+ ds
\end{aligned}$$

Consider the event $A = \{\Theta_d \text{ and } \Theta_{d+1} \text{ never change }\} \cap \{\tilde{U}_d^1 > U_d^1 + \frac{\theta_d^{\leq}}{\theta_{d+1}^{\leq}} U_{d+1}^1\} \cap \{\tilde{D}_d^1 > D_d^1 + \frac{\theta_d^{<}}{\theta_{d+1}^{<}} D_{d+1}^1\}$. This event has positive probability and on this event, for all t,

$$\begin{aligned}
\int_0^t \lambda^+_{Y,H,d}(Y_s, H_s)ds &\leq \int_0^\infty [\theta_d^{\leq}\ \partial_d U(X_s)]^+ ds + \frac{\theta_d^{\leq}}{\theta_{d+1}^{\leq}} \int_0^\infty [\theta_{d+1}^{\leq}\ \partial_{d+1} U(X_s)]^+ ds < \\
&< U_d^1 + \frac{\theta_d^{\leq}}{\theta_{d+1}^{\leq}} U_{d+1}^1 \leq \tilde{U}_d^1
\end{aligned}$$

and similarly

$$\int_0^t \lambda^-_{Y,H,d}(Y_s, H_s)ds < \tilde{D}_d^1.$$

This means that on event A the d coordinate of H_t never changes and therefore $|Y_t| \to \infty$. This contradicts the induction hypothesis $\mathcal{P}_d$ in the case where both $\theta_{d+1}^{\leq}, \theta_{d+1}^{<} \neq 0$.

The second case we treat is when $\theta_{d+1}^{\leq} = 0$ and $\theta_d^{<} \neq 0$. This also implies that $\theta_{d+1} = \theta^n$.

We use the potential V introduced in (3.37) to construct again the process (Y, H) that targets the measure that has V as minus log-likelihood. However, this time, we start the process from point $((x_1, ..., x_d); (\theta_1, ..., \theta_{d-1}, \theta_{d+1}))$. We also use the same coupling method.

Let, once again, τ be the first time one of Θ_d, Θ_{d+1} or H_d changes. Suppose that $t \leq \tau$, so until time t, $\Theta_{d+1} = H_d$ until time t and, due to the way we coupled the processes, the coordinates $\{1, 2, ..., d-1, d+1\}$ of the process X are equal to the coordinates of the process Y until time t. Then, we write

$$\int_0^t \lambda^+_{Y,H,d}(Y_s, H_s)ds = \int_0^t [H_d^{\leq}(s)\partial_d V(Y_s)]^+ ds = \int_0^t [\theta_{d+1}^{\leq}\partial_d V(Y_s)]^+ ds = 0$$

and at the same time

$$\begin{aligned}\int_0^t \lambda^-_{Y,H,d}(Y_s, H_s)ds = \int_0^t [-H_d^{\leq}(s)\partial_d V(Y_s)]^+ ds = \int_0^t [-\theta_{d+1}^{<} (\partial_d U(X_s) + \partial_{d+1} U(X_s))]^+ ds \leq \\ \leq \int_0^t [-\theta_{d+1}^{<}\partial_d U(X_s)]^+ + [-\theta_{d+1}^{<}\partial_{d+1} U(X_s)]^+ ds = \\ = \frac{\theta_{d+1}^{<}}{\theta_d^{<}} \int_0^t [-\theta_d^{<} \partial_d U(X_s)]^+ ds + \int_0^t [-\theta_{d+1}^{<} \partial_{d+1} U(X_s)]^+ ds.\end{aligned}$$

Similar to the previous case, consider the event $B = \{\Theta_d \text{ and } \Theta_{d+1} \text{ never change }\} \cap \{\tilde{D}_d^1 > \frac{\theta_{d+1}^{\leq}}{\theta^{<}} D_d^1 + D_{d+1}^1\}$. This event has positive probability and on this event, for all t,

$$\int_0^t \lambda^+_{Y,H,d}(Y_s, H_s)ds = 0 < \tilde{U}_d^1$$

and similarly

$$\begin{aligned}\int_0^t \lambda^-_{Y,H,d}(Y_s, H_s)ds \leq \frac{\theta_{d+1}^{<}}{\theta_d^{<}} \int_0^t [-\theta_d^{<} \partial_d U(X_s)]^+ ds + \int_0^t [-\theta_{d+1}^{<} \partial_{d+1} U(X_s)]^+ ds \leq \\ \leq \frac{\theta_{d+1^<}}{\theta^<} D_d^1 + D_{d+1}^1 < \tilde{D}_d^1.\end{aligned}$$

This means that on event B the d coordinate of H_t never changes and therefore $|Y_t| \to \infty$. This contradicts the induction hypothesis $\mathcal{P}_d$.

The third case we have to consider is when $\theta_{d+1}^{\leq} = \theta_d^{<} = 0$. This implies that $\theta_{d+1} = \theta^n$ and $\theta_d = -\theta^n$.

This time, we use a slightly different potential V', where we define

$$V'(y_1, .., y_d) = U(y_1, ..., y_{d-1}, -y_d, y_d). \tag{3.39}$$

We construct again the process (Y, H) such that it targets the measure that has V' as minus log-likelihood. As in the second case, we start the process from point $((x_1, ..., x_d), (\theta_1, ..., \theta_{d-1}, \theta_{d+1}))$. We also use the same coupling method. We can observe that

$$V'(y_1, .., y_d) \geq c \log(1 + \|(y_1, .., y_d, y_d)\|) - c' \geq c \log(1 + \|(y_1, .., y_d)\|) - c'$$

for some $c > d$ and $V \in C^3$ so (Y, H) must be non-evanescent by induction hypothesis.

As in the second case, if $t \leq \tau$, $\Theta_{d+1} = H_d$ until time t and, due to the way we coupled the processes, the coordinates $\{1, 2, ..., d-1, d+1\}$ of the process X are equal to the coordinates of the process Y until time t. As before, and due to the definition of V' in (3.39), we write

$$\int_0^t \lambda^+_{Y,H,d}(Y_s, H_s)ds = \int_0^t [H^{\leq}_d(s)\partial_d V'(Y_s)]^+ ds = \int_0^t [\theta^{\leq}_{d+1}\partial_d V'(Y_s)]^+ ds = 0.$$

At the same time, since $\theta_{d+1} = \theta^n$, $\theta_d = -\theta^n$ we have $\theta^{<}_{d+1} = -\theta^n = \theta^{\leq}_d$, therefore

$$\begin{aligned}\int_0^t \lambda^-_{Y,H,d}(Y_s, H_s)ds = \int_0^t [-H^{\leq}_d(s)\partial_d V'(Y_s)]^+ ds = \int_0^t [-\theta^{<}_{d+1}(-\partial_d U(X_s) + \partial_{d+1}U(X_s))]^+ ds \leq \\ \leq \int_0^t [-\theta^{<}_{d+1}(-\partial_d U(X_s))]^+ + [-\theta^{<}_{d+1}\partial_{d+1}U(X_s)]^+ ds = \\ = \int_0^t [\theta^{\leq}_d\ \partial_d U(X_s)]^+ ds + \int_0^t [-\theta^{<}_{d+1}\ \partial_{d+1}U(X_s)]^+ ds.\end{aligned}$$

Similar to the previous cases, consider the event $C = \{\Theta_d \text{ and } \Theta_{d+1} \text{ never change }\} \cap \{\tilde{D}^1_d > U^1_d + D^1_{d+1}\}$. This event has positive probability and on this event, for all t, a.s.

$$\int_0^t \lambda^+_{Y,H,d}(Y_s, H_s)ds = 0 < \tilde{U}^1_d$$

and similarly

$$\int_0^t \lambda^-_{Y,H,d}(Y_s, H_s)ds \leq \int_0^t [\theta^{\leq}_d\ \partial_d U(X_s)]^+ ds + \int_0^t [-\theta^{<}_{d+1}\ \partial_{d+1}U(X_s)]^+ ds \leq U^1_d + D^1_{d+1} < \tilde{D}^1_d.$$

This means that on event C the d coordinate of H_t never changes and therefore $|Y_t| \to \infty$. This contradicts the induction hypothesis $\mathcal{P}_d$.

The fourth case is that $\theta^{<}_{d+1} = 0$ and $\theta^{\leq}_d \neq 0$. Due to symmetry, this can be treated in the same way as case 2 (where $\theta^{<}_d \neq 0$ and $\theta^{\leq}_{d+1} = 0$). The difference is

that the d-dimensional process should start from $((x_1, ..., \theta_d); (\theta_1, ..., \theta_d))$.

The fifth and final case is that $\theta^{<}_{d+1} = 0$ and $\theta^{\leq}_{d} = 0$. This can be treated as case 3 (where $\theta^{<}_{d} = 0$ and $\theta^{\leq}_{d+1} = 0$), again the difference being that the d-dimensional process should start from $((x_1, ..., \theta_d); (\theta_1, ..., \theta_d))$. In this case, the potential V'' used to construct the d-dimensional process should, also, be

$$V''(y_1, ..., y_d) = U(y_1, ..., y_{d-1}, y_d, -y_d).$$

In all cases the d-dimensional process constructed in non-evanescent when starting from some specific point (different in each scenario). This contradicts the induction hypothesis $\mathcal{P}_d$ and concludes the proof. □

3.3.3 Proof of Theorem 3.3.3 (Ergodicity of MDCNZZ)

In this section we will use the reachability and the non-evanescence results, established in Sections 3.3.1 and 3.3.2 to prove Theorem 3.3.3 and establish convergence of the MDCNZZ to the invariant measure.
The following three Propositions are proved in manner very similar to Theorem 5 of [BRZ19].

Proposition 3.3.32. *If the Zig-Zag process is fully flippable and $U \in C^1$ then it is a T-process.*

Proof of Proposition 3.3.32. For any starting point $(x, \theta) \in E$ we can find an admissible path that switches all indices. From Continuous Component Lemma 3.3.21 and by covering $\mathbb{R}^d$ with countably many compact sets, we can find a family of open sets $(U_n)_{n\in\mathbb{N}}$ in E and $(V_n)_{n\in\mathbb{N}}$ in $\mathbb{R}^d$ and velocities $(\eta_n)_{n\in\mathbb{N}}$ and numbers $(t_n, \epsilon_n, c_n)_{n\in\mathbb{N}}$ so that

- Each $(x, \theta) \in E$ belongs to at least one and at most finitely many U_n's.
- For all $(x, \theta) \in U_n$ and $t \in [t_n, t_n + \epsilon_n]$ and functions f non-negative ,

$$\mathbb{E}_{x,\theta}[f(X_t, \Theta_t)] \geq c_n \int f(y, \eta_n) 1_{V_n}(y) dy \tag{3.40}$$

Define the lower bound Kernel by

$$K((x, \theta), A \times \{\eta\}) = \int 1_A(y) \cdot \max_{n:(x,\theta)\in U_n} \{c_n 1_{\eta_n=\eta} 1_{V_n}(y) \int_{t_n}^{t_n+\epsilon_n} \exp\{-t\} dt\} dy$$

Note that this maximum is well defined for all (x,θ) since any of them belongs to a finite number of U_n. Now, using Fubini where needed, we get for every $n : (x,\theta) \in U_n$

$$\mathbb{E}_{x,\theta}\left[\int_0^\infty \exp\{-t\}f(Z_t)dt\right] \geq \mathbb{E}_{x,\theta}\left[\int_{t_n}^{t_n+\epsilon_n} \exp\{-t\}f(X_t)dt\right] \geq$$
$$\geq \int_{t_n}^{t_n+\epsilon_n} \exp\{-t\}dt \cdot c_n \int 1_{V_n}(y)f(y,\eta_n)dy$$

This proves

$$\int_0^\infty \exp\{-t\}\mathbb{P}[Z_t \in A]dt \geq K((x,\theta),A).$$

Also $K((x,\theta),E) \geq c_n \int_{t_n}^{t_n+\epsilon_n} \exp\{-t\}dtLeb(V_n) > 0$ so $K((x,\theta),\cdot)$ is not trivial for any (x,θ).

Finally, fix $A \subset \mathbb{R}^d$ and take $x_j \xrightarrow{j\to\infty} x$ so $(x_j,\theta) \xrightarrow{j\to\infty} (x,\theta)$. The terms of x_j eventually lie in every U_n that contains (x,θ). Then, for large j $K((x_j,\theta),A\times\{\eta\}) \geq K((x,\theta),A\times\{\eta\})$ and therefore $K(\cdot,A\times\{\eta\})$ is lower semi continuous for all η. Since a finite sum of lower semi continuous functions remains lower semi continuous, we get for every $B \subset E$, $K(\cdot,B)$ lower semi continuous. This proves that MDCN Zig-Zag is a T-process. □

Proposition 3.3.33. *If the process has the property of reachability, $U \in C^1$ and there exists a probability invariant measure μ, then it is ϕ-irreducible, aperiodic and all compact sets are petite.*

Proof of Proposition 3.3.33. Suppose the process starts from (x,θ). Take $O \subset \mathbb{R}^d$ open and $y \in O$. Since for any velocity η $(x,\theta) \rightsquigarrow (y,\eta)$, from Continuity Lemma there exists c, t_0, ϵ and open V containing y so that $\mathbb{P}_{x,\theta}(X_t \in O, \Theta = \eta) \geq cLeb(O\cap V) > 0$ for all $t \in [t_0, t_0+\epsilon]$. Therefore if τ_O the first hitting time of O , $\mathbb{P}_{x,\theta}(\tau_O < \infty) = 1$ and the process is open set irreducible. As proven in Theorem 3.2 in [Twe95] this implies ϕ-irreducibility.

To prove aperiodicity, let (x,θ) any point. Since $(x,\theta) \rightsquigarrow (x,\theta)$, from Continuity Lemma, we can find neighbourhoods U', V' of x, $t_0, \epsilon > 0$ so that
$\mathbb{P}_{x',\theta}(X_t \in \cdot, \Theta_t = \theta) \geq cLeb(\cdot \cap V')$ for all $t \in [t_0, t_0+\epsilon]$ and $x' \in U'$. This means that U' is small. Consider the sets of the form $A_k = [kt_0, k(t_0+\epsilon)]$, $k \in \mathbb{N}$. Then $\inf A_{k+1} - \sup A_k = (k+1)t_0 - kt_0 - k\epsilon = t_0 - k\epsilon$ which will be negative for large k. Therefore for large k, the sets A_k will not leave any "gaps" between them and they will cover the whole interval $[T,+\infty)$ for some large T. Then for any $t \geq T$ there exists $s \in [t_0, t_0+\epsilon]$ and $k \in \mathbb{N}$ with $t = ks$ and then, for any $x' \in W = U' \cap V'$,

$$\mathbb{P}_{x',\theta}(X_t \in W, \Theta_t = \theta) \geq c'^k > 0.$$

Note that W is a small set as a subset of a small set and this proves strong aperiodicity.

In order to prove that all compact sets are petite, recall that in the proof of Lemma 3.3.29 we prove that

$$\mathbb{P}_\mu(\|X_t\| \to \infty) \xrightarrow{t\to\infty} 0.$$

More specifically, there exists $(x,\theta) \in E$ such that $\mathbb{P}_{x,\theta}(\lim_{t\to+\infty} \|X_t\| = +\infty) < 1$. From Theorem 4.1 in [MT93a] all compact sets are petite iff the process is T and ϕ-irreducible and the result follows. □

Proposition 3.3.34. *If the process is non-evanescent, has the property of reachability, $U \in C^1$ and there exists an invariant probability measure μ, then it is positive Harris recurrent and Ergodic.*

Proof of Proposition 3.3.34. If a process is T and ψ-irreducible, Harris irreducibility is equivalent to non-evanescence property (see Theorem 3.2 in [MT93a]. We have assumed non-evanescence and the other two properties are established by Propositions 3.3.32 and 3.3.33 and therefore Harris recurrence holds. Positivity holds by definition since the invariant measure is probability.

For ergodicity we will use that under positive Harris recurrence, it suffices to establish irreducibility for some skeleton chain in order to get convergence to equilibrium (see [MT93a] Theorem 6.1.).
Take (x,θ) any point and since $(x,\theta) \looparrowright (x,\theta)$ we can find ϵ, t_0, c and open neighbourhoods of x, U_0, V_0 so that for all $x' \in U$ and $t \in [t_0, t_0+\epsilon]$,

$$\mathbb{P}_{x',\theta}\left(X_t \in \cdot, \Theta_t = \theta\right) \geq cLeb(\cdot \cap V).$$

Let $(y,\eta),(y',\eta')$ two arbitrary points. Since $(y,\eta) \looparrowright (x,\theta) \looparrowright (x,\theta) \looparrowright (y',\eta')$, we find t_1, t_2, c_1, c_2 and open sets V_1, V_2, U_1 with $x \in U_2, x \in V_1, y' \in V_2$ so that

- $\mathbb{P}_{y,\eta}\left(X_{t_1} \in \cdot, \Theta_{t_1} = \theta\right) \geq c_1 Leb(\cdot \cap V_1)$
- $\mathbb{P}_{x',\theta}\left(X_{t_2}\cdot, \Theta_{t_2} = \eta'\right) \geq c_2 Leb(\cdot \cap V_2)$, $\forall x' \in U_2$.

Then, if O an open set containing y' and for any $t \in [t_0+t_1+t_2, t_0+t_1+t_2+\epsilon]$ by the Strong Markov Property we get

$$\begin{aligned}&\mathbb{P}_{y,\eta}\left(X_t \in O, \Theta_t = \eta'\right) \geq \\ &\geq \mathbb{P}_{y,\eta}\left(\Theta_{t_1} = \Theta_{t-t_2} = \theta, \Theta_t = \eta', X_{t_1} \in U_0 \cap V_1, X_{t-t_2} \in U_2 \cap V_0, X_t \in O\right)\end{aligned}$$

$$\geq \mathbb{P}_{y,\eta}\left(\Theta_{t_1} = \Theta_{t-t_2} = \theta, X_{t_1} \in U_0 \cap V_1, X_{t-t_2} \in U_2 \cap V_0\right) c_2 Leb(O \cap V_2) \geq$$
$$\geq cc_1c_2 Leb(U \cap V_1) Leb(V \cap U_2) Leb(O \cap V_2) > 0$$

This lower bound is uniform in $t \in [t_0 + t_1 + t_2, t_0 + t_1 + t_2 + \epsilon]$ and since this interval contains one multiplier of ϵ, the ϵ-chain is irreducible. This establishes ergodicity. □

We can now conclude with the proof of the main theorem of the section.

Proof of Theorem 3.3.3. Assume that assumptions 3.3.1 and 3.3.2 hold. Since assumption 3.3.2 implies assumption 3.3.5, from Theorem 3.3.6 we get that the process has the property of reachability. Assumption 3.3.2 further proves that μ is a probability measure. Non-evanescence is implied by Proposition 3.3.26. The result follows from Proposition 3.3.34. □

3.4 Geometric Ergodicity of MDCNZZ in Light Tails

In this section, we prove that, under some further growth conditions on the derivatives of U, a certain continuously differentiable function $V : \mathbb{R}^d \times \Theta \to [1, +\infty)$ is Lyapunov, i.e.

$$\mathcal{L}V(x,\theta) \leq -cV(x,\theta) + b1_{x \in C} \tag{3.41}$$

for some $c, b > 0$ and some petite set C, where recall that $\mathcal{L}$ is the weak generator of the process, given by

$$\mathcal{L}V(x,\theta) = \sum_{i=1}^{d} \theta_i \partial_i V(x,\theta) + \lambda_i^+(x,\theta)\left(V(x, F_i^+[\theta]) - V(x,\theta)\right) + \\ + \lambda_i^-(x,\theta)\left(V(x, F_i^-[\theta]) - V(x,\theta)\right). \tag{3.42}$$

Then we can use Theorem 2.1.10 to conclude that the process is geometrically ergodic, meaning that

$$||P_{x,\theta}[(X_t, \Theta_t) \in \cdot] - \pi(\cdot)||_{TV} \leq M(x,\theta)\exp\{-ct\} \tag{3.43}$$

for some $c, M > 0$. In our case, since every compact is petite due to Proposition 3.3.33, the set C will be chosen as a compact set and since V and $\mathcal{L}V$ will be bounded on compact sets, it will suffice to prove that $\mathcal{L}V < -cV$ outside a compact set. The function we pick is the same as in the original ergodicity paper [BRZ19], for the same heuristic reasons that are explained there. More specifically, we define

the function

$$V(x,\theta) = \exp\{aU(x) + \sum_{i=1}^{d} \phi(\theta_i \partial_i U(x)\} \tag{3.44}$$

where

$$\phi(s) = \frac{1}{2} sign(s) log(1 + \delta|s|).$$

For this section we will make the following assumption.

Assumption 3.4.1. *Assume that*

$$\lim_{\|x\|\to\infty} \frac{\|HessU(x)\|_1}{\|\nabla U(x)\|_1} = 0 \ , \ \lim_{\|x\|\to\infty} \frac{\|\nabla U(x)\|_1}{\|U(x)\|_1} = 0 \ , \ \lim_{\|x\|\to\infty} \|\nabla U(x)\|_1 = +\infty$$

where for a vector or a matrix A, we write $\|A\|_1$ to denote the sum of absolute values of all the entries of A.

The main theorem is the following.

Theorem 3.4.2 (Geometric Ergodicity of MDCNZZ)**.** *Assume that Assumption 3.4.1 holds, that $U \in C^3$ and has a non-degenerate local minimum. Then the MDCN Zig-Zag is geometrically ergodic, as defined in (3.43).*

Proof of Theorem 3.4.2. First we introduce some notation and for $\theta_0 = \theta^j \in \Theta_0$ write $\theta_0^+ = \theta^{j+1}$ if $j < n$ and $\theta_0^+ = \theta_0$ if $j = n$ and write $\theta_0^- = \theta^{j-1}$ if $j > -n$ and $\theta_0^- = \theta_0$ if $j = -n$. We also write $\theta_* = \min\{\theta^{j+1} - \theta^j, j = 1, 2, ..., n-1\}$ and $\theta^* = \max\{\theta^{j+1} - \theta^j, j = 1, 2, ..., n-1\}$ the smallest and largest difference between consecutive values of Θ_0. Given $\theta = (\theta_1, ..., \theta_d)$, such that $\theta_i = \theta^j$, we write $F_i^+[\theta] = (\theta_1, ..., \theta_{i-1}, \theta^{j+1}, \theta_{i+1}, ..., \theta_d)$ and $F_i^-[\theta] = (\theta_1, ..., \theta_{i-1}, \theta^{j-1}, \theta_{i+1}, ..., \theta_d)$. Using equation (3.4) for the generator of the process, along with the fact that $\phi \in C^1$ with $\phi'(s) = \frac{\delta/2}{1+\delta|s|}$, we see that for all $(x, \theta) \in E$,

$$\begin{aligned}
\mathcal{L}V(x,\theta) &= \sum_{i=1}^{d} \theta_i \partial_i V(x,\theta) + \lambda_i^+(x,\theta)\left(V(x, F_i^+[\theta]) - V(x,\theta)\right) + \\
&+ \sum_{i=1}^{d} \lambda_i^-(x,\theta)\left(V(x, F_i^-[\theta]) - V(x,\theta)\right) = \\
&= \sum_{i=1}^{d} \left[\theta_i V(x,\theta)\left(a\partial_i U(x) + \sum_{j=1}^{d} \phi'(\theta_j \partial_j U(x))\theta_j \partial_i \partial_j U(x)\right)\right] + \\
&+ \sum_{i=1}^{d} \left[\lambda_i^+(x,\theta)\left(V(x, F_i^+[\theta]) - V(x,\theta)\right) + \lambda_i^-(x,\theta)\left(V(x, F_i^-[\theta]) - V(x,\theta)\right)\right]
\end{aligned}$$

Note that from the definition of V in (3.44), for all i,

$$V(x, F_i^+[\theta]) - V(x,\theta) = V(x,\theta)\left[\exp\{\phi(\theta_i^+\partial_i U(x)) - \phi(\theta_i\partial_i U(x))\} - 1\right]$$

and

$$V(x, F_i^-[\theta]) - V(x,\theta) = V(x,\theta)\left[\exp\{\phi(\theta_i^-\partial_i U(x)) - \phi(\theta_i\partial_i U(x))\} - 1\right].$$

This means that

$$\begin{aligned}\frac{\mathcal{L}V(x,\theta)}{V(x,\theta)} = &\sum_{i,j=1}^{d}\frac{\delta/2}{1+\delta|\partial_j U(x)\theta_j|}\theta_i\theta_j\partial_{i,j}U(x) + \sum_{i=1}^{d}\theta_i a\partial_i U(x) + \\ &+\sum_{i=1}^{d}\lambda^+(x,\theta)\left[\exp\{\phi(\partial_i U(x)\theta_i^+) - \phi(\partial_i U(x)\theta_i)\} - 1\right] \\ &+\sum_{i=1}^{d}\lambda^-(x,\theta)\left[\exp\{\phi(\partial_i U(x)\theta_i^-) - \phi(\partial_i U(x)\theta_i)\} - 1\right]\end{aligned}$$

Let's call the four parts of the RHS sum A, B, C, D respectively for convenience. Our goal is to prove that this sum is bounded below 0 outside some compact set. For the first term of the sum we can bound

$$A \leq \sum_{i,j=1}^{d}\frac{\delta}{2}(\theta^n)^2|\partial_{i,j}U(x)| \tag{3.45}$$

which will be enough since the rest of the terms involve the first derivative which grows faster. For the rest of the terms we need to consider different cases. From now on, we fix $i \in \{1, .., d\}$ and let's start by assuming that $\theta_i \neq \pm\theta^1$.

Case 1: Assume that $\partial_i U(x) > 0, \theta_i > 0$.

Assume further, for now, that $\theta_i \neq \theta^n$. Then,

$$B = a\theta_i\partial_i U(x) \leq a\theta^n|\partial_i U(x)|.$$

Also,

$$\exp\{\phi(\theta_i^+\partial_i U(x)\partial_i U(x)) - \phi(\theta_i\partial_i U(x))\} - 1 = \sqrt{1 + \frac{\delta(\theta_i^+ - \theta_i)|\partial_i U(x)|}{1+\delta\theta_i|\partial_i U(x)|}} - 1 \leq$$

$$\leq 1 + \frac{\delta\theta^*|\partial_i U(x)|}{1+\delta\theta_i|\partial_i U(x)|} - 1 \leq \frac{\theta^*}{\theta^1}$$

and since in the case where $\partial_i U(x) > 0$, we have$\lambda^+(x,\theta) = \gamma(x)$, we get

$$C \leq \bar{\gamma}\frac{\theta^*}{\theta^1}.$$

Finally,

$$\exp\{\phi(\theta_i^-\partial_i U(x) - \phi(\theta_i\partial_i U(x)\} - 1 = \sqrt{1 - \frac{\delta(\theta_i - \theta_i^-)|\partial_i U(x)|}{1 + \delta|\theta_i||\partial_i U(x)|}} - 1 \leq$$

$$\leq \sqrt{1 - \frac{\delta\theta_*|\partial_i U(x)|}{1 + \delta\theta^n|\partial_i U(x)|}} - 1$$

which is negative and since $\partial_i U(x) < 0$ we have

$$\lambda^-(x,\theta) \geq [(\theta_i)^<\partial_i U(x)]^- \geq \theta^n|\partial_i U(x)|.$$

Therefore,

$$D \leq -\theta^n|\partial_i U(x)|\left(1 - \sqrt{1 - \frac{\delta\theta_*|\partial_i U(x)|}{1 + \delta\theta^n|\partial_i U(x)|}}\right)$$

So, overall

$$B + C + D \leq \bar{\gamma}\frac{\theta^*}{\theta^1} + |\partial_i U(x)|\theta^n\left(a - 1 + \sqrt{1 - \frac{\delta\theta_*|\partial_i U(x)|}{1 + \delta\theta^n|\partial_i U(x)|}}\right) \tag{3.46}$$

Note here that similar calculations can be done in the case $\theta_i = \theta^n$. The only difference is that, in that case, $\lambda_i^+(x,\theta) = 0$ and the constant $\bar{\gamma}\theta^*/\theta^1$ won't appear in the upper bound of (3.46). Since this constant only increases the upper bound, we can include here the case $\theta_i = \theta^n$ as well.

Case 2: Assume that $\partial_i U(x) < 0, \theta_i > 0$. Assume again, for now, that $\theta_i \neq \theta^n$. We have

$$B = -\theta_i a|\partial_i U(x)| \leq -\theta_1 a|\partial_U(x)|.$$

After a few calculations (and using the fact that $\theta_i\partial_i U(x) < 0$) we get

$$\exp\{\phi(\theta_i^-\partial_i U(x)) - \phi(\theta_i\partial_i U(x))\} - 1 = \sqrt{1 + \frac{\delta(\theta_i - \theta_i^-)|\partial_i U(x)|}{1 + \delta\theta_i^-|\partial_i U(x)|}} - 1 \leq \frac{\theta^*}{\theta^1}$$

and since $\lambda_i^-(x,\theta) = \gamma(x)$ we get $D \leq \bar{\gamma}\frac{\theta^*}{\theta^1}$.

Finally,

$$\exp\{\phi(\theta_i^+\partial_i U(x)) - \phi(\theta_i\partial_i U(x))\} - 1 = \sqrt{1 - \frac{\delta(\theta_i^+ - \theta_i)|\partial_i U(x)|}{1 + \delta\theta_i^+|\partial_i U(x)|}} - 1 \leq 0$$

so, $C \leq 0$. Overall,

$$B + C + D \leq \bar{\gamma}\frac{\theta^*}{\theta^1} - a\theta_1|\partial_i U(x)|. \tag{3.47}$$

Note as well that in the case $\theta_i = \theta^n$ we achieve $C = 0$ and the bound does not change.

Case 3: $\partial_i U(x) > 0, \theta_i < 0$. This case is treated in the same way as Case 2. The bound of the C quantity here is the bound of the D quantity in Case 2 and vice versa.

Case 4: $\partial_i U(x) < 0, \theta_i < 0$. This is, also, treated in the same way as Case 1. The bound of the C quantity here is the bound of the D quantity in Case 1 and the other way around.

Now let us consider the cases where $\theta_i = \pm\theta^1$. We will also have to consider four different cases.

Case 1: $\partial_i U(x) > 0, \theta_i > 0$. We have,

$$B \leq \theta^1 a|\partial_i U(x)|$$

Also,

$$\phi(\theta^2\partial_i U(x)) - \phi(\theta^1\partial_i U(x)) = \sqrt{1 + \frac{\delta(\theta^2 - \theta^1)|\partial_i U(x)|}{1 + \delta\theta_1|\partial_i U(x)|}} - 1 \leq \frac{\theta^*}{\theta^1}$$

and $\lambda_i^+(x,\theta) = \gamma(x)$ so $C \leq \bar{\gamma}\frac{\theta^*}{\theta^1}$. Also,

$$\exp\{\phi(-\theta^1\partial_i U(x)) - \phi(\theta^1\partial_i U(x))\} = \frac{1}{1 + \delta\theta^1|\partial_i U(x)|} - 1$$

which is negative and therefore, using $\lambda_i^-(x,\theta) = |\theta_1^<||\partial_i U(x)| \geq \theta^n|\partial_i U(x)|$, we get

$$D \leq -\theta^n|\partial_i U(x)|(1 - \frac{1}{1 + \delta\theta^1|\partial_i U(x)|})$$

and overall,

$$B + C + D \leq \bar{\gamma}\frac{\theta^*}{\theta^1} + \theta^n|\partial_i U(x)|\left(a - 1 + \frac{1}{1 + \delta\theta^1|\partial_i U(x)|}\right). \tag{3.48}$$

Case 2: $\partial_i U(x) < 0, \theta^{(i)} = \theta_1$ We have $B = -a\theta_1|\partial_i U(x)|$. Also,

$$\exp\{\phi(\theta^2\partial_i U(x)) - \phi(\theta^1\partial_i U(x))\} - 1 = \sqrt{1 - \frac{\delta(\theta^2 - \theta^1)|\partial_i U(x)|}{1 + \delta\theta^2|\partial_i U(x)|}} - 1 \leq 0$$

so $C \leq 0$. Also,

$$\exp\{\phi(-\theta^1\partial_i U(x)) - \phi(\theta^1\partial_i U(x))\} - 1 = \delta\theta^1|\partial_i U(x)|$$

and since $\lambda_i^-(x,\theta) = \gamma(x)$,

$$D \leq \bar{\gamma}\delta\theta^1|\partial_i U(x)|$$

Overall,

$$B + C + D \leq |\partial_i U(x)|(\bar{\gamma}\theta^1\delta - a\theta^1) \tag{3.49}$$

Case 3: $\partial_i U(x) > 0, \theta^{(i)} = -\theta_1$. This gives the same upper bound as Case 2. Quantity C in this case is bounded as quantity D in Case 2 and vice versa.

Case 4: $\partial_i U(x) < 0, \theta^{(i)} = -\theta_1$. This gives the same upper bound as Case 1. Quantity C in this case is bounded as quantity D in Case 1 and vice versa.

Overall, for all x, θ

$$B + C + D \leq \bar{\gamma}\frac{\theta^*}{\theta^1} + |\partial_i U(x)|K(|\partial_i U(x)|) \tag{3.50}$$

where,

$$K(y) = \max\left\{\theta^n\left(a - 1 + \sqrt{1 - \frac{\delta\theta_* y}{1 + \delta\theta^n y}}\right), -a\theta^1, \theta^n\left(a - 1 + \frac{1}{1 + \delta\theta^1 y}\right), \theta^1\left(\bar{\gamma}\delta - a\right)\right\}. \tag{3.51}$$

Let us pick $a > \bar{\gamma}\delta$. We observe that $K : [0, +\infty) \to \mathbb{R}$ is non-increasing. Note as well that for $0 < a < 1 - \sqrt{1 - \frac{\theta_*}{\theta^n}}$,

$$\lim_{y \to +\infty} K(y) < 0. \tag{3.52}$$

To see this consider all four terms inside the maximum. The second and the fourth are negative constants. The third converges to $\theta^n(a-1) < 0$ as $y \to +\infty$ since we have assumed that $a \in (0,1)$. The first term converges to $\theta^n(a - 1 + \sqrt{1 - \frac{\theta_*}{\theta^n}}) < 0$.

Pick $a \in (0, 1 - \sqrt{1 - \frac{\theta_*}{\theta^n}})$ and $\delta < \frac{a}{\bar{\gamma}}$. Let us pick a c such that $K(c) = 0$ and $k \in \mathbb{N}$ such that $K(kc) < 0$. We set $C = \{x \in \mathbb{R}^d : ||\nabla U(x)||_1 \leq dkc\}$ as our petite set. Note that we know that C is petite since it is compact and due to Proposition 3.3.33. Then, for any $x \notin C$, there exists a coordinate i such that

$|\partial_i U(x)| > kc$. In fact take this i to be the $argmax_i|\partial_i U(x)|$ and assume without loss of generality that $i = 1$. This further means that if $\|\nabla U(x)\|_1 = \sum_{i=1}^d |\partial_i U(x)|$, then $|\partial_1 U(x)| \geq \|\nabla U(x)\|_1/d$.

Now, if we view the quantities A, B, C, D as functions of the coordinate i as well, we have

$$\frac{\mathcal{L}V(x,\theta)}{V(x,\theta)} \leq \sum_{i=2}^{d}(B+C+D)_i + (B+C+D)_1 + \sum_{i,j=1}^{d}(\theta^n)^2\delta/2|\partial_i\partial_j U(x)| \quad (3.53)$$

Using (3.50), we see that for $\sum_{i=2}^d (B+C+D)_i$ to be maximized we must have $(B+C+D)_i > 0$ for all $i > 1$, in which case for all $i > 1$, $K(|\partial_i U(x)|) > 0$. This means that $|\partial_i U(x)| < c$ for all $i > 1$. In that case, since K is non-increasing and using (3.53) and that $|\partial_1 U(x)| \geq \|\nabla U(x)\|_1/d$,

$$\begin{aligned}\frac{\mathcal{L}V(x,\theta)}{V(x,\theta)} \leq& (d-1)cK(0) + d\frac{\theta^*}{\theta^1}\bar{\gamma} + K(kc)|\partial_1 U(x)| + \sum_{i,j=1}^{d}(\theta^n)^2\delta/2|\partial_i\partial_j U(x)| \leq \\ &\leq M + \sum_{i,j=1}^{d}(\theta^n)^2\delta/2|\partial_i\partial_j U(x)| + K(kc)\frac{\|\nabla U(x)\|_1}{d} \leq \\ &\leq (\theta^n)^2\delta/2\|HessU(x)\|_1 - h\|\nabla U(x)\|_1 + M\end{aligned}$$

for some constants $h = \frac{|K(kc)|}{d} > 0$ and $M = (d-1)cK(0) + d\frac{\theta^*}{\theta^1}\bar{\gamma} > 0$. Now, since $K(kc) < 0$ and $\lim_{|x|\to\infty} \frac{||HessU(x)||_1}{||\nabla U(x)||_1} = 0$, if we increase the petite set C appropriately, we get $\frac{\mathcal{L}V(x,\theta)}{V(x,\theta)} < -c$ for some $c > 0$ and for all $x \notin C$. Since C is bounded and V and $\mathcal{L}V$ bounded on C we get the result. □

A direct consequence of Theorem 2.1.11 (Theorem 4.4 in [GM96]) is a Central Limit Theorem.

Theorem 3.4.3 (CLT for the MDCNZZ)**.** *Suppose that Assumption 3.4.1 holds, that $U \in C^3$ and has a non-degenerate local minimum. Let $g : E \to \mathbb{R}$ such that there exists an $\epsilon > 0$ and a compact set C such that for all $x \notin C$*

$$g(x,\theta) \leq \exp\left\{\frac{1}{2}\left(1-\epsilon-\sqrt{1-\frac{\theta_*}{\theta^n}}U(x)\right)\right\}.$$

Let $(Z_t)_{t\geq 0}$ be a MDCN Zig-Zag targeting μ and for any $n \in \mathbb{N}$ define the function

$Z_n : [0,1] \to \mathbb{R}$ *by*

$$Z_n(t) = \frac{1}{\sqrt{n}} \int_0^{nt} g(X_s, \Theta_s) - \mu(g) ds.$$

Then there exists a constant $\gamma_g^2 = 2\int_E \phi(x)\left(g(x) - \pi(g)\right) dx \in [0, +\infty)$ *and* Z_n *converges weakly in the Skorokhod topology (see [Bil99]) to* $\gamma_g^2 B$*, where* B *is a standard Brownian motion on* $[0,1]$*.*

Proof of Theorem 3.4.3. We have that there exists an $a \in (1, 1 - \sqrt{1 - \frac{\theta_*}{\theta^n}})$ such that
$g^2(x,\theta) \leq MV(x,\theta)$ for all $(x,\theta) \in E$, where V as in (3.44) for this value of a and any δ small enough and for some constant M. From the proof of Theorem 3.4.2 we get that if δ is small enough, MV satisfies the drift condition (3.41) and the result follows from Theorem 2.1.11. □

3.5 Simulations

3.5.1 Two Dimensional Targets

In this section we present some simulations concerning the performance of the Multi-Directional Closest Neighbour Zig-Zag on two dimensional target distributions. We consider three different categories of target distributions.

The first category is two-dimensional normal distributions, i.e. where

$$U(x,y) = \left\langle (x\ y), A^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \right\rangle \tag{3.54}$$

for some positive definite symmetric matrix A, which serves as the covariance matrix. We will present some results of the algorithms tested against three 'sets' of positively correlated Gaussians with small, medium and large correlations. In each set we study three different Gaussians, with covariance matrices that correspond to three different principal eigenvectors, i.e. eigenvectors of the largest eigenvalue. For the first one, the principal eigenvector is $(1,1)$, for the second $(1,2)$ and for the third $(3,1)$. Our understanding is that the process would perform better if it spends more time moving in parallel to the principal eigenvector, which justifies the introduction of multi-directions. This is not reflected in the results involving Gaussians with low correlations, but becomes more clear on medium correlations. For every target distribution we use four different algorithms, namely one original Zig-Zag and three MDCNZZ with allowed velocities $\{-2,-1,1,2\}^2$, $\{-3,-1,1,3\}^2$ and $\{-3,-2,-1,1,2,3\}^2$ respectively. For all the algorithms, we set the refresh

rate to be $\gamma \equiv 0$. For each distribution and each algorithm we have simulated 25 independent realisations, starting from over-dispersed positions until time $T = 10^4$ and we use δ-skeletons of these processes as estimators, for $\delta = 0.1$. We present the average (and in a parenthesis the standard deviation) of Effective Sample Size (ESS) of these estimators over these 25 chain realisations, along with ESS per number of switches of direction for these processes. ESS per switch will be one of our main criteria for the performance of the algorithm. This is because in an ideal setting, the Poisson thinning used to sample the switching points wouldn't reject any proposed switch and the number of actual switches would be the number of gradient log-likelihood evaluations. Since ESS per gradient log-likelihood evaluations is a typical measure of efficiency, ESS per Switch can be seen as an upper bound for the efficiency of the algorithm. We decide not to use ESS per Likelihood Evaluation as our measure of efficiency, since this heavily relies on the code itself and on the ways the programmer uses Poisson thinning. On the other hand, ESS per Switch only depends on the properties of the algorithm. In our case, the Poisson thinning is carried out using constant bounding rates, which might not be optimal, especially given that the targets are Gaussians, for which the gradient of the minus log-likelihood grows linearly. We note, however, that even in our code, the processes which tend to have the largest ESS per Switch, also tend to have the largest ESS per Likelihood Evaluation. In order to evaluate the two dimensional ESS, we use the library `mcmcse` of `R` language. Furthermore, we ran a Gelman-Rubin diagnostic for all these algorithms and they all gave estimators less than 1.01. Therefore, to provide a performance measure, we calculated the average number of switches needed for the Gelman-Rubin estimator to take a value less than 1.01 and remain less than 1.01 in the future. This can serve as a proxy of how many switches of direction are needed for an algorithm to converge and consequently as a proxy of how much computational power the algorithm needs until it converges. In order to use a two-dimensional Gelman-Rubin test, library `coda` from `R` language was used.

Aside from these 25 realisations of each algorithm, we also ran one extra chain realisation for each algorithm and each target distribution, where we stop the algorithm when the number of switches becomes $N = 10^4$. We use the empirical measure to estimate the probability the target density assigns to squares centred around $(0, 0)$. We do this for squares of the form $[-l, l]^2$ for various values of l and we compare the algorithm's estimation with the actual probability. The actual probability is estimated to a very small numerical error using the library `mvtnorm` of `R`. Since, in our analysis, the number of switches serves as a proxy for the computational cost of running the algorithm, in this case we fix the number of

switches for each algorithm so that we can directly compare how good approximation of the probabilities each algorithm provides. In every table we indicate with bold letters the best performance for every criterion.

For the first set of Gaussian distributions, we consider low correlations between the two coordinates. We set $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and we consider three different cases for the covariance matrix:

$$(1)\,.\ a = c = 2, b = 1,\ (2)\,.\ a = 3, b = 2, c = 6,\ (3)\,.\ a = 12, b = 3, c = 4, \tag{3.55}$$

each one with corresponding principal eigenvector $(1,1),(1,2),(3,1)$ respectively and eigenvalue pairs $(3,1),(7,2),(13,3)$ respectively. The results for these distributions are presented on Tables 3.1 and 3.2. These results indicate that the simple Zig-Zag process performs better than the other algorithms in terms of ESS per switch, while all the algorithms perform similarly well in estimating the probabilities of the squares. This contradicts our guess that the algorithm should spend more time moving in the direction of the principal eigenvector, however, it could be the case that due to low correlations the process explores the space efficiently using only directions $\{-1,1\}^2$.

For the second set of Gaussian distributions, we consider medium correlations and once again study three different cases for the covariance matrix:

$$(1)\,.\ a = c = 51, b = 50,\ (2)\,.\ a = 41, b = 40, c = 101,\ (3)\,.\ a = 111, b = 30, c = 31. \tag{3.56}$$

Each of these targets has corresponding principal eigenvector $(1,1),(1,2)$ and $(3,1)$ respectively and eigenvalue pair $(121,21)$ in all three cases. The results for these distributions are presented on Tables 3.3 and 3.4. Here we can observe that the ESS/Switch tends to become better when we include the direction of the principal eigenvector as a direction. It is interesting that using a higher number of directions $\{\pm 1,2,3\}$ is even more efficient when the principal eigenvector is $(1,2)$ or $(3,1)$. It is also interesting to note the really high standard deviation in some of the algorithm (mainly when we use $\{\pm 1,2,3\}$). In fact, in these cases 24 out of the 25 realisations of the algorithm gave consistent ESS and one copy gave a surprisingly high ESS. This could indicate that the algorithm is not completely stable when we terminate it at time $T = 10^4$, or it could mean that we should use a different routine for calculating ESS. Finally, in Figures 3.3 and 3.4 we present the plots describing the trajectories for the original Zig-Zag and the MDCNZZ(1,2) and the QQplots of their first coordinates when targeting the Gaussian with principal eigenvector $(1,2)$.

	Gaussian $(1,1)$		
Algorithm	ESS(Standard Deviation)	ESS/Switch	Gelman-Rubin Switches
ZZ	10571.6 (9385.8)	**1.6165**	**400**
MDCNZZ({±1 2})	9599.0 (4070.21)	0.4220	600
MDCNZZ({±1 3})	13927.9 (7373.99)	0.4285	1000
MDCNZZ({±1 2 3})	17753.6 (11153.59)	0.3696	1000
	Gaussian $(1,2)$		
Algorithm	ESS(Standard Deviation)	ESS/Switch	Gelman-Rubin Switches
ZZ	3096.0 (377.20)	**0.6903**	500
MDCNZZ({±1 2})	10575.9 (1558.87)	0.6766	**400**
MDCNZZ({±1 3})	13225.4 (2801.81)	0.5994	700
MDCNZZ({±1 2 3})	19667.7 (2568.55)	0.6011	700
	Gaussian $(3,1)$		
Algorithm	ESS(Standard Deviation)	ESS/Switch	Gelman-Rubin Switches
ZZ	6069.9 (5209.53)	**1.7417**	**400**
MDCNZZ({±1 2})	5797.9 (2521.62)	0.4715	900
MDCNZZ({±1 3})	8288.6 (4576.41)	0.4765	1100
MDCNZZ({±1 2 3})	11200.53 (7159.04)	0.4370	1600

Table 3.1: *Three two-dimensional Gaussian distributions with* **low** *correlations and covariance matrices given by (3.55). Each process is labelled by its principal eigenvector and is tested against four different algorithms. For each algorithm, 25 independent realisations were simulated, each one until time* $T = 10^4$ *and the estimator is constructed using the* δ*-skeleton of the process for* $\delta = 0.1$*. We present ESS with standard deviations in a parenthesis, ESS per switch and then number of switches needed for the Gelman-Rubin estimator to take a value less than* 1.01 *and remain less than* 1.01 *until the process terminates.The best performance is indicated with bold letters.*

Gaussian $(1,1)$				
Algorithm	$l=1$	$l=2$	$l=3$	$l=5$
ZZ	0.3096	0.7391	0.9416	0.9993
MDCNZZ($\{\pm 1\ 2\}$)	0.2933	0.7270	0.9356	0.9988
MDCNZZ($\{\pm 1\ 3\}$)	0.3013	0.7354	**0.9379**	**0.9992**
MDCNZZ($\{\pm 1\ 2\ 3\}$)	**0.2972**	**0.7321**	0.9412	0.9995
Actual Value	0.2987	0.7322	0.9376	0.9992
Gaussian $(1,2)$				
Algorithm	$l=1$	$l=3$	$l=5$	$l=9$
ZZ	0.1563	0.7341	0.9569	0.9996
MDCNZZ($\{\pm 1\ 2\}$)	0.1544	0.7373	0.9599	1
MDCNZZ($\{\pm 1\ 3\}$)	**0.1538**	**0.7327**	**0.9549**	**0.9998**
MDCNZZ($\{\pm 1\ 2\ 3\}$)	0.1580	0.7403	0.9575	1
Actual Value	0.1534	0.7302	0.9559	0.9998
Gaussian $(3,1)$				
Algorithm	$l=2$	$l=4$	$l=7$	$l=10$
ZZ	**0.3135**	**0.7258**	0.9547	0.9942
MDCNZZ($\{\pm 1\ 2\}$)	0.3100	0.7148	0.9547	0.9947
MDCNZZ($\{\pm 1\ 3\}$)	0.3269	0.7369	**0.9549**	0.9945
MDCNZZ($\{\pm 1\ 2\ 3\}$)	0.3251	0.7307	0.9500	**0.9958**
Actual Value	0.3172	0.7269	0.9564	0.9961

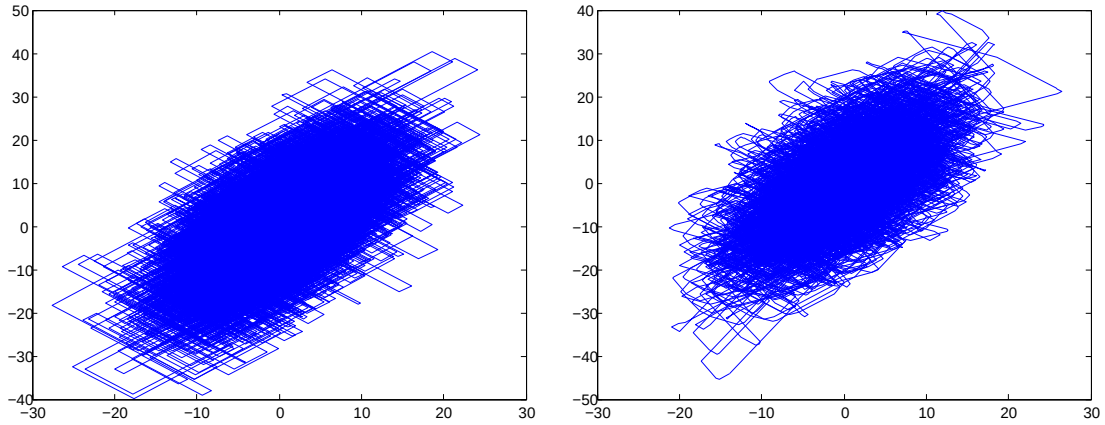
Table 3.2: *Three two-dimensional Gaussian distributions with* **low** *correlations and covariance matrices given by (3.55). Each process is labelled by its principal eigenvector and is tested against four different algorithms. For each algorithm we present estimates of the probabilities assigned to squares of the form* $[-l,l]\times[-l,l]$ *under the target Gaussian distribution, for various values of* l*. Each algorithm is simulated until* $N=10^4$ *switches of direction have occurred. The best performance is indicated with bold letters.*

Gaussian $(1,1)$			
Algorithm	ESS(Standard Deviation)	ESS/Switch	Gelman-Rubin Switches
ZZ	942.30 (479.87)	**0.1694**	**4800**
MDCNZZ($\{\pm 1\ 2\}$)	2254.28 (797.91)	0.1129	9684
MDCNZZ($\{\pm 1\ 3\}$)	3604.82 (1140.27)	0.1280	10610
MDCNZZ($\{\pm 1\ 2\ 3\}$)	4461.39 (1579.77)	0.1039	8065
Gaussian $(1,2)$			
Algorithm	ESS(Standard Deviation)	ESS/Switch	Gelman-Rubin Switches
ZZ	708.53 (66.63)	0.5483	**600**
MDCNZZ($\{\pm 1\ 2\}$)	3777.03 (1269.17)	0.8263	900
MDCNZZ($\{\pm 1\ 3\}$)	4813.87 (950.94)	0.7380	700
MDCNZZ($\{\pm 1\ 2\ 3\}$)	16138.57 (30523.83)	**1.7246**	800
Gaussian $(3,1)$			
Algorithm	ESS(Standard Deviation)	ESS/Switch	Gelman-Rubin Switches
ZZ	500.06 (21.58)	0.6195	**500**
MDCNZZ($\{\pm 1\ 2\}$)	1370.42 (90.02)	0.8537	700
MDCNZZ($\{\pm 1\ 3\}$)	6654.43 (8627.10)	1.0281	**500**
MDCNZZ($\{\pm 1\ 2\ 3\}$)	17282.59 (57456.82)	**1.8552**	800

Table 3.3: *Three two-dimensional Gaussian distributions with* **medium** *correlations and covariance matrices given by (3.56). Each process is labelled by its principal eigenvector and is tested against four different algorithms. For each algorithm, 25 independent realisations of the process were simulated, each one until time* $T = 10^4$ *and the estimator is constructed using the* δ*-skeleton of the process for* $\delta = 0.1$*. We present ESS with standard deviations, ESS per switch and then number of switches needed for the Gelman-Rubin estimator to take a value less than* 1.01 *and remain less than* 1.01 *until the process terminates. The best performance is indicated with bold letters.*

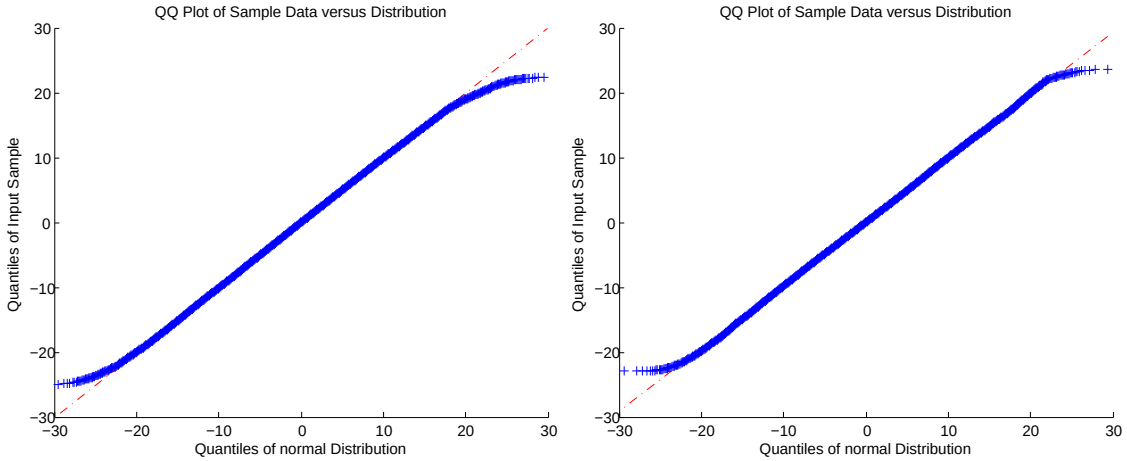
Gaussian $(1,1)$				
Algorithm	$l=10$	$l=15$	$l=20$	$l=25$
ZZ	0.8316	0.9685	0.9971	1
MDCNZZ($\{\pm 1\ 2\}$)	0.8316	0.9654	**0.9952**	0.9992
MDCNZZ($\{\pm 1\ 3\}$)	**0.8194**	**0.9578**	0.9955	**0.9993**
MDCNZZ($\{\pm 1\ 2\ 3\}$)	0.8028	0.9558	0.9907	1
Actual Value	0.8150	0.9574	0.9937	0.9994
Gaussian $(1,2)$				
Algorithm	$l=10$	$l=15$	$l=20$	$l=25$
ZZ	0.6388	0.8603	0.9546	0.9881
MDCNZZ($\{\pm 1\ 2\}$)	**0.6315**	0.8533	0.9509	**0.9868**
MDCNZZ($\{\pm 1\ 3\}$)	0.6285	**0.8534**	**0.9524**	0.9856
MDCNZZ($\{\pm 1\ 2\ 3\}$)	0.6260	0.8484	0.9474	0.9850
Actual Value	0.6350	0.8564	0.9526	0.9870
Gaussian $(3,1)$				
Algorithm	$l=10$	$l=15$	$l=20$	$l=25$
ZZ	0.6326	0.8495	0.9471	**0.9840**
MDCNZZ($\{\pm 1\ 2\}$)	**0.6267**	**0.8473**	**0.9462**	0.9847
MDCNZZ($\{\pm 1\ 3\}$)	0.6404	0.8495	0.9373	0.9849
MDCNZZ($\{\pm 1\ 2\ 3\}$)	0.6388	0.8603	0.9546	0.9881
Actual Value	0.6275	0.8423	0.9421	0.9823

Table 3.4: *Three two-dimensional Gaussian distributions with* **medium** *correlations and covariance matrices given by (3.56). Each process is labelled by its principal eigenvector and is tested against four different algorithms. For each algorithm we present estimates of the probabilities assigned to squares of the form* $[-l,l]\times[-l,l]$ *under the target Gaussian distribution, for various values of* l*. Each algorithm was simulated until* $N=10^4$ *switches of direction have occurred. The best performance is indicated with bold letters.*



(a) ZZ (b) MDCNZZ(1,2)

Figure 3.3: *Representation of the path of the two algorithms targeting a Gaussian distribution with mode* $(0,0)$*, variance* 41 *and* 101 *for the two components respectively and covariance* 40*. Both algorithms have run until* $N = 10^4$ *switches have occurred.*



(a) ZZ (b) MDCNZZ(1,2)

Figure 3.4: *QQPlots of the first coordinates of the two algorithms targeting a Gaussian distribution with mode* $(0,0)$*, variance* 41 *and* 101 *for the two components respectively and covariance* 40*. Both algorithms have run until* $N = 10^4$ *switches have occurred.*

Gaussian $(1,1)$			
Algorithm	ESS(Standard Deviation)	ESS/Switch	Gelman-Rubin Switches
ZZ	755.70 (245.80)	0.0060	23800
MDCNZZ({±1 2})	926.76 (637.89)	0.0021	272300
MDCNZZ({±1 3})	2426.73 (1000.85)	**0.0192**	12600
MDCNZZ({±1 2 3})	831.93 (347.52)	0.0066	**4000**
Gaussian $(1,2)$			
Algorithm	ESS(Standard Deviation)	ESS/Switch	Gelman-Rubin Switches
ZZ	2998.55 (1493.40)	0.3552	**600**
MDCNZZ({±1 2})	10948.54 (5629.81)	0.3732	800
MDCNZZ({±1 3})	16564.85 (7547.14)	**0.3933**	800
MDCNZZ({±1 2 3})	21848.04 (12013.90)	0.3536	**600**
Gaussian $(3,1)$			
Algorithm	ESS(Standard Deviation)	ESS/Switch	Gelman-Rubin Switches
ZZ	4039.71 (2428.03)	0.2941	**400**
MDCNZZ({±1 2})	14468.05 (9631.23)	0.3048	900
MDCNZZ({±1 3})	21695 (12915.20)	**0.3199**	700
MDCNZZ({±1 2 3})	27323.35 (17101.62)	0.2790	1000

Table 3.5: *Three two-dimensional Gaussian distributions with* **high** *correlations and covariance matrices given by (3.57). Each process is labelled by its principal eigenvector and is tested against four different algorithms. For each algorithm, 25 independent realisations were simulated, each one until time* $T = 10^4$ *and the estimator is constructed using the* δ*-skeleton of the process for* $\delta = 0.1$*. We present ESS with standard deviations, ESS per switch and then number of switches needed for the Gelman-Rubin estimator to take a value less than* 1.01 *and remain less than* 1.01 *until the process terminates. The best performance is indicated with bold letters.*

For the third set of Gaussian distributions we consider **high** correlations and we study three different cases for the covariance matrix:

$$(1)\,.\,a = c = 1, b = \frac{999}{1001},\ (2)\,.\,a = 1, b = \frac{998}{999}, c = \frac{2496}{999},\ (3)\,.\,a = 1, b = \frac{1497}{5491}, c = \frac{1499}{5491}, \tag{3.57}$$

each one with corresponding principal eigenvector $(1,1),(1,2),(3,1)$ respectively. The results for these distributions are presented on Tables 3.5 and 3.6. Contradictory to what we would expect, it is not clear which algorithm performs better for any of the three targets. This is probably due to the high correlation and maybe longer runs are necessary in order to get a more clear picture. However, due to the high correlations we encountered computational problems when we tried to run these algorithms for larger period of time.

The second target is a two-dimensional distribution whose contours have the shape of a banana. In this case we have

$$U(x,y) = (x-1)^2 + k((y-x^2))^2 \tag{3.58}$$

Gaussian $(1,1)$				
Algorithm	$l = 0.5$	$l = 1$	$l = 2$	$l = 3$
ZZ	0.4300	0.7234	0.9321	**0.9981**
MDCNZZ($\{\pm 1\ 2\}$)	0.3408	0.6364	**0.9561**	1
MDCNZZ($\{\pm 1\ 3\}$)	0.3373	**0.6665**	0.9768	1
MDCNZZ($\{\pm 1\ 2\ 3\}$)	**0.3513**	0.6828	0.9650	1
Actual Value	0.3652	0.6705	0.9518	0.9971
Gaussian $(1,2)$				
Algorithm	$l = 1$	$l = 2$	$l = 3$	$l = 5$
ZZ	**0.3726**	0.7722	0.9379	**0.9984**
MDCNZZ($\{\pm 1\ 2\}$)	0.3729	**0.7747**	0.9450	0.9991
MDCNZZ($\{\pm 1\ 3\}$)	0.3752	0.7796	**0.9416**	0.9980
MDCNZZ($\{\pm 1\ 2\ 3\}$)	0.3634	0.7726	0.9394	0.9981
Actual Value	0.3711	0.7767	0.9412	0.9984
Gaussian $(3,1)$				
Algorithm	$l = 0.5$	$l = 1$	$l = 2$	$l = 3$
ZZ	0.2821	**0.6610**	0.9524	0.9970
MDCNZZ($\{\pm 1\ 2\}$)	0.2873	0.6668	0.9588	0.9978
MDCNZZ($\{\pm 1\ 3\}$)	0.2724	0.6448	**0.9525**	0.9970
MDCNZZ($\{\pm 1\ 2\ 3\}$)	**0.2808**	0.6641	0.9577	**0.9974**
Actual Value	0.2799	0.6602	0.9544	0.9973

Table 3.6: *Three two-dimensional Gaussian distributions with* **high** *correlations and covariance matrices given by (3.57). Each process is labelled by its principal eigenvector and is tested against four different algorithms. For each algorithm we present estimates of the probabilities assigned to squares of the form* $[-l, l] \times [-l, l]$ *under the target Gaussian distribution, for various values of* l*. Each algorithm was simulated until* $N = 10^4$ *switches of direction have occurred. The best performance is indicated with bold letters.*

for some parameter $k > 0$. The global maximum of the density is $(1, 1)$ but the parameter k is an extra parameter that controls how much the mass concentrates around a parabola. The larger the value of k, the higher the tendency of the distribution to concentrate its mass around parabola $y = x^2$ and this makes the simulation harder. Again the shape of the contours gives us a reason to allow the process to move in more directions than just $\{-1, 1\}^2$. We ran simulations for Banana shaped distributions with parameter $k = 10$ and $k = 50$. In this case, each process is simulated until a fixed number $N = 2 \cdot 10^4$ of switches of direction occurred. As discussed above, fixing the number of switches, rather than the time, seems a more natural way to compare the performance of the algorithms. For each target density we use five different algorithms: the original Zig-Zag and MDCNZZ with directions $\{\pm 1, 2\}^2, \{\pm 1, 3\}^2, \{\pm 1, 5\}^2$ and $\{\pm 3, 5\}^2$. In each of the algorithms we set the refreshment to be zero. For each distribution and each algorithm we run 25 independent processes, starting from over-dispersed positions and we use δ-skeletons of these processes as samples for $\delta = 0.1$.

On Table 3.7 we summarise some of our findings. The first table refers to the case where $k = 10$ and the second on the case where $k = 50$. For each algorithm we present estimates of the probability of being inside four different rectangles, under the target distribution. The actual probability is estimated using an Adaptive MCMC algorithm implemented in library `adaptMCMC` of `R` and we have used $5 \cdot 10^6$ samples, $5 \cdot 10^4$ burn-in and average acceptance rate 0.33. We also present the average ESS over the 25 realisations (and standard deviation in a parenthesis) until $N = 2 \cdot 10^4$ switches have occurred. For the two dimensional ESS we are using the `mcmcse` library from `R` language. We note that in the case where $k = 50$ there is a lot of variance on ESS over the 25 independent chains. It is a repeated phenomenon that whenever this variance is large, typically there is typically one or two outlier chain that gave really high ESS.

In the case $k = 10$ it can be clearly seen that all the MDCNZZ algorithms outperform the original Zig-Zag. The picture is a bit more complicated in the case $k = 50$, where the original Zig-Zag and the MDCNZZ($\{\pm 1, 2\}$) have very similar behaviour. At the same time, the original Zig-Zag seems to have more ESS than the rest of the algorithms, but MDCNZZ($\{\pm 1, 3\}$) seems to estimate the rectangles' probabilities slightly better.

In Figure 3.5 we present a realisation of the path of the two algorithms when targeting the Banana shape distribution with parameter $k = 10$. Both algorithms have run until $N = 10^4$ switches have occurred.

The last family of two-dimensional distributions we target is mixtures of

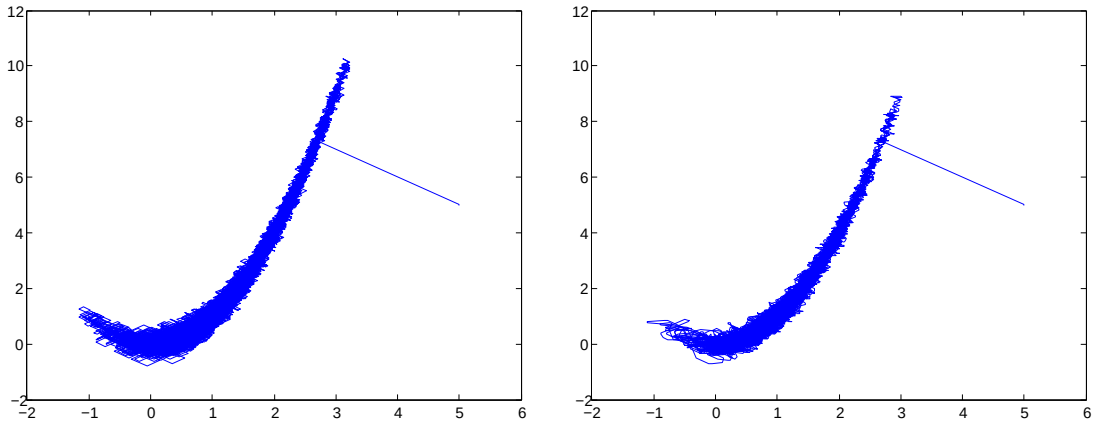
Banana target $k = 10$

Algorithm	ESS	$[1.5,3] \times [0,6]$	$[-1,1] \times [1,5]$	$[-1,1] \times [-1,1]$	$[-1,3] \times [0,6]$
ZZ	133.5 (62.98)	0.1937	0.1984	0.2808	0.7887
MDCNZZ({±1 2})	272.1 (94.23)	0.1977	0.0280	0.4987	0.8894
MDCNZZ({±1 3})	267.6 (76.35)	**0.2161**	0.0269	0.4746	**0.8909**
MDCNZZ({±1 5})	**275.0** (93.78)	0.2261	0.0261	**0.4713**	0.8934
MDCNZZ({±3 5})	224.5 (85.89)	0.2264	**0.0274**	0.4574	0.8924
Actual Value	-	0.2180	0.0277	0.4699	0.8913

Banana target $k = 50$

Algorithm	ESS	$[1.5,3] \times [0,6]$	$[-1,1] \times [1,5]$	$[-1,1] \times [-1,1]$	$[-1,3] \times [0,6]$
ZZ	93.9 (45.43)	0.2029	0.0116	0.5091	**0.9206**
MDCNZZ({±1 2})	**135.8** (129.9)	**0.2176**	0.0115	0.4960	0.9255
MDCNZZ({±1 3})	103.4 (81.6)	0.2127	**0.0119**	0.4860	0.9202
MDCNZZ({±1 5})	121.3 (130.9)	0.2523	0.0110	0.4447	0.9086
MDCNZZ({±3 5})	102.7 (72.8)	0.2201	0.0108	**0.4888**	0.9109
Actual Value	-	0.2170	0.0119	0.4878	0.9230

Table 3.7: *Two Banana targets with minus log-likelihoods given by (3.58) and parameters $k = 10, 50$ respectively. Each process is tested against five different algorithms. For each algorithm, 25 independent processes were simulated, each one until $N = 2 \cdot 10^4$ switches occurred and the estimator is constructed using the δ-skeleton of the process for $\delta = 0.1$. We present the average ESS over these 25 realisations along with the standard deviation in a parenthesis. We also present the algorithms' estimations of the probabilities the target assigns to various rectangles of $\mathbb{R}^2$, along with the actual probabilities for comparison. The best performance is indicated with bold letters.*



(a) ZZ (b) MDCNZZ(1,5)

Figure 3.5: *Representation of the path of the two algorithms targeting a Banana distribution with target minus log-likelihood given by equation (3.58) and parameter $k = 10$. Both algorithms have run until $N = 10^4$ switches have occurred.*

two two-dimensional Gaussians. These are bi-modal distributions, typically difficult targets for MCMC. We consider two different targets, mixtures of two-dimensional Gaussians. The first one has modes the vectors $(0,0)$ and $(0,6)$, while the second one has modes the vectors $(0,0)$ and $(1,4)$. In both cases we choose the covariance matrices of the Gaussians to be the identity I_2. For each one of these targets we run a simple Zig-Zag and MDCN Zig-Zag with velocity space containing the vector of the difference between the two modes. This means that for the first process with modes $(0,0),(0,6)$ we use direction $\{-6,0,6\}^2$, while for the second one with modes $(0,0),(1,4)$ we use $\{-4,-1,1,4\}^2$. The idea behind that is that the algorithm should prefer to move in parallel to the difference of the two modes, as this will help the process move from one mode to the other faster. Each algorithm was simulated 25 times independently and the results presented are the average results over these 25 realisations. All the simulations run until $N = 10^5$ switches of direction occurred. As discussed above, acting this way allows us to compare the performance of the algorithms directly. As estimators, we use the δ-skeletons of the processes every $\delta = 0.1$ time units. Our intuition seems to be verified for the first one of the two processes when we consider how capable is the algorithm approximating the probability that the target assigns to certain squares. The MDCNZZ clearly outperforms the normal Zig-Zag in that scenario. On the other hand this is not verified in the second target, where the Zig-Zag seems to outperform the MDCNZZ, although the difference in the performance does not seems to be large. We suspect that the closer the difference between the two modes is to being parallel to $(1,1)$, the less we gain by using MDCNZZ. This is due to the fact that the simple Zig-Zag will move to the direction $(1,1)$, close to parallel to the difference between the two modes and it will probably jump from one mode to the other fast enough without the need to introduce an extra direction. We summarise our results in Table 3.8, where we present the average ESS, along with the Standard Deviation in a parenthesis and the estimation of the probability the target density assigns to five different squares of $\mathbb{R}^2$. We also compare these with the actual probability, estimated using `mvtnorm` of *R*.

3.5.2 Five Dimensional Targets

In this section we will present some results on a five dimensional Gaussian target. More precisely, we target a Gaussian distribution with mode $(0,0,0,0,0)$ and

Mixture Gaussian with modes $(0,0)$ and $(0,6)$

Algorithm	ESS	$[-1,1]^2$	$[2,4]\times[4,6]$	$[2,4]^2$	$[-6,6]^2$	$[-1,1]\times[5,7]$
ZZ	**6920.2** (1109.74)	0.2275	0	0.0650	1	0
MDCNZZ($\{-6,0,6\}$)	4541.3 (715.98)	**0.2320**	**0.0055**	**0.0005**	**0.7508**	**0.2331**
Actual Value	-	0.2330	0.0054	0.0005	0.7500	0.2330

Mixture Gaussian with modes $(0,0)$ and $(1,4)$

Algorithm	ESS	$[-11]^2$	$[2,4]\times[4,6]$	$[2,4]^2$	$[-6,6]^2$	$[-1,1]\times[5,7]$
ZZ	**28859.2** (1450.2)	**0.2282**	**0.0387**	**0.0398**	**0.9885**	**0.0383**
MDCNZZ($\{\pm1,4\}$)	15744.0 (1694.7)	0.2247	0.0393	0.0402	0.9882	0.0390
Actual Value	-	0.2334	0.0375	0.0378	0.9886	0.0375

Table 3.8: *Two two-dimensional targets consisting of mixtures of two Gaussian distributions each. In the first target the Gaussians have modes* $(0,0)$ *and* $(0,6)$*, while on the second target* $(0,0)$ *and* $(1,4)$*. The covariance matrix of all Gaussians is the identity. For both targets we use the original Zig-Zag and the MDCNZZ with direction parallel to the difference between the two modes. Each algorithm was simulated 25 times independently until* $N = 10^5$ *number of switches occurred. As an estimator we use the* δ*-skeleton of the process with* $\delta = 0.1$*. We present the average ESS over these 25 realisations of the algorithms and the standard deviation in a parenthesis. We also present the algorithms' estimations of the probabilities the target assigns to various rectangles of* $\mathbb{R}^2$*, along with the actual probabilities for comparison. The best performance is indicated with bold letters.*

covariance matrix

$$A = \begin{pmatrix} 4 & 0.7 & 0.7 & 0.7 & 0.7 \\ 0.7 & 3 & 0.7 & 0.7 & 0.7 \\ 0.7 & 0.7 & 3 & 0.7 & 0.7 \\ 0.7 & 0.7 & 0.7 & 3 & 0.7 \\ 0.7 & 0.7 & 0.7 & 0.7 & 3 \end{pmatrix}. \tag{3.59}$$

We note that the principal eigenvector is (in four decimal digits accuracy)

$$v = \begin{pmatrix} -0.5632 \\ 0.8263 \\ -3.0127 \cdot 10^{-16} \\ -3.0127 \cdot 10^{-16} \\ -3.0127 \cdot 10^{-16} \end{pmatrix}. \tag{3.60}$$

Therefore, we use an original Zig-Zag sampler and a MDCNZZ with velocity space $\{-8.3, -5.6, 0, 5.6, 8.3\}^5$. We ran the algorithms for $N = 10^4$ number of switches and we use δ-skeleton with $\delta = 0.1$ as estimators and for each algorithm we ran 25 independent realisations. In Table 3.9 we present the average ESS over the 25 chains of the algorithms, along with the standard deviation in a parenthesis. We also present an estimator of the probability the target assigns to four different

Five Dimensional Gaussian Distribution

Algorithm	ESS (SD)	$[-1,1]^5$	$[-2,2]^5$	$[-4,4]^5$	$[-5,5]^5$
ZZ	3268.3 (210.23)	**0.0162**	**0.2375**	0.8815	**0.9729**
MDCNZZ($\{-8.3,-5.6,0,5.6,8.3\}$)	**4541.3** (1572.02)	**0.0162**	**0.2375**	**0.8814**	0.9730
Actual Value	-	0.0161	0.2364	0.8814	0.9728

Table 3.9: *A five dimensional Gaussian distribution with mode* $(0,0,0,0,0)$ *and covariance matrix with variances* $4,3,3,3,3$ *and covariances all equal to* 0.7 *(given by (3.59)). We use two algorithm, the original Zig-Zag and a MDCNZZ allowing a direction close to the principal eigenvector, approximately given by (3.60). Each algorithm was simulated 25 times, independently , each one ran until* $N = 10^4$ *number of switches. We use the* δ*-skeleton, with* $\delta = 0.1$*, as an estimator. We present the average ESS over the 25 realisations of the two algorithms along with the standard deviation in a parenthesis. We also present the algorithms' estimations of the probabilities the target assigns to various rectangles of* $\mathbb{R}^2$*, along with the actual probabilities for comparison.*

squares: $[-1,1]^5, [-2,2]^5, [-4,4]^5, [-5,5]^5$. The multivariate ESS and the actual probabilities are calculated using `mcmcse` and `mvtnorm` from `R` respectively. The probability estimates from both algorithms seem to be very good, while on the ESS, the MDCNZZ seems to perform better (although the standard deviation is quite large).

3.6 Discussion

In this Chapter we have generalised the Zig-Zag process so that it moves in more directions, taken from a set of the form $\{\theta_1, ..., \theta_m\}^d$. We have established conditions for the rate functions so that the process has a measure
$\mu(dx, d\theta) = \frac{1}{Z}\exp\{-U(x)\}\frac{1}{m}\sum_{i=1}^{m}\delta_{\theta_i}dx$ as invariant measure. This measure has uniform marginal distribution in the velocity space, however we, also, identified conditions for the rates so that the process can target measures with non-uniform marginals as well.

In the case of uniform marginals and when the process is only allowed to jump from one velocity to a "neighbourhood" one, we call the process Multi-directional Closest Neighbour Zig-Zag and we prove ergodicity and geometric ergodicity with assumptions very similar to the ones used in order to prove results for the original Zig-Zag. We believe, however, that the ergodicity proof can still be applied to more general algorithms, where the closest neighbour jump is no longer imposed.

Furthermore, we include some simulations that indicate that in some cases, like some banana target distributions, or some mixtures of Gaussians, one can gain by using MDCNZZ instead of original Zig-Zag. We also presented simulations on

a five dimensional Gaussian target using an appropriate MDCNZZ. We allowed the process to move in the direction of the eigenvector corresponding to the principal eigenvalue of the covariance matrix. According to these results, the MDCNZZ seems to outperform the original Zig-Zag.

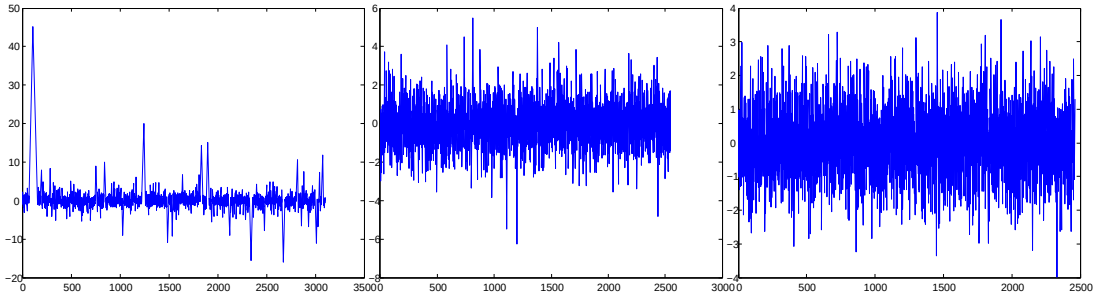
The idea that the process could perform better if it moves more in the direction of this principal eigenvector indicates that we should also target distributions with non-uniform marginals in the velocity space, but where, instead, more weight is given to directions close to parallel to the principal eigenvector. Since one will not have access to the principal eigenvector in a practical setting, it would be very interesting to create an adapted MDCNZZ algorithm that learns what type of directions it should use and how much time to spend in each direction, as it explores the space. For a review of Adapted MCMC see for example [RR09]. This adapted version of Multi Directional Zig-Zag is still work in progress.

Chapter 4

Zig-Zag Process on Heavy Tailed Target Densities

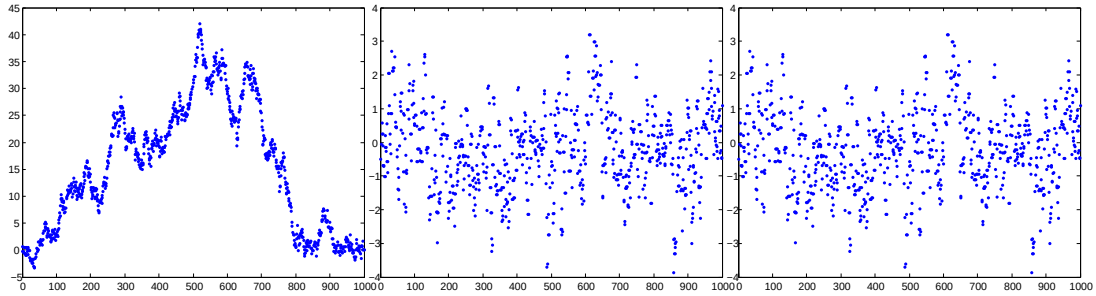
This chapter deals with the convergence behaviour of the Zig-Zag process on heavy tailed targets. It is well known that most MCMC samplers struggle in such a scenario (see for example [MT96,JH00,RT96,LBBG19]). The same behaviour is observed for the Zig-Zag sampler. To indicate this with an example, Figures 4.1 and 4.3 provide the trace plots and the QQplots respectively of one dimensional Zig-Zag processes, with minimal rates ($\gamma(x) \equiv 0$), targeting student t-distributions with three different degrees of freedom. For the trace plots, the process ran until time $N = 10^3$ switches of direction have occurred and for the QQplots until time $N = 10^4$ switches of direction have occurred. As we have explained in earlier Chapter, fixing the number of switches allows us to compare the performance of each algorithm directly.

Figures indicate that the processes become more stable when targeting a Student distribution with higher degrees of freedom, which has lighter tails. There are two main problems here, both of which can be seen in the trace plot of the Cauchy distribution. The first one is that the algorithm is very local, it can only move to neighbouring points as it always moves with unit speed and it will take a lot of time to make an appropriately large excursion to visit the tails of the distribution. The second problem is that when this excursion happens, it will be extremely large in order to compensate for the underestimation of the tails. More precisely, the switching time of the excursion will be very large because the rate of the switching process will decay to zero and even when the process does switch, it will take a lot of time to return close to the mode of the distribution, resulting in the algorithm spending a lot more time at the tails than appropriate. Overall this leads to a very unstable algorithm, having no excursions for a long period of time followed by a



(a) Cauchy distribution (b) Student(8) distribution (c) Student(500) distribution

Figure 4.1: *Trace plots Zig-Zag for three Student distributions with increasing degrees of freedom.*

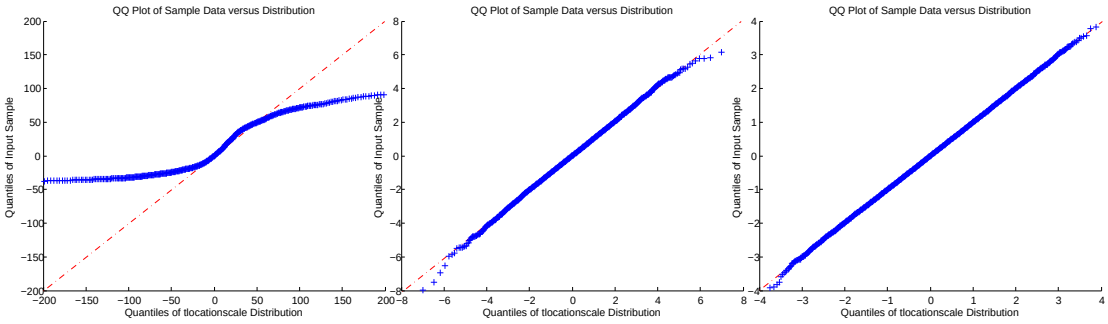


(a) Cauchy distribution (b) Student(8) distribution (c) Student(500) distribution

Figure 4.2: *Trace plots Random Walk Metropolis with Normal (*0 *mean,* 1 *variance) proposal for three Student distributions with increasing degrees of freedom.*

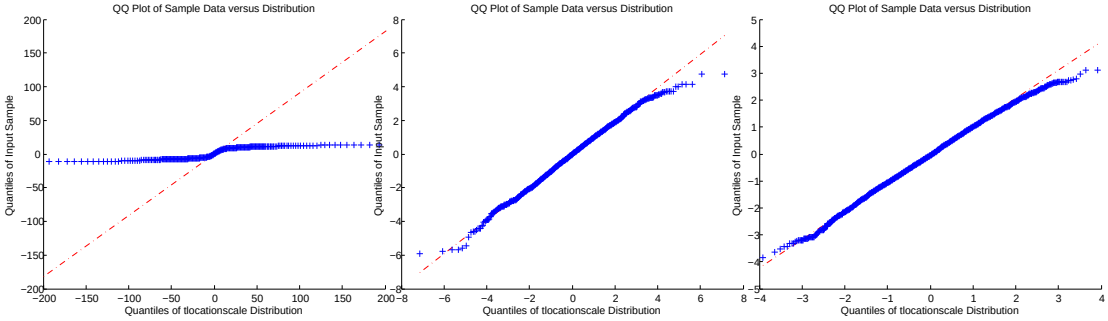
very large excursion (see for example Figure 4.2a) and means that the tails of the distribution are not adequately estimated (see for example Figure 4.3a). On the other hand, when we compare the behaviour of this process with one classic MCMC algorithm, the Random Walk Metropolis, we see that Zig-Zag performs somewhat better. To have a visual comparison, we present here trace-plots and QQplots of a Random Walk Metropolis algorithm with $N(0,1)$ proposal, targeting Cauchy, $t(8)$ and $t(500)$ distributions. The trace-plots presented have ran until $N = 10^3$ iterations and the QQplots until $N = 10^4$ iterations.

We observe that this process, lacking the notion of momentum the Zig-Zag has, struggles even more to reach the tails of the distribution (as shown in figure 4.4). At the same time, once the process reaches the tails, it has a very diffusive behaviour and struggles to return to the mode (as observed in figure 4.2). On the contrary,



(a) Cauchy distribution (b) Student(8) distribution (c) Student(500) distribution

Figure 4.3: *QQ plots Zig-Zag for three Student distributions with increasing degrees of freedom.*



(a) Cauchy distribution (b) Student(8) distribution (c) Student(500) distribution

Figure 4.4: *QQ plots of Random Walk Metropolis with Normal proposal for three Student distributions with increasing degrees of freedom.*

once the Zig-Zag process turns around and starts moving towards the mode, it will never switch direction again before hitting the mode. Overall the Zig-Zag process seems to outperform the Random Walk Metropolis with Normal Proposal.

The chapter is organised as follows. In Section 4.1 we present a simple argument that proves that the Zig-Zag is not geometrically ergodic in heavy tails. Even though the process will not converge with a geometric rate, it will, however, mix with a polynomial rate when targeting distributions with polynomial tails in one dimension. We manage to identify this rate in Section 4.2, when the refresh rate is not large enough. Then, in section 4.3 we present a second, more complicated argument which proves Zig-Zag's lack of Geometric Ergodicity after making some assumptions on the refresh rates. This argument uses equivalence between geometric ergodicity and existence of some moment generating function of hitting times of

small sets (see Theorem 2.1.5 in this work). This technique provides us the tools to prove non-geometric ergodicity for other, more complicated algorithms, where the argument of Section 4.1 can no longer be applied. Indeed, in section 4.4 we present, as an example, a variant of Zig-Zag and we establish non-geometric ergodicity using tools from section 4.3.

4.1 A Simple Argument for Non-Geometric Ergodicity of Zig-Zag

There is a simple argument to prove that any of the classic PDMP algorithms cannot be Geometrically Ergodic when targeting heavy tails. The key is to observe that all these algorithms move around the space with unit speed, which means that before time t they will not have exit a ball of radius t around the point they started, leaving all the area outside the ball unexplored. This proof was suggested to us by professor Anthony Lee during private correspondence.

Theorem 4.1.1 (Non-Geometric Ergodicity of Zig-Zag)**.** *Suppose that the Zig-Zag targets a heavy tailed distribution μ. Then the process is not Geometrically Ergodic.*

Proof. Suppose that the Zig-Zag starts from $x = 0$, $\theta \in \{-1, 1\}^d$. Fix a time $t > 0$. Let $A_t = \{x : |x| > \sqrt{d}t\}$. After time t has passed, the Zig-Zag will not have hit A_t. Therefore,

$$\|\mathbb{P}_{0,\theta}(Z_t \in \cdot) - \mu(\cdot)\|_{TV} \geq \left|\mathbb{P}_{0,\theta}(X_t \in A_t) - \mu(A_t \times \{-1, +1\}^d)\right| = \mu(A_t \times \{-1, +1\}^d).$$

Assume that the process is Geometrically ergodic. Then there exists an $M > 0$ and $\rho < 1$ such that for all $t \geq 0$

$$\mu(A_t \times \{-1, +1\}^d) \leq \|\mathbb{P}_{0,\theta}(Z_t \in \cdot) - \mu(\cdot)\|_{TV} \leq M\rho^t.$$

This contradicts the fact that μ has heavy tails. □

Note that the same proof can be used for any other algorithm with unit speed, for example the Bouncy Particle Sampler where the refresh velocity is chosen from the uniform sphere. We believe that a similar type argument can be used in any Metropolis-Hastings algorithm with a light tails proposal distribution. Note further that this proof implies that the MDCNZZ process, introduced in Chapter 3, also suffers the same problem on heavy tails distribution as the process has a maximum finite speed it can move with. We propose a solution to the non-geometric ergodicity behaviour of the process in Chapter 5.

4.2 Polynomial Ergodicity in One Dimension

Even though we have established that the Zig-Zag process will not be geometrically ergodic in heavy tails, we know from Theorem 2.4.5 that it will converge to the invariant measure under very mild assumptions. In this section we will present some results on the rate of convergence of the one dimensional Zig-Zag process, focusing on the case where the target distribution is a t-distribution. We will use the following theorem, found and proved in the presented form in the unpublished lecture notes [Hai16]. It can also be found in Theorems 3.2 and 3.4 of [DFG09] and Theorem 1.2 of [BCG08] (see also Corollary 6 in [FR05] for some earlier results on sub-geometric convergence).

Theorem 4.2.1. *Let $(X_t)_{t\geq 0}$ a continuous time Markov process on X with strong generator $\mathcal{L}$. Suppose that there exists a function $V : X \to [1, +\infty)$ and a constant K such that for all $x \in X$*

$$\mathcal{L}V(x) \leq K - f(V) \tag{4.1}$$

for a function $f : [0, +\infty) \to [0, +\infty)$ strictly concave, increasing, with $f(0) = 0$, $\lim_{s\to+\infty} f(s) = +\infty$. Suppose further that all the sub-level sets of V are pre-compact and small. Then, the following hold:

1. *There exists a unique invariant measure μ for the process such that $\int f(V(x))\mu(dx) < \infty$.*
2. *Let $H_f(u) = \int_1^u \frac{1}{f(s)} ds$, then there exists a constant $B > 0$ such that for every $x \in X$*

$$\|\mathbb{P}^t(x, \cdot) - \mu(\cdot)\|_{TV} \leq \frac{BV(x)}{H_f^{-1}(t)} + \frac{B}{f \circ H_f^{-1}(t)}. \tag{4.2}$$

Assumption 4.2.2. *Assume that $U \in C^1(\mathbb{R})$ and assume that there exists an $\epsilon > 0$ and a compact set $C \subset \mathbb{R}$ such that for all $x \notin C$,*

$$|U'(x)| \geq \frac{1+\epsilon}{|x|}. \tag{4.3}$$

Remark 4.2.3. *This assumption directly implies that there exists a c' such that for all $x \in \mathbb{R}$, $U(x) \geq (1+\epsilon)\log(|x|) - c'$. This is an assumption made in [BRZ19] in order to prove ergodicity of the Zig-Zag process and we also used it in Chapter 3 (in the form of Assumption 3.3.2) to prove non-evanescence of MDCNZZ.*

We also need the following assumption.

Assumption 4.2.4. *The refresh rates satisfy*

$$\lim_{|x|\to\infty} \frac{\gamma(x)}{|U'(x)|} = 0. \tag{4.4}$$

This assumption implies that the main reason for the Zig-Zag process to switch is that it is moving in the wrong direction, relatively to some random switches that might occur due to some extra refreshment. The higher the extra refreshment, the more the process will behave like a random walk. The polynomial rates of convergence for the Random Walk Metropolis have long been studied (see for example [JT03]). At least in one dimension the RWM looks like having a worse rate of convergence than the Zig-Zag without any refresh. This will be verified for student distributions in this section.

Proposition 4.2.5 (Polynomial Ergodicity of Zig-Zag)**.** *Suppose that U satisfies assumption 4.2.2 and let C and $\epsilon > 0$ as in (4.3). Suppose further that the refresh rate satisfies assumption 4.2.4. Let $\mathcal{L}$ be the strong generator of the one dimensional Zig-Zag process $(Z_t)_{t\geq 0}$. Then, for any $a < 1-1/(1+\epsilon)$, there exists some $\beta \in (0,1)$, such that the function $V_\beta \in C^1$ defined as*

$$V_\beta(x,\theta) = \exp\{\beta U(x) + \delta sign(x)\theta\} \tag{4.5}$$

satisfies

$$\mathcal{L}V_\beta(x,\theta) \leq K - (V_\beta(x,\theta))^a. \tag{4.6}$$

Therefore, for any $k' < \epsilon$, there exists a constant $B > 0$ such that for all $(x,\theta) \in \mathbb{R}\times\{-1,+1\}$,

$$\left\|\mathbb{P}^t_{x,\theta}(Z_t \in \cdot) - \mu(\cdot)\right\|_{TV} \leq \frac{BV_\beta(x)}{t^{1+k'}} + \frac{B}{t^{k'}}. \tag{4.7}$$

Proof of Proposition 4.2.5. Suppose that U satisfies Assumption 4.2.2 and let $a < 1-1/(1+\epsilon)$. For any $\tilde{\beta} \in (0,1)$ we have that there exists a $c_0 > 0$ such that for all $x \notin C$

$$\left(V_{\tilde{\beta}}(x,\theta)\right)^{1-a} |U'(x)| \geq c_0 \exp\{\tilde{\beta}(1-a)(1+\epsilon)\log|x|\}\frac{1+\epsilon}{|x|} = c_0(1+\epsilon)\,|x|^{\tilde{\beta}(1-a)(1+\epsilon)-1}$$

Since $(1-a)(1+\epsilon)-1 > 0$, there exists a β close to 1 such that $\beta(1-a)(1+\epsilon)-1 > 0$, so

$$\lim_{|x|\to\infty} V_\beta^{1-a}(x,\theta)|U'(x)| = +\infty. \tag{4.8}$$

Now, V_β is C^1 with $\lim_{|x|\to\infty} V_\beta(x,\theta) = +\infty$ so all the level sets are compact. Since the process is positive Harris recurrent and some skeleton is irreducible (see

[BRZ19]) we get from Proposition 6.1 of [MT93a] that the level sets are also small. Since $\lim_{|x|\to\infty} U(x) = +\infty$ it is clear that V_β is bounded below away from 0 so by multiplying with an appropriate constant we can assume that $V_\beta(x,\theta) \geq 1$ for all (x,θ). We calculate

$$\mathcal{L}V_\beta(x,\theta) = V_\beta(x,\theta)[\theta\beta U'(x) + ([\theta U'(x)]^+ + \gamma(x))[\exp\{-2\theta sign(x)\delta\} - 1]].$$

Note that due to Assumption 4.2.2 and that $U(x) \xrightarrow{|x|\to\infty} +\infty$, we get that there exists a compact set C such that for all $x \notin C$ $sign(U'(x)) = sign(x)$. Therefore, when $x \notin C$ and $\theta sign(x) > 0$,

$$\frac{\mathcal{L}V_\beta(x,\theta)}{V_\beta^a(x,\theta)} \leq V_\beta^{1-a}(x,\theta)|U'(x)| \left[\beta + \left(\frac{\gamma(x)}{|U'(x)|} + 1\right)(\exp\{-2\delta\} - 1)\right]$$

and when $\theta sign(x) < 0$

$$\frac{\mathcal{L}V_\beta(x,\theta)}{V_\beta^a(x,\theta)} \leq V_\beta^{1-a}(x,\theta)|U'(x)| \left[-\beta + \frac{\gamma(x)}{|U'(x)|}(\exp\{2\delta\} - 1)\right].$$

Overall we have for $x \notin C$

$$\frac{\mathcal{L}V_\beta(x,\theta)}{V_\beta^a(x,\theta)} \leq V_\beta^{1-a}(x,\theta)|U'(x)|\times$$
$$\max\left\{\beta + \left(\frac{\gamma(x)}{|U'(x)|} + 1\right)(\exp\{-2\delta\} - 1), \left[-\beta + \frac{\gamma(x)}{|U'(x)|}(\exp\{2\delta\} - 1)\right]\right\}.$$

Recall that $\frac{\gamma(x)}{|U'(x)|} \xrightarrow{|x|\to\infty} 0$. Fix $\delta > -1/2\log(1-\beta)$ and by possibly increasing C, we have that there exists a constant $c' > 0$ such that for all $x \notin C$

$$\max\left\{\beta + \left(\frac{\gamma(x)}{|U'(x)|} + 1\right)(\exp\{-2\delta\} - 1), -\beta + \frac{\gamma(x)}{|U'(x)|}(\exp\{2\delta\} - 1)\right\} < -c' < 0.$$

Combining this with (4.8) we get that V_β satisfies (4.6).
Therefore, all the conditions of Theorem 4.2.1 are satisfied with $f(u) = cu^a$ for some constant $c > 0$. Note that $H_f(s) = \int_1^s f(u)du = c^{-1}\int_1^s u^{-a}du = \frac{1}{c(1-a)}(s^{1-a} - 1)$ so

$$H_f^{-1}(t) = (1 + c(1-a)t)^{1/(1-a)}$$

and therefore

$$f \circ H_f^{-1}(t) = c\,(1 + c(1-a)t)^{a/(1-a)}$$

Note that the same argument can be made for any $a < 1 - 1/(1+\epsilon)$. Since $a <$

$1 - 1/(1+\epsilon)$ is equivalent to $1/(1-a) < 1+\epsilon$ and $a/(1-a) < \epsilon$, (4.7) follows. $\square$

Example 4.2.6. *Suppose π is a t-distribution with k degrees of freedom, i.e. $\pi(x) = Z^{-1}\left(1 + x^2/k\right)^{-(k+1)/2}$ and the Zig-Zag process targets μ whose marginal on $\mathbb{R}$ is π and the marginal on $\{-1,+1\}$ is the uniform distribution. Then for all $\delta' > 0$*

$$U'(x) = \frac{(k+1)\frac{|x|}{k}}{1+\frac{x^2}{k}} \geq \frac{k+1-\delta'}{|x|}$$

Therefore for every $\delta' > 0$, these distributions satisfy assumption 4.2.2 with $\epsilon = k - \delta'$. From Proposition 4.2.5, for all $k' < k$, there exists a $\beta \in (0,1)$ and $B > 0$ such that for all $(x,\theta) \in \mathbb{R} \times \{-1,+1\}$,

$$\|\mathbb{P}_{x,\theta}(Z_t \in \cdot) - \mu(\cdot)\|_{TV} \leq \frac{BV_\beta(x)}{t^{1+k'}} + \frac{B}{t^{k'}}.$$

In fact, in this case this polynomial rate is optimal, in the sense that we can prove that the total variation distance does not decay as t^{-k}. Indeed, suppose that the Zig-Zag starts from $x = 0$, $\theta = +1$. There exists a C_0 and $K > 0$ such that for all $|x| \geq K$, $\pi(x) \geq C_0|x|^{-k-1}$. Fix a time $t > K$. Let $A_t = \{x : x > t\}$. After less or equal to time t has passed, the Zig-Zag will not have hit A_t. We therefore get for all $t > K$,

$$||\mathbb{P}_{0,+1}(Z_t \in \cdot) - \mu(\cdot)||_{TV} \geq |\mathbb{P}_{0,+1}(X_t \in A_t^c) - \pi(A_t^c)| = \pi(A_t^c) = \int_{|x|>t} \pi(x)dx \geq$$
$$\geq 2C_0 \int_t^{+\infty} x^{-k-1}dx = \frac{C_0}{k}\frac{1}{t^k}.$$

It is interesting to compare these rates with one dimensional Random Walk Metropolis and MALA algorithms. In fact, it is shown in [JT03], Propositions 4.1 and 4.3 (see also [JR07]) that when targeting student distribution with k degrees of freedom with any finite variance proposal RWM or with MALA, one gets polynomial order of convergence $(k/2)^-$. This means that, for any $\epsilon > 0$, the polynomial rate of convergence is at least $k/2-\epsilon$, but it is proven not to be $k/2$. In [JR07] the authors provide these lower bounds for the convergence rates, while in [JT03] they provide the upper bound. As proven in this Chapter, the one-dimensional Zig-Zag has polynomial rate of convergence k^- in the same setting, which is better than Random Walk Metropolis or MALA. This phenomenon was also observed in simulations in [BD17]. We conjecture that the advantage of the Zig-Zag is due to the momentum, which, in a one-dimensional, uni-modal setting with zero refresh rate, will force the process to

never switch direction before it hits the mode. This diminishes any possible diffusive behaviour of the process at the tails and helps the algorithm converge faster. We should note here that better polynomial rates (and more precisely, arbitrarily better rates) can be achieved for the Random Walk Metropolis if one introduces a proposal with heavier tails and this will in fact outperform the Zig-Zag. However, the natural analogue of this modification is to allow the Zig-Zag to speed up and move faster in areas of lower density as will be introduced in Chapter 5.

4.3 Random Walk Construction for Zig-Zag Excursions

The purpose of this section is to introduce theoretical tools that will be used to progressively prove the Zig-Zag non-geometric ergodicity result, Theorem 4.3.3 in various settings. The result is already proven in Section 4.1, however these tools will later be used to prove non-Geometric ergodicity for some variants of the Zig-Zag, which are not covered by the proof of Theorem 4.1.1.

The strategy for this section is to use the second part of Theorem 2.1.5 in chapter 2 and prove that the process is not geometrically ergodic by proving that no petite set can have the SGR property. This will involve studying the structure of the movement of the process. For the one dimensional Zig-Zag, which will be our main focus, the process moves up and down with up-steps and down-steps. Consider the following scenario. Assume that the process starts from $(y, +1)$ and we want to study the time it takes for the process to hit a petite set that lies left of y. The time period between the start of an up-step and the end of the down-step forms an excursion of the process and studying the hitting time of the petite set is equivalent to study the length of these excursions along with the number of excursions that will take place before the process hits the set. Let us consider the number of excursions. If we consider the chain $(S_n)_{n\geq 0}$ that marks the end point after each down step, the number of excursions occurring before the Zig-Zag process hits the petite set C is equal to the hitting time of the set C by the chain S_n. This chain forms a random walk (see for example Figure 4.5) and part of this section will be to control its behaviour.

Throughout the section we will consider $U \in C^1(\mathbb{R}^d \times \{-1, +1\}^d)$ and make the following assumption.

Assumption 4.3.1 (Log-Likelihood Uniform Decay)**.** *For one of the coordinates* $i \in \{1, .., d\}$ $\lim_{x_i \to +\infty} \partial_i U(x) = 0$ *uniformly in the rest of the coordinates, i.e.*

for all $\epsilon > 0$ there exists $x_0 = x_0(\epsilon)$ such that for all $x = (x_1, ..., x_d)$ with $x_i > x_0$,

$$|\partial_i U(x)| < \epsilon.$$

Remark 4.3.2. *Although this looks a technical assumption it is always satisfied when $\lim_{|x|\to\infty} \|\nabla U(x)\| = 0$, which is the most common definition of a heavy tailed distribution in the MCMC literature. It will become evident that the following results also hold if on the assumption above the limit is taken as $x_i \to -\infty$, using symmetric arguments.*

Our main goal of the section will be to prove the following theorem.

Theorem 4.3.3. *Let $(Z_t)_{t\geq 0}$ be a d-dimensional Zig-Zag process and $U \in C^1$ be the minus log-likelihood of the invariant distribution μ of the process. Assume that Assumption 4.3.1 holds. Assume further that for the coordinate i introduced in the Assumption 4.3.1, one of the following two propositions holds for the refresh rate γ_i,*

1. *$\lim_{x\to+\infty} \gamma_i(x) = 0$.*
2. *We have $\gamma_i(x) = \gamma_i(x_i)$ and there exist $0 < m < M < \infty$ and an $x_0 \in \mathbb{R}$ such that for all $x \geq x_0$*

$$m \leq \gamma_i(x_i) \leq M. \tag{4.9}$$

Then, the Zig-Zag process is not Geometrically Ergodic.

Note that the assumption on the refresh rates is satisfied in the case where one allows a constant refreshment rate, the same for all coordinates. This is usually the case in the literature so far. The reason we ask for γ_i to only depend on the i position coordinate is that we want to project this process to a one dimensional one, which is much easier to analyse. In terms of assumption (4.9) on the refresh rates, the upper bound seems a natural restriction for the techniques we will use, as these take advantage of the piecewise deterministic nature of the process, whereas an unbounded refresh rate would make the process look more like a diffusion. On the other hand assuming a lower bound on the refresh rate seems arbitrary given that we can study the behaviour of the process when the refresh rate decays to zero (as in the first assumption of Theorem 4.3.3).

Before we begin our analysis, note that every petite set C is bounded, due to the fact the the process can only have a finite speed.

Lemma 4.3.4. *For any Zig-Zag process in $\mathbb{R}^d$, any petite set C is bounded.*

Proof. Suppose that C is small and not bounded so there exists $(x_n, \theta_n) \in C$ with $|x_n| \to \infty$. Let c, ν, a as in the (2.10). Fix $m \in \mathbb{N}$, consider the ball A_m centred at 0, with radius m and for every n, consider the Euclidean distance between x_n and A_m, $d(x_n, A_m)$. Let $t_n = \dfrac{d(x_n, A_m)}{\sqrt{d}}$, the least amount of time the process needs to hit A_m, starting from x_n.
If $t < t_n$, then $P^t((x_n, \theta_n), A_m) = 0$ so

$$c\,\nu(A_m) \leq \int_0^{+\infty} P^t((x_n, \theta_n), A_m)(a(dt)) = \int_{t_n}^{+\infty} P^t((x_n, \theta_n), A_m)(da(t)) \leq a(t_n, +\infty).$$

Let $n \to \infty$ and we get that $\nu(A_m) = 0$. Since this is true for all m and $(A_m)_{m\geq 0}$ cover $\mathbb{R}^d$, we get that ν is the trivial measure, which contradicts our assumption. □

Recall from Theorem 2.1.5 that in order to prove that the process is not geometrically ergodic, we may prove that no petite set has the SGR property (introduced in (2.7)). Lemma 4.3.4 implies that in order to prove that every petite set does not satisfy the SGR property, we only need to consider the first hitting time of bounded sets. For that reason, we may start the process from a point in the boundary of the set and prove that the first hitting time of the set has no finite Moment Generating Function. We give the following definition, which is similar to the one given in (2.7) in Chapter 2.

Definition 4.3.5. *Let $Z_t = (X_t, \Theta_t)_{t\geq 0} = (X_t^{(1)}, ..., X_t^{(d)}; \Theta_t^{(1)}, ..., \Theta_t^{(d)})_{t\geq 0}$ a ZZ process on $\mathbb{R}^d \times \{-1, +1\}^d$. For a point $(x, \theta) = (x_1, ..., x_d; \theta_1, ..., \theta_d)$ with $\theta_i = +1$ and for a coordinate i, we define*

$$\tau_x^i = \inf\{t \geq \delta : X_t^{(i)} \leq x_i\}.$$

The point (x, θ) is said to have the Self Geometric Recurrence (SGR) property if there exists a $\delta > 0$, a coordinate i and a $b > 1$ such that

$$\mathbb{E}_{(x,\theta)}[b^{\tau_x^i}] < \infty$$

The equivalent definition is given for point (x, θ) when $\theta_i = -1$, where in that case $\tau_x^i = \inf\{t \geq \delta : X_t^{(i)} \geq x_i\}$.

Note that technically τ_x^i depends on the quantity δ. However, for presentation reasons we will omit writing this dependence. At the same time, throughout most of this chapter we will focus on the one-dimensional process, therefore, we will also omit the dependence of τ_x on the coordinate i. Using this terminology, in order to

prove that the process is not Geometrically ergodic, we need to prove that no point $(x, \theta) \in \mathbb{R} \times \{-1, +1\}$ has the SGR property. The remaining section is organised as follows. In subsection 4.3.1 we prove that the one dimensional Zig-Zag process has no point with the SGR property when the refresh rate decays to zero. We extend this result to the case where there is a non-zero, constant refresh rate in subsection 4.3.2. We then generalise to non-constant refresh rates, bounded above and below away from 0 in subsection 4.3.3. In subsection 4.3.4 we conclude the proof that the one-dimensional process lacks Geometric ergodicity and in subsection 4.3.5 we generalise the result to higher dimensions.

4.3.1 No Refreshment

We first assume that the process has a negligible refresh rate, i.e. $\gamma(x) \xrightarrow{\|x\| \to \infty} 0$. In this case it is not hard to establish the following.

Proposition 4.3.6 (No Refresh)**.** *Assume that* $U'(x) \xrightarrow{x \to +\infty} 0$ *and* $\gamma(x) \xrightarrow{x \to +\infty} 0$*. Then, for the one dimensional Zig-Zag process, no point* $(x, +1)$ *has the SGR property.*

Proof. Suppose that there exists a $b > 1$, $\delta > 0$ and $x \in \mathbb{R}$, such that $\mathbb{E}_{(x,+1)}[b^{\tau_x}] < \infty$. Let us start the process from $(x, +1)$. Pick $\epsilon < \log b$ and take y large enough so that $U'(z) < \epsilon$ for all $z \geq y$. From boundedness of U' on $[x, y]$ there is a positive probability that the process will reach $(y, +1)$ without switching and in this event the process will return to y before returning to x. Using the Strong Markov Property we get that $\mathbb{E}_{(y,+1)}[b^{\tau_y}] < \infty$.

Let T be the first time the process starting from $(y, +1)$ switches the direction to -1. Then $\tau_y \geq T$ and therefore it suffices to prove that $\mathbb{E}_{(y,+1)}[b^T] = \infty$. Since up to time T the process is on $[y, +\infty)$ we can write, for all $n \in \mathbb{N}$

$$\mathbb{E}_{(y,+1)}[b^T] \geq b^n \mathbb{P}_{(y,+1)}(T \geq n) = b^n \exp\{-\int_y^{y+n} \lambda(z, +1) dz\} \geq$$
$$\geq b^n \exp\{-n\epsilon\} = (\frac{b}{\exp\{\epsilon\}})^n \xrightarrow{n \to \infty} +\infty$$

since $\dfrac{b}{\exp\{\epsilon\}} > 1$. This proves the result. □

Remark 4.3.7. *Due to symmetry, the same lack of SGR would hold for points* $(x, -1)$ *if*
$\lim_{x \to -\infty} |U'(x)| = 0$ *and* $\lim_{x \to -\infty} \gamma(x) = 0$.

4.3.2 Constant Refresh Rate in One Dimension

Here we will prove that, in a heavy tail scenario, the one-dimensional Zig-Zag lacks SGR when we allow a constant, non zero refresh rate on switching direction. We have the following

Proposition 4.3.8 (Constant Refresh Rate)**.** *Assume that* $\lim_{x\to+\infty} U'(x) = 0$ *and that there exists an* x_0 *such that for all* $x \geq x_0$, $\gamma(x) = \gamma > 0$*. Then, for the Zig-Zag process, no point* $(x, +1)$ *has the SGR property.*

As in the previous section it will become evident from the proof that the same result holds if the $\lim_{x\to-\infty} U'(x) = 0$ and $\gamma(x) = \gamma$ for $x < -x_0$.

The difference to the case considered in subsection 4.3.1, which makes the analysis harder, is that the extra refresh rate γ ensures that, even if the process is far to the right of the petite set C, it will switch after a reasonable period of time and start moving back towards C. However, the same rate γ will apply again while the process moves towards C and after some period of time the process will switch again and start moving away from C. Ultimately the process will move up and down many times before returning to C (as illustrated in Figure 4.5). We will prove the lack of SGR by trying to control the number of times the process switches direction before it hits C. To get a better understanding of how the process behaves, we will start the next paragraph by considering the simpler case where $U'(x) = 0$ for all $x \geq x_0$. In this special case, we can use results from the theory of symmetric Random Walks. Although this simple setting never arises in practice, we will use it to study the model where $U'(x) = \epsilon$ for some small ϵ, for all $x \geq x_0$. Having established that, we will then use these results to prove lack of SGR.

A Bounding Model, The Case where U' is Constant, with Constant Refresh Rate

To get a better understanding of how the process will behave, let's assume that $U'(x) = 0$ for all $x \geq x_0$. Let us, also, assume that the process starts from $(x, +1)$. In this simple case, the process moves upwards until the point $x + U_1$ for some $U_1 \sim \exp(\gamma)$ and then proceeds to move downwards until the point $x + U_1 - D_1$ for some $D_1 \sim \exp(\gamma)$. Then it goes up until the point $x + U_1 - D_1 + U_2$ for some $U_2 \sim \exp\{\gamma\}$ and in general the process keeps moving upwards and downwards according to iid exponentials U_k, D_k. Note that

$$S_n = x + \sum_{k=1}^{n} (U_k - D_k) \tag{4.10}$$

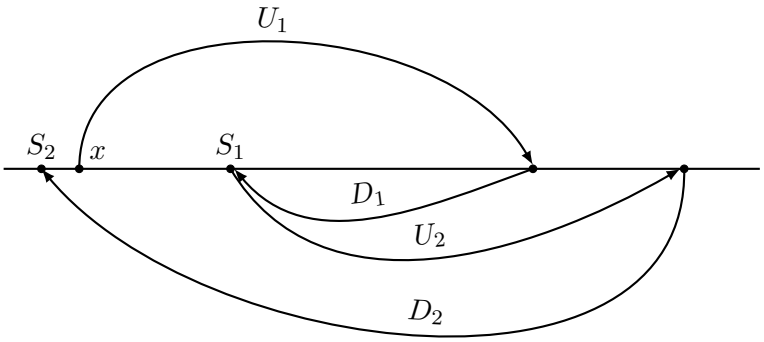


Figure 4.5: *A representation of the movement of the process. U_1 is the first up-step, D_1 the first down-step , U_2 the second up-step etc. The random walk S_n is the position of the process after the n'th down-step. T^0 is the first time the walk becomes less than the starting point x. In this configuration $T^0 = 2$.*

is a Random walk with increments the difference of two independent exponentials. Figure 4.5 provides an illustration of the up-steps and down-steps of the process. We observe that the time it takes for the Zig-Zag process to return to x is analogous to the time it takes the random walk to return to x. Our goal is to study the total time the process needs to return back to x, therefore we define $T^0 = \{n \geq 1 : S_n \leq x\}$ the first hitting time of $(-\infty, x]$ of the Random Walk. There is a vast literature on the behaviour of Random walk and the following Theorem is Theorem 2 of [And55].

Theorem 4.3.9 (Andersen 1955)**.** *Consider a Random Walk of the form $S_n = \sum_{k=1}^{n} X_k$ where $X_k \in \mathbb{R}$ i.i.d. such that $\mathbb{P}(S_k > 0) = 1/2$ for all $k \in \mathbb{N}$. Then, for all n,*

$$\mathbb{P}(S_1 > 0, ..., S_n > 0) = \frac{1}{4^n}\binom{2n}{n}. \tag{4.11}$$

Remark 4.3.10. *Note the surprising result that the probability the walk will stay on the positive side through the first n steps does not depend at all on the distribution of the jumps, as long as the probability of being positive remains $1/2$ in each step. In our case the jumps are the difference of two independent exponential distributions with the same parameter, which clearly satisfies the conditions of the Theorem.*

In a practical setting, U' cannot be identically equal to 0 as $\pi(dx) = \exp\{-U(x)\}dx$ wouldn't be a finite measure in that case. Let us now consider the case where $U'(x) = \epsilon > 0$ for all $x \geq x \geq x_0$ and for some very small ϵ. The behaviour of this Zig-Zag is encaptured by what we shall call **ϵ-model**. In this model, we have

a Random Walk starting from x with increments $U_k - D_k$, where $U_k \sim \exp(\gamma + \epsilon)$, $D_k \sim \exp(\gamma)$. Let T^ϵ be the first time this Random Walk hits the set $(-\infty, x]$. For $\epsilon = 0$ the distribution of T^0 was fully characterised by the previous theorem. The addition of a positive ϵ induces a drift that forces the walk towards the negative numbers. However, we can still control the law of the first return time with the following.

Lemma 4.3.11 (Hitting Lower Bound)**.** *If T^ϵ is the first hitting time of $(-\infty, x]$ for the ϵ-model, then there exists a constant $C > 0$ such that for all $n \in \mathbb{N}$.*

$$\mathbb{P}\left(T^\epsilon \geq n+1\right) \geq C \frac{1}{n^{3/2}} \left(\frac{\gamma}{\gamma+\epsilon}\right)^{n+1}. \tag{4.12}$$

Proof of Lemma 4.3.11. In order to simulate each one of the down-steps of the ϵ-model we need to draw an $\exp(\gamma)$. Since this distribution is the first arrival time of a Poisson Process with rate γ, we can use Poisson Thinning, with the bounding process having a rate $\gamma + \epsilon$. Each time this process generates a point we accept it with probability $\dfrac{\gamma}{\gamma+\epsilon}$ and arrival times between any two consecutive points follow an $\exp(\gamma + \epsilon)$. Therefore, we can draw a down-step D_k by first drawing an $N_k \sim Geom\left(\dfrac{\gamma}{\gamma+\epsilon}\right)$ and then draw N_k i.i.d. random variables $B_{k,1}, .., B_{k,N_k} \sim \exp(\gamma + \epsilon)$ and set $D_k = \sum_{i=1}^{N_k} B_{k,i}$ (see Figure 4.6 for example). All the $B_{k,i}$ and N_k are independent to each other. We also draw the up-steps $U_k \sim \exp(\gamma + \epsilon)$.

Consider the event A_n defined as: *$B_{k,1}$ and U_k for $k = 1, .., n$ took values such that the Random Walk with increments $\{U_k - B_{k,1}\ , k = 1, 2, ...\}$ does not hit $(-\infty, x]$ on the first n steps.* Since $B_{k,1}, U_k \sim \exp(\gamma + \epsilon)$, this is a Random Walk that satisfies the conditions of Theorem 4.3.9, therefore we know that there exists a constant $C > 0$ such that

$$\mathbb{P}(A_n) = \frac{1}{4^n}\binom{2n}{n} \geq C\frac{1}{n^{1/2}},$$

where in the second equality we use the Stirling formula

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

where we write $a_n \sim b_n$ to denote that $\lim_{n\to\infty} a_n/b_n = c > 0$ for some constant c. Then, the event that the ϵ-model does not hit $(-\infty, x]$ on the first n steps includes the event $\{N_1 = ... = N_n = 1\} \cap A_n$. Since $N_k \sim Geom\left(\dfrac{\gamma}{\gamma+\epsilon}\right)$ are i.i.d. and

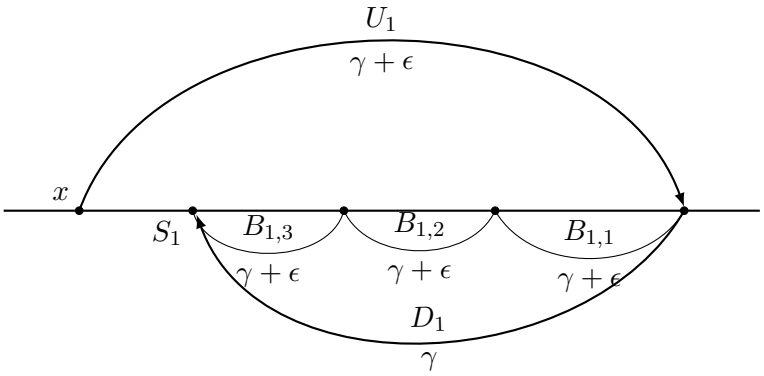


Figure 4.6: *A representation of the Poisson thinning construction for the down-steps used in the proof of Lemma 4.3.11. The first up-step U_1 is simulated according to* $\exp\{\gamma+\epsilon\}$. *A first down-step is proposed to be $B_{1,1}$. It follows an* $\exp\{\gamma+\epsilon\}$ *and is accepted with probability $\gamma/(\gamma+\epsilon)$, or rejected. If accepted, then the first down-step is set to be $D_1 = B_{1,1}$. In this configuration it was rejected, so a second down-step $B_{1,2}$ was proposed, independently of the previous one according to* $\exp\{\gamma+\epsilon\}$. *This was, again, accepted with probability $\gamma/(\gamma+\epsilon)$, or rejected. If rejected, a third proposed down-step is simulated, etc. Overall, $N_1 \sim Geom\left(\frac{\gamma}{\gamma+\epsilon}\right)$ proposed down-steps are drawn and the first down-step of the process is set to be $D_1 = \sum_{i=1}^{N_1} B_{1,i}$. Overall, $D_1 \sim \exp\{\gamma\}$. In this configuration $N_1 = 3$.*

independent of A_n we get

$$\mathbb{P}(T^\epsilon \geq n+1) \geq \left(\frac{\gamma}{\gamma+\epsilon}\right)^n \mathbb{P}(A_n) \geq C\frac{1}{n^{1/2}}\left(\frac{\gamma}{\gamma+\epsilon}\right)^n \tag{4.13}$$

which proves the Lemma. □

Before we proceed to prove lack of SGR for the Zig-Zag process, we need a way to compare the Zig-Zag process for which $U'(x) < \epsilon$ for all x large enough with the ϵ-model. This will allow us to use Lemma 4.3.11 in the context of Zig-Zag. The following proposition proves that the return time of the ϵ-model is stochastically dominated by the one of the Zig-Zag process in the case where $U'(x) < \epsilon$.

Lemma 4.3.12. *Suppose that two Zig-Zag processes Z_1, Z_2 evolve in $\mathbb{R}$, starting from the same point $(x, +1)$. Let a_i, b_i be the rate functions of the up-steps/down-steps of the process Z_i. Assume that $a_1(y) \leq a_2(y)$ and $b_1(y) \geq b_2(y)$ for all $y \geq x$. Let T^i be the first positive time the Random Walk induced by the i'th Zig-Zag hits $(-\infty, x]$. Then, for all $n \in \mathbb{N}$, $\mathbb{P}(T^1 \geq n) \geq \mathbb{P}(T^2 \geq n)$*

Proof of Lemma 4.3.12. Let $A_1^{(i)}, A_2^{(i)}, ..$ and $B_1^{(i)}, B_2^{(i)}, ..$ be the Up-steps/Down-steps of Z_i. Let $S_n^{(i)} = x + \sum_{k=1}^n \left(A_k^{(i)} - B_k^{(i)}\right)$. From the Exponential Representation of the Poisson process, Proposition 2.2.2, for some $E_1, E_2, ...$ and $F_1, F_2, .. \sim$ exp(1) i.i.d. and for all $n \in \mathbb{N}$,

$$A_n^{(i)} = \inf\{t \geq 0 : \int_{S_{n-1}^{(i)}}^t a_i(s)ds = E_n\} \tag{4.14}$$

and

$$B_n^{(i)} = \inf\{t \geq 0 : \int_{S_{n-1}^{(i)}+A_n^{(i)}-t}^{S_{n-1}^{(i)}+A_n^{(i)}} b_i(s)ds = F_n\}. \tag{4.15}$$

We will prove by induction that $S^{(2)}(n) \leq S^{(1)}(n)$ for all n which implies the result. More precisely, the induction statement will be:

$$S_n^{(1)} \geq S_n^{(2)} \text{ and } S_n^{(1)} + A_{n+1}^{(1)} \geq S_n^{(2)} + A_{n+1}^{(2)} \text{ for all } n \in \mathbb{N}.$$

For $n = 0$, $S_0^{(1)} = x = S_0^{(2)}$ and since $a_1(s) < a_2(s)$ we have from (4.14), $A_1^{(1)} \geq A_1^{(2)}$.

Now, assume the induction hypothesis. That is, assume that for some n, $S_{n-1}^{(1)} \geq S_{n-1}^{(2)}$ and $S_{n-1}^{(1)} + A_n^{(1)} \geq S_{n-1}^{(2)} + A_n^{(2)}$. We have from (4.15),

$$\int_{S_n^{(2)}}^{S_{n-1}^{(2)}+A_n^{(2)}} b_2(s)ds = F_n = \int_{S_n^{(1)}}^{S_{n-1}^{(1)}+A_n^{(1)}} b_1(s)ds \geq \int_{S_n^{(1)}}^{S_{n-1}^{(1)}+A_n^{(1)}} b_2(s)ds \geq \int_{S_n^{(1)}}^{S_{n-1}^{(2)}+A_n^{(2)}} b_2(s)ds$$

So $S_n^{(1)} \geq S_n^{(2)}$. Also, from (4.14),

$$\int_{S_n^{(2)}}^{S_n^{(2)}+A_{n+1}^{(2)}} a_2(s)ds = E_{n+1} = \int_{S_n^{(1)}}^{S_n^{(1)}+A_{n+1}^{(1)}} a_1(s)ds \leq$$
$$\leq \int_{S_n^{(1)}}^{S_n^{(1)}+A_{n+1}^{(1)}} a_2(s)ds \leq \int_{S_n^{(2)}}^{S_n^{(1)}+A_{n+1}^{(1)}} a_2(s)ds$$

therefore $S_n^{(2)} + A_{n+1}^{(2)} \leq S_n^{(1)} + A_{n+1}^{(1)}$ and the induction is complete. □

Let us assume that $\lim_{x\to+\infty} U'(x) = 0$ and $\gamma(x) = \gamma$ for x large enough. The use of the previous Lemma will become more clear once we observe that for any $\epsilon > 0$, there exists some x_0 such that for all $x \geq x_0$ we have

$$\lambda(x, +1) = [U'(x)]^+ + \gamma < \epsilon + \gamma$$

and

$$\lambda(x, -1) \geq \gamma.$$

The rates on the RHS these two equations are exactly the rates of the ϵ-model. Using Lemma 4.3.12 we can compare the probability that the Zig-Zag has not returned to the starting position x after n down-steps, with the respective probability of an ϵ-model (for any ϵ we choose). We already know how to control this probability due to Lemma 4.3.11. In the next paragraph, we will use these results to establish SGR for the Zig-Zag process with constant refresh rate, but where we allow U' to have a general form, as long as it decays to zero.

The Case where U' is Not Constant, with Constant Refresh Rate

In the previous paragraph, we studied the behaviour of the Zig-Zag process, when both the refresh rate $\gamma(x)$ and $U'(x)$ were constants. This is a very restrictive assumption for U'. In this paragraph, we will prove that every point $(x, +1)$ lacks the SGR property in the general U' case, where we assume that

$$\lim_{x\to+\infty} U'(x) = 0.$$

We will, however, still impose the condition that the refresh rate is a constant, i.e. there exists an $x_0, \gamma > 0$ such that $\gamma(x) = \gamma > 0$ for all $x \geq x_0$. This will finish the proof of Proposition 4.3.8, which is the main goal of this subsection.

The first step is to use the FKG inequality to get a lower bound for the moment generating function of the return time in terms of the law of T^ϵ for some ϵ

small enough. The inequality was first proved in [FKG71] for probability measures on finite distributive lattices. Heuristically, it says the following. Consider the state space of all possible $0-1$ labellings of a finite set Λ. The FKG inequality proves the natural observation that if two real functions of the state space take larger values on larger labellings, they should be positively correlated. This was later extended in [Pre74] to the case where the state space is X^{Λ}, for some continuous set of labels X. This includes the case $X = [0, +\infty)$, which will be the case in this sub-section. Although [Pre74] don't directly prove FKG, they prove Holley's inequality, which is used in the proof of FKG in [Hol74]. The continuous labelling FKG can then be deduced using the same arguments as in [Hol74]. We begin by introducing the setting of the theorem.

Suppose that Λ is a finite set and for all $t \in \Lambda$, $(X_t, \mathcal{F}_t, \omega_t)$ is probability measure space. Let $X = \prod_{t\in\Lambda} X_t$, $\mathcal{F} = \prod_{t\in\Lambda} \mathcal{F}_t$, $\omega = \prod_{t\in\Lambda} \omega_t$. For $x = (x_1, .., x_n), y = (y_1, .., y_n) \in X$ write $x \wedge y = (\min\{x_1, y_1\}, ..., \min\{x_n, y_n\})$ and $x \vee y = (\max\{x_1, y_1\}, ..., \max\{x_n, y_n\})$. Assume that there is a total ordering $\leq$ in every X_t which induces a partial ordering on X, defining $x \leq y \iff x_t \leq y_t$ for all $t \in \Lambda$.

Definition 4.3.13. *We call a function $g : X \to \mathbb{R}$ increasing if for all $x \leq y$, $g(x) \leq g(y)$.*

Theorem 4.3.14 (FKG Inequality for Continuous Spins, Preston 1974)**.** *Suppose μ is the probability measure on X with density $f : X \to \mathbb{R}$ with respect to ω and assume that f satisfies the Lattice Condition*

$$f(x \vee y)f(x \wedge y) \geq f(x)f(y), \text{ for all } x, y \in X \tag{4.16}$$

Then, for any increasing functions $g, h : X \to \mathbb{R}$ where g is bounded and $\mu(|h|) < \infty$ we have

$$\mathbb{E}_{\mu}[gh] \geq \mathbb{E}_{\mu}[g]\mathbb{E}_{\mu}[h]. \tag{4.17}$$

An immediate Corollary of the Generalised FKG inequality is the following proposition. This proposition gives a lower bound on the moment generating function of the up-steps of the Zig-Zag, in terms of the probability that the Random Walk will not have returned to its origin. Heuristically, the fact that the Random Walk, starting from x_0, has not hit the set $(-\infty, x_0]$ in the first n steps, means that the first n up-steps must have been larger than the typical up-steps the Zig-Zag would have had. We have the following.

Proposition 4.3.15. *Assume that for a one-dimensional Zig-Zag process there exist $x_0, \gamma, \epsilon > 0$ such that for all $y \geq x_0$ the refresh rate satisfies $\gamma(y) = \gamma > 0$ and $U'(y) < \epsilon$. Assume that the process starts from $(x, +1)$, with $x \geq x_0$. Let U_k, D_k denote that up-steps and down-steps of the Zig-Zag process (illustrated in Figure 4.5) and suppose T is the first time the process $S_n = x + \sum_{k=1}^{n} (U_k - D_k)$ hits the interval $(-\infty, x]$. Then, for any $b > 1$ and any $n \in \mathbb{N}$,*

$$\mathbb{E}\left[b^{\sum_{k=1}^{T-1} U_k}\right] \geq \left(1 - \frac{\log b}{\gamma + \epsilon}\right)^{-n} \mathbb{P}\left(T > n\right). \tag{4.18}$$

Proof. Since the up-steps and the down-steps of the Zig-Zag are the first arrival times of Poisson Processes, we can use the exponential representation to generate the process using i.i.d. $E_1, E_2, ...$ and $F_1, F_2, ... \sim \exp(1)$ for the up-steps and the down-steps respectively. Then, $U_k = \inf\{t \geq 0 : \int_{S_{k-1}}^{S_{k-1}+t} \lambda(s, +1)ds \geq E_k\}$ and $D_k = \inf\{t \geq 0 : \int_{S_{k-1}+U_k-t}^{S_{k-1}+U_k} \lambda(s, -1)ds \geq F_k\}$. Consider also the Random Variables $A_k = \inf\{t \geq 0 : \int_{S_{n-1}}^{S_{n-1}+t} \gamma + \epsilon \, ds = E_k\} \leq \int_{S_{n-1}}^{S_{n-1}+t} \lambda(s, +1)ds = E_k\} = U_k$. Note that $A_k \sim \exp(\gamma + \epsilon)$. Then, for every $n \in \mathbb{N}$, we write

$$\mathbb{E}_x\left[b^{\sum_{k=1}^{T-1} U_k}\right] \geq \mathbb{E}_x\left[b^{\sum_{k=1}^{T-1} U_k} 1_{T>n}\right] \geq \mathbb{E}_x\left[b^{\sum_{k=1}^{n} U_k} 1_{T>n}\right] \geq \mathbb{E}_x\left[b^{\sum_{k=1}^{n} A_k} 1_{T>n}\right].$$

Note that the functions inside the last expectation are functions of $E_1, ..., E_n, F_1, ..., F_n$, in that if one has access to the values of these $2n$ random variables,then the value of $b^{\sum_{k=1}^{n} A_k} 1_{T>n}$ is deterministic. Now, let us condition on some realisation of the random values $F_1, ..., F_n$. Using the Exponential Representation, we observe that both $b^{\sum_{k=1}^{n} A_k}$ and $1_{T>n}$ are increasing functions on the underlying probability space $([0, +\infty)^n, \mathcal{B}([0, +\infty)^n, \prod_{k=1}^{n} \exp(1))$, the one induced by $E_1, ..., E_n$. The Lattice Condition holds for this space as the measure is the product of the laws of the independent exponentials $E_1, ..., E_n$. FKG inequality, i.e. Theorem 4.3.14 gives

$$\begin{aligned}
&\mathbb{E}_x\left[b^{\sum_{k=1}^{n} A_k} 1_{T>n} | F_1, .., F_n\right] \geq \mathbb{E}_x\left[b^{\sum_{k=1}^{n} A_k} | F_1, .., F_n\right] \mathbb{E}_x\left[1_{T>n} | F_1, .., F_n\right] = \\
&= \mathbb{E}_x\left[b^{\sum_{k=1}^{n} A_k}\right] \mathbb{E}_x\left[1_{T>n} | F_1, .., F_n\right] = \left(1 - \frac{\log b}{\gamma + \epsilon}\right)^{-n} \mathbb{E}_x\left[1_{T>n} | F_1, .., F_n\right].
\end{aligned}$$

where the first equality is due to independence between A_k and $F_1, .., F_k$. Therefore, on integrating over $F_1, .., F_n$ we get

$$\mathbb{E}_x\left[b^{\sum_{k=1}^{n} U_k} 1_{T^\epsilon>n}\right] \geq \left(1 - \frac{\log b}{\gamma + \epsilon}\right)^{-n} \mathbb{P}(T > n). \qquad \square$$

Now we are ready to prove Proposition 4.3.8.

Proof of Proposition 4.3.8. Suppose that there exists an x, $b > 1$ and some $\delta > 0$ such that, if $\tau_x = \inf\{t \geq \delta : X_t \in (-\infty, x]\}$, then

$$\mathbb{E}_{x,+1}[b^{\tau_x}] < \infty. \tag{4.19}$$

Pick $\epsilon < \log b$. Using the same argument as in the proof of Proposition 4.3.6 we may assume without loss of generality that for all $x \geq x_0$ and therefore for all $y \geq x$, $|U'(y)| < \epsilon < \log b$ and $\gamma(y) = \gamma$. Assume that the process starting from $(x, +1)$ and is decomposed to its up-steps and down-steps $U_1, U_2, ...$ and $D_1, D_2, ...$ Let T be the number of steps needed for the Random Walk $S_n = x + \sum_{k=1}^{n} (U_k - D_k)$ to hit $(-\infty, x]$. Then, note that $\sum_{k=1}^{T-1} (U_k + D_k) \leq \tau_x$. From equation (4.19) we get

$$\mathbb{E}\left[b^{\sum_{k=1}^{T-1}(U_k+D_k)}\right] < \infty.$$

Now, since $|U'(y)| < \epsilon$ on $[x, \infty)$, note that $\lambda(y, +1) = \gamma + [U'(y)]^+ \leq \gamma + \epsilon$ and $\lambda(y, -1) = \gamma + [-U'(y)]^+ \geq \gamma$. Using Lemma 4.3.12 we get that, if T^ϵ is the first return time of the ϵ-model, then for any $n \in \mathbb{N}$, $\mathbb{P}(T > n) \geq \mathbb{P}(T^\epsilon > n)$. Using Proposition 4.3.15 we get, for any $n \in \mathbb{N}$,

$$\mathbb{E}[b^{\sum_{k=1}^{T-1} U_k}] \geq \left(1 - \frac{\log b}{\gamma + \epsilon}\right)^{-n} \mathbb{P}(T > n) \geq \left(1 - \frac{\log b}{\gamma + \epsilon}\right)^{-n} \mathbb{P}(T^\epsilon > n).$$

Using Lemma 4.3.11 we can find a $C > 0$ such that for all $n \in \mathbb{N}$,

$$\mathbb{E}[b^{\tau_x}] \geq \mathbb{E}[b^{\sum_{k=1}^{T-1} U_k}] \geq C\left(1 - \frac{\log b}{\gamma + \epsilon}\right)^{-n} \left(\frac{\gamma}{\gamma + \epsilon}\right)^{n} \frac{1}{n^{1/2}} \geq C\left(\frac{\gamma}{\gamma + \epsilon - \log b}\right)^{n} \frac{1}{n^{1/2}}.$$

Since $\epsilon < \log b$, the last quantity diverges to $+\infty$ as we let $n \to \infty$. This proves the contradiction. □

4.3.3 Bounded Refresh Rate

So far we have considered the case where the refresh rate is a constant and we have proven that no point $(x, +1)$ has the SGR property. We will now generalise the lack of SGR result in the more general case where for the refresh rate γ, there exist $0 < m < M$ and $x_0 > 0$: for all $x \geq x_0$,

$$m \leq \gamma(x) \leq M$$

as assumed in the second assumption of Theorem 4.3.3. Before this, we need the following lemma that will allow us to use FKG inequality again.

Lemma 4.3.16. *Consider a Zig-Zag process $Z(t)$ starting from $(x, +1)$. Suppose that the up-steps and the down-steps (illustrated in figure 4.5) are generated using the exponential representation Proposition 2.2.2, with i.i.d. $E_1, E_2, \dots$, $F_1, F_2, .. \sim \exp(1)$ respectively. Consider U_k, D_k the k'th up-step and down-step respectively, let $S_n = x + \sum_{k=1}^{n} (U_k - D_k)$ and $T = \inf\{n \geq 1 : S_n \leq x\}$. Suppose that we condition on the value of $F_1, ..., F_n$ for some $n \in \mathbb{N}$. Then, the random variable $1_{T \geq n+1}$ is an increasing function of $(E_1, ..., E_n)$.*

Proof. The proof will use induction and is very similar to the proof of Lemma 4.3.12. For $i = 1, 2$ let $E_1^i, E_2^i, ...$ a sequence of positive real numbers with $E_n^1 \leq E_n^2$ for all $n \in \mathbb{N}$. Let $F_1, F_2, ...$ another sequence of positive real numbers. Let $S_0^i = x$ and inductively define the sequences $U_1^i, U_2^i, ...$, $D_1^i, D_2^i, ...$ and $S_1^i, S_2^i, ...$as follows:

$$U_{n+1}^i = \inf\left\{ t \geq 0 : \int_{S_n^i}^{S_n^i + t} \lambda(s, +1) ds \geq E_{n+1}^i \right\}$$

and

$$D_{n+1}^i = \inf\left\{ t \geq 0 : \int_{S_{n+1}^i + U_{n+1}^i - t}^{S_{n+1}^i + U_{n+1}^i} \lambda(s, -1) ds \geq F_{n+1} \right\}$$

and

$$S_n^i = x + \sum_{k=1}^{n} \left(U_k^i - D_k^i \right).$$

We prove by induction the following statement:

$$\text{For all } n \in \mathbb{N},\ S_n^1 \leq S_n^2 \text{ and } S_n^1 + U_{n+1}^1 \leq S_n^2 + U_{n+1}^2.$$

For $n = 0$, $S_0^1 = x_0 = S_0^2$ and $S_0^1 + U_1^1 \leq S_0^2 + U_1^2$, since $E_1^1 \leq E_1^2$. Assume that the statement holds for some n. Then,

$$\int_{S_{n+1}^1}^{S_n^1 + U_{n+1}^1} \lambda(s, -1) ds = F_{n+1} = \int_{S_{n+1}^2}^{S_n^2 + U_{n+1}^2} \lambda(s, -1) ds \geq \int_{S_{n+1}^2}^{S_n^1 + U_{n+1}^1} \lambda(s, -1) ds$$

and therefore $S_{n+1}^1 \leq S_{n+1}^2$. Then, let $A = \inf\{t \geq 0 : \int_{S_{n+1}^1}^{S_{n+1}^1 + t} \lambda(s, +1) ds \geq E_{n+2}^2\}$. Then,

$$\int_{S_{n+1}^1}^{S_{n+1}^1 + A} \lambda(s, +1) ds = E_{n+2}^2 = \int_{S_{n+1}^2}^{S_{n+1}^2 + U_{n+2}^2} \lambda(s, +1) ds$$

and since $S_{n+1}^1 \leq S_{n+1}^2$ we have

$$S_{n+1}^1 + A \leq S_{n+1}^2 + U_{n+2}^2.$$

Also,

$$\int_{S^1_{n+1}}^{S^1_{n+1}+U^1_{n+2}} \lambda(s,+1)ds = E^1_{n+2} \leq E^2_{n+2} = \int_{S^1_{n+1}}^{S^1_{n+1}+A} \lambda(s,+1)ds$$

from which we conclude that

$$S^1_{n+1} + U^1_{n+2} \leq S^1_{n+1} + A \leq S^2_{n+1} + U^2_{n+2}$$

which proves the induction. This proves that for all $n \in \mathbb{N}$, $S^1_n \leq S^2_n$ which proves the result. □

We can now state and prove the main proposition of this subsection.

Proposition 4.3.17 (Bounded Rates)**.** *Assume that* $\lim_{x\to+\infty} U'(x) = 0$ *and there exists an* $x_0 > 0$ *and* $M > m > 0$ *such that for all* $x \geq x_0$*,* $m \leq \gamma(x) \leq M$*. Then, for the one-dimensional Zig-Zag process, no point* $(x, +1)$ *has the SGR property.*

Proof of Proposition 4.3.17. Assuming that SGR property is satisfied for some point $(x, +1)$, we can find $b > 1$ and $\delta > 0$ such that if $\tau_x = \inf\{t \geq \delta : X_t \leq x\}$, then $\mathbb{E}_{(x,+1)}[b^{\tau_x}] < \infty$.

Fix $\epsilon > 0$ small enough, to be defined later. Using similar arguments to the proof of Proposition 4.3.1 we can assume without loss of generality that x is large enough, therefore for all $y \geq x$, $U'(y) < \epsilon$ and $m \leq \gamma(y) \leq M$.

We will use the same argument as in Proposition 4.3.15 to find an appropriate lower bound for the $\mathbb{E}_{(x,+1)}[b^{\tau_x}]$. A difference will be that, since $\lambda(y,+1) = [U'(y)]^+ + \gamma(y) < \epsilon + M$, the up-steps will stochastically dominate $\exp(M+\epsilon)$ distributions. More precisely, we follow the proof of 4.3.15 as follows:

Let $S_0 = x$. We generate the up-steps and down-steps using i.i.d. $E_1, E_2, ...$ and $F_1, F_2, ... \sim \exp(1)$ respectively. Then, inductively we define

$$U_{n+1} = \inf\{t \geq 0 : \int_{S_n}^{S_n+t} \lambda(s,+1)\, ds \geq E_{n+1}\}$$

and

$$D_{n+1} = \inf\{t \geq 0 : \int_{S_n+U_{n+1}-t}^{S_n+U_{n+1}} \lambda(s,-1)\, ds \geq F_{n+1}\}$$

and

$$S_{n+1} = x + \sum_{k=1}^{n+1} U_k - D_k.$$

Consider also the random variables

$$A_n = \inf\{t \geq 0 : \int_{S_{n-1}}^{S_{n-1}+t} M + \epsilon \, ds \geq E_n\} \leq \int_{S_{n-1}}^{S_{n-1}+t} \lambda(s, +1) \, ds \geq E_n\} = U_n.$$

Note that $A_n \sim \exp(M + \epsilon)$. Then, for every $n \in \mathbb{N}$, we write

$$\mathbb{E}_x \left[b^{\sum_{k=1}^{T-1} U_k} \right] \geq \mathbb{E}_x \left[b^{\sum_{k=1}^{T-1} U_k} 1_{T>n} \right] \geq \mathbb{E}_x \left[b^{\sum_{k=1}^{n} U_k} 1_{T>n} \right] \geq \mathbb{E}_x \left[b^{\sum_{k=1}^{n} A_k} 1_{T>n} \right]. \tag{4.20}$$

Note that due to Lemma 4.3.16, conditioning on the values of $F_1, F_2, ..., F_n$, both $b^{\sum_{k=1}^{n} A_k}$ and $1_{T>n}$ are increasing functions of $E_1, ..., E_n$. Using the FKG inequality in the same fashion as in the previous section we get

$$\mathbb{E}_x \left[b^{\sum_{k=1}^{n} A_k} 1_{T>n} | F_1, .., F_n \right] \geq \mathbb{E}_x \left[b^{\sum_{k=1}^{n} A_k} | F_1, .., F_n \right] \mathbb{E}_x \left[1_{T>n} | F_1, .., F_n \right] =$$
$$= \mathbb{E}_x \left[b^{\sum_{k=1}^{n} A_k} \right] \mathbb{E}_x \left[1_{T>n} | F_1, .., F_n \right] = \left(\frac{M + \epsilon}{M + \epsilon - \log b} \right)^n \mathbb{E}_x \left[1_{T>n} | F_1, .., F_n \right]$$

and on integrating both sides over $F_1, .., F_n$ we get

$$\mathbb{E} \left[b^{\sum_{k=1}^{T-1} U_k} \right] \geq \left(\frac{M + \epsilon}{M + \epsilon - \log b} \right)^n \mathbb{P} \left(T > n \right)$$

Using the fact that $\tau_x \geq \sum_{k=1}^{T-1} U_k$, we get

$$\mathbb{E} \left[b^{\tau_x} \right] \geq \mathbb{E} \left[b^{\sum_{k=1}^{T-1} U_k} \right] \geq \left(\frac{M + \epsilon}{M + \epsilon - \log b} \right)^n \mathbb{P} \left(T > n \right). \tag{4.21}$$

Now, consider a bounding Zig-Zag process with up-stes $U_1', U_2', ...$ and down-steps $D_1', D_2', ...$ generated with refresh rate

$$\lambda'(y, +1) = \epsilon + \gamma(y)$$

and

$$\lambda'(y, -1) = \gamma(y)$$

respectively. Assume that the process starts from $(x, +1)$ and let $S_n' = x + \sum_{k=1}^{n} (U_k' - D_k')$. Let T^ϵ be the first hitting time of $(-\infty, x]$ for the process S_n'. We know that for all $y \geq x$

$$\lambda(y, +1) = [U'(y)]^+ + \gamma(y) \leq \epsilon + \gamma(y) = \lambda'(y, +1)$$

while

$$\lambda(y, -1) = [-U'(y)]^+ + \gamma(y) \geq \gamma(y) = \lambda'(y, -1).$$

Using Lemma 4.3.12 we get for all $n \in \mathbb{N}$, $\mathbb{P}(T > n) \geq \mathbb{P}(T^\epsilon > n)$ and due to (4.21) we get

$$\mathbb{E}\left[b^{\tau_x}\right] \geq \mathbb{E}\left[b^{\sum_{k=1}^{T-1} U_k}\right] \geq \left(\frac{M+\epsilon}{M+\epsilon-\log b}\right)^n \mathbb{P}(T^\epsilon > n). \tag{4.22}$$

Now, the goal is to bound from below the probability of the RHS. We will construct this process following the idea of the proof of Lemma 4.3.11. We draw the first up-step U_1' as the first time of the Poisson Process with intensity $\{\epsilon + \gamma(z), z \geq x\}$. The first down-step D_1' is the first time of the Poisson Process with intensity $\{\gamma(x + U_1 - z), z \geq 0\}$. Instead, we will use Poisson Thinning and we will draw the down-step using the following algorithm (see also Figure 4.7):

1. Set $D_{prop} = 0, D_1 = 0$ and $D^* = x + U_1$.
2. Draw a new D_{prop} according to the first arrival of the Poisson Process with rate $\epsilon + \gamma(D^* - D_1 - z),\ z \geq 0$.
3. Set $D_1 = D_1 + D_{prop}$.
4. Accept D_{prop} with probability $\dfrac{\gamma(D^* - D_1)}{\gamma(D^* - D_1) + \epsilon}$.
5. If we reject, draw a new D_{prop} according to step 2.

After the down-step D_1 is completed, the process is at point $x + U_1 - D_1$ and we start again by drawing U_2 according to $\{\gamma(x + U_1 - D_1 + z), z \geq 0\}$ and D_2 using the above algorithm but with $D^* = x + U_1 - D_1 + U_2$. Define U_n, D_n in a similar way.

Note that if after the proposed down-step D_{prop} the process moves to z, then this proposed down-step is accepted with probability $\gamma(z)/(\gamma(z) + \epsilon) \geq m/(m + \epsilon)$. Note as well that a process that has lower acceptance probability of the proposed down-steps will tend to have larger down-steps. This in turn means that a process with lower acceptance probability will return to the starting point faster. Since we want to bound from below the probability that the process will not have returned to the starting point after a certain number of down-steps, we can assume without loss of generality that the acceptance probability of all the down-steps is $m/(m + \epsilon)$ and independent of the proposed down-steps.

Now consider a process with symmetric up-steps and down-steps. This process one that starts from x and makes successive up-steps U_n'' according to intensity

$$\lambda''(x, +1) = \gamma(x) + \epsilon \tag{4.23}$$

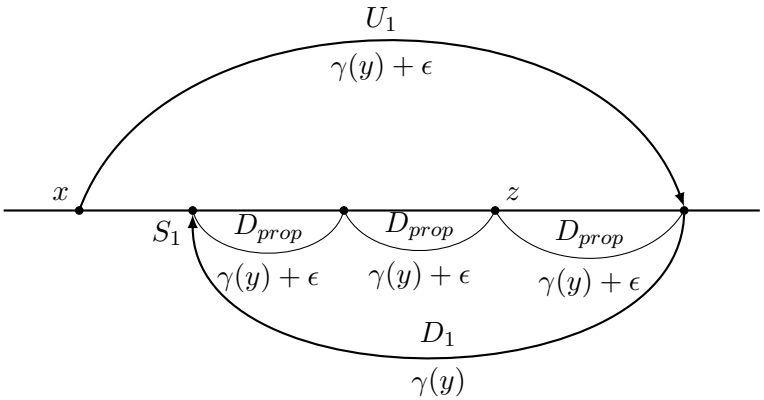


Figure 4.7: *Representation of the Poisson thinning construction of the down-steps. The procedure is very similar to the procedure discussed in Figure 4.6 except that the rates of the Poisson processes are not constant. The first up-step U_1 is simulated according to the hazard rate $\gamma(y)+\epsilon$. For the first down-step we use the following procedure. A proposed down-step D_{prop} with rate hazard $\gamma(y)+\epsilon$ is simulated and is accepted with probability $\frac{\gamma(z)}{\gamma(z)+\epsilon}$, where z is the position of the process after this proposed down-step is completed. If this proposed down-step is accepted, then this is set to be the first down-step. If the down-step is rejected, the process starts again from z and a new proposed down-step D_{prop} is simulated according to the hazard rate $\gamma(y)+\epsilon$. This new proposed down-step is either accepted or rejected as before. This procedure continues until some proposed down-step is accepted. The process' first down-step D_1 is the sum of all the rejected down-steps with the addition of the accepted one and overall D_1 is the first arrival time of the Poisson process with hazard rate $\gamma(y)$. In this configuration, two proposed down-steps got rejected and the third one was accepted.*

and down-steps D''_n according to the same intensity

$$\lambda''(x,-1) = \gamma(x) + \epsilon \tag{4.24}$$

and let $S''_n = x + \sum_{k=1}^{n} (U''_k - D''_k)$. Consider, as usual the random variables

$$E_n^+ = \int_{S''_{n-1}}^{S'_{n-1}+U''_n} \gamma(z) + \epsilon \, dz \tag{4.25}$$

and

$$E_n^- = \int_{S''_n}^{S''_{n-1}+U''_n} \gamma(z) + \epsilon \, dz. \tag{4.26}$$

Then from the Exponential Representation of Poisson Process, $E_k^+, E_k^- \sim \exp(1)$ i.i.d. Consider the walk

$$Y_n = \sum_{k=1} \left(E_k^+ - E_k^- \right)$$

Because of (4.25) (4.26) we have

$$Y_n = \int_x^{S'_n} \gamma(x) + \epsilon \, dx.$$

Then the first time T^0 when the process S''_n hits $(-\infty, x]$ is the same as the first time when the process Y_n, starting from 0, hits $(-\infty, 0]$. Now, Y_n is a symmetric random walk, therefore, using Proposition 4.3.9 for the one dimensional random walk without drift, we get that there exists a constant $C > 0$ such that for every $n \in \mathbb{N}$,

$$\mathbb{P}(T^0 > n) \geq C \frac{1}{n^{1/2}}$$

Fix $n \in \mathbb{N}$. Recall that we have assumed that for our bounding Random Walk S'_n, every proposed down-step is accepted with probability $m/(m+\epsilon)$. The probability that the first n down-steps were all generated by accepting the first proposed down-steps is $(m/(m+\epsilon))^n$. On this event, all the down-steps were generated using rates given by (4.24), while we know that the up-steps are created using rates given by (4.23). Therefore, we get

$$\mathbb{P}(T^\epsilon > n) \geq \left(\frac{m}{m+\epsilon} \right)^n \mathbb{P}(T^0 > n) \geq C \left(\frac{m}{m+\epsilon} \right)^n \frac{1}{n^{1/2}}$$

Therefore, from (4.22) we get for all $n \in \mathbb{N}$,

$$\mathbb{E}_{(x,+1)}\left[b^{\tau_x}\right] \geq \left(\frac{M+\epsilon}{M+\epsilon-\log b}\right)^n \mathbb{P}\left(T^\epsilon > n\right) \geq C\left(\frac{M+\epsilon}{M+\epsilon-\log b}\right)^n \left(\frac{m}{m+\epsilon}\right)^{n+1} \frac{1}{n^{1/2}} \geq$$
$$\geq C\left(\frac{m}{m+\epsilon}\frac{M+\epsilon}{M+\epsilon-\log b}\right)^n \frac{1}{n^{1/2}}.$$

If we set $\epsilon < 1/2\left(\sqrt{(M-\log b)^2+4m\log b}-(M-\log b)\right)$, we have $\frac{m}{m+\epsilon}\frac{M+\epsilon}{M+\epsilon-\log b} > 1$ and if we let $n \to \infty$ we get that the RHS diverges to $+\infty$ which proves the contradiction. □

Remark 4.3.18. *As will be seen in Section 4.4, one can relax the assumption that* $\gamma(x) \geq m > 0$ *to the slightly more general assumption that* $\lim_{x\to+\infty} \frac{U'(x)}{\gamma(x)} = 0$.

4.3.4 Non-Geometric Ergodicity in One Dimension

So far, in Sections 4.3.1 and 4.3.3, we managed to prove that under various assumptions on the refresh rate, for a one-dimensional Zig-Zag process targeting a distribution with $\lim_{x\to+\infty} U'(x) = 0$, no point $(x,+1)$ has the SGR property. In this section we will combine the results to prove that the one dimensional Zig-Zag process is not geometrically ergodic, therefore proving Theorem 4.3.3 in dimension one.

Proof of Theorem 4.3.3 in Dimension One. We will prove that for any $\delta > 0$ the δ-skeleton is not Geometrically ergodic which also implies that the continuous time process is not Geometrically ergodic. From Theorem 2.1.5 if a δ-skeleton was geometrically ergodic, then there would exist some petite set C and $b > 1$ with $\sup_{(z,\theta)\in C} \mathbb{E}[b^{\sigma_C}] < \infty$ where $\sigma_C = \inf\{n \geq 1 : Z_{n\delta} \in C\}$. We distinguish between the following two cases for C.

First, consider the case where there exists a point $(x_0,+1) \in C$ and let $y_1 = \sup\{y \in \mathbb{R} : (y,+1) \in C\}$, $y_2 = \sup\{y \in \mathbb{R} : (y,-1) \in C\}$. Note that due to Lemma 4.3.4, $y_1, y_2 \in \mathbb{R}$ and assume that we start the process from $(y_1,+1)$. Assume first that $y_1 \geq y_2$. Then, in order for the process to hit C, it first has to hit $(y_1,-1)$ so from strong Markov property, $\mathbb{E}_{(y_1,+1)}[b^{\sigma_{y_1}}] < \infty$ where $\sigma_{y_1} = \inf\{n \geq 1 : X_{n\delta} \leq y_1\}$. On the other hand, if $y_1 < y_2$, due to boundedness of U' and γ on $[y_1, y_2]$, there is a positive probability for the process not to switch until it reaches $(y_2,+1)$. In this event the process will only hit C after it hits $(y_2,+1)$ and then $(y_2,-1)$ therefore, using the strong Markov Property $\mathbb{E}_{(y_2,+1)}[b^{\sigma_{y_2}}] < \infty$.

Now, consider the case where $C = B \times \{-1\}$ for some bounded $B \subset \mathbb{R}$. Write $z = \inf B$ and $y = \sup B$ and start the process from $(z, -1)$. Since the process is non-evanescent (see for example [BRZ19]) the process will switch to $+1$ at some point $z_0 < z$ and, due to boundedness of the rates on $[z_0, y]$, there is a positive probability that it will not switch to -1 until it reaches $(y, +1)$. Throughout this path the process will not have hit C and it will only do that after it hits $(y, -1)$. Using again the Strong Markov Property, we get that, $\mathbb{E}_{(y,+1)}[b^{\sigma_y}] < \infty$.

In all cases, assuming that a δ-skeleton is Geometrically ergodic implies that there exists a $b > 1$ and $y \in \mathbb{R}$, such that $\mathbb{E}_{(y,+1)}[b^{\sigma_y}] < \infty$. Now, observe that if $\tau_y = \inf\{t \geq \delta : X_t \leq y\}$ then $\tau_y \leq \delta\sigma_y$ which means that

$$\mathbb{E}_{(y,+1)}\left[\left(b^{1/\delta}\right)^{\tau_y}\right] \leq \mathbb{E}_{(y,+1)}\left[\left(b^{1/\delta}\right)^{\delta\sigma_y}\right] \leq \mathbb{E}_{(y,+1)}\left[b^{\sigma_y}\right] < \infty$$

and this means that $(y, +1)$ has the SGR property, which contradicts Propositions 4.3.6 and 4.3.17. □

4.3.5 Non-Geometric Ergodicity in Higher Dimensions

Proof of Theorem 4.3.3. Let us assume without loss of generality that for the coordinate i appearing in Assumptions 4.3.1, $i = 1$. As in the previous subsection, we will prove that every skeleton of the process is non geometrically ergodic. Assume otherwise and suppose that for some δ, the δ-skeleton is geometrically ergodic. We get from Theorem 2.1.5 that there would exist some small and therefore bounded set C and $b > 1$ with $\sup_{(z,\theta)\in C} \mathbb{E}_{z,\theta}[b^{\sigma_C}] < \infty$. The idea is to project the process in the first coordinate and observe that if we start the process from a point $(x, \theta) = ((x_1, ..., x_d); (\theta_1, ..., \theta_d))$ with $\theta_1 = +1$ and $x_1 > 0$ very large, then in order for the process to return to the small set C, it first needs to hit a point whose first coordinate is less than x_1. This way, we can study the time it takes the first coordinate process to return to the starting point, for which we can use the one-dimensional techniques we have already established.

Let $\epsilon < \log(b^{1/\delta})$. Let $\left(X_t^{(1)}\right)_{t\geq 0}$ be the first coordinate of the space component of the Zig-Zag process. For $x = (x_1, ..., x_d)$, let $\sigma_{x_1} = \inf\{n \geq 1 : X_{n\delta}^{(1)} \leq x_1\}$. Then, considering all possibilities for the structure of C, as in the proof of Theorem 4.3.3 in the previous section, we can assume that for any $x_1 > 0$ large enough, there exists a point $(x, \theta) = (x_1, ..., x_d, \theta_1, ..., \theta_d)$ with $\theta_1 = +1$ such that $|\partial_1 U(z)| < \epsilon$ for all z with first coordinate $z_1 \geq x_1$ and such that

$$\mathbb{E}_{(x,\theta)}[b^{\sigma_{x_1}}] < \infty.$$

Consider the first coordinate process $Z^{(1)}(t) = \left(X_t^{(1)}, \Theta_t^{(1)}\right)$, and let $\tau_{x_1} = \inf\{t \geq \delta : X^{(1)}(t) \leq x_1\}$ be the first time that this process hits $(-\infty, x_1]$. Then, arguing as in the proof of Theorem 4.3.3 in dimension one, we have $\tau_{x_1} \leq \delta\sigma_{x_1}$, meaning that

$$\mathbb{E}_{(x,\theta)}[\left(b^{1/\delta}\right)^{\tau_{x_1}}] < \infty.$$

Let as consider the process $\{Z^{(1)}(t)\}_{t\geq 0}$, starting from $(x_1, +1)$. Since for any $y = (y_1, ..., y_d)$ with $y_1 \geq x_1$ we have $\gamma_1(y) = \gamma_1(y_1)$ and $\partial_1 U(y) \leq \epsilon$, the rate of switching of this process is $\lambda_1(y_1, \eta_1) = [\eta_1 \partial_1 U(y)]^+ + \gamma_1(y_1)$. Therefore, the hazard rate of the up-steps is bounded above by

$$\tilde{\lambda}(y_1, +1) = \epsilon + \gamma_1(y_1) \tag{4.27}$$

and the hazard rate of the down-steps is bounded below by

$$\tilde{\lambda}(y_1, -1) = \gamma_1(y_1), \tag{4.28}$$

Therefore, using the proof of Lemma 4.3.12 we see that τ_{x_1} stochastically dominates the stopping time $\tilde{\tau_{x_1}}$, which is defined to be the first return time to x_1 for the one-dimensional Zig-Zag process, starting from $(x_1, +1)$ and having hazard rate given by (4.27) and (4.28). Therefore,

$$\mathbb{E}_{(x_1,+1)}\left[\left(b^{1/\delta}\right)^{\tilde{\tau}_{x_1}}\right] < \infty.$$

We know, however, from the Proof of Proposition 4.3.17, that when $\epsilon < \log b^{1/\delta}$, the $b^{1/\delta}$-moment generating function of $\tilde{\tau}_{x_1}$ for the one dimensional Zig-Zag process with rates given by (4.27) and (4.28) will not be finite. We therefore reach a contradiction and this concludes the proof of Theorem 4.3.3. □

4.4 Random Velocity Refreshment

In the literature on Bouncy Particle Sampler it is often the case that the process is formulated so that when a refreshing event occurs, a new velocity is picked according to a Gaussian distribution. In this section we will consider a variant of the one-dimensional Zig-Zag process in which, if the process stops due to a refresh event, instead of just switching the sign of the velocity, it picks a new velocity, of sign opposite to the current one, from a probability distribution, with possibly unbounded support. We will study the question of geometric ergodicity of this process in heavy

tails, which is a bit more complicated than the question for the original Zig-Zag. The process no longer moves with bounded speed and one cannot guarantee that it will not have visited some area by a given time. This means that one can no longer use an argument similar to the one in Section 4.1.

We begin by formally defining the process and prove that the measure of interest is invariant. Then we use the techniques developed in the previous section to prove that this process is also non-geometrically ergodic.

The state space is $E = \mathbb{R} \times \mathbb{R}$ where the first component is space and the second is the velocity. Given that the process $(Z_t)_{t\geq 0} = (X_t, \Theta_t)_{t\geq 0}$ is at $(y, \theta) \in E$, it moves with constant velocity θ and when the Poisson process with rate $m(t) = \lambda(y + t\theta, \theta)$ generates a point, it changes the velocity θ. Here

$$\lambda(x, \theta) = [\theta U'(x)]^+ + \gamma(x) \tag{4.29}$$

for some non-negative locally integrable function γ and some $U \in C^1$. The mechanism of switching the velocity is the following. If the process stops at point (x, θ), the new velocity θ' is set to be $-\theta$, just as in the normal Zig-Zag case, with probability $\frac{[\theta U'(x)]^+}{\lambda(x,\theta)}$. Otherwise, with probability $\frac{\gamma(x)}{\lambda(x,\theta)}$, the sign of the velocity switches, but in absolute value the new velocity is picked according to a probability measure q on $[0, +\infty)$. Here we assume that $q(\{0\}) = 0$ and we will be writing Q to denote the tail distribution of q, i.e. for all $a > 0$ $Q(a) = q([a, +\infty))$. Using Poisson thinning we observe that the first possibility of velocity change occurs at first arrival time of a Poisson process with hazard rate $[\theta U'(x)]^+$, while the second possibility occurs at first arrival time of a Poisson process with hazard rate $\gamma(x)$. Since the hazard rate of the second possibility does not depend on θ, we will call the events generated from γ refresh events, as in the original Zig-Zag. We will call the process **Random Velocity Refreshment Zig-Zag (RVRZZ)**. The process in algorithmic terms is described as follows.

Algorithm 4.4.1.

1. *Set* $t_{curr} = 0$.
2. *Start from point* $(x, \theta) \in E = \mathbb{R} \times \mathbb{R}$.
3. *Set* $(X_{t_{curr}}, \Theta_{t_{curr}}) = (x, \theta)$
4. *The* x*-component moves along the line* $\{x + t\theta, t \geq 0\}$.
5. *We define the function* $\lambda(y, \eta) = [\eta U'(y)]^+ + \gamma(y)$.

6. *We construct a Poisson Process with intensity* $\{m(t) = \lambda(x + t\theta, \theta), t \geq 0\}$.

7. *Suppose that the first arrival time of the process is* T.

8. *For all* $t \in (t_{curr}, t_{curr} + T)$ *set* $X_t = x + t\theta$, $\Theta_t = \theta$.

9. *Set* $x = x + T\theta$, $t_{curr} = t_{curr} + T$ *and*

   (a) *With probability* $\frac{[\theta U'(x)]^+}{\lambda(x,\theta)}$, *set* $\theta = -\theta$.

   (b) *Else, pick a random* $\theta' \sim q$ *and set* $\theta = -sign(\theta)\,\theta'$.

10. *Repeat from Step 2.*

We have the following.

Proposition 4.4.2 (Invariant Measure of RVRZZ)**.** *The RVRZZ has generator given by*

$$\begin{aligned}\mathcal{L}f(x,\theta) = \theta f'(x,\theta) + [\theta U'(x)]^+ \left(f(x,-\theta) - f(x,\theta)\right) + \\ + \gamma(x)\left(\int_0^{+\infty} f(x, -sign(\theta)\theta')q(d\theta') - f(x,\theta)\right)\end{aligned}$$

for all $f \in C_c^1(E)$. *Furthermore, if* p *is the probability measure that extends* q *symmetrically to the whole real line, i.e.* $p(A) = \frac{1}{2}q(A\cap(0,+\infty)) + \frac{1}{2}q(-A\cap(0,+\infty))$, *then the process has the probability measure* $\mu(dx, d\theta) = \frac{1}{Z}\exp\{-U(x)\}dx\; p(d\theta)$ *as invariant.*

Proof of Proposition 4.4.2. The equation for the generator comes from Theorem 2.3.3. Using similar arguments as in the proof of Proposition 3.1.5, involving integration by parts and rearranging of sums, we write for any $f \in C_c^1(E)$

$$\begin{aligned}
&2Z\int_E \mathcal{L}f(x,\theta)\mu(dx,d\theta) = \\
&= \int_{\theta\in\mathbb{R}^+}\int_{\mathbb{R}} \exp\{-U(x)\}\left(\theta f'(x,\theta) + [\theta U'(x)]^+\left(f(x,-\theta) - f(x,\theta)\right)\right)dx\; q(d\theta) + \\
&+ \int_{\theta\in\mathbb{R}^+}\int_{\mathbb{R}} \exp\{-U(x)\}\gamma(x)\int_{\theta'\in\mathbb{R}^+}\left(f(x,-\theta') - f(x,\theta)\right)q(d\theta')dx\; q(d\theta) + \\
&+ \int_{\theta\in\mathbb{R}^+}\int_{\mathbb{R}} \exp\{-U(x)\}\left(-\theta f'(x,-\theta) + [-\theta U'(x)]^+\left(f(x,\theta) - f(x,-\theta)\right)\right)dx\; q(d\theta) + \\
&+ \int_{\theta\in\mathbb{R}^+}\int_{\mathbb{R}} \exp\{-U(x)\}\gamma(x)\int_{\theta'\in\mathbb{R}^+}\left(f(x,\theta') - f(x,-\theta)\right)q(d\theta')dx\; q(d\theta) = \\
&= \int_{\theta\in\mathbb{R}^+}\int_{\mathbb{R}} f(x,\theta)\exp\{-U(x)\}\left(\theta U'(x) - [\theta U'(x)]^+ + [-\theta U'(x)]^+\right)dx\; q(d\theta) +
\end{aligned}$$

$$+\int_{\mathbb{R}}\exp\{-U(x)\}\gamma(x)\left(\int_{\theta'\in\mathbb{R}^+}f(x,-\theta')q(d\theta')-\int_{\theta\in\mathbb{R}^+}f(x,\theta)q(d\theta)\right)dx+$$
$$+\int_{\theta\in\mathbb{R}^+}\int_{\mathbb{R}^d}f(x,-\theta)\exp\{-U(x)\}\left(-\theta U'(x)-[-\theta U'(x)]^+ +[\theta U'(x)]^+\right)dx\,q(d\theta)$$
$$+\int_{\mathbb{R}}\exp\{-U(x)\}\gamma(x)\left(\int_{\theta'\in\mathbb{R}^+}f(x,\theta')q(d\theta')-\int_{\theta\in\mathbb{R}^+}f(x,-\theta)q(d\theta)\right)dx=0.$$

The result follows from Proposition 2.1.9, since $C_c^1(E)$ is core for the generator (see for example Corollary 19 of [DGM18]). □

The main result of the section is the lack of geometric ergodicity of the process on heavy tailed targets and is shown in the following theorem.

Theorem 4.4.3 (Non-Geometric Ergodicity of RVRZZ)**.** *Assume that $U\in C^1$ and $\frac{1}{Z}\exp\{-U(x)\}$ is the density of a probability measure on $\mathbb{R}$ for some $Z>0$. Let q be a probability measure on $[0,+\infty)$ and $q(\{0\})=0$. Also, assume that the potential U satisfies*

$$U'(x)\xrightarrow{x\to+\infty}0. \tag{4.30}$$

Assume further that refresh rate γ is a non-negative, bounded and locally integrable function of $\mathbb{R}$. Consider a RVRZZ process with rates given by (4.29). Assume that one of the following two propositions holds for the refresh rate γ:

1. $\lim_{x\to+\infty}\gamma(x)=0$.
2. $\lim_{x\to+\infty}\frac{U'(x)}{\gamma(x)}=0$.

Then, the RVRZZ process is not Geometrically Ergodic.

Remark 4.4.4. *As in Section 4.3, due to the symmetry of the arguments, the theorem also holds if the assumptions on γ hold for $x\in(-\infty,0)$.*

The idea of the proof, as in the previous section, will be to prove that the petite sets do not possess the SGR property. Since the process can refresh the velocity to an arbitrarily large velocity, we cannot guarantee that the process starting from a point $x>0$ very large, will not have explored the space fast enough. This means that we can no longer rule out the possibility that some small set is unbounded, as we did in Lemma 4.3.4, in Section 4.3. However, we can still pose some constraints in the structure of petite sets, as will be seen in the following Lemma.

Lemma 4.4.5. *Assume that $C\subset\mathbb{R}\times\mathbb{R}$ is a petite set for the RVRZZ where the refresh rate γ is bounded above. If the sequence $(x_n,\theta_n)\in C$ satisfies $x_n\xrightarrow{n\to+\infty}+\infty$ then there exists an $M>0$ such that $x_n/|\theta_n|\leq M$ for all $n\in\mathbb{N}$.*

Remark 4.4.6. *It will become evident from the proof that the same result holds when* $x_n \xrightarrow{n\to\infty} -\infty$.

Proof of Lemma 4.4.5. Assume that there exists a petite set C with $(x_n, \theta_n) \in C$, $x_n \to +\infty$ and $\frac{x_n}{|\theta_n|}$ unbounded. Then there exists $c > 0$, a probability measure ν on $\mathbb{R} \times \mathbb{R}$ and a probability measure a on $[0, +\infty)$ such that for all $(x, \theta) \in C$ and all $A \in \mathcal{B}(\mathbb{R} \times \mathbb{R})$,

$$\int_0^{+\infty} \mathbb{P}_{x,\theta} \left(Z_t \in A \right) a(dt) \geq c\, \nu(A),$$

where $Z_t = (X_t, \Theta_t)$ is the RVRZZ process. Let $N > 0$ such that for all $x \in \mathbb{R}$, $\gamma(x) \leq N$. Fix $m \in \mathbb{N}$ and set $G_m = [-m, m] \times \mathbb{R}$. Fix $T > 0$ and since $x_n/|\theta_n|$ is unbounded, pick n large enough such that $x_n - m > a = a(T) > T|\theta_n|$ for some value of a that will be specified later. This means than if the process starts from (x_n, θ_n) it cannot reach G_m before time T by retaining the same speed θ_n. The only way to reach G_m in time less than T is for a refresh event to occur and for a new velocity to be picked with absolute value at least a/T. Write $Q(a/T) = q(a/T, +\infty)$. Since $\gamma(x) \leq N$ and since more refresh events make it more likely to refresh to a large velocity, we may assume without loss of generality that $\gamma(x) = N$. This further means that the refresh events are first arrival times of a PP with rate N and each of these events refreshes to a large velocity with probability $Q(a/T)$. Therefore, from Poisson thinning, the refreshments to large velocities are generated according to a Poisson process with rate $NQ(a/T)$. We then bound, for all $t \leq T$,

$$\mathbb{P}_{x_n,\theta_n} \left(Z_t \in G_m \right) \leq 1 - \exp\left\{ -tNQ(a/T) \right\} \leq 1 - \exp\left\{ -TNQ(a/T) \right\}.$$

Then, we can further bound

$$\begin{aligned} c\, \nu(G_m) &\leq \int_0^{+\infty} \mathbb{P}_{x_n,\theta_n} \left(Z_t \in G_m \right) a(dt) \leq \\ &\leq \int_0^{T} \mathbb{P}_{x_n,\theta_n} \left(Z_t \in G_m \right) a(dt) + \int_T^{+\infty} \mathbb{P}_{x_n,\theta_n} \left(Z_t \in G_m \right) a(dt) \leq \\ &\leq \left(1 - \exp\left\{ -TNQ\left(\frac{a}{T} \right) \right\} \right) + a((T, +\infty)). \end{aligned} \tag{4.31}$$

If, given T, we pick $a = a(T)$ such that

$$Q\left(\frac{a}{T} \right) \leq -\frac{1}{NT} \log\left(1 - \frac{1}{T} \right),$$

then we get

$$1 - \exp\left\{ -TNQ\left(\frac{a}{T} \right) \right\} \leq \frac{1}{T}.$$

On letting $T \to \infty$ the RHS of (4.31) converges to 0 and therefore $\nu(G_m) = 0$. Since this is true for all $m \in \mathbb{N}$, by covering $\mathbb{R} \times \mathbb{R}$ with countably many G_m's, we get that ν is the trivial measure, which is a contradiction. □

To sum up, we want to show that the return time to any petite set has no finite moment generating function. We just proved that, even though a petite set C may be unbounded, if $(x, \theta) \in C$ for some $x > 0$ large enough, the velocity θ must also be very large in absolute value. Therefore, if we restrict ourselves to the event that the process will never refresh the speed to a very large value, as long as the process stays in some area where the space component x is very large, we can guarantee that the process will not be hitting the petite set. We will use this later to bound the return time to C from below, by the first time the process returns to the starting point x, which was the approach we also adopted in Subsection 4.3.

We can, then, focus our attention on the behaviour of the space coordinate x of the RVRZZ process. This will move in a similar fashion to the original Zig-Zag, having up-steps and down-steps. If we mark the position at the end of each down-step, this will induce a random, albeit not necessarily symmetric, walk. In order to control the first time the space coordinate returns to not large values, we could try to control the first time this random walk returns to not large values. If the walk was symmetric we could just use the universality Random Walk Theorem 4.3.9. Recall that in the previous section, this theorem was used to study the law of the first hitting time of the origin for the Random Walk in the case where up-steps had hazard rate $\gamma(x) + \epsilon$ for some $\epsilon > 0$ small and down-steps had hazard rate $\gamma(x)$. The next Lemma we will present here, will allow us to use Theorem 4.3.9 in the new setting for a process similar to RVRZZ that tends to return to low values of x faster than the RVRZZ. We begin by defining this process.

Let q^* be a probability measure on $(0, \theta^*]$ for some $\theta^* > 0$ and let $\epsilon > 0$. Let $U \in C^1$ and γ a locally integrable function such that there exist $M > 0$ such that $\gamma(x) \leq M$ for all $x \in \mathbb{R}$. Assume that

$$U'(x) \leq \epsilon \gamma(x) \tag{4.32}$$

and consider the following process. The process starts from $S'_0 = x^*$. We draw a speed $\theta_1 \sim q^*$ and the first up-step time T_1 as the first arrival time of a Poisson process with intensity $m(t) = \theta_1 |U'(t\theta_1)| + \gamma(t\theta_1)$. The process then moves to point $x^* + \theta_1 T_1$, which is the end of the up-step and a new velocity θ'_1 is picked, having the opposite sign of θ_1 and being independent of θ_1, such that $-\theta'_1 \sim q^*$. After the new velocity θ'_1 is chosen, we simulate the first down-step time D_1 as the first arrival time

of a Poisson process with intensity $m(t) = \gamma(\theta_1 T_1 + \theta_1' t)$. The process then moves to the position after the down-step $x^* + T_1\theta_1 + D_1\theta_1'$ and we set $S_1' = x^* + T_1\theta_1 + D_1\theta_1'$ (see Figure 4.8).

Assume that the process is at S_{n-1}' after $n-1$ steps. We draw a new velocity $\theta_n \sim q^*$ independently of the previous velocities and we simulate the n'th up-step time T_n as the first arrival time of a Poisson process with intensity $m(t) = \theta_n |U'(S_{n-1} + t\theta_n)| + \gamma(S_{n-1}' + t\theta_n)$. The process then moves up to point after the up-step, $S_{n-1}' + \theta_n T_n$ and a new velocity θ_n' is picked, having the opposite sign and being independent of θ_n, such that $-\theta_n' \sim q^*$. Then, we simulate the n'th down-step time D_n as the first arrival time of a Poisson process with intensity $m(t) = \gamma(S_{n-1}' + \theta_n T_n + \theta_n' t)$. The process then moves to the position after the down-step and we set $S_n' = S_{n-1}' + T_n\theta_n + D_n\theta_n'$.

We shall be referring to this process $(S_n')_{n\geq 0}$ as the **Random Velocity Bounding Model**. The name comes from the fact that the up-steps of the RVRZZ (having hazard rate $[\theta U'(x)]^+ + \gamma(x)$ while moving with velocity $\theta > 0$) stochastically dominate the up-steps of this Bounding model (having hazard rate $|\theta U'(x)| + \gamma(x)$) and the down-steps of the RVRZZ (having hazard rate $[\theta U'(x)]^+ + \gamma(x)$ while moving with velocity $\theta < 0$) are stochastically dominated by the ones of this bounding model (having hazard rate $\gamma(x)$). This means that in order to bound from below the probability that the Random Walk induced by the up-steps and down-steps of the RVRZZ will not have returned to the origin during the first n steps, we only need to do this for the respective probability of the Random Velocity Bounding Model. This will be done using Theorem 4.3.9 from paragraph 4.3.2.

Lemma 4.4.7. *Consider a Random Velocity Bounding Model with parameter ϵ, as described above in equation (4.32) and let $T^\epsilon = \inf\{n \geq 1 : S_n' \leq x^*\}$. Then there exists a constant $C > 0$ such that for all $n \in \mathbb{N}$,*

$$\mathbb{P}\left(T^\epsilon \geq n+1\right) \geq C \left(\frac{1}{1+\epsilon}\right)^{2n} \frac{1}{n^{1/2}}. \tag{4.33}$$

Proof of Lemma 4.4.7. Recall that we only draw velocities from measure q^* with support bounded above by θ^*. Let us consider an up-step of the Random Velocity Bounding Model which, while having velocity $\theta > 0$, has hazard rate $\gamma(x) + |\theta||U'(x)| \leq \gamma(x) + \theta^*|U'(x)|$. This means that in order to simulate any up-step time of the Bounding Model, we can use Poisson thinning with $\gamma(x) + \theta^*|U'(x)|$ being the bounding hazard rate, similarly to the Algorithm presented in the proof of Proposition 4.3.17 in Section 4.3.3. More precisely, assume that the process is at state S_{n-1}' and picks a new speed of $\theta_n \sim q^*$. We can simulate a proposed up-step as the first ar-

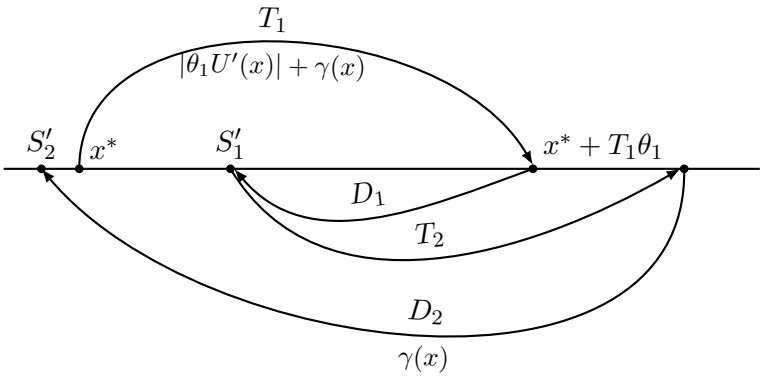


Figure 4.8: *A representation of the movement of the Random Velocity Bounding Model. T_1 is the time it takes the first up-step to be completed, D_1 the corresponding time of the first down-step , T_2 the respected time of the second up-step etc. While moving with velocity θ, the up-steps have hazard rate $|\theta U'(x)|+\gamma(x)$, while the down-steps have $\gamma(x)$. S_n' is the position of the process after the end of the n'th down-step. T^ϵ is the first time the Walk S_n' takes a value less than x^*. In this configuration $T^\epsilon = 2$.*

rival time T_n of a Poisson process with rate $m(t) = \theta^*|U'((S_{n-1}'+t\theta_n)|+\gamma(S_{n-1}'+t\theta_n)$ and then accept this time on the event that

$$u_n \leq \frac{\theta_n|U'((S_{n-1}' + T_n\theta_n)| + \gamma(S_{n-1}' + T_n\theta_n t)}{\theta^*|U'((S_{n-1}' + T_n\theta_n)| + \gamma(S_{n-1}' + T_n\theta_n)} \tag{4.34}$$

where $u_n \sim unif(0,1)$ is independent of the rest of the process' randomness. If the process accepts the proposed up-step, then T_n is the time of the n'th up-step. If it rejects, then it starts the procedure again from $S_{n-1}' + T_n\theta_n$ and simulates a new proposed up-step time according to the bounding rate function
$m(t) = \theta^*|U'(S_{n-1}'+T_n\theta_n+t\theta_n)|+\gamma(S_{n-1}'+T_n\theta_n+t\theta_n)$. The new proposed up-step time is accepted or rejected according to an event of the form (4.34), formulated by a new independent uniform distribution, where U' and γ on the RHS of (4.34) are evaluated at the ending point of the new proposed up-step. This procedure continues until there is an acceptance. The same procedure can be carried out to simulate the down-steps. This is because the hazard rate for the down-steps is always $\gamma(x) \leq \gamma(x)+\theta^*|U'(x)|$, therefore we can use the same bounding process to simulate the down-steps. More precisely, if the end of the n'th up-step is at $S_{n-1}' + T_n\theta_n$ and a velocity $\theta_n' \sim q^*$ is chosen, one can simulate a proposed down-step time as the first arrival time of the PP with rate $m(t) = \theta^*|U'(S_{n-1}'+T_n\theta_n+t\theta_n')|+\gamma(S_{n-1}'+T_n\theta_n+t\theta_n')$. If after the proposed down-step the process is at x, one accepts or reject that

down-step on the event that

$$u_n' \leq \frac{\gamma(x)}{\theta^*|U'(x)| + \gamma(x)}. \tag{4.35}$$

where $u_n' \sim unif(0,1)$ independent of the rest of the process' randomness.

We emphasise that all the uniform distributions used to accept of reject the proposed up-steps or down-step times are all independent between themselves and the rest of the process' randomness. Note as well, that in equation (4.32), the quantities appearing on the RHS of the equations (4.34), defining the events in which the proposed up-steps are accepted, are of the form

$$\frac{\gamma(x) + \theta_1|U'(x)|}{\gamma(x) + \theta^*|U'(x)|} \geq \frac{1}{1 + \theta^*\epsilon}. \tag{4.36}$$

The same lower bound holds for the RHS of equations (4.35), defining the events in which the proposed down-steps are accepted and we have

$$\frac{\gamma(x)}{\gamma(x) + \theta^*|U'(x)|} \geq \frac{1}{1 + \theta^*\epsilon}. \tag{4.37}$$

Let us consider the following event, which we call event B: *for all the first n up-steps and all the first n down-steps, the $2n$ i.i.d. uniform distributions that decide whether to accept or reject the first proposed up-step/down-step time, all took values less than* $1/(1 + \theta^*\epsilon)$. This event has probability

$$\left(\frac{1}{1 + \theta^*\epsilon}\right)^{2n}.$$

Note that when event B occurs, due to equations (4.36) and (4.37), throughout the first n steps, the process is forced to accept the first proposed up-steps or down-steps, simulated with hazard rate $|\theta^*U'(x)| + \gamma(x)$. This hazard rate does not depend on the current velocity (and more specifically, does not depend on the sign of the current velocity), therefore the Random Walk induced by these hazard rates is a symmetric one. Note as well that these uniform distributions are all independent of the rest of the process' randomness and event B involves these uniform distributions taking values less than a specified constant. This means that conditioning event B does not affect the law of the process, except for forcing the first proposed up-steps or down-steps to be accepted. From now on, we will condition on event B occurring. The rest of the process will move in the following manner.

Write $b(x) = \gamma(x) + \theta^*|U'(x)|$. If the Random Velocity Bounding Model is

at S'_{n-1} after $n-1$ steps, a new speed $\theta_n \sim q^*$ is picked and the next up-step time T_n is the first arrival time of a Poisson process with intensity $m(t) = b(S'_{n-1} + t\theta_n)$. The up-step stops at point $S'_{n-1} + T_n\theta_n$. From exponential representation of Poisson Process, this means that for some $E_n \sim \exp(1)$ independent of all the other process randomness,

$$E_n = \int_0^{T_n} b(S'_{n-1}+\theta_n t)dt = \frac{1}{\theta_n}\int_{S'_{n-1}}^{S'_{n-1}+T_n\theta_n} b(u)du \iff \theta_n E_n = \int_{S'_{n-1}}^{S'_{n-1}+T_n\theta_n} b(u)du.$$

Then the process picks a new velocity θ'_n, independently of the other velocities, such that $-\theta'_n \sim q^*$. The next down-step time D_n is picked according to the first arrival time of a Poisson process with intensity $m(t) = b(S'_{n-1} + \theta_n T_n + \theta'_n t)$. The next state S'_n is defined to be the point where the down-step stops, so $S'_n = S'_{n-1} + \theta_n T_n + \theta'_n D_n$. This means that for some $F_n \sim \exp(1)$ independent of all the other process randomness,

$$F_n = \int_0^{D_n} b(S'_{n-1} + \theta_n T_n + \theta'_n)dt = \frac{1}{\theta'_n}\int_{S'_{n-1}+\theta_n T_n}^{S'_{n-1}+T_n\theta_n+D_n\theta'_n} b(u)du \iff$$

$$\iff \theta'_n F_n = \int_{S'_{n-1}+\theta_n T_n}^{S'_n} b(u)du.$$

Overall we get

$$\int_{S'_{n-1}}^{S'_n} b(u)du = \theta_n E_n + \theta'_n F_n$$

and this further means that

$$\int_{S'_0}^{S'_n} b(u)du = \sum_{k=1}^{n} \left(\theta_k E_k + \theta'_k F_k\right) \tag{4.38}$$

where $E_1, ..., E_n, F_1, ..., F_n$ are i.i.d. $\exp(1)$ and $\theta_1, ..., \theta_n, -\theta'_1, ..., -\theta'_n$ i.i.d samples from q^*. Therefore, conditioning on event B we get

$$T^\epsilon = \inf\{n \geq 1 : S'_n \leq S'_0\} = \inf\{n \geq 1 : \sum_{i=1}^{n} \left(\theta_i E_i + \theta'_i F_i\right) \leq 0\} \quad a.s. \tag{4.39}$$

Now, for the symmetric random walk with increments $\theta_i E_i + \theta'_i F_i$ where $E_i, F_i \sim \exp(1)$ and $\theta_i, -\theta'_i \sim q^*$ are all i.i.d., Theorem 4.3.9 applies and therefore, there exists a constant $C > 0$ such that for all $n \in \mathbb{N}$,

$$\mathbb{P}\left(T^\epsilon \geq n+1 \middle| B\right) = \frac{1}{4^n}\binom{2n}{n} \geq C\frac{1}{n^{1/2}},$$

Overall this gives

$$\mathbb{P}\left(T^{\epsilon} \geq n+1\right) \geq \mathbb{P}\left(\{T^{\epsilon} \geq n+1\} \cap B\right) \geq C\left(\frac{1}{1+\epsilon}\right)^{2n} \frac{1}{n^{1/2}}.$$

which is the statement of the Lemma. □

To sum up, in Lemma 4.4.5 we provide a restriction the structure of the petite sets of the RVRZZ. This will allow us to bound from below the first return time to the petite set by the first time the x-coordinate returns to the starting point. Focusing on the behaviour of the x-coordinate, which will move with up-steps and down-steps, the next step was to control the behaviour the x-coordinate has at the end of each down-step. In Lemma 4.4.7 we provide such a control in a specific case of the RVRZZ process, which we call Random Velocity Bounding Model. We will see later that this process tends to return to the starting value of x faster than the RVRZZ. This will provide us control on the behaviour of RVRZZ. We can now prove the main theorem of the Section.

Proof of Theorem 4.4.3. Let Z_t be an RVRZZ process satisfying the conditions of Theorem 4.4.3. We will prove that for any petite set C and any $b > 1$ and any $\delta_1 > 0$, there exists some point $(x, \theta) \in \bar{C}$ such that if $\tau_C = \inf\{t \geq \delta_1 : Z_t \in C\}$ is the first return time to C after a small time δ_1, then $\mathbb{E}_{x,\theta}[b^{\tau_C}] = +\infty$. The result will follow using similar arguments as in the case of the original Zig-Zag in the previous section, which proves that no skeleton chain has the SGR property.

From now on, we will assume for contradiction that there exists a petite set C, some $b > 1$ and some $\delta_1 > 0$, such that if $\tau_C = \inf\{t \geq \delta_1 : Z_t \in C\}$,

$$sup_{(x,\theta)\in C}\mathbb{E}_{x,\theta}[b^{\tau_C}] < +\infty. \tag{4.40}$$

We consider the following possibilities for C.

The first possibility is that for every $\theta > 0$ and every $x \in \mathbb{R}$, $(x, \theta) \notin C$. Then $(y_1, -\theta) \in C$ for some $y_1 \in \mathbb{R}$, $\theta > 0$. Let $y_0 = \inf\{x \in \mathbb{R} : (x, -\theta) \in C\}$ which is in $\mathbb{R}$ due to Lemma 4.31. Then we know from (4.40) that $\mathbb{E}_{y_0,-\theta}[b^{\tau_C}] < +\infty$. Let us start the process from $(y_0, -\theta)$. Since $U \in C^1$ and $1/Z \exp\{-U(x)\}$ is the density of a probability measure, we get that for some $y_3 < y_2 < y_0$, U' is negative, bounded away from zero on (y_3, y_2) and therefore the rate $[-\theta U']^+$ is strictly positive, bounded away from zero on (y_3, y_2). This means that there is a positive probability that the process will switch from $-\theta$ to θ while on $(y_3, y_2) \subset (-\infty, y_0)$ and from the definition of y_0 the process will not have hit C until that time. After this switch,

since the rates are locally bounded, for any $x^* > 0$ large enough there exists a positive probability that the process will reach (x^*, θ) without switching velocity. Throughout this procedure, the process will not have hit C (since for any $x \in \mathbb{R}$, $(x, \theta) \notin C$) and therefore, the time the process hits (x^*, θ) is less than τ_C. From strong Markov property, $\mathbb{E}_{x^*,\theta}[b^{\tau_C}] < \infty$.

The second possibility for C is that there exists $(y_0, \theta) \in C$ for some $\theta > 0$. In that case, from Lemma 4.31, $a = \sup\{x : (x, \theta) \in C\} < \infty$, so we can start the process from (a, θ) and, due to (4.40), we have that $\mathbb{E}_{a,\theta}[b^{\tau_C}] < +\infty$. Since the rates are locally bounded, for any $x^* > a$, there is a positive probability that the process will not have switched velocity until it reaches (x^*, θ) and until then it will not have hit C, by definition of a. From the Strong Markov property, $\mathbb{E}_{x^*,\theta}[b^{\tau_C}] < +\infty$ for any x^* large enough.

We conclude that in both cases for C, if equation (4.40) holds, then for any x^* large enough and for some $\theta > 0$,

$$\mathbb{E}_{x^*,\theta}[b^{\tau_C}] < +\infty. \tag{4.41}$$

We will ultimately prove that equation (4.41) will lead to contradiction. We treat the two cases for the refreshed rate γ, as introduced in Theorem 4.4.3 differently.

We first consider the case where $\gamma(x) \xrightarrow{x\to+\infty} 0$. Let $\epsilon < \log b$. We can pick x^* as in (4.41) large enough such that for all $x \geq x^*$, $|U'(x)| + \gamma(x) \leq \epsilon$. Furthermore, the set $\{x \in \mathbb{R} : (x, \theta) \in C\}$ is bounded above, due to Lemma 4.31, so we can pick $x^* > \sup\{x \in \mathbb{R} : (x, \theta) \in C\}$. This means that if the process starts from (x^*, θ), before switching a velocity the process will take values (y, θ) for $y >^* x$ and therefore $(y, \theta) \notin C$. Therefore, the process cannot hit C before first switching velocity. Now, if T is the first time the velocity switches, we have that T stochastically dominates an exponential random variable with parameter ϵ. Therefore, as in the proof of Proposition 4.3.6, for any $n \in \mathbb{N}$ we write,

$$\mathbb{E}_{(x^*,\theta)}[b^{\tau_C}] \geq \mathbb{E}_{(x^*,\theta)}[b^T] \geq b^n \mathbb{P}_{(x^*,\theta)}(T \geq n) = b^n \exp\left\{-\int_{x^*}^{x^*+n} \lambda(z, \theta) dz\right\} \geq$$

$$\geq b^n \exp\{-n\epsilon\} = \left(\frac{b}{\exp\{\epsilon\}}\right)^n \xrightarrow{n\to\infty} +\infty$$

which is a contradiction.

For the rest of the proof we focus on the case where there exists an $M > 0$

such that for all $x \in \mathbb{R}$, $\gamma(x) \leq M$ and

$$\lim_{x\to\infty} \frac{U'(x)}{\gamma(x)} = 0. \tag{4.42}$$

Let us pick $\epsilon > 0$ small enough, to be specified later. We can pick (x^*, θ) so that (4.41) holds and x^* is large enough so that for all $x \geq x^*$, $|U'(x)| \leq \epsilon$ and

$$\frac{|U'(x)|}{\gamma(x)} \leq \epsilon. \tag{4.43}$$

Let us also pick $\delta > 0$ small enough, to be specified later and take $\theta^* > \theta$ such that $Q(\theta^*) = q(\theta^*, +\infty) < \delta$. From Lemma 4.31 the set $A = \{x \in \mathbb{R} : (x, \eta) \in C \text{ for some } |\eta| \leq \theta^*\}$ is bounded so we can further assume that we pick x^* such that (4.41) holds in such a way so that $x^* > \sup A$. Note that for any η with $|\eta| \leq \theta^*$ and any $x \geq x^*$, $(x, \eta) \notin C$. We will prove that the process starting from (x^*, θ) will spend a lot of time inside the set $B = \{(x, \eta) : x > x^*, |\eta| \leq \theta^*\}$ and this will be a contradiction as the process needs to hit C fast enough for (4.41) to hold. Note that there are two ways for the RVRZZ process $Z_t = (X_t, \Theta_t)$ to exit B. The first one is for X_t to become less or equal to x^* and the other is for a refresh event to occur and for a new velocity η to be picked with $|\eta| > \theta^*$. Picking such a velocity has probability less than δ and as long as such an event does not occur, the process starting from (x^*, θ) needs to return to x^* before it returns to C, i.e. $\tau_{x^*} = \inf\{t \geq \delta_1 : X_t \leq x^*\} \leq \tau_C$.

Let us consider, for now, the event A defined as: *that for the first n up-steps and down-steps, every occurred velocity refreshment chose new velocity that did not have an absolute value larger than θ^*.* Therefore the new velocities refreshments are chosen according to measure to the measure q^* which is q conditioned on having a value less than θ^*.

We can decompose the movements of the process in up-steps and down-steps as in the original Zig-Zag and use the tools from the previous section. We write for any $k \in \mathbb{N}$, U_k for the time it takes to complete the k'th up-step and D_k the corresponding time for the k'th down-step. The main difference with the simple Zig-Zag is what happens to velocities at the end of an up-step or a down-step. More specifically, in RVRZZ, when the process moves with velocity θ and ends an up-step or a down-step at point x, then a $v \sim unif(0, 1)$ that is independent of the rest of the process randomness is sampled and on the event

$$v \leq \frac{\gamma(x)}{[\theta U'(x)]^+ + \gamma(x)}, \tag{4.44}$$

we pick a velocity θ' of opposite sign and independent of θ, such that $-\theta' \sim q^*$. Otherwise, we set $\theta' = -\theta$. Note here that as long as $x \geq x^*$, the quantity on the RHS of (4.44), can be bounded from below, similarly to equation (4.36) in the proof of Lemma 4.4.7 by

$$\frac{\gamma(x)}{[\theta U'(x)]^+ + \gamma(x)} \geq \frac{1}{1+\theta^*\epsilon}. \tag{4.45}$$

As a next step, we will further restrict to the event B defined as: *at the end of each of the first n up-steps and down-steps, all the i.i.d. uniform distributions $v_1, ..., v_{2n}$ that decide how the process will update the velocity after the first n up-steps and down-steps (as described on (4.44)), all took values less than $1/(1+\theta^*\epsilon)$.* This event has probability

$$\left(\frac{1}{1+\theta^*\epsilon}\right)^{2n}$$

and due to (4.45), on this event, at the end of each one of the first n up-steps and down-steps, the new velocity is refreshed independently of the previous velocities and according to q^*. Note that event B involves only the evaluation of $v_1, ..., v_n$ which are independent of the rest of the process' randomness, therefore by conditioning on A and B we do not influence the law of the process, except by forcing all the velocity changes to be refreshments from q^*.

We write S_n for the position of the process after the n'th down-step, which plays the role of the random walk, as in the previous section. Let $T = \inf\{n \geq 1 : S_n \leq x^*\}$. When we restrict to $A \cap B$ the process will not hit C before the x-component takes a value less than x^*. We bound, as in equation (4.20), for all $n \in \mathbb{N}$

$$\begin{aligned}\mathbb{E}_{x^*,\theta}[b^{\tau_C}] &\geq \mathbb{E}_{x^*,\theta}[b^{\sum_{k=1}^n U_k} 1_{T\geq n+1} 1_B 1_A] \\ &\geq (1-\delta)^{2n} \left(\frac{1}{1+\theta^*\epsilon}\right)^{2n} \mathbb{E}_{x^*,\theta}[b^{\sum_{k=1}^n U_k} 1_{T\geq n+1} | B, A]. \end{aligned} \tag{4.46}$$

From now on we condition on events A and B, meaning that every time a velocity change occurs, this is accomplished by refreshing and a new velocity, with different sign from the previous one, is picked with absolute value coming from q^*, the conditional measure q truncated at θ^*.

We note here that due to Lemma 4.3.12, if the RHS of (4.46) was finite, then the equivalent expectation for the process with higher hazard rates on up-steps and lower hazard rates on down-steps would be finite as well. Since $0 \leq [\theta U'(x)]^+ \leq |\theta||U'(x)|$, we can focus on the case where the hazard rate during an up-step with

velocity $\theta > 0$ is

$$\lambda(x, \theta) = |U'(x)|\theta + \gamma(x), \tag{4.47}$$

and for a down-step with speed $-\theta$ for $\theta > 0$ we can assume that the hazard rate is

$$\lambda(x, -\theta) = \gamma(x). \tag{4.48}$$

This is exactly what was described before as the Random Velocity Bounding Model. Conditioning on the down-steps and on how the velocities will be refreshed, using FKG inequality and then taking expectation over the configurations of the down-steps and the refreshing velocities, just as in the proof of Proposition 4.3.17 in the previous section, we can bound

$$\mathbb{E}_{x^*,\theta}[b^{\sum_{k=1}^n U_k} 1_{T \geq n+1} | A, B_n] \geq \mathbb{E}_{x^*,\theta}[b^{\sum_{k=1}^n A_k} | A, B_n] \mathbb{P}(T^\epsilon \geq n+1 | A, B_n) \tag{4.49}$$

where $A_k \sim \exp(M + \epsilon\theta_k)$ and $\theta_k \sim q^*$ i.i.d and T^ϵ is the first time S_n becomes less than x^* when the hazard rates satisfy (4.47) and (4.48). Now,

$$\mathbb{E}_{x^*,\theta}[b^{\sum_{k=1}^n A_k} | A, B_n] = \left(\int_0^{\theta^*} \frac{M + \epsilon\theta'}{M + \epsilon\theta' - \log b} q^*(d\theta') \right)^n \geq \left(\frac{M + \epsilon\theta^*}{M + \epsilon\theta^* - \log b} \right)^n.$$

Therefore, using (4.46)

$$\mathbb{E}_{x^*,\theta}[b^{\tau_C}] \geq (1-\delta)^{2n} \left(\frac{1}{1 + \theta^*\epsilon} \right)^{2n} \left(\frac{M + \epsilon\theta^*}{M + \epsilon\theta^* - \log b} \right)^n \mathbb{P}(T^\epsilon \geq n+1 | A, B_n). \tag{4.50}$$

Using Lemma 4.4.7 we get that for all $\epsilon, \delta > 0$ there exists a $C > 0$ such that for all $n \in \mathbb{N}$

$$\mathbb{E}_{x^*,\theta}[b^{\tau_C}] \geq \left((1-\delta)^2 \frac{M + \epsilon\theta^*}{M + \epsilon\theta^* - \log b} \left(\frac{1}{1 + \epsilon\theta^*} \right)^2 \left(\frac{1}{1 + \theta^*\epsilon} \right)^2 \right)^n C \frac{1}{n^{1/2}}.$$

When δ and ϵ are chosen appropriately small, the RHS diverges to $+\infty$ as we let $n \to \infty$ and this is a contradiction. □

Chapter 5

Speed Up Zig-Zag

In order to address the problem of slow mixing on heavy tails we introduce a variant of the Zig-Zag process, called Speed Up Zig-Zag (SUZZ). Instead of allowing the particle to move with unit velocity, we allow it to have a positive speed depending on the current position. Since in high dimensions this might create a system of ODEs that is not easy to solve, making such a process non-implementable, we only allow the process to move in directions $\{\pm 1\}^d$ as the original Zig-Zag does. What is different is the speed in which the process moves along these straight lines.

We begin this chapter by introducing the process in Section 5.1. Then, in Section 5.2 we first establish sufficient conditions for the process to be non-explosive. Afterwards, in Section 5.3, we prove that if we pick the underlying Poisson Process rates appropriately, we recover the measure of interest as invariant measure of the process. We continue by establishing conditions that imply geometric ergodicity for the speed up process in Section 5.4. Subsequently, in Section 5.5 we relate the one dimensional SUZZ process with a one dimensional Zig-Zag and in Section 5.7 we provide some criteria on how to choose the speed function for the process. Finally, in Section 5.8 we provide some simulation results, comparing the speed up process with the original one.

5.1 Definition of the Algorithm

In this section we will introduce a variation of the Zig-Zag process that can help overcome the problem that one encounters in heavy tails. The state space will, again, be $E = \mathbb{R}^d \times \{\pm 1\}^d$. However, when the process is at point $(x, \theta) \in E$, it will move along the path $\{x + \theta t, t \geq 0\}$ with speed that depends on the current position of the particle. Typically, this speed will increase the further the process is from

the mode. After a random time that will depend on a Poisson process, as in the original Zig-Zag, the process will stop, one of the coordinates of θ will switch and the process will start moving again to the new direction. This will create excursions that tend to leave the area of high density and visit the tails quite often. At the same time, when the process is at the tails of the distribution, it can speed up and return to the centre fast enough. We define the **Speed Up Zig-Zag (SUZZ)** on $E \cup \{\partial\}$ (where ∂ is a graveyard state we need for technical reasons), with C^1 **speed function** $s : \mathbb{R}^d \to (0, \infty)$ and rate functions $\lambda_i : E \to [0, +\infty)$ for all $i \in \{1, ..., d\}$ as follows.

Let $(E_n^i)_{i\in\{1,...,d\},n\in\mathbb{N}}$ i.i.d. random variables following an exp(1) distribution. Suppose that the process starts from $(x, \theta) \in E$. Let $\{\Phi_{(x,\theta)}(t)\}_{t\geq 0}$ be the flow on $\mathbb{R}^d$ that solves the ODE system

$$\begin{cases} \dfrac{d\Phi_{(x,\theta)}(t)}{dt} = \theta s(\Phi_{(x,\theta)}(t)), t \in [0, t^*(x, \theta)) \\ \Phi_{(x,\theta)}(0) = x, \end{cases} \tag{5.1}$$

where $t^*(x, \theta) = \sup\{t \geq 0 : \Phi_{(x,\theta)}(t) \in \mathbb{R}^d\}$ is the explosion time of the flow, with the convention that if the flow does not explode this is set $t^*(x, \theta) = +\infty$. Note that if $s \in C^1$, the ODE system (5.1) has a unique solution. Note as well that this solution flow moves in a straight line parallel to $\theta \in \{\pm 1\}^d$. For all $i \in \{1, ..., d\}$ let

$$\tau_1^i = \inf\{t \in [0, t^*(x, \theta)) : \int_0^t \lambda_i(\Phi_{(x,\theta)}(s), \theta)ds \geq E_1^i\},$$

where we will always use the convention that $\inf \emptyset = +\infty$. Let $\tau_1 = \min\{\tau_1^i, i \in \{1, ..., d\}\}$, $T_1 = \tau_1$ and $i_1 = argmin\{\tau_1^i, i \in \{1, ..., d\}\}$.
If $\tau_1 = \infty$ we set $(X_t, \Theta_t) = (\Phi_{(x,\theta)}(t), \theta)$ for all $t < t^*(x, \theta)$ and $(X_t, \Theta_t) = \partial$ for $t \geq t^*(x, \theta)$.
If $\tau_1 < \infty$ set $(X_t, \Theta_t) = (\Phi_{(x,\theta)}(t), \theta)$ for all $0 \leq t < T_1$. Then set $(X_{T_1}, \Theta_{T_1}) = (\Phi_{(x,\theta)}(T_1), F_{i_1}[\theta])$, where $F_{i_1}[\theta]) \in \{\pm 1\}^d$ such that if $\theta = (\theta_1, ..., \theta_d)$,

$$\begin{cases} (F_{i_1}[\theta])_k = -\theta_k \text{ for } k = i_1 \\ (F_{i_1}[\theta])_k = \theta_k \text{ for } k \neq i_1. \end{cases} \tag{5.2}$$

We then continue the construction inductively, for any $n \in \mathbb{N}$. If $T_n < \infty$ we then consider the flow $\{\Phi_{(X_{T_n},\Theta_{T_n})}(t)\}_{t\geq 0}$ and for all $i \in \{1, ..., d\}$ and let

$$\tau_{n+1}^i = \inf\{t \in [0, t^*(X_{T_n}, \Theta_{T_n})) : \int_0^t \lambda_i(\Phi_{(X_{T_n},\Theta_{T_n})}(s), \Theta_{T_n})ds \geq E_{n+1}^i\}$$

and let $\tau_{n+1} = \min\{\tau^i_{n+1}, i \in \{1, ..., d\}\}$, $T_{n+1} = T_n + \tau_{n+1}$ and $i_{n+1} = argmin\{\tau^i_{n+1}, i \in \{1, ..., d\}\}$.

If $\tau_{n+1} = \infty$, we set $(X_{T_n+t}, \Theta_{T_n+t}) = (\Phi_{(X_{T_n}, \Theta_{T_n})}(t), \Theta_{T_n})$ for all $t < t^*(X_{T_n}, \Theta_{T_n})$ and $(X_{T_n+t}, \Theta_{T_n+t}) = \partial$ for $t \geq t^*(X_{T_n}, \Theta_{T_n})$.

If $\tau_{n+1} < \infty$ set $(X_{T_n+t}, \Theta_{T_n+t}) = (\Phi_{(X_{T_n}, \Theta_{T_n})}(t), \Theta_{T_n})$ for all $0 \leq t < \tau_{n+1}$. Then set $\left(X_{T_{n+1}}, \Theta_{T_{n+1}}\right) = \left(\Phi_{(X_{T_n}, \Theta_{T_n})}(\tau_{n+1}), F_{i_{n+1}}[\Theta_{T_n}]\right)$.
This defines the process until time $\xi = \lim_{n \to +\infty} T_n$ which is the first time that infinitely many jumps occur. We define $(X_t, \Theta_t) = \partial$ for all $t \geq \xi$.

The process is therefore defined as a Piecewise Deterministic Process in [Dav84] would be. The difference is that we allow the deterministic dynamics to have a finite explosion type, which Davis in [Dav84] does not. We therefore need to be more careful in the analysis of the process. Recall that O_m is the ball of radius m centred around the origin 0. We define $\zeta_m = \inf\{t \geq 0 : X_t \notin O_m\}$ and let $\zeta = \lim_{m \to \infty} \zeta_m$. These two random variables ξ and ζ quantify two types of explosion that can occur for the process. The first is that the process could have infinitely many switches and the second, that the process might diverge to infinity in finite time. Note that the way we have constructed the process, we have $Z_t = (X_t, \Theta_t) \in E$ only for $t < \xi \wedge \zeta$. For formality let us define $Z_t = \partial$ for $t \geq \xi \wedge \zeta$.

In algorithmic terms the process in the event where $\{\zeta = +\infty\}$ is described as follows.

Algorithm 5.1.1 (Speed Up Zig-Zag)**.**

1. *Set* $t_{current} = 0$
2. *Start from point* $(X_{t_{current}}, \Theta_{t_{current}}) = (x, \theta) \in E$.
3. *The process* (X_t, Θ_t) *moves according to the deterministic ODE system*

$$\begin{cases} \frac{d}{dt} X_{t_{current}+t} = \frac{d\Phi_{(x,\theta)}(t)}{dt} = \theta s(X_{t_{current}+t}), t \geq 0\} \\ \frac{d}{dt} \Theta_{t_{current}+t} = 0, t \geq 0 \\ X_{t_{current}} = x, \Theta_{t_{current}} = \theta \end{cases} \tag{5.3}$$

4. *For every coordinate* $i \in \{1, ..., d\}$ *construct a Poisson Process with intensity* $\{m_i(s) = \lambda_i(\Phi_{(x,\theta)}(s), \theta), s \geq 0\}$, *for some* $\lambda_i : E \to [0, +\infty)$.
5. *Let* τ_i *be the first arrival time of the* i *Poisson Process, i.e. for all* $s \geq 0$, $\mathbb{P}(\tau_i \geq s) = \exp\{-\int_0^t m_i(u)du\}$. *Let* $j = argmin\{\tau_i, i = 1, ..., d\}$ *and* $\tau = \tau_j$ *the first arrival time of all the processes.*

6. *For* $t \in [t_{current}, t_{current} + \tau)$ *set* $X_t = \Phi_{(x,\theta)}(t - t_{current})$ *and* $\Theta_t = \theta$.

7. *Set* $x = \Phi_{(x,\theta)}(\tau)$, $X_{t_{current}+\tau} = x$, $\Theta_{t_{current}+\tau} = F_j[\theta]$ *and* $t_{current} = t_{current} + \tau$.

8. *Repeat from the Step 2.*

Note that the system of ODEs (5.3) leads to dynamics that are straight lines moving along the directions $\{\pm 1\}^d$ and the function λ_i plays the role of the intensity function that will switch the i component of the direction.

We will now present some first properties of the process. We begin by noting that the only way for the process to explode in finite time is to reach infinity in finite time and until that point, only finitely many switches occur. We have the following.

Lemma 5.1.2. *Assume that the speed function* $s \in C^1$ *is strictly positive and the rate functions* λ_i *are locally bounded for all* i. *Let* ζ_m, ζ *and* ξ *defined as in the discussion before Algorithm 5.1.1. Then almost surely* $\zeta_m < \xi$ *for all* $m \in \mathbb{N}$. *Therefore, a.s.* $\zeta \leq \xi$.

Proof of Lemma 5.1.2. Let $\bar{\lambda}$ an upper bound for all λ_i on O_m and let $E_1, E_2, ...$ a configuration of the i.i.d. exponential random variables used to construct the $(X_t, \Theta_t)_{t\geq 0}$ process such that $\xi \leq \zeta_m$. Let $T_1 < T_2 < ...$ the switching times of the process. Then for all $t < \xi$, we have $t < \zeta_m$ therefore $X_t \in O_m$. By the definition of the switching times T_k (and writing $T_0 = 0$) we get

$$\sum_{k=1}^{\infty} E_k = \sum_{k=1}^{\infty} \int_{T_{k-1}}^{T_k} \lambda_{i_k}(X_t, \Theta_t) dt \leq \xi \bar{\lambda}.$$

and therefore

$$\mathbb{P}\left(\{\xi < \infty\} \cap \{\xi \leq \zeta_m\}\right) \leq \mathbb{P}\left(\sum_{k=1}^{\infty} E_k \leq \xi \bar{\lambda} < \infty\right) = 0.$$

Let t_m be the maximum time it takes for a flow that starts from inside O_m and solves (5.1) to exit O_m. For any m, on the event $\{E_n \geq \bar{\lambda} t_m\}$, if the process has not escaped the ball O_m until the $n-1$'th switch, it does so following the dynamics before the n'th switch happens. Since $\mathbb{P}\left(E_n \geq \bar{\lambda} t_m\right) = a > 0$ we have for all n, $\mathbb{P}\left(\zeta_m > n\right) \leq (1-a)^n$ and therefore $\mathbb{P}\left(\zeta_m = +\infty\right) = 0$.

Overall this gives $\mathbb{P}\left(\xi \leq \zeta_m\right) = 0$. □

Lemma 5.1.3 (Local Lemma)**.** *If the speed function* $s \in C^1$ *and the rate functions are locally bounded, then for all* $x \in \mathbb{R}^d$ *there exists a neighbourhood* U_x *of* x *and a*

time $t > 0$ such that for any $\theta \in \{\pm 1\}^d$ if the SUZZ starts from (x, θ), then $X_s \in U_x$ for all $0 \leq s \leq t$.

Proof of Lemma 5.1.3. For $x \in \mathbb{R}^d$ we find a small neighbourhood U_x so that s is bounded on U_x by $\bar{s}$. Take t small enough so that $t\sqrt{d}\bar{s} < dist(x, \partial U_x)$ so that a path starting from x and following a straight line with speed $s(x) \leq \bar{s}$ in each coordinate for time less than t, will not have exit U_x. Then any path moving in directions $\{\pm 1\}^d$ with speed $s(x)$ in each component that switches direction finitely many times will not have exit U_x. From Lemma 5.1.2 a.s. the Zig-Zag process will switch direction finitely many times until it exits the bounded set U_x and this proves that the process a.s. stays inside U_x until time t. □

Finally, we have the following proposition, which gives us intuition on what the generator of the process is.

Proposition 5.1.4. *For any function $f \in C^1(E)$ and any $(x, \theta) \in E$,*

$$\lim_{t \to 0} \frac{\mathbb{E}_{x,\theta}[f(Z_t)] - f(x,\theta)}{t} = \mathcal{L}f(x,\theta) = \sum_{i=1}^{d} \theta_i s(x) \partial_i f(x,\theta) + \lambda_i(x,\theta)(f(x, F_i[\theta]) - f(x,\theta)). \tag{5.4}$$

Proof of Proposition 5.1.4. For given (x, θ) we know from Lemma 5.1.3 that for small t the quantity $\mathbb{E}_{x,\theta}[f(Z_t)]$ is well defined so the limit makes sense. Write
$S_i(t) = \{$the i coordinate switches before time t and is the first coordinate to switch$\}$, for $i = 1, .., d$ and
$S_0(t) = \{$ no coordinate switches until time $t\}$. Note that if the process starts from (x, θ) and if T_i is the first arrival time of the Poisson process with intensity $t \to \lambda_i(\Phi_t(x,\theta), \theta)$, then $\mathbb{P}(T_i \geq t) = \exp\{-\int_0^t \lambda_i(\Phi_u(x,\theta),\theta)\}du$ therefore the density of T_i is

$$f_{T_i}(t) = \lambda_i(\Phi_t(x,\theta),\theta) \exp\left\{-\int_0^t \lambda_i(\Phi_u(x,\theta),\theta)\right\} du$$

therefore

$$\mathbb{P}(S_i(t)) = \int_0^t \mathbb{P}(\text{ no switches until time } t \text{ for any component } j \neq i)\, f_{T_i}(u)du =$$
$$= \int_0^t \lambda_i(\Phi_u(x,\theta),\theta) \exp\left\{-\int_0^u \lambda_i(\Phi_{u'}(x,\theta),\theta)du'\right\} \prod_{j \neq i} \exp\left\{-\int_0^u \lambda_j(\Phi_{u'}(x,\theta),\theta)du'\right\} du =$$
$$\int_0^t \lambda_i(\Phi_u(x,\theta),\theta) \exp\left\{-\int_0^u \lambda(\Phi_{u'}(x,\theta),\theta)du'\right\} du.$$

Therefore, from a L'Hospital rule

$$\lim_{t\to 0}\frac{\mathbb{P}\left(S_i(t)\right)}{t}=\lambda_i(x,\theta).$$

For $f\in C^1(E)$, conditioning on the coordinate which was the first to switch before t (or whether no switch happened)

$$\begin{aligned}
&\lim_{t\to 0}\frac{\mathbb{E}_{x,\theta}[f(X_t,\Theta_t)]-f(x,\theta)}{t}=\\
&=\lim_{t\to 0}\mathbb{P}_{x,\theta}\left(S_0(t)\right)\frac{f(\Phi_t(x,\theta),\theta)-f(x,\theta)}{t}+\sum_{i=1}^{d}\lim_{t\to 0}\frac{\mathbb{P}(S_i(t))}{t}\left(\mathbb{E}_{x,\theta}[f(X_t,\Theta_t)|S_i(t)]-f(x,\theta)\right)=\\
&=\sum_{i=1}^{d}\lim_{t\to 0}\exp\left\{-\int_0^t\lambda(\Phi_s(x,\theta),\theta)ds\right\}\partial_i f(x,\theta)\theta_i s(x)+\\
&+\sum_{i=1}^{d}\lambda_i(x,\theta)\left(\mathbb{E}_{x,\theta}\left[f(X_t,\Theta_t)|S_i(t)\right]-f(x,\theta)\right)=\\
&=\sum_{i=1}^{d}\theta_i s(x)\partial_i f(x,\theta)+\lambda_i(x,\theta)[f(x,F_i[\theta])-f(x,\theta)].
\end{aligned}$$

□

Assume that we are trying to target the measure

$$\mu(dx)=\frac{1}{Z}\exp\{-U(x)\}d\mu_0 \tag{5.5}$$

where μ_0 is the product between the Lebesgue measure in $\mathbb{R}^d$ and the uniform measure on $\{-1,+1\}^d$. We will later see that in order to target this measure using a SUZZ process with speed function s, the rates must satisfy the following assumption.

Assumption 5.1.5. *Assume that $U,s\in C^2$ and the rates of the process satisfy*

$$\lambda_i(x,\theta)=[\theta_i A_i(x)]^+ +\gamma_i(x,\theta_{-i}) \tag{5.6}$$

where γ_i is a non-negative, locally bounded, integrable function that does not depend on the i'th component of θ (and θ_{-i} is θ without the i'th component) and

$$A_i(x)=s(x)\partial_i U(x)-\partial_i s(x). \tag{5.7}$$

Note that if we use $s\equiv 1$ we retrieve the normal Zig-Zag rates when targeting μ. A usual way to prove that a measure μ is invariant for a PDMP is to prove that when combined with the generator $\mathcal{L}$, the operator $\mu\mathcal{L}$ defined by $\mu\mathcal{L}f=\mu(\mathcal{L}f)$ is zero everywhere. To motivate our choice of rates in assumption 5.1.5 we present the

following Lemma.

Lemma 5.1.6. *Assume that $s, U \in C^2(E)$ and the rates λ_i satisfy Assumption 5.1.5 for all $i = 1, ..., d$. Let $\mathcal{L}$ be the operator defined in (5.4). Then for the measure μ introduced in (5.5) and for every function $f \in C_c^1(E)$ we have $\mu(\mathcal{L}f) = 0$.*

Proof of Lemma 5.1.6. Assume that the rates satisfy Assumption 5.1.5. Following the argument in [BR17] we get for $f \in C_c^1(E)$

$$
\begin{aligned}
&2^d Z \mathbb{E}_\mu[\mathcal{L}f(X,\Theta)] = \\
&= \sum_{i=1}^{d} \sum_{\theta\in\{\pm1\}^d} \int_{\mathbb{R}^d} \exp\{-U(x)\}[\theta s(x)\partial_i f(x,\theta) + \lambda_i(x,\theta)(f(x,F_i[\theta]) - f(x,\theta))]dx = \\
&= \sum_{i=1}^{d} \sum_{\theta\in\{\pm1\}^d} \int_{\mathbb{R}^d} \exp\{-U(x)\}\theta s(x)\partial_i f(x,\theta)dx + \\
&+ \sum_{i=1}^{d} \sum_{\theta\in\{\pm1\}^d} \int_{\mathbb{R}^d} \exp\{-U(x)\}\lambda_i(x,\theta)(f(x,F_i[\theta]) - f(x,\theta))dx = \\
&= \sum_{i=1}^{d} \sum_{\theta\in\{\pm1\}^d} \int_{\mathbb{R}^d} \partial_i(-\exp\{-U(x)\}\theta s(x))f(x,\theta)dx + \\
&+ \sum_{i=1}^{d} \sum_{\theta\in\{\pm1\}^d} \int_{\mathbb{R}^d} \exp\{-U(x)\}f(x,\theta)(\lambda_i(x,F_i[\theta]) - \lambda_i(x,\theta)) = \\
&= \sum_{i=1}^{d} \sum_{\theta\in\{\pm1\}^d} \int_{\mathbb{R}^d} f(x,\theta)\exp\{-U(x)\}\left[\theta(s(x)\partial_i U(x) - \partial_i s(x)) - (\lambda_i(x,\theta) - \lambda_i(x,F_i[\theta]))\right]dx \\
&= 0.
\end{aligned}
$$

where the third equation is an integration by parts and the boundary terms are 0 since f has compact support. □

5.2 Non-Explosivity

Most MCMC algorithms do not suffer from explosiveness, since by construction they can guarantee that the process will not reach infinity after finitely many steps. However, in SUZZ, even in one dimension, picking a large enough speed function s can lead to deterministic dynamics that explode in finite time. For example, the

ODE

$$\begin{cases} \dfrac{dX_t}{dt} = X(t)^{1+\epsilon} \\ X(0) = 1 \end{cases}$$

admits solution $X(t) = (1-\epsilon t)^{-1/\epsilon}, t \leq \epsilon^{-1}$ which explodes at time $t^* = \epsilon^{-1}$. Reaching infinity in finite time will lead to very unstable algorithms that cannot be implemented in the computer. However, this type of dynamics have the interesting property that if we reverse the time, they can return close to the mode of the target distribution very fast, even if they start from very far apart. The main idea is that one can allow the deterministic dynamics to be explosive, as long as a switching Poisson process is also introduced, having a very large intensity. This intensity will switch the direction the process is moving towards before it reaches the explosion time. In this section we provide conditions the rates should satisfy for the process to be non-explosive. To do this we will use techniques from [MT93b].

Definition 5.2.1. *Let $\zeta = \lim_{m\to\infty} \zeta_m$ the random variable introduced in the previous section. The process is called non-explosive if $\zeta = +\infty$ a.s.*

We begin with the most essential assumption for the speed function, in that if this is not satisfied, there will be a positive probability that the process might reach the finite explosion time without switching dynamics and therefore explode.

Assumption 5.2.2 (Speed Growth)**.**

$$\lim_{\|x\|\to\infty} \exp\{-U(x)\}s(x) = 0. \tag{5.8}$$

Proposition 5.2.3. *If the process satisfies Assumption 5.2.2 then for any starting point (x,θ) the process will switch dynamics before the explosion time $t^*(x,\theta)$.*

Proof of Proposition 5.2.3. Suppose that the process starts from $(x,\theta) = (x_1,..,x_d;\theta_1,..,\theta_d)$. The process evolves under the deterministic dynamics given in (5.3) until the explosion time $t^* = t^*(x,\theta)$. Note that under these dynamics, which are a straight line parallel to $\theta = (\theta_1,...,\theta_d) \in \{\pm 1\}^d$, $X_t = (X_t^{(1)},..,X_t^{(d)})$ satisfies $\left(X_t^{(k)} - x_k\right)\theta_k = \left(X_t^{(1)} - x_1\right)\theta_1$ and solving for $X_t^{(k)}$ we get

$$X_t^{(k)} = x_k - \theta_1\theta_k x_1 + \theta_1\theta_k X_t^{(1)} = c_k + \theta_1\theta_k X^{(1)}(t)$$

for all $k \in \{1,..,d\}$, where $c_k = x_k - \theta_1\theta_k x_1$. Consider the Poisson process with rate

$\{m(t) = \lambda(X_t, \theta) = \sum_{k=1}^{d} \lambda_k(X_t, \theta), t \geq 0\}$ and note that

$$\int_0^{t^*} m(t)dt = \int_0^{t^*} \lambda(X_t^{(1)}, c_2 + \theta_1\theta_2 X_t^{(1)}, ..., c_d + \theta_1\theta_d X_t^{(1)})dt =$$
$$= \int_{x_1}^{sign(\theta_1)\infty} \lambda(u, c_2 + \theta_1\theta_2 u, .., c_d + \theta_1\theta_d u) \frac{1}{s(u, c_2 + \theta_1\theta_2 u, .., c_d + \theta_1\theta_d u)} \theta_1 du \geq$$
$$\geq \int_{x_1}^{sign(\theta_1)\infty} \sum_{i=1}^{d} \theta_1\theta_i \partial_i U(u, c_2 + \theta_1\theta_2 u, .., c_d + \theta_1\theta_d u) -$$
$$- \theta_1\theta_i \frac{\partial_i s(u, c_2 + \theta_1\theta_2 u, .., c_d + \theta_1\theta_d u)}{s(u, c_2 + \theta_1\theta_2 u, .., c_d + \theta_1\theta_d u)} du =$$
$$= \int_{x_1}^{sign(\theta_1)\infty} \frac{d}{du}[U(u, c_2 + \theta_1\theta_2 u, .., c_d + \theta_1\theta_d u) - \log s(u, c_2 + \theta_1\theta_2 u, .., c_d + \theta_1\theta_d u)]du =$$
$$= \lim_{u \to sign(\theta_1)\infty} U(u, c_2 + \theta_1\theta_2 u, .., c_d + \theta_1\theta_d u) - \log s(u, c_2 + \theta_1\theta_2 u, .., c_d + \theta_1\theta_d u) - C =$$
$$+ \infty.$$

by assumption 5.2.2 and so $\mathbb{P}_{x,\theta}(\text{no switches until time } t^*(x, \theta)) = \exp\{-\int_0^{t^*(x,\theta)} m(t)dt\} = 0$. □

Remark 5.2.4. *Note that Assumption 5.2.2 is essential for the result of Proposition 5.2.3 to hold. Consider a one dimensional SUZZ with speed s targeting a distribution that has U as minus log-likelihood. Assume that there exists a x_0 such that for all $x \geq x_0$, $(U(x) - \log s(x))' > 0$, as would be the case for any Gaussian target Gaussian and $s(x)$ a polynomial, for example. Then doing the same calculations as in the proof of Proposition 5.2.3 we get*

$$\int_{x_0}^{t^*} m(t)dt = \lim_{x \to +\infty} U(x) - \log s(x) - (U(x_0) - \log s(x_0)).$$

Therefore assuming that $s(x)\exp\{-U(x)\} \xrightarrow{x \to +\infty} a > 0$ we get that

$$\lim_{x \to +\infty} U(x) - \log(s(x)) = \log(a^{-1}) < \infty$$

and therefore

$$\mathbb{P}_{x_0,+1}(\text{no switches until time } t^*(x_0, +1)) = \exp\left\{-\int_0^{t^*(x_0,+1)} m(t)dt\right\} > 0.$$

Therefore, if $t^*(x_0, +1) < \infty$ then the process has a positive probability to explode. Even if $t^*(x_0, +1) = \infty$ the process still has a positive probability of never switching

direction towards -1. The same situation is experienced in higher dimensions assuming that for all coordinates i, $\partial_i(U(x) - \log s(x)) > 0$ for all $x = (x_1, ..., x_d)$ for which x_i is positive and very large. This forces us to adopt assumption 5.2.2.

Except for Assumption 5.2.2, in order to prove non-explosivity of the process we will, also, make the following assumption.

Assumption 5.2.5. *Assume that all the refresh rates depend only on x and are bounded, i.e. there exists $\bar{\gamma}$ such that for all $i \in \{1, ..., d\}, \gamma(x, \theta_{-i}) = \gamma_i(x) \leq \bar{\gamma}$. Furthermore, the gradient rates do not decay to zero, i.e. there exists $R > 0$ and $A > 0$ so that for all $x \notin B(0, R)$*

$$\sum_{i=1}^{d} |A_i(x)| > A > \max\{3d\bar{\gamma}, 4d(d-1)\bar{\gamma}\}. \tag{5.9}$$

Remark 5.2.6. *In the setting of original Zig-Zag the authors of [BRZ19] make the assumption that $\|\nabla U(x)\|_1 \xrightarrow{\|x\| \to \infty} \infty$ in order to prove Geometric Ergodicity. Note that in the setting of the normal Zig-Zag $\sum_{i=1}^{d} |A_i(x)| = \|\nabla U(x)\|_1$, therefore Assumption 5.2.5 can be seen as a more relaxed version of the assumption of [BRZ19].*

When all refresh rates are zero, then $\bar{\gamma} = 0$ and the condition in equation (5.9) is that the overall switching rate is bounded away from zero. Heuristically, function $|A_i|$ should never be lower than γ_i. This is because the former function describes the intention of the algorithm to switch from a direction leading to lower density areas, while the latter describes the intention of the algorithm to switch direction for a random reason. This somehow explains equation (5.9), although we believe that our lower bound on A in terms of γ is not optimal.

Our next assumption seems more arbitrary and probably more difficult to verify in practice.

Assumption 5.2.7. *For all $i \in \{1, ..., d\}$*

$$\lim_{\|x\| \to \infty} \frac{s(x)}{\sum_{k=1}^{d} |A_k(x)|} \sum_{j=1}^{d} \frac{\sum_{i=1}^{d} |\partial_i A_j(x)|}{(1 + |A_j(x)|)(1 + \log(1 + |A_j(x)|))} = 0. \tag{5.10}$$

Furthermore,

$$\lim_{\|x\| \to \infty} \frac{\log |\log |\nabla (U(x) - \log s(x))||}{U(x) - \log s(x)} = 0$$

and

$$\lim_{\|x\| \to \infty} \frac{\log |\log |s(x)||}{U(x) - \log s(x)} = 0$$

Remark 5.2.8. *This assumption, when seen in the setting of normal Zig-Zag where speed $s \equiv 1$, is a more relaxed version of the last two equations in (2.34) of Theorem 2.4.7. More precisely, in view of Assumption 5.2.2, in the case of original Zig-Zag, assumption 5.2.7 writes*

$$\lim_{\|x\|\to\infty} \frac{\|Hess(U(x))\|}{\|\nabla U(x)\|\|\log \nabla U(x)\|} = 0$$

and

$$\lim_{\|x\|\to\infty} \frac{\log|\log\|\nabla U(x)\||}{U(x))} = 0$$

which is a slightly weaker assumption than $\lim_{\|x\|\to\infty} \frac{\|\nabla U(x)\|}{U(x)} = 0$ *and* $\lim_{\|x\|\to\infty} \frac{\|Hess(U(x))\|}{\|\nabla U(x)\|} = 0$ *respectively, at least assuming that* $\lim_{\|x\|\to\infty} \|\nabla U(x)\| = +\infty$.

In order to show that Assumption 5.2.7 is very likely to be satisfied in practice we consider the following example which involves targeting t-distributions.

Example 5.2.9. *Suppose that $d = 1$ the density is of the form $\pi(x) = \frac{1}{Z}\frac{1}{1+|x|^k}$ for some $k \in \mathbb{N}$ so that $U(x) = \log(1+|x|^k)$. This is asymptotically the same density as a student distribution with $k-1$ degrees of freedom. Suppose that we use a SUZZ algorithm with speed $s(x) = \max\{|x|^a, 1\}$, $a \in [1, k)$. Then, for $x > 0$*

$$A(x) = s(x)U'(x) - s'(x) = x^{a-1}\frac{(k-a)x^k - a}{1+x^k} = O(x^{a-1})$$

and

$$A'(x) = \frac{[(k-a)(a-1)]x^{a+2k-2} + [(k-a)(k+a-1) - a(a-1) - k]x^{a+k-2} - a(a-1)x^{a-2}}{(1+x^k)^2}.$$

So $A'(x)$ is $O(x^{a-2})$ when $a > 1$ and $O(x^{-k-1})$ when $a = 1$ and assumption 5.2.7 is equivalent to ask that

$$\lim_{|x|\to\infty} \frac{|x|^a A'(x)}{A^2(x)\log(1+|A(x)|)} = 0$$

which is the case.

We then have the following, which is the main result of the section.

Proposition 5.2.10 (Non-Explosion)**.** *Assume that $s \in C^2$ and Assumptions 5.2.2, 5.2.5, and 5.2.7 hold. Then the process is non-explosive.*

The rest of this section is devoted in the proof of Proposition 5.2.10. To do this we use techniques from [MT93b] which depend on the generator of the process.

Although we can define the operator $\mathcal{L}$ in Proposition 5.1.4, we cannot immediately conclude that this is the strong generator of the process. This is because the strong generator is defined usually through uniform convergence and we only defined $\mathcal{L}$ in (5.4) as a point-wise limit. However, the techniques in [MT93b] only require to use the generator of the process restricted in a bounded domain, which we introduce now.

Definition 5.2.11. *Let $m \in \mathbb{N}$ and consider the ball centred around 0, having radius m, O_m. Recall that $\zeta_m = \inf\{t \geq 0 : X_t \notin O_m\}$ is the first exit time of the O_m for the SUZZ process. Starting from $(x, \theta) \in E$ define the stopped m-process as the restriction of the SUZZ on O_m, stopped when exiting O_m, i.e. $(Z_t^m)_{t\geq 0} = (X_t^m, \Theta_t^m)_{t\geq 0} = (X_{t\wedge\zeta_m}, \Theta_{t\wedge\zeta_m})_{t\geq 0}$.*

Let $\Gamma^* = \{(y, \eta) \in E : y \in \partial O_m, \text{ there exists } \epsilon > 0 : \text{ for all } t \in (0, \epsilon], y - t\eta \in O_m\}$. Then Z^m is defined on $E_m = \left(O_m \times \{\pm 1\}^d\right) \cup \Gamma^*$. We define $(P_t^m)_{t\geq 0}$ to be the transition semigroup of Z^m and we allow it to act on a continuous function $f : E \to \mathbb{R}$ by considering the restriction of that function on E_m. Since the switching rate of Z^m is bounded as the process is defined on a bounded set and λ_i are locally bounded, we have that for any $T > 0$, if N_T is the number of switching events before time T, then $\mathbb{E}_{x,\theta}[N_T] < \infty$ for any $(x, \theta) \in E_m$. Therefore Z^m is a PDMP that can be seen in the setting of [Dav84] and we have the following as a result of Theorem 5.5 of [Dav84].

Proposition 5.2.12. *Let $\mathcal{L}$ the operator defined on (5.4). The extended generator $\mathcal{L}^m$ for Z^m has domain $\mathcal{D}(\mathcal{L}^m) \supset C^1(E)$ and for any function $f \in C^1(E)$ we have for all $x \in O_m$.*

$$\mathcal{L}^m f(x, \theta, t) = \mathcal{L} f(x, \theta, t) 1_{x \in O_m}.$$

Consider the function

$$V(x, \theta) = \exp\{aU(x) - a \log s(x) + \sum_{i=1}^{d} \phi(\theta_i A_i(x))\} \tag{5.11}$$

where

$$\phi(s) = \frac{1}{2} sign(s) \log(1 + \log(1 + \delta|s|)), \tag{5.12}$$

for some $a \in (0, 1)$ and $\delta > 0$. The proof of non-explosion relies on the following lemma.

Lemma 5.2.13. *Assume that Assumptions 5.1.5, 5.2.2, 5.2.5 and 5.2.7 hold. Let $\mathcal{L}$ be the operator defined in (5.4). Then, there exist $a \in (0, 1)$ and $\delta > 0$ for which*

V introduced in (5.11) is norm-like function, i.e. $\lim_{\|x\|\to\infty} V(x,\theta) = +\infty$ *and there exists a compact set* C *and* $b, c > 0$ *and for all* $(x,\theta) \in E$

$$\mathcal{L}V(x,\theta) \leq -cV(x,\theta) + b1_{(x,\theta)\in C}. \tag{5.13}$$

Proof of Lemma 5.2.13. One can verify that $V \in C^1$ therefore $V \in D(\mathcal{L})$. Note that

$$V(x, F_i[\theta]) - V(x,\theta) = V(x,\theta)\left(\exp\{\phi(-\theta_i A_i(x)) - \phi(\theta_i A_i(x))\} - 1\right)$$

We then calculate

$$\begin{aligned}\frac{\mathcal{L}V(x,\theta)}{V(x,\theta)} = \sum_{i=1}^{d}\{\theta_i a A_i(x) + \sum_{j=1}^{d}\theta_i\theta_j s(x)\partial_i A_j(x)\phi^{'}(\theta_j A_j(x)) + \\ + [(\theta_i A_i(x))^{+} + \gamma_i(x)](\exp\{\phi(-\theta_i A_i(x)) - \phi(\theta_i A_i(x))\} - 1)\}\end{aligned} \tag{5.14}$$

Note that $\phi(-s) - \phi(s) = (1 + \log(1+\delta|s|))^{-sign(s)}$ and

$$\phi^{'}(s) = \frac{\delta}{2}\frac{1}{(1+\delta|s|)(1+\log(1+\delta|s|))}.$$

Consider the i'th component of the sum in the RHS of (5.14) and the following cases.

Case 1: $\theta_i A_i(x) \geq 0$. Then the i'th component of the sum in the RHS of (5.14) can be written as

$$\begin{aligned}&a|\theta_i A_i(x)| + \sum_{j=1}^{d}\theta_i\theta_j s(x)\partial_i A_j(x)\phi^{'}(\theta_j A_j(x)) + \\ &+ [(\theta_i A_i(x))^{+} + \gamma_i(x)]\left(\frac{1}{1+\log(1+\delta|A_i(x)|)} - 1\right) \leq \\ &\leq |A_i(x)|\left[a - 1 + \frac{1}{1+\log(1+\delta|A_i(x)|)}\right] + \sum_{j=1}^{d}\theta_i\theta_j s(x)\partial_i A_j(x)\phi^{'}(\theta_j A_j(x))\end{aligned} \tag{5.15}$$

where we used that $\gamma_i(x) \geq 0$.

Case 2: $\theta_i A_i(x) < 0$. Then the i'th component of the sum in the RHS of

(5.14) can be written as

$$a\theta_i A_i(x) + \sum_{j=1}^{d} \theta_i\theta_j s(x)\partial_i A_j(x)\phi^{'}(\theta_j A_j(x)) + [(\theta_i A_i(x))^{+} + \gamma_i(x)]\log(1+\delta|A_i(x)|) \leq$$

$$\leq -a|A_i(x)| + \bar{\gamma}\log(1+\delta|A_i(x)|) + \sum_{j=1}^{d} \theta_i\theta_j s(x)\partial_i A_j(x)\phi^{'}(\theta_j A_j(x)) \leq$$

$$\leq |A_i(x)|[-a+\bar{\gamma}\delta] + \sum_{j=1}^{d} \theta_i\theta_j s(x)\partial_i A_j(x)\phi^{'}(\theta_j A_j(x))$$

where we used that $\log(1+x) \leq x$ for $x \geq 0$. Overall, we get,

$$\frac{\mathcal{L}V(x,\theta)}{V(x,\theta)} \leq \sum_{i=1}^{d} |A_i(x)| \max\{a - 1 + \frac{1}{1+\log(1+\delta|A_i(x)|)}, -a + \delta\bar{\gamma}\} +$$
$$+ \sum_{i=1}^{d}\sum_{j=1}^{d} \theta_i\theta_j s(x)\partial_i A_j(x)\phi^{'}(\theta_j A_j(x)). \tag{5.16}$$

Let us set $\epsilon = \bar{\gamma} > 0$, if $\bar{\gamma} > 0$ or $0 < \epsilon < A/(2d^2)$, if $\bar{\gamma} = 0$. Since we have assumed that $A > 3d\bar{\gamma}$ in Assumption 5.2.5, we get that $\frac{A}{d} > \bar{\gamma} + 2\epsilon$. We will choose $\delta > 0$ small enough to be specified later and given δ, we set $a = \delta\bar{\gamma} + \delta\epsilon$. Then the second part of the maximum of (5.16) is equal to $-\delta\epsilon < 0$.

Consider the function

$$f(x) = \max\left\{-\delta\epsilon, \delta\epsilon + \delta\bar{\gamma} - 1 + \frac{1}{1+\log(1+\delta x)}\right\}.$$

Our goal will be to show that $\sum_{i=1}^{d} |A_i(x)| f(|A_i(x)|) < 0$ for $\|x\|$ large enough. One can verify that

$$f(A) < 0 \iff A \leq P(\delta) = \frac{1}{\delta}\left[\exp\left\{\frac{\delta\bar{\gamma} + \delta\epsilon}{1 - \delta\bar{\gamma} - \delta\epsilon}\right\} - 1\right]$$

and

$$f(A) = -\delta\epsilon \iff A \geq M(\delta) = \frac{1}{\delta}\left[\exp\left\{\frac{\delta\bar{\gamma} + 2\delta\epsilon}{1 - \delta\bar{\gamma} - 2\delta\epsilon}\right\} - 1\right].$$

Now,

$$\lim_{\delta\to 0} M(\delta) = \bar{\gamma} + 2\epsilon \ , \ \lim_{\delta\to 0} P(\delta) = \bar{\gamma} + \epsilon$$

so if we choose δ small enough $M(\delta) < A/d$. Suppose $k \in \{1, .., d\}$ is the coordinate

with the maximum value of $|A_i(x)|$, so that

$$|A_k(x)| \geq \frac{\sum_{i=1}^d |A_i(x)|}{d} > \frac{A}{d} > M(\delta).$$

Therefore

$$|A_k(x)|f(|A_k(x)|) \leq -|A_k(x)|\delta\epsilon \leq -\frac{\sum_{i=1}^d |A_i(x)|}{d}\delta\epsilon. \tag{5.17}$$

For any other coordinate i, the contribution to the sum $\sum_{i=1}^d |A_i(x)|f(|A_i(x)|)$ will be positive if and only if $|A_i(x)| \leq P(\delta)$. Then, using that $\dfrac{1}{1+\log(1+\delta|A_i(x)|)} \leq 1$ we can bound

$$f(x) \leq \delta\epsilon + \delta\bar{\gamma}$$

and therefore

$$\sum_{i\neq k}^d |A_i(x)|f(|A_i(x)|) \leq (d-1)P(\delta)(\delta\epsilon + \delta\bar{\gamma}). \tag{5.18}$$

Recall that when $\bar{\gamma} > 0$ we have picked $\epsilon = \bar{\gamma}$, so due to (5.9) we get

$$A > d(d-1)\frac{(\bar{\gamma}+\epsilon)^2}{\epsilon}. \tag{5.19}$$

On the other hand, if $\bar{\gamma} = 0$ we have picked $\epsilon < A/(2d^2)$ so (5.19) holds in this case as well. Therefore

$$\lim_{\delta\to 0} \frac{1}{\sum_{i=1}^d |A_i(x)|}(d-1)(\bar{\gamma}+\epsilon)P(\delta) - \frac{\epsilon}{d} = \frac{1}{\sum_{i=1}^d |A_i(x)|}(d-1)(\bar{\gamma}+\epsilon)^2 - \frac{\epsilon}{d} < 0$$

Combining this with (5.17) and (5.18) we get

$$\begin{aligned}
&\sum_{i=1}^d |A_i(x)|f(|A_i(x)|) = |A_k(x)|f(A_k(x)) + \sum_{i\neq k} |A_i(x)|f(|A_i(x)|) \leq \\
&\leq \delta \sum_{i=1}^d |A_i(x)| \left[-\frac{\epsilon}{d} + \frac{1}{\sum_{i=1}^d |A_i(x)|}(d-1)(\bar{\gamma}+\epsilon)P(\delta) \right] \leq -c\sum_{i=1}^d |A_i(x)|
\end{aligned} \tag{5.20}$$

for some $c > 0$, assuming δ is small enough.

To finish the proof, let us consider the last term of the RHS in (5.16). Here, due to (5.10) and assuming that $x \notin C$ for some compact set large enough, we can write

$$\sum_{j=1}^d \theta_i\theta_j s(x)\partial_i A_j(x)\phi'(\theta_j A_j(x)) \leq \sum_{i,j=1}^d \frac{\delta}{2}\frac{s(x)|\partial_i A_j(x)|}{(1+\delta|A_j(x)|)(1+\log(1+\delta|A_j(x)|))} =$$

$$= \left(\sum_{k=1}^{d} |A_k(x)| \right) \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{s(x)}{\sum_{k=1}^{d} |A_k(x)|} \frac{\delta |\partial_i A_j(x)|}{(1 + \delta |A_j(x)|)(1 + \log(1 + \delta |A_j(x)|))} \leq$$

$$\leq \sum_{i=1}^{d} |A_i(x)| \frac{c}{2}. \tag{5.21}$$

Then, combining (5.16), (5.20) and (5.21) we get for $x \notin C$

$$\frac{\mathcal{L}V(x,\theta)}{V(x,\theta)} \leq -\frac{c}{2} \sum_{i=1}^{d} |A_i(x)| \leq -\frac{c}{2} A$$

and this proves (5.13) since V and $\mathcal{L}V$ are bounded on C.

Finally, we need to prove that $\lim_{\|x\| \to \infty} V(x,\theta) = +\infty$. We can assume without loss of generality that for all i, $\lim_{\|x\| \to \infty} A_i(x) = \infty$ else the result holds from Assumption 5.2.2. Due to the last two equations of Assumption 5.2.7 we can write for some constants $C, C' > 0$ and $\|x\|$ large enough,

$$\begin{aligned} V(x,\theta) &\geq C \exp\{aU(x) - a \log s(x) - \sum_{i=1}^{d} \frac{1}{2} \log(|A_i(x)|)\} \geq \\ &\geq C' \exp\{aU(x) - a \log s(x)\} \prod_{i=1}^{d} (\log |A_i(x)|)^{-1/2} \geq \\ &\geq C' \exp\{aU(x) - a \log s(x)\} (\log s(x) + \log |\nabla (U(x) - \log s(x))|)^{-d} = +\infty. \end{aligned}$$

This completes the proof. □

Proof of Propositions 5.2.10. Under the assumption of Proposition 5.2.10, Lemma 5.2.13 along with Proposition 5.2.12 prove that there exists a norm-like function V and constants $c, b > 0$ such that $\mathcal{L}^m V(x,\theta) \leq cV(x,\theta) + b$ for all $m \in \mathbb{N}$. The assumption of Theorem 2.1 in [MT93b] is satisfied and this proves that the process is non-explosive. □

Note that in Lemma 5.2.13 we prove the drift condition $\mathcal{L}^m V(x,\theta) \leq -cV(x,\theta) + b$ for some positive quantity c. This is an even stronger drift condition as compared to the one required by Theorem 2.1 in [MT93b] to prove non-explosivity. This drift equation could allow us to conclude existence and uniqueness of an invariant measure and Geometric convergence towards it. To conclude with that, however, we still need to prove that every compact subset of E is small for some skeleton of the process. We postpone this until Section 5.4. Before that, in the next section, we give a formal proof that when we pick the switching rates according to (5.6), the

process targets the right distribution.

5.3 Invariant Measure of Speed Up Zig-Zag

In this section we will show that if we pick the switching rates according to (5.6) then our non-explosive process leaves the target distribution of interest invariant. For this we need to make the following assumption in the case where the deterministic dynamics of the process are explosive.

Assumption 5.3.1.

$$\lim_{\|x\|\to\infty} \|x\|^{d-1} s(x) \exp\{-U(x)\} = 0. \tag{5.22}$$

Note that in the original Zig-Zag for $s \equiv 1$, (5.22) is assumed to hold as it is assumed that the potential U satisfies the growth condition $U(x) \geq (d+\epsilon)\log(\|x\|) - c'$ for some $\epsilon > 0, c \in \mathbb{R}$. This is used to prove non-evanescence.

The main result of this section is the following.

Proposition 5.3.2 (Invariant Measure)**.** *Assume that Assumptions 5.1.5, 5.2.5 , 5.2.7 and 5.3.1 hold. Then, the SUZZ process has the measure μ in (5.5) as invariant.*

The rest of the section is devoted to the proof of Proposition 5.3.2.

Lemma 5.3.3. *Assume that Assumptions 5.1.5, 5.2.5 , 5.2.7 and 5.3.1 hold and the $(Z_t)_{t\geq 0}$ is a SUZZ process with speed function s. If $f \in C_c^1(E)$ then f is in the domain of the strong generator of Z (see equation (2.11)) and that generator is equal to $\mathcal{L}f$, where $\mathcal{L}$ is the operator defined in (5.4).*

Proof of Lemma 5.3.3. Let K be a compact set that contains the support of f and let $K' = \{(x,\theta) : \text{ there exists a } y, \text{ with } \|x-y\| < \epsilon, (y,\theta) \in K\} = K + \epsilon B(0,1)$ for some $\epsilon > 0$. Let $\bar{s}$ be an upper bound on the speed s on K'. Then for all $t < t_1 = \epsilon/(\sqrt{d}\bar{s})$, if the process starts from any $(x,\theta) \notin K'$, then the process will not have hit K until time t, and since the support of f is contained in K, $P^t f(x,\theta) = \mathbb{E}_{x,\theta}[f(X_t)] = 0$, for all $(x,\theta) \notin K'$. Note as well that for all $(x,\theta) \notin K'$, $\mathcal{L}f(x,\theta) = 0$.
Now, let us focus on $(x,\theta) \in K'$. Pick $K'' = K' + \epsilon B(0,1)$ and let $\bar{\bar{s}}$ be an upper bound of s on K''. Then for all $t \leq t_2 = \epsilon/(\sqrt{d}\bar{\bar{s}}) < t_1$, the process starting from K' will not have exited K'' by time t and if we cover K'' by some O_m for some large

m, then a.s. $Z(t) = Z^m(t)$ for all $t \leq t_2$ as long as we start from somewhere in K'. Then, for any $(x, \theta) \in K'$ and any $t < t_2$

$$\frac{\mathbb{E}_{x,\theta}[f(Z_t)] - f(x,\theta)}{t} - \mathcal{L}f(x,\theta) = \frac{\mathbb{E}_{x,\theta}[f(Z_t^m)] - f(x,\theta)}{t} - \mathcal{L}^m f(x,\theta)$$

so overall for all $t < t_2$

$$\begin{aligned}
&\sup_{(x,\theta)\in E} \left| \frac{\mathbb{E}_{x,\theta}[f(Z_t)] - f(x,\theta)}{t} - \mathcal{L}f(x,\theta) \right| \leq \\
&\leq \sup_{(x,\theta)\in K''} \left| \frac{\mathbb{E}_{x,\theta}[f(Z_t^m)] - f(x,\theta)}{t} - \mathcal{L}^m f(x,\theta) \right| \xrightarrow{t\to 0} 0
\end{aligned}$$

which proves the result. □

The following Lemma is the key to prove Proposition 5.3.2.

Lemma 5.3.4. *Assume that Assumptions 5.1.5, 5.2.5 , 5.2.7 and 5.3.1 hold and the $(Z_t)_{t\geq 0}$ is a SUZZ process with speed function s. Let $(P^t)_{t\geq 0}$ be the transition semi-group of a SUZZ process with speed function s and let $\mathcal{L}$ be the operator defined in (5.4). If $f \in C_c^1(E)$ then for all $s > 0$,*

$$\int_E \mathcal{L}P^s f(x,\theta)\mu(dx, d\theta) = 0. \tag{5.23}$$

Note that if $f \in C_c^1(E)$, we have already proven in Lemma 5.1.6 that $\int_E \mathcal{L}f(x,\theta)\mu(dx,d\theta) = 0$. If we could guarantee that for any $f \in C_c^1(E)$, for all $s > 0$, $P^s f \in C_c^1(E)$ the result would be evident. However, since we allow explosive deterministic dynamics, the process can come down from infinity in finite time. Therefore, for any $s > 0$, the process may have hit the compact support of f in time less than s no matter how far away we start the process from and this means that $P^s f(x,\theta) = \mathbb{E}_{x,\theta}[f(Z_s)]$ is not necessarily compactly supported. Furthermore, we cannot even guarantee that $P^s f \in C^1$. In [DGM18], the authors manage to prove differentiability for various PDMPs, but the assumptions they are making include non-explosive deterministic dynamics. This creates a problem since the technique used so far in the literature to prove that $\int_E \mathcal{L}P^s f(x,\theta)\mu(dx,d\theta) = 0$ is an integration by parts on $P^s f$, which cannot be used if we cannot guarantee the differentiability of $P^s f$

Heuristically, the way to go around this problem is the following. Integrating $\mathcal{L}P^s f$ over $\mathbb{R}^d$ can be well approximated by integrating over a large ball of radium m, O_m. We can prove that $P^s f$ is in the domain of the extended generator of the stopped process Z^m. Furthermore, the stopped process is in Davis PDMP

setting [Dav84] and a complete characterisation of the functions in the domain of the extended generator was given in that setting (see Theorem 2.3.3). Corollary 2.3.4 guarantees that the function $P^s f$ must have a weak derivative along lines parallel to the vectors $\{-1,+1\}^d$. Therefore, the fundamental theorem of calculus and an integration by parts can be used along such lines. When we will integrate $\mathcal{L}P^s f$ over the ball O_m, we may do the integration over many different lines parallel to some vector $\{-1,+1\}^d$ and apply the integration by parts technique in each of these lines to get the result. This is the main idea of the following proof.

Proof of Lemma 5.3.4. We begin by noticing that if $E_m = O_m \times \{\pm 1\}^d$

$$\begin{aligned}
&\left| \int_{E_m} \mathcal{L}^m P^s f(x,\theta)\mu(dx,d\theta) - \int_E \mathcal{L}P^s f(x,\theta)\mu(dx,d\theta) \right| \leq \\
&\leq \left| \int_{E_m} \mathcal{L}^m P^s f(x,\theta)\mu(dx,d\theta) - \int_{E_m} \mathcal{L}P^s f(x,\theta)\mu(dx,d\theta) \right| + \\
&+ \left| \int_{E_m} \mathcal{L}P^s f(x,\theta)\mu(dx,d\theta) - \int_E \mathcal{L}P^s f(x,\theta)\mu(dx,d\theta) \right| \xrightarrow{m\to\infty} 0
\end{aligned}$$

since $\mathcal{L}^m P^s f(x,\theta) = \mathcal{L}P^s f(x,\theta)$ for all $(x,\theta) \in E_m$ so the first part is zero and the second part converges to zero since $E_m \nearrow E$ and $\mathcal{L}P^s f = P^s \mathcal{L} f$ is bounded. Therefore, it suffices to prove that

$$\lim_{m\to\infty} \int_{E_m} \mathcal{L}^m P^s f(x,\theta)\mu(dx,d\theta) = 0. \tag{5.24}$$

Note that we know that f is in the domain of the strong generator of Z, therefore, using standard results (for example [EK86]), it is also in the domain of the extended generator of Z. Therefore, the process is also in the domain $\mathcal{D}(\mathcal{L}^m)$ of the extended generator of the stopped Z^m. Since $\mathcal{D}(\mathcal{L}^m)$ is the domain of a PDMP in Davis setting, due to Theorem 2.3.3 and Corollary 2.3.4, we have the following. Let us write $g = P^s f$ to ease the notation. Then, there exists a function $\nabla g : E_m \to \mathbb{R}$ such that if X_t satisfies ODE (5.3) with starting point (x,θ) then for all $t \geq 0$

$$g(X_t,\theta) - g(x,\theta) = \int_0^t \nabla g(X_s,\theta) ds$$

and for all $(x,\theta) \in E_m$,

$$\mathcal{L}^m g(x,\theta) = \nabla g(x,\theta) + \sum_{i=1}^d \lambda_i(x,\theta) \left(g(x, F_i[\theta]) - g(x,\theta) \right)$$

Our goal is to use an integration by parts technique to control the first part of the sum of the generator. We fix a $\theta \in \{-1,+1\}^d$. Use a linear, invertible transformation A on $\mathbb{R}^d$ such that $A\theta = \sqrt{d}e_1$, $AO_m = O_m$ and $detA = 1$. We use the transformation $y = (y_1, ..., y_d) = Ax$. Also, given $y_2, y_3, ..., y_d$ with $y_2^2 + ... + y_d^2 < m^2$ we write

$$y_1^* = \sqrt{m^2 - y_2^2 - ... - y_d^2} \text{ if } \theta_1 = 1$$

or

$$y_1^* = -\sqrt{m^2 - y_2^2 - ... - y_d^2} \text{ if } \theta_1 = -1$$

and we omit the dependence on $y_2, ..., y_d$ for ease of notation. We write $x_0 = A^{-1}(-y_1^*, y_2, ..., y_d)'$.

We further consider the solution X_t to the ODE (5.3) starting from (x_0, θ) and we write $Y_t = (Y_t^1, ..., Y_t^d) = AX_t$ so that Y_t starts from $(-y_1^*, y_2, ..., y_d)$ and solves the ODE $dY_t/dt = \sqrt{d}s(A^{-1}Y_t)e_1$, where we denote $e_1 = (1, 0, ..., 0) \in \mathbb{R}^d$. Also write t^* such that $Y_{t^*}^1 = y_1^*$. Then we can write,

$$\int_{O_m} \nabla g(x,\theta)\exp\{-U(x)\}dx = \int_{O_m} \nabla g(A^{-1}y,\theta)\exp\{-U(A^{-1}y)\}dy =$$
$$= \int_{-m}^{m}\int_{-\sqrt{m^2-y_2^2}}^{\sqrt{m^2-y_2^2}} ... \int_{-y_1^*}^{y_1^*} \nabla g(A^{-1}y,\theta)\exp\{-U(A^{-1}y)\}dy_1 dy_d ... dy_2 =$$
$$= \int_{-m}^{m}\int_{-\sqrt{m^2-y_1^2}}^{\sqrt{m^2-y_1^2}} ... \int_{0}^{t^*} \nabla g(A^{-1}Y_t,\theta)\exp\{-U(A^{-1}Y_t)\}s(A^{-1}Y_t)\sqrt{d}dt dy_d ... dy_2$$

Having fixed $y_2, ..., y_d$, write $z_1 = (-y_1^*, y_2, ..., y_d), z_2 = (y_1^*, y_2, ..., y_d) \in \partial O_m$ and $y = (y_1, ..., y_d)$ and using integration by parts we get

$$\int_0^{t^*} \nabla g(A^{-1}Y_t,\theta)\exp\{-U(A^{-1}Y_t)\}s(A^{-1}Y_t)\sqrt{d}\ dt =$$
$$= g(A^{-1}z_2,\theta)\exp\{-U(A^{-1}z_2)\}s(A^{-1}z_2)\sqrt{d} - g(A^{-1}z_1,\theta)\exp\{-U(A^{-1}z_1)\}s(A^{-1}z_1)\sqrt{d} -$$
$$- \int_0^{t^*} g(A^{-1}Y_t,\theta)\frac{d}{dt}\left[\exp\{-U(A^{-1}Y_t)\}s(A^{-1}Y_t)\right]\sqrt{d}\ dt =$$
$$= g(A^{-1}z_2,\theta)\exp\{-U(A^{-1}z_2)\}s(A^{-1}z_2)\sqrt{d} - g(A^{-1}z_1,\theta)\exp\{-U(A^{-1}z_1)\}s(A^{-1}z_1)\sqrt{d} -$$
$$- \int_0^{t^*} g(A^{-1}Y_t,\theta)\exp\{-U(A^{-1}Y_t)\}\sum_{i=1}^{d}\{-\partial_i U(A^{-1}Y_t)\theta_i s(A^{-1}Y_t) +$$
$$+\partial_i s(A^{-1}Y_t)\theta_i\}\ s(A^{-1}Y_t)\sqrt{d}\ dt =$$
$$= g(A^{-1}z_2,\theta)\exp\{-U(A^{-1}z_2)\}s(A^{-1}z_2)\sqrt{d} - g(A^{-1}z_1,\theta)\exp\{-U(A^{-1}z_1)\}s(A^{-1}z_1)\sqrt{d} -$$

$$- \int_{-y_1^*}^{y_1^*} g(A^{-1}y, \theta) \exp\{-U(A^{-1}y)\} \sum_{i=1}^{d} \left(-\partial_i U(A^{-1}y)\theta_i s(A^{-1}y) + \partial_i s(A^{-1}y)\theta_i\right) dy_1.$$

Overall

$$\sum_{\theta\in\{\pm 1\}^d} \int_{O_m} \nabla g(x,\theta) \exp\{-U(x)\} dx =$$

$$\sum_{\theta\in\{\pm 1\}^d} \int_{-m}^{m} \int_{-\sqrt{m^2-y_2^2}}^{\sqrt{m^2-y_2^2}} \dots \int_{-\sqrt{m^2-y_2^2-\dots-y_{d-1}^2}}^{\sqrt{m^2-y_2^2-\dots-y_{d-1}^2}} g(A^{-1}z_2, \theta) \exp\{-U(A^{-1}z_2)\} \times$$

$$s(A^{-1}z_2)\sqrt{d}\; dy_d \dots dy_3 dy_2 -$$

$$- \sum_{\theta\in\{\pm 1\}^d} \int_{-m}^{m} \int_{-\sqrt{m^2-y_2^2}}^{\sqrt{m^2-y_2^2}} \dots \int_{-\sqrt{m^2-y_2^2-\dots-y_{d-1}^2}}^{\sqrt{m^2-y_2^2-\dots-y_{d-1}^2}} g(A^{-1}z_1, \theta) \exp\{-U(A^{-1}z_1)\} \times$$

$$s(A^{-1}z_1)\sqrt{d} dy_d \dots dy_3 dy_2 +$$

$$+ \sum_{\theta\in\{\pm 1\}^d} \int_{O_m} g(x,\theta) \exp\{-U(x)\} \sum_{i=1}^{d} \left(\theta_i \partial_i U(x) s(x) - \theta_i \partial_i s(x) s(x)\right) dx. \tag{5.25}$$

On the other hand, following a rearrangement of the sum over θ, as in the proof of Lemma 5.1.6

$$\sum_{i=1}^{d} \sum_{\theta\in\{\pm 1\}^d} \int_{O_m} \lambda_i(x,\theta) \left(g(x, F_i[\theta]) - g(x,\theta)\right) \exp\{-U(x)\} dx =$$

$$- \sum_{i=1}^{d} \sum_{\theta\in\{\pm 1\}^d} \int_{O_m} g(x,\theta) \exp\{-U(x)\} \left(\lambda_i(x,\theta) - \lambda_i(x, F_i[\theta])\right) dx \tag{5.26}$$

Recall that from Assumption 5.1.5 we have

$$\lambda_i(x,\theta) - \lambda_i(x, F_i[\theta]) = \theta_i \partial_i U(x) s(x) - \theta_i \partial_i s(x) s(x).$$

When we integrate $\int_{E_m} \mathcal{L} g d\mu$, we get the sum of the RHS of equations (5.25) and (5.26). On this sum, only the boundary parts remain and we have

$$\left| \int_{E_m} \mathcal{L} g(x,\theta) \mu(dx, d\theta) \right| = \left| \sum_{\theta\in\{\pm 1\}^d} \int_{O_m} \mathcal{L} g(x,\theta) dx \right| \leq$$

$$\leq | \sum_{\theta\in\{\pm 1\}^d} \int_{-m}^{m} \int_{-\sqrt{m^2-y_2^2}}^{\sqrt{m^2-y_2^2}} \dots \int_{-\sqrt{m^2-y_2^2-\dots-y_d^2}}^{\sqrt{m^2-y_2^2-\dots-y_d^2}} g(A^{-1}z_2, \theta) \exp\{-U(A^{-1}z_2)\} \times$$

$$
\begin{aligned}
&s(A^{-1}z_2)\sqrt{d}\ dy_d...dy_3dy_2|+\\
&+|\sum_{\theta\in\{\pm1\}^d}\int_{-m}^{m}\int_{-\sqrt{m^2-y_2^2}}^{\sqrt{m^2-y_2^2}}...\int_{-\sqrt{m^2-y_2^2-...-y_d^2}}^{\sqrt{m^2-y_2^2-...-y_d^2}} g(A^{-1}z_1,\theta)\exp\{-U(A^{-1}z_1)\}\times\\
&s(A^{-1}z_1)\sqrt{d}\ dy_d...dy_3dy_2|\leq\\
&\leq 2\sqrt{d}\|g\|_\infty\sup_{x\in\partial_{O_m}}\{\exp\{-U(x)\}s(x)\}\int_{x\in\partial O_m}1dx\leq\\
&\leq C2\sqrt{d}\|g\|_\infty\sup_{x\in\partial_{O_m}}\{\exp\{-U(x)\}s(x)\}m^{d-1}\xrightarrow{m\to\infty}0
\end{aligned}
$$

where the converge holds due to Assumption 5.3.1. Here $\|g\|_\infty$ is well defined since $g = P^s f$ is bounded since f is bounded. This completes the proof. □

Remark 5.3.5. *Note that we only used Assumption 5.22 in order to ensure that the boundary terms appearing in the integration by parts will decay as $\|x\|$ goes to infinity. If the deterministic dynamics are non-explosive, the path of the process until time s has a bounded length, therefore the function $g = P^s f$ has compact support and all the boundary terms disappear as $\|x\| \to \infty$. This means that when the deterministic dynamics are non-explosive we do not need to make Assumption 5.3.1, as long as we still impose Assumption 5.2.2.*

Now, we can conclude with the proof of invariance.

Proof of Proposition 5.3.2. Let $(P^t)_{t\geq 0}$ be the transition semi-group of the process and $\mathcal{L}$ the operator defined in (5.4). Let $f \in C_c^1(E)$. From Lemma 5.3.3, $\mathcal{L}f$ is the strong generator of f. Because of Dynkin's formula,

$$P^t f(x,\theta) - f(x,\theta) = \int_0^t \mathcal{L}P^s f(x,\theta)ds$$

Since s and λ are bounded on compact sets, for any $f \in C_c^\infty(E)$ we have that $\mathcal{L}f$ is bounded and after integrating both sides over μ and using Fubini's theorem, we get

$$
\begin{aligned}
&\int_E P^t f(x,\theta)\mu(dx,d\theta) - \int_E f(x,\theta)\mu(dx,d\theta) = \int_E\int_0^t \mathcal{L}P^s f(x,\theta)ds\mu(dx,d\theta) =\\
&= \int_0^t\int_E \mathcal{L}P^s f(x,\theta)\mu(dx,d\theta)ds = 0
\end{aligned}
$$

where the last equality follows from Lemma 5.3.4. Therefore, for all $f \in C_c^\infty$

$$\int P^t f(x,\theta)\mu(dx,d\theta) = \int f(x,\theta)\mu(dx,d\theta). \tag{5.27}$$

Since, as a simple application of Stone-Weierstrass, C_c^∞ is dense in C_c, we get that (5.27) holds for all $f \in C_c$ and this further extends to all bounded measurable functions f from Lusin's theorem [Fol13]. This proves the result. □

5.4 Geometric Ergodicity of Speed Up Zig-Zag

In this section we prove that the SUZZ process can be geometrically ergodic, even in heavy tailed targets, under some assumptions on the speed function s. This can further provide a Central Limit Theorem for the process.

The main technical work of this section is to follow the techniques of [BRZ19] to prove that all compact sets are petite for the SUZZ process. Combining this with (5.13) we have the following.

Theorem 5.4.1 (Geometric Ergodicity)**.** *Let $(Z_t)_{t\geq 0} = (X_t, \Theta_t)_{t\geq 0}$ a SUZZ process with speed function s. Suppose that assumptions 5.1.5, 5.2.5, 5.2.7 and 5.3.1 hold. Assume further that the function $U - \log s \in C^3$ and has a non-degenerate local minimum, i.e. there exists an $x_0 \in \mathbb{R}^d$ local minimum for $U - \log s$ such that the Hessian matrix $Hess(U - \log s)(x_0)$ is strictly positive definite. Finally, assume that μ introduced in (5.5) is a probability measure. Then the Speed Up Zig-Zag process is geometrically ergodic, i.e. there exists some $M > 0, \rho < 1$ such that for V introduced in (5.11) and any $(x, \theta) \in E$,*

$$\|\mathbb{P}_{x,\theta}(Z_t \in \cdot) - \mu(\cdot)\|_{TV} \leq MV(x, \theta)\rho^t. \tag{5.28}$$

An immediate result due to Theorem 2 of [CG94] (see also Theorem 1.2 of [H05]) is the following CLT.

Theorem 5.4.2. *Suppose that all the assumptions of Theorem 5.4.1 hold. Let $\{Y_n, n \geq 0\}$ any skeleton of the SUZZ process (i.e. $Y_n = Z_{n\delta}$ for some $\delta > 0$) and let $f : E \to \mathbb{R}$ such that there exists an $\epsilon > 0$ with $\mathbb{E}_\mu[f^{2+\epsilon}] < \infty$, then there exists a $\gamma_f^2 \in [0, \infty)$ such that*

$$\frac{\sum_{k=1}^n f(Y_k) - \mu(f)}{\sqrt{n}} \xrightarrow[D]{n\to\infty} Z \tag{5.29}$$

for some $Z \sim \mathcal{N}(0, \gamma_f^2)$.

Furthermore, due to Theorem 2.1.11 we can have the following result for the original Zig-Zag. This is a slight improvement of Theorem 2.4.9, since we do not ask the existence of $\eta \in (0, 1)$ such that $\int_{\mathbb{R}^d} \exp\{-\eta U(x)\}dx < \infty$.

Corollary 5.4.3. *Consider an original Zig-Zag process (Z_t) for which $U \in C^3$, $\lim_{\|x\|\to\infty} U(x) = +\infty$ and U has a non-degenerate local minimum. Assume further that $\lim_{\|x\|\to\infty} \|\nabla U(x)\| = +\infty$, $\lim_{\|x\|\to\infty} \frac{\|\nabla U(x)\|}{U(x)} = 0$ and $\lim_{\|x\|\to\infty} \frac{\|Hess(U)(x)\|}{\|\nabla U(x)\|} = 0$.*

Let $g : E \to \mathbb{R}$ such that there exists $b \in (0, 1/2)$ and a compact set C such that for all $x \notin C$ and $\theta \in \{-1, 1\}^d$, $g(x, \theta) \leq \exp\{bU(x)\}$. Define for any $n \in \mathbb{N}$ the function $Z_n : [0, 1] \to \mathbb{R}$ by

$$Z_n(t) = \frac{1}{\sqrt{n}} \int_0^{nt} g(X_s, \Theta_s) - \mu(g)ds.$$

and let $B = (B_t)_{t\in[0,1]}$ a standard Brownian motion on $[0, 1]$. Then there exists $\gamma_g \in [0, +\infty)$ such that the sequence Z_n converges in distribution in Skorokhod topology (see [Bil99]) in $\mathcal{D}[0, 1]$ to $\gamma_g B$.

The rest of this section is devoted to the proof of Theorems 5.4.1 and Corollary 5.4.3. For the proof of Theorem 5.4.1 we will use the following result (Theorem 6.1 in [MT93b]).

Theorem 5.4.4 (Meyn-Tweedie 1993)**.** *Assume that a Markov process $(Z_t)_{t\geq 0}$ on E is cad-lag, and all compact subset of E are petite for some skeleton chain of Z. Assume further that if $\mathcal{L}^m$ is the extended generator of the process $(Z_t^m)_{t\geq 0}$, which is the process Z, stopped upon exiting O_m, then there exists a function $V : E \to [1, +\infty)$ and $c, b > 0$ and a compact set C such that for all $m \in \mathbb{N}$ and $z \in E$,*

$$\mathcal{L}^m V(z) \leq -cV(z) + b1_C(z).$$

Then the process Z is Geometrically ergodic, i.e. there exists a constant M and $\rho < 1$ such that for all $z \in E$

$$\|\mathbb{P}_z (Z_t \in \cdot) - \pi(\cdot)\|_{TV} \leq MV(z)\rho^t.$$

From the proof of non-explosivity we have that the function V introduced in (5.11) satisfies the drift condition (5.13) for some compact set C, which is the same as the drift condition assumption of Theorem 5.4.4. Therefore, in order to prove Theorem 5.4.1 we need to prove that the speed up process has all the compact sets as petite for some skeleton chain. The focus of this section is to find conditions that guarantee this property for the SUZZ.

In order to do this we need to establish the reachability property, similar to Chapter 3. We recall here some definitions, very similar to the ones in Chapter 3.

Given a speed function s, generating the family of deterministic flows $\{\Phi_t(x,\theta), t \geq 0\}$ for every $(x,\theta) \in E$, we define as control sequence an object $u = (t, \iota)$, where $t = (t_0, ..., t_m) \in (0, +\infty)^{m+1}$, $\iota = (i_1, ..., i_m) \in \{1, ..., n\}^m$ for some $m \in \mathbb{N}$. Starting from $(x,\theta) \in E$, a control sequence u gives rise to a SUZZ trajectory (X_t, Θ_t) as follows: Start from (x,θ) and follow direction θ for t_0 time, i.e. set $X_t = \Phi_{(x,\theta)}(t), \Theta_t = \theta$ for $t \in [0, t_0)$. Then, switch the i_1'th component of θ to $F_{i_1}[\theta]$ and follow that direction for t_1 time, i.e. set $X_t = \Phi_{(\Phi_{(x,\theta)}(t_0), F_{i_1}[\theta])}(t - t_0), \Theta_t = F_{i_1}[\theta]$ for $t \in [t_0, t_0 + t_1)$. Continue similarly until time $t_0 + ... + t_m$. Write $\tau_k = \sum_{i=0}^{k-1} t_i$ for the time of the k'th switch and denote the final position $(X_{\tau_{m+1}}, \Theta_{\tau_{m+1}})$ of the path by $\Psi_u(x,\theta)$.

Definition 5.4.5. *Given a starting point $(x,\theta) \in E$, a control sequence $u = (t, \iota)$ is admissible if for all $k \in \{1, ..., m\}$ we have $\lambda_{i_k}(X_{\tau_k}, \Theta_{\tau_k}) > 0$.*
Given two points $(x,\theta), (y,\eta) \in E$ we say that (y,η) is reachable from (x,θ) and write $(x,\theta) \to (y,\eta)$ if there exists a control sequence u admissible from (x,θ) such that $\Psi_u(x,\theta) = (y,\eta)$.
We write $(x,\theta) \not\hookrightarrow (y,\eta)$ if $(x,\theta) \to (y,\eta)$ and for the admissible sequence $u = (t, \iota)$ connecting the two points, we have that every index of $\{1, ..., d\}$ appears in ι.

We now focus on proving that for any two points $(x,\theta), (y,\eta) \in E$ we have $(x,\theta) \to (y,\eta)$. Note that we can assume that the SUZZ has minimal rates (i.e. $\gamma_i \equiv 0$ for all $i \in \{1, ..., d\}$) as higher rates make admissible paths more likely. In the case of normal Zig-Zag there is the following result (Theorem 4 [BRZ19]).

Theorem 5.4.6 (Bierkens-Roberts-Zitt 2019)**.** *Assume that $U \in C^3$, $\lim_{\|x\| \to \infty} U(x) = +\infty$ and there exists an x_0 local minimum for U such that $Hess(U)(x_0)$ is strictly positive definite. Then the normal Zig-Zag process targeting the potential U satisfies that for all $(x,\theta), (y,\eta) \in E$, $(x,\theta) \not\hookrightarrow (y,\eta)$.*

We can then generalise these results on the SUZZ using the following Lemma.

Lemma 5.4.7. *Suppose $s \in C^2$ and $s(x) > 0$ for $x \in \mathbb{R}^d$. Then, for any $(x,\theta), (y,\eta) \in E$, $(x,\theta) \to (y,\eta)$ in a SUZZ with speed s, targeting a potential U with minimal rates if and only if $(x,\theta) \to (y,\eta)$ in a normal Zig-Zag, targeting a potential $U - \log s$ with minimal rates.*

Proof of Lemma 5.4.7. Consider a normal Zig-Zag process targeting the potential $U - \log s$ with minimal rates. The rates of this process for the i coordinate are $\lambda_i^0(x', \theta') = [\theta_i' \partial_i (U(x') - \log s(x'))]^+$.
On the other hand, a SUZZ process with minimal rates targeting the potential U

has rates for the i coordinate given by

$$\lambda_i(x', \theta') = [\theta_i'(s(x')\partial_i U(x') - \partial_i s(x'))]^+ = s(x)\lambda_i^0(x', \theta').$$

Therefore for any $(x', \theta') \in E$ and any $i \in \{1, ..., d\}$

$$\lambda_i(x', \theta') > 0 \iff \lambda_i^0(x', \theta') > 0. \tag{5.30}$$

Assume $(x, \theta) \to (y, \eta)$ with some admissible control sequence
$u = (t, \iota) = (t_0, ..., t_m, i_1, ..., i_m)$ for the normal Zig-Zag process, targeting the potential $U - \log s$ with minimal rates. Let (X_t, Θ_t) be the configuration of that Zig-Zag path and let $\tau_k = \sum_{i=0}^{k-1} t_i$ be the times of the switches. We have $\lambda_{i_k}^0(X_{\tau_k}, \Theta_{\tau_k}) > 0$ for all k.
Note that since s is continuous and strictly positive, for any $(x', \theta') \in E$, $\lim_{t\to+\infty} \|\Phi_{(x',\theta')}(t)\| = +\infty$. Therefore, there exists an $s_0 > 0$ such that $\Phi_{(x,\theta)}(s_0) = x + t_0\theta = X_{\tau_1}$. Likewise, there exists an $s_1 > 0$ such that $\Phi_{(X_{\tau_1},\Theta_{\tau_1})}(s_1) = X_{\tau_2}$ and via induction we can construct for all $k \in \{0, ..., m\}$ an s_k such that $\Phi_{(X_{\tau_k},\Theta_{\tau_k})}(s_k) = X_{\tau_{k+1}}$. Then the control sequence $\tilde{u} = (s, \iota) = (s_0, ..., s_m, i_1, ..., i_m)$ is an admissible sequence starting from (x, θ) for the SUZZ targeting the potential U with minimal rates. Furthermore, the ending point of $\tilde{u}$ starting from (x, θ) is (y, η).

The other way around, i.e. that an admissible path for the SUZZ process targeting U implies existence of an admissible path for the ZZ process targeting $U - \log s$ follows using similar arguments. □

Combining Theorem 5.4.6 and Lemma 5.4.8 we can prove the following.

Proposition 5.4.8. *Assume that $s \in C^2$ is a strictly positive function such that the Speed Growth Assumption 5.2.2 holds (i.e. $\lim_{\|x\|\to\infty} U(x) - \log s(x) = +\infty$), $U - \log s \in C^3$ and there exists an $x_0 \in \mathbb{R}^d$ such that $U - \log s$ has a local minimum in x_0 with $Hess(U - \log s)(x_0)$ being strictly positive definite. Then, for every $(x, \theta), (y, \eta) \in E$, $(x, \theta) \leftsquigarrow (y, \eta)$.*

As in [BRZ19] we can use this to prove that from any starting point and given any other point in $z \in E$, the process has a positive probability of visiting a neighbourhood of z. The following lemma is the same as Lemma 8 in [BRZ19].

Lemma 5.4.9 (Continuous Component)**.** *If $(x, \theta) \leftsquigarrow (y, \eta)$ and the rates λ_i are continuous. Then there exist $U_x, V_x \subset \mathbb{R}^d$ open with $x \in U_x, y \in V_x$ and $\epsilon, t_0, c > 0$ such that for all $x' \in U_x, t \in [t_0, t_0 + \epsilon]$*

$$\mathbb{P}_{x',\theta}(X_t \in \cdot, \Theta_t = \eta) \geq cLeb(\cdot \cap V_x) \tag{5.31}$$

where Leb is the Lebesgue measure on $\mathbb{R}^d$.

The proof is very similar in spirit to the proof of Lemma 8 (Continuous Component) of [BRZ19] and to the proof of Lemma 3.3.21 of this work. The main idea is that since there is an admissible path from $(x, \theta) \rightsquigarrow (y, \eta)$, the process has a positive probability to follow some path very close to the admissible path and therefore starting from somewhere close to (x, θ) to end up somewhere close to (y, η). The difference is that this time we allow explosive deterministic dynamics. Therefore, the process is not guaranteed to stay inside a fixed ball until time $t_0 + \epsilon$, as in the original Zig-Zag. This means that the Poisson thinning construction that the authors propose to get the result cannot be applied here directly. However, in order to get the result, we only need to consider paths that are close to the admissible path. These paths need to lie on a fixed ball where the hazard rates are bounded and therefore we can use Poisson thinning to construct this type of paths. This gives us the result as shown in the next proof.

Proof of Lemma 5.4.9. Since $(x, \theta) \rightsquigarrow (y, \eta)$ there exists an admissible control sequence $u = (t, \iota) = (t_0, ..., t_m, i_1, ..., i_m)$ starting from (x, θ) with $\Psi_u(x, \theta) = (y, \eta)$. Let U' a small neighbourhood of x and $B(0, R)$ a ball large enough to contain the paths induced by u starting from (x', θ) for every $x' \in U'$. Also, pick ϵ small enough to ensure that for any $x' \in U'$, if the process starts from the end of any path induced by u and (x', θ), i.e. from $\Psi_u(x', \theta) = (y, \eta)$, then the process a.s. stays inside the ball $B(0, R)$ until time ϵ. Such an ϵ can be picked due to Lemma 5.1.3. Since the rates of the process are continuous functions, they are bounded above in $B(0, R)$ say by $\bar{\lambda}$. Let $\tau_k = \sum_{i=0}^{k-1} t_i$ and let $\underline{\lambda} > 0$ such that

$$\lambda_{min}(t, \iota) = \min_{k \in \{0,...,m-1\}} \lambda_{i_k}(X_{\tau_k}, \Theta_{\tau_k}) > \underline{\lambda},$$

i.e. $\underline{\lambda}$ be a lower bound on the rates at the switching points of the path. Then from continuity of the rates (and possibly by making U' smaller) we can find $(U_k)_{k=0}^{m-1}$ non-intersecting small neighbourhoods of τ_k such that for any control sequence (s, ι) with the property that $\sum_{j=0}^{k} s_j \in U_k$ for all $k \in \{0, ..., m-1\}$ and for any starting point (x', θ) such that $x' \in U'$, we have

$$\lambda_{min}(s, \iota) \geq \underline{\lambda} > 0. \tag{5.32}$$

Let $T_1, T_2, ...$ be the switching times of the process. On the event that there are m

switches before time t, we introduce the function

$$\Omega(x, \theta, t, T_1, ..., T_m) = x + T_1\theta + (T_2 - T_1)F_{i_1}[\theta] + ... + (t - T_m)F_{i_1,...,i_m}[\theta]$$

that is the end point of the path until time t when the order in which the coordinates change are $i_1, i_2, ..., i_m$.

The random variable $\Omega(x, \theta, T_1, ..., T_m)$ can be simulated using Poisson thinning. As a bounding process we can use the hazard rate $\Lambda(x, \theta)$, defined to be equal to $d\bar{\lambda}$ for all $x \in B(0, R)$ and defined to be equal to $\lambda(x, \theta)$ for all $x \notin B(0, R)$. Using the exponential representation of Poisson process along with Poisson thinning, we can construct the first m switches of the process using i.i.d. $E_1, E_2, ... \sim \exp(1)$ and i.i.d. $u_1, u_2, ... \sim unif(0, 1)$. We then simulate T_1^{prop} from the bounding process Λ such that

$$T_1^{prop} = \inf\{t \geq 0 : \int_0^t \lambda(X_{t'}, \theta)dt' \geq E_1\}.$$

We then accept this time on the event that $Acc_1 = \{u_1 \leq \lambda(X_{T_1^{prop}}, \theta)/\Lambda(X_{T_1^{prop}}, \theta)\}$. If we accept, we set $T_1 = T_1^{prop}$. Else, we start again the process from $(X_{T_1^{prop}}, \theta)$, we pick a new T_2^{prop} according to E_2 and we add it to the previous T_1^{prop}. We keep doing that until we accept. Suppose that the j_1'th proposal was the one that was accepted, then we decide which coordinate of the velocity to change according to the value that u_{j_1} took and more specifically, we change the i_1 coordinate (which is the first coordinate to be switched according to to control sequence u) when the event B_{j_1} occurs, where $B_{j_1} = \{u_{j_1} \leq \lambda_{i_1}(X_{T_1}, \theta)/\Lambda(X_{T_1}, \theta)\} \subset Acc_{j_1}$.

Using the same construction, we construct the second switching time T_2 and if j_2 is the time to accept the switch, then we decide which coordinate to switch according to the value u_{j_2} took and we change the i_2 coordinate (which is the second coordinate to be switched according to control sequence u) on the event that $B_{j_2} = \{u_{j_2} \leq \lambda_{i_2}(X_{T_2}, F_{i_1}[\theta])/\Lambda(X_{T_2}, F_{i_1}[\theta])\} \subset Acc_{j_2}$.

Now, let us condition on the event B that *the first m, $u_1, ..., u_m$ took values less than $\underline{\lambda}/d\bar{\lambda}$ and the first m $E_1, ..., E_m$ took values such that for all $k \in \{1, ..., m\}$, $\frac{1}{d\bar{\lambda}} \sum_{i=1}^{k} E_i \in U_{k-1}$, where U_k are the open subsets of $\mathbb{R}^+$ defined just before equation (5.32). Furthermore, if τ_m is the ending time of the admissible path given by the control sequence u, then E_{m+1} took a value larger than $(\tau_m + \epsilon - \inf U_{m-1})\, d\bar{\lambda}$, where ϵ was introduced in the beginning of the proof.*

Note that B has a positive probability c to occur that does not depend on the starting point of the path. Let us consider what happens if event B occurs. In the beginning, the process moves in a straight line with velocity θ and stays inside the ball of radius $B(0, R)$. This means that the bounding process is equal

to $d\bar{\lambda}$ by construction. Since $a_1 = \frac{1}{d\bar{\lambda}} E_1 \in U_1$, this means that $E_1 = \int_0^{a_1} d\bar{\lambda}\, ds = \int_0^{a_1} \Lambda(X_s, \theta) ds$ and therefore $T_1^{prop} = a_1 \in U_1$. At the same time $u_1 \leq \underline{\lambda}/\bar{\lambda}$ therefore the first switch was accepted and the coordinate to be switched was i_1.

Using the same line of argument and since $\frac{1}{d\bar{\lambda}} (E_1 + E_2) \in U_2$ we can guarantee that until the second proposal time T_2^{prop} the process will remain inside the ball $B(0, R)$ so the bounding rate will be $d\bar{\lambda}$ and therefore the proposed switching time will occur inside U_2. Since $u_2 \leq \underline{\lambda}/\bar{\lambda}$ the switch is accepted and the coordinate to switch is i_2.

Using induction we see that for all $k \in \{1, ..., m\}$ the k'th switching time T_k occurred inside U_k and the i_k coordinate was the one to switch.

Furthermore, from time T_m until $\tau_m + \epsilon$, the process is guaranteed to not leave the ball $B(0, R)$ by construction of ϵ. Therefore, the bounding process until $\tau_m + \epsilon$ is $d\bar{\lambda}$ and since $E_{m+1} > (\tau_m + \epsilon - \inf U_{m-1})\, d\bar{\lambda}$ the process is guaranteed to not switch the velocity until time $\tau_{m+1} + \epsilon$.

This means that given event B occurs and if $T_1, T_2, ...$ are the switching events, then for any $t \in [\tau_m, \tau_m + \epsilon]$, $X_t = \Omega(x, \theta, t, T_1, ..., T_m)$. Furthermore, conditioning on B all times $T_1, ..., T_m$ are distributed according to first arrival times of the homogeneous bounding Poisson process, conditioned on taking values on the sets $U_0, U_1, ..., U_{m-1}$. Therefore, $T_k \sim unif(U_{k-1})$ for all $k = 1, ..., m$. Conditioning on event B occurring, we write

$$\mathbb{P}_{x,\theta} \left(X_t \in \cdot, \Theta_t = \eta \right) \geq c\, \mathbb{P}_{x,\theta} \left(\Omega(x, \theta, t, T_1, ..., T_m) \in \cdot, \Theta_t = \eta \right) \tag{5.33}$$

where $T_k \sim unif(U_{k-1})$. Recall, that c does not depend on the starting position x, θ. We can use the same argument for every starting point (x', θ) for any $x' \in U'$.

Now, since $(x, \theta) \looparrowright (y, \eta)$ we can assume that $\{1, ..., d\} \subset \{i_1, ..., i_m\}$. Therefore, for all $t \in [\tau_m, \tau_m + \epsilon]$, the (up to translation) linear map
$(u_1, ..., u_m) \to \Omega(x, \theta, t, u_1, ..., u_m)$ is of full rank since its matrix has column vectors

$$\left\{ \theta - F_{i_1}[\theta], ..., F_{i_1,...,i_{m-1}}[\theta] - F_{i_1,...,i_m}[\theta] \right\} = \{ \pm 2e_{i_1}, ..., \pm 2e_{i_m} \} = \{ \pm 2e_1, ..., \pm 2e_d \}.$$

From Lemma 6.3 of [BLBMZ15] we get that there exists a neighbourhood U_t of x and $c' > 0$ and a neighbourhood V_t of y such that for all $x' \in U_t$,

$$\mathbb{P}_{x',\theta} \left(\Omega(x, \theta, t, T_1, ..., T_m) \in \cdot, \Theta_t = \eta \right) \geq c' \lambda(\cdot \cap V_t).$$

Now, since $\Omega(x, \theta, t, T_1, ..., T_m) = x + T_1\theta + (T_2 - T_1)F_{i_1}[\theta] + ... + (t - T_m)F_{i_1,...,i_m}[\theta]$, a change in t effects on Ω as a translation in direction $F_{i_1,...,i_m}[\theta]$. This means that

for every $t \in [\tau_m, \tau_m + \epsilon]$ if we pick a starting point (x', θ) with $x' \in U_{\tau_m}$ we get for all $A \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}
&\mathbb{P}_{x',\theta}\left(\Omega(x,\theta,t,T_1,...,T_m) \in A, \Theta_t = \eta\right) = \\
&= \mathbb{P}_{x',\theta}\left(\Omega(x,\theta,\tau_m,T_1,...,T_m) \in A - (t-\tau_m)F_{i_1,...,i_m}[\theta], \Theta_t = \eta\right) \\
&\geq c'\lambda((A - (t-\tau_m)F_{i_1,...,i_m}[\theta]) \cap V_{\tau_m}) \geq c'\lambda(A \cap (V_{\tau_m} + (t-\tau_m)F_{i_1,...,i_m}[\theta])).
\end{aligned}$$

If $\epsilon > 0$ is picked small enough then $\cap_{t\in[\tau_m,\tau_m+\epsilon]}\left(V_{\tau_m} + (t-\tau_m)F_{i_1,...,i_m}[\theta]\right)$ is not empty and contains an open set V_x. Then, for all $x' \in U_{\tau_m}$, and all $A \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}_{x',\theta}\left(\Omega(x,\theta,t,T_1,...,T_m) \in A, \Theta_t = \eta\right) \geq c'\lambda(A \cap V_x).$$

Overall, using (5.33), for all $x' \in U_{\tau_m} \cap U'$, for all $t \in [\tau_m, \tau_m + \epsilon]$ and all $A \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}_{x',\theta}[X_t \in A, \Theta_t = \eta] \geq c\mathbb{P}_{x',\theta}\left(\Omega(x,\theta,t,T_1,...,T_m) \in A, \Theta_t = \eta\right) \geq cc'\lambda(A \cap V_x).$$

which proves the result. □

Lemma 5.4.9 allows us to prove the following stability properties for the process. The proof is the same as the proof of Theorem 5 in [BRZ19] and the proof of Propositions 3.3.32, 3.3.33 and 3.3.34 of this work, but we include it for completeness.

Proposition 5.4.10. *Let $(Z_t)_{t\geq 0} = (X_t, \Theta_t)_{t\geq 0}$ a SUZZ process with speed s. Assume that for all $(x,\theta),(y,\eta) \in E$, $(x,\theta) \nrightarrow (y,\eta) \in E$. Then, the process is T, ϕ-irreducible and aperiodic. If in addition the invariant measure μ is a probability measure then all compact sets are petite, they are also small for any skeleton chain and the process is ergodic.*

Proof of Proposition 5.4.10. We first prove that SUZZ is a T-process. For all $(x,\theta) \in E$ we know that for any other (y,η), $(x,\theta) \nrightarrow (y,\eta)$. From Lemma 5.4.9 we can find U_x, V_x open subsets of $\mathbb{R}^d$ with $x \in U_x$, $y \in V_x$, $c_x, t_x, \epsilon_x > 0$ such that for all $t \in [t_x, t_x + \epsilon_x]$, $x' \in U_x$ and $f : E \to [0, +\infty)$

$$\mathbb{E}[f(X_t, \Theta_t)] \geq c_x \int f(y,\eta) 1_{V_x}(y) dy.$$

Cover $\mathbb{R}^d$ with a sequence of compact sets, for example $\bar{B}(0,1), \bar{B}(0,2)\backslash B(0,1), \bar{B}(0,3)\backslash B(0,2)$ etc. Since they are all compact, each one of them can be covered by finitely many U_x. Therefore, we can construct a sequence $(U_n)_{n=1}^{\infty}, (V_n)_{n=1}^{\infty}, (c_n)_{n=1}^{\infty}, (t_n)_{n=1}^{\infty}, (\epsilon_n)_{n=1}^{\infty}, (\eta_n)_{n=1}^{\infty}$ where $\eta_n \in \{\pm 1\}^d$ such that

- For every $x \in \mathbb{R}^d$, x belongs to at least one and at most finitely many of the U_n's.
- For all $(x, \theta) \in E$ with $x \in U_n$, $t \in [t_n, t_n + \epsilon_n]$ and f non-negative

$$\mathbb{E}_{x,\theta}[f(X_t, \Theta_t)] \geq c_n \int f(y, \eta_n) 1_{V_n}(y) dy$$

Then, we can define the kernel

$$K((x, \theta), A \times \{\eta\}) = \int 1_A(y) \max_{n: x \in U_n} \left\{ c_n 1_{\eta=\eta_n} 1_{V_n}(y) \int_{t_n}^{t_n+\epsilon_n} e^{-t} dt \right\} dy.$$

Then for all $(x, \theta) \in E$, for all $\eta \in \{-1, 1\}^d$ and all $A \in \mathcal{B}(\mathbb{R}^d)$

$$\int_0^{+\infty} \mathbb{P}_{x,\theta}((X_t, \Theta_t) \in A \times \{\eta\}) \exp\{-t\} dt \geq K((x, \theta), A \times \{\eta\}).$$

Summing over all η we see that for any $B \in \mathcal{B}(E)$, $\int_0^{+\infty} \mathbb{P}_{x,\theta}((X_t, \Theta_t) \in B) \exp\{-t\} dt \geq K((x, \theta), B)$.
Also, if $x \in U_n$, $K((x, \theta), E) \geq c_n \int_{t_n}^{t_n+\epsilon_n} e^{-t} dt > 0$.
Finally, let $(x_k)_{k=1}^{\infty} \subset \mathbb{R}^d$ with $x_k \xrightarrow{k\to\infty} x$. Then, eventually x_k will belong to every U_n in which x belongs, so for any A , for all η, $\liminf_{k\to\infty} K((x_k, \theta), A \times \eta) \geq K((x, \theta), A \times \{\eta\})$. This means that $K((\cdot), A \times \{\eta\})$ is lower semi-continuous. Summing over all η and since a finite sum of lower semi-continuous functions is lower semi-continuous, we get that $K(\cdot, B)$ is lower semi-continuous for all $B \in \mathcal{E}$. Therefore, SUZZ is a T-process.

For ϕ-irreducibility, let $O \subset \mathbb{R}^d$ an open set. For any $\eta \in \{\pm 1\}^d$ and $y \in O$, since $(x, \theta) \leadsto (y, \eta)$ an application of Lemma 5.4.9 shows that there exists a time t, $c > 0$ and $V \subset O$ with $y \in V$ and $\mathbb{P}_{x,\theta}(Z_t \in O \times \{\eta\}) \geq cLeb(V) > 0$ and therefore if $\tau_{0\times\{\eta\}}$ is the first hitting time of $0 \times \{\eta\}$, we get $\mathbb{P}_{x,\theta}(\tau_{0\times\{\eta\}} < \infty) > 0$. This means that the process is open set irreducible (see [Twe95]) and from Theorem 3.2 in [Twe95] an open set irreducible and T-process is ϕ-irreducible.

To prove aperiodicity, let $(x, \theta) \in E$. Since $(x, \theta) \leadsto (x, \theta)$, from Lemma 5.4.9 there exist U_x, V_x open neighbourhoods of x, $t_0, \epsilon, c > 0$ such that for all $t \in [t_0, t_0 + \epsilon], x' \in U_x$

$$\mathbb{P}_{x',\theta}(X_t \in \cdot, \Theta_t = \theta) \geq cLeb(\cdot \cap V_x). \tag{5.34}$$

By making U_x, V_x smaller we can assume without loss of generality that $U_x = V_x$

and (5.34) implies that $U_x \times \{\theta\}$ is small and for all $x' \in U_x$, $t \in [t_0, t_0 + \epsilon]$

$$\mathbb{P}_{x',\theta}(X_t \in U, \Theta_t = \theta) \geq c' > 0.$$

Take $N = \lceil t_0/\epsilon \rceil$ and $T = Nt_0$. Let $t \geq T$ and define $n = \lfloor t/t_0 \rfloor$. Then $n \leq t/t_0$ so $nt_0 \leq t$. Also, $n = \lfloor t/t_0 \rfloor \geq \lfloor Nt_0/t_0 \rfloor = N = \lceil t_0/\epsilon \rceil \geq t_0/\epsilon$. Therefore $n\epsilon \geq t_0$. At the same time $t \leq t_0(n+1) = t_0 n + t_0 \leq t_0 n + \epsilon n = n(t_0 + \epsilon)$. Overall, $t \in [nt_0, n(t_0 + \epsilon)]$. From the Markov Property for any $x' \in U$ and $t \geq T$

$$\mathbb{P}_{x',\theta}((X_t, \Theta_t) \in U \times \{\theta\}) \geq (c')^n > 0$$

This proves aperiodicity.

Now, we prove that every compact set is petite. A standard argument as in [BRZ19] shows that for μ-almost all starting points (x, θ) we have $\mathbb{P}_{x,\theta}(\lim_{t\to+\infty} \|X_t\| = +\infty) = 0$. Indeed, for any compact set K, $1_{\{X_t \text{ eventually leaves } K\}} = 1 \liminf_{t\to+\infty} 1_{X_t \notin K}$ and by Fatou's lemma

$$\mathbb{P}_\mu(X_t \text{ eventually leaves } K) \leq \liminf_{t\to+\infty} \mathbb{P}_\mu(X_t \notin K) = 1 - \mu(K)$$

and by exhausting E with compact sets we get $\mathbb{P}_\mu(\lim_{t\to+\infty} \|X_t\| = +\infty) = 0$. More specifically, there exists $(x, \theta) \in E$ such that $\mathbb{P}_{x,\theta}(\lim_{t\to+\infty} \|X_t\| = +\infty) < 1$. From Theorem 4.1 in [MT93a] all compact sets are petite iff the process is T and ϕ-irreducible. The result follows.

Furthermore, the process is positive Harris recurrent from an application of Theorem 4.4 of [MT93b].

Next, we want to prove that some skeleton of the process is irreducible. For this we use again Lemma 5.4.9 and since for any (x, θ) we have $(x, \theta) \nrightarrow (x, \theta)$ we find U_x, V_x neighbourhoods of x, t_0, ϵ, c such that (5.31) holds. For any two points $(y, \eta), (y', \eta')$ we have $(y, \eta) \nrightarrow (x, \theta) \nrightarrow (x, \theta) \nrightarrow (y', \eta')$. Therefore there exist t_1, c_1 and a neighbourhood V_1 of x such that

$$\mathbb{P}_{y,\eta}((X_{t_1}, \Theta_{t_1}) \in \cdot \times \{\theta\}) \geq c_1 Leb(\cdot \cap V_1)$$

and constants t_2, c_2 and two neighbourhoods U_2, V_2 of x, y' respectively such that for all $x' \in U_2$

$$\mathbb{P}_{x',\theta}((X_{t_2}, \Theta_{t_2}) \in \cdot \times \{\theta\}) \geq c_2 Leb(\cdot \cap V_2)$$

Then for any $t \in [t_0 + t_1 + t_2, t_0 + t_1 + t_2 + \epsilon]$ we get for any open neighbourhood O

of y'

$$\begin{aligned}
&\mathbb{P}_{y,\eta}((X_t,\Theta_t)\in O\times\{\eta'\})\geq\\
&\geq\mathbb{P}_{y,\eta}\left(\Theta_{t_1}=\Theta_{t-t_2}=\theta,\Theta_t=\eta',X_{t_1}\in U_x\cap V_1,X_{t-t_2}\in V_x\cap U_2,X_t\in O\cap V_2\right)\geq\\
&\mathbb{P}_{y,\eta}\left(\Theta_{t_1}=\theta,\Theta_t=\eta',X_{t_1}\in U_x\cap V_1,X_{t-t_2}\in V_x\cap U_2\right)c_2Leb(O\cap V_2)\geq\\
&\mathbb{P}_{y,\eta}(\Theta_{t_1}=\theta,X_{t_1}\in U_x\cap V_1)cLeb(V_x\cap U_2)c_2Leb(O\cap V_2)\geq\\
&\geq cc_1c_2Leb(U_x\cap V_1)Leb(V_x\cap U_2)Leb(O\cap V_2)>0.
\end{aligned}$$

The time interval $[t_0+t_1+t_2,t_0+t_1+t_2+\epsilon]$ contains at least one multiple of ϵ so the ϵ-chain starting from any (y,η) has a positive probability of reaching any O open neighbourhood of any (y',η') making the chain open set irreducible and therefore irreducible.

From Theorem 6.1 of [MT93a] we get that the process is ergodic.

Finally, all compacts are small for any skeleton chain from Proposition 6.1 in [MT93a]. □

Proof of Theorem 5.4.1. From Propositions 5.2.10 and 5.3.2 we know that the process is non-explosive and μ introduced in (2.26) is invariant for the SUZZ. By Proposition 5.4.8 for all $(x,\theta),(y,\eta)\in E$ we have $(x,\theta)\nrightarrow(y,\eta)$ and therefore, by Proposition 5.4.10 the process is ϕ-irreducible , aperiodic and all compact sets are small for some skeleton chain. Also, from Lemma 5.2.13, V as in (5.11) satisfies the drift condition (5.13). All the conditions of Theorem 5.4.4 are satisfied and the result follows. □

Proof of Corollary 5.4.3. Since $g(x,\theta)\leq\exp\{bU(x)\}$ for some $b<1/2$ outside a compact set, we can find $a\in(2b,1)$ and if we consider the function V defined in (5.11) (where we set $s(x)=1$ since we are in the original Zig-Zag case) and for all x outside some compact set $g^2(x,\theta)\leq V(x,\theta)$. Since both functions are continuous and V is bounded away from zero, there exists a constant C such that for all (x,θ), $g^2(x,\theta)\leq CV(x,\theta)$. Function V satisfies the drift condition (5.13) and the result is an immediate corollary of Theorem 2.1.11. □

5.5 Relationship Between Speed Up Zig-Zag and Zig-Zag in One Dimension

When we focus on one dimensional process, we can prove geometric ergodicity results in a more straightforward way. We can prove that any one dimensional Speed

Up process is a space transformation of an original one dimensional Zig-Zag process and then we can use known results for the original case to pass to the Speed Up case. Furthermore, using this line of argument we can prove that if we use explosive deterministic dynamics we can create algorithms that are uniformly ergodic, meaning that the mixing rate of the algorithm does not depend on the starting point. This is because when we use explosive deterministic dynamics we can reach infinity in finite time, which means that if we reverse the time, the process can return close to the mode from any starting point "close to infinity" in a small amount of time.

Theorem 5.5.1 (Uniform Ergodicity in One Dimension)**.** *Consider an one dimensional Speed Up Zig-Zag process $Z_t = (X_t, \Theta_t)_{t\geq 0}$ with speed function $s(x)$, targeting a measure μ as in (5.5). Assume that for any (x, θ) the deterministic flow of the process $\{\Phi_{(x,\theta)}(t), t \geq 0\}$ has a finite explosive time $t^*(x, \theta)$ and that Assumptions 5.1.5, 5.2.2, 5.2.5 and 5.2.7 hold. Then the process is non-explosive and uniformly ergodic, i.e. there exists a $M > 0$ and $\rho < 1$ such that for any $(x, \theta) \in E$*

$$\|\mathbb{P}(Z_t \in \cdot) - \mu(\cdot)\|_{TV} \leq M\rho^t.$$

Proof of Theorem 5.5.1. Introduce the space transformation $f(x) = \int_0^x 1/s(u)du$ and note that from a separation of variables technique the solution of the ODE

$$\begin{cases} \dfrac{dX_t}{dt} = \theta s(x) \\ X_0 = x_0 \end{cases} \tag{5.35}$$

satisfies

$$X_t = f^{-1}(f(x_0) + \theta t). \tag{5.36}$$

Therefore, having finite explosive time when $\theta = +1$ means that for some $M^+ > 0$, $\lim_{y\to M^+} f^{-1}(y) = +\infty$ and for $\theta = -1$ means that for some $-M^- < 0$, $\lim_{y\to -M^-} f^{-1}(y) = -\infty$. This means that f maps $\mathbb{R}$ to some finite interval $(-M^-, M^+)$. Consider the process $(Y_t, \Theta_t)_{t\geq 0}$ where $Y_t = f(X_t)$. Then the process $(Y_t, \Theta_t)_{t\geq 0}$ is a one dimensional normal Zig-Zag with unit speed and intensity rates $\lambda_Y(y, \theta) = \lambda(f^{-1}(y), \theta)$. To see this, note that when the process starts from point $(y, \theta) = (f(x), \theta)$, it follows deterministic dynamics

$$\frac{dY_t}{dt} = f'(X_t)\frac{dX_t}{dt} = \frac{1}{s(X_t)}\Theta_t s(X_t) = \Theta_t$$

and

$$\frac{d\Theta_t}{dt} = 0.$$

Furthermore, the random time T when the sign of Θ_t changes, is the same time when the direction changes for the $(X_t, \Theta_t)_{t\geq 0}$ process. Therefore,

$$\mathbb{P}(T \geq t) = \exp\left\{-\int_0^t \lambda(X_s, \theta)ds\right\} = \exp\left\{-\int_0^t \lambda(f^{-1}(Y_s), \theta)ds\right\}.$$

This proves that $(Y_t, \Theta_t)_{t\geq 0}$ is a one dimensional normal Zig-Zag with unit speed and intensity rates $\lambda_Y(y, \theta) = \lambda(f^{-1}(y), \theta)$.

Note that the flow $\{\Phi_{(x,\pm 1)}(t), t \geq 0\}$ will have finite explosive time iff $\left|\int_0^{\pm\infty} 1/s(u)du\right| = M^{\pm} < \infty$ and in that case the transformed process $(Y_t, \Theta_t)_{t\geq 0}$ is defined on the space $(M^-, M^+) \times \{\pm 1\}$.

Due to Assumption 5.2.5, we have $\lim_{y\to\pm M^{\pm}} \lambda_Y(y, \pm 1) = \lim_{x\to\pm\infty} \lambda(x, \pm 1) > 0$. Following the argument for the proof of geometric ergodicity in [BR17] we get that there exist $a^{\pm}, b^{\pm} > 0$, $y_1 \in (0, \min\{M^+, M^-\})$ such that the function V defined on $(-M^-, -y_1] \cup [y_1, M^+)$ by

$$V(y, \theta) = \begin{cases} \exp\{a^+ y + b^+\theta\}, & y \in [y_1, M^+] \\ \exp\{-a^- y - b^-\theta\}, & y \in [-M^-, -y_1] \end{cases}$$

and extended as C^1 and positive in $[-y_1, y_1]$ satisfies the drift condition for the generator $\mathcal{L}_Y$ of (Y_t, Θ_t),

$$\mathcal{L}_Y V(y, \theta) \leq -cV(y, \theta) + b1_{[-y_1, y_1]}$$

for some $c, b > 0$. By using the technique of [MT93b], of restricting the process onto sets O_m and stopping the process upon hitting O_m^c, as in the previous section, where we prove Geometric Ergodicity in any dimension, we get for the transformed process $(f(X_t), \Theta_t)$ that there exists a constant $C > 0$ and $\rho < 1$ and a measure ν such that for all $y \in (-M^-, M^+)$ and any $\theta \in \{\pm 1\}$,

$$\|\mathbb{P}_{f(x),\theta}((f(X_t), \Theta_t) \in \cdot) - \nu(\cdot)\|_{TV} \leq CV(y, \theta)\rho^t.$$

Since f is $1-1$ we get that for all $x \in \mathbb{R}$, $\theta \in \{\pm 1\}$,

$$\|\mathbb{P}_{x,\theta}(Z_t \in \cdot) - \nu(f^{-1}(\cdot))\|_{TV} \leq CV(f(x), \theta)\rho^t.$$

Since μ is invariant for the process Z_t due to Proposition 5.3.2, we get $\nu(f^{-1}(\cdot)) = \mu(\cdot)$. Note that V is bounded on $(-M^-, M^+)$ so there exists an M such that

$V(f(x),\theta) \leq M$ for all $(x,\theta) \in \mathbb{R} \times \{\pm 1\}$. This means that

$$\|\mathbb{P}_{x,\theta}(Z_t \in \cdot) - \mu(\cdot)\|_{TV} \leq CM\rho^t.$$

This proves the result. □

Note that the same proof can be used in one dimension when the deterministic dynamics are non-explosive. The difference is that the space transformation f maps $\mathbb{R}$ onto $\mathbb{R}$ instead of a finite interval. In that way, the Lyapunov function on the transformed space $\mathbb{R}$ is not necessarily bounded and we do not obtain uniform ergodicity.

Theorem 5.5.2. *Consider an one dimensional Speed Up Zig-Zag process $Z_t = (X_t, \Theta_t)_{t\geq 0}$ with speed function $s(x)$ once weakly differentiable, targeting a measure μ as in (5.5). Assume that Assumptions 5.1.5, 5.2.2, 5.2.5 and 5.2.7 hold. Then the process is non-explosive, has μ as invariant measure and is geometrically ergodic.*

From the proof of Theorem 5.5.1 we see that the one dimensional SUZZ process is a space transformation of a Zig-Zag process. More precisely, we have the following.

Proposition 5.5.3. *Consider an one dimensional Speed Up Zig-Zag process $Z_t = (X_t, \Theta_t)_{t\geq 0}$ with speed function $s(x)$, targeting a measure μ as in (5.5).Assume that Assumptions 5.1.5, 5.2.2, 5.2.5 and 5.2.7 hold. Let $f(x) = \int_0^x 1/s(u)du$ and $\pm M^{\pm} = \lim_{x\to\pm\infty} f(x) \in \mathbb{R} \cup \{-\infty, +\infty\}$. Then, the process $(Y_t, \Theta)_{t\geq 0}$, where $Y_t = f(X_t)$, is an one dimensional Zig-Zag process, defined on $(-M^-, M^+) \times \{-1, +1\}$ and has measure ν invariant where*

$$\nu(dy, d\theta) = \frac{1}{Z_0} \exp\{-V(y)\}dyd\theta \tag{5.37}$$

and

$$V(y) = U(f^{-1}(y)) - \log s(f^{-1}(y)). \tag{5.38}$$

Proof. The fact that (Y_t, Θ_t) is a Zig-Zag process comes from inspecting the Proof of Theorem 5.5.1. In that proof we also prove that if λ are the rates of the SUZZ process, then the rates λ_Y of the transformed Zig-Zag process satisfy

$$\begin{aligned}\lambda_Y(y,\theta) &= \lambda(f^{-1}(y),\theta) = [\theta\left(s(f^{-1}(y))U'(f^{-1}(y)) - s'(f^{-1}(y))\right)]^+ = \\ &= \left[\theta\left(\left(U(f^{-1}(y)) - \log s(f^{-1}(y))\right)'\right)\right]^+\end{aligned}$$

where we used that $(f^{-1})'(y) = s(f^{-1}(y))$. Since, from well known results for the Zig-Zag process we have

$$\lambda_Y(y,\theta) = [\theta V(y)]^+$$

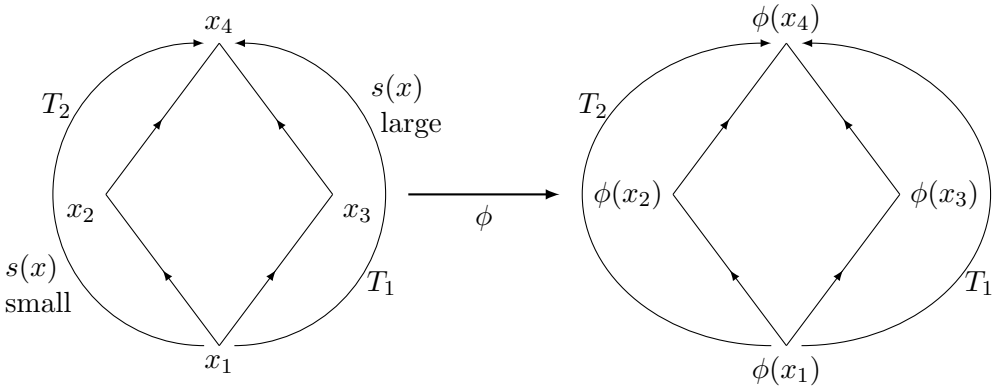


Figure 5.1: *Figure explaining why a two dimensional SUZZ is not in general a space transformation ϕ of a normal Zig-Zag process. The times, T_1 and T_2, to traverse the two paths from x_1 to x_4 on the left figure, depending on the speed function, do not have to be the same. The same times, T_1, T_2, would be the times to traverse the two paths from $\phi(x_1)$ to $\phi(x_4)$ on the ϕ-transformed right figure. However, if the ϕ-transformed right figure was an ordinary Zig-Zag, moving with constant unit speed, these two times would had to be the same.*

we get the result. □

We emphasise, however, that the SUZZ algorithm is no longer a space transformation of a simple Zig-Zag in higher dimensions. This justifies the introduction of the process in this work, since, if it was just a space transformation, one could, instead, run an original Zig-Zag in the transformed space. To see that a two dimensional SUZZ is no longer a space transformation of an ordinary Zig-Zag, see Figure 5.1. If we assume that it is a space transformation, the left figure represents the movement of a two-dimensional SUZZ process starting from x_1 and ending at x_4. The right figure represents the movement of the ϕ-space transformed process, assumed to be a simple Zig-Zag, starting from $\phi(x_1)$ and ending at $\phi(x_4)$. There are two paths from x_1 to x_4, passing through x_2 or x_3 respectively. If the speed function $s(x)$ takes smaller values on the path via x_2, then the process arrives at x_4 faster via the x_3 path than via the x_2 path. However, the transformed process moves with constant unit speed, as it is a simple Zig-Zag process and the two paths from $\phi(x_1)$ to $\phi(x_4)$, passing either via $\phi(x_2)$ or via $\phi(x_3)$ have the same length. Therefore the time it takes for the transformed process to traverse either of the two paths is the same, which is a contradiction.
We will discuss more on the choice of speed function s in the next two sections.

5.6 A First Proposal for the Choice of Speed Function in One Dimension

The next natural question would be what is an appropriate speed function s to be used in applications. We note that from the proof of Theorems 5.5.1 and 5.5.2 the SUZZ process $(X_t, \Theta_t)_{t\geq 0}$ can be seen as the space transformation of a one dimensional Zig-Zag process targeting a different measure. More specifically, as we saw in Theorem 5.5.1 and Proposition 5.5.3 in the previous section, if $f(x) = \int_0^x 1/s(u)du$, then $(Y_t, \Theta_t) = (f(X_t), \Theta_t)$ is a normal Zig-Zag process. The rates of the Zig-Zag process are $\lambda_Y(y, \theta) = [\theta(U(f^{-1}(y)) - \log s(f^{-1}(y)))']^+$ and this means that this process is targeting the potential $V(y) = U(f^{-1}(y)) - \log s(f^{-1}(y))$ defined on a subset of $\mathbb{R}$. Therefore, instead of using SUZZ, one could simply use the normal Zig-Zag (Y_t, Θ_t), target the potential V and then use the path of $f^{-1}(Y_t)$ as a way to sample from potential U. In order for this to work, we need to be able to evaluate the function f^{-1} point wise, but note that, due to (5.36), this is a necessary and sufficient condition in order to be able to simulate from the deterministic dynamics of the SUZZ given by the ODE (5.3). Therefore, all the discussion of choosing the best speed s boils down to choosing the best possible space transformation f so that f^{-1} can be evaluated point wise and the potential

$$V(y) = U(f^{-1}(y)) - \log s(f^{-1}(y)) = U(f^{-1}(y)) + \log f'(f^{-1}(y))$$

leads to an efficient normal Zig-Zag algorithm. If we view SUZZ in that sense and ignoring Assumption 5.2.2 for the moment, then when we want to target $\pi(dx) = \exp\{-U(x)\}dx$ the choice of $s(x) = \exp\{U(x)\}$ seems reasonable. This is because the space transformation is

$$f(x) = \int_0^x \exp\{-U(x)\}dx = F(x) - \pi((-\infty, 0]),$$

where F is the CDF of π, and running this SUZZ is equivalent to run a normal Zig-Zag on the potential $V \equiv 0$ (i.e. the Lebesgue measure) and then transform the values back according to the function $f^{-1}(y)$. This is basically inverse sampling.

We do note, however, that this choice of s is not admissible in our setting as it violates Assumption 5.2.2 and leads to explosive algorithms. This can also be seen in terms of the transformed original Zig-Zag process. The process will have state space $(-M^-, M^+) \times \{\pm 1\}$ where $f^{-1}(\pm M^{\pm}) = \pm\infty$, so we would need a mechanism to ensure that the process will not hit the boundary $\{\pm M^{\pm}\}$ in order for the algorithm not to explode. Furthermore, even if Assumption 5.2.2 was not violated,

assuming that one can simulate directly the deterministic dynamics induced by a speed function $s(x) = \exp\{U(x)\}$ is not a realistic assumption.

The fact, however, that the choice of speed $s(x) = \exp\{U(x)\}$ would be attractive, motivates us to pick a speed function s such that $s(x)\exp\{-U(x)\}$ decays to zero as $|x| \to \infty$ (so that Assumption 5.2.2 holds), but does so as slowly as possible. For example, if the non-normalised invariant distribution $\exp\{-U(x)\}$ decays like $|x|^{-(1+k)}$ at the tails, it seems reasonable to pick a speed function s that grows at the tails like $|x|^{1+\epsilon}$ for some ϵ very close to k, but slightly less than k. Note that this polynomial speed function induces deterministic dynamics that can be simulated exactly. This idea of picking a speed function so that $s(x)\exp\{-U(x)\}$ decays as slowly as possible could serve as a motivation in higher dimensions as well, even though, as argued earlier, the SUZZ process is no longer a Zig-Zag space transformation.

5.7 Computational Efficiency

So far we have established convergence properties for the SUZZ process. The original motivation for introducing the process was that the original Zig-Zag is not geometrically ergodic for heavy tailed targets. We expect a SUZZ process to converge much faster if we pick a speed s that satisfies conditions 5.1.5, 5.2.5, 5.2.7 and 5.3.1. However, the question of how efficient these algorithms are, relative to the normal Zig-Zag is a bit more subtle. For example, one could try to speed up the original Zig-Zag process by a constant factor of two, setting $s(x) = 2$. The mixing time of the process would be half the original one. However, this does not constitute any real improvement, since the actual cost of implementing an MCMC algorithm does not come from the time it has to run until it mixes, but from the number of operations that need to be performed. In our setting the cost comes from evaluating the gradient of the log-likelihood of the target and if we were to speed up the process by a constant factor of two, we would at the same time increase the number of gradient log-likelihood evaluations, again, by a factor of two and ultimately would not gain anything in terms of efficiency. Therefore, in order to compare the efficiency of the algorithms, we need to compare the number of the gradient log-likelihood evaluations need to be performed until we get a good approximation of the quantity of interest. In the SUZZ setting, these evaluations appear in the implementation of the algorithm when one uses Poisson thinning to simulate the times when the process switches direction. In a perfect setting the Poisson thinning would not generate any switches finally rejected and the number of gradient log-likelihood evaluations

would be equal to the actual number of switches. Therefore, since the number of gradient log-likelihood evaluations depends on the programmer's coding skills, we find it more convenient to use the number of direction switches as a unit for measuring the efficiency of the algorithm. For the remainder of this section, we focus on dimension one and closely follow [BD17].

Let N_T be the expected number of switches until time T. Then $N_T = \mathbb{E}[\int_0^T \lambda(X_s, \Theta_s)ds]$. Since the process is Harris recurrent, we have a Law of Large numbers and

$$\begin{aligned} N_0 := \lim_{n\to\infty} \frac{N_T}{T} = \int \lambda(x,\theta)d\mu(x,\theta) &= \frac{1}{2Z}\sum_{\theta=\pm 1}\int_{\mathbb{R}} \exp\{-U(x)\}\lambda(x,\theta)dx \quad (5.39) \\ &= \frac{1}{2Z}\int_{\mathbb{R}} \exp\{-U(x)\}|s(x)U'(x) - s'(x)|dx \end{aligned}$$

Consider a functional $g \in L^2(\mu)$ whose expectation with respect to the targeting measure we are trying to approximate. For simplicity of notation, for the rest of the section, we will assume that $\mu(g) = 0$. Suppose further that a SUZZ process is used to approximate this expectation and when run until time T, the process generates the estimator g_T. If SUZZ could generate completely independent samples from μ until time T, then $Var(g_T) = \frac{Var_\mu(g)}{T}$. Since the generated samples from SUZZ are correlated, the expression emerging when solving with respect to T can be viewed as an approximation of the number of independent samples that the process can generate when run until time T and we then define the quantity of Effective Sample Size (ESS) of SUZZ for this g as

$$ESS(T) = \frac{Var_\mu(g)}{Var(g_T)}.$$

Overall, if we view ESS as a function of the number of switches, it seems natural to consider the quantity of ESS per number of direction switches in order to evaluate the efficiency of the algorithm. It turns out that for the one dimensional SUZZ we can approximate this quantity reasonably well assuming the existence of a CLT for the functional g. In that case, if γ_g^2 is the asymptotic variance of the SUZZ process, we can estimate $Var(g_T) \approx \gamma_g^2/T$ and therefore

$$ESS(T) = \frac{Var_\mu(g)}{Var(g_T)} \approx \frac{Var_\mu(g)}{\gamma_g^2}T \approx \frac{Var_\mu(g)}{\gamma_g^2}\frac{N_T}{N_0}$$

and finally

$$\lim_{T\to\infty} \frac{ESS(T)}{N_T} = \frac{Var_\mu(g)}{\gamma_g^2 N_0}.$$

Therefore, in order to choose the optimal s that makes the algorithm the most efficient we need to minimize the quantity $\gamma_g^2 N_0$ over different speed functions. The following proposition, describes the asymptotic variance γ_g^2 in terms of s. Before that, we need to make an assumption.

Assumption 5.7.1. *Let $\mathcal{L}$ be the operator defined in (5.4) for the one-dimensional SUZZ, i.e. for all $f \in C^1(E)$,*

$$\mathcal{L}f(x,\theta) = \theta s(x) f'(x,\theta) + \left([\theta U'(x)]^+ + \gamma(x)\right)\left(f(x,-\theta) - f(x,\theta)\right).$$

Let V defined as in (5.11) for $d = 1$. Assume that there exists a $c > 0$ such that for all $g \in L^1(E)$ satisfying $g(x,\theta) \leq V(x,\theta)$ for all $(x,\theta) \in E$, there exists a ϕ such that

$$-\mathcal{L}\phi = g$$

and such that for all $(x,\theta) \in E$

$$\phi(x,\theta) \leq c\ V(x,\theta).$$

This is not an intuitive assumption, however it is a result proven in [GM96] in the case where $\mathcal{L}$ is the extended generator of a process. In our case, due to the fact that we allow the process to have explosive deterministic dynamics, $\mathcal{L}$ is only the generator of the process stopped when exiting a ball. This basically means that for any $f \in C^1(E)$ the process

$$M_t = f(X_t) - f(X_0) - \int_0^t \mathcal{L}f(X_s)\ ds$$

is a Local Martingale, instead of a Martingale (as would be the case for $\mathcal{L}$ to be the extended generator of the process). We note here that in [GM96] the authors claim that Assumption 5.7.1 holds in our case as well, i.e. when $\mathcal{L}$ only induces a Local Martingale. However, to the best of our knowledge this is not something proven in the literature. Therefore, we make this assumption here and we present the following result under Assumption 5.7.1.

Proposition 5.7.2. *Assume that s satisfies Assumptions 5.1.5, 5.2.2, 5.2.5 and*

5.2.7. Let $a < 1$ and $\delta > 0$ be small enough, such that for the function

$$V(x,\theta) = \exp\{aU(x) - a\log s(x) + \sum_{i=1}^{d} \frac{1}{2} sign(\theta A(x)) \log(1 + \log(1 + \delta|A(x)|))\}$$

the drift condition (5.13) holds. Let $g : E \to \mathbb{R}$ with $\mu(g) = 0$ and assume that there exists a constant $C > 0$ such thatg$(x,\theta) \leq CV(x,\theta)$ for all $(x,\theta) \in E$. Finally, assume that Assumption 5.7.1 holds. Then, if $Z_t = (X_t, \Theta_t)$ is the SUZZ process with speed s, starting from the invariant measure μ, then we have

$$\frac{1}{\sqrt{T}} \int_0^T g(Z_s)ds \xrightarrow{T\to\infty} N(0, \gamma_g^2)$$

in distribution where

$$\gamma_g^2 = \frac{1}{2Z} \int_{\mathbb{R}} \left|s(x)U'(x) - s'(x)\right| \frac{1}{s^2(x)\exp\{-U(x)\}} k^2(x)dx \tag{5.40}$$

and

$$k(x) = \int_x^{+\infty} (g(y,+1) + g(y,-1)) \exp\{-U(y)\}dy. \tag{5.41}$$

Proposition 5.7.2, equation (5.39) and the above discussion indicates that for a given function $g : E \to [0, +\infty)$ such that $g(x,\theta) \leq \exp\{aU(x) - a\log s(x)\}$ for some value a small enough and $\mu(g) = 0$, we need to pick a speed function s in order to minimize the quantity

$$J[r] = \gamma_g^2 N_0 = \int_{\mathbb{R}} |r'(x)|dx \int_{\mathbb{R}} \frac{|r'(x)|}{r^2(x)} k^2(x)dx \tag{5.42}$$

where

$$r(x) = s(x)\exp\{-U(x)\}$$

and we need to impose the condition

$$\lim_{|x|\to\infty} r(x) = 0$$

so that Assumption 5.2.2 holds.

Note that the functional J to be minimized is invariant under constant scaling of function s. This is in accordance to the fact that we don't gain any efficiency by speeding up Zig-Zag with a constant speed, for example by having velocities of the form $\{\pm 2\}^d$.

Remark 5.7.3. *So far we have assumed that $\mu(g) = 0$. In the general case, the*

only difference would be that the function k used to define $J[r]$ is

$$k(x) = \int_x^{+\infty} \left(g(y,+1) + g(y,-1) - \frac{1}{2}\mu(g) \right) \exp\{-U(y)\} dy.$$

Of course in general we do not know $\mu(g)$. However, the quantity J can be used as an approximation of the inverse of efficiency of a one dimensional SUZZ algorithm when one tests the algorithms on specific examples where the quantity of interest, $\mu(g)$, is known. It would be interesting to find a way to estimate J without knowing $\mu(g)$.

Note that minimising (5.42) is not a well-posed problem. Indeed, let r_0 be a function such that $J[r_0] < \infty$ and $\lim_{|x|\to\infty} r_0(x) = 0$ and for any $n \in \mathbb{N}$, let

$$\begin{cases} r_n(x) = 1, |x| \leq n \\ r_n(x) = r_0(x-n), x > n \\ r_n(x) = r_0(x+n), x < -n \end{cases} . \tag{5.43}$$

Then $J[r_n] \xrightarrow{n\to\infty} 0$. At the same time, the only functions that satisfy $J[r] = 0$ are the constant ones and since we impose the condition that $\lim_{|x|\to\infty} r(x) = 0$ this is not a possible choice for a strictly positive r.

Note, however, that the n'th term of the minimising sequence r_n is equal to 1 on $[-n, n]$ and this means that $s(x) = \exp\{U(x)\}$ for $x \in [-n, n]$. Heuristically, and as discussed in the previous section, one could expect good performance in the ideal case where $s(x)$ could be set equal to $\exp\{U(x)\}$ for $x \in [-n, n]$ for some large n.

Below, we present three examples where the quantity J can be approximated sufficiently well.

Example 5.7.4 (Example 1)**.** *Let us suppose that we are targeting a Normal distribution with mean 0 and variance 1 and we are using SUZZ with speed function $s(x) = \left(1 + x^2\right)^{(1+\epsilon)/2}$ to estimate the mean, i.e. $g(x, \theta) = x$.*

Then $U(x) = 1/2x^2$ so $r(x) = s(x)\exp\{-U(x)\} = \left(1 + x^2\right)^{(1+\epsilon)/2} \exp\{-1/2x^2\}$ so

$$\frac{|r'(x)|}{r^2(x)} = \exp\{1/2x^2\} \left(1 + x^2\right)^{-(1+\epsilon)/2} |x| \left(\frac{1+\epsilon}{1+x^2} - 1 \right).$$

Furthermore, $k(x) = \int_x^{+\infty} (g(y,+1) + g(y,-1)) \exp\{-U(y)\} dy = 2\exp\{-/2x^2\}$ and after a few calculations and a change of variables $u = x^2$, we get

$$\int_{\mathbb{R}} \frac{|r'(x)|}{r^2(x)} k^2(x) dx = 2 \int_0^{+\infty} \exp\{-1/2u\} (1+u)^{(-3+\epsilon)/2} |\epsilon - u| du.$$

Also, simple calculations give

$$\int_{\mathbb{R}} |r'(x)|dx = 2\left(2\exp\{-1/2\epsilon\}\,(1+\epsilon)^{(1+\epsilon)/2} - 1\right)$$

and overall

$$J[r] = 4\left(2\exp\{-1/2\epsilon\}\,(1+\epsilon)^{(1+\epsilon)/2} - 1\right)\int_0^{+\infty} \exp\{-1/2u\}\,(1+u)^{(-3+\epsilon)/2}\,|\epsilon - u|du$$

and the integral on the RHS can be numerically approximated.

Example 5.7.5 (Example 2)**.** *Let us suppose that we are targeting a double Exponential distribution with mean* 0 *and we are using SUZZ with speed function* $s(x) = \left(1+x^2\right)^{(1+\epsilon)/2}$ *to estimate the mean, i.e.* $g(x,\theta) = x$*.*

Then $U(x) = |x|$ *so* $r(x) = s(x)\exp\{-U(x)\} = \left(1+x^2\right)^{(1+\epsilon)/2}\exp\{-|x|\}$

so

$$\frac{|r'(x)|}{r^2(x)} = \exp\{|x|\}\left(1+x^2\right)^{-(1+\epsilon)/2-1}|x|\left(|x|(1+\epsilon) - 1 - x^2\right).$$

Furthermore, $k(x) = \int_x^{+\infty}(g(y,+1)+g(y,-1))\exp\{-U(y)\}dy = 2\exp\{-|x|\}(1+|x|)$ *and after a few calculations, we get*

$$\int_{\mathbb{R}} \frac{|r'(x)|}{r^2(x)}k^2(x)dx = 4\int_0^{+\infty} \exp\{x\}\left(1+x^2\right)^{(-3+\epsilon)/2}(1+x)^2|x(1+\epsilon) - 1 - x^2|du.$$

Also, simple calculations give for $\epsilon \in (0,1)$

$$\int_{\mathbb{R}} |r'(x)|dx = 2$$

and overall, for $\epsilon \in (0,1)$

$$J[r] = 8\int_0^{+\infty} \exp\{-x\}\left(1+x^2\right)^{(-3+\epsilon)/2}(1+x)^2\left|x(1+\epsilon) - 1 - x^2\right|dx$$

and the integral on the RHS can be numerically approximated.

Example 5.7.6 (Example 3)**.** *Let us suppose that we are targeting a Student distribution with* v *degrees of freedom. Assume for this example that* $\nu > 1$ *so that the expectation exists, it is equal to* 0 *and we are trying to estimate it using SUZZ with speed function* $s(x) = \left(1+x^2\right)^{(1+\epsilon)/2}$*, therefore* $g(x,\theta) = x$*.*
Then $U(x) = (\nu+1)/2\log(1+x^2/\nu)$ *so*
$r(x) = s(x)\exp\{-U(x)\} = \left(1+x^2\right)^{(1+\epsilon)/2}\left(1+x^2/\nu\right)^{-(\nu+1)/2}$

so after a few calculations

$$\frac{|r'(x)|}{r^2(x)} = |x| \left(1+x^2\right)^{-(3+\epsilon)/2} \left(1+x^2/\nu\right)^{(\nu-1)/2}.$$

Furthermore, $k(x) = \int_x^{+\infty} (g(y,+1)+g(y,-1)) \exp\{-U(y)\} dy = 4\nu/(\nu-1) \left(1+x^2/\nu\right)^{(\nu-1)/2}$ *and after a few calculations, we get*

$$\int_{\mathbb{R}} \frac{|r'(x)|}{r^2(x)} k(x) dx = 16 \int_0^{+\infty} \frac{\nu}{(\nu-1)^2} \left(1+x^2/\nu\right)^{-\nu/2+1/2} \left(1+x^2\right)^{-(\epsilon+3)/2} x \left| (1+\epsilon)(1+\frac{x^2}{\nu}) - \frac{\nu+1}{\nu}(1+x^2) \right| dx.$$

Also, simple calculations give for $\epsilon \in (0, 1/\nu]$

$$\int_{\mathbb{R}} |r'(x)| dx = 2$$

and for $\epsilon \in (1/\nu, \nu)$

$$\int_{\mathbb{R}} |r'(x)| dx = 2 \left(2 \left(\frac{\nu^2-1}{\nu(\nu-\epsilon)} \right)^{-(\nu+1)/2} \left(1 + \frac{\epsilon\nu - 1}{\nu - \epsilon} \right)^{(1+\epsilon)/2} - 1 \right)$$

and overall, for $\epsilon \in (0, 1/\nu]$

$$J[r] = 32 \int_0^{+\infty} \frac{\nu^2}{(\nu-1)^2} \left(1+x^2/\nu\right)^{-\nu/2+1/2} \left(1+x^2\right)^{-(\epsilon+3)/2} x \cdot \left| (1+\epsilon)(1+\frac{x^2}{\nu}) - \frac{\nu+1}{\nu}(1+x^2) \right| dx$$

and for $\epsilon \in [1/\nu, \nu)$

$$\begin{aligned} J[r] =& 32 \left(2 \left(\frac{\nu^2-1}{\nu(\nu-\epsilon)} \right)^{-(\nu+1)/2} \left(1 + \frac{\epsilon\nu - 1}{\nu - \epsilon} \right)^{(1+\epsilon)/2} - 1 \right) \\ & \int_0^{+\infty} \frac{\nu^2}{(\nu-1)^2} \left(1+x^2/\nu\right)^{-\nu/2+1/2} \left(1+x^2\right)^{-(\epsilon+3)/2} x \cdot \\ & \left| (1+\epsilon)(1+\frac{x^2}{\nu}) - \frac{\nu+1}{\nu}(1+x^2) \right| dx. \end{aligned}$$

The integral on the RHS can be numerically approximated.

In Tables 5.1 and 5.2 we present some examples, comparing the efficiency of different algorithms. As target distribution we consider a Normal with mean zero and vari-

ance one, an exponential with parameter one symmetrically extended to the negative real numbers and student distributions with $\nu = 1, 2, 10, 100$ degrees of freedom. For each of these densities, except for the Cauchy distribution ($t(1)$), we are estimating the expectation of the distribution, so the function is $g(x) = x$, whereas for the Cauchy distribution, since the expectation does not exist, we are estimating the expectation of the function $g(x) = sign(x)\log(1+|x|)$. As regards the speed function, we use either $s(x) = \left(1+x^2\right)^{(1+\epsilon)/2}$ or $s(x) = \max\{1, |x|^{1+\epsilon}\}$ for $\epsilon \in [0, 1)$ and we compare the efficiencies over different values of ϵ. In order to numerically estimate the integrals arising in the definition of $J[r]$ we use the `integrate` function of `R`. We should emphasize that since we do not take into account any normalisation constant and since in the case of Cauchy distribution we are estimating a different observable, the comparison in Tables 5.1 and 5.2 should only be made column-wise (i.e. for a given distribution compare different algorithms).

For the first choice of speed function, $s(x) = \left(1+x^2\right)^{(1+\epsilon)/2}$, on the light tails (normal and exponential) distributions, the algorithm with optimal performance seems to be the one when ϵ is close to the value of 0.5. The efficiency function values seem to be "quadratic" with respect to ϵ. On the other hand, for Cauchy distribution the performance of the algorithm tends to get better as we increase the values of ϵ. As the degrees of freedom increase and the target distribution tends to approximate a normal, the "quadratic" behaviour with respect to ϵ seems to arise.

For the second choice of speed function, $s(x) = \max\{1, |x|^{1+\epsilon}\}$, the "quadratic" behaviour with respect to ϵ seems to take place in all distributions except for the student with two degrees of freedom. This quadratic behaviour is verified for Normal, Exponential and Cauchy distributions in simulations which we present in later section. There, we see that the algorithm performs better when $\epsilon = 0.5$.

It is interesting to observe that for any target distribution and any form of speed function, any of the SUZZ algorithms provides better results than the normal ZZ algorithm.

5.7.1 Proof of Proposition 5.7.2

Proof of Proposition 5.7.2. The existence of ϕ such that $\phi(x,\theta) \leq c_0(V(x,\theta)+1)$ and such that $\mathcal{L}\phi(x,\theta) = -g(x,\theta)$ is guaranteed by Assumption 5.7.1. Let

$$M_t = \phi(Z_t) - \phi(Z_0) + \int_0^t g(Z_s)ds$$

which is a local martingale by Dynkin's formula. Assume that Z starts from the invariant measure μ. Under $\mathbb{P}_\mu$, M_t is also a martingale since for any $t > 0$ and any

	Algorithmic Inverse Efficiency					
Algorithms	Normal	Exponential	$t(1)$	$t(2)$	$t(10)$	$t(100)$
Zig-Zag	4	20	$+\infty$	51025.26	31.04	16.1935
SUZZ(0)	1.2454	1.2454	23.1865	61.2736	9.4975	5.0451
SUZZ(0.1)	1.0732	1.0732	17.3233	50.1058	8.1268	4.3450
SUZZ(0.2)	0.9517	0.9517	13.0671	40.9789	6.9797	3.8502
SUZZ(0.3)	0.8714	0.8714	7.4232	33.4034	6.1353	3.5220
SUZZ(0.4)	0.8256	0.8256	5.4917	27.1874	5.5384	3.3334
SUZZ(0.5)	**0.8097**	**0.8097**	3.9415	22.4808	5.1503	**3.2658**
SUZZ(0.6)	0.8208	0.8208	2.6760	18.9692	4.9439	3.3068
SUZZ(0.7)	0.8567	0.8567	1.6760	16.4042	**4.9008**	3.4484
SUZZ(0.8)	0.9166	0.9166	1.6275	14.6148	5.0093	3.6865
SUZZ(0.9)	0.9998	0.9998	**0.7474**	**13.4862**	5.2632	4.0182

Table 5.1: *J values over various algorithms for different target distributions ,* $s(x) = \left(1+x^2\right)^{(1+\epsilon)/2}$*; Smallest value for every column in bold*

	Algorithmic Inverse Efficiency					
Algorithms	Normal	Exponential	$t(1)$	$t(2)$	$t(10)$	$t(100)$
Zig-Zag	4	20	$+\infty$	51025.26	9.4975	5.3070
SUZZ(0)	0.2972	8.2279	25.2317	17.9439	3.7677	2.4377
SUZZ(0.1)	0.2556	7.6968	19.6768	15.4059	3.5475	2.3287
SUZZ(0.2)	0.2255	7.3666	16.0537	13.5510	3.3969	2.2599
SUZZ(0.3)	0.2047	7.2039	13.7565	12.2140	3.3054	2.2250
SUZZ(0.4)	0.1918	**7.1839**	12.4349	11.2829	**3.2654**	**2.2197**
SUZZ(0.5)	0.1856	7.2890	**11.8899**	10.6805	3.2711	2.2406
SUZZ(0.6)	**0.1854**	7.5070	12.0240	10.3540	3.3185	2.2855
SUZZ(0.7)	0.1904	7.8300	12.8231	**10.2708**	3.4048	2.3526
SUZZ(0.8)	0.2002	8.2542	14.3705	10.3934	3.5282	2.4409
SUZZ(0.9)	0.2144	8.7791	16.9353	10.7219	3.6878	2.5514

Table 5.2: *J values over various algorithms for different target distributions ,* $s(x) = \max\{1, |x|^{1+\epsilon}\}$*; Smallest value for every column in bold*

$s \leq t$ we have

$$\mathbb{E}_\mu[M_s] \leq \mathbb{E}_\mu[|\phi(Z_s)|] + \mathbb{E}_\mu[|\phi(Z_0)|] + \int_0^s \mathbb{E}_\mu\left[|-\mathcal{L}\phi(Z_s)|\right] ds \leq 2\mathbb{E}_\mu[\phi] + t\mathbb{E}_\mu[g] < \infty.$$

Furthermore M_t has stationary increments. From Theorem 2.1 of [KLO12] we have

$$\frac{M_t}{\sqrt{t}} \xrightarrow{t\to\infty} N(0, \mathbb{E}[M_1^2])$$

in distribution under $\mathbb{P}_\mu$. Also, under $\mathbb{P}_\mu$

$$\frac{\phi(Z_t) - \phi(Z_0)}{\sqrt{t}} \xrightarrow{t\to+\infty} 0$$

since $\phi(Z_t)$ has the same law as $\phi(Z_0)$ and $\mathbb{E}_\mu[\phi[Z_0]] = \mathbb{E}_\mu[\phi[Z_0]] < \infty$. Therefore

$$\frac{1}{\sqrt{T}} \int_0^T g(Z_s) ds \xrightarrow{T\to\infty} N(0, \mathbb{E}_{Z_0\sim\mu}[M_1^2])$$

under $\mathbb{P}_\mu$. It suffices to prove that $\mathbb{E}_{Z_0\sim\mu}[M_1^2]$ admits the expression in (5.40). Let K_t be the number of switches before time t and let $T_1, T_2, ...$ the times of the switches. We write

$$\begin{aligned}
M_t &= \phi(Z_t) - \phi(Z_0) - \int_0^t \mathcal{L}\phi(Z_s) ds = \\
&= \int_0^t \Theta_s s(X_s)\phi'(Z_s) ds + \sum_{i=1}^{K_t} \phi(Z(T_i)) - \phi(Z(T_i^-)) - \\
&- \int_0^t \Theta_s s(X_s)\phi(Z_s) + \lambda(Z_s)(\phi(X_s, \Theta_s) - \phi(X_s, -\Theta_s) ds = \\
&= \sum_{i=1}^{K_t} \phi(X_{T_i}, \Theta_{T_i}) - \phi(X_{T_i}, -\Theta_{T_i}) + \int_0^t \lambda(Z_s)(\phi(X_s, \Theta_s) - \phi(X_s, -\Theta_s)) ds
\end{aligned}$$

and therefore (see [Kal02], Theorem 23.6) the predictable quadratic variation of M is

$$\langle M\rangle_t = \int_0^t \lambda(Z_t)\left(\phi(X_s, \Theta_s) - \phi(X_s, -\Theta_s)\right)^2 = 4\int_0^t \lambda(Z_t)\left(\psi(X_s)\right)^2 \, ds$$

where $\psi(x) = \frac{1}{2}\left(\phi(x, +1) - \phi(x, -1)\right)$. Therefore, from the definition of predictable quadratic variation, under $\mathbb{P}_\mu$,

$$\gamma_g^2 = \mathbb{E}_{Z_0\sim\mu}[M_1^2] = \mathbb{E}_{Z_0\sim\mu}\langle M\rangle_1 = 4\int_E \lambda(x,\theta)\psi^2(x) d\mu(x,\theta). \tag{5.44}$$

It remains to write ψ in terms of g, this can be done since for all x, θ $\mathcal{L}\phi(x,\theta) = -g(x,\theta)$ and therefore

$$\theta s(x)\phi'(x,\theta) + \lambda(x,\theta)(\phi(x,-\theta) - \phi(x,\theta)) = -g(x,\theta).$$

Writing down the two equations for $\theta = \pm 1$ and adding them up we get

$$s(x)(\phi'(x,+1)-\phi(x,-1))-(\lambda(x,+1)-\lambda(x,-1))(\phi(x,+1)-\phi(x,-1)) = -(g(x,+1)+g(x,-1))$$

and therefore

$$s(x)\psi'(x) - (s(x)U'(x) - s'(x))\psi(x) = -\frac{g(x,+1) + g(x,-1)}{2}.$$

Solving this first order linear ODE we get

$$\psi(x) = \frac{1}{2\exp\{-U(x)\}s(x)} \int_x^{+\infty} (g(y,+1) + g(y,-1)) \exp\{-U(y)\} dy$$

which when combined with (5.44) gives (5.40). □

5.8 Simulations

In this section we will present some computational results that aim to highlight the behaviour of SUZZ and compare it with ZZ. We will present results in one, two and five dimensions. As already suggested in Section 5.7, the one dimensional Speed Up Zig-Zag can vastly outperform the original Zig-Zag. However, it will be seen that there are significant advantages in using a speed function in higher dimensions as well.

5.8.1 One Dimension

We are concerned with five different versions of Speed Up Zig-Zag. We shall be referring to the SUZZ process with speed

$$s(x) = \max\left\{1, |x|^{1+\epsilon}\right\} \tag{5.45}$$

as SUZZ(ϵ) and we will run simulations for different values of ϵ. More specifically, ϵ will take the values $0, 0.1, 0.5, 0.9$. Note that $\epsilon = 0$ induces non-explosive deterministic dynamics, whereas positive values of ϵ induce explosive ones.

We are targeting standard Normal distributions, exponential distributions

extended to the negative real line symmetrically, i.e. with density

$$f(x) = \frac{1}{2}\exp\{-|x|\} \tag{5.46}$$

and Cauchy distributions with density $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. For every realisation of a process we created a skeleton with step size $\delta = 0.1$.

All the simulated values were produced using `MATLAB`. To analyse the performance of the algorithms, various diagnostic tests found in [CC96] have been used, namely Effective Sample Size (ESS), Geweke, Gelman-Rubin, Raftery, Traceplots and QQPlots have been used. All these diagnostics have been computed using `coda` of `R`.

For the Effective Sample Size we have sampled 22 independent chains and let them ran for time $T = 10^4$. For the Normal and the Exponential the ESS is for estimating the expectation of the distribution. Since the Cauchy does not have a finite expectation, we first transformed the samples via the transformation $f(x) = sign(x)\log(1+|x|)$, which is smooth, and then we calculated the ESS when estimating the expectation of $f(Z)$ where Z follows a Cauchy. In Table 5.3, we report the average ESS over 22 independent realisations of the process and the standard deviation in parenthesis.

For all three distributions, the algorithm that maximizes the quantity $ESS/Switch$ (also highlighted in bold) is the SUZZ with speed function $s(x) = \max\left\{1, |x|^{1.5}\right\}$, i.e. $\epsilon = 0.5$. The same conclusion is reached if we use ESS per Likelihood Evaluation as measure of efficiency. However, since the likelihood evaluations depend on the way we do Poisson Thinning, which in turn depends on the programmers ability, we tend to use ESS per Switch as the main measure of Efficiency, also discussed in Section 5.7.

Next we run 22 processes for Cauchy, Exponential and Normal until time $T = 10^4$ and we use Geweke method to test the H_0 hypothesis that the first 10% of the process has the same expectation with the last 50%. This is done via a z-test and we reject at a 5% significant level. Here we have once again applied the transformation $f(x) = sign(x)\log(1+|x|)$ to the samples targeting the Cauchy distribution. The number of rejections out of 22 tests are presented in Table 5.4.

Then, we present results based on Gelman-Rubin diagnostic, for which we run 16 chains until time $T = 2 \cdot 10^3$ and we take the δ-skeleton for $\delta = 0.1$, after we transformed the samples from the Cauchy distribution so that they target a normal. Such a transform can be for example $f(x) = \Phi^{-1}\left(\frac{\arctan(x)+\frac{\pi}{2}}{\pi}\right)$, where Φ is the CDF of the standard normal distribution. This is because if X is a Cauchy

Cauchy target

Algorithms	Lik Eval	Switches	Effective Sample Size	ESS/Switch	ESS/Lik Eval
Zig-ZaZ	149970	3161	317.26 (180.2)	0.100	$2.1 \cdot 10^{-3}$
SUZZ(0)	261021	3185	2022.26 (147.3)	0.635	$7.7 \cdot 10^{-3}$
SUZZ(0.1)	261356	3198	2584.26 (153.1)	0.808	$9.9 \cdot 10^{-3}$
SUZZ(0.5)	290236	3672	4924.30 (455.5)	**1.341**	$17 \cdot 10^{-3}$
SUZZ(0.9)	610997	5173	4682.15 (193.3)	0.905	$7.7 \cdot 10^{-3}$

Double Exponential target

Algorithms	Lik Eval	Switches	Effective Sample Size	ESS/Switch	ESS/Lik Eval
Zig-Zag	200590	4980	2025.87 (173.23)	0.4068	10^{-2}
SUZZ(0)	309030	4990	4926.24 (338.87)	0.9872	$1.5 \cdot 10^{-2}$
SUZZ(0.1)	310810	5017	5257.22 (277.38)	1.0479	$1.7 \cdot 10^{-2}$
SUZZ(0.5)	317880	5420	5864.86 (291.03)	**1.0821**	$1.8 \cdot 10^{-2}$
SUZZ(0.9)	365530	6352	5519.48 (338.44)	0.8690	$1.5 \cdot 10^{-2}$

Normal target

Algorithms	Lik Eval	Switches	Effective Sample Size	ESS/Switch	ESS/Lik Eval
Zig-Zag	190150	3993	6356.1 (331.44)	1.61	0.0334
SUZZ(0)	298140	3988	10617.0 (752.6)	2.66	0.0356
SUZZ(0.1)	299470	3993	11354.92 (702.1)	2.84	0.0379
SUZZ(0.5)	329940	4266	12547.3 (986.6)	**2.93**	0.0380
SUZZ(0.9)	344380	4842	11986.72 (893.4)	2.48	0.0348

Table 5.3: *Five different algorithms (original Zig-Zag and four speed-up algorithms) targeting Cauchy, Double Exponential and Normal distributions. The four different SUZZ algorithms correspond to four different speed functions, all of the form of (5.45), for different choice of ϵ. Each algorithm is denoted as $SUZZ(\epsilon)$, where ϵ is the parameter value of (5.45). Each process was simulated 22 times independently, each one until time $T = 10^4$ and the estimator constructed using the δ-skeleton of the process, where $\delta = 0.1$. For each algorithm, we present the average ESS over the 22 realisation, along with the standard deviation in a parenthesis. We also present the number of switches of direction of each algorithm along with the likelihood evaluations, occurred in our code where we use constant bounds for the Poisson thinning. As a main comparison tool we use ESS/Switch, with the bold letter indicating the best performance and we accompany these with the ESS per likelihood evaluation results.*

		Cauchy			
Algorithms	Original	SUZZ(0)	SUZZ(0.1)	SUZZ(0.5)	SUZZ(0.9)
Rejections	0/22	1/22	2/22	0/22	2/22
		Exponential			
Algorithms	Original	SUZZ(0)	SUZZ(0.1)	SUZZ(0.5)	SUZZ(0.9)
Rejections	1/22	1/22	1/22	0/22	0/22
		Normal			
Algorithms	Original	SUZZ(0)	SUZZ(0.1)	SUZZ(0.5)	SUZZ(0.9)
Rejections	0/22	0/22	1/22	0/22	2/22

Table 5.4: *Geweke Diagnostics over 22 simulated processes targeting Cauchy, Double Exponential and Normal. $SUZZ(\epsilon)$, denotes the SUZZ process with speed function as in (5.45) and ϵ the parameter introduced in that equation. All algorithms were simulated, independently, 22 times, each until time $T = 10^4$. For each algorithm, we present the number of realisations, out of total 22, that did not pass the z-test, testing whether the first* 10% *of the process has the same expectation as the final* 50%.

then $\arctan(X) \sim unif(-\frac{\pi}{2}, \frac{\pi}{2})$ and if $U \sim unif(0,1)$ then $\Phi^{-1}(U)$ is a standard normal. The table contains a point estimator for the Gelman-Rubin diagnostic and an upper confidence interval for this estimator. We also present the first time when the upper confidence interval becomes less than 1.01 and stays less than 1.01 in all subsequent time periods. The results are presented in Table 5.5.

Furthermore, we calculated a Raftery diagnostic to check how many iterations are needed for the process to estimate various percentiles of the distributions. For Cauchy, double Exponential and Normal distributions we study quantiles for the values $0.025, 0.01, 0.001, 0.975, 0.99, 0.999$, approximated to an error of $r = 0.005$, with probability 0.95. For all three targets, we present the average over 22 chains and standard deviation in a parenthesis for the following quantities: Number of Iterations to estimate quantile , Dependence factor and Burn in period. The results are presented on Tables 5.6 , 5.7 and 5.8.

Finally, we present some Traceplots and QQPlots for one chain for each one of the five algorithms studied above. All the algorithms target the Cauchy distribution. The Traceplots are ran until time $T = 10^4$ while the QQPlots are ran until the number of switches becomes $N = 10^4$. It can be observed from the Traceplots in Figure 5.2 that the SUZZ algorithms explore the space faster as they have these regular long excursions from which they return to the mode fast enough. The better performance of the SUZZ algorithms is even more visible on the QQPlots in Figure 5.3, where SUZZ(0.5) and SUZZ(0.9) are clearly performing significantly better that the other algorithms.

The diagnostic tests show that all the Speed Up algorithms used outperform

	Cauchy		
Algorithms	Point Estimator	Upper CI	Time Upper CL < 1.01
Original	1.03	1.05	$> 2 \cdot 10^3$
SUZZ(0)	1.0008	1.001	700
SUZZ(0.1)	1.0009	1.002	200
SUZZ(0.5)	0.9999	1.0003	100
SUZZ(0.9)	1.0002	1.0008	200
	Exponential		
Algorithms	Point Estimator	Upper CI	Time Upper CL < 1.01
Original	1.03	1.05	$> 2 \cdot 10^3$
SUZZ(0)	1.0008	1.001	700
SUZZ(0.1)	1.0009	1.002	200
SUZZ(0.5)	0.9999	1.0003	100
SUZZ(0.9)	1.0002	1.0008	200
	Normal		
Algorithms	Point Estimator	Upper CI	Time Upper CL < 1.01
Original	0.9999	1.0004	100
SUZZ(0)	0.9998	1.0001	100
SUZZ(0.1)	0.9997	0.9998	50
SUZZ(0.5)	0.9998	1.0001	50
SUZZ(0.9)	0.9999	1.0002	50

Table 5.5: *Gelman-Rubin Diagnostics for processes targeting Cauchy, Double Exponential and Normal.* $SUZZ(\epsilon)$*, denotes the SUZZ process with speed function as in (5.45) and* ϵ *the parameter introduced in that equation. For this diagnostic* 16 *realisations of each algorithm were simulated, starting from over-dispersed starting positions. Each realisation was simulated until time* $T = 2 \cdot 10^3$*. For each algorithm, we present the point estimator and the upper level of the confidence interval of Gelman-Rubin. We also present the first time the upper level of the confidence interval took a value less than* 1.01*.*

Number of Iterations						
Algorithms	0.025	0.01	0.001	0.975	0.99	0.999
Original	446643 (580920)	141842 (102571)	5767 (14135)	209168 (3746)	77727 (60930)	2742.4 (619)
SUZZ(0)	14928 (1444)	5972 (729)	826 (765)	15528 (1410)	6164 (836)	586 (240)
SUZZ(0.1)	11305 (943)	4282 (462)	383 (186)	11295 (785)	4411 (498)	332 (91.08)
SUZZ(0.5)	3640 (943)	1493 (0)	156 (0)	3632 (49.39)	1493 (0)	156 (0)
SUZZ(0.9)	4452 (1466)	1542 (43.48)	156 (0)	4876 (1991)	1549 (40.25)	156 (0)
IID	3746	1522	154	3746	1522	154
Dependence Factor						
Algorithms	0.025	0.01	0.001	0.975	0.99	0.999
Original	119.27 (155.11)	93.11 (67.25)	37.48 (91.76)	55.85 (101.16)	51.06 (39.97)	17.85 (4.04)
SUZZ(0)	3.9 (0.39)	3.92 (0.48)	5.37 (4.98)	4.15 (0.38)	4.05 (0.55)	3.80 (1.55)
SUZZ(0.1)	3.02 (0.25)	2.81 (0.30)	2.49 (1.20)	3.02 (0.21)	2.90 (0.33)	2.16 (0.59)
SUZZ(0.5)	0.97 (0.02)	0.981 (0)	1.01 (0)	0.97 (0.01)	0.98 (0)	1.01 (0)
SUZZ(0.9)	1.19 (0.39)	1.01 (0.03)	1.01 (0)	1.30 (0.53)	1.02 (0.03)	1.01 (0)
Burn In						
Algorithms	0.025	0.01	0.001	0.975	0.99	0.999
Original	409 (533.74)	314.09 (228.99)	3194 (14699)	192.50 (347.76)	176.40 (137.77)	60.73 (13.72)
SUZZ(0)	14 (1.34)	13.68 (1.73)	18.18 (17.00)	14.50 (1.37)	14.10 (2.09)	12.86 (5.42)
SUZZ(0.1)	10.45 (10.01)	9.73 (1.16)	8.23 (4.37)	10.45 (0.74)	10.00 (1.20)	7.14 (2.10)
SUZZ(0.5)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)
SUZZ(0.9)	2.68 (1.39)	1.73 (0.55)	2 (0)	2.91 (1.72)	1.64 (0.58)	2 (0)

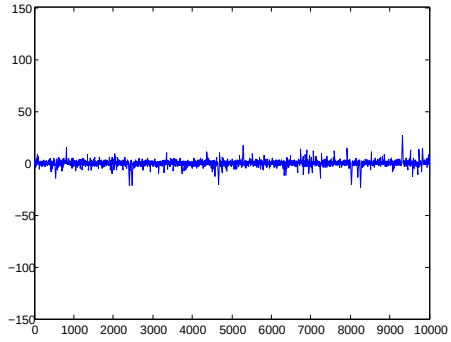
Table 5.6: *Raftery Diagnostics for five different algorithms targeting* **Cauchy** *distribution. All the processes were simulated 22 times, independently, until time* $T = 10^4$*. As estimators we use the* δ*-skeleton for* $\delta = 0.1$*. We estimate six different quantiles of the distribution,* $0.025, 0.01, 0.001, 0.975, 0.99, 0.999$*. We present the average number of iterations needed to approximate the quantiles to an error of* 0.005 *with probability* 0.95 *and we also present the standard deviation in a parenthesis. We also present the number of i.i.d. observations from the target density are needed to estimate the same quantiles with the same accuracy. Furthermore, we present the average dependence factors (defined as the number of iterations of the algorithm divided by the number of i.i.d. iterations), and the average burn-in periods, with the standard deviations in parenthesis.*

Number of Iterations						
Algorithms	0.025	0.01	0.001	0.975	0.99	0.999
Original	15234.9 (1593.3)	6465.0 (1435.6)	746.7 (366.2)	14573.6 (903.2)	6004.0 (793.4)	583.9 (313.9)
SUZZ(0)	5557.4 (1175.4)	1730.9 (67.0)	159.2 (10.3)	5098.5 (836.6)	1750.8 (62.8)	165.5 (16.0)
SUZZ(0.1)	4335.1 (132.4)	1607.0 (56.7)	156.0 (0)	4313.1 (175.7)	1614.4 (55.9)	157.6 (7.5)
SUZZ(0.5)	3568.6 (17.7)	1491.1 (4.9)	156.0 (0)	3579.9 (25.6)	1491.1 (4.9)	156 (0)
SUZZ(0.9)	3863.1 (95.7)	1525.2 (31.2)	156.0 (0)	3922.6 (113.7)	1553.5 (49.0)	156 (0)
IID	3746	1522	154	3746	1522	154
Dependence Factor						
Algorithms	0.025	0.01	0.001	0.975	0.99	0.999
Original	4.07 (0.42)	4.25 (0.94)	4.85 (2.38)	3.89 (0.24)	3.95 (0.52)	3.79 (2.04)
SUZZ(0)	1.48 (0.31)	1.14 (0.04)	1.031 (0.07)	1.36 (0.22)	1.15 (0.04)	1.07 (0.10)
SUZZ(0.1)	1.16 (0.04)	1.06 (0.04)	1.01 (0)	1.15 (0.05	1.06 (0.04)	1.02 (0.05)
SUZZ(0.5)	0.95 (0.01)	0.98 (0)	1.01 (0)	0.96 (0.01)	0.98 (0.00)	1.01 (0)
SUZZ(0.9)	1.03 (0.03)	1.03 (0.03)	1.01 (0)	1.05 (0.03)	1.02 (0.03)	1.01 (0)
Burn In						
Algorithms	0.025	0.01	0.001	0.975	0.99	0.999
Original	14.09 (1.41)	14.91 (3.21)	16.46 (8.15)	13.54 (0.91)	13.73 (1.72)	12.73 (7.05)
SUZZ(0)	3.95 (0.21)	2.95 (0.38)	2.09 (0.29)	3.91 (0.29)	3.05 (0.38)	2.27 (0.46)
SUZZ(0.1)	3.09 (0.29)	2.32 (0.57)	2 (0)	3.00 (0.31)	2.27 (0.55)	2.05 (0.21)
SUZZ(0.5)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)	2(0)
SUZZ(0.9)	2.05 (0.21)	1.55 (0.51)	2 (0)	2.27 (0.46)	1.86 (0.56)	2 (0)

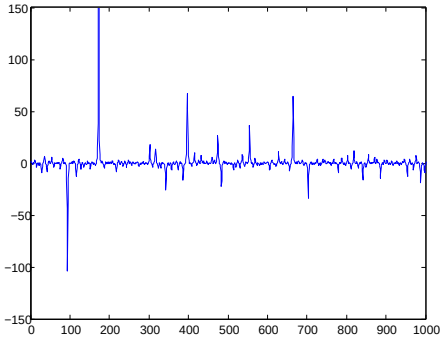
Table 5.7: *Raftery Diagnostics for five different algorithms targeting a* **double Exponential** *distribution, with density as in (5.46). All the processes were simulated 22 times, independently, until time* $T = 10^4$*. As estimators we use the* δ*-skeleton for* $\delta = 0.1$*. We estimate six different quantiles of the distribution,* $0.025, 0.01, 0.001, 0.975, 0.99, 0.999$*. We present the average number of iterations needed to approximate the quantiles to an error of* 0.005 *with probability* 0.95 *and we also present the standard deviation in a parenthesis. We also present the number of i.i.d. observations from the target density are needed to estimate the same quantiles with the same accuracy. Furthermore, we present the average dependence factors (defined as the number of iterations of the algorithm divided by the number of i.i.d. iterations), and the average burn-in periods, with the standard deviations in parenthesis.*

Number of Iterations needed						
Algorithms	0.025	0.01	0.001	0.975	0.99	0.999
Original	6444.9 (637.2)	2398.7 (190.1)	210.1 (43.5)	6298.1 (362.9)	2393.0 (205.9)	217.7 (33.6)
SUZZ(0)	3749.1 (92.7)	154.0 (27.1)	152.9 (6.7)	3738.8 (59.8)	1509.8 (27.6)	152.1 (7.3)
SUZZ(0.1)	3661.3 (47.8)	1498.7 (16.8)	152.9 (6.7)	3639.8 (55.0)	1494.6 (13.5)	152.9 (6.7)
SUZZ(0.5)	3570.9 (22.7)	1489.2 (6.4)	152.1 (7.3)	3578.9 (26.0)	1489.2 (6.4)	152.1 (7.3)
SUZZ(0.9)	3863.5 (74.9)	1508.6 (28.7)	156.9 (6.7)	3872.7 (91.0)	1524.0 (36.3)	156 (6.7)
IID	3746	1522	154	3746	1522	154
Dependence Factor						
Algorithms	0.025	0.01	0.001	0.975	0.99	0.999
Original	1.72 (0.17)	1.58 (0.13)	1.37 (0.10)	1.68 (0.10)	1.57 (0.14)	1.42 (0.99)
SUZZ(0)	1.00 (0.03)	0.99 (0.02)	0.99 (0.04)	1.00 (0.02)	0.99 (0.02)	0.99 (0.05)
SUZZ(0.1)	0.98 (0.01)	0.99 (0.01)	0.99 (0.05)	0.97 (0.04)	0.98 (0.01)	0.99 (0.04)
SUZZ(0.5)	0.95 (0.01)	0.98 (0.00)	0.99 (0.04)	0.96 (0.01)	0.98 (0.00)	0.99 (0.05)
SUZZ(0.9)	1.03 (0.02)	0.99 (0.02)	1.01 (0)	0.99 (0.04)	1.00 (0.02)	0.99 (0.04)
Burn In Period						
Algorithms	0.025	0.01	0.001	0.975	0.99	0.999
Original	5.55 (0.51)	5.23 (0.53)	4.09 (1.66)	5.59 (0.50)	5.09 (0.75)	4.41 (1.26)
SUZZ(0)	1.96 (0.21)	1.68 (0.48)	1.82 (0.40)	1.82 (0.40)	1.59 (0.50)	1.77 (0.43)
SUZZ(0.1)	1.96 (0.21)	1.68 (0.48)	1.82 (0.40)	2 (0)	1.82 (0.40)	1.82 (0.40)
SUZZ(0.5)	2 (0)	2 (0)	1.77 (0.43)	2 (0)	2 (0)	1.77 (0.43)
SUZZ(0.9)	1.96 (0.21)	1.86 (0.35)	1.82 (0.40)	2.09 (0.29)	1.59 (0.50)	1.82 (0.40)

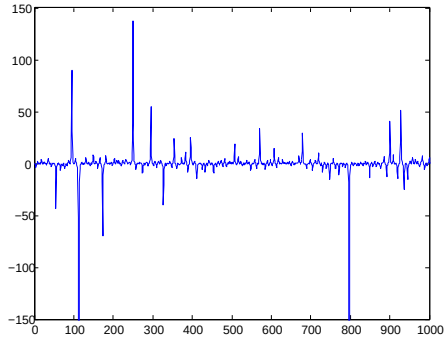
Table 5.8: *Raftery Diagnostics for five different algorithms targeting a* **Normal** *with* 0 *expectation and* 1 *variance. All the processes were simulated 22 times, independently, until time* $T = 10^4$*. As estimators we use the* δ*-skeleton for* $\delta = 0.1$*. We estimate six different quantiles of the distribution,* $0.025, 0.01, 0.001, 0.975, 0.99, 0.999$*. We present the average number of iterations needed to approximate the quantiles to an error of* 0.005 *with probability* 0.95 *and we also present the standard deviation in a parenthesis. We also present the number of i.i.d. observations from the target density, needed to estimate the same quantiles with the same accuracy. Furthermore, we present the average dependence factors (defined as the number of iterations of the algorithm divided by the number of i.i.d. iterations), and the average burn-in periods, with the standard deviations in parenthesis.*



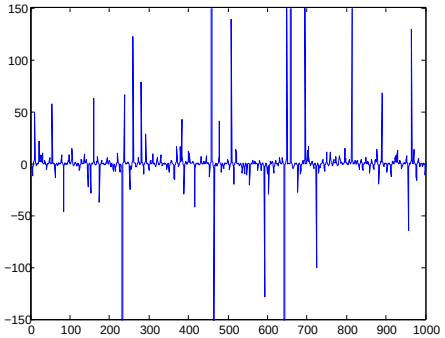
(a) ZZ



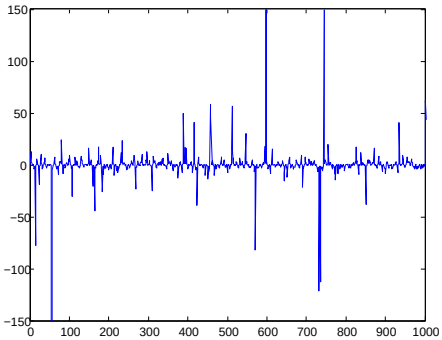
(b) SUZZ(0)



(c) SUZZ(0.1)

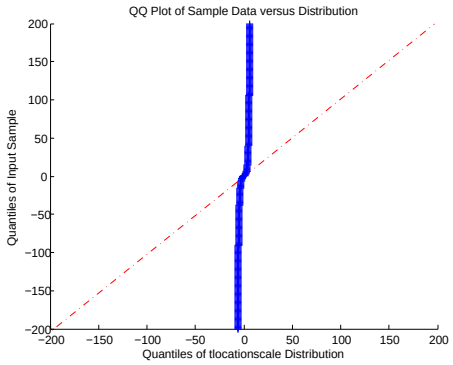


(d) SUZZ(0.5)

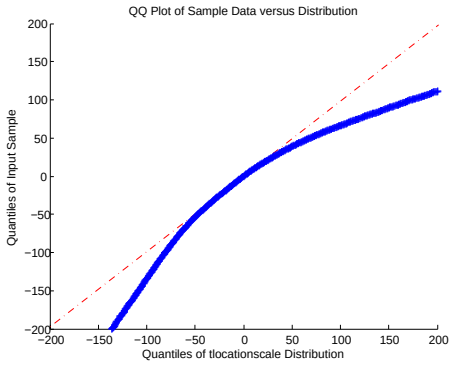


(e) SUZZ(0.9)

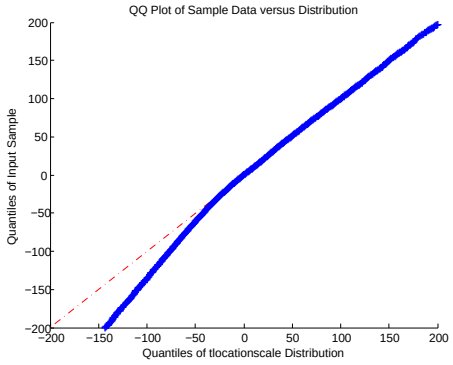
Figure 5.2: *Traceplots of five different one-dimensional algorithms targeting a Cauchy distribution. As $SUZZ(\epsilon)$ we denote the SUZZ algorithm with speed of the form (5.45) and ϵ the parameter appearing in the equation.*



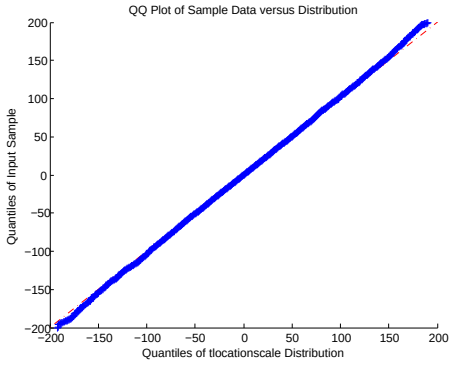
(a) ZZ



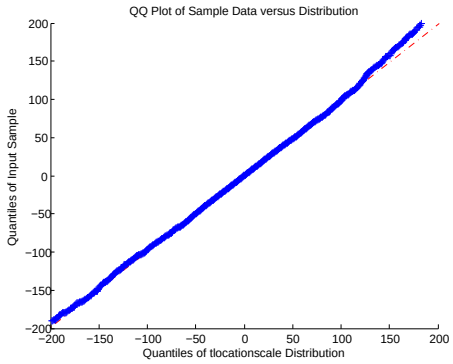
(b) SUZZ(0)



(c) SUZZ(0.1)



(d) SUZZ(0.5)



(e) SUZZ(0.9)

Figure 5.3: *QQ plots of various one-dimensional Speed Up algorithms targeting a Cauchy distribution. The algorithms have ran until* $N = 10^4$ *switches of direction have happened. The sample is created using the* δ*-skeleton of the process for* $\delta = 0.1$*. As* $SUZZ(\epsilon)$ *we denote the SUZZ algorithm with speed of the form (5.45) and* ϵ *the parameter appearing in the equation.*

the Zig-Zag algorithm, not only on the Cauchy distribution, but on the Gaussian and the Exponential as well. This could suggest that SUZZ algorithms can be used outside the regime of heavy tails as well. It is also interesting to observe the $SUZZ(0.5)$ is consistently performing better than all the other algorithms through every test, which is in accordance with Table 5.2 in the previous Section.

5.8.2 2 Dimensions

Here we simulate from the SUZZ on two dimensional targets. We will use the speed function

$$s(x) = \sqrt{1 + \|x\|^2} \tag{5.47}$$

which leads to deterministic dynamics. One can find a closed formula for the deterministic dynamics this speed induces. Indeed, if the process starts from (x, θ), $x = (x_1, x_2)$, $\theta = (\theta_1, \theta_2)$, then at time t is has position $X_t = (X_t^{(1)}, X_t^{(2)})$ given by the ODE system

$$\begin{cases} \dfrac{dX_t^{(1)}}{dt} = \theta_1 s(X_t) \\ \dfrac{dX_t^{(2)}}{dt} = \theta_2 s(X_t) \\ X_0 = x \end{cases} . \tag{5.48}$$

To solve this, note that $\{X_t, t \geq 0\}$ is a straight line in $\mathbb{R}^2$ so

$$X_t^{(2)} = x_2 + \theta_1\theta_2(X_t^{(2)} - x_1) = y_2 + \theta_1\theta_2 X_t^{(1)}$$

where

$$y_2 = x_2 - \theta_1\theta_2 x_1.$$

Therefore we need to solve

$$\frac{dX_t^{(1)}}{dt} = \theta_1 s(X_t^{(1)}, y_2 + \theta_1\theta_2 X_t^{(1)})$$

Separating the variables we get

$$\theta_1 t = \int_{x_1}^{X_t^{(1)}} \frac{1}{s(u, y_2 + \theta_1\theta_2 u)} du = \int_{x_1}^{X_t^{(1)}} \frac{1}{\sqrt{1 + u^2 + (y_2 + \theta_1\theta_2 u)^2}} du. \tag{5.49}$$

For the integral on the RHS we write

$$\int_{x_1}^{X_t^{(1)}} \frac{1}{\sqrt{1+u^2+(y_2+\theta_1\theta_2 u)^2}}du = \frac{1}{\sqrt{2}}\int_{x_1}^{X_t^{(1)}} \frac{1}{\sqrt{(u+\frac{\theta_1\theta_2 y_2}{2})^2+\frac{1}{2}+\frac{y_2^2}{2}-(\frac{\theta_1\theta_2 y_2}{2})^2}}du$$

$$\frac{1}{\sqrt{2}}\int_{x_1}^{X_t^{(1)}} \frac{1}{\sqrt{(u+\frac{\theta_1\theta_2 y_2}{2})^2+\frac{1}{2}+\frac{y_2^2}{4}}}du = \frac{1}{\sqrt{2}}\log\left|\frac{X_t^{(1)}+\frac{\theta_1\theta_2 y_2}{2}+\sqrt{(X_t^{(1)}+\frac{\theta_1\theta_2 y_2}{2})^2+a}}{x_1+\frac{\theta_1\theta_2 y_2}{2}+\sqrt{(x_1+\frac{\theta_1\theta_2 y_2}{2})^2+a}}\right|$$

where

$$a = \frac{1}{2}+\frac{y_2^2}{4}.$$

If we write for simplicity $Y_t = X_t^{(1)} + \frac{\theta_1\theta_2 y_2}{2}$, so

$$Y_0 = x_1 + \frac{\theta_1\theta_2 y_2}{2}$$

then from (5.49) we get

$$\frac{Y_t+\sqrt{Y_t^2+a}}{Y_0+\sqrt{Y_0^2+a}} = \exp\{\sqrt{2}\theta_1 t\} \iff Y_t+\sqrt{Y_t^2+a} = (Y_0+\sqrt{Y_0^2+a})\exp\{\sqrt{2}\theta_1 t\}.$$

Write

$$b(t) = (Y_0+\sqrt{Y_0^2+a})\exp\{\sqrt{2}\theta_1 t\}$$

and solving for Y_t we get

$$Y_t = \pm\frac{b^2(t)-a}{2b(t)}.$$

Studying the monotonicity of Y_t and $\frac{b^2(t)-a}{2b(t)}$ we see that they are both increasing with respect to t when $\theta_1 = +1$ and both decreasing when $\theta_1 = -1$. Therefore

$$Y_t = \frac{b^2(t)-a}{2b(t)}$$

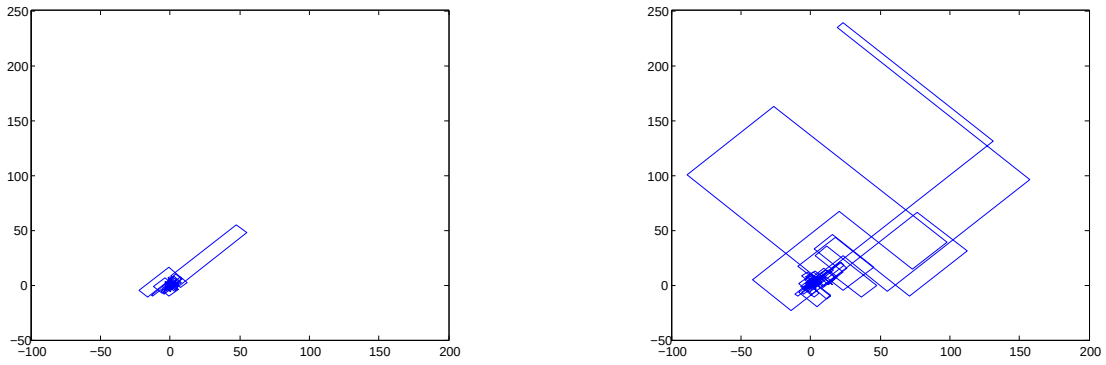
and therefore

$$X_t^{(1)} = \frac{b^2(t)-a}{2b(t)} - \frac{\theta_1\theta_2 y_2}{2}. \tag{5.50}$$

We are using Speed Up Zig-Zag with speed function

$$s(x) = \left(1+x_1^2+x_2^2\right)^{1/2}$$

and original Zig-Zag to target a two dimensional Cauchy distribution, with correla-



(a) ZZ

(b) SUZZ(0)

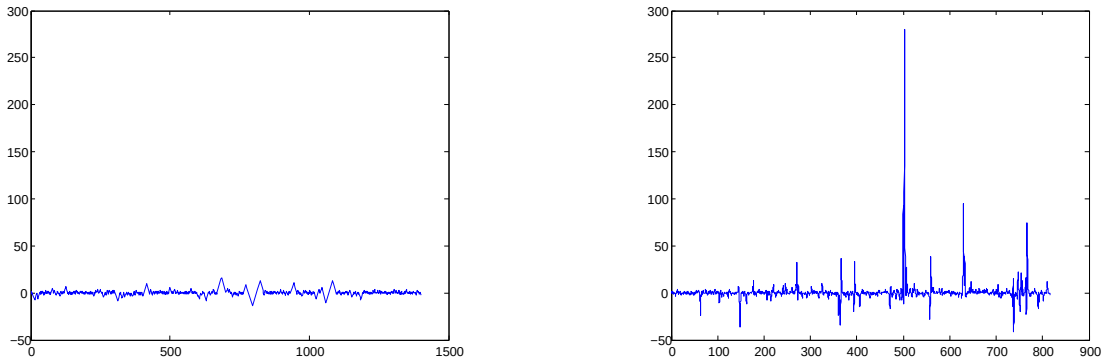
Figure 5.4: *Representation of the path for two algorithms, targeting a two-dimensional Cauchy with mode* $(0,0)$ *and covariance matrix composed by variances* 1 *in both coordinates and positive covariance* 0.5 *between the two coordinates. With SUZZ(0) we denote the two-dimensional SUZZ algorithm with speed given by (5.47). Each process was simulated until* $N = 10^3$ *number of switches.*

tion matrix $A = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, that is the density is of the form

$$\pi(x) = \frac{1}{Z}\left(1 + x'A^{-1}x\right)^{-3/2}.$$

We present and compare the trace plots of the first coordinate of the two processes in Figure 5.5 We further present and compare the plots where the two coordinates of the process are plotted against each other in Figure 5.4. This represents the path of the process, evolving with time. For these figures, we run both processes until the number of switches reaches $N = 10^3$. Is seems that the Speed Up process can explore better the areas away from the mode.

Furthermore, in Table 5.9 we present some estimates of the probabilities the target distribution assigns to four different squares. We consider the squares $[-1,1]^2, [-2,2]^2, [-10,10]^2$ and $[-20,30] \times [-50,40]$. The results for the two processes in Table 5.9. We, also, include the actual value of the probability of each square, which was estimated using `mvtnorm` of `R`. We consider two cases, where the processes have ran until N switches have occurred, where N can be either 10^3, or 10^4. It seems that in the second case, both processes have converged fairly well. On the other hand, in the first case, even though none of the processes provides a very reliable estimation, the SUZZ estimates are better than those of the orig-



(a) ZZ

(b) SUZZ(0)

Figure 5.5: *Traceplots for the first coordinates of two algorithms, targeting a two-dimensional Cauchy with mode* $(0,0)$ *and covariance matrix composed by variances* 1 *in both coordinates and positive covariance* 0.5 *between the two coordinates. With SUZZ(0) we denote the two-dimensional SUZZ algorithm with speed given by (5.47). Each process was simulated until* $N = 10^3$ *number of switches.*

inal Zig-Zag. Especially in the case of the square $[-20, 30] \times [-50, 40]$ the SUZZ gives a much more reliable estimation and we conjecture that for very rare events, SUZZ will provide substantially better probability estimates for the same number of switches.

	Switches $N = 10^3$			
Algorithms	$[-1,1]^2$	$[-2,2]^2$	$[-10,10]^2$	$[-20,30] \times [-50,40]$
ZZ	0.3755	0.6383	0.9264	1
SUZZ(0)	0.3312	0.5846	0.9254	0.9726
Actual	0.3505	0.6033	0.9134	0.9703
	Switches $N = 10^4$			
Algorithms	$[-1,1]^2$	$[-2,2]^2$	$[-10,10]^2$	$[-20,30] \times [-50,40]$
ZZ	0.3534	0.6027	0.9149	0.9698
SUZZ(0)	0.3458	0.5949	0.9109	0.9708
Actual	0.3505	0.6033	0.9134	0.9703

Table 5.9: *Estimation of probabilities assigned to various rectangles of* $\mathbb{R}^2$ *by the two-dimensional Cauchy distribution with covariance* 0.5 *and variance* 1*. The first Table shows the results of SUZZ and ZZ ran until* $N = 10^3$ *switches have occurred and the second until* $N = 10^4$*. The actual probabilities are also presented.*

5 Dimensional SUZZ on Cauchy

Algorithms	ESS(SD)	Median ESS	$l = 2.2577$	$l = 12.4788$	$l = 125.3256$	$l = 1325.867$
ZZ	1027.4 (1123.58)	676.3	0.4883	0.9019	1	1
SUZZ(0)	**4859.53** (717.45)	**4710.4**	**0.5026**	**0.9000**	**0.9900**	**0.9996**
Actual	-	-	0.5	0.9	0.99	0.999

Table 5.10: *SUZZ and ZZ algorithm targeting a five dimensional Cauchy distribution with minus log-likelihood given by (5.51). For SUZZ, we use* $s(x) = \sqrt{1 + x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2}$. *The algorithms ran until* $N = 10^4$ *switches occur and the average ESS (with Standard deviation in a parenthesis) along with the median ESS is presented. In both cases, the ESS is applied on the target transformed via the function (5.52). An estimation of probabilities assigned to various rectangles of* $\mathbb{R}^5$*, by the five-dimensional Cauchy distribution is, also, presented. The squares are of the form* $[-l, l]^5$ *for various values of* l*. For these estimations, the algorithms ran until* $N = 10^5$ *switches occurred. The actual probabilities are also presented.*

5.8.3 5 Dimensions

We finally, present some results on a SUZZ process targeting a five dimensional Cauchy distribution, without correlations. This means that

$$U(x) = 3\log(1 + x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2) \tag{5.51}$$

and we use speed function

$$s(x) = \sqrt{1 + x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2}.$$

We also ran an original Zig-Zag and we compare the results. We simulated 25 processes, independently, for each algorithm, until $N = 10^4$ number of switches occurred. As an estimator we use the δ-skeletons for $\delta = 0.1$. In Table 5.10, we present the average ESS and standard deviation in a parenthesis along with the median ESS (since the Standard Deviation seems to be large), after we transform the sample via the function

$$f(x) = sign(x)\log(1 + |x|), \tag{5.52}$$

so that the expectation is finite. We also ran one more realisation of the algorithms until $N = 10^5$ switches and we use this to estimate the probability the distribution assigns to some squares in $\mathbb{R}^5$. All the squares are of the form $[-l, l]^5$ for various values of l. We, also present the actual probability values, estimated using `mvtnorm` of `R`. We observe that the SUZZ process clearly outperforms the original Zig-Zag, both in ESS (and in variance of ESS) and in estimation of probabilities.

Chapter 6

Conclusion

In this work we were concerned with studying the Zig-Zag process, a process recently proposed as an MCMC method.

In Section 2 we summarised the existing literature around the area.

In Section 3 we proposed a generalisation of the Zig-Zag process that allows the process to move in more directions than just parallel to $\{-1, +1\}^d$. We call this process Multi Directional Zig-Zag and we managed to identify conditions on the rates of switching directions so that the process targets the distribution of interest. For a specific class of Multi Directional Zig-Zag, we proved ergodicity and geometric ergodicity and we provided simulations, indicating that, in difficult distribution settings such as some bimodal or banana shaped distributions, we can have computational advantages by using this process over simple Zig-Zag.

In Section 4 we studied the behaviour of the Zig-Zag process when targeting a heavy tailed distribution. We provided two different proofs that the process is not geometrically ergodic. Furthermore, we proved that in one dimension, this lack of geometric ergodicity cannot be solved by changing the process so that it randomly chooses a new velocity every time a refresh event occurs. Finally, we studied the convergence rate of the one dimensional Zig-Zag process in heavy tails and we proved that it converges polynomially fast in total variation under reasonable assumptions satisfied for example by Student distributions. When the target density decays like a polynomial, we can exactly identify the rate of convergence of the Zig-Zag process, which is better than the rate of convergence of state of the art MCMC algorithms like MALA or Random Walk Metropolis.

Finally, in Section 5 we proposed a solution to the problem of non geometric ergodicity of the Zig-Zag process in heavy tails. We proposed an extension of the Zig-Zag process that, given a function s, it allows the process to move with non

constant speed s, depending on its current position. This allows the process to have larger excursions towards the tails of the distribution and return to the centre of the mass fast enough. We, also, allowed the deterministic dynamics of the process to explode in finite time. We call this process Speed Up Zig-Zag. We proved that under some assumptions on the speed function the process is geometrically ergodic and that if we pick the switching rates appropriately, the process targets the distribution of interest. We further develop a quantity that describes the efficiency of the one dimensional algorithm and show that one gains a lot of computational efficiency when using Speed Up Zig-Zag, compared to normal Zig-Zag. This is also supported by simulations.

There are a lot of interesting areas where this work could lead, as the theory and methodology of PDMPs for MCMC is in a rather early stage. First of all, choosing the optimal speed function in the setting of high dimensional Speed Up Zig-Zag is not investigated yet. Although a suggestion was made in the thesis, we are currently working on this direction. At the same time, we are, also, working on developing a Speed Up version of the Bouncy Particle Sampler, another PDMP with a lot of potential.

The Multi Directional Zig-Zag could have a greater potential if the directions allowed can be chosen in the right way. We believe that an adapted version of the algorithm, that progressively learns what kind of directions to use and how much time it should spend in each direction, would be a very interesting project.

Finally, the polynomial rates of convergence of the Zig-Zag process in heavy tails are only studied in one dimension in this work. It would be interesting to get a higher dimension result for the process. Another direction for this work would be to study the polynomial ergodicity of the higher dimensional Bouncy Particle Sampler in heavy tails.

Main Notation Table

1. π: The target probability measure on $\mathbb{R}^d$. Also used for the density of that measure, page 1.
2. $\mathbb{E}_x$: the expectation when the process starts from $x \in E$, page 7.
3. $\|\cdot\|_{TV}$: Total variation distance, page 7.
4. μ: The target probability measure on the extended space E, pages 7, 20.
5. $\mathbb{P}_x$: the law of the process starting from $x \in E$, page 8.
6. V: Lyapunov function, pages 9, 12.
7. $(P^t)_{t\geq 0}$: Transition semi-group of the process, page 10.
8. $\mathcal{L}$: Strong or extended generator of the process, page 10.
9. $\mathcal{B}(E)$: Borel functions of E, page 11.
10. E: The state space of the process, typically of the form $\mathbb{R}^d \times \Theta_0^d$, pages 15, 19.
11. λ: The switching rates of the process, page 16.
12. ∂_i: The operator of the partial derivative of the i coordinate, page 17.
13. ∇A: The gradient of A, page 17.
14. $(Z_t)_{t\geq 0} = (X_t, \Theta_t)_{t\geq 0}$: The process on E, page 19.
15. $(X_t)_{t\geq 0}$: The space component of the Z process on $\mathbb{R}^d$, page 19.
16. $(\Theta_t)_{t\geq 0}$: The velocity component of the Z process on Θ_0^d, page 19.
17. F_i: The operator of flipping the i coordinate of the velocity, page 19.
18. θ_i: The i coordinate of velocity i, page 19.

19. θ: A velocity, an element of Θ, page 19.

20. U: the potential, the function such that the target probability measure on $\mathbb{R}^d$, π, satisfies $\pi(dx) = \frac{1}{Z}\exp\{-U(x)\}dx$ for some $Z > 0$, page 20.

21. a^+: The positive part operator, page 20.

22. γ: The refresh rate of switching, page 20.

23. $HessA$: The Hessian matrix of function A, page 28.

24. Θ_0: The set of available values for the coordinates of the velocity, page 41.

25. Θ: The set of available velocities, page 41.

26. θ^j: An element of Θ_0, page 41.

27. F_i^j: The operator of switching the i coordinate of the velocity to θ^j, page 42.

28. $\Phi_{x,\theta}(t)$: The flow of the deterministic dynamics starting from (x, θ) at time t pages 43, 154.

29. $\theta_0^{\leq}$, page 50.

30. $\theta_0^{<}$, page 50.

31. λ_i^+: The upwards rates of MDCNZZ, page 50.

32. λ_i^-: The downwards rates of MDCNZZ, page 50.

33. $\rightarrow$, page 52

34. $\not\rightarrow$, page 53

35. Leb: The Lebesgue measure on $\mathbb{R}^d$, page 67.

36. s: Speed function of the SUZZ process, page 154.

37. $t^*(x, \theta)$: The explosion time of the deterministic dynamics starting from (x, θ), page 154.

38. A_i page, page 158.

39. $(Z_t^m)_{t\geq 0} = (X_t^m, \Theta_t^m)_{t\geq 0}$: The SUZZ process, stopped upon hitting a ball of radium m, page 164.

Bibliography

[ADNR21] C. Andrieu, A. Durmus, N. Nüsken, and J. Roussel. Hypocoercivity of piecewise deterministic markov process-monte carlo. *To appear in Annals of Applied Probability*, 2021.

[AL19] C. Andrieu and S. Livingstone. Peskun-tierney ordering for markov chain and process monte carlo: beyond the reversible scenario, 2019.

[And55] E. S. Andersen. On the fluctuations of sums of random variables ii. *Mathematica Scandinavica*, 2(2):195–223, 1955.

[AR09] C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient monte carlo computations. *Ann. Statist.*, 37(2):697–725, 04 2009.

[BBCD+18] J. Bierkens, A. Bouchard-Côté, A. Doucet, A. B. Duncan, P. Fearnhead, T. Lienart, G. O. Roberts, and S. G. Vollmer. Piecewise deterministic markov processes for scalable monte carlo on restricted domains. *Statistics & Probability Letters*, 136:148 – 154, 2018. The role of Statistics in the era of big data.

[BCG08] D. Bakry, P. Cattiaux, and A. Guillin. Rate of convergence for ergodic continuous markov processes: Lyapunov versus poincaré. *Journal of Functional Analysis*, 254(3):727 – 759, 2008.

[BCVD18] A. Bouchard-Côté, S. J. Vollmer, and A. Doucet. The bouncy particle sampler: A nonreversible rejection-free markov chain monte carlo method. *Journal of the American Statistical Association*, 113(522):855–867, 2018.

[BD17] J. Bierkens and A. Duncan. Limit theorems for the zig-zag process. *Advances in Applied Probability*, 49(3):791–825, 2017.

[Bes94] J. Besag. Comments on "representations of knowledge in complex systems" by ulf grenander and michael i. miller. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(4):549–603, 1994.

[BFR19] J. Bierkens, P. Fearnhead, and G. O. Roberts. The zig-zag process and super-efficient sampling for bayesian analysis of big data. *Ann. Statist.*, 47(3):1288–1320, 06 2019.

[BGKR20] J. Bierkens, S. Grazzi, K. Kamatani, and G. O. Roberts. The boomerang sampler. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 908–918, Virtual, 13–18 Jul 2020. PMLR.

[BGvdMS20] J. Bierkens, S. Grazzi, F. van der Meulen, and M. Schauer. A piecewise deterministic monte carlo method for diffusion bridges, 2020.

[Bie14] J. Bierkens. Non-reversible metropolis-hastings. *Statistics and Computing*, 26, 01 2014.

[Bil99] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[BKR18] J. Bierkens, K. Kamatani, and G.O. Roberts. High-dimensional scaling limits of piecewise deterministic sampling algorithms, 2018.

[BLBMZ15] M. Benaïm, S. Le Borgne, F. Malrieu, and P. A. Zitt. Qualitative properties of certain piecewise deterministic markov processes. *Ann. Inst. H. Poincaré Probab. Statist.*, 51(3):1040–1075, 08 2015.

[BNC19] J. Bierkens, P. Nyquist, and Schlottke M. C. Large deviations for the empirical measure of the zig-zag process, 2019.

[BPR+13] A. Beskos, N. Pillai, G. O. Roberts, J. M. Sanz-Serna, and A. Stuart. Optimal tuning of the hybrid monte carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 11 2013.

[BR17] J. Bierkens and G. O. Roberts. A piecewise deterministic scaling limit of lifted metropolis–hastings in the curie–weiss model. *The Annals of Applied Probability*, 27(2):846–882, Apr 2017.

[BRB19] J. Birrell and L. Rey-Bellet. Concentration inequalities and performance guarantees for hypocoercive mcmc samplers, 2019.

[BRSS17] N. Bou-Rabee and J. M. Sanz-Serna. Randomized hamiltonian monte carlo. *Ann. Appl. Probab.*, 27(4):2159–2194, 08 2017.

[BRZ19] J. Bierkens, G. O. Roberts, and P. A. Zitt. Ergodicity of the zigzag process. *Ann. Appl. Probab.*, 29(4):2266–2301, 08 2019.

[BVL19] J. Bierkens and S. M. Verduyn Lunel. Spectral analysis of the zigzag process, 2019.

[Cam86] L.L. Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer series in statistics. Springer My Copy UK, 1986.

[CC96] M. K. Cowles and B. P. Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[CG94] K. S. Chan and C. J. Geyer. Discussion: Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1747–1758, 12 1994.

[CH13] T. L. Chen and C. R. Hwang. Accelerating reversible markov chains. *Statistics & Probability Letters*, 83(9):1956 – 1962, 2013.

[CHP20] S. Cotter, T. House, and F. Pagani. The nuzz: Numerical zigzag sampling for general models, 2020.

[CLP99] F. Chen, L. Lovász, and I. Pak. Lifting markov chains to speed up mixing. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, STOC '99, page 275–281, New York, NY, USA, 1999. Association for Computing Machinery.

[Dav84] M. H. A. Davis. Piecewise-deterministic Markov processes: a general class of nondiffusion stochastic models. *J. Roy. Statist. Soc. Ser. B*, 46(3):353–388, 1984. With discussion.

[Dav18] M.H.A. Davis. *Markov Models and Optimization.* Chapman and Hall/CRC Monographs on Statistics and Applied Probability. Routledge, 2018.

[DBCD19] G. Deligiannidis, A. Bouchard-Côté, and A. Doucet. Exponential ergodicity of the bouncy particle sampler. *Ann. Statist.*, 47(3):1268–1287, 06 2019.

[DFG09] R. Douc, G. Fort, and A. Guillin. Subgeometric rates of convergence of f-ergodic strong markov processes. *Stochastic Processes and their Applications*, 119(3):897 – 923, 2009.

[DGM18] A. Durmus, A. Guillin, and P. Monmarché. Piecewise deterministic markov processes and their invariant measure, 2018.

[DGM20] A. Durmus, A. Guillin, and P. Monmarché. Geometric ergodicity of the bouncy particle sampler. *Ann. Appl. Probab.*, 30(5):2069–2098, 2020.

[DHN00] P. Diaconis, S. Holmes, and R. M. Neal. Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Probab.*, 10(3):726–752, 2000.

[DKPR87] S. Duane, A. D. Kennedy, B.J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics Letters B*, 195(2):216 – 222, 1987.

[DLP15] A. Duncan, T. Lelièvre, and G. Pavliotis. Variance reduction using nonreversible langevin samplers. *Journal of Statistical Physics*, 163, 06 2015.

[DLPD12] P. Dupuis, Y. Liu, N. Plattner, and J. D. Doll. On the infinite swapping limit for parallel tempering. *Multiscale Model. Simul.*, 10(3):986–1022, 2012.

[DMS09] J. Dolbeault, C. Mouhot, and C. Schmeiser. Hypocoercivity for kinetic equations with linear relaxation terms. *Comptes Rendus Mathematique*, 347(9):511 – 516, 2009.

[DMS15] J. Dolbeault, C. Mouhot, and C. Schmeiser. Hypocoercivity for linear kinetic equations conserving mass. *Trans. Amer. Math. Soc.*, 367(6):3807–3828, 2015.

[DMT95] D. Down, S. P. Meyn, and R. L. Tweedie. Exponential and uniform ergodicity of markov processes. *The Annals of Probability*, 23(4):1671–1691, 1995.

[DPBCD21] G. Deligiannidis, D. Paulin, A. Bouchard-Côté, and A. Doucet. Randomized hamiltonian monte carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates. *to appear in the Annals of Applied Probability*, 2021.

[Dyn65] E. B. Dynkin. *Markov processes. Vols. I, II.* Die Grundlehren der Mathematischen Wissenschaften, Bände 121. Academic Press Inc., 1965.

[EK86] S. N. Ethier and T. G. Kurtz. *Markov processes: Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986.

[FBPR18] P. Fearnhead, J. Bierkens, M. Pollock, and G. O. Roberts. Piecewise deterministic Markov processes for continuous-time Monte Carlo. *Statist. Sci.*, 33(3):386–412, 2018.

[FGM12] J. Fontbona, H. Guérin, and F. Malrieu. Quantitative estimates for the long-time behavior of an ergodic variant of the telegraph process. *Adv. in Appl. Probab.*, 44(4):977–994, 12 2012.

[FGM16] J. Fontbona, H. Guérin, and F. Malrieu. Long time behavior of telegraph processes under convex potentials. *Stochastic Processes and their Applications*, 126(10):3077 – 3101, 2016.

[FKG71] C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre. Correlation inequalities on some partially ordered sets. *Comm. Math. Phys.*, 22(2):89–103, 1971.

[Fol13] G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013.

[FR05] G. Fort and G. O. Roberts. Subgeometric ergodicity of strong markov processes. *Ann. Appl. Probab.*, 15(2):1565–1589, 05 2005.

[GG84] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[GM96] P. W. Glynn and S. P. Meyn. A liapounov bound for solutions of the poisson equation. *Ann. Probab.*, 24(2):916–931, 04 1996.

[GN20] A. Guillin and B. Nectoux. Low-Lying Eigenvalues and Convergence to the Equilibrium of Some Piecewise Deterministic Markov Processes Generators in the Small Temperature Regime. *Ann. Henri Poincaré*, 21(11):3575–3608, 2020.

[Gol51] S. Goldstein. On Diffusion By Discontinuous Movements, And On The Telegraph Equation. *The Quarterly Journal of Mechanics and Applied Mathematics*, 4(2):129–156, 01 1951.

[GRS96] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov chain Monte Carlo in practice*. Interdisciplinary Statistics. Chapman & Hall, London, 1996.

[Gus98] P. Gustafson. A guided walk metropolis algorithm. *Statistics and Computing*, 8, 1998.

[HÖ5] O. Häggström. On the central limit theorem for geometrically ergodic Markov chains. *Probab. Theory Related Fields*, 132(1):74–82, 2005.

[Hai16] M. Hairer. Convergence of markov processes. *Unpublished Lecture Notes*, 2016. Also available in http://www.hairer.org/Teaching.html.

[Has70] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[HHMS93] C. R. Hwang, S. Y. Hwang-Ma, and S. J. Sheu. Accelerating Gaussian diffusions. *Ann. Appl. Probab.*, 3(3):897–913, 1993.

[Hol74] R. Holley. Remarks on the fkg inequalities. *Comm. Math. Phys.*, 36(3):227–231, 1974.

[Jac06] M. Jacobsen. *Point process theory and applications*. Probability and its Applications. Birkhäuser Boston, Inc., Boston, MA, 2006. Marked point and piecewise deterministic processes.

[JH00] S. F. Jarner and E. Hansen. Geometric ergodicity of metropolis algorithms. *Stochastic Processes and their Applications*, 85(2):341 – 361, 2000.

[JR07] S. Jarner and G. O. Roberts. Convergence of heavy-tailed monte carlo markov chain algorithms. *Scandinavian Journal of Statistics*, 34(4):781–815, 2007.

[JT03] S. F. Jarner and R. L. Tweedie. Necessary conditions for geometric and polynomial ergodicity of random-walk-type. *Bernoulli*, 9(4):559–578, 08 2003.

[Kac74] M. Kac. A stochastic model related to the telegrapher's equation. *Rocky Mountain J. Math.*, 4(3):497–510, 09 1974.

[Kal02] O. Kallenberg. *Foundations of Modern Probability*. Probability and Its Applications. Springer New York, 2002.

[Kin92] J.F.C. Kingman. *Poisson Processes*. Oxford Studies in Probability. Clarendon Press, 1992.

[KLO12] T. Komorowski, C. Landim, and S. Olla. *Fluctuations in Markov processes*, volume 345 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Heidelberg, 2012. Time symmetry and martingale approximation.

[LBBG19] S. Livingstone, M. Betancourt, S. Byrne, and M. Girolami. On the geometric ergodicity of Hamiltonian Monte Carlo. *Bernoulli*, 25(4A):3109–3138, 2019.

[LLP07] D. Levin, M. Luczak, and Y. Peres. Glauber dynamics for the mean-field ising model: cut-off, critical power law, and metastability. *Probability Theory and Related Fields*, 146:223–265, 2007.

[LNP13] T. Lelièvre, F. Nier, and G. A. Pavliotis. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *Journal of Statistical Physics*, 152(2):237–274, 2013.

[LPW09] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2009. With a chapter by James G. Propp and David B. Wilson.

[LS79] P. A. W. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

[LS19] M. Ludkin and C. Sherlock. Hug and hop: a discrete-time, non-reversible markov chain monte carlo algorithm, 2019.

[LW20] J. Lu and L. Wang. On explicit l^2-convergence rate estimate for piecewise deterministic markov processes, 2020.

[Mir01] A. Mira. Ordering and improving the performance of monte carlo markov chains. *Statist. Sci.*, 16(4):340–350, 11 2001.

[MPS12] J. C. Mattingly, N. S. Pillai, and A. M. Stuart. Diffusion limits of the random walk metropolis algorithm in high dimensions. *Ann. Appl. Probab.*, 22(3):881–930, 06 2012.

[MRR+53] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[MS18] J. Markovic and A. Sepehri. Bouncy hybrid sampler as a unifying device, 2018.

[MT93a] S. P. Meyn and R. L. Tweedie. Stability of markovian processes ii: Continuous-time processes and sampled chains. *Advances in Applied Probability*, 25(3):487–517, 1993.

[MT93b] S. P. Meyn and R. L. Tweedie. Stability of markovian processes iii: Foster-lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3):518–548, 1993.

[MT96] K. L. Mengersen and R. L. Tweedie. Rates of convergence of the hastings and metropolis algorithms. *Ann. Statist.*, 24(1):101–121, 02 1996.

[MT09] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, 2 edition, 2009.

[Nea11] R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 113–162. CRC Press, Boca Raton, FL, 2011.

[Oks03] B. K. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, sixth edition edition, 2003.

[Pak17] A. Pakman. Binary bouncy particle sampler. *arXiv: Computation*, 2017.

[PdW12] E. A. J. F. Peters and G. de With. Rejection-free monte carlo sampling for general potentials. *Phys. Rev. E*, 85:026703, Feb 2012.

[Pes73] P. H. Peskun. Optimum monte-carlo sampling using markov chains. *Biometrika*, 60(3):607–612, 1973.

[PFJR20] M. Pollock, P. Fearnhead, A. M. Johansen, and G. O. Roberts. Quasi-stationary Monte Carlo and the ScaLE algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 82(5):1167–1221, 2020.

[PGCP17] A. Pakman, D. Gilboa, D. Carlson, and L. Paninski. Stochastic bouncy particle sampler. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2741–2750, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[Pre74] C. J. Preston. A generalization of the fkg inequalities. *Comm. Math. Phys.*, 36(3):233–241, 1974.

[RBS15] L. Rey-Bellet and K. Spiliopoulos. Irreversible langevin samplers and variance reduction: a large deviations approach. *Nonlinearity*, 28(7):2081–2103, May 2015.

[RGG97] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 02 1997.

[RR98] G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

[RR01] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various metropolis-hastings algorithms. *Statist. Sci.*, 16(4):351–367, 11 2001.

[RR09] G. O. Roberts and J. S. Rosenthal. Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[RT96] G. O. Roberts and R. L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 12 1996.

[SSG10] Y. Sun, J. Schmidhuber, and F. Gomez. Improving the asymptotic performance of markov chain monte-carlo by inserting vortices. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23, pages 2235–2243. Curran Associates, Inc., 2010.

[SSLD20] D. Sen, M. Sachs, J. Lu, and D. B. Dunson. Efficient posterior sampling for high-dimensional imbalanced logistic regression. *Biometrika*, 06 2020. asaa035.

[ST17] C. Sherlock and A. H. Thiery. A discrete bouncy particle sampler, 2017.

[TCV11] K. S. Turitsyn, M. Chertkov, and M. Vucelja. Irreversible monte carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240(4):410 – 414, 2011.

[Tie98] L. Tierney. A note on metropolis-hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9, 02 1998.

[Twe95] R. L. Tweedie. Topological conditions enabling use of harris methods in discrete and continuous time. *Acta Applicandae Mathematicae*, 34, 11 1995.

[Vas17] G. Vasdekis. Nonreversible Markov Processes and Zig-Zag. Master's thesis, University of Warwick, United Kingdom, 2017.

[VBCDD17] P. Vanetti, A. Bouchard-Côté, G. Deligiannidis, and A. Doucet. Piecewise-deterministic markov chain monte carlo, 2017.

[Vil06] C. Villani. Hypocoercivity. *Mem. Amer. Math. Soc*, 202, 10 2006.

[WR19] C. Wu and C. P. Robert. Generalized bouncy particle sampler. *arXiv: Computation*, 2019.

[WR20] C. Wu and C. P. Robert. Coordinate sampler: a non-reversible Gibbs-like MCMC sampler. *Stat. Comput.*, 30(3):721–730, 2020.