Approximations and transformations of piecewise deterministic Monte Carlo algorithms

Bertazzi, A.

DOI
10.4233/uuid:c53da6a5-948a-490d-9061-1f650f7a6125

Publication date
2023

Document Version
Final published version

Citation (APA)
Bertazzi, A. (2023). *Approximations and transformations of piecewise deterministic Monte Carlo algorithms*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:c53da6a5-948a-490d-9061-1f650f7a6125

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

APPROXIMATIONS AND TRANSFORMATIONS OF PIECEWISE DETERMINISTIC MONTE CARLO ALGORITHMS

ANDREA BERTAZZI

$$\mathcal{L}V \le -b_1 V + b_2 \mathbb{1}_C \qquad \left|\mathbb{E}_z[g(Z_{t_n})] - \mathbb{E}_z[g(\bar{Z}_{t_n})]\right| \le C\delta H(z)$$

$$X_t = Y_{\tau(t)} \quad \text{for } \tau(t) = \int_0^t s(X_u)\,du \qquad \mathcal{L}_s f = s\mathcal{L}f$$

$$\int_E \mathcal{L}f\,d\mu = 0 \qquad \|P_t(x,\cdot) - \pi(\cdot)\|_V \le \rho^t V(x)\, b$$

$$P_t(x,\cdot) \gg b\nu(\cdot) \quad \text{for all } x \in C$$

$$\mu(dz)P(z,dz') = \mu(dz')SPS(z',dz)$$

$$\frac{1}{T}\int_0^T f(X_t)\,dt \to \pi(f) \qquad \mu_\delta = \mu\left(1 - \delta^2 f_2 + \mathcal{O}(\delta^4)\right)$$

APPROXIMATIONS AND TRANSFORMATIONS OF PIECEWISE DETERMINISTIC MONTE CARLO ALGORITHMS

ANDREA BERTAZZI

COLOPHON

Andrea Bertazzi
Approximations and transformations of piecewise deterministic Monte Carlo algorithms, Delft, 2023

Printed by: Proefschriftspecialist

An electronic version of this dissertation is available at:

`http://repository.tudelft.nl/`

APPROXIMATIONS AND TRANSFORMATIONS OF PIECEWISE DETERMINISTIC MONTE CARLO ALGORITHMS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof.dr.ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Tuesday 13 June 2023 at 10 o'clock

by

Andrea BERTAZZI

Master of Science in Applied Mathematics
Delft University of Technology, the Netherlands
born in Milano, Italy

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof.dr.ir. G. Jongbloed	Delft University of Technology, promotor
Dr.ir. G.N.J.C. Bierkens	Delft University of Technology, copromotor

Independent members:

Prof.dr. A. Durmus	Ecole Polytechnique, France
Prof.dr. G.O. Roberts	University of Warwick, United Kingdom
Prof.dr. J.S. Rosenthal	University of Toronto, Canada
Prof.dr. F.H. van der Meulen	Vrije Universiteit Amsterdam
Prof.dr. F.H.J. Redig	Delft University of Technology
Prof.dr. A.W. van der Vaart	Delft University of Technology, reserve member

The research in this dissertation was funded by NWO *as part of the research programme 'Zigzagging through computational barriers' with project number 016.Vidi.189.043.*

Ai miei nonni, Beppe e Lely,
e alla mia mamma

Contents

Part I

Introduction

Chapter 1

Introduction

This thesis treats methods to approximate integrals of the form

$$\pi(f) := \mathbb{E}_\pi[f(X)] = \int_{\mathsf{X}} f(x)\pi(\mathrm{d}x), \tag{1.1}$$

where π is a probability distribution on a measurable space X and $f : \mathsf{X} \to \mathbb{R}^d$ is a function of interest that is integrable with respect to π. This problem arises in several fields such as Bayesian methods in statistics and machine learning, statistical physics, cosmology, as well as in the applied sciences. In the Bayesian paradigm, a statistical parametric model is selected together with a distribution on the parameters, which should in principle reflect our beliefs on the phenomenon of interest *before* observing any data and is thus called *prior* distribution. The objective is to study the probability distribution of the parameters *after* having observed some realisations of the phenomenon. This distribution is obtained by Bayes rule and is called *posterior distribution*. It is of the utmost interest for the Bayesian statistician to have an understanding of the posterior distribution, as this can bring clarity on which values of the parameter are deemed most likely by the model and thus enables some uncertainty quantification e.g. in the predictions of the model. When the parameter is very high dimensional, a situation that is very common in contemporary approaches, direct inspection of the posterior is impossible and thus the only way to obtain this understanding is to compute quantities such as its mean, covariance matrix, as well as its quantiles, the probabilities it assigns to certain regions, and so on. All these properties correspond to computing the expectation of a corresponding function, e.g. $f(x) = x$ in the case of the mean, with respect to the posterior distribution. Therefore, the problem fits in our framework of Equation (1.1), where π coincides with the posterior distribution of the parameter. Without reliable techniques to obtain approximations of (1.1), Bayesian statistics could only be applied on a limited set of simple, tractable models which cannot give a reasonable representation of reality,

and hence it would merely be a theoretical paradigm with little applicability. On the other hand, statistical physics concerns the modelling of molecular dynamics for large systems of particles. The formulation of a suitable model usually corresponds to considering the pair composed of the positions of particles and their momenta. The dynamics of the particles typically preserve the *Hamiltonian*, which is defined as the sum of a *potential energy*, accounting for the interactions between particles, and of a *kinetic energy*, usually dependent only on the momenta. The common practice in statistical physics is to associate to each possible configuration of the system a probability proportional to the exponential of the negative Hamiltonian of such configuration. The resulting probability distribution is known as the *Boltzmann distribution*. In this framework, states with large Hamiltonian are less likely than those with small Hamiltonian, respecting the physical intuition. It is important to have an understanding of how a model of this kind evolves in time, with particular attention on certain properties of the system of particles such as the relation between microscopic and macroscopic aspects, or thermodynamic properties. These can usually be represented as expectations of the form (1.1), where π is the Boltzmann distribution of the system. Similarly to Bayesian statistics, analytic solutions are typically available only in simple models, which do not model accurately the true physical system. While a way to address this is to approach this experimentally, this can be costly and/or unfeasible. Therefore, methods to approximate such quantities are necessary to give physicists a way to study models of interest.

The Monte Carlo approach is perhaps the most well known method to approximate (1.1) and has originated a research area that is still extremely active today, of which this thesis is an addition. The Monte Carlo method was proposed in 1949 in the foundational paper [103], coauthored by Metropolis and Ulam, two physicist of the Los Alamos Laboratory in New Mexico, USA. However, historical accounts report that the physicist Enrico Fermi had independently invented the Monte Carlo method as early as 15 years before [103], which was anyway the result of the collaboration of several scientists including the mathematician John von Neumann [102]. The Monte Carlo approach approximates (1.1) with the mean of the function evaluated in a random sample of data points from the distribution π, essentially relying on the law of large numbers. Building on the Monte Carlo approach, in [104] Metropolis and coauthors proposed to obtain this sample by accepting or rejecting states from a Markov chain with stationary distribution π, as under simple conditions the law of such Markov chain converges to the target distribution π independently of its initial condition. The resulting approach is known as Markov chain Monte Carlo (MCMC). In 1970, Hastings [80] generalised further the algorithm of [104], nowadays known as the Metropolis-Hastings (MH) algorithm. The MH algorithm was for several years predominantly used by physicists, who are still some of the main users of MCMC algorithms, until in the 1990s it became popular also among statisticians. The combination of this popularity and the availability of the first computers sparked a true revolution that made Bayesian statistics finally widespread, more than 200 years after the works by Thomas Bayes. Nowadays, the MH and related MCMC algorithms are discussed in

any Bayesian statistics textbook and are of crucial importance for anyone who wishes to follow the Bayesian paradigm. For an overview of MCMC algorithms we refer to [134] or the book [33], while for a historical account we recommend [130, 101] though many other good papers exist.

The MH algorithm is based on Markov chains that satisfy the *detailed balance* condition with respect to π, that is a sufficient condition for stationarity. The detailed balance condition expressing that, assuming the chain is initialised from π, for any states x and y the probability that the Markov chain moves from x to y is as likely to a move from y to x. Chains that satisfy detailed balance are time reversible, and thus commonly referred to as *reversible*. Reversible Markov chains then exhibit a backtracking behaviour as a consequence of detailed balance and thus are typically expected to have a slow, diffusive convergence to their stationary distribution π. This consideration lead to efforts to design algorithms based on Markov chains *non-reversible*, i.e. violating the detailed balance condition, but still have π as stationary distribution. The intuition that non-reversible chains can lead to better performing algorithms was supported by mathematical results from the end of the 1990s [55, 34], which showed that non-reversible chains can indeed have considerably faster convergence. One of the major techniques to obtain non-reversible chains is called *lifting* and relies on a clever augmentation of the state space to speed up the chain. The first idea from [55] was to add a momentum variable expressing the direction in which the Markov chain can move at a given iteration. The additional variable can be used to give some persistence to the chain, thus avoiding an inefficient, random-walk-like behaviour. Similar ideas lead to other non-reversible chains as [82, 77, 116, 151, 157], which empirically showed improved convergence properties compared to standard reversible approaches. As we shall see in Chapter 2, many non-reversible algorithms satisfy a modified balance condition, known as *skew detailed balance*.

This line of research lead the statistical physics [125, 13, 108] and the mathematical statistics [21] communities to (re)discover a class of continuous time stochastic processes known as piecewise deterministic Markov processes (PDMPs), which are the subject of this thesis. Such interest is explained by the fact that PDMPs arise naturally as limits of lifted Markov chains when time and space are suitably rescaled, and can be thus thought of as a continuous time analogue of non-reversible MCMC algorithms such as those of [55, 151]. However, PDMPs were discovered and studied well before the interest from the MCMC community in the seminal paper [49] by Mark Davis, who laid the foundations of this class of processes and also wrote the book [48], now a standard reference for mathematicians approaching the topic. A PDMP can be thought of as a stochastic process with degenerate noise, as it evolves according to the prescribed deterministic dynamics for an exponential random time, at which it jumps to a new state according to a prescribed kernel. Naturally, it is possible to design PDMPs that have π as stationary distribution, thus positioning them as novel MCMC algorithms. In this sense, PDMPs are different than other MCMC approaches, in most cases based on the MH algorithm, as they can be *rejection free*, i.e. it is unnecessary to include an acceptance/rejection step. As we shall discuss later on in this thesis,

this can be achieved only when the PDMP can be simulated exactly, which is possible only in simple settings. In addition, two other properties make PDMPs of particular interest for MCMC purposes: their natural non-reversibility and the possibility for *exact subsampling*. Non-reversibility of PDMPs is a byproduct of the deterministic dynamics, which are typically not time reversible. Hence PDMPs constitute a general framework to design non-reversible processes. Exact subsampling refers to the fact that, when π is a posterior distribution obtained by some Bayesian statistics model, under suitable conditions the random event times can be computed while accessing only a subset of the data-set. This property makes PDMPs stand out, as subsampling usually leads to the introduction of a bias in the estimation of integral (1.1). Thus PDMPs can lead to faster, more efficient MCMC algorithms when dealing with data sets with a large number of observations. The analogue of this property in the statistical physics context is that only a random subset of the interactions between particles should be computed at each iteration, as opposed to having to compute them all. Hence this can be beneficial when the number of particles is large.

The remarkable properties of PDMPs lead to considerable efforts from the mathematics and statistics communities to gain a deeper understanding of their properties and efficiency. At the time of writing of this thesis, several PDMPs have been proposed in the literature [23, 32, 159, 30, 25, 152], and modifications of such processes have been discussed [110, 22, 15, 26, 37, 148, 109, 153]. We point to [69] for an accessible introduction to PDMPs in the context of sampling. Theoretical investigations on various aspects related to the convergence of PDMPs to their stationary distributions have been extensively studied with various techniques [24, 64, 51, 5, 4, 56, 2, 19, 27, 20, 52, 65, 137, 154].

Contributions and organisation of the thesis

In Chapter 2 we illustrate the main mathematical concepts that brought us from the classical Monte Carlo method to MCMC algorithms based on PDMPs. It will become clear that a crucial concept is that of *lifted* Markov chains. We shall emphasise that most non-reversible algorithms, including PDMPs, are indeed instances of the lifting approach. While illustrating the main ideas, we introduce some concepts that will be useful throughout the thesis.

In the following chapters of the thesis, we focus on two different aspects of PDMPs. **Part II:** in this part of the thesis, we study *approximations of PDMPs*, a topic which was before mostly ignored in the literature. Our goal is to design discretisation schemes which satisfy suitable theoretical guarantees and can thus be used when PDMPs cannot be implemented exactly. This situation is very common in practice and is especially important in the context of MCMC algorithms. Chapter 3 discussed a general framework to define approximations of PDMPs of any order and is based on the paper

[16] A. Bertazzi, J. Bierkens, and P. Dobson. Approximations of Piecewise Deterministic Markov Processes and their convergence properties. *Stochastic Processes and their Applications*, 154:91–153, 2022

In Chapter 4 we investigate approximations given by the classical approach of *splitting schemes*, which we confirm gives second order schemes when given suitable symmetry. This is based on the work

[17] A. Bertazzi, P. Dobson, and P. Monmarché. Splitting schemes for second order approximations of piecewise-deterministic Markov processes. *arXiv.2301.02537*, 2023

Part III: in this part of the thesis we study how the velocity and speed of time of PDMPs can be tuned in order to improve their convergence properties. In Chapter 5, we introduce adaptive algorithms that learn suitable velocity vectors *on the go*, i.e. they use the path of the process to update their transition kernel. This chapter is based on

[15] A. Bertazzi and J. Bierkens. Adaptive schemes for piecewise deterministic Monte Carlo algorithms. *Bernoulli*, 28(4):2404 – 2430, 2022

In the last chapter, Chapter 6, we study time transformed Markov processes, with emphasis on PDMPs. The fundamental idea is that we can improve the exploration of the state space by speeding up time when the process is in low density regions for the target distribution. This can be done leaving π stationary by modifying the paths of the process suitably. We obtain novel processes applying this idea to PDMPs commonly used for MCMC purposes. Rigorous theoretical statements are proved to justify this approach.

[14] A. Bertazzi. Time transformations of piecewise-deterministic Markov processes. *In preparation*, 2023

Chapter 2

A mathematical introduction

In this chapter we start with the basics of Monte Carlo methods (Section 2.1), we describe the foundational ideas behind reversible and non-reversible MCMC algorithms (Section 2.2), and finally we introduce and explain the basics of PDMPs (Section 2.3). The goal of the chapter is to discuss the key ideas that drove the research from 1949 until today, as well as on the meaning of crucial techniques and conditions such as the (skew) detailed balance and the lifting approach.

2.1 Overview of Monte Carlo methods

While there are several approaches to approximate integrals of the form (1.1), the Monte Carlo method has revolutionised many fields of research. The idea is simple: given an independent sample from the target distribution π, that is

$$X_1, \dots, X_n \overset{iid}{\sim} \pi,$$

the Monte Carlo estimator is given by

$$\hat{\pi}_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i). \tag{2.1}$$

This estimator enjoys several important properties. First of all, it is unbiased, i.e. $\mathbb{E}[\hat{\pi}_n(f)] = \pi(f)$, where the expectation is over the samples. Moreover, it is supported by standard results in probability theory. The strong law of large numbers (SLLN) implies that $\hat{\pi}_n(f)$ converges almost surely to $\pi(f)$ as $n \to \infty$ if f is integrable wrt π, while, assuming that the variance $\sigma_f^2 = \mathrm{Var}_\pi(f(X))$ is finite, the central limit theorem (CLT) gives that

$$\sqrt{n}(\hat{\pi}_n(f) - \pi(f)) \to N(0, \sigma_f^2),$$

where the limit is in distribution. Thus the CLT gives that the Monte Carlo estimator converges with rate $\sqrt{n}$ and clarifies the importance of the asymptotic variance σ_f^2.

The main difficulty of this procedure is that we should be able to obtain samples from π. This is in general a challenging task which lead to the development of specialised methods to either generate samples from a given probability distribution or to go around this problem. In the next sections we give a brief overview of some of the main methods that rely on the Monte Carlo idea. In Sections 2.1.1 and 2.1.2 we discuss two different approaches to obtain samples from the target: sampling via the *inverse transform* and *rejection sampling*. As we shall see, the applicability of these methods is limited and motivates the search for alternative algorithms. In the following two sections, 2.1.3 and 2.1.4, we discuss the two main classes of algorithms that are nowadays used in practice: *importance sampling* and *Markov chain Monte Carlo* (MCMC) algorithms. An excellent reference that is used throughout this section is the book [129]. We shall work on a measurable space $(\mathsf{X}, \mathcal{X})$, where X is the state space and $\mathcal{X}$ is a σ-algebra on X.

Remark 2.1. In this thesis we shall often assume the following setting: the target distribution π has density wrt Lebesgue measure where with an abuse of notation (common in the literature) we denote by π also the density, i.e. $\pi(\mathrm{d}x) = \pi(x)\mathrm{d}x$. Moreover, we often write

$$\pi(x) = \frac{1}{Z} e^{-\psi(x)},$$

where $Z = \int_{\mathsf{X}} e^{-\psi(x)} \mathrm{d}x$. The function ψ is called *potential*.

2.1.1 Sampling with the inverse transform

The first approach to generate a sample from π leverages the fact that all one-dimensional distributions can be obtained by a suitable transformation of the uniform distribution on $[0, 1]$, which we denote as $\mathrm{Unif}([0, 1])$. Let us assume that the target π is defined in $\mathbb{R}$ and admits density which we also denote by π. We denote its cumulative distribution function by F, that is $F(x) = \int_{-\infty}^{x} \pi(\mathrm{d}y)$. Then we observe that, assuming F is invertible, for $U \sim \mathrm{Unif}([0, 1])$ it holds

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x),$$

which is obtained applying F to both sides of the inequality $F^{-1}(U) \leq x$ and used that F is strictly increasing. This shows that $F^{-1}(U) \sim \pi$. The same equality holds if F is not invertible by considering the generalised inverse $F^{-1}(y) = \inf\{x \in \mathbb{R} : F(x) \leq y\}$, which is equivalent to the standard inverse function when F is invertible. Therefore, knowledge of the inverse cdf implies that we can transform a draw from the uniform distribution into a draw from π. Here the implicit assumption is that we can draw samples from the uniform distribution; this can be achieved via pseudo random number generators, though it has to be noted that these techniques are deterministic and hence the samples will not be truly random, but will be accepted according to a set

of statistical tests. It is now straightforward to take advantage of this technique in the context of the Monte Carlo estimator (2.1): first we generate a sample $U_1, \dots, U_n \sim \text{Unif}([0,1])$ and then we compute the estimator

$$\hat{\pi}_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(F^{-1}(U_i)).$$

As a simple example which will play a role in subsequent chapters, consider the exponential distribution with parameter $\lambda > 0$, denoted as $\text{Exp}(\lambda)$. In this case $F(x) = 1 - \exp(-\lambda x)$ for $x > 0$ and the inverse cdf is $F^{-1}(u) = -\lambda^{-1} \ln(1-u)$ and hence it is straightforward to generate a sample $X \sim \text{Exp}(\lambda)$. This approach can be generalised to higher dimensional targets taking advantage of a factorisation such as

$$\pi(x) = \pi(x_1)\pi(x_2|x_1)\pi(x_3|x_1,x_2) \cdots \pi(x_d|x_1,\dots,x_{d-1}),$$

where we used the notation $x = (x_1, \dots, x_d)$ (note the order of components is irrelevant) and $\pi(x_k|x_1 \dots, x_{k-1})$ are the conditional distributions given by π. Assuming it is possible to compute the inverse cdf for each of the terms above, we can obtain a sample X from π by sampling X_1 from the first marginal, then X_2 from the conditional distribution $\pi(x_2|X_1)$, and so on for other components. Other factorisations of the target can be used (think e.g. of Bayesian networks where each component is a node of a directed acyclic graph and is independent of other components given its parents).

The issue with sampling with the inverse transform is that F^{-1} is unknown in closed form even for simple distributions such as the one-dimensional standard Gaussian, for which one would have to invert the function $x \mapsto \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \mathrm{d}y$, or the beta distribution. Hence the applicability of this method is very limited and other approaches must be sought.

2.1.2 Rejection sampling

Rejection sampling is a method that obtains independent samples from π by accepting or rejecting samples from another distribution g. Suppose we can draw samples from a probability distribution with density g which satisfies $\pi(x) \leq Mg(x)$ for all x, where M is a constant (note it must be $M \geq 1$ because g integrates to 1). In rejection sampling we start by drawing a sample $Y \sim g$ and an auxiliary random variable $U \sim \text{Unif}([0,1])$. Starting from Y and U, we can then obtain a sample $X \sim \pi$ by accepting or rejecting the proposal Y. In particular, the proposal Y is accepted if $U \leq \pi(Y)/(Mg(Y))$, and rejected otherwise. In case of acceptance we set $X = Y$, which is our sample from π. In case of rejection the procedure is repeated by drawing new samples Y' and U' and checking in the same fashion if Y' is accepted, and so on until the first acceptance. Let us show that this indeed gives that $\mathbb{P}(X \in A) = \pi(A)$ for any measurable set A, which means $X \sim \pi$ and thus we have obtained a sample from π. We can compute $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A | Y \text{ is accepted})$ by Bayes' rule as the

ratio of

$$\mathbb{P}(Y \in A, Y \text{ is accepted}) = \int_A \int_0^{\frac{\pi(y)}{Mg(y)}} \mathrm{d}u g(y) \mathrm{d}y = M^{-1}\pi(A)$$

and

$$\mathbb{P}(Y \text{ is accepted}) = \int_{\mathsf{X}} \int_0^{\frac{\pi(y)}{Mg(y)}} \mathrm{d}u g(y) \mathrm{d}y = \frac{1}{M}.$$

Therefore, accepted samples are distributed according to π.

Starting from $Y_1, \ldots, Y_n \overset{iid}{\sim} g$ and $U_1, \ldots, U_n \overset{iid}{\sim} \mathrm{Unif}([0,1])$, the rejection sampling method outputs an independent sample $X_1, \ldots, X_N \sim \pi$, where the sample size N is random. Assuming at least one sample is accepted, we can write the estimator of this procedure as

$$\hat{\pi}_n^{RS}(f) = \frac{\sum_{i=1}^n f(Y_i) \mathbb{1}_{U_i \leq \frac{\pi(Y_i)}{Mg(Y_i)}}}{\sum_{i=1}^n \mathbb{1}_{U_i \leq \frac{\pi(Y_i)}{Mg(Y_i)}}}.$$

Clearly, rejection sampling can only be applied when a suitable probability distribution g is available. In particular, g cannot have lighter tails than π since it must be that $\pi(x) \leq Mg(x)$ for all $x \in \mathsf{X}$. The computational cost of rejection sampling is related to the constant M, which is the inverse of the fraction of accepted proposals. Clearly, a small M, i.e. close to 1, can be chosen only if g is close to π, a requirement which is typically at odds with the fact that it should also be easy to sample from g. Finally, recalling that $\pi(x) = \exp(-\psi(x))/Z$ we observe that this procedure requires knowledge of the normalising constant of π, Z, as this is essential to determine a suitable M such that the bound $\pi(x) \leq Mg(x)$ holds. All these reasons hinder the applicability of rejection sampling to general targets.

2.1.3 Importance sampling

Importance sampling (IS) is an approach to obtain (unbiased) estimators of expectations wrt π by *weighting* samples from another probability distribution g. In particular, importance sampling does *not* give a sample from π. Consider a sample $X_1, \ldots, X_n$ from a probability distribution g which is different from the target distribution π, but has the same (or larger) support as π. In IS a weight is attached to each sample point to adjust for the fact that it is not from the right distribution. The importance sampling estimator of the integral (1.1) is then given by

$$\hat{\pi}_n^{IS}(f) = \frac{1}{n} \sum_{i=1}^n w(X_i) f(X_i), \tag{2.2}$$

where $w(x) = \frac{\pi(x)}{g(x)}$ is the weight function. This choice of weights is explained by the following equation:

$$\pi(f) = \int_{\mathsf{X}} f(x)\pi(x)\mathrm{d}x = \int_{\mathsf{X}} f(x) \frac{\pi(x)}{g(x)} g(x) \mathrm{d}x, \tag{2.3}$$

which is sometimes known as the importance sampling fundamental identity. Equation (2.3) expresses that the integral $\pi(f)$ can be rephrased as an integral with respect to any other probability distribution and thus estimating $\mathbb{E}_\pi[f(X)]$ is the same as estimating $\mathbb{E}_g[W(X)f(X)]$. It follows that the estimator (2.2) is unbiased and we have the usual guarantees on its convergence by the SLLN and CLT. Interestingly, IS was originally designed as a variance reduction technique, that is its goal was to obtain decrease the asymptotic variance of the estimator, hence speeding up convergence. To see that variance reduction is possible, we observe that $\mathrm{Var}_g(w(X)f(X)) = \mathbb{E}_g[w^2(X)f^2(X)] - \mathbb{E}_g[w(X)f(X)]^2$, which is in general different from $\mathrm{Var}_\pi(f)$. For weights $w(x) = \pi(x)/g(x)$ we obtain

$$\mathrm{Var}_g(w(X)f(X)) = \mathbb{E}_\pi\left[\frac{\pi(X)}{g(X)}f^2(X)\right] - \pi(f)^2. \tag{2.4}$$

Using Jensen's inequality for the first term we find that the choice of g which gives the smallest variance is $g(x) \propto |f(x)|\pi(x)$, where omitted factor ensures g integrates to 1. A simple computation shows that, if f is non-negative, this choice of g gives $\mathrm{Var}_g(w(X)f(X)) = 0$, which is a remarkable result. Unfortunately, this comes at the price that the normalising constant of such g is exactly $\pi(f)$, the unknown we started with, and hence this choice is not of practical interest. Nonetheless, this shows that indeed the asymptotic variance can be decreased with suitable choices of g. It should be observed that the variance of the IS estimator can also become infinite if g is chosen poorly, thus great care should be placed in the choice of g. A guideline is given by Equation (2.4): g should have thicker tails than π, as this implies that $\mathbb{E}_\pi\left[(\pi(X)/g(X))f^2(X)\right]$ is finite.

In most applications the normalising constant of π, which we denoted by Z, is unknown, hence the weight function w cannot be computed. This can be solved by estimating Z with the IS trick:

$$Z = \int_{\mathsf{X}} \exp(-\psi(x))\mathrm{d}x = \int_{\mathsf{X}} \frac{\exp(-\psi(x))}{g(x)} g(x)\mathrm{d}x.$$

This motivates what is known as the self normalised IS (SNIS) estimator:

$$\hat{\pi}_n^{SNIS}(f) = \frac{\sum_{i=1}^n \tilde{w}(X_i)f(X_i)}{\sum_{i=1}^n \tilde{w}(X_i)}, \tag{2.5}$$

where the weights are $\tilde{w}(x) = \exp(-\psi(x))/g_0(x)$ and g_0 is the unnormalised version of g. By the SLLN we have that the numerator converges to $Z\pi(f)$, while the denominator converges to Z, and hence asymptotically this converges to $\pi(f)$, though the estimator is biased for finite sample sizes.

Importance sampling is a method that has some interesting properties, such as variance reduction and the natural estimator of the normalising constant Z. However, poor choices of g can cause the asymptotic variance of the estimator to explode. More

in general, in order to have a good performance of the IS estimator one would like to choose g as close to the optimal choice as possible, though this is hardly achievable as designing and sampling from such g can be difficult.

2.1.4 The Markov chain Monte Carlo idea

The methods we have discussed so far use a sample from an a distribution other than π and proceed applying either a transformation, an acceptance-rejection step, or a weighting of each sample. The approach we discuss in this section, called Markov chain Monte Carlo, is based on a different idea. Essentially, in MCMC algorithms we design a Markov chain in such a way that its law converges asymptotically to the target π, and moreover the Monte Carlo estimator (2.1) converges to the right value.

Let us explain in more detail the overarching idea. Consider a Markov chain $(X_n)_{n\geq 0}$, that is a stochastic process taking values on the state space X such that the law of X_n conditional on the previous state X_{n-1} is independent of the previous states $X_0, X_1, \ldots, X_{n-2}$. The law of the Markov chain is specified by a transition kernel $P : (\mathsf{X}, \mathcal{X}) \to [0, 1]$, which gives $\mathbb{P}(X_{n+1} \in A | X_n = x) = P(x, A)$. A probability measure π is said *stationary* for the Markov chain when drawing the initial state of the chain from π, that is taking $X_0 \sim \pi$, implies that $X_n \sim \pi$ for all $n \geq 1$. In mathematical terms, the probability distribution π is stationary if for all measurable sets A

$$\pi(A) = \int_{\mathsf{X}} \pi(\mathrm{d}x) P(x, A) \tag{2.6}$$

where $\pi(A) = \int_A \pi(\mathrm{d}x)$. Stationarity means that π is the marginal of the y variable for the measure on the product space $\pi(\mathrm{d}x)P(x, \mathrm{d}y)$. This is a key concept, because a Markov chain which has π as its unique stationary distribution verifies a SLLN: for all initial conditions $X_0 = x \in \mathsf{X}$ it holds that

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \to \pi(f) \tag{2.7}$$

almost surely with respect to the law of the process (see [105, Theorem 17.0.1]). Moreover, a Markov chain version of the CLT holds under suitable conditions (again see [105, Theorem 17.0.1]), where the asymptotic variance is given by

$$\sigma_f^2 = \mathrm{Var}_\pi(f(X_0)) + 2 \sum_{i=1}^{\infty} \mathrm{Cov}_\pi(f(X_0), f(X_i)).$$

It follows that the MCMC estimator enjoys the same rate of convergence as the standard Monte Carlo estimator (2.1), but where the asymptotic variance now takes into account the correlations within the chain. Therefore, it is clear that we can approximate our integral $\pi(f)$ designing a Markov chain with stationary distribution

π[1]. The simplicity of applying this idea made MCMC methods an extremely popular choice and has resulted in intensive research for over thirty years. In the next section we delve into the details of the design of such chains.

2.2 Markov chain Monte Carlo algorithms

As discussed in the previous section, π-stationarity (2.6) is a key notion in the design of MCMC algorithms. Here we start our journey with a sufficient condition known as *detailed balance* (Equation 2.8). The detailed balance is at the core of the Metropolis-Hastings algorithm and is described in Section 2.2.1. In Section 2.2.2 we describe two of the most well known reversible algorithms, the Metropolis adjusted Langevin algorithm (MALA) and the classical version of the Hamiltonian Monte Carlo (HMC) algorithm. These two algorithms give us the chance to introduce our first discretisation schemes for stochastic processes, which we extensively treat in Chapters 3 and 4. The backtracking behaviour motivates looking for alternatives to the detailed balance condition. In Section 2.2.3 we discuss the *lifting* approach, which builds Markov chains on an augmented state space in such a way that π is a marginal stationary distribution. We shall show in Section 2.2.4 that typically lifted chains satisfy a *skew detailed balance* condition (Equation (2.17)). In Section 2.2.4.1 we discuss how to design a Metropolis-Hastings algorithm based on such condition. We conclude our journey into the basics of MCMC in Section 2.2.5 with a discussion on scaling limits of certain Markov chains that satisfy the skew detailed balance. As we shall see, these chains converge to a piecewise deterministic Markov process, a class of processes which is the protagonist of Section 2.3 and of the rest of this thesis.

2.2.1 Reversibility and the Metropolis-Hastings algorithm

As we have seen, it is crucial to build Markov chains which have the target, π, as stationary distribution. It is useful to rewrite (2.6) as

$$\int_{y\in A}\int_{x\in\mathsf{X}} P(y,\mathrm{d}x)\pi(\mathrm{d}y) = \int_{x\in\mathsf{X}}\int_{y\in A} P(x,\mathrm{d}y)\pi(\mathrm{d}x) \quad \text{for all } A\in\mathcal{X}.$$

This global balance condition has the physical interpretation that (in stationarity) for any set $A\in\mathcal{X}$ the flow of probability exiting A is the same that is entering A. This expression suggests the sufficient condition

$$\pi(\mathrm{d}x)P(x,\mathrm{d}y) = \pi(\mathrm{d}y)P(y,\mathrm{d}x) \quad \text{for all } x,y\in\mathsf{X}, \tag{2.8}$$

which is known as *detailed balance* (DB). The mathematical meaning of the DB condition is that the lhs and rhs, which can be seen as probability measures on the product

[1]Here we omitted some details for simplicity. Convergence to the stationary distribution formally requires notions such as irreducibility and positive Harris recurrence. See [105] for the details.

space $(\mathsf{X}\times\mathsf{X}, \mathcal{X}\times\mathcal{X})$, should coincide, in the sense that for all measurable sets A, B

$$\int_{x\in A}\int_{y\in B}\pi(\mathrm{d}x)P(x,\mathrm{d}y) = \int_{x\in A}\int_{y\in B}\pi(\mathrm{d}y)P(y,\mathrm{d}x).$$

Clearly, π is stationary for transition kernels P such that (2.8) holds. DB expresses that the flow of probability between points x and y should be the same in both directions when the chain is stationary. From the point of view of the theory of Markov chains, DB implies that the chain is *reversible*, that is starting the chain at π it holds that the law of X_{n+1} given $X_n = x$ is the same of the law of X_{n+1} given $X_{n+2} = x$. Reversibility in fact gives that changing the direction of time does not alter the law of the process. Indeed, taking two $\mathcal{X}$-measurable sets A, B and assuming the chain is in stationarity (i.e. $X_0 \sim \pi$) and time homogeneous (i.e. for all n $X_{n+1} \sim P(X_n, \cdot)$ and P does not depend on n) we obtain by DB and Bayes' rule

$$\mathbb{P}_\pi(X_{n+1}\in A | X_{n+2}\in B) = \mathbb{P}_\pi(X_{n+1}\in A | X_n \in B)$$

i.e. the direction of time is irrelevant. This property is also known as *time reversibility*.

The Metropolis-Hastings (MH) algorithm builds a Markov chain that is reversible with respect to π. The strategy is to consider a transition kernel P that accepts or rejects proposals from a kernel Q, where the probability of acceptance is chosen to ensure DB. Denoting by $\alpha : \mathsf{X}\times\mathsf{X}\to[0,1]$ the acceptance probability (where the first argument is the initial state and the second is the proposed state), this means P is of the form $P(x,A) = \int_A P(x,\mathrm{d}y)$ where

$$P(x,\mathrm{d}y) = \alpha(x,y)Q(x,\mathrm{d}y) + \delta_x(\mathrm{d}y)\int_{\mathsf{X}} Q(x,\mathrm{d}x')(1-\alpha(x,x')).$$

The first term corresponds to acceptance of a new state and the second term corresponds to rejection, in which case the state of the process is unchanged and the chain does not move. DB holds for P if

$$\alpha(x,y)\pi(\mathrm{d}x)Q(x,\mathrm{d}y) = \alpha(y,x)\pi(\mathrm{d}y)Q(y,\mathrm{d}x) \quad \text{for all } x\neq y,\ x,y\in\mathsf{X}.$$

In general, as prescribed by [150, Theorem 2], we should find the density

$$r(x,y) = \frac{\pi(\mathrm{d}x)Q(x,\mathrm{d}y)}{\pi(\mathrm{d}y)Q(y,\mathrm{d}x)}$$

restricting our attention to points (x,y) for which both measures are positive, that is considering states x, y in a set $R\subset\mathsf{X}\times\mathsf{X}$ such that the measures $\pi(\mathrm{d}x)Q(x,\mathrm{d}y)$ and $\pi(\mathrm{d}y)Q(y,\mathrm{d}x)$ are strictly positive[2]. Then it is clear that one can choose any

[2]Letting $h(x,y) = \pi(\mathrm{d}x)Q(x,\mathrm{d}y)/(\pi(\mathrm{d}x)Q(x,\mathrm{d}y) + \pi(\mathrm{d}y)Q(y,\mathrm{d}x))$, we have $R = \{(x,y)\in\mathsf{X} : h(x,y)>0,\ h(y,x)>0\}$.

Algorithm 1: Metropolis-Hastings algorithm

Input : Number of iterations N, initial condition X_0.
Output: Markov chain $(X_n)_{n=0}^N$.
Set $n = 0$;
while $n < N$ **do**
  Draw proposal $Y \sim Q(X_n, \cdot)$;
  Draw $U \sim \text{Unif}([0,1])$;
  Compute $\alpha(X_n, Y) = 1 \wedge r(Y, X_n)$;
  if $U \leq \alpha(X_n, Y)$ **then**
    Set $X_{n+1} = Y$;
  else
    Set $X_{n+1} = X_n$;
  end
end

acceptance probability which satisfies $\alpha(x,y) = r(y,x)\alpha(y,x)$. Among this class[3], the MH algorithm has acceptance probability

$$\alpha(x,y) = \begin{cases} 1 \wedge r(y,x) & \text{for all } (x,y) \in R, \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

This choice achieves the smallest asymptotic variance among the class of algorithms for a given proposal Q [124, 150]. Thus an iteration of the MH algorithm proceeds by first generating a proposal $Y \sim Q(x, \cdot)$, where x is the state at the previous iteration, and then either accepting Y as new state with probability $\alpha(x, Y)$, or rejecting it and staying at x. We describe this procedure in Algorithm 1.

In order to give instances where the acceptance probability is more explicit, we now consider three common situations:

a) *π and Q have a common dominating measure*: denote the respective densities as $\pi(\mathrm{d}x) = \pi(x)\eta(\mathrm{d}x)$ and $Q(x, \mathrm{d}y) = q(x,y)\eta(\mathrm{d}y)$, where e.g. η can be the Lebesgue measure and q is the density of a centred Gaussian distribution. Then

$$\alpha(x,y) = 1 \wedge \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}.$$

This setting is the basis of several classical algorithms, where typically Q has density wrt the Lebesgue measure with density given by $q(x,y) = q(y,x)$ (*symmetric MH*), $q(x,y) = q(|x-y|)$ (*random walk Metropolis*), and $q(x,y) = q(y)$ (*independent sampler*).

[3] In general it is enough to ensure $\alpha(x,y) = g(r(x,y))$ for any function $g(r) = rg(1/r)$. A well known alternative to MH is Barker's proposal [10], which takes $g(r) = r/(1+r)$.

b) *Q is reversible wrt another measure η*: in this case

$$\eta(\mathrm{d}x)Q(x,\mathrm{d}y) = \eta(\mathrm{d}y)Q(y,\mathrm{d}x)$$

and assuming η has a common dominating measure with π we have

$$\alpha(x,y) = 1 \wedge \frac{\pi(y)\eta(x)}{\pi(x)\eta(y)}.$$

This can be an advantageous framework when evaluating $Q(x,\mathrm{d}y)$ is difficult, while constructing a sample from it is achievable (think for instance of Q corresponding to a sequence of operations which can be carried out, but the overall probability is hard to write analytically).

c) *$Q(x,\cdot)$ has support on a finite number of states*: consider as proposal mechanism a kernel which, for a given initial state x, has support on a finite set of states. In this case we can write such kernel for $B \in \mathcal{X}$ as $Q(x,B) = \sum_{i=1}^{n} p_i(x)\mathbb{1}_{R_i(x)\in B}$ with $R_i : \mathsf{X} \to \mathsf{X}$ and $\sum_{i=1}^{n} p_i(x) = 1$ for all $x \in \mathsf{X}$. Here R_i are the possible maps that are applied to the state x, and $p_i(x)$ is the corresponding probability that $Q(x,\cdot)$ proposes a state $R_i(x)$. We assume that all R_i's are involutions, that is $R_i(R_i(x)) = x$ and also that $\pi(\mathrm{d}x) = \pi(x)\mathrm{d}x$. We find that a move from x to $R_i(x)$ is accepted with probability

$$\alpha(x,R_i(x)) = 1 \wedge \frac{\pi(R_i(x))p_i(R_i(x))|\det \nabla R_i(x)|}{\pi(x)p_i(x)}.$$

This setting covers the case of deterministic proposals, which we will discuss again in Section 2.2.2.2.

We conclude by observing that the MH algorithm does not need knowledge of the normalisation constant of π. This feature, together with its simplicity, has made for several decades the MH algorithm the workhorse behind most sampling algorithms.

2.2.2 Two classical MCMC algorithms

In this section we describe two MCMC algorithms that are nowadays the standard choice for most users: the Metropolis adjusted Langevin algorithm (MALA) and the Hamiltonian Monte Carlo (HMC) algorithm. There are naturally other important approaches in the MCMC literature that would deserve to be discussed, such as the Gibbs sampler [74, 73], the (elliptical) slice sampler [132, 117, 115]. For the sake of brevity we limit our attention to MALA and HMC, which we encounter in later chapters of this thesis and give us examples of discretisation schemes.

2.2.2.1 The Metropolis adjusted Langevin algorithm

MALA, introduced in [139], is based on the d-dimensional *overdamped Langevin diffusion*:

$$\mathrm{d}x_t = -\nabla\psi(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \tag{2.10}$$

where W_t is a d-dimensional Brownian motion. The interest in the Langevin diffusion is due to the fact that the stochastic process $(X_t)_{t\geq 0}$ that solves this stochastic differential equation (SDE) can be shown to have stationary distribution with density $\pi(x) \propto \exp(-\psi(x))$ wrt Lebesgue measure under mild assumptions. The fact that X_t is π-stationary makes it a natural candidate to obtain samples from π, since by the ergodic theorem one finds

$$\frac{1}{T}\int_0^T f(X_t)\mathrm{d}t \to \pi(f) \quad \text{a.s.}$$

which is the continuous time analogue of the discrete time empirical averages (2.1). Because it is not possible to compute the solution X_t analytically, the Langevin diffusion needs to be approximated using numerical approaches such as the *Euler discretisation*, which defines a Markov chain $(\overline{X}_n)_{n\geq 0}$ by

$$\overline{X}_{n+1} = \overline{X}_n - \delta\nabla\psi(\overline{X}_n) + \sqrt{2\delta}Z_n, \tag{2.11}$$

where $\delta > 0$ is the time step size and Z_n is a sequence of d-dimensional standard Gaussian random variables. The transition (2.11) is obtained by *freezing* the coefficients of (2.10) to their value at the beginning of each time step. In Chapter 3 we follow similar ideas to design discretisations of PDMPs. Naturally, this approach introduces a bias that depends on δ. In particular the stationary distribution of the chain $(\overline{X}_n)_{n\geq 0}$ will no longer be π and $n^{-1}\sum_{i=1}^n f(\overline{X}_i) \to \pi(f)$ does not hold. However, MALA uses the Euler discretisation as a proposal within the MH algorithm, correcting it with a suitable acceptance probability. Indeed Equation (2.11) corresponds to a proposal kernel Q given by

$$Q(x,\cdot) = \mathcal{N}(x - \delta\nabla\psi(x), 2\delta I_d).$$

Proposals from Q are then normally distributed with mean around the gradient descent path with step size δ and isotropic covariance with variance 2δ. This procedure corresponds to setting in Algorithm 1

$$r(y,x) = \frac{\pi(y)\nu(x; y - \delta\nabla\psi(y), 2\delta I_d)}{\pi(x)\nu(y; x - \delta\nabla\psi(x), 2\delta I_d)},$$

where $\nu(\cdot;\mu,\Sigma)$ is the density of $\mathcal{N}(\mu,\Sigma)$.

The choice of the step size δ in MALA has been extensively studied in the literature, mostly with the approach of scaling limits [131, 133], which led to guidelines on the acceptance probability of the proposals. Finally, we also observe that in many settings where evaluating the target is expensive it is common practice to omit the acceptance-rejection step and use the biased algorithm, known as the unadjusted Langevin algorithm (ULA). ULA has been the subject of substantial theoretical investigations, see e.g. [61, 62].

2.2.2.2 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is arguably the most well known MCMC algorithm to date. The main idea dates back to [58], but since then several variants to the basic algorithm have been proposed, e.g. [82, 75, 30]. A standard reference is the book chapter by Radford Neal [118].

As suggested by the name, HMC is based on *Hamilton's equations*, which we now describe. Consider a particle with position $x \in \mathbb{R}^d$ and momentum $p \in \mathbb{R}^d$, i.e. essentially its mass multiplied by its velocity. Hamilton's equations define deterministic dynamics that preserve the *Hamiltonian* $H(x,p) = \psi(x) + K(p)$, in which $\psi(x)$ and $K(p)$ play the role of the particle's *potential energy* and *kinetic energy*. A typical choice is $K(p) = p^T M^{-1} p/2$, where M is the mass matrix. In the following we take $M = I_d$, and hence p is to be interpreted as the velocity of the particle, which we then denote by v. Hamilton's equations associated to H give the system of ordinary differential equations (ODEs)

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_t, \quad \frac{\mathrm{d}v_t}{\mathrm{d}t} = -\nabla\psi(x_t), \tag{2.12}$$

with initial condition (x_0, v_0). Hamiltonian dynamics have several important properties. First of all, solutions of (2.12) preserve the Hamiltonian, in the sense that $H(x_t, v_t) = H(x_0, v_0)$ for all t. Also, the solution is volume preserving and reversible in a physical sense, meaning that the process goes back along its original path if the velocity is negated. For MCMC purposes it is particularly interesting that both volume and the Hamiltonian are preserved, as this means that the probability measure $\mu(x,v) \propto \exp(-H(x,v))$ is preserved and is thus a stationary distribution of the dynamics. It follows that it is possible to obtain samples from the target density $\pi(x) \propto \exp(-\psi(x))$ as it is the marginal of the position part, while the momentum variable has Gaussian stationary distribution.

Two issues should be addressed in order to use Hamiltonian dynamics for MCMC purposes. First, (x_t, v_t) can only explore states with same Hamiltonian as the initial value $H(x_0, v_0)$. A solution to this is to *refresh* the velocity, drawing a new value $v \sim \mathcal{N}(0, I_d)$ at either deterministic or random times. We shall encounter this strategy plenty of times in this thesis. The second issue is that Hamiltonian dynamics cannot be simulated exactly if not for a very limited number of choices of ψ, thus it is necessary to use discretisation schemes. The most used approach to obtain an approximate solution of (2.12) is to use a *splitting scheme*, a classical method which we shall study in the context of PDMPs in Chapter 4. The first step to define a splitting schemes is to *split* the dynamical system into sub-parts that can be solved exactly. For instance, consider an ODE $\dot{z}_t = f(z_t) = f_A(z_t) + f_B(z_t)$ with solution $\varphi_t(z)$. We can split it into two ODEs $\dot{z}_t = f_A(z_t)$ and $\dot{z}_t = f_B(z_t)$ of which we assume we can compute the solutions φ_t^A and φ_t^B. Then an approximation of φ_t is given by composing φ_t^A and φ_t^B, for instance in the order $\varphi_{\delta/2}^A \circ \varphi_\delta^B \circ \varphi_{\delta/2}^A$, where δ is the step size of the discretisation. It turns out that this is a second order approximation of the true solution, in the

sense that

$$\varphi_\delta(z) = \varphi^A_{\delta/2} \circ \varphi^B_\delta \circ \varphi^A_{\delta/2} + \mathcal{O}(\delta^2).$$

Following this philosophy, we can split the Hamiltonian system (2.12) as

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_t, \quad \frac{\mathrm{d}v_t}{\mathrm{d}t} = 0,$$

and

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = 0, \quad \frac{\mathrm{d}v_t}{\mathrm{d}t} = -\nabla\psi(x_t),$$

both of which can be easily solved exactly. To obtain an approximation of the solution at time δ, the splitting scheme known as *velocity Verlet* then proceeds by doing a velocity update, followed by a position update, and finally another velocity update:

$$\begin{aligned} v_{\delta/2} &= v - \nabla\psi(x)\frac{\delta}{2}, \\ x_\delta &= x + v_{\delta/2}\delta, \\ v_\delta &= v_{\delta/2} - \nabla\psi(x_\delta)\frac{\delta}{2}. \end{aligned}$$

We denote by $(x_\delta, v_\delta) = \overline{\varphi}_\delta(x, v)$ the solution given by the splitting scheme at time δ starting at (x, v). Importantly, this splitting scheme preserves volume and also $(y, w) = \overline{\varphi}_\delta(x, v)$ implies that $\overline{\varphi}_\delta(y, -w) = (x, -v)$, i.e. the trajectory is reversed by flipping the sign of the velocity. Clearly, the Hamiltonian is not preserved by $\overline{\varphi}_\delta$ because of the discretisation error. Moreover, it is common to apply $\overline{\varphi}_\delta$ a number $K \geq 1$ of times to obtain trajectories which are potentially further away from the initial state. We denote the obtained state as $\overline{\varphi}^K_t(x, v)$.

Therefore, splitting schemes give a deterministic mapping which approximately preserves the Hamiltonian, and we know that velocity refreshments are needed to ensure ergodicity. With these ingredients, the HMC algorithm builds upon the MH algorithm to define a Markov chain $(X_n)_{n\geq 0}$ by accepting or rejecting proposals from the kernel

$$Q(x, A) = \int_{\mathbb{R}^d} \mathbb{1}_{\overline{\varphi}^K_\delta(x,v)\in\overline{A}}\, \nu(v)\mathrm{d}v,$$

where $\overline{A} = A \times \mathbb{R}^d$ and ν is the density of the standard Gaussian distribution. Q corresponds to drawing a momentum $v \sim \mathcal{N}(0, I_d)$, and then obtaining a new state y as the x-marginal of $\overline{\varphi}^K_t(x, v)$. The corresponding acceptance probability is given by

$$\alpha(x, y) = 1 \wedge \exp(H(x, v) - H(\overline{\varphi}^K_\delta(x, v))), \tag{2.13}$$

where v is then seen as the momentum that gives proposal y. Indeed, to go back from y to x we must sample a velocity which is equal to the v-marginal of $\overline{\varphi}^K_\delta(x, v)$, but with flipped sign. In this classical version of HMC, v is fully refreshed at each iteration, hence it is unnecessary to keep track of its value. In Example 2.6 we describe a generalisation of this algorithm in which the velocity is only *partially refreshed* [82], in which case HMC gives a chain (X_n, V_n) where the velocity cannot be ignored.

Remark 2.2. When $K = 1$ the HMC algorithm we described coincides with MALA. Indeed, in this case the proposal generated by the leap-frog starting at x and with $v \sim \mathcal{N}(0, I_d)$ is

$$x_\delta = x + v_{\delta/2}\delta \sim \mathcal{N}(x - \nabla\psi(x)\delta^2/2, \delta^2 I_d),$$

which is the same proposal of MALA with step size $\delta^2/2$.

2.2.3 Lifted Markov chains

The detailed balance condition served as foundation of the MCMC revolution because of its simplicity and straightforward application via the MH algorithm. However, DB is only a sufficient condition for stationarity and intuitively gives Markov chains exhibiting a *backtracking behaviour*, in the sense that a move from $X_n = x$ to $X_{n+1} = y$ is followed by an equally likely move from $X_{n+1} = y$ to $X_{n+2} = x$. This is an undesirable property since it leads to an inefficient exploration of the state space. For this reason, it is interesting to study Markov chains that violate DB while having the correct stationary distribution. Various ideas have been proposed in the literature, e.g. non-reversible perturbations of Langevin diffusions [92, 59], inserting vortices within the MH algorithm [18], constructing a chain which keeps track of the previous state in order to avoid backtracking [116], or introducing a momentum variable to speed up mixing [82, 77]. Several of the non-reversible methods actually follow the same principle, which goes by the name of *lifting*, an approach that was the subject of several papers towards the end of the 20-th century [82, 55, 77, 34]. The foundational idea of lifting is simple: starting from a (reversible) π-invariant chain, we wish to define a related Markov chain on an augmented state space with the hope that the mixing time of this chain is smaller than that of the chain we started with. We call the chain on the augmented space *lifted* chain, as opposed to the original *collapsed* chain. Typically the augmentation of the state space aims to give the chain a sense of direction, for instance through the introduction of a *momentum* which improves over the *diffusive* behaviour of reversible chains. Importantly, the lifted chain should have π as marginal stationary distribution as this remains our target distribution. Let us now give a mathematical definition of a lifted chain, adapting from [34].

Definition 2.3 (Lifted chain)**.** *Starting from a Markov chain with transition kernel $\tilde{P}$ and state space X which is the support of the target distribution π, a corresponding lifted chain is a Markov chain with transition kernel P, state space $E \supset \mathsf{X}$, and stationary distribution μ for which there exists a surjective, measurable mapping $f : E \to \mathsf{X}$ such that*

$$\begin{aligned} &\pi(A) = \int_{f^{-1}(A)} \mu(\mathrm{d}z), \\ &\int_{x\in A}\int_{y\in B} \pi(\mathrm{d}x)\tilde{P}(x,\mathrm{d}y) = \int_{z\in f^{-1}(A)}\int_{u\in f^{-1}(B)} \mu(\mathrm{d}z)\ P(z,\mathrm{d}u) \end{aligned} \tag{2.14}$$

for all measurable sets A, B.

The first condition ensures that π is re-obtained when integrating out the additional variables, while the second condition gives that in stationarity the transition kernels $P, \tilde{P}$ assign the same probability of going from x to $\mathrm{d}y$ for all $x, y \in \mathsf{X}$, that is the following measures coincide

$$\pi(\mathrm{d}x)\tilde{P}(x, \mathrm{d}y) = \int_{z \in f^{-1}(x)} \int_{u \in f^{-1}(y)} \mu(\mathrm{d}z)\ P(z, \mathrm{d}u). \tag{2.15}$$

When $E = \mathsf{X} \times \mathcal{V}$, in which case the mapping f is $f(x, v) = x$, with $f^{-1}(x) = \{(x, v) : v \in \mathcal{V}\}$, and $\mu(\mathrm{d}x, \mathrm{d}v) = \pi(\mathrm{d}x)\nu_x(\mathrm{d}v)$, the conditions (2.14) become

$$\begin{aligned} &\int_{v \in \mathcal{V}} \nu_x(\mathrm{d}v) = 1, \\ &\int_{y \in A} \tilde{P}(x, \mathrm{d}y) = \int_{y \in A} \int_{v, w \in \mathcal{V}} \nu_x(\mathrm{d}v) P((x, v), (\mathrm{d}y, \mathrm{d}w)), \end{aligned} \tag{2.16}$$

for all measurable sets A and all $x \in \mathsf{X}$. In this case π is the marginal distribution of the lifted chain. Conditions (2.14) or (2.16) do not give a concrete recipe on how to construct a lifted chain, but rather lay a framework to think of such approach. Before giving examples of specific lifted chains, let us comment on why this is an approach worth considering. First of all, in Section 2.2.5 we shall see how, under suitable conditions, PDMPs appear as limits of lifted chains. Therefore, this entire thesis is a result of the lifting approach, which then gives us a clear picture of the reasons that sparked research on PDMPs. Moreover, several results indicate that lifting can indeed improve the speed of convergence over a reversible collapsed chain [34, 55] (we also refer the interested reader to [88] for a recent overview, and to [6, 128] for a more recent analysis).

Now, we give examples of lifted chains which were shown to have either theoretical or empirical improvements over their collapsed chains. In each example we emphasise a common structure, which allows us to obtain a modified balance condition in Section 2.2.4. After the examples, we briefly discuss what type of improvements can be expected from such lifted chains.

Example 2.4 (Diaconis, Holmes, Neal [55])**.** *Consider the state space $\mathsf{X} = \{1, \dots, n\}$ on which we wish to target the uniform distribution $\pi(i) = 1/n$ for $i \in \mathsf{X}$ (in Section 2.2.5.1 we describe the case of a generic π as given in [55, Section 5]). A random walk on X evolves by moving from state $i \in \mathsf{X}$ to $i - 1$ or $i + 1$ both with probability $1/2$, whereas if $i = 1$ or $i = n$ it stays where it is with probability $1/2$ and moves respectively to node 2 or $n - 1$ with probability $1/2$. This chain satisfies DB wrt π and hence is π-stationary; moreover it converges after a number of iterations of order n^2 (observe e.g. that the mean travelled distance after m iterations is of order $\sqrt{m}$). Now consider the augmented state space $E = \mathsf{X} \times \{+1, -1\}$ and let us define a Markov chain on E which has stationary distribution $\mu(x, v) = \pi(x)/2$. The idea is that the variable $v \in \{\pm 1\}$ denotes two* replicas *of the Markov chain with different behaviours. The chain moves to either $(x + v, v)$ or $(x + v, -v)$, hence the index v indicates the*

direction in which it should move. Starting from state (x, v) *and assuming for the moment* $x + v \in \mathsf{X}$*, for a free parameter* $p \in (0, 1)$ *the chain evolves following the two steps:*

1. *generate the proposal* $(y, w) = (x+v, -v)$ *and accept it with probability* p*, where in case of rejection* $(y, w) = (x + v, v)$*;*
2. *apply a deterministic flip to the variable* w*, that is output* $(y, -w)$*.*

If the chain is at a boundary point, that is either $(x, v) = (n, +1)$ *or* $(x, v) = (1, -1)$*, it has probability* $1 - p$ *of staying at* (x, v) *and alternatively moves to* $(x, -v)$*. Both steps preserve* μ *as they are* μ*-reversible for any* $p \in [0, 1]$*, though their composition is non-reversible. Hence this chain can be used to obtain draws from* π *by simply discarding the additional* v *variable. It is simple to verify that* this chain is a lifting of the random walk on X *as per the condition* (2.16)*: denoting respectively by* $\tilde{P}$ *and* P *the transition kernels of the random walk and the lifted walk we see for instance that for all* p

$$\tilde{P}(x, x + 1) = \frac{1}{2}\left(P((x, 1), (x + 1, 1)) + P((x, 1), (x + 1, -1))\right) = \frac{1}{2} \quad \text{for } x < n$$

and similarly for other transitions, including boundary terms. The interesting aspect of this chain is that it has momentum when $p > 1/2$*, in the sense that it keeps going in the same direction with probability* p*. The larger* p *the more this behaviour is observed, where in the case* $p = 1$ *the chain deterministically sweeps the state space. In [55, Theorem 1] it is shown that when* $p = 1 - 1/n$ *this chain mixes in order* n *iterations, which corresponds to a* speed up of order n *compared to the random walk. In [81] it is shown that the improvement obtained with the Diaconis-Holmes-Neal lifted chain is marginal for a* V *or* W *shaped target, in the sense that the speed-up is only logarithmic in the size of the state space. A generalisation allowing moves between states* (i, v) *and* (j, v) *according to a transition matrix for all* i, j *can be found in [151].*

Example 2.5 (Guided random Walk Metropolis, Gustafson [77])**.** *Let* $\mathsf{X} = \mathbb{R}$ *and* π *a generic target distribution on* X*. A* π*-stationary Markov chain* $(X_n)_{n\geq 0}$ *can be obtained with the MH algorithm using proposals from the Gaussian distribution with variance* σ^2*. This is the analogue of the random walk of the previous example, but this time defined on the real line. Gustafson [77] proposed a way to modify this chain to encode some momentum, with an approach that closely resembles that of [55, Section 5]. As in the previous example, let* $E := \mathsf{X} \times \{+1, -1\}$ *and* $\mu(x, v) = \pi(x)/2$*. From state* $(X_n, V_n) = (x, v) \in E$*,* (X_{n+1}, V_{n+1}) *is obtained as follows:*

1. *generate a proposal* $(y, w) = (x + v|Z|, -v)$ *for* $Z \sim \mathcal{N}(0, \sigma^2)$*; accept it with the MH probability* $1 \wedge \pi(y)/\pi(x)$*, and in case of rejection set* $(y, w) = (x, v)$*;*
2. *deterministically flip the velocity, that is output* $(y, -w)$*.*

As above, both steps are μ*-reversible, while their composition is not. The chain keeps moving in the direction* v *as long as proposals in such direction are accepted, e.g.*

when π increases. When π is unimodal this gives a chain that accepts all proposals that push the chain towards the mode and can reject proposals only when the chain is heading away from it. This is intuitively what we would like to see and closely resembles the behaviour of some PDMPs we encounter later on. We can see that the second condition in (2.16) *holds taking wlog $y > x$*

$$\tilde{P}(x, \mathrm{d}y) = \frac{1}{2} P((x, +1), (\mathrm{d}y, +1)),$$
$$\tilde{P}(x, x) = \frac{1}{2}(P((x, +1), (x, -1)) + P((x, -1), (x, +1)))$$

and thus $\tilde{P}$ is the transition kernel of the RWM. Therefore, the guided RWM is a lifted RWM. *Numerical simulations in [77] show that the guided random walk we described shows faster mixing. This is supported by the theoretical results of [2].*

Example 2.6 (HMC with partial momentum refreshments, Horowitz [82])**.** *In Section 2.2.2.2 we have seen that the basic HMC algorithm of [58] is reversible and gives a chain $(X_n)_{n\geq 0}$. Here we discuss how to lift this chain using* partial *refreshments as proposed by Horowitz [82]. In this algorithm, the lifted chain $(X_n, V_n)_{n\geq 0}$ has state space $E = \mathbb{R}^d \times \mathbb{R}^d$ and target $\mu(\mathrm{d}x, \mathrm{d}v) = \pi(\mathrm{d}x)\nu(\mathrm{d}v)$, where ν is the standard Gaussian measure. Starting from $(X_n, V_n) = (x, v)$, the idea is now to draw $W \sim \mathcal{N}(0, I_d)$ and take $V = \alpha v + \sqrt{1-\alpha^2} W$, i.e. a partial refreshment of v, as velocity to be used as input for the Verlet integrator. This gives a sense of direction to the chain, depending on the parameter $\alpha \in (0, 1)$. The lifted HMC algorithm follows the next two steps:*

1. *generate (y, w') from the kernel*

$$Q((x, v), A) = \int_{\mathbb{R}^d} \mathbb{1}_{\overline{\varphi}_\delta^K(x, \alpha v + \sqrt{1-\alpha^2} w) \in A} \, \nu(w)\mathrm{d}w,$$

   and set $(y, w) = (y, -w')$ with the MH probability $1 \wedge \exp(H(x, v) - H(y, w))$, or else set $(y, w) = (x, -v)$;

2. *deterministically flip the momentum, that is output $(y, -w)$.*

It is clear that the chain $(X_n, V_n)_{n\geq 0}$ is a lifting of the chain defined by the HMC chain with full refreshments, i.e. with $\alpha = 0$. Indeed, the chain with $\alpha = 0$ respects the conditions (2.16)*. Therefore we can see HMC with partial refreshments as a lifting of the HMC algorithm with full refreshment, or also as a lifting of MALA in the $K = 1$ case.*

Example 2.7 (Neal [116])**.** *An alternative approach to incorporate a sense of direction is to keep track also of the second to last state, obtaining a* second order *Markov chain, as opposed to the usual* first order *Markov chains. This approach gives a chain $(X_n, Y_n)_{n\geq 0}$ with state space $E = \mathsf{X} \times \mathsf{X}$. As in [116] we assume that X is finite. Let T be a π-reversible transition kernel. The second order chain of [116] is*

designed to have stationary distribution $\mu(x,y) := \pi(x)T(x,y)$, which by reversibility of T satisfies $\mu(x,y) = \mu(y,x)$. Since π is the marginal stationary distribution for both components, it is enough to compute the Monte Carlo estimator using either, e.g. the second component. Starting from (X_n, Y_n), an iteration of the chain proceeds as follows:

1. *set $(\tilde{X}_{n+1}, \tilde{Y}_{n+1}) = (Y_n, X_n)$;*
2. *draw $(X_{n+1}, Y_{n+1}) \sim Q((\tilde{X}_{n+1}, \tilde{Y}_{n+1}), \cdot)$, where Q is a μ-reversible transition kernel of the form*

$$Q((x_1,y_1),(x_2,y_2)) = \mathbb{1}_{x_1=x_2} U(x_1, y_2|y_1).$$

Overall, the one-step transition kernel of this procedure is given by

$$P((x_1,y_1),(x_2,y_2)) = \mathbb{1}_{y_1=x_2} U(y_1, y_2|x_1).$$

While both steps are μ-reversible, their composition P is not. Taking $U(y_1,y_2|x_1) = T(y_1,y_2)$ gives a chain for which the component $(Y_n)_{n\geq 0}$ evolves just like a first order chain with kernel T and thus does not reduce backtracking, though still breaks DB. We would like to choose U in such a way that the probability of rejection is smaller than for the kernel T. The reason is simple: since the first step above preserves μ, a rejection in the second step corresponds to a backtracking move, in the sense that upon rejection the chain can show a path of the form $(X_{n-1}, X_n) \to (X_n, Y_n) \to (Y_n, X_n)$, i.e. the second component has gone back to the state X_n. Based on a suggestion from [97], [116] discusses the following choice: for $y_2 \neq y_1$

$$U(x_1,y_2|y_1) = T(x_1,y_2) \min\left\{ \frac{1}{1-T(x_1,y_1)}, \frac{1}{1-T(y_2,y_1)} \right\},$$

while with remaining probability U proposes $y_2 = y_1$. It is clear that such U corresponds to decreasing the probability of rejection and hence of backtracking. This choice of U satisfies the conditions of [116, Theorem 2] and gives that Q is μ-reversible and moreover the asymptotic variance is at least as small as that of the first order chain with kernel T. This relation between asymptotic variances is shown for general state space in [2]. Second order chains were also studied in [54]. It is not difficult to see that the second order Markov chain is obtained by lifting the original Markov chain with transitions T.

The examples we discussed should give the impression that augmenting the state space in a smart way can steer the lifted chain to good directions of the state space, hence exploring it in a more systematic way. In order to better understand when lifting can be expected to give improvements, we briefly discuss some of the findings of [34]. A key quantity to study the mixing time of Markov chains is the *conductance*, which gives an indication of the worst bottleneck in the exploration of the target π.

Defining the conductance of the collapsed chain $\tilde{P}$ as

$$\Phi(\tilde{P}) := \min_{S:\pi(S)<1/2} \Phi(\tilde{P}, S), \qquad \Phi(\tilde{P}, S) := \frac{\sum_{x\in S, y\in S^C} \pi(x)\tilde{P}(x,y)}{\pi(S)},$$

we observe that the definition of lifted chain (2.14) gives that the set S^* that minimises $\Phi(\tilde{P}, S)$ can be translated to a set $f^{-1}(S^*)$ for which

$$\Phi(P) := \min_{\overline{S}:\mu(\overline{S})<1/2} \frac{\sum_{z_1\in\overline{S}, z_2\in\overline{S}^C} \mu(z_1)P(z_1,z_2)}{\mu(\overline{S})} \leq \Phi(P, f^{-1}(S^*)) = \Phi(\tilde{P}).$$

This shows that the conductance is monotone under lifting and thus we cannot hope to increase it with this approach. For this reason, lifting should not be seen as an approach that improves convergence for multimodal targets, as communication between modes does not get better for the lifted chain. For example, when π is multimodal a simple computation shows that jumps between modes are as likely for the lifted RWM of Example 2.5 as for the RWM. Therefore, in such cases one should use other techniques such as *simulated tempering* [100], perhaps in conjunction with lifting.

Remark 2.8. Simulated tempering is also an augmentation of the state space, which becomes $(x, \beta) \in \mathsf{X}\times\{\beta_0, \ldots, \beta_n\}$. The variable β plays the role of inverse temperature and a chain is designed to have stationary distribution

$$\mu(x, \beta) \propto \nu(\beta)\pi^\beta(x).$$

It is easy to see that μ and π do not satisfy the first condition in Definition 2.3 and therefore simulated tempering does not fit into the lifting framework.

However, lifting can lead to considerable speed-ups as seen in Example 2.4. To see that this is the case, consider the set time

$$\mathcal{A} = \max_S \pi(S)\mathcal{H}(\pi, S)$$

where $\mathcal{H}(\pi, S)$ is the expected hitting time of set S starting the chain in stationarity. Small values of $\mathcal{A}$ correspond to faster convergence of the chain. [34, Lemma 2.1] shows that for any Markov chain on a finite state space

$$\frac{1}{4\Phi} \leq \mathcal{A} \leq \frac{20}{\Phi^2}.$$

A simple application of such bounds relates the set times of P and $\tilde{P}$:

$$\mathcal{A}(P) \geq \frac{1}{4\Phi(P)} \geq \frac{1}{4\Phi(\tilde{P})} \geq \frac{1}{8\sqrt{5}}\sqrt{\mathcal{A}(\tilde{P})},$$

which shows that the *set time of the lifted chain P can improve up to square root of the set time of $\tilde{P}$* (see [34, Theorem 3.1] for a similar result for the mixing time). In

the context of a state space with n states, this corresponds to a remarkable best case speed up of order n, which importantly can be achieved as witnessed by Example 2.4. An upper bound for the mixing time of the optimal lifted chain of $\tilde{P}$ is given in [34, Theorem 3.2], though it is not clear when this bound is informative. As argued in [88], lifted chains can lead to improved convergence properties in the sense that the set and mixing times are closer to the lower bounds compared to the collapsed chain. It also follows that only marginal improvements can be achieved when the collapsed chain is already close to the lower bounds (see e.g. [81]).

2.2.4 The skew detailed balance condition

The examples of the previous section follow a similar structure, which we now analyse. Given a target π on X, we enlarge the state space to $E = \mathsf{X} \times \mathcal{V}$, on which we define the distribution $\mu(\mathrm{d}x, \mathrm{d}v) = \pi(\mathrm{d}x)\nu_x(\mathrm{d}v)$. In all examples, we applied an involution s, which in particular preserves ν_x and volume. We denote as S the transition kernel that applies the involution s, that is $S(z, \mathrm{d}z') = \delta_{s(z)}(\mathrm{d}z')$. Then the lifted chains have transition kernel $P((x, v), \cdot)$ which executes the following steps:

1. draw $(y, w) \sim Q((x, v), \cdot)$, where Q is a μ-reversible transition kernel;
2. apply the involution kernel S and output $s(y, w)$.

As we have stressed throughout the examples, this gives a chain that breaks DB though each individual operation is reversible. Examples 2.4, 2.5, and 2.6 exactly fit into this framework and choose as involution $s(x, v) = (x, -v)$ and transition kernel $P = QS$[4]. Example 2.7 differs slightly in that the involution is applied before the transition kernel Q, hence $P = SQ$, and the involution is $s(x, y) = (y, x)$. In fact, all these lifted chains satisfy what is called a *skew detailed balance* condition:

$$\mu(\mathrm{d}z)P(z, \mathrm{d}z') = \mu(\mathrm{d}z')SPS(z', \mathrm{d}z), \tag{2.17}$$

which is to be interpreted as equality of the two measures: for all measurable sets A, B

$$\int_{z\in A}\int_{z'\in B} \mu(\mathrm{d}z)P(z, \mathrm{d}z') = \int_{z\in A}\int_{z'\in B} \mu(\mathrm{d}z')SPS(z', \mathrm{d}z).$$

Taking e.g. $B = E$ we see that skew-DB implies μ-stationarity of the Markov chain P. This type of condition is sometimes referred to as (μ, S)-reversibility [2]. Note that for all $f, g \in L^2(\mu)$ it holds $\int gPf d\mu = \int fSPSg d\mu$, thus SPS is the adjoint of P. Interestingly, [2, Proposition 1] shows that a skew reversible transition operator P is always given by the composition of two μ-reversible operators, one of which is S. A Peskun-Tierney ordering for skew reversible chains was obtained in [2, Theorem 2]. In the next proposition, we show that under a simple condition a skew reversible lifted chain originates from a reversible collapsed chain.

[4]Here QP is to be intended as composition of kernels, that is according to the notation $QP(z, A) = \int_{z'\in A}\int_{u\in E} Q(z, du)P(u, \mathrm{d}z')$.

Proposition 2.9. *Consider two Markov chains $\tilde{P}$, P which are respectively the collapsed and lifted chains. Assume $z \in f^{-1}(x)$ is equivalent to $s(z) \in f^{-1}(x)$ for any $x \in \mathsf{X}$. Then P satisfies the skew detailed balance condition* (2.17) *if and only if $\tilde{P}$ satisfies the detailed balance condition* (2.8)*.*

Proof. For all measurable sets A, B

$$\begin{aligned}
&\int_{x\in A,\, y\in B} \pi(\mathrm{d}x)\tilde{P}(x,\mathrm{d}y) = \\
&= \int_{z\in f^{-1}(A),\, u\in f^{-1}(B)} \mu(\mathrm{d}z)\; P(z,\mathrm{d}u) && \text{(Conditions (2.14))}\\
&= \int_{z\in f^{-1}(A),\, u\in f^{-1}(B)} \mu(\mathrm{d}u)\; P(s(u),\mathrm{d}s(z)) && \text{(Skew-DB of } P\text{)}\\
&= \int_{u\in f^{-1}(B),\, z\in f^{-1}(A)} \mu(\mathrm{d}u)\; P(u,\mathrm{d}z) && (\nu\text{-invariance of } S)\\
&= \int_{y\in B,\, x\in A} \pi(\mathrm{d}y)\tilde{P}(y,\mathrm{d}x).
\end{aligned}$$

In the last equality we also used our assumption that for all $x \in \mathsf{X}$ it holds $z \in f^{-1}(x)$ iff $s(z) \in f^{-1}(x)$. □

The main assumption of the proposition is satisfied e.g. when $E = \mathsf{X} \times \mathcal{V}$, s is $s(x, v) = (x, -v)$, and ν is independent of the sign of v, while it does not hold in Neal's algorithm we encountered in Example 2.7.

As a final observation, we consider the relation of a skew reversible chain to its time reversal. It is simple to obtain that in stationarity (e.g. $Z_0 \sim \mu$) it holds that for all measurable sets A, B

$$\mathbb{P}_\mu(Z_{n+1} \in A |\, Z_n \in B) = \mathbb{P}_\mu(Z_{n+1} \in s(A) |\, Z_{n+1} \in s(B)).$$

Therefore, the chain SPS is the time reversal of P.

2.2.4.1 The skew-reversible MH algorithm

In this section, we discuss how the MH algorithm can be modified to give a skew reversible Markov chain. This more general setting is discussed e.g. in [2, 119, 149] and is useful for proposal kernels Q that are not time reversible, but for which SQS is time reversible, e.g. Hamiltonian dynamics given by the leap-frog scheme. Suppose P accepts or rejects proposals from a kernel Q according to probabilities prescribed by $\alpha : E \times E \to [0, 1]$. Once again, we denote as s an involution which preserves the target measure μ. We define the kernel P by

$$P(z, A) = \int_A \alpha(z, z')Q(z, \mathrm{d}z') + \mathbb{1}_{s(z)\in A}\left(1 - \int_E Q(z, \mathrm{d}z')\alpha(z, z')\right),$$

Algorithm 2: Skew-reversible Metropolis-Hastings algorithm

Input : Number of iterations N, initial condition Z_0.
Output: Markov chain $(Z_n)_{n=0}^N$.
Set $n = 0$;
while $n < N$ **do**
  Draw proposal $Y \sim Q(Z_n, \cdot)$;
  Draw $U \sim \mathrm{Unif}([0,1])$;
  if $U \leq r(Y, Z_n)$ **then**
    Set $Z_{n+1} = Y$;
  else
    Set $Z_{n+1} = s(Z_n)$;
  end
end

which corresponds to applying the involution s in case of rejection. Enforcing the skew-DB condition for the non-diagonal terms we find that α should satisfy

$$\alpha(z, z')\mu(\mathrm{d}z)Q(z, \mathrm{d}z') = \alpha(s(z'), s(z))\mu(\mathrm{d}z')Q(s(z'), s(\mathrm{d}z)) \quad \text{for all } z, z' \in E.$$

Denoting as R the set in which $\mu(\mathrm{d}z)Q(z, \mathrm{d}z')$ and $\mu(\mathrm{d}z')Q(s(z'), s(\mathrm{d}z))$ are equivalent, we define the density

$$r(z, z') = \frac{\mu(\mathrm{d}z)Q(z, \mathrm{d}z')}{\mu(\mathrm{d}z')Q(s(z'), s(\mathrm{d}z))},$$

which satisfies $r(z', z) = 1/r(s(z), s(z'))$. The skew reversible MH algorithm corresponds to the acceptance probability[5]

$$\alpha(z, z') = \begin{cases} 1 \wedge r(z', z) & \text{for all } (z, z') \in R, \\ 0 & \text{otherwise.} \end{cases}$$

This procedure is detailed in Algorithm 2.

Let us give the explicit expression of α in three cases:

a) *μ and Q have a common dominating measure*: denoting the respective densities as $\mu(z)$ and $q(z, z')$, we have

$$\alpha(z, z') = 1 \wedge \frac{\mu(z')q(s(z'), s(z))}{\mu(z)q(z, z')}.$$

This is the setting of the guided RWM of Example 2.5.

[5]This choice satisfies the condition $\alpha(z, z') = r(z', z)\alpha(s(z'), s(z))$

b) *Q is skew-reversible wrt another measure η with involution s*: in this case $\eta(\mathrm{d}z)Q(z,\mathrm{d}z') = \eta(s(\mathrm{d}z'))Q(s(z'),s(\mathrm{d}z))$ and assuming η and μ have a common dominating measure we have

$$\alpha(x,y) = 1 \wedge \frac{\mu(z')\eta(z)}{\mu(z)\eta(s(z'))}. \tag{2.18}$$

c) *Q has support on a finite number of states*: consider a kernel Q which, from an initial state x, has support on a finite set of states, i.e.

$$Q(z,B) = \sum_{i=1}^{n} p_i(z)\mathbb{1}_{R_i(z)\in B}$$

with $R_i : E \to E$ and $\sum_{i=1}^n p_i(z) = 1$ for all $z \in E$. Assume $R_i^{-1} = s \circ R_i \circ s$ for all i, and also that $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$. We find that the move from z to $R_i(z)$ is accepted with probability

$$\alpha(z, R_i(z)) = 1 \wedge \frac{p_i(s(R_i(z)))\mu(R_i(z))|\det \nabla R_i(z)|}{\mu(z)p_i(z)}. \tag{2.19}$$

This setting covers the case of deterministic proposals, such as the skew-reversible HMC algorithm of Section 2.6, but also proposals that are considered in Chapter 4. Indeed, in Section 4.1.2 we shall see that approximations of PDMPs obtained with splitting schemes correspond to proposals of the form (2.18).

2.2.5 Piecewise deterministic scaling limits of discrete lifted chains

The study of scaling limits of Markovian processes is widespread in the literature of MCMC algorithms as a useful tool to gain understanding of the properties of these processes. This approach consists in studying a process of interest (with an appropriate rescaling of time and space) in some limit (usually taking the dimension of the process to infinity). In the important paper [21], Bierkens and Roberts obtained the scaling limits of suitable rescalings of the MH algorithm and its lifted counterpart as given in [151] in the context of the Curie-Weiss model, one of the most well known models from statistical physics. In particular, the Curie-Weiss model has state space $\{+1,-1\}^n$ and the limit considered in [21] is for n going to infinity. Their main theorems show how in the limit the behaviours of MH and lifted MH differ substantially: the standard MH converges to the (reversible) Langevin diffusion (2.10), while the lifted MH converges with $\sqrt{n}$ faster rate to a different, non-reversible stochastic process belonging to the class of piecewise deterministic Markov processes (PDMPs). The limiting PDMP, known as the Zig-Zag process (ZZP), is a process $(X_t, V_t) \in \mathbb{R} \times \{+1,-1\}$, where X_t plays the role of the position of the process and $V_t \in \{+1,-1\}$ denotes what replica the process is currently in, or similarly the direction of the process. The dynamics of the ZZP combine a deterministic motion with

constant velocity with deterministic changes of replica at exponentially distributed random times with a suitable rate λ. This process can be truly seen as a continuous time and continuous state space analogue of the lifted MH algorithm. For this reason it is natural to suspect that the Zig-Zag process has the potential to exhibit faster mixing to its stationary distribution compared to the Langevin diffusion. In the next section, we give an informal proof that the scaling limits of the lifted chain of [55] converges to the Zig-Zag process.

2.2.5.1 An illustrative example

Consider the finite state space

$$\mathsf{X} = \{-n, -n+1, \dots, -1, 0, 1, \dots, n-1, n\}$$

on which we define a target distribution $\pi_n(y) = \exp(-\psi(y))/Z_n$, where Z_n is a suitable normalisation constant. Now we define a lifted walk with transition kernel P following [55]. First of all, our chain is of the form (X_n, V_n), where $X_n \in \mathsf{X}$ and $V_n \in \{\pm 1\}$ indicates the current velocity/replica. We define P as follows: starting from any $(x, v) \in \mathsf{X} \times \{\pm 1\}$ apart from $(n, 1)$ and $(-n, -1)$

$$\begin{aligned} P((x,v),(x+v,v)) &= 1 \wedge \frac{\pi(x+v)}{\pi(x)}, \\ P((x,v),(x,-v)) &= 0 \vee \left(1 - \frac{\pi(x+v)}{\pi(x)}\right), \end{aligned}$$

while for the two boundary states we take any suitable choice, as it does not matter when taking the limit. This chain satisfies skew-DB with involution $s(x, v) = (x, -v)$ and deterministic proposal kernel $Q((x, v), (x + v, v)) = 1$. Now we shall consider a rescaled version of this chain, in particular taking $y = n^{-\alpha}x$ for some $\alpha \in (0, 1)$. We denote the resulting transition kernel as P_n, which proposes moves from (y, w) to $(y + n^{-\alpha}w, w)$ with same probability as moves from (x, v) to $(x + v, v)$ according to P. Finally, we define a continuous time, pure jump process which at random times with distribution $\mathrm{Exp}(n^\alpha)$ updates its states drawing from P_n. In order to give an informal proof that this process converges to the one-dimensional Zig-Zag process as $n \to \infty$, we show that the generator converges to the generator of the ZZP[6]. The generator of the jump process we have defined acts on suitable functions as

$$\mathcal{L}_n f(y, w) = n^\alpha (P_n f(y, w) - f(y, w)).$$

For $P_n f(y, w) = \mathbb{E}_{(y,w)}[f(Y, W)]$ we find

$$\begin{aligned} P_n f(y,w) &= f(y + n^{-\alpha}w, w)\mathbb{P}_{(y,w)}(W = w) + f(y, -w)\mathbb{P}_{(y,w)}(W = -w) \\ &= f(y + n^{-\alpha}w, w) + \mathbb{P}_{(y,w)}(W = -w)(f(y, -w) - f(y + n^{-\alpha}w, w)) \end{aligned}$$

[6]See Section 2.3.3 for an introduction on the generator of a Markov process and in particular of a PDMP.

and thus we can rewrite the generator as

$$\mathcal{L}_n f(y,w) = \underbrace{n^\alpha \left(f(y+n^{-\alpha}w,w) - f(y,w)\right)}_{\text{drift part}} + \underbrace{n^\alpha \, \mathbb{P}_{(y,w)}(W=-w)(f(y,-w)-f(y+n^{-\alpha}w,w))}_{\text{jump part}}.$$

Applying Taylor's theorem to the *drift part* we find that as $n \to \infty$ it converges to $wf'(y,w)$, which corresponds to the deterministic motion part of the Zig-Zag process. For the jump part it is sufficient to notice that for our process it holds

$$\mathbb{P}_{(y,w)}(W=-w) = 0 \vee \left(1 - \frac{\pi(y+wn^{-\alpha})}{\pi(y)}\right)$$

and in particular because $\pi(y) \propto \exp(-\psi(y))$ that

$$1 - \frac{\pi(y+wn^{-\alpha})}{\pi(y)} \approx 1 - \exp(-wn^{-\alpha}\psi'(y)) \approx wn^{-\alpha}\psi'(y),$$

which gives that in the limit as $n \to \infty$ we have that the *jump part* of the generator converges to

$$(0 \vee w\psi'(y))(f(y,-w) - f(y,w)).$$

Therefore as $n \to \infty$ we have found that informally $\mathcal{L}_n$ converges to

$$\mathcal{L}f(y,w) = wf'(y,w) + (0 \vee w\psi'(y))(f(y,-w) - f(y,w)). \tag{2.20}$$

As we shall see in Section 2.3.3, this is the generator of the one dimensional Zig-Zag process with invariant distribution $\mu(x,v) = \exp(-\psi(x))/2$. Therefore, the lifted chain informally converges to the Zig-Zag process.

Remark 2.10. An informal scaling limit of the RWM algorithm can be obtained with similar computations. The result is that the RWM converges to the overdamped Langevin diffusion (2.10) with target π if the time is sped up by a factor n^α and space is rescaled by $n^{-\beta}$ with $\alpha > 0$, $\beta \in (0,1)$, and $\alpha - 2\beta = 0$. Indeed, one finds

$$\mathcal{L}_n f(x) = \frac{1}{2} n^{\alpha-2\beta}\big(-\psi'(x)f'(x) + f''(x)\big) + \mathcal{O}(n^{\alpha-3\beta}).$$

It is then important to observe that the rescaling of time and space is different than for the lifted chain, which corresponded to $\alpha - \beta = 0$ and thus needs a slower time rescaling. For such choice, the RWM converges to the Markov process which has generator $\mathcal{L}f(x) = 0$, that is a process that does not move.

2.3 Piecewise deterministic Monte Carlo algorithms

In Section 2.2 we gave an introduction to MCMC algorithms with an overview of reversibility and skew reversibility, and discussing the intuitive interest in processes that break the detailed balance condition. We concluded with an informal study of the scaling limit of the lifted random walk Metropolis algorithm, which led to a piecewise deterministic Markov process that we called Zig-Zag process. Because the ZZP arises as a continuous time version of a lifted chain, it seems natural to wonder if it has a better performance than reversible algorithms such as MALA. For this reason, processes such as the ZZP have received substantial attention from the MCMC community. In this section we give an introduction to PDMPs and to their use in the context of MCMC algorithms.

PDMPs were popularised by Mark Davis [49, 48] and are a large class of processes which can be designed to have a wide array of dynamics. In general, a PDMP defined on a space $(E, \mathcal{E})$ is identified by three *characteristics*:

1. a flow map $\varphi_t : E \times [0, \infty) \to E$ governing the deterministic motion;
2. a jump rate $\lambda : E \to [0, \infty)$ governing the random times of jumps;
3. a jump kernel $Q : E \times \mathcal{E} \to [0, 1]$ which is applied at event times and defines the new location of the process.

Starting at z, a PDMP with characteristics (φ_t, λ, Q) moves in space according to the flow map φ_t, which is the solution of an ODE

$$\frac{d\varphi_t(z)}{\mathrm{d}t} = \Phi(\varphi_t(z)), \quad \varphi_0(z) = z, \tag{2.21}$$

for random time τ, which has law

$$\mathbb{P}_z(\tau > t) = \exp\left(-\int_0^t \lambda(\varphi_u(z))\mathrm{d}u\right).$$

At time τ the state of the process is drawn from the jump kernel Q evaluated at $Z_{\tau-}$, that is

$$Z_\tau \sim Q(\varphi_{\tau-}(z), \cdot).$$

Clearly, the name of these processes comes from the fact that between two events the process appears deterministic, and the randomness appears only at random times. Applying this procedure we can obtain the path of the process between any two event times, thus giving the dynamics of the process up to a fixed time horizon. The one dimensional Zig-Zag process encountered in Section 2.2.5 corresponds to the particular choice $\varphi_t(x, v) = (x + vt, v)$, $\lambda(x, v) = \max(0, v\psi'(x))$, and $Q((x, v), (x, -v)) = 1$. Notice that we can retrieve the entire path of a PDMP as long as we store a *skeleton chain* $(Z_{t_k}, t_k)_{k\geq 0}$, which contains the states of the process right after every random

Algorithm 3: Piecewise deterministic Markov process

Input : Initial condition z.
Output: Skeleton chain $(Z_n, t_n)_{n\in\mathbb{N}}$.
Set $(Z_0, t_0) = (z, 0)$;
for $n = 1, 2, \ldots$ **do**
  Simulate next event time as

$$\tau = \inf\left\{r > 0 : \int_0^r \lambda(\varphi_u(Z_n))\mathrm{d}u \geq E\right\}$$

  where $E \sim \mathrm{Exp}(1)$;
  Simulate $Z_n \sim Q(\varphi_{\tau-}(Z_{n-1}), \cdot)$;
  Set $t_n = t_{n-1} + \tau$;
end

event, together with the time of such event. Indeed the state at any time point between two events is readily obtained applying the flow map. In Algorithm 3 we illustrate the general procedure to simulate a PDMP.

2.3.1 Some PDMPs from the MCMC literature

In this section we present some of the several PDMPs that have been introduced in the MCMC literature, with particular attention to those that we encounter later on in this thesis. Examples 2.11, 2.12, 2.13 can be seen as multi-dimensional generalisations of the Zig-Zag process we encountered in Section 2.2.5.1, while Examples 2.14 and 2.15 are based on Hamiltonian dynamics. All PDMPs we describe are of the form $(X_t, V_t)_{t\geq 0}$ and have $\pi \propto \exp(-\psi)$ as marginal stationary distribution for the position part. In particular, X_t takes in general values on $\mathbb{R}^d$, while the state space for the velocity component differs for the various processes. [152] gives general conditions on the three characteristics to ensure a PDMP has a wanted stationary distribution.

Example 2.11 (Zig-Zag sampler [23])**.** *The Zig-Zag sampler (ZZS) extends the process of Section 2.2.5.1 by taking state space $E = \mathbb{R}^d \times \{+1, -1\}^d$. For any $z \in E$, we write $z = (x, v)$ for $x \in \mathbb{R}^d$, $v \in \{+1, -1\}^d$, where x is interpreted as the position of the particle and v denotes the corresponding velocity. The deterministic motion of ZZS is determined by $\Phi(x, v) = (v, 0)^T$, i.e. the particle travels with constant velocity v. For $i = 1, \ldots, d$ we define the jump rates*

$$\lambda_i(x, v) := (v_i \partial_i \psi(x))_+ + \gamma_i(x, v), \tag{2.22}$$

where $\gamma_i(x, v)$ should satisfy

$$\gamma_i(x, v) = \gamma_i(x, R_i v) \quad \text{for all } (x, v) \in E,\ i = 1, \ldots, d \tag{2.23}$$

(see [23, Proposition 2.3]). The excess switching rates are often chosen to be zero, as this gives smaller asymptotic variance [19, 2]. The corresponding (deterministic) jump kernels are given by $Q_i((x,v),(\mathrm{d}y,\mathrm{d}w)) = \delta_{(x,R_iv)}(\mathrm{d}y,\mathrm{d}w)$*, where* δ_z *denotes the Dirac delta measure centred at* z *and* R_i *is the operator that flips the sign of the* i*-th component of the vector it is applied to, that is*

$$R_iv = (v_1 \dots, v_{i-1}, -v_i, v_{i+1}, \dots, v_d).$$

Hence the i*-th component of the velocity is flipped with rate* λ_i*. The ZZS falls in our definition of PDMP taking*

$$\lambda(x,v) = \sum_{i=1}^{d} \lambda_i(x,v), \quad Q((x,v),(\mathrm{d}y,\mathrm{d}w)) = \sum_{i=1}^{d} \frac{\lambda_i(x,v)}{\lambda(x,v)} \delta_{(x,R_iv)}(\mathrm{d}y,\mathrm{d}w).$$

Example 2.12 (Bouncy Particle Sampler [32])**.** *The bouncy particle sampler (BPS) can also be seen as a multidimensional extension of ZZS. In this case, the state space is* $E = \mathbb{R}^d \times \mathbb{R}^d$*. The deterministic motion of the BPS is the same as ZZS:* $\Phi(x,v) = (v,0)^T$*, while BPS has two types of random events:* reflections *and* refreshments*. These respectively have rates* $\lambda_1(x,v) = (\langle v, \nabla_x \psi(x)\rangle)_+$ *and* $\lambda_2(x,v) = \lambda_r$ *for* $\lambda_r > 0$*, and corresponding jump kernels*

$$Q_1((x,v),(\mathrm{d}y,\mathrm{d}w)) = \delta_{(x,R(x)v)}(\mathrm{d}y,\mathrm{d}w), \quad Q_2((x,v),(\mathrm{d}y,\mathrm{d}w)) = \delta_x(\mathrm{d}y)\nu(\mathrm{d}w),$$

where ν *is a rotation-invariant probability measure on* $\mathbb{R}^d$ *(usual choices are the standard Gaussian measure or the uniform measure on the unit sphere* $\mathbb{S}^{d-1}$*), and*

$$R(x)v = v - 2\frac{\langle v, \nabla_x \psi(x)\rangle}{|\nabla_x \psi(x)|^2} \nabla_x \psi(x).$$

The operator R reflects *the velocity* v *off the hyperplane that is tangent to the contour line of* ψ *passing though point* x*. Importantly, the norm of the velocity is unchanged by the application of* R*, and this gives the interpretation that* R *is an elastic collision of the particle off such hyperplane. As observed in [32], a strictly positive* λ_r *is needed to ensure ergodicity of the BPS. Notice also that the switching rate of BPS satisfies* $\lambda_1(x,v) \leq \sum_{i=1}^{d} (v_i \partial_i \psi(x))_+$ *and thus BPS has fewer events compared to ZZS, ignoring refreshments. Finally, it is possible to consider partial refreshments similarly to Example 2.6.*

Example 2.13 (Coordinate sampler [159])**.** *The coordinate sampler has* $V_t \in \mathcal{V}$*, where* $\mathcal{V} = \{\pm e_i : i = 1, \dots, d\}$ *and* e_i *is the* i*-th vector of the canonical basis. The deterministic dynamics of the coordinate sampler are as ZZS and BPS,* $\Phi(x,v) = (v,0)^T$*, and thus in this case only one component of the position at a time is changed. The switching rate is* $\lambda(x,v) = \langle v, \nabla\psi(x)\rangle_+ + \lambda_r$*, where* λ_r *plays the role of the excess switching rate similarly to* γ_i *for ZZS. At event times, the jump kernel is*

$$Q((x,v),(\mathrm{d}x',\mathrm{d}v')) = \sum_{w \in \mathcal{V}} \frac{\lambda(x,-w)}{\lambda(x)} \delta_x(\mathrm{d}x')\delta_w(\mathrm{d}v'),$$

where $\lambda(x) = \sum_{w\in\mathcal{V}} \lambda(x,v)$ is the total switching rate at position x. Therefore, Q leaves the position unchanged and updates the velocity from v to $w \in \mathcal{V}$ with probability $\lambda(x,-w)/\lambda(x)$. As shown in [159], under suitable conditions the coordinate sampler is geometrically ergodic as long as $\lambda_r > 0$, though it is unclear if this condition is necessary. The authors show that this algorithm can lead to improvements of the performance in challenging targets when compared to the ZZS.

Example 2.14 (Boomerang sampler [25])**.** *The Boomerang sampler has state space $\mathbb{R}^d \times \mathbb{R}^d$ and deterministic motion $\Phi(x,v) = (v, -(x-x_*))^T$, where x_* is an arbitrary reference point. These are the dynamics of a Hamiltonian system with Gaussian target. The corresponding ODE with initial condition $(x,v) \in \mathbb{R}^d \times \mathbb{R}^d$ has analytic solution*

$$x_t = x_* + (x - x_*)\cos t + v \sin t,$$
$$v_t = -(x - x_*)\sin t + v \cos t,$$

which gives elliptical trajectories. It is a simple exercise to check that these dynamics preserve the Gaussian measure $\mathcal{N}(x_*, \Sigma) \times \mathcal{N}(0, \Sigma)$ for any symmetric matrix Σ. Since in general we do not want to target a Gaussian, but a target $\pi \propto \exp(-\psi)$, it is necessary to add a suitable jump mechanism. Similarly to BPS, the Boomerang sampler can have both reflections and refreshments. Reflections have rate $\lambda_1(x,v) = (v^T \nabla_x U(x))_+$, where $U(x) = \psi(x) - (x-x_*)^T\Sigma^{-1}(x-x_*)/2$, and the jump kernel is $Q_1((x,v),(\mathrm{d}y,\mathrm{d}w)) = \delta_{(x,R(x)v)}(\mathrm{d}y,\mathrm{d}w)$ for the reflection operator

$$R(x)v = v - 2\frac{\langle v, \nabla_x \psi(x)\rangle}{|\Sigma^{1/2}\nabla_x\psi(x)|^2}\Sigma\nabla_x\psi(x).$$

Refreshments have $\lambda_2(x,v) = \lambda_r > 0$ and correspond to drawing a new value of the velocity vector from $\mathcal{N}(0,\Sigma)$. A factorised version of Boomerang in with componentwise velocity flips instead of full reflections can be obtained as discussed in [25].

Example 2.15 (Randomised Hamiltonian Monte Carlo)**.** *The randomised HMC algorithm presented in [31] is based on Hamiltonian dynamics described in Section 2.2.2.2, where refreshments take place at rate λ_r as opposed to at deterministic times. Hence, also in this case $\mathcal{V} = \mathbb{R}^d$. In fact, this setup corresponds to $\lambda(x,v) = \lambda_r$, while Q corresponds to drawing a fresh velocity from the standard Gaussian distribution. Naturally, this can be modified to have partial refreshments of the velocity choosing Q appropriately.*

It is worth observing that similar PDMPs can be designed with the same idea for general Hamiltonian dynamics given by the system of ODEs

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = \nabla K(p_t), \quad \frac{\mathrm{d}p_t}{\mathrm{d}t} = -\nabla\psi(x_t), \tag{2.24}$$

where x_t is the position and p_t the momentum, while K is the kinetic energy. An interesting choice is that of [120], where $K(p) = \sum_{i=1}^d |p_i|$ and gives the Laplace

distribution as stationary distribution for the p marginal. For this choice Hamilton's equation for x becomes

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = \mathrm{sign}(p),$$

where $\mathrm{sign}(p)$ *corresponds to the application of the sign function component-wise. It follows that the position vector evolves with velocity* ± 1 *in each direction similarly to the Zig-Zag process. However, this Hamiltonian-based process appears to possess more momentum than the ZZP, as less velocity flips per time unit take place.*

2.3.2 Tricks for the exact simulation of event times of PDMPs

As we have seen, a PDMP is a stochastic process characterised by the triple (φ_t, λ, Q). Clearly, in order to simulate a PDMP it is necessary to be able to perform the following tasks:

- solve the ODE (2.21). This is the case of ZZS, BPS, Boomerang sampler, but it does not hold for RHMC;
- simulate the event times by some suitable algorithm, which corresponds to simulating an exponential random variable with non-homogeneous rate;
- draw from the jump kernel Q. This is usually possible for all the PDMP based samplers commonly encountered in the literature.

In the ideal case, one would choose the characteristics in such a way that all three tasks can be solved. The exact simulation of the ODE and the jump kernel is case specific and there is no recipe to obtain exact simulation if e.g. an analytic expression for φ_t is unavailable. There are however techniques for the simulation of the event times, which is typically the most difficult task. In the following sections we discuss these methods, which are based on properties of Poisson processes and of the exponential distribution. All of these are used several times throughout this thesis.

2.3.2.1 Poisson thinning

The main technique to simulate a PDMP exactly (assuming φ_t and Q can be easily simulated) is to obtain the random events applying Poisson thinning [95]. Suppose we want to simulate a random time with non-homogeneous rate $\lambda(t)$, that is $\mathbb{P}(\tau > t) = \exp(-\int_0^t \lambda(u)du)$ and we are able to upper bound the rate as

$$\lambda(t) \leq \Lambda(t) \quad \text{for all } t \geq 0.$$

Then by [95, Theorem 1] we can obtain the next event time by iteratively generating proposals $\tilde{\tau}$ according to the rate $\Lambda(t)$, that is $\mathbb{P}(\tilde{\tau} > t) = \exp(-\int_0^t \Lambda(u)du)$, until one is accepted with probability $\lambda(\tilde{\tau})/\Lambda(\tilde{\tau})$. Rejected proposals correspond to "phantom" events and the following proposals are summed to the rejected ones. We give the procedure in Algorithm 4.

Algorithm 4: Poisson thinning

Input : Rate λ, upper bound Λ.
Output: Event time τ.
Set $\tau = 0$ and accept = false;
while accept == false **do**
  Simulate proposal $\tilde{\tau}$ with law $\mathbb{P}(\tilde{\tau} > t) = \exp(-\int_0^t \Lambda(u)\mathrm{d}u)$;
  Draw $U \sim \mathrm{Unif}[0, 1]$;
  if $U \leq \frac{\lambda(\tilde{\tau})}{\Lambda(\tilde{\tau})}$ **then**
    Set accept = true;
  end
  Set $\tau = \tau + \tilde{\tau}$;
end

In the context of PDMPs, Poisson thinning is applicable when we can bound the switching rates along the deterministic dynamics, that is when it is possible to obtain Λ such that for a fixed initial condition $z \in E$

$$\lambda(t) := \lambda(\varphi_t(z)) \leq \Lambda(t) \quad \text{for all } t \geq 0.$$

In this case Λ can (and in general does) depend on z. This technique allows exact simulation of PDMPs for sufficiently well-behaved contexts. However, it is only useful in practice when we can find an upper bound Λ which is sharp, i.e. $\Lambda(s)$ is not too large compared to $\lambda(s)$. Indeed, a loose bound corresponds to a big number of rejected proposals, which increases the computational cost of the algorithm. Finding sharp bounds is a very challenging problem in the MCMC setting. For ZZS, it is immediate to obtain Λ when e.g. ψ is gradient Lipschitz or has bounded Hessian (though these bounds can be inefficient), but otherwise there is not a general recipe to obtain the computational bounds.

2.3.2.2 Superposition

Another simple technique that can be helpful is *superposition*. Suppose the switching rate is of the form

$$\lambda(z) = \lambda_1(z) + \cdots + \lambda_N(z).$$

Then we can simulate the next event times by obtaining proposals τ_j for $j = 1 \ldots, N$ with distribution $\mathbb{P}(\tau_j > t) = \exp(-\int_0^t \lambda_j(\varphi_u(z))\mathrm{d}u)$ and then take $\tau = \min_j \tau_j$. It is a simple property of the exponential distribution that τ has the right distribution, that is $\mathbb{P}(\tau > t) = \exp(-\int_0^t \lambda(\varphi_u(z))\mathrm{d}u)$. This procedure is applicable for instance to the ZZS, in which the total switching rate is given by the sum of the rates of flipping each component of the velocity.

In Algorithm 5 we describe how to obtain a realisation of the skeleton chain (with phantom events) of ZZS with Poisson thinning and superposition.

Algorithm 5: The Zig-Zag sampler with Poisson thinning and superposition

Input : Initial condition (x, v).
Output: Skeleton chain $(X_n, V_n, t_n)_{n\in\mathbb{N}}$.
Set $(X_0, V_0, t_0) = (x, v, 0)$;
for $n = 1, 2, \ldots$ **do**
  Simulate proposals τ_i with distribution $\mathbb{P}(\tau_i > t) = \exp(-\int_0^t \Lambda_i(u)\mathrm{d}u)$ for $i = 1, \ldots, d$;
  Set $i^* = \arg\min_i \tau_i$;
  Set $X_n = X_{n-1} + V_{n-1}\,\tau_{i^*}$;
  Draw $U \sim \mathrm{Unif}([0, 1])$;
  if $U \leq \frac{\lambda_{i^*}(X_n, V_{n-1})}{\Lambda_i(\tau_i^*)}$ **then**
    Set $V_n = F_{i^*} V_{n-1}$;
  else
    Set $V_n = V_{n-1}$;
  end
  Set $t_n = t_{n-1} + \tau_{i^*}$;
end

2.3.2.3 Subsampling

Subsampling is a simple trick based on Poisson thinning which gives PDMPs one of their most remarkable properties. Consider the setting in which the switching rates admit the decomposition

$$\lambda(z) = \frac{1}{N}\sum_{n=1}^{N} \lambda_n(z), \tag{2.25}$$

and moreover that for an initial condition $z \in E$ it holds that

$$\lambda_n(t) := \lambda_n(\varphi_t(z)) \leq \Lambda(t) \quad \text{for all n=1,...,N and } t \geq 0.$$

Here it is important that Λ upper bounds all the terms in (2.25). Subsampling gives a way to generate the next event time according to the rate λ with $\mathcal{O}(1)$ computations, as opposed to the $\mathcal{O}(N)$ computations that would be necessary with the standard Poisson thinning procedure. Here by $\mathcal{O}(1)$ computations we mean that only a user defined number of λ_n's need to be evaluated, as opposed to the full sum (2.25). This is achieved by first generating a proposal $\tilde{\tau}$ with law $\mathbb{P}(\tilde{\tau} > t) = \exp(-\int_0^t \Lambda(\varphi_u(z))\mathrm{d}u)$, then drawing independently an index $J \in \{1, \ldots, N\}$ uniformly at random, and finally accepting $\tilde{\tau}$ with probability $\lambda_J(\tilde{\tau})/\Lambda(\tilde{\tau})$. The only modification to Algorithm 4 is then that the acceptance probability depends on the random rate λ_J as opposed to the total rate λ. The law of the accepted sample is then the correct one. In fact, subsampling is a computational trick that can be applied during the Poisson thinning step.

In the next example we describe how the subsampling procedure can be implemented within ZZS in the context of Bayesian inference, where subsampling corresponds to using only a random subset of the data to compute event times.

Example 2.16 (ZZS with subsampling)**.** *A typical situation in Bayesian statistics is that of modeling the observed data $Y_1, \dots, Y_N$ as realisations of a probability distribution $p(y|x)$, where x is a parameter with prior distribution $p(x) \propto \exp(-\psi_0(x))$. In this case the posterior distribution given the data is of the form*

$$\pi(x) = p(x) \prod_{i=1}^{N} p(Y_i|x).$$

We can rewrite $\pi(x)$ as

$$\pi(x) \propto \exp\left(-\psi_0(x) + \sum_{j=1}^{N} \log p(Y_j|x)\right)$$

and therefore as $\pi(x) \propto \exp(-\psi(x))$ where

$$\psi(x) = \frac{1}{N} \sum_{j=1}^{N} (\psi_0(x) - N \log p(Y_j|x)) =: \frac{1}{N} \sum_{j=1}^{N} \psi_j(x). \tag{2.26}$$

Now, using the definition of the ZZS on a target of this form, we must choose the switching rates such that (2.22) *is satisfied, but also keeping in mind that we would like to take advantage of subsampling. The canonical switching rates, i.e. corresponding to zero excess switching rates, are given by*

$$\lambda_i(x, v) = \max\left(0, \frac{1}{N} \sum_{j=1}^{N} v_i \partial_i \psi_j(x)\right), \tag{2.27}$$

which are not of the form (2.25)*, therefore suggesting to add a suitable excess rate. A choice that achieves both goals is the following: for all $i = 1, \dots, d$ we define the switching rate corresponding to flips of the i-th component of the velocity vector as*

$$\lambda_i(x, v) = \frac{1}{N} \sum_{j=1}^{N} \max(0, v_i \partial_i \psi_j(x)). \tag{2.28}$$

Indeed, this corresponds to choosing excess switching rate

$$\gamma_i(x, v) = \frac{1}{N} \sum_{j=1}^{N} \max(0, v_i \partial_i \psi_j(x)) - \frac{1}{N} \max\left(0, \sum_{j=1}^{N} v_i \partial_i \psi_j(x)\right)$$

which satisfies the condition (2.23). *Clearly,* $\gamma_i(x,v) \geq 0$ *as a consequence of the property* $\max(0, a+b) \leq \max(0,a) + \max(0,b)$. *Hence the switching rates* (2.28) *give a process with the correct stationary distribution and also allows for subsampling as they are of the form* (2.25). *The drawback is that positive excess switching rates* γ_i *increase the number of unnecessary events and make the process more diffusive (or "more reversible"). In order to apply subsampling it is now sufficient to have for all* $i = 1, \dots, d$ *bounds of the form*

$$\lambda_i^j(x + vt, v) \leq \Lambda_i(t) \quad \textit{for all } j = 1, \dots, N.$$

The ZZS with subsampling then coincides with Algorithm 5 in that the proposal events are obtained using the upper bounding rates Λ_i*, while the acceptance probability is obtained using the rate* $\lambda_{i^*}^J$ *for* $J \sim \mathrm{Unif}(1 \dots, N)$.

Remark 2.17. In fact, the ZZS with subsampling works by using unbiased estimates of the partial derivatives of ψ, where the estimate $\partial_i \psi_j(x)$ is obtained drawing $J \sim \mathrm{Unif}(\{1, \dots, N\})$. Indeed, we can rewrite the canonical rates in (2.27) as

$$\lambda_i(x,v) = \max\left(0, v_i \mathbb{E}_J[\partial_i \psi_J(x)]\right),$$

and thus $\partial_i \psi_J$ is an unbiased estimator of $\mathbb{E}_J[\partial_i \psi_J(x)]$. Naturally, this approach introduces variance due to the randomness in J. A simple approach to decrease the variance is to draw a set $\mathcal{M}$ of $M < N$ indices without repetition and use the estimator $N \sum_{j \in \mathcal{M}} \psi_j(x)/M$ instead. Alternatively, as described in [23] one can use *control variates*, which is a standard variance reduction method.

2.3.3 The generator of PDMPs

Most properties of discrete time Markov chains can be established by analysing the one-step transition kernel P. This is possible because P typically can be precisely characterised e.g. through its effect on test functions, or anyway the law of the process after one step is simple enough to be used for manipulations. This approach is in general not applicable to continuous time Markov chains, i.e. *Markov processes*. The reason for this is that the semigroup P_t, which gives the law of the process at time t, is typically a very complicated object which cannot be written in a nice form. It turns out that we can study another object, the *generator* of the process, instead of the semigroup. We shall make use of this operator countless times throughout this thesis and thus we now given an informal introduction. First we give the expression for the generator of PDMPs, then we show how this arises in Section 2.3.3.1 and the type of properties that we show manipulating it in Section 2.3.3.2.

The generator is the operator that describes the change of the law of the process in the limit as time goes to zero. This intuition translates to the concept of derivative, and thus the generator is the operator $\mathcal{L}$ such that its action on test functions is given by

$$\mathcal{A}f(z) = \lim_{t \to 0} \frac{\mathbb{E}_z[f(Z_t)] - f(z)}{t} \tag{2.29}$$

where the limit is uniform in supremum norm, i.e. $\sup_{z\in E}|t\mathcal{A}f(z) - (\mathbb{E}_z[f(Z_t)] - f(z))| \to 0$ as $t \to 0$. Therefore we can think of the generator applied to f as the derivative of $\mathbb{E}_z[f(Z_t)]$ evaluated at $t = 0$. The *domain of the generator* $\mathcal{D}(\mathcal{A})$ is the set of functions that are continuous and vanishing at infinity for which such limit exists. This is a set that is often hard to characterise, and therefore in the context of PDMPs it is common to talk about the *extended generator*, which is an operator that coincides with $\mathcal{A}$ on functions in $\mathcal{D}(\mathcal{A})$, but that is defined on a larger set of functions. The extended generator, which we denote by $\mathcal{L}$, is defined as the operator that makes the process

$$C_t^f = f(Z_t) - f(z) - \int_0^t \mathcal{L}f(Z_s)\mathrm{d}s \tag{2.30}$$

a local martingale. The set of functions for which this holds is denoted as $\mathcal{D}(\mathcal{L})$. In the case of PDMPs, the extended generator $(\mathcal{L}, \mathcal{D}(\mathcal{L}))$ was fully characterised in [48, Theorem 26.14], which gives that

$$\mathcal{L}f(z) = \langle \Phi(z), \nabla_z f(z)\rangle + \lambda(z)\int_E (f(y) - f(z))Q(z, \mathrm{d}y). \tag{2.31}$$

The first term, $\langle \Phi(z), \nabla_z f(z)\rangle$, corresponds to the deterministic motion of the process according to the ODE (2.21), while the second term, $\lambda(z)\int_E (f(y) - f(z))Q(z, \mathrm{d}y)$, corresponds to the jump part of the process, where λ is the rate of a jump taking place and Q is the kernel determining the new state if a jump takes place. As examples, consider ZZS (Example 2.11) and RHMC (Example 2.15). In the case of ZZS, we find that the extended generator is

$$\mathcal{L}f(x, v) = \langle v, \nabla_x f(x, v)\rangle + \sum_{i=1}^{d} \lambda_i(x, v)[f(x, R_i v) - f(x, v)]. \tag{2.32}$$

while for RHMC with Gaussian velocity we have

$$\mathcal{L}f(x, v) = \langle v, \nabla_x f(x, v)\rangle - \langle \nabla_x \psi(x), \nabla_v f(x, v)\rangle + \lambda_r \int \big(f(x, v') - f(x, v)\big)\nu(\mathrm{d}v').$$

2.3.3.1 An informal derivation of the generator of a PDMP

In order to obtain the generator of a PDMP with general characteristics (φ_t, λ, Q) we can then start by writing $\mathbb{E}_z[f(Z_t)]$ and then discarding all terms that are of order t^2 or higher, as those will disappear when taking the limit in (2.29). The first observation is that the PDMP can have any number of random events by time t, and thus we can write

$$\begin{aligned}\mathbb{E}_z[f(Z_t)] &= \mathbb{E}_z\left[\sum_{n=0}^{\infty} f(Z_t)\mathbb{1}_{n \text{ events}}\right] \\ &= \mathbb{E}_z\left[f(Z_t)\mathbb{1}_{0 \text{ events}} + f(Z_t)\mathbb{1}_{1 \text{ event}}\right] + \mathcal{O}(t^2).\end{aligned} \tag{2.33}$$

In the second equation we used that the probability of 2 or more events is an order t^2 event and thus we can ignore it in our informal derivation. Now, the probability of 0 events is given by $\exp(-\int_0^t \lambda(\varphi_u(z))du)$ and therefore

$$\begin{aligned}\mathbb{E}_z[f(Z_t)\mathbb{1}_{0 \text{ events}}] &= f(\varphi_t(z))\exp\left(-\int_0^t \lambda(\varphi_u(z))\mathrm{d}u\right)\\ &= \big(f(z) + t\langle\Phi(z), \nabla f(z)\rangle\big)\,(1 - t\lambda(z)) + \mathcal{O}(t^2),\end{aligned}$$

which is obtained Taylor expanding both factors around $t = 0$ and neglecting higher order terms. Hence we can further discard the remaining second order term to find

$$\mathbb{E}_z[f(Z_t)\mathbb{1}_{0 \text{ events}}] = f(z)(1 - t\lambda(z)) + t\langle\Phi(z), \nabla f(z)\rangle + \mathcal{O}(t^2).$$

The term corresponding to one event, $\mathbb{E}_z\left[f(Z_t)\mathbb{1}_{1 \text{ event}}\right]$, can be written as

$$\int_0^t \lambda(\varphi_u(z))e^{-\int_0^u \lambda(\varphi_r(z))\mathrm{d}r}\int_E f(\varphi_{t-u}(z'))Q(\varphi_u(z), \mathrm{d}z')e^{-\int_0^{t-u}\lambda(\varphi_r(z'))\mathrm{d}r}\mathrm{d}u.$$

The integral makes this at least an order t term and hence we can Taylor expand all terms as

$$\begin{aligned}\mathbb{E}_z\left[f(Z_t)\mathbb{1}_{1 \text{ event}}\right] &= \int_0^t \lambda(z)e^{-u\lambda(z)}\int_E f(z')Q(z, \mathrm{d}z')\mathrm{d}u + \mathcal{O}(t^2)\\ &= t\lambda(z)\int_E f(z')Q(z, \mathrm{d}z') + \mathcal{O}(t^2).\end{aligned}$$

We have thus obtained that

$$\mathbb{E}_z[f(Z_t)] = f(z) + t\langle\Phi(z), \nabla f(z)\rangle + t\lambda(z)\left(Qf(z') - f(z)\right) + \mathcal{O}(t^2)$$

and therefore as $t \to 0$

$$\frac{1}{t}(\mathbb{E}_z[f(Z_t)] - f(z)) = \mathcal{L}f(z) + \mathcal{O}(t) \to \mathcal{L}f(z).$$

2.3.3.2 Using the generator to study Markov processes

Let us discuss two properties that are crucial in the MCMC context which can be obtained using the generator. The first concerns the stationary distribution of the process. It is shown in e.g. [68, Proposition 9.2] that a probability distribution μ is stationary for the process if and only if

$$\int \mathcal{L}f(z)\mu(\mathrm{d}z) = 0 \tag{2.34}$$

for all functions f in (a core of) the domain of the generator. This is the analogue of showing for a discrete time chain $\int Pf(z)\mu(\mathrm{d}z) = \int f(z)\mu(\mathrm{d}z)$ for all functions in a class that separates measures. Indeed, in continuous time we can informally obtain the

condition (2.34) by differentiating wrt t the equation $\int t^{-1}(P_t f(z) - f(z))\mu(\mathrm{d}z) = 0$. A second type of question that can be addressed studying the generator is the uniqueness of the stationary distribution and the rate of convergence of the law of the process to it. The approach that is commonly used to obtain such properties is to prove *drift* and *minorisation* conditions. In this thesis we shall use this approach several times, see e.g. Theorems 4.26, 5.14, 5.18, 6.20. In the case of exponential convergence, [57] gives that a sufficient drift condition is the existence of a function $V \geq 1$ such that for constants $a, b > 0$ and a measurable set C it holds

$$\mathcal{L}V(z) \leq -aV(z) + b\mathbb{1}_C(z). \tag{2.35}$$

The discrete time analogue of this condition is for $\lambda \in (0,1)$ and $c > 0$

$$PV(z) \leq \lambda V(z) + c\mathbb{1}_C(z).$$

Similar drift conditions exist to obtain polynomial convergence of the process, see e.g. [70, 78]. The set C should satisfy a suitable property, which is connected to the ability of the process to regenerate itself when inside such set. The requirement on C is that it should be a *petite set*, which means there exists a non-trivial measure ν such that for all $z \in C$

$$\int P_t(z, A)\eta(\mathrm{d}t) \geq \nu(A), \tag{2.36}$$

where η is some distribution on the positive line. We shall encounter several times in this thesis a specific version of this notion, which is called *small set* condition and corresponds to $\eta(\mathrm{d}t) = \delta_{t^*}(\mathrm{d}t)$ for some $t^* > 0$. Typically, C is a compact set. Conditions (2.35) and (2.36), together with aperiodicity and irreducibility of the Markov process, imply that there exists a unique invariant measure μ and moreover there exist $\rho \in (0,1)$ and $D > 0$ such that

$$\|P_t(z, \cdot) - \mu\|_V \leq \rho^t DV(z).$$

Here $\|\cdot\|_V$ is the V-norm, defined as $\|\mu(z)\|_V := \sup_{|f| \leq V} |\mu(V)|$. The idea behind these conditions is that when the chain is in C, which for simplicity we assume to be a small set, we can draw from the measure ν with suitable probability to refresh the chain, while outside of C we can use the drift condition to obtain contractivity. Similar conditions in discrete and continuous time have been extensively studied, in addition to [57] see e.g. [107, 79, 64].

2.3.4 The skew detailed balance condition for PDMPs

It is interesting to understand in what sense PDMPs are non-reversible. As it turns out, one can choose the characteristics of the PDMP to satisfy the continuous time version of the skew detailed balance (2.17). In this section we work in the space $L^2(\mu) := \{f : \int f^2(z)\mu(\mathrm{d}z) < \infty\}$, where in particular we consider the generator as the operator obtained by taking the limit (2.29) with respect to the norm $\|f\|_\mu = \int f(z)^2 \mu(\mathrm{d}z)$.

Let us start by stating the detailed balance condition for a Markov process in terms of its semigroup P_t. This emulates the condition (2.8), where the one step transition kernel is substituted by P_t. This is then equivalent to asking for all $f, g \in L^2(\mu)$ that

$$\int \int f(z')g(z)P_t(z, \mathrm{d}z')\mu(\mathrm{d}z) = \int \int f(z)g(z')P_t(z, \mathrm{d}z')\mu(\mathrm{d}z).$$

Note that this implies that μ is stationary, which corresponds to $g = 1$. In other words, we have that the semigroup is self adjoint in $L^2(\mu)$, that is for all $t > 0$ and all $f, g \in L^2(\mu)$ it holds that $\langle P_t f, g\rangle_{L^2(\mu)} = \langle f, P_t g\rangle_{L^2(\mu)}$, where $\langle f, g\rangle_{L^2(\mu)} = \int f(z)g(z)\mu(\mathrm{d}z)$. Recalling that the time derivative of the semigroup is the generator, i.e. $\partial_t P_t = \mathcal{L}$, we can rephrase this condition in terms of generator as $\langle \mathcal{L}f, g\rangle_{L^2(\mu)} = \langle f, \mathcal{L}g\rangle_{L^2(\mu)}$ for all $f, g \in L^2(\mu) \cap \mathcal{D}(\mathcal{L})$, that is the generator is self-adjoint in $L^2(\mu)$. It can be shown that this property is sufficient and necessary for the detailed balance condition to hold (for a more precise statement, see [123, Theorem 4.5]).

Similarly, we can write the skew-detailed balance for a Markov process with semigroup P_t, recalling that we denote as s an involution which preserves μ and volume, and S is the corresponding transition kernel. The skew-DB condition reads

$$\int \int f(z')g(z)P_t(z, \mathrm{d}z')\mu(\mathrm{d}z) = \int \int f(z)g(z')SP_tS(z, \mathrm{d}z')\mu(\mathrm{d}z'),$$

which means $\langle P_t f, g\rangle_{L^2(\mu)} = \langle f, SP_tSg\rangle_{L^2(\mu)}$ and therefore in this case the adjoint of P_t is SP_tS. Using the terminology of [2], we say P_t is (μ, S)-self adjoint. Differentiating with respect to t we can rephrase this in terms of the generator obtaining the condition $\langle \mathcal{L}f, g\rangle_{L^2(\mu)} = \langle f, S\mathcal{L}Sg\rangle_{L^2(\mu)}$. Under suitable conditions this condition on the generator implies that the semigroup is (μ, S)-self adjoint (see [2, Theorem 9]). As observed in [29], we can also rewrite this condition as

$$\int g(z)\mathcal{L}f(z)\mu(\mathrm{d}z) = \int f \circ s(z)\mathcal{L}(g \circ s)(z)\mu(\mathrm{d}z). \tag{2.37}$$

This implies that μ is stationary for the process as long as $1 \in \mathcal{D}(\mathcal{L})$ [7] and $f \in \mathcal{D}(\mathcal{L})$ implies $f \circ s \in \mathcal{D}(\mathcal{L})$. Given this condition on the generator, in [2] it is established that the ZZP is skew-reversible wrt $\mu(\mathrm{d}x, \mathrm{d}v) = \pi(\mathrm{d}x)/2^d$ with involution $s(x, v) = (x, -v)$, while [29] gives conditions on the characteristics of PDMPs to ensure that skew reversibility wrt a wanted measure holds. In particular, also BPS and RHMC are (μ, S)-reversible [29].

2.3.5 Lifted Markov processes

It seems natural to wonder if the PDMPs we have introduced in the previous section are in some sense lifted Markov processes, and if so what is the corresponding collapsed

[7] This is actually not a trivial assumption to verify for non-compact state spaces, but here we do not discuss this technical aspect.

process. In order to address this question, we should first give a sensible definition of lifted Markov process, resembling the discrete time case of Definition 2.3. The first condition in (2.39) should not be different from the discrete time case. Let us now focus on the second condition. Let $g : \mathsf{X} \to \mathbb{R}$ be a function such that its extension to the augmented space E it satisfies $g(z) = g(y)$ for all $z \in f^{-1}(y)$. Then we can multiply by $g(y)$ in (2.15) and use the semigroups instead of the one-step transition kernels to obtain

$$\pi(\mathrm{d}x)\tilde{P}_t(x, \mathrm{d}y)g(y) = \int_{z\in f^{-1}(x)} \int_{u\in f^{-1}(y)} \mu(\mathrm{d}z)P_t(z, \mathrm{d}u)g(y).$$

This is to be interpreted for very small t, in the sense that in continuous time we expect the collapsed process to evolve by continuously drawing the additional variables in E from μ. Using that for $u \in f^{-1}(y)$ we have $g(y) = g(u)$ and integrating y on X we find

$$\pi(\mathrm{d}x)\tilde{P}_t g(x) = \int_{z\in f^{-1}(x)} \mu(\mathrm{d}z)P_t g(x).$$

Finally, we take the derivative of both sides wrt t to find the condition

$$\pi(\mathrm{d}x)\, \tilde{\mathcal{L}}g(x) = \int_{z\in f^{-1}(x)} \mu(\mathrm{d}z)\, \mathcal{L}g(x) \qquad \text{for all } x \in \mathsf{X}. \tag{2.38}$$

We are now ready to give the definition of lifted Markov process.

Definition 2.18. *Let $\tilde{\mathcal{L}}$ be the generator of a Markov process with state space X and stationary distribution π. Let $E \supset \mathsf{X}$ and suppose there exists a surjective, measurable mapping $f : E \to \mathsf{X}$ connecting the two spaces. Let μ be a probability distribution on E. A* lifted process *is any Markov process with generator $\mathcal{L}$, state space E, and stationary distribution μ such that for any measurable set A and any measurable function $g : \mathsf{X} \to \mathbb{R}$ for which the extension to E satisfies $g(u) = g(y)$ for all $u \in f^{-1}(y)$ it holds*

$$\begin{aligned} &\pi(A) = \mu(f^{-1}(A)), \\ &\int_A \pi(\mathrm{d}x)\, \tilde{\mathcal{L}}g(x) = \int_{z\in f^{-1}(A)} \mu(\mathrm{d}z)\, \mathcal{L}g(z). \end{aligned} \tag{2.39}$$

As we have seen, all lifted chains and processes we have encountered have state space $E = \mathsf{X} \times \mathcal{V}$ and stationary distribution $\mu(\mathrm{d}x, \mathrm{d}v) = \pi(\mathrm{d}x)\nu_x(\mathrm{d}v)$. In this setting, according to Definition 2.18 we find that ν_x should be a probability distribution, which is obvious, and more importantly that for all A

$$\int_A \pi(\mathrm{d}x)\, \tilde{\mathcal{L}}g(x) = \int_A \pi(\mathrm{d}x) \int_{v\in\mathcal{V}} \nu_x(\mathrm{d}v)\, \mathcal{L}g(x, v),$$

and thus that

$$\tilde{\mathcal{L}}g(x) = \int_{v\in\mathcal{V}} \nu_x(\mathrm{d}v)\, \mathcal{L}g(x, v), \tag{2.40}$$

where again g satisfies $g(x,v) = g(x)$ for all $x \in \mathsf{X}$. This condition respects our intuition: every infinitesimal step of the collapsed process starts by drawing v from ν_x and then updating x, thus mimicking the discrete time case.

The PDMPs we introduced in Section 2.3.1 have generators of the form

$$\mathcal{L}g(x,v) = \langle \Phi_1(x), \nabla_v g(x,v) \rangle + \langle \Phi_2(v), \nabla_x g(x,v) \rangle + \lambda(x,v)(Qg(x,v) - g(x,v)),$$

where Q affects only the v-component of the process. Moreover, the stationary distribution of these processes is of the form $\mu(\mathrm{d}x, \mathrm{d}v) = \pi(\mathrm{d}x)\nu(\mathrm{d}v)$. It is therefore clear that for g that is independent of v we obtain that (2.40) gives

$$\tilde{\mathcal{L}}g(x) = \langle\, \nabla_x g(x,v), \int_{v \in \mathcal{V}} \Phi_2(v)\nu(\mathrm{d}v) \,\rangle.$$

For all the PDMPs of Section 2.3.1 the expectation of Φ_2 wrt ν equals zero and thus these processes are obtained by lifting a process with generator $\tilde{\mathcal{L}}g(x) = 0$. This fact might seem surprising at first, but it can actually be explained by what mentioned in Remark 2.10, which we now elaborate further. The Zig-Zag process was obtained as a scaling limit of the lifted random walk, where the time is sped up by the same rate at which the space is contracted. On the other hand, the Langevin diffusion is the scaling limit of the random walk Metropolis algorithm if time is sped up at a rate that is the square of the rate of space contraction. Applying the same rescaling of the lifted chain to the RWM we find that the chain converges to a degenerate process with generator $\tilde{\mathcal{L}}g(x) = 0$. In view of this it is then not surprising that the Zig-Zag process is obtained by lifting such a degenerate process. Another interesting result which gives more clarity is [19, Theorem 4.1] and shows that the one-dimensional Zig-Zag process converges weakly to the overdamped Langevin diffusion as the excess switching rate goes to infinity. We illustrate these relations between the Zig-Zag process and the

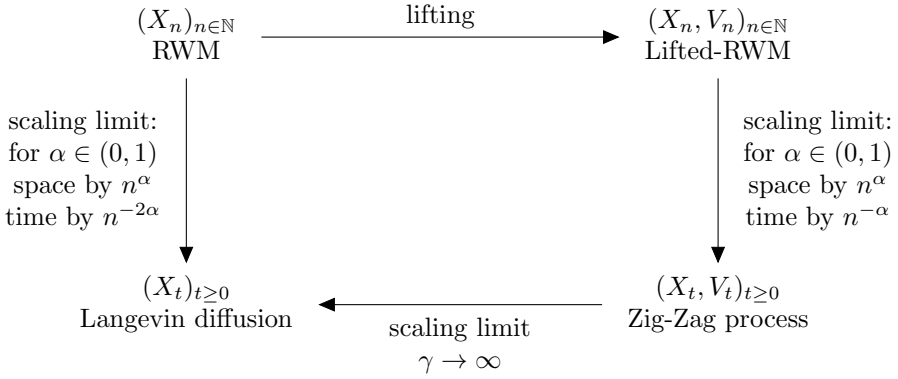


Figure 2.1: Relations between the Zig-Zag process and the overdamped Langevin diffusion in terms of lifting and scaling limits.

Langevin diffusion in Figure 2.1.

It is not clear whether processes obtained by lifting a non-degenerate, continuous time process have slower mixing times when compared to the ZZP or other PDMPs. What is certain is that most of the effort in the continuous time setting has followed the intuition given by the scaling limit approach, rather than the concept of lifting. Indeed, it is a simple computation to verify that the underdamped Langevin diffusion is also obtained by lifting a degenerate collapsed process and hence many of the most important processes of the form $(X_t, V_t)_{t\geq 0}$ from the MCMC literature share this property.[8] Are then scaling limits the right approach to understand the relations between these processes, or has continuous time lifting been overlooked? We leave this as a subject for future research, convinced that the quest for ever faster MCMC algorithms has not ended.

[8]We refer to [127] for a non-reversible version of MALA which can be easily adapted to become a lifting of the overdamped Langevin diffusion (2.10).

Part II

Approximations of piecewise deterministic Markov processes

Chapter 3

Approximations of PDMPs and their convergence properties

3.1 Introduction

Piecewise Deterministic Markov Processes (PDMPs) [49, 48] are nowadays widely used in mathematical modelling in fields such as mathematical biology [12, 39, 143], biochemistry [146], insurance risk theory [47, 66], materials science [1], neuroscience [122], and neutron transport [83]. Mathematical properties of PDMPs such as stability and stationarity have been extensively investigated in the mathematics community, see e.g. [11, 43, 65]. Moreover, in recent years these processes have also quickly gained in popularity for purposes of Monte Carlo computation in statistical physics [108, 125, 151] and in Bayesian statistics [69, 152], for example in the form of the Bouncy Particle Sampler (BPS) and the Zig-Zag Sampler (ZZS) [32, 23]. Several papers have further investigated the use of PDMPs in this area, e.g. [5, 2, 15, 24, 27, 64, 76, 99].

PDMPs are continuous time Markov processes which move along deterministic trajectories (typically in Euclidean space) on a time interval of random length, after which a (possibly random) transition occurs to a new state, followed by another deterministic motion, etc. The deterministic motion is prescribed by the integral curves, φ_t, of a vector field $\Phi : \mathbb{R}^d \to \mathbb{R}^d$, the length of the random time intervals between transitions is governed by a transition rate $\lambda : \mathbb{R}^d \to [0, \infty)$, and the transitions are described by a Markov kernel $Q : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0, 1]$. Together the vector field Φ, transition rate λ and transition kernel Q comprise the *characteristics* of the PDMP.

These processes are relatively easy to understand from a conceptual point of view and in some special cases their simulation can be performed exactly. In particular, if (i) the vector field Φ is explicitly integrable, (ii) it is possible to generate random times exactly as prescribed by λ, and (iii) it is possible to simulate from the transition kernel Q, then the iterative computation of trajectories of the associated PDMP is relatively straightforward.

Simulation of trajectories becomes problematic if one or more of these conditions are not met. Let us discuss possible problems that may arise. Concerning (i), the vector field Φ, as is well known in the field of differential equations, explicit solutions to the ODE $\dot{\varphi}_t = \Phi(\varphi_t)$ are only available in special cases, for example when Φ is affine, or when it is has some other special structure or symmetry. Concerning (ii), the transition rate λ, it is easy to simulate when the rate λ is constant or globally bounded. If λ is constant, then the random times between transitions are simply Exponential(λ)-distributed and thus easily simulated. If λ is globally bounded, say by a constant M, we may use a technique called *Poisson thinning* [95], which allows us to first simulate the random times according to an Exponential(M)-distribution and then accept a proposed transition time as a true transition with a probability governed by the ratio between $\lambda(\cdot)$ and M. The use of Poisson thinning may be extended to cases with non-constant bounds $M(s)$ along trajectories under the condition that it is simple to simulate from an inhomogeneous Poisson process with rate $M(s)$. However, finding a sharp bound $M(s)$ can be an extremely challenging problem in most practical settings. Moreover, the looser the bound the greater the computational cost of the simulation of the PDMP. For more extensive descriptions of Poisson thinning we refer to, e.g., [32, 23]. Finally problems with (iii), the simulation of transitions according to Q, may arise in various ways. For instance it may be interesting to approximate the transition kernel of the BPS (see Section 2.4 of [145]).

In this chapter we propose several schemes to approximate a PDMP in cases that are otherwise not straightforward to simulate, and we accompany these schemes by a detailed analysis of the convergence of the approximate process towards its exact, theoretical counterpart as the parameter governing the numerical precision, δ, converges to zero. Moreover, in the setting in which the PDMP is geometrically ergodic with a specified invariant measure, we investigate the theoretical convergence of the law of the approximate scheme to the invariant measure of the PDMP.

We introduce the *Fully Discrete PDMP* (FD-PDMP) Algorithm, the *Partially Discrete PDMP* (PD-PDMP) Algorithm and the Higher Order Partially Discrete PDMP Algorithm (Algorithms 7, 8 and 10, respectively). The FD-PDMP algorithm (Algorithm 7) defines a Markov chain $\{\overline{Z}_{t_n}\}_{n\in\mathbb{N}}$ on a mesh $0 = t_0 < t_1 < t_2 < \dots$ that moves deterministically between time steps, and a random event may occur at each of the mesh points with suitable probability. The PD-PDMP algorithm (Algorithm 8) defines a Markov chain that moves deterministically with exception of at most one random event in each interval of the form $[t_{n-1}, t_n]$. In contrast to the FD-PDMP the random event does not need to occur at mesh points. This difference motivates

the choice of name of the two algorithms. By allowing at most p random events per time step, the higher order algorithm (Algorithm 10) constructs an approximation of the PDMP of order p.

Naturally these algorithms are designed to be straightforward to simulate. Both the FD-PDMP and the PD-PDMP algorithms rely on first order approximations of the characteristics of the PDMP. A wide range of approximations for φ_t, λ, Q is allowed, see Assumptions 3.10, 3.11, 3.12 for the formal requirements. As a simple yet important example, consider the case in which we are interested in simulating a PDMP for which the event times are hard to obtain. With an Euler-type approach, we can use an approximation of λ that is constant between mesh points, based on the state of the process at the initial point of each time step. For such approximation, the next event time in the case of PD-PDMP is simply exponentially distributed with constant rate, which is straightforward to simulate. Similarly, in the case of the FD-PDMP a random event takes place at the end of the time interval according to a Bernoulli distributed random variable. In comparison to the simulation of the continuous time PDMP, both algorithms do not require an upper bound to the switching rates, which is required to apply Poisson thinning. In a similar fashion, simple approximations of φ_t and Q can be employed. We refer to Section 3.3 for a detailed description of the algorithms.

We study convergence of these algorithms as a function of the step size and of the time horizon. Under very broad assumptions on the approximation, in particular allowing for approximations of all three λ, φ_t, and Q, in Theorem 3.15 we are able to show convergence in a Wasserstein distance to the PDMP as the step size tends to zero. In the case in which it is possible to simulate φ_t and Q exactly, we obtain convergence of the PD-PDMP algorithm in the stronger metric of total variation (see Theorem 3.23). In this setting weaker assumptions on the continuous time PDMP are required. For instance we show in Examples 3.36 and 3.39 that BPS satisfies the assumptions of Theorem 3.23 but not those of Theorem 3.15. Moreover, both Theorems establish convergence of order p as long as the approximations of φ_t, λ, and Q are of order p. The proofs of both these theorems rely on couplings of the continuous time PDMP with its approximation and are described respectively in Couplings 3.54 and 3.57.

In many areas it is important to understand the long time behaviour of the approximation schemes. In the field of Markov chain Monte Carlo (MCMC) algorithms the goal is to simulate a process that converges in law to the correct probability measure, which is the posterior distribution in Bayesian statistics and the Boltzmann-Gibbs distribution in statistical physics. In this context, such a probability measure is the invariant distribution of the PDMP. In Theorem 3.30 we prove uniform in time convergence of the weak error between the PDMP and the approximations given by the FD-PDMP or the PD-PDMP algorithms. In particular, we obtain convergence in law of the approximation and its time average to the invariant measure of the PDMP in the joint limit as time tends to infinity and step size tends to zero (see Corollary 3.33).

We confirm the applicability of our theorems on a variety of examples. ZZS and BPS are instances of PDMPs for which exact simulation of the random event times is not always possible. In Example 3.2 we discuss how to approximate the ZZS, and in Examples 3.35, 3.45 we show that our Theorems apply to the proposed approximation. An attractive feature of ZZS is that it allows for exact *subsampling* (see [23]), which means that in a Bayesian statistics setting for each "iteration" of the algorithm only a subset of the data has to be accessed. In Example 3.41 we propose an approximation of ZZS with subsampling which also has the property of accessing a batch of the data over each time step and prove convergence in total variation as the step size tends to zero. For BPS we construct an approximation and prove convergence as step size tends to zero and time tends infinity in Examples 3.3, 3.39 and 3.51. In contrast to ZZS and BPS, randomised Hamiltonian Monte Carlo (RHMC) (see [30]) is an example of a PDMP in which it is typically not possible to simulate the flow φ_t exactly. In Examples 3.4, 3.37, and 3.42 we discuss approximations of RHMC and show convergence as the step size tends to zero and time tends to infinity. We also discuss continuous time approximation schemes of a PDMP in Examples 3.40 and 3.52.

Related works

Whereas discretisations of stochastic differential equations such as the Langevin equation have been studied extensively in the literature, see e.g. the book [87] or recent papers [61, 62, 144], the same has not been done for PMDPs. Here we give a brief overview of works that are to some extent related to the present manuscript.

An approximation scheme for PDMPs suitable for a specific setting was proposed in [94]. The authors consider the case in which the ODE describing the deterministic motion can only be solved numerically, a global upper bound for the switching rates is available, and the kernels Q can be simulated exactly. Their proposal is to move deterministically according to a numerical integrator and to draw a proposal for the following event time according to the upper bound of the switching rates and then accept or reject it by Poisson thinning. The framework we propose in Algorithms 8 and 10 is more general as approximations of all characteristics are possible. Moreover, the approximations in this manuscript do not require existence or knowledge of an upper bound for the switching rates. As discussed in Example 3.38 it is possible to closely resemble the proposal of [94] using our framework. Moreover, we obtain similar finite time strong and weak error results as in [94] by applying Theorem 3.15.

In [152] the authors focus on how to design a discrete time PDMP with a specific invariant measure. This is a fundamentally different approach to the focus of this chapter. A related work is [145], which defines a discrete time chain that resembles a BPS.

The book [40] discusses approximations of PDMP based on finite volume schemes for the Chapman-Kolmogorov type equations. Such schemes approximate the law of

the process and are thus different in nature compared to this manuscript.

Finally, we discuss papers that deal with continuous time approximations of the ZZS. In [121] the authors propose an approximation that relies on an integrator and a root finding method to generate the random event times. The paper [84] discusses the effect of approximate switching rates $\{\tilde{\lambda}\}_{i=1}^d$ on the stationary measure of the ZZS. This approximation relies on the availability of suitable $\{\tilde{\lambda}\}_{i=1}^d$ for which it is possible to (efficiently) simulate the corresponding ZZS. In Examples 3.40 and 3.52 we discuss applications of our theory to these approximation schemes. A similar setting is considered in [65, Theorems 11 and 25], where the authors establish bounds in total variation distance between (the invariant measures of) two PDMPs with same deterministic dynamics, but different switching rates and jump kernels. The authors prove such bounds by a coupling of the two continuous time PDMPs that is similar in spirit to our Coupling 3.57. In this chapter, in particular in Section 3.4.2, we bound the TV distance between a PDMP and a discrete time approximation. Thus the statements and proofs differ from [65] in this sense.

Organisation of the chapter

The chapter is organised as follows. In Section 3.2 we define notation that we use throughout the chapter. In Section 3.3 we describe the setting and the proposed algorithms. In particular in Section 3.3.1 we discuss first order schemes and in Section 3.3.3 we consider higher order schemes. Section 3.4 contains the main results together with the required assumptions. This section is divided into three parts. Section 3.4.1 is devoted to convergence in Wasserstein distance, which is established in Theorem 3.15. Section 3.4.2 concerns convergence in total variation as stated in Theorem 3.23. Section 3.4.3 gives conditions for uniform in time convergence of the weak error, as expressed by Theorem 3.30. In Section 3.5 we gather examples to demonstrate the when the assumptions of the main theorems are satisfied. The proofs of the three main theorems can be found respectively in Section 3.6, Section 3.7 and Section 3.8. All other results as well as all auxiliary lemmas from Sections 3.4.1, 3.4.2, and 3.4.3 can be found respectively in Appendix 3.A, Appendix 3.B, and Appendix 3.C.

3.2 Notation

We denote the semigroup of the continuous time PDMP, $\{Z_t\}_{t\geq 0}$, as $\mathcal{P}_t$ which acts on suitable functions by

$$\mathcal{P}_t f(z) = \mathbb{E}_z[f(Z_t)].$$

Here the subscript z denotes that the process Z_t has initial position $Z_0 = z$. Note that the semigroup is related to the transition probability of Z_t, which is denoted by $\mathcal{P}_t(z, A)$. These concepts are related for any function f and measurable set $A \subseteq E$ by

$$\mathcal{P}_t f(z) = \int f(y)\mathcal{P}_t(z, \mathrm{d}y), \quad \mathcal{P}_t(z, A) = (\mathcal{P}_t \mathbb{1}_A)(z).$$

Similarly we denote the transition probability of the approximation processes $\{\overline{Z}_{t_n}\}_{n\in\mathbb{N}}$ as $\overline{\mathcal{P}}_{t_n}$.

Consider a metric $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ and let P, Q be probability measures on $\mathbb{R}^d$. Then we define the Wasserstein distance of order 1 with respect to the metric d as

$$\mathcal{W}_1(P,Q) = \inf_{R\in\Pi(P,Q)} \left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} d(x,y) R(\mathrm{d}x, \mathrm{d}y) \right\}, \tag{3.1}$$

where $\Pi(P,Q)$ is the set of couplings of the two probability measures P, Q, that is the set of probability measures R on $\mathbb{R}^d \times \mathbb{R}^d$ such that $R(A, \mathbb{R}^d) = P(A)$ and $R(\mathbb{R}^d, B) = Q(B)$.

We will denote a norm by $\|\cdot\|$. The maximum between $a \in \mathbb{R}$ and 0 is denoted by $(a)_+ = \max\{a, 0\}$.

Let us define the space $\mathcal{C}^k$ to be the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ which are k times continuously differentiable. $\mathcal{C}^k_b$ ($\mathcal{C}^k_c$ respectively) denotes the subset of $\mathcal{C}^k$ to functions which are bounded (resp. have compact support) with bounded and continuous derivatives up to order k. We endow the space $\mathcal{C}_b$ with the supremum norm $\|\cdot\|_\infty$ and the space $\mathcal{C}^1_b$ is endowed with the norm

$$\|f\|_{\mathcal{C}^1_b} = \|f\|_\infty + \sum_{i=1}^d \|\partial_i f\|_\infty.$$

Consider a random variable I with values in $\{1, \dots, m\}$ such that $\mathbb{P}(I = i) = w_i$ for $i = 1, \dots, m$. Then we say I has a discrete distribution with probabilities w_i and we denote this as $I \sim \mathrm{Discrete}(\{w_i\}_{i=1}^m)$.

Given a measure π we define for any $f \in L^1_\pi$

$$\pi(f) = \int f(z)\pi(\mathrm{d}z).$$

Similarly for a probability kernel $Q(x, \mathrm{d}y)$ we write

$$Qf(x) = \int f(y) Q(x, \mathrm{d}y) \tag{3.2}$$

for f measurable and integrable with respect to $Q(x, \mathrm{d}y)$ for all x. Note that (3.2) allows us to consider a probability kernel as a map from the space of bounded and measurable functions, B_b, to B_b.

Let us define the *total variation distance* between two probability measures μ, and ν as

$$\|\mu - \nu\|_{TV} = \sup_{f\in\mathcal{C}_b : \|f\|_\infty \le 1} |\mu(f) - \nu(f)|.$$

Given a vector field Φ we can view this as a map, $\mathbf{\Phi}$, which acts on $\mathcal{C}^1$ functions as

$$\mathbf{\Phi}(f)(x) = \Phi(x)^T \nabla f(x), \quad \text{for } f \in \mathcal{C}^1.$$

Given two maps $X, Y : \mathcal{C}^\infty \to \mathcal{C}^\infty$ we shall define the commutator of X and Y, $[X, Y]$ to be the map $[X, Y] : \mathcal{C}^\infty \to \mathcal{C}^\infty$ by

$$[X, Y]f = XYf - YXf, \qquad \text{for } f \in \mathcal{C}^\infty.$$

We will use this with the maps $\mathbf{\Phi}$ and Q. If we assume that Q preserves $\mathcal{C}^1$ and $\mathbf{\Phi}$ is bounded then we can view $[\mathbf{\Phi}, Q] : \mathcal{C}_b^1 \to B_b$ defined by

$$[\mathbf{\Phi}, Q]f = \mathbf{\Phi}(Qf) - Q\mathbf{\Phi}(f), \qquad \text{for } f \in \mathcal{C}_b^1.$$

Note that although the commutator was defined for smooth vector fields the above definition makes sense for all C_b^1-functions since $\mathbf{\Phi}$ is a map from $\mathcal{C}_b^1$ to $\mathcal{C}_b$ and we have $Q(\mathcal{C}_b^1) \subseteq \mathcal{C}_b^1$, $Q(\mathcal{C}_b) \subseteq \mathcal{C}_b$ so both the operations $Q(\mathbf{\Phi} f)$ and $\mathbf{\Phi}(Qf)$ are well defined for $f \in \mathcal{C}_b^1$.

3.3 Algorithms

Consider a PDMP $(Z_t)_{t \geq 0}$ taking values on a state space E, which is a subset of a finite dimensional vector space. Examples are $E = \mathbb{R}^d \times \mathbb{R}^d$ or $E = \mathbb{R}^d \times \{-1, +1\}^d$. The dynamics of the process are described by the generator $\mathcal{L}$, which applied on a function in the domain of the extended generator gives

$$\mathcal{L}f(z) = \langle \Phi(z), \nabla_z f(z) \rangle + \sum_{i=1}^m \lambda_i(z) \int_E (f(y) - f(z)) Q_i(z, \mathrm{d}y). \tag{3.3}$$

The generator here is understood to be the extended generator, see [48, Theorem 26.14] for the exact description of the domain of the extended generator. Note, in particular that functions that are differentiable in the direction Φ and bounded are included in the domain. Here Φ is a smooth and globally Lipschitz vector field, $\lambda_i : E \to [0, \infty)$ are continuous functions and Q_i are probability kernels. Let φ_t denote the integral curve of Φ. Note that φ_t exists since Φ is globally Lipschitz. We assume that φ_t leaves E invariant. Define the total switching rate

$$\lambda(z) = \sum_{i=1}^m \lambda_i(z).$$

As shown in [48, Section 26] (3.3) corresponds to a PDMP where the next event time is distributed as

$$\mathbb{P}_z(\tau \leq t) = 1 - \exp\left(- \int_0^t \lambda(\varphi_s(z)) \mathrm{d}s \right), \tag{3.4}$$

Algorithm 6: Pseudo-code for the simulation of a PDMP

Input : Time horizon T, initial condition z.
Set $t = 0$, $Z_0 = z$;
while $t < T$ **do**
  simulate next event time as

$$\tau = \inf\left\{r > 0 : 1 - \exp\left(-\int_0^r \lambda(\varphi_s(Z_t))\mathrm{d}s\right) \geq U\right\}$$

  where $U \sim \mathrm{Unif}[0,1]$;
  simulate $Z_{t+s} = \varphi_s(Z_t)$ for $s \in (0, \tau)$;
  draw $I \sim \mathrm{Discrete}\big(\big\{\frac{\lambda_i(Z_{t+\tau-})}{\lambda(Z_{t+\tau-})}\big\}_{i=1}^m\big)$;
  simulate $Z_{t+\tau} \sim Q_I(Z_{t+\tau-}, \cdot)$;
  set $t = t + \tau$;
end

and that between two random events the process follows the flow-map φ_t, i.e. $Z_t = \varphi_t(z)$. At event time, τ, the process jumps according to probability kernel Q_I, where I is distributed according to the following discrete distribution

$$I \sim \mathrm{Discrete}\left(\left\{\frac{\lambda_i(\varphi_\tau(z))}{\lambda(\varphi_\tau(z))}\right\}_{i=1}^m\right).$$

Algorithm 6 describes the simulation procedure for a PDMP with generator (3.3).

Remark 3.1. It is possible to rewrite (3.3) to the form

$$\mathcal{L}f(z) = \langle \Phi(z), \nabla_z f(z)\rangle + \lambda(z)\int_E (f(y) - f(z))Q(z, \mathrm{d}y) \tag{3.5}$$

for some continuous function $\lambda : E \to [0, \infty)$ and probability kernel Q. Indeed this can be achieved by setting

$$\lambda(z) = \sum_{i=1}^m \lambda_i(z), \quad Q(z, \mathrm{d}y) = \sum_{i=1}^m \frac{\lambda_i(z)}{\lambda(z)} Q_i(z, \mathrm{d}y). \tag{3.6}$$

Therefore there is no loss of generality for the PDMP to take $m = 1$. However we will see in Section 3.4.1 that allowing $m \geq 1$ leads to weaker assumptions for our convergence results, in particular we will see in Example 3.35 a case where the assumptions are satisfied with $m > 1$ but would not be satisfied when written in the form (3.5).

The focus of this chapter is to define and analyse approximations of PDMPs that can be employed in settings where their simulation cannot be performed exactly. As explained in the introduction, there are three quantities which characterise a PDMP

and may be difficult to simulate. These are the flow map φ_t, the random event times with rates λ_i, and the Markov kernels Q_i. The idea is then to introduce p-th order approximations of the three characteristics for some $p \geq 1$. Precise conditions on the approximations are given in Assumptions 3.10, 3.11, 3.12, but here we provide a heuristic description. The flow map $\varphi_t(z)$ can be approximated with a numerical integrator, which is denoted as $\overline{\varphi}_t(z;\delta,p)$. The parameters δ, p have the meaning that for $s \in [0,\delta]$ we have that $\overline{\varphi}_s(z;\delta,p)$ is an approximation of order δ^p of $\varphi_s(z)$. Classical examples of numerical integrators from the ODE literature include the Euler discretisation, the leap frog scheme, and higher order numerical schemes. Then we want to approximate the switching rates in such a way that the random times (3.4) can be simulated easily at the cost of a small error. This can be done by using order δ^p approximations of $\lambda(\varphi_s(z))$, i.e. the switching rate along the deterministic flow. We denote the corresponding approximation as $\overline{\lambda}(z,s;\delta,p) : E \times [0,\infty) \to [0,\infty)$. The motivation is to ensure the following as an approximation of order δ^p for $t \leq \delta$:

$$\mathbb{P}_z(\tau \leq t) \approx 1 - \exp\left(-\int_0^t \overline{\lambda}(z,s;\delta,p)\mathrm{d}s\right).$$

Here

$$\overline{\lambda}(z,s;\delta,p) = \sum_{i=1}^{m} \overline{\lambda}_i(z,s;\delta,p).$$

Let us give some examples with $p = 1$. A possible choice is to "freeze" the switching rate, thus taking $\overline{\lambda}_i(z,s;\delta,1) = \lambda_i(z)$. This is supported by the intuition that $\lambda(\varphi_s(z)) \approx \lambda(z)$ for small s. In this case $\mathbb{P}_z(\tau \leq t)$ is approximately equal to $1 - \exp(-t\lambda(z))$, which is the cumulative distribution function of the exponential distribution with constant rate $\lambda(z)$. We refer to the $\overline{\lambda}_i$ as *frozen switching rates* and to the corresponding approximation process as *Euler approximation*. Alternatively one could take $\overline{\lambda}_i(z,s;\delta,1) = \lambda_i(\varphi_\delta(z))$, or the switching rates along the trajectory given by the numerical integrator $\overline{\lambda}_i(z,s;\delta,1) = \lambda_i(\overline{\varphi}_s(z;\delta,1))$, or more generally $\overline{\lambda}_i(z,s;\delta,p) = \lambda_i(\overline{\varphi}_s(z;\delta,p))$. Finally, consider the Markov kernels Q_i. We define a function F_i which describes a choice of implementation of Q_i. Let $F_i : E \times \mathcal{U} \to E$ be a deterministic map such that $F_i(z,U)$ is distributed according to $Q_i(z,\cdot)$ when U is distributed according to a probability distribution $\nu_{\mathcal{U}}$. We can then approximate each map F_i by a map $\overline{F}_i(\cdot;\delta,p) : E \times \mathcal{U} \to E$, where once again δ, p denotes the order of accuracy of our estimate. To simplify the notation, when we consider first order schemes, i.e. $p = 1$, we shall suppress the p-dependence and write $\overline{\varphi}_s(z;\delta)$, $\overline{\lambda}_i(z,s;\delta)$, $\overline{F}_i(z,U;\delta)$.

Now that we have introduced the problem and the various approximations we wish to exploit, we illustrate how to design first order and higher order approximation schemes for PDMPs. By an *order p scheme* we mean an approximation process for which the *local error*, i.e. the error between the PDMP and the approximation over a step of size δ with identical initial conditions, is proportional to δ^{p+1}. Therefore after n steps of size δ the *global error* is proportional to $t_n\delta^p$ where $t_n = n\delta$, which motivates the term order p scheme.

3.3.1 First order schemes

Let us introduce a mesh $\{t_n\}_{n\in\mathbb{N}}$ for the time variable where $t_n = \sum_{\ell=1}^{n} \delta_\ell$, and δ_ℓ are step sizes. For example if the step size is constant $\delta_\ell = \delta$ then $t_n = n\delta$ for all $n \in \mathbb{N}$. In this section we introduce two alternative first order schemes: the FD-PDMP algorithm and the PD-PDMP algorithm. We define the FD-PDMP approximation $\{\overline{Z}_{t_n}\}$ on the mesh $\{t_n\}_{n\in\mathbb{N}}$ by setting $\overline{Z}_0 = z$ and then following the procedure

$$\begin{aligned}\tilde{Z}_{t_{n+1}} &= \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n}; \delta_{n+1}),\\ \overline{Z}_{t_{n+1}} &= \alpha_{n+1}\overline{F}_{\overline{I}_{n+1}}(\tilde{Z}_{t_{n+1}}, U_{n+1}; \delta_{n+1}) + (1-\alpha_{n+1})\tilde{Z}_{t_{n+1}}.\end{aligned}$$

Here we have $U_{n+1} \sim \nu_{\mathcal{U}}$. The value of α_{n+1} is determined as follows. We simulate $\overline{\tau}$ which has distribution conditional on $\overline{Z}_{t_n}$ given by

$$\mathbb{P}_z(\overline{\tau} \le t | \overline{Z}_{t_n}) = 1 - \exp\left(-\int_0^t \overline{\lambda}(\overline{Z}_{t_n}, s; \delta_{n+1})\mathrm{d}s\right).$$

Then $\alpha_{n+1} = 1$ if and only if $\overline{\tau} \le \delta_{n+1}$, otherwise $\alpha_{n+1} = 0$. We then draw

$$\overline{I}_{n+1} \sim \text{Discrete}\left(\left\{\frac{\overline{\lambda}_i(\overline{Z}_{t_n}, \overline{\tau}; \delta_{n+1})}{\overline{\lambda}(\overline{Z}_{t_n}, \overline{\tau}; \delta_{n+1})}\right\}_{i=1}^{m}\right). \tag{3.7}$$

The resulting Markov chain $\overline{Z}_{t_n}$ is thus updated by first following the approximate flow map and then establishing whether a random event takes place at the end of the current time interval. This procedure is written in pseudo-code form in Algorithm 7. Note that if $\overline{\lambda}(z, s; \delta_{n+1})$ is independent of s, i.e. $\overline{\lambda}(z, s; \delta_{n+1}) = \overline{\lambda}(z; \delta_{n+1})$, then we do not need to simulate $\overline{\tau}_{n+1}$ and we have that α_{n+1} is a Bernoulli random variable with success rate $1 - \exp(-\delta_{n+1}\overline{\lambda}(z; \delta_{n+1}))$ and $\overline{I}_{n+1}$ is distributed as

$$\overline{I}_{n+1} \sim \text{Discrete}\left(\left\{\frac{\overline{\lambda}_i(\overline{Z}_{t_n}; \delta_{n+1})}{\overline{\lambda}(\overline{Z}_{t_n}; \delta_{n+1})}\right\}_{i=1}^{m}\right).$$

This is for instance the case of frozen switching rates.

A different approach is shown in Algorithm 8, which describes the PD-PDMP approximation. Here the idea is to simulate the switching time $\overline{\tau}$ with rate $\overline{\lambda}(\overline{Z}_{t_n}, s; \delta_{n+1})$, then if $\overline{\tau}$ is before the end of the current time step set $t = t_n + \overline{\tau}$, draw $\overline{I}_{n+1}$ as in (3.7), and follow the procedure below:

$$\begin{aligned}\overline{Z}_t \quad &= \overline{F}_{\overline{I}_{n+1}}(\tilde{Z}_t, U_{n+1}; \delta_{n+1}), \quad \text{where } \tilde{Z}_t = \overline{\varphi}_{\overline{\tau}}(\overline{Z}_{t_n}; \delta_{n+1}),\\ \overline{Z}_{t_{n+1}} &= \overline{\varphi}_{t_{n+1}-t}(\overline{Z}_t; \delta_{n+1}).\end{aligned}$$

On the other hand, when $\overline{\tau} > \delta_{n+1}$ the process is simply moving deterministically according to the approximate flow map, i.e. $\overline{Z}_{t_{n+1}} = \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n}; \delta_{n+1})$. Only one random jump per time step is allowed, and in this case it happens at time $\overline{\tau}$ instead

Algorithm 7: Fully Discrete Approximation of a PDMP

Input : Number of iterations N, initial condition z, step sizes $(\delta_n)_{n=0}^N$.
Output: Chain $(\overline{Z}_{t_n})_{n=0}^N$.
Set $n = 0$, $\overline{Z}_0 = z$;
while $n < N$ **do**
  simulate $\tilde{Z} = \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n}; \delta_{n+1})$;
  simulate

$$\overline{\tau} = \inf\left\{ r > 0 : 1 - \exp\left(- \int_0^r \overline{\lambda}(\overline{Z}_{t_n}, s; \delta_{n+1}) \mathrm{d}s \right) \geq U \right\}$$

  where $U \sim \mathrm{Unif}[0,1]$;
  if $\overline{\tau} \leq \delta_{n+1}$ **then**
    draw $U_{n+1} \sim \nu_{\mathcal{U}}$ and $\overline{I}_{n+1} \sim \mathrm{Discrete}\left(\left\{ \frac{\overline{\lambda}_i(\overline{Z}_{t_n}, \overline{\tau}; \delta_{n+1})}{\overline{\lambda}(\overline{Z}_{t_n}, \overline{\tau}; \delta_{n+1})} \right\}_{i=1}^m \right)$;
    set $\tilde{Z} = \overline{F}_{\overline{I}_{n+1}}(\tilde{Z}, U_{n+1}; \delta_{n+1})$;
  end
  set $\overline{Z}_{t_{n+1}} = \tilde{Z}$;
  set $n = n + 1$;
end

of at the end of the time step. This choice comes with advantages and disadvantages if compared to Algorithm 7. As we shall see in Sections 3.4.2 and 3.4.3 it is possible to obtain stronger results under weaker assumptions on the PDMP in the setting of Algorithm 8 compared to Algorithm 7. However this may come at a larger computational cost (see e.g. Example 3.3).

3.3.2 Examples

In this section we introduce several examples, which will be revisited as illustrative applications of our results. In the first three examples, i.e. Examples 3.2, 3.3, 3.4, we discuss MCMC samplers which target a probability measure with density $\pi(x) \propto \exp(-\psi(x))$ for $x \in \mathbb{R}^d$.

Example 3.2 (Zig-Zag sampler [23])**.** *Recall the ZZS we described in Example 2.11, with generator given in* (2.32) *and rates* $\lambda_i(x, v) = (v_i \partial_i \psi(x))_+$*. Simulating the event times with rates of this form is in general a very challenging problem as the integral in* (3.4) *cannot be computed for general potentials* ψ*.*

We can apply Algorithm 7 to the ZZS to obtain $(\overline{X}_{t_{n+1}}, \overline{V}_{t_{n+1}})$ *given the previous state by first simulating the next switching time* $\overline{\tau}$ *with rate* $\overline{\lambda}((\overline{X}_{t_n}, \overline{V}_{t_n}), s; \delta_{n+1})$ *and then*

$$\overline{X}_{t_{n+1}} := \overline{X}_{t_n} + \overline{V}_{t_n} \delta_{n+1}$$

Algorithm 8: Partially Discrete Approximation of a PDMP

Input : Number of iterations N, initial condition z, step sizes $(\delta_n)_{n=0}^N$.
Output: Chain $(\overline{Z}_{t_n})_{n=0}^N$.
Set $n = 0$, $\overline{Z}_0 = z$;
while $n < N$ **do**
  simulate

$$\overline{\tau} = \inf\left\{ r > 0 : 1 - \exp\left(- \int_0^r \overline{\lambda}(Z_t, s; \delta_{n+1}) \mathrm{d}s \right) \geq U \right\}$$

  where $U \sim \mathrm{Unif}[0,1]$;
  if $\overline{\tau} < \delta_{n+1}$ **then**
    set $t = t_n + \overline{\tau}$;
    simulate $\tilde{Z}_t = \overline{\varphi}_{\overline{\tau}}(\overline{Z}_{t_n}; \delta_{n+1})$;
    draw $U_{n+1} \sim \nu_{\mathcal{U}}$ and $\overline{I}_{n+1} \sim \mathrm{Discrete}\left(\left\{ \frac{\overline{\lambda}_i(\overline{Z}_{t_n}, \overline{\tau}; \delta_{n+1})}{\overline{\lambda}(\overline{Z}_{t_n}, \overline{\tau}; \delta_{n+1})} \right\}_{i=1}^m \right)$;
    set $\overline{Z}_t = \overline{F}_{\overline{I}_{n+1}}(\tilde{Z}_t, U_{n+1}; \delta_{n+1})$;
    simulate $\overline{Z}_{t_{n+1}} = \overline{\varphi}_{t_{n+1}-t}(\overline{Z}_t; \delta_{n+1})$;
  else
    simulate $\overline{Z}_{t_{n+1}} = \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n}; \delta_{n+1})$;
  end
  set $n = n + 1$;
end

$$\overline{V}_{t_{n+1}} := \begin{cases} R_{\overline{I}_{n+1}} \overline{V}_{t_n} & \text{if } \overline{\tau} \leq \delta_{n+1}, \\ \overline{V}_{t_n} & \text{if } \overline{\tau} > \delta_{n+1}, \end{cases}$$

where $\overline{\lambda}(z, s; \delta) = \sum_{i=1}^d \overline{\lambda}_i(z, s; \delta)$, *and*

$$\overline{I}_{n+1} \sim \mathrm{Discrete}\left(\left\{ \frac{\overline{\lambda}_i((\overline{X}_{t_n}, \overline{V}_{t_n}), \overline{\tau}; \delta_{n+1})}{\overline{\lambda}((\overline{X}_{t_n}, \overline{V}_{t_n}), \overline{\tau}; \delta_{n+1})} \right\}_{i=1}^d \right).$$

The only approximation concerns the switching rates, whereas it is straightforward to simulate the linear dynamics and the jumps at event times. As mentioned above, a simple choice is to take $\overline{\lambda}_i((x,v), s; \delta) = \lambda_i(x,v)$, *which results in an Euler approximation of the ZZS. An alternative choice is*

$$\overline{\lambda}_i((x,v), s; \delta) = \frac{1}{\delta} \left(\psi(x + v_i e_i \delta) - \psi(x) \right)_+ + \gamma_i(x,v), \tag{3.8}$$

which is obtained by a finite difference scheme approximation for $\partial_i \psi$. *Here* e_i *is the* i*-th vector of the canonical basis. Observe that with this choice of* $\overline{\lambda}_i$ *the approximation is gradient free, as it does not require computing* $\nabla \psi$. *An approximation given by Algorithm 8 may be introduced analogously.*

Example 3.3 (Bouncy Particle Sampler [32, 125])**.** *Recall the BPS we introduced in Example 2.12. The BPS has generator*

$$\mathcal{L}f(x,v)=\langle v,\nabla_x f(x)\rangle+\lambda_1(x,v)[f(x,R(x)v)-f(x,v)]+\lambda_2\int\big(f(x,w)-f(x,v)\big)\nu(\mathrm{d}w),$$

where $\lambda_1(x,v)=(v^T\nabla_x\psi(x))_+$*. For the same reasons of ZZS, it is in general not possible to simulate the event times of BPS.*

For this process we introduce an approximation based on Algorithm 8. Let $U_{n+1}=(\mathscr{Z}_{n+1},\mathscr{U}_{n+1})$ *for* $\mathscr{Z}_{n+1}$ *distributed according to the standard Gaussian distribution* ν *and* $\mathscr{U}_{n+1}\sim\mathrm{Unif}([0,1])$ *is an independent uniform random variable. For* $n\geq 0$ *we define the next state of the approximation* $(\overline{X}_{t_{n+1}},\overline{V}_{t_{n+1}})$ *given the previous state by first simulating* $\overline{\tau}$ *with distribution* $\mathbb{P}_z(\overline{\tau}>t)=\exp(-\int_0^t\overline{\lambda}((\overline{X}_{t_n},\overline{V}_{t_n}),s;\delta_{n+1})\mathrm{d}s)$ *and then*

$$\begin{aligned}\overline{X}_{t_{n+1}}&\coloneqq\overline{X}_{t_n}+\overline{\tau}\overline{V}_{t_n}+(\delta_{n+1}-\overline{\tau})\overline{V}_{t_{n+1}},\\ \overline{V}_{t_{n+1}}&\coloneqq\begin{cases}F((\overline{X}_{t_n+\overline{\tau}},\overline{V}_{t_n}),U_{n+1}) & \text{if } \overline{\tau}\leq\delta_{n+1},\\ \overline{V}_{t_n} & \text{if } \overline{\tau}>\delta_{n+1}.\end{cases}\end{aligned}$$

Here $\overline{\lambda}((x,v),t;\delta_{n+1})=\overline{\lambda}_1((x,v),t;\delta_{n+1})+\lambda_r$ *where* $\overline{\lambda}_1((x,v),t;\delta_{n+1})$ *approximates* $\lambda_1(x+vt,v)$ *and*

$$F((\overline{X}_{t_n+\overline{\tau}},\overline{V}_{t_n}),U_{n+1})=\begin{cases}R(\overline{X}_{t_n+\overline{\tau}})\overline{V}_{t_n} & \text{if } \mathscr{U}_{n+1}>\frac{\lambda_r}{\overline{\lambda}((\overline{X}_{t_n},\overline{V}_{t_n}),\overline{\tau};\delta_{n+1})},\\ \mathscr{Z}_{n+1} & \text{if } \mathscr{U}_{n+1}\leq\frac{\lambda_r}{\overline{\lambda}((\overline{X}_{t_n},\overline{V}_{t_n}),\overline{\tau};\delta_{n+1})}.\end{cases}$$

It is thus clear that applying Algorithm 8 rather than Algorithm 7 can be more computationally expensive in the case of BPS, as when an event takes place $\nabla\psi$ *has to be evaluated at some midpoint* $\overline{X}_{t_n+\overline{\tau}}$ *in order to compute the reflection operator. In contrast,* $\nabla\psi$ *has to be computed only at gridpoints in Algorithm 7. We shall see in Section 3.4 that our theoretical results can only be applied to approximations of the BPS based on Algorithm 8, motivating the need for that algorithm.*

Similarly to the case of the ZZS described in Example 3.2, possible approximations of $\lambda_1(x+vt,v)$ *are* $\overline{\lambda}_1((x,v),t;\delta)=\lambda(x,v)$ *or*

$$\overline{\lambda}_1((x,v),t;\delta)=\frac{1}{\delta}\left(\psi(x+v\delta)-\psi(x)\right)_+.$$

The latter choice is not enough to not make the simulation of $(\overline{X}_t,\overline{V}_t)$ *gradient free because* $\nabla\psi$ *is needed in the computation of the reflection operator.*

Example 3.4 (Randomized Hamiltonian Monte Carlo algorithm)**.** *The randomized Hamiltonian Monte Carlo algorithm (see [30]) is defined on* $E=\mathbb{R}^d\times\mathbb{R}^d$ *by the generator*

$$\mathcal{L}f(q,p)=\langle p,\nabla_q f(q,p)\rangle-\langle\nabla_q\psi(q),\nabla_p f(q,p)\rangle+\lambda_r\int\big(f(q,p')-f(q,p)\big)\nu(\mathrm{d}p'),$$

where ν is a Gaussian measure on $\mathbb{R}^d$. The Hamiltonian flow cannot be simulated exactly in most cases, and thus it becomes necessary to approximate it by a numerical integrator $\overline{\varphi}_s$. Then according to Algorithm 7 we obtain the next state by first denoting $(\tilde{Q}_{t_{n+1}}, \tilde{P}_{t_{n+1}}) = \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_{n+1}}; \delta_{n+1})$ and thus

$$(\overline{Q}_{t_{n+1}}, \overline{P}_{t_{n+1}}) = \begin{cases} (\tilde{Q}_{t_{n+1}}, \tilde{P}_{t_{n+1}}) & \text{with probability } \exp(-\lambda_r \delta_{n+1}), \\ (\tilde{Q}_{t_{n+1}}, \mathscr{Z}) & \text{with probability } 1 - \exp(-\lambda_r \delta_{n+1}), \end{cases}$$

where $\mathscr{Z} \sim \nu$. We remark that the most efficient implementation is to simulate the next refreshment time and then follow the numerical integrator until then, without drawing a new refreshment time at each iteration.

Example 3.5 (Modelling the size of a cell)**.** *Following Section 1.5 in [143], denote the size of a cell by $z \in \mathbb{R}$. The cell grows in time with deterministic flow φ_t, and splits into two daughter cells with division rate $\lambda(z)$. Then denote as τ_n the time when a cell from the n-generation splits. The size of a daughter cell is half of the parent cell, and thus $Z_{\tau_n} = \frac{1}{2} Z_{\tau_n -}$. We can characterise the resulting process with its generator:*

$$\mathcal{L}f(z) = \langle \Phi(z), \nabla f(z) \rangle + \lambda(z) \left(f\left(\frac{z}{2}\right) - f(z) \right).$$

Therefore it may not be possible to simulate such a process if the desired φ and λ are complicated functions. An approximation can be obtained applying the ideas above introducing a numerical integrator $\overline{\varphi}$ and approximate division rate $\overline{\lambda}$.

Example 3.6 (Chemotaxis in Escherichia coli)**.** *It was shown in [12] that the bacteria Escheria coli have two types of behaviour describing their motion, which are called "runs" and "twiddles". When the bacteria is "running" it moves with near uniform speed. However when "twiddling" the bacteria changes direction very abruptly. We will describe this using the stochastic model as given in [147]. We describe the bacteria by giving its position $x \in \mathbb{R}^3$ and velocity $v \in \mathbb{S}^2$ at each time, where $\mathbb{S}^2$ is the sphere in $\mathbb{R}^3$. Then there exists a function $\lambda : [0, \infty) \times \mathbb{R}^3 \times \mathbb{S}^2 \to (0, \infty)$ which describes the next time the bacteria twiddles; at such a twiddle the velocity changes according to some probability measure μ_v on $\mathbb{S}^2 \setminus \{v\}$ where v is the velocity before the twiddle. The dynamics of the bacteria are given as a PDMP described by the backward equation*

$$\frac{\partial u}{\partial t}(t, x, v) + \langle v, \nabla_x u(t, x, v) \rangle + \lambda(t, x, v) \int_{\mathbb{S}^2} [u(t, x, \eta) - u(t, x, v)] \mu_v(\mathrm{d}\eta) = 0.$$

Note if λ is independent of t then we can describe this process by writing a generator in the form (3.3); otherwise we can extend the space to include a time variable and then write a corresponding generator in the form (3.3) which is given by

$$\mathcal{L}f(t, x, v) = \partial_t f(t, x, v) + \langle v, \nabla_x f(t, x, v) \rangle + \lambda(t, x, v) \int_{\mathbb{S}^2} [f(t, x, \eta) - f(t, x, v)] \mu_v(\mathrm{d}\eta).$$

We can introduce an approximation of this process by using frozen switching rates.

Algorithm 9: Second order Partially Discrete Approximation of a PDMP

Input : Number of iterations N, initial condition z, step sizes $(\delta_n)_{n=0}^N$.
Output: Chain $(\overline{Z}_{t_n})_{n=0}^N$.

Set $n = 0$, $\overline{Z}_0 = z$;
while $n < N$ **do**
  set $\tilde{Z} = \overline{Z}_{t_n}$;
  draw $U^1 \sim \text{Unif}[0,1]$ and simulate

$$\overline{\tau}_1 = \inf\left\{r > 0 : 1 - \exp\left(-\int_0^r \overline{\lambda}(\tilde{Z}, s; \delta_{n+1}, 2)\mathrm{d}s\right) \geq U^1\right\}$$

  if $\overline{\tau}_1 < \delta_{n+1}$ **then**
    draw $U_{n+1}^1 \sim \nu_{\mathcal{U}}$ and $\overline{I}_1 \sim \text{Discrete}\left(\left\{\frac{\overline{\lambda}_i(\tilde{Z}, \overline{\tau}_1; \delta_{n+1}, 2)}{\overline{\lambda}(\tilde{Z}, \overline{\tau}_1; \delta_{n+1}, 2)}\right\}_{i=1}^m\right)$;
    set $\tilde{Z} = \overline{\varphi}_{\overline{\tau}_1}(\tilde{Z}; \delta_{n+1}, 2)$;
    set $\tilde{Z} = \overline{F}_{\overline{I}_1}(\tilde{Z}, U_{n+1}^1; \delta_{n+1}, 2)$;
    draw $U^2 \sim \text{Unif}[0,1]$ and simulate

$$\overline{\tau}_2 = \inf\left\{r > 0 : 1 - \exp\left(-\int_0^r \overline{\lambda}(\tilde{Z}, s; \delta_{n+1}, 1)\mathrm{d}s\right) \geq U^2\right\}$$

    if $\overline{\tau}_2 < t_{n+1} - \overline{\tau}_1$ **then**
      draw $U_{n+1}^2 \sim \nu_{\mathcal{U}}$ and $\overline{I}_2 \sim \text{Discrete}\left(\left\{\frac{\overline{\lambda}_i(\tilde{Z}, \overline{\tau}_2; \delta_{n+1}, 1)}{\overline{\lambda}(\tilde{Z}, \overline{\tau}_2; \delta_{n+1}, 1)}\right\}_{i=1}^m\right)$;
      set $\tilde{Z} = \overline{F}_{\overline{I}_2}(\tilde{Z}, U_{n+1}^2; \delta_{n+1}, 1)$;
      simulate $\overline{Z}_{t_{n+1}} = \overline{\varphi}_{t_{n+1} - \overline{\tau}_2 - \overline{\tau}_1}(\tilde{Z}; \delta_{n+1}, 1)$;
    else
      simulate $\overline{Z}_{t_{n+1}} = \overline{\varphi}_{t_{n+1} - \overline{\tau}_1}(\tilde{Z}; \delta_{n+1}, 1)$;
    end
  else
    set $\overline{Z}_{t_{n+1}} = \overline{\varphi}(\tilde{Z}; \delta_{n+1}, 2)$;
  end
end

3.3.3 Higher order schemes

A natural question is how to obtain higher order schemes. The first important observation is that that the probability that a PDMP has more than one jump in a time interval of length δ is of order δ^2. Therefore in order to construct higher order schemes it is natural to allow multiple jumps in the same time step.

A detailed implementation of a higher order approximation scheme can be found in Algorithm 10. Let us first describe a second order algorithm. Starting at state

$\overline{Z}_{t_n} = z$, the proposed time for the first event is given by $\overline{\tau}$ where

$$\mathbb{P}(\overline{\tau} > r) = \exp\left(-\int_0^r \overline{\lambda}(z, s; \delta_{n+1}, 2)\mathrm{d}s\right).$$

If $\overline{\tau} < \delta_{n+1}$, the process moves according to the numerical flow $\overline{\varphi}_s(z; \delta_{n+1}, 2)$ for time $\overline{\tau}$, and at time $t_n + \overline{\tau}$ the random event takes place according to $\overline{F}_{\overline{I}}(\overline{Z}_{t_n+\overline{\tau}}, \cdot; \delta_{n+1}, 2)$, where $\overline{I}$ has discrete distribution. In this case, a second jump is allowed in the current time step. The simulation of this event can be made using first order approximations $\overline{\lambda}(\cdot, \cdot; \delta_{n+1}, 1), \overline{\varphi}_s(\cdot; \delta_{n+1}, 1)$, and $\overline{F}_i(\cdot, \cdot; \delta_{n+1}, 1)$.

Algorithm 10: Order p Partially Discrete Approximation of a PDMP

Input : Number of iterations N, initial condition z, step sizes $(\delta_n)_{n=0}^N$.
Output: Chain $(\overline{Z}_{t_n})_{n=0}^N$.
Set $n = 0$, $\overline{Z}_0 = z$;
while $n < N$ **do**
  set $q = p$, $\tilde{Z} = \overline{Z}_{t_n}$;
  set $t_{\text{left}} = \delta_{n+1}$;
  while $q > 0$ **do**
    simulate

$$\overline{\tau} = \inf\left\{r > 0 : 1 - \exp\left(-\int_0^r \overline{\lambda}(\tilde{Z}, s; \delta_{n+1}, q)\mathrm{d}s\right) \geq U\right\}$$

    where $U \sim \text{Unif}[0, 1]$;
    if $\overline{\tau} < t_{\text{left}}$ **then**
      draw $U_{n+1} \sim \nu_{\mathcal{U}}$ and $\overline{I} \sim \text{Discrete}\left(\left\{\frac{\overline{\lambda}_i(\tilde{Z}, \overline{\tau}; \delta_{n+1}, q)}{\overline{\lambda}(\tilde{Z}, \overline{\tau}; \delta_{n+1}, q)}\right\}_{i=1}^m\right)$;
      set $\tilde{Z} = \overline{\varphi}_{\overline{\tau}}(\tilde{Z}; \delta_{n+1}, q)$;
      set $\tilde{Z} = \overline{F}_{\overline{I}}(\tilde{Z}, U_{n+1}; \delta_{n+1}, q)$;
      set $q = q - 1$ and $t_{\text{left}} = t_{\text{left}} - \overline{\tau}$;
    else
      set $\tilde{Z} = \overline{\varphi}_{\delta_{n+1}}(\tilde{Z}; \delta_{n+1}, q)$;
      set $q = 0$;
    end
  end
  set $\overline{Z}_{n+1} = \tilde{Z}$, $n = n + 1$;
end

Let us consider as an example how to obtain a second order approximation for smooth switching rates. For $s \leq \delta$ the first order Taylor approximation of $\lambda_i(\varphi_s(z))$ is given by

$$\overline{\lambda}_i(z, s; \delta_{n+1}, 2) = \lambda_i(z) + s\langle\Phi(z), \nabla\lambda_i(z)\rangle.$$

Because the integral in (3.4) is with respect to s, this choice of $\overline{\lambda}_i(z, s; \delta_{n+1}, 2)$ is such that computing the corresponding switching time is equivalent to computing the root

of a second order polynomial. The downside is that an evaluation of the gradient of λ_i is needed and may be unavailable or expensive to compute. However, we can further approximate the product $\langle \Phi(z), \nabla \lambda_i(z) \rangle$ with a finite difference scheme to obtain for $s \leq \delta_{n+1}$ the expression

$$\overline{\lambda}_i(z, s; \delta_{n+1}, 2) = \lambda_i(z) + \frac{s}{\delta_{n+1}}(\lambda_i(\varphi_{\delta_{n+1}}(z)) - \lambda_i(z)), \tag{3.9}$$

which is a second order approximation provided λ is sufficiently smooth. The algorithm for $p = 2$ is given by Algorithm 9.

Similarly, it is possible to obtain an order $p > 2$ approximation. The simulation up to and counting the first event of each time step should be made according to approximations of order δ^p of the flow map, switching rates, and jump kernels. After the first event it is then possible to use approximations of order $p - 1$, then of order $p - 2$, and so on until one reaches the end of the current time interval, with the constraint that at most p events take place. Finally, it is clearly possible to use approximations of order δ^p for the simulation of all events in the same time step, although such approximations can be in general more expensive to compute.

3.4 Main results

3.4.1 Error bounds in Wasserstein distance

The main result of this section is Theorem 3.15, which shows convergence of the Wasserstein distance between the approximation and the continuous process as the step size goes to 0. We consider the Wasserstein distance of order 1 with respect to any normed distance, that is we take $d(x, y) = \|x - y\|$ in Equation (3.1) for any vector norm $\|\cdot\|$. For convenience we assume that for all $n \in \mathbb{N}$ we have an upper bound $\delta_n \leq \delta_0$.

Let us now state the assumptions on the process and on the various approximations that are required to show Theorem 3.15. We start with assumptions on the continuous time PDMP, and specifically from a condition on the deterministic dynamics. In particular, we require that Φ is Lipschitz.

Assumption 3.7. *For the vector field Φ there exists a constant $C > 0$ such that for all $z, z' \in E$ it holds that*

$$\|\Phi(z) - \Phi(z')\| \leq C\|z - z'\|.$$

We now shift our focus to the jump part of the process. In particular, we need the kernel $Q(z, \cdot)$ to satisfy the next conditions.

Assumption 3.8. *There exist constants $D_1, D_2, D_3 > 0$ such that for $\tilde{U} \sim \nu_{\mathcal{U}}$ the following conditions hold for all $i \in \{1, \ldots, m\}$:*

(a) For any $z \in E$

$$\mathbb{E}[\|z - F_i(z, \tilde{U})\|] \leq D_1.$$

(b) For all $z, z' \in E$

$$\mathbb{E}[\|F_i(z, \tilde{U}) - F_i(z', \tilde{U})\|] \leq D_2 \|z - z'\|.$$

(c) For all $z \in E$ *and all* $s \leq \delta \leq \delta_0$

$$\mathbb{E}\left[\|\varphi_{\delta - s}(F_i(\varphi_s(z), \tilde{U})) - F_i(\varphi_\delta(z), \tilde{U})\|\right] \leq D_3 \delta.$$

The first assumption asks that after a random jump the process is in expectation at bounded distance to its previous state, while condition (b) states that a Lipschitz condition with respect to the previous state holds for coupled jumps. Finally, condition (c) asks that the error committed by switching at the end of the time step or at an earlier time is of order δ if the two jumps are coupled. Moreover, the following Lipschitz condition for the switching rates is required.

Assumption 3.9. *There exists* $D_4 > 0$ *such that for all* $z, z' \in E$ *and* $i = 1, \ldots, m$

$$|\lambda_i(z) - \lambda_i(z')| \leq D_4 \|z - z'\|.$$

Let us now focus on the required assumptions on the various approximations employed in the approximation process. We state the assumptions for a general order of accuracy $p \geq 1$, with $p \in \mathbb{N}$. Starting from the deterministic dynamics, we assume that the numerical integrator for the flow map is an approximation of order p.

Assumption 3.10. *There exists* $\tilde{C} \geq 0$ *such that for any* $z \in E$ *and any* $0 \leq s \leq \delta \leq \delta_0$

$$\|\varphi_s(z) - \overline{\varphi}_s(z; \delta, p)\| \leq \tilde{C} s^{p+1}.$$

In case the flow map can be simulated exactly, one can simply take $\overline{\varphi}_s = \varphi_s$ and $\tilde{C} = 0$. Next we focus on the approximate jump kernels $\overline{F}_i$.

Assumption 3.11. *The approximate jump kernels* $\overline{F}_i : E \times \mathcal{U} \times [0, \delta_0] \to E$, *satisfy for any* $z \in E$ *and* $\delta \in (0, \delta_0]$

$$\mathbb{E}_z[\|\overline{F}_i(z, \tilde{U}; \delta, p) - F_i(z, \tilde{U})\|] \leq M_1 \delta^p$$

for all $i = 1, \ldots, m$.

Let us now state the requirement on the approximate switching rates $\overline{\lambda}_i$.

Assumption 3.12. *The following conditions hold:*

(a) There exists $\overline{M}_2(z)$ *such that for all* $0 \le s \le \delta \le \delta_0$ *and* $i \in \{1, \dots, m\}$

$$|\overline{\lambda}_i(z, s; \delta, p) - \lambda_i(\varphi_s(z))| \le \delta^p \overline{M}_2(z).$$

(b) For any $n \in \mathbb{N}$ *there is a function* $M_2(t, z)$ *such that*

$$\mathbb{E}_z \left[\overline{M}_2(\overline{Z}_{t_n}) \right] \le M_2(t_n, z) < \infty.$$

As a final assumption, we require that both the continuous time PDMP and the approximation process have almost surely bounded norm for a finite time horizon. This assumption is verified for instance if the state space is compact, or if the processes travel with bounded velocity.

Assumption 3.13. *For any* $t > 0$ *there exists* $B(t, z) > 0$ *such that almost surely both* $\|Z_t\| \le B(t, z)$ *and* $\|\overline{Z}_t\| \le B(t, z)$*, where* $Z_0 = \overline{Z}_0 = z$*.*

Remark 3.14. Let us comment on these assumptions:

- It is worth observing that conditions such as Assumption 3.9 can be weakened to forms such as

  $$|\lambda_i(z) - \lambda_i(z')| \le D_4 \|z - z'\|(1 + \|z\|^q + \|z'\|^{q'}),$$

  for some $q, q' \in \mathbb{N}$. This is because by Assumption 3.13 the norms at time t of the two processes are bounded almost surely and therefore for some $M(t)$ we have

  $$(1 + \|Z_t\|^q + \|\overline{Z}_t\|^{q'}) \le M(t) < \infty$$

  almost surely. A similar reasoning can be applied to other assumptions that have this structure. For simplicity we will not consider this set of weakened assumptions in the proof of Theorem 3.15, but we remark that the extension is straightforward.

- In both Example 3.2 on the ZZS and Example 3.3 on the BPS we can write λ of the form

  $$\lambda(x, v) = f(r)$$

  where $r = \partial_i \psi(x) v_i$ for ZZS or $r = \langle \nabla \psi(x), v \rangle$ for BPS and $f(r) = r_+$. Note that it is possible to take a smooth function f for which the process still has the desired invariant measure, see [2]. We will demonstrate some choices of $\overline{\lambda}$ for ZZS which satisfy Assumption 3.12, and analogous choices hold for BPS. For smooth λ we can use (3.9) to obtain a second order approximation or similarly a p-th order finite difference scheme to have an order p approximation. However if $\lambda(x, v) = (v_i \partial_i \psi(x))_+$ is only Lipschitz then this approximation is no longer valid; instead we can write

  $$\overline{\lambda}_i((x, v), s; \delta, p) = (\overline{\partial_i \psi}((x, v), s; \delta, p)\, v_i)_+$$

where $\overline{\partial_i \psi}((x,v),s;\delta,p)$ is a p-th order approximation in s of $\partial_i\psi(x+sv)$ and can be obtained either by a truncated Taylor expansion or using a finite difference scheme. Then using that $(\cdot)_+$ is 1-Lipschitz

$$|\overline{\lambda}_i((x,v),s;\delta,p) - \lambda_i(\varphi_s(x,v))| \leq |\overline{\partial_i\psi}((x,v),s;\delta,p) - \partial_i\psi(x+sv)| \leq \overline{M}_2\delta^p.$$

For example, for ψ sufficiently smooth, we can take

$$\overline{\lambda}_i((x,v),s;\delta,p) = \left(\sum_{q=0}^{p-1} \frac{(sv_i)^q}{q!}\Delta^{q+1}_{i,\delta,p-q}\psi(x)\right)_+,$$

where $\Delta^q_{\delta,p-q}\psi$ denotes the δ^{p-q}-th order approximation of the q-th derivative of ψ in the variable x_i.

We are ready to state the main result of this section.

Theorem 3.15. *Let $p \geq 1$. Denote by $\{\mathcal{P}_t\}_{t\geq 0}$ the semigroup of a PDMP with generator* (3.3)*, which satisfies Assumptions 3.7-3.9. Denote by $\overline{\mathcal{P}}_t$ the transition probability of the Markov chain described by either Algorithms 7 or 8 in the case $p = 1$, or by Algorithm 10 for $p > 1$. Suppose that $\overline{\varphi}_t(\cdot;\delta,q)$, $\overline{\lambda}(\cdot,\cdot;\delta,q)$, $\overline{F}_i(\cdot;\delta,q)$ satisfy Assumptions 3.10-3.13 for some $\delta_0 > 0$ and for every $1 \leq q \leq p$ with $q \in \mathbb{N}$. Then for a fixed $T > 0$ there exist $K_1 = K_1(T)$, $K_2 = K_2(T)$ such that for any mesh $0 = t_0 < t_1 < \ldots < t_N = T$ with $\delta_n = t_n - t_{n-1}$ and $\delta_n \leq \delta_0$ for any $n \leq N$*

$$\mathcal{W}_1(\mathcal{P}_T(z,\cdot),\overline{\mathcal{P}}_T(z,\cdot)) \leq K_2 \sum_{k=1}^{N} \delta_k^{p+1} \left(\prod_{\ell=k}^{N}(1+\delta_\ell K_1)\right).$$

If the step size is uniform, i.e. $\delta_n = \delta$ and $t_n = n\delta$, then

$$\mathcal{W}_1(\mathcal{P}_T(z,\cdot),\overline{\mathcal{P}}_T(z,\cdot)) \leq \delta^p \left(e^{TK_1} - 1\right)\frac{K_2}{K_1}.$$

Proof of Theorem 3.15. The proof of Theorem 3.15 can be found in Section 3.6. □

We now give a setting in which Assumption 3.8 simplifies. This is motivated by and includes the ZZS. Let us now consider a PDMP $Z_t = (X_t, V_t) \in \mathbb{R}^n \times \mathcal{V}$, where X_t and V_t should be interpreted as the position and velocity at time t. Here $\mathcal{V}$ is some subset of Euclidean space. Consider the case in which the deterministic dynamics with initial condition (x,v) are of the form

$$\begin{cases} \dot{x} = \Phi(v), \\ \dot{v} = 0. \end{cases}$$

Therefore the deterministic motion is $X_t = \varphi_t(x,v)$ and $V_t = v$ if $(X_0,V_0) = (x,v)$. Then assume that the random events affect only the velocity, and leave the position unchanged, i.e. $F_i((x,v),U) = (x, F_i^v((x,v),U))$. This is the setting for example of the ZZS and BPS. Consider the following assumption.

Assumption 3.16. *The space $\mathcal{V}$ is such that for all $v, w \in \mathcal{V}$ with $v \neq w$ it holds that*

$$0 < V_{min} \leq \|v - w\| \leq V_{max} < \infty.$$

Assume also that there exists $D > 0$ such that for any $x, y \in \mathbb{R}^n$, $i \in \{1, \dots, m\}$ and $v \in \mathcal{V}$

$$\mathbb{E}_{(x,v)}[\|F_i^v((x,v),U) - F_i^v((y,v),U)\|] \leq D\|x - y\|.$$

The next corollary states that in this setting Assumption 3.16 implies Assumption 3.8.

Corollary 3.17. *Consider a PDMP of the particular form described above. Suppose Assumptions 3.7, 3.9-3.12, as well as Assumption 3.16 hold. Then Theorem 3.15 applies.*

Proof. The proof can be found in Appendix 3.A.2. □

Finally, we consider the setting in which we have a deterministic upper bound for the switching rates, but the process is not almost surely bounded as was required by Assumption 3.13. This is the case for instance of the Randomized HMC algorithm [30]. We shall show that in this case Theorem 3.15 holds as long as for a finite time horizon the processes are bounded in expectation. The formal condition is the following.

Assumption 3.18. *There exists a constant $\lambda_{max} > 0$ such that $\lambda(z) \leq \lambda_{max}$ for all $z \in E$. Moreover there exists $L(t, z) < \infty$ such that*

$$\max\{\mathbb{E}_z[\|Z_t\|], \mathbb{E}_z[\|\overline{Z}_t\|]\} \leq B(t, z).$$

Proposition 3.19. *Suppose Assumptions 3.7-3.12 and 3.18 hold. Then Theorem 3.15 applies.*

Proof. The proof is given in Appendix 3.A.3. □

3.4.2 Error bounds in total variation distance

In this section we show that a bound of order δ^p on the total variation distance between the approximation and the PDMP can be derived for Algorithm 10 assuming it is possible to simulate exactly the flow φ_t and the Markov kernels Q_i. Interestingly, this result can be proved under considerably weaker assumptions on the PDMP compared to what is considered in Section 3.4.1. We remark in particular that no assumption on the maps F_i is needed, which was the case in Assumption 3.8. Moreover the process needs not be bounded almost surely for finite time horizons, as described in Assumption 3.13. The main result of this section is proved by coupling the event times of the PDMP and of the approximations via Poisson thinning. It follows that

with a positive probability the processes, which are initialised at the same point, will remain together during a time step.

Let us state the required assumptions on the switching rates and on the continuous time process. Recall that for first order approximations of the characteristics we drop the specific order of accuracy, e.g. for switching rates we have $\overline{\lambda}_i(z,s;\delta) = \overline{\lambda}_i(z,s;\delta,1)$ for $i = 1,\dots,m$. We distinguish the assumptions between the setting $p = 1$ and $p > 1$. In the case $p = 1$ we impose the following assumption.

Assumption 3.20. *Each of the approximate switching rates $\overline{\lambda}_i(\cdot;\delta)$ for $i = 1,\dots,m$ satisfies Assumption 3.12(a) with $p = 1$ for some $\overline{M}_2(z)$. Furthermore for $z \in E$ and $s \geq 0$ define $\overline{\lambda}(z,s;\delta) = \sum_{i=1}^m \overline{\lambda}_i(z,s;\delta)$, and $\lambda_{tot}(z,s;\delta) = \lambda(z) + \overline{\lambda}(z,s;\delta) + m$. Let $T > 0$. Then there exist $L_1(T,z)$, $L_2(T,z)$, $L_3(T,z) < \infty$ such that for any mesh $0 = t_0 < t_1 < \cdots < t_N = T$ with $t_{k+1} - t_k = \delta_{k+1}$ and $N \in \mathbb{N}$ the following conditions hold:*

$$\sup_{n\leq N}\sup_{i=1,\dots,m}\sup_{s\in[0,\delta_n]}\sup_{r\in[s,\delta_n]}\mathbb{E}_z\Big[\lambda(\varphi_s(F_i(\varphi_r(Z_{t_{n-1}}),\tilde{U}_n)\lambda_{tot}(Z_{t_{n-1}},s;\delta)\Big] \leq L_1(T,z),$$

$$\sup_{n\leq N}\sup_{s\in[0,\delta_n]}\mathbb{E}_z\left[\overline{M}_2(Z_{t_{n-1}})\lambda_{tot}(Z_{t_{n-1}},s;\delta)\right] \leq L_2(T,z),$$

$$\sup_{n\leq N}\sup_{s\in[0,\delta_n]}\sup_{r\in[s,\delta_n]}\mathbb{E}_z\Big[\left(\lambda(\varphi_r(Z_{t_{n-1}})) + \overline{\lambda}(Z_{t_{n-1}},r;\delta)\right)\lambda_{tot}(Z_{t_{n-1}},s;\delta)\Big] \leq L_3(T,z).$$

For the case $p > 1$ we make the following assumption. Recall in the case $p > 1$ if in a single time step there have been q jumps then we use $\overline{\lambda}_i(\cdot;\delta,p-q)$ to simulate the next jump time. As the probability of there having been q jumps in a time interval is order δ^q the conditions required on $\overline{\lambda}_i(\cdot;\delta,q)$ are lessened, for this reason there are different requirements for each q.

Assumption 3.21. *Each of the approximate switching rates $\overline{\lambda}_i(\cdot;\delta,q)$ for $i = 1,\dots,m$ and $q = 1,\dots,p$ satisfies Assumption 3.12(a) for some $\overline{M}_2(z)$. When $q = 1$ the approximate switching rates $\overline{\lambda}_i(\cdot;\delta,1)$ for $i = 1,\dots,m$ satisfy Assumption 3.20. We make the additional moment bound for any $1 \leq q \leq p$*

$$\sup_{n\leq N}\sup_{s\in[0,\delta_n]}\mathbb{E}_z\left[(1+\overline{M}_2(Z_{t_{n-1}}))\lambda_{tot}(Z_{t_{n-1}},s;\delta,q) + \lambda_{tot}(Z_{t_{n-1}},s;\delta,q)^{q+1}\right] \leq L_4(T,z).$$

Remark 3.22. The moment bounds in Assumption 3.20 are rather technical, but also general. For instance Assumption 3.20 holds if Assumptions 3.12 and 3.13 hold, i.e. when the process has bounded norm for any finite time horizon. Furthermore, as Assumption 3.20 does not depend on moment bounds for the approximate process $\{\overline{Z}_{t_n}\}_{n\geq 1}$, one can verify Assumption 3.20 by finding a suitable Lyapunov function for the PDMP. Indeed if there exists a Lyapunov function which bounds the functions appearing in Assumption 3.20 then Assumption 3.20 holds with L_1, L_2, L_3 independent of T. This is the case for instance of the ZZS and BPS, see Example 3.39.

Alternatively, one can take advantage of Holder's inequality to reduce the problem to bounding polynomial moments of the various quantities. In Section 3.5.2 we show that the assumption holds for several examples. Finally we remark that in Assumption 3.20 it is possible to substitute $Z_{t_{n-1}}$ with $\overline{Z}_{t_{n-1}}$ and Theorem 3.23 still holds.

Theorem 3.23. *Denote as $\overline{\mathcal{P}}_t(z,\cdot)$ the transition probability of the approximation process obtained by Algorithm 8 for $p=1$ or by Algorithm 10 for $p>1$. Denote by $\{\mathcal{P}_t\}_{t\geq 0}$ the semigroup of a PDMP with generator (3.3) satisfying Assumption 3.7. Let $p\geq 1$ and suppose the approximations $\overline{\lambda}_i(z,s;\delta,q)$ for $q\leq p$ satisfy Assumption 3.20 if $p=1$ or Assumption 3.21 if $p>1$. Suppose the mesh $t_n=\sum_{i=1}^n \delta_n$ is such that $\delta_n<\delta_0$ for δ_0 as in Assumption 3.12(a). Suppose that $\overline{\varphi}_s=\varphi_s$ and $\overline{F}_i=F_i$ for all $i=1,\dots,m$. Then for any $z\in E$ and any mesh $0=t_0<t_1<\ldots<t_N=T$ with $\delta_n=t_n-t_{n-1}$ and $\delta_n\leq\delta_0$ for any $n\leq N$*

$$\|\mathcal{P}_T(z,\cdot)-\overline{\mathcal{P}}_T(z,\cdot)\|_{TV}\leq\sum_{i=1}^{N}\delta_i^{p+1}D(T,z)\prod_{\ell=i+1}^{N}(1-D(T,z)\delta_\ell),$$

where $D(t,z)$ is a non-decreasing function of t. If $\delta_n=\delta$ for all $n\in\mathbb{N}$ then

$$\|\mathcal{P}_T(z,\cdot)-\overline{\mathcal{P}}_T(z,\cdot)\|_{TV}\leq 1-e^{-D(T,z)T\delta^p}.$$

Proof. The proof can be found in Section 3.7. □

Remark 3.24. Let us for simplicity consider the constant step size case. If we fix a time horizon t, then the theorem shows that $\|\mathcal{P}_t(z,\cdot)-\overline{\mathcal{P}}_t(z,\cdot)\|_{TV}\to 0$ as $\delta\to 0$. On the other hand, the upper bound tends to 1 as $T\to\infty$ if the step size δ is fixed. Moreover, because $1-\exp(-D(t_n,z)t_n\delta)\leq D(t_n,z)t_n\delta$ we have

$$\|\mathcal{P}_{t_n}(z,\cdot)-\overline{\mathcal{P}}_{t_n}(z,\cdot)\|_{TV}\leq D(t_n,z)\,t_n\,\delta^p$$

and therefore we have convergence of order δ^p as $\delta\to 0$.

Remark 3.25. In a similar fashion to [65], it is possible to obtain a bound as that in Theorem 3.23 also when the jump kernel is approximated. To prove such result it is sufficient to define a coupled jump kernel that keeps the two processes together with strictly positive probability if they are together right before the jump.

3.4.3 Convergence to the invariant measure

In this section we give conditions for the approximation process $\{\overline{Z}_{t_n}\}_{n\geq 1}$ to converge to μ, the invariant measure of the PDMP, which we shall assume to exist and be unique. We do this by showing convergence in law to the PDMP uniformly in time and requiring that the PDMP converges to its invariant measure. In the following we consider the case of geometric convergence as it is verified for a range of PDMPs, however convergence with any rate $r(t)$ which is integrable over $[0,\infty)$ is sufficient.

The strategy of this proof is inspired by [46], which uses derivative estimates to obtain uniform in time convergence of an Euler Scheme for an SDE. In that case the authors rely on having exponential decay of the derivatives of the semigroup for the SDE of interest, for which conditions are given in [45].

Assumption 3.26. *Let $\{Z_t\}_{t\geq 0}$ be a PDMP with corresponding generator (3.5). Recall the definition of Q given by (3.6). We assume the following:*

(a) There exists an invariant measure, μ, for the PDMP, $\{Z_t\}_{t\geq 0}$, and μ is invariant under Q, that is

$$\mu(Qf) = \mu(f)$$

for any f measurable and integrable.

(b) The Markov process $\{Z_t\}_{t\geq 0}$ is geometrically ergodic with invariant measure μ. Specifically fix $\overline{G} : E \to [1, \infty)$ and define $\mathcal{G} = \{\text{measurable } g : E \to \mathbb{R}, |g| \leq \overline{G}\}$. Assume that $\overline{G}(Z_t)$ is integrable for all $t \geq 0$. For some $R_1 > 0$, $\omega > 0$

$$\sup_{g\in\mathcal{G}} |\mathbb{E}_z[g(Z_t)] - \mu(g)| \leq R_1 e^{-\omega t}\overline{G}(z). \tag{3.10}$$

This Assumption has been shown in a variety of cases, for example for the 1-dimensional Zig-Zag process this was shown in [21, Theorem 5] and for higher dimensions in [24, Theorem 2]. For BPS this was shown in [64, 51]. For RHMC see [30, Theorem 3.9].

The following assumption is required for Algorithm 7, but not for Algorithm 8, for the reasons explained in Note 3.31. In general, derivative estimates on the semigroup are useful for proving convergence of approximations as they control the effect of a small error in the initial condition of a stochastic process. In this case we are not using explicitly a derivative estimate but instead the operator $[\mathbf{\Phi}, Q]$. The role of this commutator is to describe the difference in the direction of the process over an infinitesimal time interval if the process first jumps then follows the flow map or first follows the flow map and then jumps.

Assumption 3.27. *Let $\{\mathcal{P}_t\}_{t\geq 0}$ denote the semigroup corresponding to the generator $\mathcal{L}$ given by (3.5). Recall the notation $[\mathbf{\Phi}, Q]$ defined in Section 3.2. Let $\overline{G}$ and $\mathcal{G}$ be given as in Assumption 3.26. There exist some $R_2 > 0$, $\omega > 0$ and set $\mathcal{G}_1 \subseteq \mathcal{G}$, such that for all $t \geq 0$ we have*

$$\sup_{g\in\mathcal{G}_1} \sup_{\delta\in(0,\delta_0), s\in[0,\delta]} [\mathbf{\Phi}, Q](\mathcal{P}_t g \circ \varphi_{\delta-s})(\varphi_s(z)) \leq R_2 e^{-\omega t}\overline{G}(z).$$

In Example 3.45 we show that this assumption is satisfied for ZZS with a non-trivial set $\mathcal{G}_1$.

Finally, we require the following moment bounds.

Assumption 3.28. *Let $\{\overline{Z}_t\}_{t\geq 0}$ be the process described by Algorithm i where i is either 7 or 8 and suppose Assumption 3.26 holds for the function $\overline{G}$. Recall the definition of λ and Q given by (3.6). Define for $i = 2$ or 3 (corresponding to Algorithm i)*

$$\begin{aligned}\overline{G}_i(z,r,s) = K_i(z,r,s) + \lambda(\varphi_r(z))Q((Q\overline{G}+\overline{G})\lambda)(\varphi_{s-r}(z)) \\ + \lambda(\varphi_r(z))\lambda(\varphi_s(z))(Q\overline{G}(\varphi_s(z)) + \overline{G}(\varphi_s(z)))\end{aligned} \tag{3.11}$$

where

$$\begin{aligned}K_7(z,r,s) &= \left(\overline{G}(z)\overline{\lambda}(z,s;\delta) + K_8(z,r,s)\right), \\ K_8(z,r,s) &= \left((Q\overline{G}(\varphi_s(z)) + \overline{G}(\varphi_s(z)))(\overline{\lambda}(z,s;\delta)\overline{\lambda}(z,r;\delta) + \overline{M}_2(z))\right).\end{aligned}$$

For $i = 2$ or 3 there exist a function $H_i(z)$ such that for any mesh $0 = t_0 \leq t_1 \leq ...$ with $\delta_k = t_k - t_{k-1} < \delta_0$ for any k

$$\mathbb{E}_z\left[\sup_{0\leq r<s\leq\delta_0} \overline{G}_i(\overline{Z}_{t_k}, r, s)\right] \leq CH_i(z).$$

Remark 3.29. Observe that since $\overline{G}$ is a Lyapunov function for the PDMP $\{Z_t\}$ we have that $\mathbb{E}_z[\overline{G}(Z_t)]$ is bounded in t for any z. Since $\{\overline{Z}_{t_n}\}_{n\geq 0}$ is designed to be a good approximation of $\{Z_t\}_{t\geq 0}$ we may hope that $\mathbb{E}_z[\overline{G}(\overline{Z}_{t_n})]$ is also bounded in n. We confirm this for ZZS and BPS in 1 dimension in Lemma 3.68 and test numerically in a higher dimensional setting.

In each of the references discussed after Assumption 3.26 there is some freedom in the choice of parameters in the Lyapunov function. By adjusting these parameters we can bound the terms in $\overline{G}_i(z)$ appearing in Assumption 3.28 by using a different choice of the parameters of the Lyapounov function. Confirming Assumption 3.28 then reduces to showing that, for a fixed Lyapunov function $\overline{G}$ for the PDMP, we have

$$\sup_n \mathbb{E}_z[\overline{G}(\overline{Z}_{t_n})] < \infty.$$

Theorem 3.30. *Let $\{Z_t\}_{t\geq 0}$ be the PDMP with generator given by (3.5). Let $\{\overline{Z}_t\}_{t\geq 0}$ be the process described by Algorithm 7 or 8, with $\overline{\varphi} = \varphi$ and $\overline{F} = F$. Suppose that Assumption 3.12 (a), 3.26, 3.28 holds and that if $\{\overline{Z}_t\}_{t\geq 0}$ is described by Algorithm 7 that Assumption 3.27 holds also. Let $\mathcal{G}_1 \subseteq \mathcal{G}$ be given as in Assumption 3.27 if this assumption is required and $\mathcal{G}_1 = \mathcal{G}$ otherwise.*

Then there exists $K > 0$ which depends only on R_1, R_2 and C such that for any $g \in \mathcal{G}_1$, $n \in \mathbb{N}, z \in E$ we have

$$\left|\mathbb{E}_z[g(Z_{t_n})] - \mathbb{E}_z[g(\overline{Z}_{t_n})]\right| \leq KS_nH_i(z). \tag{3.12}$$

Here

$$S_n = \sum_{k=0}^{n-1} \delta_{k+1}^2 e^{-\omega(t_n - t_{k+1})}. \tag{3.13}$$

Proof of Theorem 3.30. The proof of this theorem is given in Section 3.8. □

The choice of the set $\mathcal{G}_1$ here determines the type of convergence that we obtain. For example if $\mathcal{G}_1$ contains the set of continuous functions with supremum norm bounded by 1 then this corresponds to convergence in the total variation distance. On the other hand, if $\mathcal{G}$ contains the set of functions with Lipschitz constant less than 1 then we have convergence in the Wasserstein distance of order 1. Since we do not require Assumption 3.27 to hold when we use Algorithm 8 we can typically take $\mathcal{G}_1 = \mathcal{G}$ in that case and hence we have convergence in a metric that is stronger than total variation. However for Algorithm 7 we need an additional bound on the derivatives of the function so we have convergence in a weaker metric, see Example 3.45.

Remark 3.31. An important estimate in the proof of Theorem 3.30 will be obtaining a bound between the law of the first jump of the PDMP, τ, and of the approximation process, $\overline{\tau}$. This is done in Lemma 3.67. In this lemma we need to treat Algorithm 7 differently to Algorithm 8. In particular, we show convergence as $\delta \to 0$ by considering $\mathbb{E}[h(\tau)] - \mathbb{E}[h(\overline{\tau})]$ for a class $\mathcal{C}$ of test functions h. In the case of Algorithm 8 we use the set $\mathcal{C} = \mathcal{C}_b([0,\delta])$ of test functions whereas in the case of Algorithm 7 we use the set $\mathcal{C} = \mathcal{C}_b^1([0,\delta])$. The result of using this weaker convergence is that we need a form of derivative estimate. The derivative estimate we require is given by Assumption 3.27 and is needed only if we are considering Algorithm 7.

Remark 3.32. To simplify the exposition we have only considered the case when we can simulate the flow exactly. We can extend this proof to allow also for the use of a numerical integrator provided we have a suitable derivative bound. More precisely we require that for some $R_1 > 0$, $\omega > 0$ and some set $\mathcal{G}_1 \subseteq \mathcal{G}$ and for any $\delta \leq \delta_0, t > 0, z \in E, i \in \{1, \dots, m\}$

$$\sup_{g \in \mathcal{G}_1} |\mathcal{P}_t g(\overline{\varphi}_\delta(z)) - \mathcal{P}_t g(\varphi_\delta(z))| \leq \delta^2 R_3 e^{-\omega t} \overline{G}(z), \tag{3.14}$$
$$\sup_{g \in \mathcal{G}_1} |Q_i \mathcal{P}_t g(\overline{\varphi}_\delta(z)) - Q_i \mathcal{P}_t g(\varphi_\delta(z))| \leq \delta^2 R_3 e^{-\omega t} \overline{G}(z).$$

Now using the uniform in time weak error estimate (3.12) and exponential ergodicity (3.10) we can show convergence to the invariant measure of the PDMP.

Corollary 3.33. *Suppose that the conclusion of Theorem 3.30 holds. Set $\delta_k = \delta$ for all $k \in \mathbb{N}$. Then for $g \in \mathcal{G}_1$ we have*

$$\left|\mathbb{E}_z[g(Z_{t_n})] - \mathbb{E}_z[g(\overline{Z}_{t_n})]\right| \leq C\delta H(z), \tag{3.15}$$
$$\left|\mathbb{E}_z[g(\overline{Z}_{t_n})] - \mu(g)\right| \leq CH(z)(\delta + e^{-\omega t_n}) \tag{3.16}$$
$$\left|\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_z[g(\overline{Z}_{t_n})] - \mu(g)\right| \leq CH(z)\left(\delta + \frac{1}{t_N}\right). \tag{3.17}$$

Proof of Corollary 3.33. The proof of this corollary is given in Appendix 3.C.1. □

Corollary 3.34. *Suppose that the assumptions of Theorem 3.30 holds. Assume that* $\delta_k \to 0$ *as* $k \to \infty$ *and* $\sum_{k=1}^{\infty} \delta_k = \infty$. *For any* $g \in \mathcal{G}_1$ *we have*

$$\lim_{n\to\infty} \left|\mu(g) - \mathbb{E}_{x,v}[g(\overline{X}_{t_n}, \overline{V}_{t_n})]\right| = 0.$$

Proof of Corollary 3.34. The proof of this corollary is given in Appendix 3.C.1. □

3.5 Examples

3.5.1 Examples for Section 3.4.1

Example 3.35 (Zig-zag sampler continued)**.** *We continue Example 3.2 checking that the conditions of the previous section are satisfied. Let us check that approximations of the ZZS based on Algorithm 7 or 8 satisfy Corollary 3.17. Assumption 3.7 clearly holds. Assumption 3.13 holds because the process travels with bounded velocity, so we can apply the reasoning in Note 3.14 to verify Assumption 3.9. In particular, Assumption 3.8 holds as long as* $\psi \in \mathcal{C}^2$ *and* γ_i *is locally Lipschitz for all* $i \in \{1, \dots, m\}$. *Assumptions 3.10 and 3.11 follow from the fact that we can simulate exactly the flow and the kernels. Assumption 3.12(a) is satisfied for* $p = 1$ *both for* $\overline{\lambda}_i(z, s; \delta) = \lambda_i(z)$ *and* (3.8) *for* $\psi \in \mathcal{C}^2$ *. Assumption 3.12(b) follows from Assumption 3.13. Finally Assumption 3.16 clearly holds for any* $D > 0$.

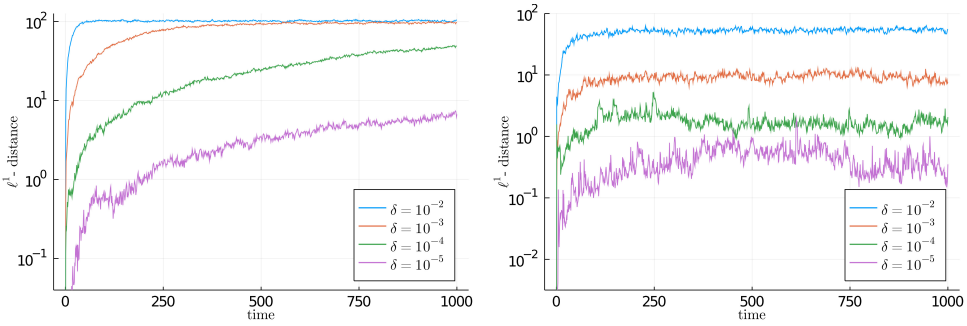
Note that we could define the same algorithm with $m = 1$ *according to Note 3.1. However, in this case neither Assumption 3.8 nor Assumption 3.16 hold as the function* F *is not Lipschitz.*

In Figure 3.1a we demonstrate numerically the difference between the ZZS with 50-dimensional Gaussian target and the approximation scheme corresponding to Algorithm 7 with constant step size δ *and frozen switching rates.* [1] *In this plot the two processes have been coupled according to Coupling 3.54, which is a synchronous coupling that is used in Section 3.6 to prove Theorem 3.15. In the figure we see that as* δ *tends to zero that the distance between the two processes converges to zero. We also observe there is an upper bound on how large the error can get, which roughly corresponds to the velocities having the opposite sign, i.e.* $\overline{V}_{t_n} = -V_{t_n}$.

Example 3.36 (Bouncy Particle Sampler continued)**.** *We continue Example 3.3 and discuss the assumptions of this section in this context. We show that* $x \mapsto R(x)v$ *need not be Lipschitz. Indeed, for a Gaussian example with* $d > 1$ *fix* $v \in \{1, -1\}^d$ *and take* $y \in \mathbb{R}^d$ *orthogonal to* v *then for any* $s > 0$ *consider*

$$\|R(sv)v - R(y)v\| = 2\left\|\frac{\langle v, sv\rangle}{\|sv\|^2} sv\right\| = 2\|v\|.$$

[1] The codes for all experiments in this paper can be found in a dedicated GitHub repository at `https://github.com/andreabertazzi/Euler_PDMC_algorithms`



(a) Results for the ZZS with $\gamma(x, v) = 0$.

(b) Results for the BPS with $\lambda_r = 1$.

Figure 3.1: Plots of the distance between the continuous time PDMPs and their approximations given by Algorithm 7 for several values of the step size. The x-axis shows continuous time units, i.e. the time of $\overline{Z}_{t_n}$ is $t_n = n\delta$. The distance is $\|x-y\|_1 = \sum_{i=1}^d |x_i - y_i|$ The plots show the average of 50 experiments. The processes are coupled according to Coupling 3.54. The continuous PDMPs have a 50-dimensional standard Gaussian as stationary measure. Here we choose $\overline{\lambda}((x, v), s; \delta) = \lambda(x, v)$.

Letting s tend to zero we see that $x \mapsto R(x)v$ is not Lipschitz at zero and hence Assumption 3.8 does not hold.

In Figure 3.1b we demonstrate numerically the difference between the BPS with 50-dimensional Gaussian target, Gaussian refreshments with rate 1 and the approximation scheme corresponding to Algorithm 7 with constant step size δ, and frozen switching rates when coupled according to Coupling 3.54. We see that although the assumptions of the theory do not hold the error appears to tend to zero as $\delta \to 0$. Indeed in Section 3.4.2 we obtain theory supporting this observation. Moreover, from the plot it appears that the error converges as $\delta \to 0$ uniformly in time. We will investigate this property further in Section 3.4.3.

Example 3.37 (Randomized Hamiltonian Monte Carlo algorithm continued)**.** *We continue Example 3.4. As long as $\nabla\psi$ is Lipschitz, Proposition 3.19 can be applied to the approximations based on Algorithms 7 and 8.*

Example 3.38 (PDMP two-dimensional Morris-Lecar model [94])**.** *Let us consider the PDMP defined on $E = \{0, \dots, N_K\} \times \mathbb{R}$ whose characteristics are given by*

$$\Phi(\theta, \nu) = \begin{pmatrix} 0 \\ \frac{1}{C}\left(1 - g_{\text{Leak}}(\nu - V_{\text{Leak}}) - g_{\text{Ca}} M_\infty(\nu)(\nu - V_{\text{Ca}}) - g_K \frac{\theta}{N_K}(\nu - V_K)\right) \end{pmatrix},$$

$$\lambda(\theta, \nu) = (N_K - \theta)\alpha_K(\nu) + \theta\beta_K(\nu),$$

$$Q((\theta, \nu), (\theta + 1)) = \frac{(N_K - \theta)\alpha_K(\nu)}{\lambda(\theta, \nu)}, \quad Q((\theta, \nu), \{\theta - 1\}) = \frac{\theta\beta_K(\nu)}{\lambda(\theta, \nu)},$$

$$M_\infty(\nu) = (1 + \tanh((\nu - V_1)/V_2))/2,$$
$$\alpha_K(\nu) = \lambda_K(\nu)N_\infty(\nu), \quad \beta_K(\nu) = \lambda_K(\nu)(1 - N_\infty(\nu)),$$
$$N_\infty(\nu) = (1 + \tanh((\nu - V_3)/4))/2, \quad \lambda_K(\nu) = \overline{\lambda}_K \cosh((\nu - V_3)/2V_4).$$

This model was given in [94] and is a PDMP version of the deterministic Morris-Lecar model introduced in [114] to explain the dynamics of the barnacle muscle fibre. Here ν denotes the membrane potential, θ is number of open Potassium channels, $g_{\text{Leak}}, g_{\text{Ca}}, g_K$ is maximum conductance value for leak, Calcium, and Potassium respectively, C is the membrane capacitance, $V_{\text{leak}}, V_{\text{Ca}}, V_K$ is the equilibrium potential of relevant ion channels, $M_\infty(\nu)$ ($N_\infty(\nu)$ respectively) is the fraction of open Calcium (Potassium resp.) channels at steady state. V_1 (V_3 respectively) is the potential at which $M_\infty = 0.5$ ($N_\infty = 0.5$ resp.). V_2 (respectively V_4) is the reciprocal is the slope of the voltage dependence of M_∞ (N_∞ resp.).

We will consider a PD-PDMP approximation of this PDMP. Note that in this case the flow does not have an explicit solution so a numerical integrator is required. Therefore we will set $\overline{\varphi}_t$ to be an Euler approximation of φ_t and $\overline{\lambda}((\theta,\nu),t;\delta) = \lambda(\overline{\varphi}_t(\theta,\nu;\delta)) = \lambda((\theta,\nu) + t\Phi(\theta,\nu))$. Note we can simulate jump times with this approximate rate using Poisson thinning. Since the kernel Q can be simulated exactly we do not need to approximate this. This algorithm is very similar to the approximation proposed in [94] and we confirm their results in our framework. Indeed, one can verify that Assumptions 3.7, 3.9-3.12, as well as Assumption 3.16 hold so by Theorem 3.15 we have that the approximation converges as $\delta \to 0$ to the PDMP.

3.5.2 Examples for Section 3.4.2

It is straightforward to verify Assumption 3.20 for either ZZS or BPS. Below we give details for BPS.

Example 3.39 (Bouncy Particle Sampler continued)**.** *We continue Examples 3.3 and 3.36 and discuss when we may apply Theorem 3.23 in this setting. Recall in Example 3.36 we showed that we can not expect BPS to satisfy the assumptions of Theorem 3.15 because the reflection operator is in general not Lipschitz. However, we do not need any assumption of this type for Theorem 3.23. Consider for instance a simple example in which $\overline{\lambda}_1((x,v),s) = \lambda_1(x,v)$. Then Assumption 3.12 (a) follows provided $\psi \in \mathcal{C}^2$. If in particular ψ has bounded Hessian, then $\overline{M}_2(x,v) \leq \|v\|\|\nabla^2 U\|_\infty$. It remains to verify the moment bounds in Assumption 3.20 hold. It is clear that these are satisfied if the velocities are bounded, as for instance when refreshments are from $\mathscr{Z}_{n+1} \sim \text{Unif}(\mathbb{S}^{d-1})$ where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. On the other hand, for the BPS with Gaussian refreshments we observe that outside of a compact set one can bound the moments in Assumption 3.20 by the expectation of the Lyapunov functions derived in [51] or [64]. Therefore we can apply Theorem 3.23 to obtain convergence as $\delta \to 0$.*

Let us derive a rough estimate on the dimensional dependence of Theorem 3.23 in the $p = 1$ case. In particular we focus on the dependence of $D(t_n, z)$ in the dimension. Observe from the proof of the theorem that $D(t, z)$ depends linearly on L_1, L_2, L_3, and thus it is sufficient to check the dimensional dependence of such constants. By applying Cauchy-Schwartz inequality the interesting terms are of the form $\mathbb{E}_z[(\lambda_1(X_t, V_t))^2]$. We approximate this expectation with its value in stationarity. Let us restrict to the case of Gaussian refreshments and a standard Gaussian invariant measure. Then in stationarity we obtain

$$\mathbb{E}_\pi[(\lambda_1(X, V))^2] = \mathbb{E}_\pi[\langle V, X\rangle_+^2] \leq d^2.$$

Therefore we expect $D(t, z)$ to have a quadratic dependence in the dimension of the PDMP. In order to obtain a fixed error in total variation distance one should then choose δ such that $D(t, z)\delta$ is constant, and thus δ of order d^{-2}. Observe that taking refreshments on the unit sphere the dependence would be linear in d.

Example 3.40 (Continuous time approximations of PDMP)**.** *So far we have concentrated on discrete time approximations of PDMP, however it is also possible to apply our results to approximate a PDMP with a second continuous time PDMP. Similarly to the setting of [84] suppose we have an approximation $\tilde{\lambda}_i$ of λ_i, i.e. there exists $\varepsilon > 0$ such that for all $i \in \{1, \dots, m\}$ we have*

$$|\tilde{\lambda}_i(z) - \lambda_i(z)| \leq \tilde{M}(z)\varepsilon. \tag{3.18}$$

A possible motivation for this approach is the case of ZZS when we either can not evaluate $\partial_i\psi$ exactly or it is too expensive to do so. Then we can use an approximation $\overline{\partial_i\psi}$ to obtain an approximation of λ, i.e. $\tilde{\lambda}_i = (v_i\overline{\partial_i\psi})_+$. Now we define a PDMP with approximated rates $\tilde{\lambda}_i$ which moves according to the generator $\tilde{\mathcal{L}}$ acting on sufficiently smooth functions by

$$\tilde{\mathcal{L}}f(z) = \langle \phi(z), \nabla_z f(z)\rangle + \sum_{i=1}^m \tilde{\lambda}_i(z) \int (f(z') - f(z))Q_i(z, \mathrm{d}z'). \tag{3.19}$$

We are interested in comparing this process to the PDMP with generator $\mathcal{L}$ given by (3.3). In order to use Theorem 3.23 we introduce a discrete time process which we can use to compare to both the PDMPs corresponding to $\mathcal{L}$ and $\tilde{\mathcal{L}}$. Set $\delta = \varepsilon$ and $t_n = n\delta$, define $\{\overline{Z}_{t_n}\}_{n\in\mathbb{N}}$ to be given by Algorithm 8 with rates $\overline{\lambda}_i(z, t) = \tilde{\lambda}_i(\varphi_t(z))$, and according to the exact flow $\overline{\varphi}_t = \varphi_t$ and Markov kernels Q_i, i.e. $\overline{F}_i = F_i$. We assume that the moment bounds of Assumption 3.20 are satisfied which is clear for example if the processes move with bounded velocity. We may apply Theorem 3.23 both when the PDMP is given by $\mathcal{L}$ and by $\tilde{\mathcal{L}}$. Therefore for any $t > 0$ there exist a constant $D(t, z) > 0$ such that

$$\|\mathcal{P}_t(z, \cdot) - \tilde{\mathcal{P}}_t(z, \cdot)\|_{TV} \leq 2(1 - e^{-D(t,z)\,t\,\varepsilon}) \leq 2D(t, z)\,t\,\varepsilon.$$

Here $\{\mathcal{P}_t\}_{t\geq 0}$ denotes the semigroup of a PDMP with generator $\mathcal{L}$ and $\{\tilde{\mathcal{P}}_t\}_{t\geq 0}$ denotes the semigroup of a PDMP with generator $\tilde{\mathcal{L}}$. We remark that the analysis above could be adapted to the setting of the Numerical Zig-zag algorithm introduced in [121].

Example 3.41 (ZZS with subsampling)**.** *In Bayesian statistics the posterior distribution $\pi(x) \propto \exp(-\psi(x))$ is often of the form $\psi(x) = \sum_{j=1}^{N} \psi_j(x)$, where $\psi_j(x)$ depends only on a subset of the data. As described in [23], the ZZS allows for exact subsampling, which means that the simulation of each event time is calculated using ψ_J for some $J \sim \mathrm{Unif}\{1, \dots, N\}$ instead of the full negative log-density ψ. This is achieved by defining switching rates $\lambda_i^j(x, v) = (v_i \partial_i \psi_j(x))_+$ and computational bounds $M_i(t)$ such that $\lambda_i^j(x + vt, v) \leq M_i(t)$. Then starting at state (x, v) at time t, a proposal for the next event time is found by taking $\tau_{i^*} = \min \tau_i$, where τ_i has rate $M_i(t)$ for $i = 1, \dots, d$. Then the proposal is accepted with probability $\lambda_{i^*}^J(x + v\tau_{i^*}, v)/M_{i^*}(\tau_{i^*})$ for $J \sim \mathrm{Unif}\{1, \dots, N\}$ and in case of acceptance we set $V_{t+\tau_{i^*}} = R_{i^*} V_t$.*

Motivated by the fact that the bounds $M_i(t)$ may be unavailable or hard to compute, we can approximate this process as follows. Here we restrict to the case of frozen switching rates, that is

$$\overline{\lambda}_i^j((x, v), s; \delta) = \lambda_i^j(x, v).$$

We apply the same idea behind Algorithm 8 to obtain $(\overline{X}_{t_{n+1}}, \overline{V}_{t_{n+1}})$ given the previous state by first drawing $J \sim \mathrm{Unif}\{1, \dots, N\}$, and then simulating the next switching time $\overline{\tau} = \overline{\tau}_{i^*} = \min \overline{\tau}_i$ with rates $\lambda_i^J(\overline{X}_{t_n}, \overline{V}_{t_n})$. Finally

$$\begin{aligned} \overline{X}_{t_{n+1}} &\coloneqq \overline{X}_{t_n} + (\delta_{n+1} - \overline{\tau})\overline{V}_{t_n} + \overline{\tau}\overline{V}_{t_{n+1}} \\ \overline{V}_{t_{n+1}} &\coloneqq \begin{cases} R_{i^*}\overline{V}_{t_n} & \text{if } \overline{\tau} \leq \delta_{n+1}, \\ \overline{V}_{t_n} & \text{if } \overline{\tau} > \delta_{n+1}, \end{cases} \end{aligned}$$

where the operator R_i was defined in Example 3.2.

The fundamental difference with respect to the setting of Algorithm 8 lies in the additional level of randomness introduced by the random variable J. We can adapt the proof of Theorem 3.23 to also allow this additional randomness. Indeed in each step we use a synchronous coupling of the random variable J, then conditional on J we can apply Coupling 3.57 with λ_i replaced by λ_i^J in (3.34)-(3.35)*, and setting*

$$\lambda_{tot}^i((x, v), t; \delta) = \sum_{j=1}^{N} \lambda_i^j(x + vt, v) + \sum_{j=1}^{N} \lambda_i^j(x, v) + 1.$$

Thus provided $\psi \in \mathcal{C}^2$ for any $(x, v) \in E$ and $t > 0$ there exists $D = D(t, (x, v)) > 0$ such that

$$\|\mathcal{P}_t((x, v), \cdot) - \overline{\mathcal{P}}_t((x, v), \cdot)\|_{TV} \leq 1 - e^{-Dt\delta},$$

where $\{\mathcal{P}_t\}_{t \geq 0}$ is the semigroup of ZZS with subsampling and $\{\overline{\mathcal{P}}_t\}_{t \geq 0}$ is the transition probability of the approximation. We remark that a similar reasoning can be applied to the BPS with subsampling (see [32]).

3.5.3 Examples for Section 3.4.3

Example 3.42 (Randomized Hamiltonian Monte Carlo algorithm continued)**.** *We continue Example 3.4 by verifying the various assumptions for Theorem 3.30. For this example we assume that $\psi \in \mathcal{C}^2$, is strongly convex and has bounded Hessian, i.e. for some $K, L > 0$*

$$KI_d \preceq \nabla^2 \psi(q) \preceq LI_d. \tag{3.20}$$

When this holds we have that Assumption 3.26 is satisfied by [30, Theorem 3.9] with $\overline{G} = H$, where H is the Hamiltonian function

$$H(q,p) = \psi(q) + \frac{1}{2}\|p\|^2.$$

Since we consider the approximation based on Algorithm 8 we do not need Assumption 3.27. As λ is constant in this case Assumption 3.28 is satisfied provided

$$\sup_{n\in\mathbb{N}} \mathbb{E}_{(q,p)}\left[\psi(\overline{Q}_{t_n}) + \frac{1}{2}\|\overline{P}_{t_n}\|^2\right] < \infty. \tag{3.21}$$

Because ψ has bounded second order derivative this reduces to showing that the second moment of the approximation is bounded uniformly in time. This condition depends on the choice of the numerical integrator and should be checked depending on the specific choice.

As mentioned in Note 3.32 in order to apply Theorem 3.30 with a numerical error we need to verify (3.14) *holds. It is sufficient to show that the derivative of the semigroup decays exponentially. In order to prove this we shall rely on two Lipschitz conditions for the Hamiltonian flow φ_t: there exist $\nu, K_1, C \geq 1, \gamma \in (0,1)$ such that for any $q, \overline{q}, p, \overline{p} \in \mathbb{R}^d$*

$$\sup_{t>0} \|\varphi_t(q,p) - \varphi_t(\overline{q},\overline{p})\| \leq C\|(q,p) - (\overline{q},\overline{p})\| \tag{3.22}$$

$$\|\varphi_t(q,p) - \varphi_t(\overline{q},p)\| \leq \gamma\|q - \overline{q}\| \quad \text{for } \nu < t \leq K_1. \tag{3.23}$$

It is shown in [31] that under (3.20) *the contraction* (3.23) *holds for some ν, γ, K_1. Indeed the authors prove a stronger result under which $\nu = 0$, but γ depends on t. There are also extensions to non-convex functions ψ, however here we only consider the convex setting. On the other hand,* (3.22) *is for instance satisfied for linear flows since the flow preserves the Hamiltonian and the Hamiltonian is equivalent to the norm. To simplify the exposition we will restrict to the case where* (3.22) *and* (3.23) *hold.*

Proposition 3.43. *Let $\{\mathcal{P}_t\}_{t\geq 0}$ denote the semigroup of RHMC. Suppose that* (3.22) *and* (3.23) *hold. Moreover assume that*

$$0 < \kappa := C(1 - (e^{-\lambda\nu} - e^{-\lambda K_1})(1 - \gamma C^{-1})) < 1.$$

Then

$$\|\nabla_{q,p}\mathcal{P}_t f(q,p)\| \leq C^2 e^{-\kappa t}\|f\|_{C_b^1}.$$

Proof. The proof is deferred to Appendix 3.C.2. □

A case where it is easy to see that $\kappa < 1$ is the standard Gaussian case, i.e. $\psi(q) = \|q\|^2/2$, since in this case we have that $C = 1$.

Theorem 3.44. *Suppose that ψ satisfies* (3.20), (3.22), (3.23), *and the numerical integrator satisfies* (3.21). *Then the conclusions of Theorem 3.30 hold.*

Example 3.45 (Zig Zag Sampler continued)**.** *Recall the notation of Example 3.2 and 3.35. Let us verify the assumptions of Theorem 3.30 for the ZZS.*

We will make the following assumptions on ψ. Assume $\psi \in \mathcal{C}^2$ and

$$\lim_{\|x\|\to\infty} \frac{\max(1, \|\nabla_x^2 \psi(x)\|)}{\|\nabla_x \psi(x)\|} = 0 \quad \text{and} \quad \lim_{\|x\|\to\infty} \frac{\|\nabla_x \psi(x)\|}{\psi(x)} = 0. \tag{3.24}$$

Verifying Assumption 3.26:

Geometric ergodicity of the ZZS was established in [24] under (3.24) with Lyapunov function

$$\overline{G}_{\alpha,\epsilon}(x, v) = \exp\left(\alpha\psi(x) + \sum_{i=1}^{d} \phi_\epsilon(v_i \partial_i \psi(x))\right). \tag{3.25}$$

Here $\phi_\epsilon(s) = \text{sign}(s)\log(1 + \epsilon|s|)/2$, $\alpha \in (0, 1)$, $\epsilon > 0$, $\alpha > \epsilon\overline{\gamma}$, where $\overline{\gamma}$ upper bounds the excess switching rate $\gamma : E \to \mathbb{R}_+$.

Verifying Assumption 3.27:

When dealing with this assumption it is convenient to use a smooth choice of λ_i, so for this section we will set $\phi(r) = r(1 + r)^{-1}$ and

$$\lambda_i(x, v) = -\log\left(\phi(\exp(-v_i \partial_i \psi(x)))\right) \tag{3.26}$$

which was shown in [2] to be a smooth choice of λ_i for which the ZZS has the correct invariant measure. Note that for this choice of λ_i the excess switching rate γ takes values between 0 and $\overline{\gamma} = \log(2)$.

Lemma 3.46. *Let $\{\mathcal{P}_t\}_{t\geq 0}$ denote the semigroup corresponding to the ZZS as described in Example 3.2. Assume $\psi \in \mathcal{C}^2$ and has bounded Hessian. For λ_i given by* (3.26) *there exist a constant C depending on the Hessian of ψ and on d such that for any $f \in \mathcal{C}^1$ we have*

$$\begin{aligned}[\mathbf{\Phi}, Q](f \circ \varphi_{\delta - s})(x + sv, v) \leq C \sup_{i\in\{1,\dots,d\}} \{&|f(x + sv + (\delta - s)F_i v, F_i v)| \\ &+ |\partial_{x_i} f(x + sv + (\delta - s)F_i v, F_i v)|\}.\end{aligned}$$

Proof of Lemma 3.46. The proof is deferred to Appendix 3.C.3. □

We apply this Lemma with $f = \mathcal{P}_t g$ with $g \in \mathcal{G}_1$ where

$$\mathcal{G}_1 = \{g : E \to \mathbb{R} : x \mapsto g(x,v) \in \mathcal{C}^1, \mu(g) = 0, |g| \leq \overline{G}_{\overline{\alpha},\epsilon}, \|\nabla_x g\| \leq \overline{G}_{\overline{\alpha},\epsilon}\} \tag{3.27}$$

where $\overline{\gamma}\epsilon < \overline{\alpha} < \alpha < 1$. Such an $\overline{\alpha}$ can always be found by taking ϵ sufficiently small. It remains to show that $\nabla_x \mathcal{P}_t g$ converges to zero for $g \in \mathcal{G}_1$.

Theorem 3.47. *Let $\{\mathcal{P}_t\}_{t\geq 0}$ denote the semigroup of the ZZS with generator given by* (2.32) *and with λ_i such that $x \mapsto \lambda_i(x,v) \in \mathcal{C}^1$ for each v and has bounded derivative $\nabla_x \lambda_i(x,v)$. Fix $\epsilon\overline{\gamma} < \overline{\alpha} < \alpha$, and let $\mathcal{G}_1$ be given by* (3.27)*. Then there exists a constant C such that for any $g \in \mathcal{G}_1$*

$$\|\nabla_x \mathcal{P}_t g(x,v)\| \leq C(1+t)e^{-\omega t}\overline{G}_{\alpha,\epsilon}(x,v).$$

Proof of Theorem 3.47. The proof is deferred to Appendix 3.C.3. □

Note that by adjusting C, κ and α we can show that

$$\sup_{\delta\in(0,\delta),s\in[0,\delta]} \|\nabla_x \mathcal{P}_t g \circ \varphi_{\delta-s}(x+sv,v)\| \leq Ce^{-\omega t}\overline{G}_{\alpha,\epsilon}(x,v).$$

Therefore Assumption 3.27 holds by combining Lemma 3.46, Theorem 3.47 and Assumption 3.26.

Verifying Assumption 3.28: Note that since λ grows at most linearly it is sufficient to show that there exist a function $H(x,v)$ for any mesh $0 = t_0 \leq t_1 \leq ...$ with $\delta_k = t_k - t_{k-1} < \delta_0$ for any k

$$\mathbb{E}_z\left[\|\overline{X}_{t_k}\|^2 \overline{G}_{\alpha,\epsilon}(\overline{X}_{t_k}, \overline{V}_{t_k})\right] \leq H(x,v).$$

Note that $\overline{G}_{\alpha,\epsilon}$ is dominated by the $e^{\alpha\psi}$ term so we can bound $\|x\|^2\overline{G}_{\alpha,\epsilon}(x,v)$ by $e^{\alpha_1\psi}$ for any $\alpha_1 > \alpha$ so it remains to show there exist a function $H(x,v)$ for any mesh $0 = t_0 \leq t_1 \leq ...$ with $\delta_k = t_k - t_{k-1} < \delta_0$ for any k

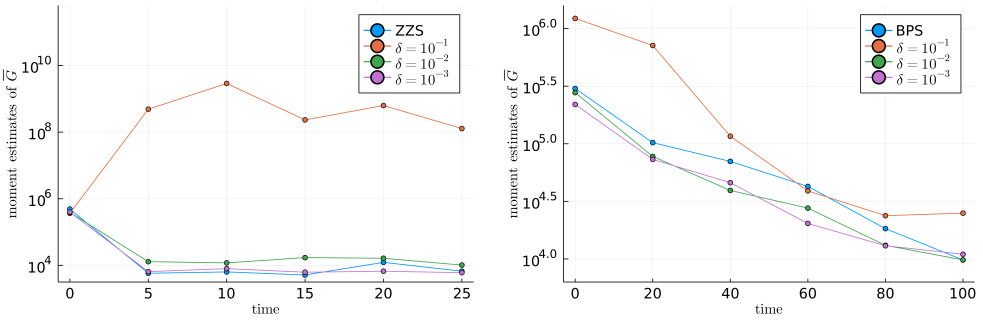
$$\mathbb{E}_z\left[e^{\alpha_1\psi(\overline{X}_{t_k})}\right] \leq H(x,v). \tag{3.28}$$

Then in the 1-dimensional case we prove that the required bound holds.

Lemma 3.48. *Suppose* (3.24) *hold and $d = 1$. Then* (3.28) *holds for both Algorithms 7 and 8, hence Assumption 3.28 is satisfied.*

This follows from Lemmas 3.68 and 3.69 which can be found in Appendix 3.C.3. The generalisation to the d-dimensional setting is challenging and is thus left as a conjecture, supported by the experiments in Figure 3.2a.

Conjecture 3.49. *Suppose ψ satisfies* (3.24)*. Then inequality* (3.28) *holds for Algorithms 7 and 8.*



(a) Results for the ZZS and its approximations given by Algorithm 7. Here $\gamma(x, v) = 0$.

(b) Results for the BPS and its approximations given by Algorithm 8. Here $\lambda_r = 1$.

Figure 3.2: Plots of the estimates of $\mathbb{E}[\overline{G}(X_t, V_t)]$ and $\mathbb{E}[\overline{G}(\overline{X}_t, \overline{V}_t)]$, which are respectively for the continuous time PDMPs and their approximations for several values of the step size. The plots show the average of 10^5 experiments. The continuous PDMPs have a 25-dimensional standard Gaussian as stationary measure. Here we choose $\overline{\lambda}((x, v), s; \delta) = \lambda(x, v)$. For each experiment $X_0, \overline{X}_0$ are given by an independent realisation of the sum of a 25-dimensional standard Gaussian and a uniform random variable on $[0, 1]^{25}$, while $V_0, \overline{V}_0$ are drawn from the stationary distribution of the continuous time PDMP.

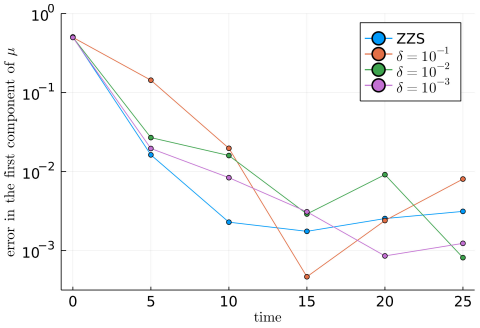
Theorem 3.50. *Let $\{(X_t, V_t)\}_{t\geq 0}$ be the ZZS. Let $\{(\overline{X}_t, \overline{V}_t)\}_{t\geq 0}$ be the process described in Example 3.2. Assume that ψ satisfies* (3.24) *and that Conjecture 3.49 holds. Let $\mathcal{G}_1$ be given by* (3.27)*. Then the conclusions of Theorem 3.30 hold.*

In Figure 3.3 we show some numerical results in the case of a Gaussian target. We observe that the error in the estimation of the first component of the mean of the approximations is similar to that of the continuous ZZS, while the error for the radius statistic, i.e. $t(x) = \sum_{i=1}^{d} x_i^2$, obtained with the approximations decreases to that of the ZZS as δ becomes smaller.

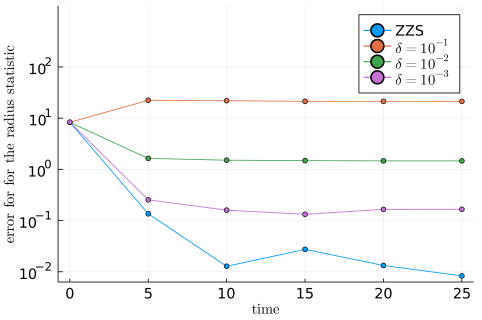
Example 3.51 (BPS continued)**.** *Let us discuss the assumptions of Theorem 3.30 for the approximation of BPS as given in Example 3.3, 3.36, and 3.39. Since this approximation is based on Algorithm 8 it is sufficient to check Assumptions 3.26 and 3.28. Conditions under which Assumption 3.26 holds are given in [51] and [64]. To be concrete we concentrate on [51], in which the Lyapunov function is given by*

$$\overline{G}(x, v) = \frac{e^{\frac{1}{2}\psi(x)}}{\sqrt{\lambda(x, -v)}}.$$

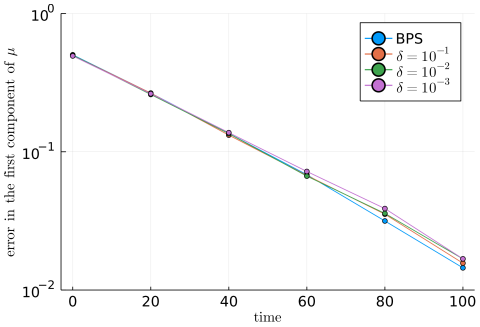
Here at refreshment times a new velocity vector is drawn from the uniform distribution on the unit sphere. In Figure 3.2b we estimate the moments of $\overline{G}$ for the continuous time BPS with a 25-dimensional standard Gaussian target and compare it



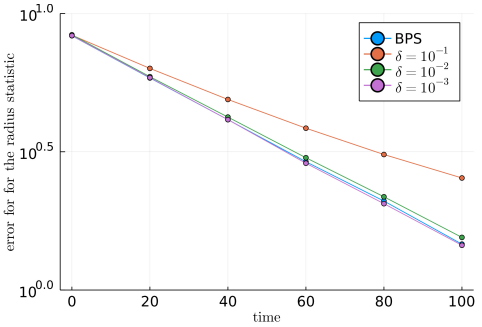
(a) Error for the mean for the ZZS and its approximations given by Algorithm 7.



(b) Error for the radius for the ZZS and its approximations given by Algorithm 7.



(c) Error for the mean for the BPS and its approximations given by Algorithm 8.



(d) Error for the radius for the BPS and its approximations given by Algorithm 8.

Figure 3.3: Errors in the estimation of the first component of the mean and radius statistic in the context of Figure 3.2. For the ZZS we take $\gamma(x, v) = 0$, while for the BPS we have $\lambda_r = 1$.

to the approximations obtained by applying Algorithm 8 for several values of δ. We observe that the moments of the approximations resemble the continuous BPS and $\mathbb{E}[\overline{G}(\overline{X}_t, \overline{V}_t)]$ appears to be bounded uniformly in time. Therefore we conjecture that Theorem 3.30 holds under the assumptions of [51] for approximations of the BPS according to Algorithm 8.

In Figure 3.3 we compare the errors of the BPS and its approximations given by Algorithm 8 in the case of a Gaussian target. We observe that the approximations perform similarly to the BPS. Note that for this target measure the BPS and the approximation are both rotationally invariant so they both have mean zero and hence in Figure 3.3 (c) we do not see the effect of the bias of the approximation.

Example 3.52 (Continuous time approximation of a PDMP)**.** *We continue our analysis of the setting introduced in Example 3.40. We wish to extend the conclusions of*

Theorem 3.30 to the continuous PDMP with generator $\tilde{\mathcal{L}}$ given by (3.19).

Theorem 3.53. *Suppose both the PDMPs with generators $\mathcal{L}$ and $\tilde{\mathcal{L}}$ satisfy Assumption 3.26 with Lyapunov functions $\overline{G}$ and $\tilde{G}$ respectively, and with invariant measures μ and $\tilde{\mu}$. Assume* (3.18) *holds for some $\varepsilon > 0$. Moreover, suppose the approximation of the PDMP with generator $\mathcal{L}$ described in Example 3.40 satisfies Assumption 3.28 both for $\overline{G}$ and $\tilde{G}$. Set $\mathcal{G}_1 = \{g \in \mathcal{C}(E) : |g(x,v)| \leq \min\{\overline{G}(x,v), \tilde{G}(x,v)\}\}$. Then for all $g \in \mathcal{G}_1$*

$$\left|\mathbb{E}_z[g(Z_t)] - \mathbb{E}_z[g(\tilde{Z}_t)]\right| \leq C\varepsilon H(z).$$

Moreover, letting $t \to \infty$ we have

$$|\mu(g) - \tilde{\mu}(g)| \leq D\varepsilon.$$

Hence, in the case of ZZS we recover the result obtained in Theorem 6.2 of [84].

3.6 Proof of Theorem 3.15

We shall first prove the case of $p = 1$ in Section 3.6.1, and then in Section 3.6.2 we will use the $p = 1$ setting as a base case in a proof by induction to obtain the result for $p > 1$.

3.6.1 The case of $p = 1$

To prove Theorem 3.15 in this setting we define a coupling of Z_{t_n} and $\overline{Z}_{t_n}$ that satisfies the bounds in the statement. Then because the Wasserstein distance is defined as an infimum over all couplings we immediately obtain

$$\mathcal{W}_1(\mathcal{P}_T(z,\cdot), \overline{\mathcal{P}}_T(z,\cdot)) \leq \mathbb{E}_z[\|Z_T - \overline{Z}_T\|],$$

where the expectation in the right hand side is with respect to the specific coupling we consider.

Let us now introduce a general framework that contains both Algorithm 7 and Algorithm 8. Denote the approximation process as $\overline{Z}_{t_n}$ with initial state $\overline{Z}_0 = z$. Then given the previous state $\overline{Z}_{t_n}$ define

$$\begin{aligned} \tilde{\tau}^i_{n+1} &\coloneqq \inf\left\{t \geq 0 : 1 - \exp\left(-\int_0^t \overline{\lambda}_i(\overline{Z}_{t_n}, s; \delta_{n+1})\mathrm{d}s\right) \geq \tilde{U}^i_{n+1}\right\}, \\ \tilde{\tau}_{n+1} &\coloneqq \min_{i=1,\ldots,m} \tilde{\tau}^i_{n+1}, \quad I_{n+1} = \operatorname*{arg\,min}_{i=1,\ldots,m} \tilde{\tau}^i_{n+1} \end{aligned} \tag{3.29}$$

where $\tilde{U}^1_{n+1}, \ldots, \tilde{U}^m_{n+1} \overset{iid}{\sim} \mathrm{Unif}[0,1]$. Then the switching time of the process is

$$\overline{\tau}_{n+1} = \overline{\tau}_{n+1}(\tilde{\tau}_{n+1}, \delta_{n+1}),$$

with the requirement that $\overline{\tau}_{n+1} \leq \delta_{n+1}$ if and only if $\tilde{\tau}_{n+1} \leq \delta_{n+1}$. In particular Algorithm 7 corresponds to the choice

$$\overline{\tau}_{n+1}(\tilde{\tau}_{n+1}, \delta_{n+1}) = \delta_{n+1} \mathbb{1}_{\{\tilde{\tau}_{n+1} \leq \delta_{n+1}\}} + \infty \mathbb{1}_{\{\tilde{\tau}_{n+1} > \delta_{n+1}\}},$$

while Algorithm 8 corresponds to

$$\overline{\tau}_{n+1}(\tilde{\tau}_{n+1}, \delta_{n+1}) = \tilde{\tau}_{n+1}.$$

The process can now be defined as follows:

$$\overline{Z}_{t_{n+1}} = \begin{cases} \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n}; \delta_{n+1}) & \text{if } \overline{\tau}_{n+1} > \delta_{n+1}, \\ \overline{\varphi}_{\delta_{n+1}-\overline{\tau}_{n+1}}(\overline{F}_{I_{n+1}}(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}; \delta_{n+1}), U_{n+1}; \delta_{n+1}); \delta_{n+1}) & \text{if } \overline{\tau}_{n+1} \leq \delta_{n+1}, \end{cases}$$

where $U_{n+1} \sim \nu_{\mathcal{U}}$ takes values in $\mathcal{U}$.

Let us now define the coupling of Z_{t_n} and $\overline{Z}_{t_n}$ that we will use to prove Theorem 3.15.

Coupling 3.54. *Fix both processes up to time t_n and let $(Z_{t_{n+1}}, \overline{Z}_{t_{n+1}})$ evolve as follows. Let $U_{n+1} \sim \nu_{\mathcal{U}}$ and $\tilde{U}^1_{n+1}, \dots, \tilde{U}^m_{n+1} \overset{iid}{\sim} \mathrm{Unif}([0,1])$ be independent of each other and of Z_{t_n} and $\overline{Z}_{t_n}$. The coupling evolves as follows:*

- *Define the next switching time of the continuous process as*

$$\begin{aligned} \tau^i_{n+1} &:= \inf\left\{ t \geq 0 : 1 - \exp\left(- \int_0^t \lambda_i(\varphi_s(Z_{t_n}))\mathrm{d}s \right) \geq \tilde{U}^i_{n+1} \right\}, \\ \tau_{n+1} &:= \min_{i=1,\dots,m} \tau^i_{n+1}. \end{aligned} \tag{3.30}$$

  Then there are two cases:

  - *if $\tau_{n+1} \leq \delta_{n+1}$, then set $Z_{t_n+s} = \varphi_s(Z_{t_n})$ for $s \in (0, \tau_{n+1})$ and*

$$Z_{t_n+\tau_{n+1}} = F_{I_{n+1}}(\varphi_{\tau_{n+1}}(Z_{t_n}), U_{n+1})$$

    where $I_{n+1} = \arg\min_i \tau^i_{n+1}$. Then simulate the process independently of the rest for the remaining time $\delta_{n+1} - \tau_{n+1}$.

  - *if $\tau_{n+1} > \delta_{n+1}$, then set $Z_{t_n+s} = \varphi_s(Z_{t_n})$ for $s \in (0, \delta_{n+1}]$.*

- *Define $\tilde{\tau}_{n+1}$ as in Equation (3.29), where $\tilde{U}^1_{n+1}, \dots, \tilde{U}^m_{n+1}$ is the same random variable used in (3.30). Compute $\overline{\tau}_{n+1} = \overline{\tau}_{n+1}(\tilde{\tau}_{n+1}, \delta_{n+1})$. Then the approximation process evolves as:*

  - *if $\overline{\tau}_{n+1} \leq \delta_{n+1}$, then set*

$$\overline{Z}_{t_{n+1}} = \overline{\varphi}_{\delta_{n+1}-\overline{\tau}_{n+1}}(\overline{F}_{\overline{I}_{n+1}}(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}; \delta_{n+1}), U_{n+1}; \delta_{n+1}); \delta_{n+1})$$

    where $\overline{I}_{n+1} = \arg\min_i \tilde{\tau}^i_{n+1}$.

– *if* $\overline{\tau}_{n+1} > \delta_{n+1}$, *then set* $\overline{Z}_{t_{n+1}} = \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n}; \delta_{n+1})$.

Therefore the first switching times of the two processes are coupled, and so is the eventual random jump. Once $Z_{t_{n+1}}$ *and* $\overline{Z}_{t_{n+1}}$ *have been obtained, repeat the same procedure to obtain* $Z_{t_{n+2}}$ *and* $\overline{Z}_{t_{n+2}}$.

We remark that the marginal distributions of each process is the correct one, and thus this is indeed a valid coupling of the two processes.

In the proof that follows we simplify the notation denoting the approximations as $\overline{\varphi}_z(z)$, $\overline{\lambda}_i(z,s)$, and $\overline{F}_i(z,U)$, instead of $\overline{\varphi}_z(z;\delta_{n+1})$, $\overline{\lambda}_i(z,s;\delta_{n+1})$, and $\overline{F}_i(z,U;\delta_{n+1})$.

Proof (Theorem 3.15). We begin by partitioning the space as

$$\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|] = \mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|(\mathbb{1}_{E_{00}} + \mathbb{1}_{E_{11}} + \mathbb{1}_{E_{10}} + \mathbb{1}_{E_{01}})],$$

where E_{ij} for $i,j = 0,1$ denotes the event in which there are i random events for the approximation process, while $j = 0$ denotes that no events take place for the continuous process, and $j = 1$ that at least one event for the original process happens in the time interval $s \in [t_n, t_{n+1})$. The four events are considered respectively in Lemmas 3.63, 3.64, 3.65, and 3.66. Since the upper bounds in these results are non-decreasing functions of the time t_n, we combine the results of the Lemmas to obtain that there exist constants $K_1 = K_1(t_{n+1})$ and $K_2 = K_2(t_{n+1})$ such that

$$\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|] \leq (1 + \delta_{n+1}K_1)\mathbb{E}_z[\|Z_{t_n} - \overline{Z}_{t_n}\|] + \delta_{n+1}^2 K_2.$$

Since the two processes start at the same point this implies, by recursion,

$$\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|] \leq \sum_{k=1}^{n+1} K_2\delta_k^2 \left(\prod_{\ell=k}^{n+1}(1 + \delta_\ell K_1)\right). \tag{3.31}$$

In the setting when $\delta_n = \delta$ the right hand side of (3.31) becomes a geometric series which leads to the estimate

$$\begin{aligned}\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|] &\leq \delta^2 \frac{(1+\delta K_1)^{n+1} - 1}{(1+\delta K_1) - 1} K_2 \\ &\leq \delta \left(e^{K_1(n+1)\delta} - 1\right) \frac{K_2}{K_1}.\end{aligned} \tag{3.32}$$

□

3.6.2 The case of $p > 1$

In order to simply the notation we shall restrict to the case $\delta_n = \delta$. To prove the result we reason by induction on p. In particular, we consider the following inductive hypothesis. Fix $p \geq 1$ and $n \geq 1$.

Inductive Hypothesis 3.55. *Suppose the PDMP satisfies Assumptions 3.7-3.9. Moreover suppose the approximation given by Algorithm 10 satisfies Assumptions 3.10-3.13 hold for some $\delta_0 > 0$. Given $(Z_{t_n}, \overline{Z}_{t_n})$ there exist a coupling $(Z_{t_{n+1}}, \overline{Z}_{t_{n+1}})$ with respective marginals corresponding to $\mathcal{P}_\delta(Z_{t_n}, \cdot), \overline{\mathcal{P}}_\delta(\overline{Z}_{t_n}, \cdot; \delta, p)$ there exist $A = A(T)$, $B = B(T)$ independent of n such that for any $0 < \delta \leq \delta_0$*

$$\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|] \leq A\delta^{p+1} + (1 + B\delta)\mathbb{E}_z[\|Z_{t_n} - \overline{Z}_{t_n}\|]. \tag{3.33}$$

It is sufficient to show that the Inductive hypothesis holds and then the statement of the Theorem follows by recursion in n as done in (3.32). Observe that the case $p = 1$, which corresponds to Algorithm 8, holds by the proof of Section 3.6.1. Suppose the Inductive Hypothesis holds for some $p \geq 1$, let us consider the case of $p + 1$. Let us define the following coupling of $(Z_{t_{n+1}}, \overline{Z}_{t_{n+1}})$ given $(Z_{t_n}, \overline{Z}_{t_n})$.

Coupling 3.56. *Define for $0 \leq t \leq \delta$*

$$\lambda^i_{tot}(z, \overline{z}, t; \delta, p + 1) = \overline{\lambda}_i(\overline{z}, t; \delta, p + 1) + \lambda_i(\varphi_t(z)) + 1.$$

Then for $i = 1, \ldots, m$ draw the proposed event times T_i with distribution given by

$$\mathbb{P}(T_i > t) = \exp\left(-\int_0^t \lambda^i_{tot}(Z_{t_n}, \overline{Z}_{t_n}, r; \delta, p + 1)\mathrm{d}r\right).$$

Let $T_{i^*} = \min_{i=1,\ldots,m} T_i$ and let i^* be the argument that minimises T_i. If $T_{i^*} \geq \delta$, then let $Z_{t_{n+1}} = \varphi_\delta(Z_{t_n})$, $\overline{Z}_{t_{n+1}} = \overline{\varphi}_\delta(\overline{Z}_{t_n}; \delta, p + 1)$.

Consider now the case in which $T_{i^*} < \delta$. Let $U \sim \nu_{\mathcal{U}}$ and $\overline{U} \sim \mathrm{Unif}([0, 1])$ independent of the T_i's and independent of each other. Then set

$$\tau_* = T_{i^*} \quad \text{if } \overline{U} \leq \frac{\lambda_{i^*}(\varphi_{T_{i^*}}(Z_{t_n}))}{\lambda^{i^*}_{tot}(Z_{t_n}, \overline{Z}_{t_n}, T_{i^*}; \delta, p + 1)},$$

i.e. the proposed event time is accepted for the continuous time process. Alternatively set $\tau_*(z) = R$ for some constant $R > \delta$. Similarly let

$$\overline{\tau}_* = T_{i^*} \quad \text{if } \overline{U} \leq \frac{\overline{\lambda}_{i^*}(\overline{Z}_{t_n}, T_{i^*}; \delta, p + 1)}{\lambda^{i^*}_{tot}(Z_{t_n}, \overline{Z}_{t_n}, T_{i^*}; \delta, p + 1)},$$

and thus conditional on acceptance $\overline{\tau}_*$ is the next event time for the approximation process. In case of rejection set $\overline{\tau}_* = R$ for some constant $R > \delta$ as done above. Set $Z_{t_n+s} = \varphi_s(z)$ and $\overline{Z}_{t_n+s} = \overline{\varphi}_s(\overline{Z}_{t_n}; \delta, p + 1)$ for $s \in [0, T_{i^*})$. We distinguish three scenarios:

(1) The proposed switching time T_{i^*} is accepted by both processes. Then set

$$\begin{aligned} Z_{t_n+T_{i^*}} &= F_{i^*}(\varphi_{T_{i^*}}(Z_{t_n}), U), \\ \overline{Z}_{t_n+T_{i^*}} &= \overline{F}_{i^*}(\overline{\varphi}_{T_{i^*}}(\overline{Z}_{t_n}; \delta, p + 1), U; \delta, p + 1). \end{aligned}$$

To get from time $t_n + T_{i^*}$ to t_{n+1} we apply the coupling given by the Inductive Hypothesis 3.55.

(2) *The proposed switching time T_{i^*} is accepted for one process, but rejected for the other. To get from time $t_n + T_{i^*}$ to t_{n+1} we let the two processes evolve independently according to their marginal distributions.*

(3) *The proposed switching time T_{i^*} is rejected by both processes. Then set $Z_{T_{i^*}} = \varphi_{T_{i^*}}(Z_{t_n})$ and $\overline{Z}_{T_{i^*}} = \overline{\varphi}_{T_{i^*}}(\overline{Z}_{t_n}; \delta, p+1)$. To get from time $t_n + T_{i^*}$ to t_{n+1} we repeat this procedure starting at time $t_n + T_{i^*}$ and with δ replaced with $\delta - T_{i^*}$.*

Proof of Theorem 3.15. Assume $Z_0 = \overline{Z}_0 = z$. Suppose that (3.33) holds for some $p \geq 1$. We will show that (3.33) then follows for p replaced $p+1$ by using Coupling 3.56.

Suppose first that $T_{i^*} > \delta$. Then the two processes follow the deterministic flow and by Assumption 3.10 with order $p+1$ and Lemma 3.61 we have

$$\begin{aligned}\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\| \mathbb{1}_{T_{i^*} > \delta}] &= \mathbb{E}_z[\|\varphi_\delta(Z_{t_n}) - \overline{\varphi}_\delta(\overline{Z}_{t_n}; \delta, p+1)\| \mathbb{1}_{T_{i^*} > \delta}] \\ &\leq (1 + CC'\delta)\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|] + \tilde{C}^{p+2}.\end{aligned}$$

Let us consider the case (1) in Coupling 3.56 and denote the corresponding event as E_1. Then using the Inductive Hypothesis 3.55

$$\begin{aligned}&\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\| \mathbb{1}_{E_1}] = \mathbb{E}_z \mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\| \mathbb{1}_{E_1} | T_{i^*}] \\ &\leq \mathbb{E}_z[(A\delta^{p+1} + (1 + B\delta)\|Z_{t_n + T_{i^*}} - \overline{Z}_{t_n + T_{i^*}}\|) \mathbb{1}_{E_1}] \\ &= \mathbb{E}_z[(A\delta^{p+1} + (1 + B\delta)\|F_{i^*}(\varphi_{T_{i^*}}(Z_{t_n}), U) \\ &\quad - \overline{F}_{i^*}(\overline{\varphi}_{T_{i^*}}(\overline{Z}_{t_n}; \delta, p+1), U; \delta, p+1)\|) \mathbb{1}_{E_1}] \\ &\leq \mathbb{E}_z[(A\delta^{p+1} + (1 + B\delta)(M_1\delta^{p+1} + D_2\|\varphi_{T_{i^*}}(Z_{t_n}) - \overline{\varphi}_{T_{i^*}}(\overline{Z}_{t_n}; \delta, p+1)\|)) \mathbb{1}_{E_1}] \\ &\leq \mathbb{E}_z[(A\delta^{p+1} + (1 + B\delta)(M_1\delta^{p+1} + D_2(1 + CC'\delta)\|Z_{t_n} - \overline{Z}_{t_n}\| + D_2\tilde{C}\delta^{p+2})) \mathbb{1}_{E_1}].\end{aligned}$$

Here we used Assumption 3.8(b), and Lemma 3.61. Then we take advantage of

$$\mathbb{P}_z(T_{i^*} < \delta) \leq 1 - \exp(-\delta(2L(t_{n+1}, z, p+1) + m)) \leq \delta(2L(t_{n+1}, z, p+1) + m)$$

to get

$$\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\| \mathbb{1}_{E_1}] \leq \tilde{A}_1\delta^{p+2} + (1 + \tilde{B}_1\delta)\mathbb{E}_z[\|Z_{t_n} - \overline{Z}_{t_n}\|]$$

for suitable constants $\tilde{A}_1, \tilde{B}_1$ and taking advantage of $\delta < \delta_0$.

Now consider the case (2) in Coupling 3.56 and denote the corresponding event as E_2. Note that

$$\mathbb{P}_z(E_2 | Z_{t_n}, \overline{Z}_{t_n}) = \delta(2L(t_{n+1}, z, p+1) + m) \frac{|\lambda_{i^*}(\varphi_{T_{i^*}}(Z_{t_n})) - \overline{\lambda}_{i^*}(\overline{Z}_{t_n}, T_{i^*}; \delta, p+1)|}{\lambda^{i^*}_{tot}(Z_{t_n}, \overline{Z}_{t_n}, T_{i^*}; \delta, p+1)}.$$

Using Assumptions 3.7, 3.9, 3.12, and the triangle inequality we obtain

$$\mathbb{P}_z(E_2|Z_{t_n}, \overline{Z}_{t_n}) \leq \delta(2L(t_{n+1}, z, p+1) + m)(D_4 C' \|Z_{t_n} - \overline{Z}_{t_n}\| + \delta^{p+1}\overline{M}_2(\overline{Z}_{t_n})).$$

Therefore using Assumption 3.13

$$\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\| \mathbb{1}_{E_2}] \leq \tilde{A}_2 \delta^{p+2} + (1 + \tilde{B}_2\delta)\mathbb{E}_z[\|Z_{t_n} - \overline{Z}_{t_n}\|]$$

for some constants $\tilde{A}_2, \tilde{B}_2$.

Let us consider the case (3) in Coupling 3.56 and denote the corresponding event as E_3. Note that since case (3) involves repeating the coupling we may have to repeat this step an arbitrary number of times. Let q denote the number of times we propose a candidate jumping time. If $q < p+2$ then we must have reached case (1) or (2), so it is sufficient to use the respective estimate derived above to get the desired result. On the other hand, the probability that $q \geq p+2$ is bounded by $(2L(t_{n+1}, z, p+1) + m)^{p+2}\delta^{p+2}$, which gives us the correct order.

□

3.7 Proof of Theorem 3.23

3.7.1 The case of $p = 1$

To prove the result we define a coupling of the continuous process with the approximation process. The intuitive idea is that, assuming the two processes are equal at the beginning of the current time step, we can use Poisson thinning [53, 95] to simulate a proposal for the next event time that is common to both processes. This is achieved by simulating a Poisson process with rate given by the sum of the rates of the two processes. The proposal is then accepted or rejected individually for each process based on the correct switching rates. For this acceptance-rejection step a common uniform random variable is used. If the proposal is accepted for both processes, then a coupled event takes place, thus ensuring that the processes are equal after the event has happened. If the thinning step is successful it follows that the processes are equal for all $s \in (t_n, t_{n+1}]$ unless a second event takes place for the continuous time process in the current time interval, which is an event with $\mathcal{O}(\delta_{n+1}^2)$ probability. Let us now give the formal definition of the coupling.

Coupling 3.57. *Let t_n be the current time and assume $Z_{t_n} = \overline{Z}_{t_n} = z_n$. Define $\lambda^i_{tot}(z, t; \delta_{n+1}) = \overline{\lambda}_i(z, t; \delta_{n+1}) + \lambda_i(\varphi_t(z)) + 1$. Then for $i = 1, \ldots, m$ draw the proposed event times $T_i(z_n)$ with distribution*

$$\mathbb{P}(T_i(z_n) \leq t) = 1 - \exp\left(-\int_0^t \lambda^i_{tot}(z_n, r; \delta_{n+1})\mathrm{d}r\right).$$

Let $T_{i^*}(z) = \min_{i=1,\dots,m} T_i(z)$. Now let $U_{n+1} \sim \nu_{\mathcal{U}}$ and $\overline{U} \sim \mathrm{Unif}([0,1])$ independent of the T_i's and of Z_{t_n}. Then set

$$\tau(z_n) = T_{i^*}(z_n) \quad \text{if } \overline{U} \leq \frac{\lambda_{i^*}(\varphi_{T_{i^*}(z_n)}(z_n))}{\lambda^{i^*}_{tot}(z_n, T_{i^*}(z_n); \delta_{n+1})}, \tag{3.34}$$

hence upon acceptance the proposed event time is the next switching time for the continuous time process. Alternatively set $\tau(z_n) = R > \delta_{n+1}$ for some constant $R \neq T_{i^*}(z_n)$. Similarly let

$$\overline{\tau}(z_n) = T_{i^*}(z_n) \quad \text{if } \overline{U} \leq \frac{\overline{\lambda}_{i^*}(z_n, T_{i^*}(z_n); \delta_{n+1})}{\lambda^{i^*}_{tot}(z_n, T_{i^*}(z_n); \delta_{n+1})}, \tag{3.35}$$

and thus conditional on acceptance $\overline{\tau}(z_n)$ is the next event time for the approximation process. In case of rejection set $\overline{\tau}(z_n) = R > \delta_{n+1}$ for some constant $R \neq T_{i^*}(z_n)$ as done above.

If $T_{i^*} \geq \delta_{n+1}$, then let $Z_{t_n+s} = \overline{Z}_{t_n+s} = \varphi_s(z_n)$ for $s \in (0, \delta_{n+1}]$. In this case the two processes are equal at time t_{n+1}.

Alternatively, we have $T_{i^*} < \delta_{n+1}$ and thus we set $Z_{t_n+s} = \overline{Z}_{t_n+s} = \varphi_s(z_n)$ for $s \in (0, T_{i^*}(z_n))$. Then the continuous process evolves as follows:

- *if $\tau(z_n) = T_{i^*}(z_n)$, then set*

$$Z_{t_n+\tau(z_n)} = F_{i^*}(\varphi_{\tau(z_n)}(z_n), U_{n+1}).$$

  Then let the process evolve independently of the approximation until time t_{n+1}.

- *if $\tau(z_n) \neq T_{i^*}(z_n)$, the proposed event time is rejected and we let the process evolves independently of the approximation until time t_{n+1}.*

On the other hand, the approximation process evolves as follows:

- *if $\overline{\tau}(z_n) = T_{i^*}(z)$, set*

$$Z_{t_n+\overline{\tau}(z_n)} = F_{i^*}(\varphi_{\overline{\tau}(z_n)}(z_n), U_{n+1}),$$

  and finally $\overline{Z}_{t_n+s} = \varphi_s(Z_{t_n+\overline{\tau}(z_n)})$ for $s \in (\overline{\tau}(z_n), \delta_{n+1}]$.

- *if $\overline{\tau}(z_n) \neq T_{i^*}(z_n)$, then repeat this procedure from the beginning starting at time $t_n + T_{i^*}(z_n)$ and with step $\delta_{n+1} - T_{i^*}(z_n)$.*

Lemma 3.58. *Under Assumption 3.20, there exists $D(t_n, z) > 0$ such that*

$$\mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n} | Z_{t_{n-1}} = \overline{Z}_{t_{n-1}}) \leq D(t_n, z)\delta_n^2,$$

for $D(t_n, z) = (L_1(t_n, z)/2 + L_2(t_n, z) + L_3(t_n, z)/2)$.

Proof. The proof is postponed to Appendix 3.B.1. □

Proof of Theorem 3.23. By the coupling inequality we have

$$\|P_{t_n}(z,\cdot) - \overline{P}_{t_n}(z,\cdot)\|_{TV} \leq \mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n})$$

and thus it is sufficient to bound the right hand side. Apply Lemma 3.58 to obtain

$$\begin{aligned}\mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n}) &= \mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n} | Z_{t_{n-1}} \neq \overline{Z}_{t_{n-1}})\mathbb{P}_z(Z_{t_{n-1}} \neq \overline{Z}_{t_{n-1}}) \\ &\quad + \mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n} | Z_{t_{n-1}} = \overline{Z}_{t_{n-1}})(1 - \mathbb{P}_z(Z_{t_{n-1}} \neq \overline{Z}_{t_{n-1}})) \\ &\leq \mathbb{P}_z(Z_{t_{n-1}} \neq \overline{Z}_{t_{n-1}}) + D(t_n,z)\delta_n^2(1 - \mathbb{P}_z(Z_{t_{n-1}} \neq \overline{Z}_{t_{n-1}})) \\ &= (1 - D(t_n,z)\delta_n^2)\mathbb{P}_z(Z_{t_{n-1}} \neq \overline{Z}_{t_{n-1}}) + D(t_n,z)\delta_n^2.\end{aligned} \tag{3.36}$$

Thus by recursion and since $\overline{Z}_0 = Z_0 = z$ it follows that

$$\mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n}) \leq \sum_{i=1}^{n} D(t_n,z)\delta_i^2 \prod_{\ell=i+1}^{n} (1 - D(t_n,z)\delta_\ell^2).$$

In particular if $\delta_n = \delta$ for all $n \in \mathbb{N}$ we have that

$$\begin{aligned}\mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n}) &\leq D(t_n,z)\delta^2 \sum_{\ell=0}^{n-1} (1 - D(t_n,z)\delta^2)^\ell \\ &\leq 1 - (1 - D(t_n,z)\delta^2)^n \\ &\leq 1 - e^{-D(t_n,z)t_n\delta}.\end{aligned}$$

□

3.7.2 The case of $p > 1$

In order to simplify the notation we shall restrict to the case $\delta_n = \delta$. To prove the result we reason by induction on p similarly to Section 3.6.2. In particular, we consider the following inductive hypothesis. Fix $p \geq 1$ and $n \geq 1$.

Inductive Hypothesis 3.59. *Suppose $\overline{\lambda}$ satisfies Assumption 3.20 for some $\delta_0 > 0$. Given $Z_{t_n} = \overline{Z}_{t_n}$ there exist a coupling $(Z_{t_{n+1}}, \overline{Z}_{t_{n+1}})$ with respective marginals corresponding to $P_\delta(Z_{t_n},\cdot)$, $\overline{P}_\delta(\overline{Z}_{t_n},\cdot;\delta,p)$, and constants $A = A(T)$, $B = B(T)$ independent of n such that for any $0 < \delta \leq \delta_0$*

$$\mathbb{P}_z(Z_{t_{n+1}} \neq \overline{Z}_{t_{n+1}} | Z_{t_n} = \overline{Z}_{t_n}) \leq A\delta^{p+1}.$$

It is sufficient to show that the Inductive hypothesis holds and the statement of the Theorem follows by recursion in n as done in (3.36). Observe that the case $p = 1$ holds by the proof of Section 3.7.1. To obtain the result we use Coupling 3.56 but with $\overline{\varphi} = \varphi$, $\overline{F}_i = F_i$, and replacing Inductive Hypothesis 3.55 with Inductive Hypothesis 3.59. Because the strategy is similar to that in Section 3.6.2 we postpone the formal proof to Appendix 3.B.2.

3.8 Proof of Theorem 3.30

Recall in Section 3.6.1 we introduced a general framework which includes both Algorithms 7 and 8. We now introduce some further notation. Let $p_{\overline{\tau}}^{z,\delta,i}$ be a probability measure on $[0,\infty]$ which denotes the law of $\overline{\tau}$ for Algorithm i with initial condition at z for a time step of length δ. Note that for Algorithm 7 we have $p_{\overline{\tau}}^{z,\delta,i}$ is a point measure with

$$\begin{aligned} p_{\overline{\tau}}^{z,\delta,7}(\{\delta\}) &= 1 - e^{-\int_0^\delta \overline{\lambda}(z,s;\delta)\mathrm{d}s}, \\ p_{\overline{\tau}}^{z,\delta,7}(\{+\infty\}) &= e^{-\int_0^\delta \overline{\lambda}(z,s;\delta)\mathrm{d}s}. \end{aligned} \tag{3.37}$$

On the other hand, in the case of Algorithm 8 $p_{\overline{\tau}}^{z,\delta,8}$ is admits a density which is given by

$$p_{\overline{\tau}}^{z,\delta,8}(\mathrm{d}s) = \overline{\lambda}(z,s;\delta)\exp\left(-\int_0^s \overline{\lambda}(z,r;\delta)\mathrm{d}r\right)\mathrm{d}s. \tag{3.38}$$

Proof of Theorem 3.30. Fix $g \in \mathcal{G}_1$. Then by a telescoping sum we have

$$\mathbb{E}_z[g(\overline{Z}_{t_n})] - \mathbb{E}_z[g(Z_{t_n})] = \sum_{k=0}^{n-1}(\mathbb{E}_z[\mathcal{P}_{t_n-t_{k+1}}g(\overline{Z}_{t_{k+1}})] - \mathbb{E}_z[\mathcal{P}_{t_n-t_k}g(\overline{Z}_{t_k})]).$$

For each $k \in \{0,\dots,n-1\}$, set $f_k(y,s) = \mathcal{P}_{t_n-t_k-s}g(y)$ then we have

$$\mathbb{E}_z[g(\overline{Z}_{t_n})] - \mathbb{E}_z[g(Z_{t_n})] = \sum_{k=0}^{n-1}\mathbb{E}_z[f_k(\overline{Z}_{t_{k+1}},\delta_{k+1}) - f_k(\overline{Z}_{t_k},0)].$$

By conditioning on $\overline{Z}_{t_k}$ it is sufficient to prove that

$$|\mathbb{E}_z[f_k(\overline{Z}_{\delta_{k+1}},\delta_{k+1})] - f_k(z,0)| \le Re^{-\omega(t_n-t_{k+1})}\overline{G}_i(z)\delta_{k+1}^2. \tag{3.39}$$

Here with an abuse of notation we have denoted as $\overline{Z}_{\delta_{k+1}}$ the approximation process with initial condition at z and step size δ_{k+1}. Indeed if we have that (3.39) holds then by Assumption 3.28 we have

$$\begin{aligned} |\mathbb{E}_z[g(Z_{t_n})] - \mathbb{E}_z[g(\overline{Z}_{t_n})]| &\le R\sum_{k=0}^{n-1} e^{-\omega(t_n-t_{k+1})}\delta_{k+1}^2\mathbb{E}_z[\overline{G}_i(\overline{Z}_{t_k})] \\ &\le RCS_nH_i(z). \end{aligned}$$

Which gives the desired result. It remains to show that (3.39) holds.

Using that the approximation process jumps according to Q at a time determined by $p_{\overline{\tau}}^{z,\delta_{k+1},i}$ we can evaluate the expectations

$$\begin{aligned} &\mathbb{E}_z[f_k(\overline{Z}_{\delta_{k+1}},\delta_{k+1})] - f_k(z,0)] = \\ &\quad = \mathbb{E}_z[f_k(\overline{Z}_{\delta_{k+1}},\delta_{k+1})] - f_k(\varphi_{\delta_{k+1}}(z),\delta_{k+1}) + f_k(\varphi_{\delta_{k+1}}(z),\delta_{k+1}) - f_k(z,0) \end{aligned}$$

$$= \int_0^{\delta_{k+1}} Q(f_k(\varphi_{\delta_{k+1}-s}(\cdot), \delta_{k+1}))(\varphi_s(z)) - f_k(\varphi_{\delta_{k+1}}(z), \delta_{k+1}) p_{\overline{\tau}}^{z,\delta_{k+1},i}(\mathrm{d}s)$$
$$+ f_k(\varphi_{\delta_{k+1}}(z), \delta_{k+1}) - f_k(z, 0).$$

Recall $p_{\overline{\tau}}^{z,\delta,7}$ ($p_{\overline{\tau}}^{z,\delta,8}$ respectively) is defined in (3.37) (resp. (3.38)). Using the fundamental Theorem of calculus we can rewrite this as

$$\mathbb{E}_z[f_k(\overline{Z}_{\delta_{k+1}}, \delta_{k+1})] - f_k(z, 0) =$$
$$= \int_0^{\delta_{k+1}} Q(f_k(\varphi_{\delta_{k+1}-s}(\cdot), \delta_{k+1}))(\varphi_s(z)) - f_k(\varphi_{\delta_{k+1}}(z), \delta_{k+1}) p_{\overline{\tau}}^{z,\delta_{k+1},i}(\mathrm{d}s)$$
$$+ \int_0^{\delta_{k+1}} \frac{\mathrm{d}}{\mathrm{d}r} f_k(\varphi_r(z), r)\mathrm{d}r$$
$$= \int_0^{\delta_{k+1}} Q(f_k(\varphi_{\delta_{k+1}-s}(\cdot), \delta_{k+1}))(\phi_s(z)) - f_k(\varphi_{\delta_{k+1}}(z), \delta_{k+1}) p_{\overline{\tau}}^{z,\delta_{k+1},i}(ds)$$
$$+ \int_0^{\delta_{k+1}} \langle \Phi(\varphi_r(z)), \nabla f_k(\varphi_r(z), r)\rangle + (\partial_s f_k)(\varphi_r(z), r)\mathrm{d}r.$$

Note that $\partial_s f_k(y, s) = -\mathcal{L} f_k(y, s)$

$$\mathbb{E}_z[f_k(\overline{Z}_{\delta_{k+1}}, \delta_{k+1})] - f_k(z, 0) =$$
$$= \int_0^{\delta_{k+1}} Q(f_k(\varphi_{\delta_{k+1}-s}(\cdot), \delta_{k+1}))(\varphi_s(z)) - f_k(\varphi_{\delta_{k+1}}(z), \delta_{k+1}) p_{\overline{\tau}}^{z,\delta_{k+1},i}(\mathrm{d}s)$$
$$+ \int_0^{\delta_{k+1}} \langle \Phi(\varphi_r(z)), \nabla f_k(\varphi_r(z), r)\rangle - \mathcal{L} f_k(\varphi_r(z), r)\mathrm{d}r.$$

Recall $\mathcal{L}$ is given by (3.5) so we can write the above as

$$\mathbb{E}_z[f_k(\overline{Z}_{\delta_{k+1}}, \delta_{k+1})] - f_k(z, 0) =$$
$$= \int_0^{\delta_{k+1}} Q(f_k(\varphi_{\delta_{k+1}-s}(\cdot), \delta_{k+1}))(\varphi_s(z)) - f_k(\varphi_{\delta_{k+1}}(z), \delta_{k+1}) p_{\overline{\tau}}^{z,\delta_{k+1},i}(\mathrm{d}s)$$
$$+ \int_0^{\delta_{k+1}} -\lambda(\varphi_r(z))[Q(f_k(\cdot, r))(\varphi_r(z)) - f_k(\varphi_r(z), r)]\mathrm{d}r.$$

We rewrite this as

$$\mathbb{E}_z[f_k(\overline{Z}_{\delta_{k+1}}, \delta_{k+1})] - f_k(z, 0) =$$
$$= \int_0^{\delta_{k+1}} \big(Q(f_k(\varphi_{\delta_{k+1}-s}(\cdot), \delta_{k+1}))(\varphi_s(z)) - f_k(\varphi_{\delta_{k+1}}(z), \delta_{k+1})\big)(p_{\overline{\tau}}^{z,\delta_{k+1},i}(\mathrm{d}s) - \lambda(\varphi_s(z))\mathrm{d}s)$$
$$- \int_0^{\delta_{k+1}} \lambda(\varphi_r(z))[Q(f_k(\cdot, r))(\varphi_r(z)) - Q(f_k(\varphi_{\delta_{k+1}-r}(\cdot), \delta_{k+1}))(\varphi_r(z))]\mathrm{d}r \quad (3.40)$$
$$- \int_0^{\delta_{k+1}} \lambda(\varphi_r(z))[f_k(\varphi_{\delta_{k+1}}(z), \delta_{k+1}) - f_k(\varphi_r(z), r)]\mathrm{d}r.$$

We will divide the remainder of the proof into 3 steps:

Step (i): For this step we distinguish between Algorithm 7 and 8. Let

$$h_s = Q(f_k(\varphi_{\delta_{k+1}-s}(\cdot), \delta_{k+1}))(\varphi_s(z)) - f_k(\varphi_{\delta_{k+1}}(z), \delta_{k+1}).$$

Then we will show that there exists a constant $R > 0$ such that for any $h \in C_b^1([0,\delta])$ (for Algorithm 8 we only need $h \in C_b([0,\delta])$) we have

$$\left| \int_0^{\delta_{k+1}} h_s(p_{\overline{\tau}}^{z,\delta_{k+1},i}(\mathrm{d}s) - \lambda(\varphi_s(z))\mathrm{d}s) \right| \leq Re^{-\omega(t_n - t_{k+1})}\delta_{k+1}^2 \sup_{s,r\in[0,\delta_0]} K_i(z,s,r) \tag{3.41}$$

where K_i is as in Assumption 3.28.

Step (ii): For any $z \in E, r \in [0, \delta_{k+1}]$ we have

$$\begin{aligned} &|f_k(\varphi_{\delta_{k+1}}(z), \delta_{k+1}) - f_k(\varphi_r(z), r)| \leq \\ &\quad \leq R_1 \int_r^{\delta_{k+1}} e^{-\omega(t_n - t_k - s)} \lambda(\varphi_s(z))(Q\overline{G}(\varphi_s(z)) + \overline{G}(\varphi_s(z)))\mathrm{d}s. \end{aligned} \tag{3.42}$$

Step (iii): For any $z \in E, r \in [0, \delta_{k+1}]$ we have

$$\begin{aligned} &|Q(f_k(\varphi_{\delta_{k+1}-r}(\cdot), \delta_{k+1}))(z) - Q(f_k(\cdot, r))(z)| \leq \\ &\quad \leq R_1 \int_r^{\delta_{k+1}} e^{-\omega(t_n - t_k - s)} Q((Q\overline{G} + \overline{G})\lambda)(\varphi_{s-r}(z))\mathrm{d}s. \end{aligned} \tag{3.43}$$

Equation (3.39) follows from Step (i), (ii), (iii) and (3.40), as this gives

$$\begin{aligned} \mathbb{E}_z[f_k(\overline{Z}_{\delta_{k+1}}, \delta_{k+1})] - f_k(z,0) &\leq Re^{-\omega(t_n - t_{k+1})}\delta_{k+1}^2 \sup_{s,r\in[0,\delta_0]} K_i(z,r,s) \\ &+ \int_0^{\delta_{k+1}} \lambda(\varphi_r(z)) R_1 \int_r^{\delta_{k+1}} e^{-\omega(t_n - t_k - s)} Q(\lambda[Q\overline{G} + \overline{G}])(\varphi_{s-r}(z))\mathrm{d}s\mathrm{d}r \\ &+ \int_0^{\delta_{k+1}} \lambda(\varphi_r(z)) R_1 \int_r^{\delta_{k+1}} e^{-\omega(t_n - t_k - s)} \lambda(\varphi_s(z))[Q\overline{G}(\varphi_s(z)) + \overline{G}(\varphi_s(z))]\mathrm{d}s\mathrm{d}r. \end{aligned}$$

Recall that $\overline{G}_i(z,r,s)$ is given by (3.11), then we have

$$|\mathbb{E}_z[f_k(\overline{Z}_{\delta_{k+1}}, \delta_{k+1})] - f_k(z,0)| \leq Re^{-\omega(t_n - t_{k+1})}\delta_{k+1}^2 \sup_{s,r\in[0,\delta_0]} \overline{G}_i(z,r,s).$$

Proof of Step (i): This step follows from Lemma 3.67. It remains to find a bound for $|h_s|$ and $|\partial_s h_s|$ for the case of Algorithm 7. By Assumption 3.26

$$|h_s| = \left|Q(f_k(\varphi_{\delta_{k+1}-s}(\cdot), \delta_{k+1}))(\varphi_s(z)) - f_k(\varphi_{\delta_{k+1}}(z), \delta_{k+1})\right|$$

$$\begin{aligned}
&\leq \left|Q(f_k(\varphi_{\delta_{k+1}-s}(\cdot),\delta_{k+1}))(\varphi_s(z)) - \mu(f_k(\varphi_{\delta_{k+1}-s}(\cdot),\delta_{k+1}))\right| \\
&\quad + \left|\mu(f_k(\varphi_{\delta_{k+1}-s}(\cdot),\delta_{k+1})) - f_k\circ\varphi_{\delta_{k+1}-s}(\varphi_s(z),\delta_{k+1})\right| \\
&\leq R_1 e^{-\omega(t_n-t_{k+1})}[Q\overline{G}(\varphi_s(z)) + \overline{G}(\varphi_s(z))]. \qquad (3.44)
\end{aligned}$$

For Algorithm 7 we also require to control $|\partial_s h_s|$ for which we require a bound on the derivative of f, for this case we use Assumption 3.27. Note that

$$\begin{aligned}
\partial_s h_s &= \Big|\langle\Phi(\varphi_s(z)), \nabla_z(Q(f_k(\varphi_{\delta_{k+1}-s}(\cdot),\delta_{k+1})))(\varphi_s(z))\rangle \\
&\qquad - Q(\langle\Phi, \nabla_z(f_k(\varphi_{\delta_{k+1}-s}(\cdot),\delta_{k+1}))\rangle)(\varphi_s(z)))\Big| \\
&= [\mathbf{\Phi}, Q](f_k(\varphi_{\delta_{k+1}-s}(\cdot),\delta_{k+1}))(\varphi_s(z)).
\end{aligned}$$

Recall here we have defined the commutator in Section 3.2 and we are denoting by $\mathbf{\Phi}$ the differential operator corresponding to Φ. This term is bounded by Assumption 3.27 and we have

$$|\partial_s h_s| \leq R_2 e^{-\omega(t_n-t_{k+1})}\overline{G}(z). \qquad (3.45)$$

Combining Lemma 3.67 with (3.44) and (3.45) we have that (3.41) holds.

Proof of Step (ii): Observe that since

$$\partial_s f_k(y,s) + \langle\Phi(y), \nabla f_k(y,s)\rangle = -\lambda(y)[Qf_k(y,s) - f_k(y,s)]$$

we have

$$\begin{aligned}
&|f_k(\varphi_{\delta_{k+1}}(z),\delta_{k+1}) - f_k(\varphi_r(z),r)| = \left|\int_r^{\delta_{k+1}} \frac{\mathrm{d}}{\mathrm{d}s} f_k(\varphi_s(z),s)\mathrm{d}s\right| \\
&= \left|\int_r^{\delta_{k+1}} \lambda(\varphi_s(z))[Qf_k(\varphi_s(z),s) - f_k(\varphi_s(z),s)]\mathrm{d}s\right| \\
&\leq \int_r^{\delta_{k+1}} \lambda(\varphi_s(z))[|Qf_k(\varphi_s(z),s) - \mu(Q(f_k(\cdot,s)))| + |\mu(f_k(\cdot,s)) - f_k(\varphi_s(z),s)|]\mathrm{d}s.
\end{aligned}$$

We can bound this using Assumption 3.26 we obtain (3.42).

Proof of Step (iii): Applying (3.42) with z replaced by $\varphi_{-r}(y)$ and applying Q we have (3.43).

□

3.A Proofs of Section 3.4.1

3.A.1 Proof of Theorem 3.15

In this section we prove the lemmas that are used to prove Theorem 3.15 in the case $p = 1$. In the proofs that follow we simplify the notation denoting the approximations as $\overline{\varphi}_z(z)$, $\overline{\lambda}_i(z,s)$, and $\overline{F}_i(z,U)$, instead of $\overline{\varphi}_z(z;\delta_{n+1})$, $\overline{\lambda}_i(z,s;\delta_{n+1})$, and

$\overline{F}_i(z, U; \delta_{n+1})$. Before proving bounds on the events E_{ij}, let us state three simple lemmas which will be used multiple times in the proof. The proofs are omitted as they are a straightforward consequence of the assumptions.

Lemma 3.60. *Assumption 3.7 implies that for any $\delta_0 > 0$ there exists a constant $C' = C'(\delta_0) > 0$ such that for any $z, z' \in E$ and $t \in (0, \delta_0)$ it holds that*

$$\|\varphi_t(z) - \varphi_t(z')\| \leq C'\|z - z'\|. \tag{3.46}$$

Moreover, for any $t \in (0, \delta_0)$ and any $z, z' \in E$ we have the alternative bound

$$\|\varphi_t(z) - \varphi_t(z')\| \leq (1 + CC't)\|z - z'\|.$$

Lemma 3.61. *Suppose Assumptions 3.7 and 3.10 hold. Then for any $p \geq 1$, $s \geq 0$ and $z, z' \in E$ it holds that*

$$\|\varphi_s(z) - \overline{\varphi}_s(z')\| \leq C'\|z - z'\| + \tilde{C}s^{p+1}.$$

Moreover, using Lemma 3.60 we can replace C' with $1 + CC'\delta$.

Lemma 3.62. *Under Assumptions 3.9, 3.12, and 3.13, for any $t \geq 0$, $p \geq 1$ there exists a positive constant $L(t, z, p)$ such that*

$$\sup_{r \in [0,t], s \in [0, \delta_0]} \max\{\lambda(Z_r), \overline{\lambda}(\overline{Z}_r, s; \delta_{n+1}, p)\} \leq L(t, z, p) \quad \text{a.s.}$$

where in particular $z = Z_0 = \overline{Z}_0$. Note if $p = 1$ we write $L(t, z, 1) = L(t, z)$.

We can now start showing a bound on event E_{00}, followed by the other events.

Lemma 3.63. *Under Assumptions 3.7 and 3.10, it holds that*

$$\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\| \mathbb{1}_{E_{00}}] \leq (1 + \delta_{n+1}CC')\mathbb{E}_z[\|Z_{t_n} - \overline{Z}_{t_n}\|] + \tilde{C}\delta_{n+1}^2.$$

Proof. On E_{00} we are interested only in the error introduced by the integrator $\overline{\varphi}$. We have

$$\begin{aligned}
\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\| \mathbb{1}_{E_{00}}] &= \mathbb{E}_z\left[\left\|\varphi_{\delta_{n+1}}(Z_{t_n}) - \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n})\right\| \mathbb{1}_{E_{00}}\right] \\
&\leq \mathbb{E}_z\left[\left\|\varphi_{\delta_{n+1}}(Z_{t_n}) - \varphi_{\delta_{n+1}}(\overline{Z}_{t_n})\right\|\right] \\
&\quad + \mathbb{E}_z\left[\left\|\varphi_{\delta_{n+1}}(\overline{Z}_{t_n}) - \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n})\right\|\right],
\end{aligned}$$

Then one can directly apply to the first term Assumption 3.7 and thus Lemma 3.60, together with the assumption that $\delta_n \leq \delta_0$, and to the second term Assumption 3.10 to obtain the wanted result for $C' = C'(\delta_0)$. □

Lemma 3.64. *Under Assumption 3.7, parts (b) and (c) of Assumption 3.8, as well as Assumptions 3.9-3.13, it holds that*

$$\begin{aligned}\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|\mathbb{1}_{E_{11}}] \leq \delta_{n+1}^2\big(mK_2 + m(m-1)\tilde{K}_2 + 2B(t_{n+1},z)(L(t_{n+1},z))^2\big)\\ + \delta_{n+1}(mK_1 + m(m-1)\tilde{K}_1)\mathbb{E}_z\left[\|Z_{t_n} - \overline{Z}_{t_n}\|\right],\end{aligned}$$

where $L(t_{n+1},z)$ was defined in Lemma 3.62, while

$$\begin{aligned}K_1 &= D_2(C')^2 L(t_{n+1},z),\\ K_2 &= (D_2\tilde{C}C' + L(t_{n+1},z)(2D_3 + M_1C')),\\ \tilde{K}_1 &= D_2(C')^2(L(t_{n+1},z))^2,\\ \tilde{K}_2 &= (D_2\tilde{C}C' + (L(t_{n+1},z))^2)(2D_3 + M_1C').\end{aligned}$$

Proof. Let us first restrict to the event that Z_s has only one event for $s \in (t_n, t_{n+1}]$ and denote such event as $\overline{E}$. For any $i,j \in \{1,\dots,m\}$, let A_{ij} be the event that Z_{t_n} jumps according to F_i and $\overline{Z}_{t_n}$ jumps according to $\overline{F}_j$ and no other jumps occur. Note that $\{A_{ij}\}_{i,j=1}^m$ is a partition of $E_{11} \cap \overline{E}$ so we may write

$$\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|\mathbb{1}_{E_{11}\cap\overline{E}}] = \sum_{i,j=1}^m \mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|\mathbb{1}_{A_{ij}}].$$

Let us first consider event A_{ii}, i.e. the processes have a switch according to kernels F_i and $\overline{F}_i$. Considering Coupling 3.54, we first observe that A_{ii} is an order δ_{n+1} event. This follows from the fact that in this case we require

$$\tilde{U}_{n+1}^i \leq \min\left\{1 - \exp\left(-\int_0^{\delta_{n+1}} \overline{\lambda}_i(\overline{Z}_{t_n}, s)\mathrm{d}s\right), 1 - \exp\left(-\int_0^{\delta_{n+1}} \lambda_i(\varphi_s(Z_{t_n}))\mathrm{d}s\right)\right\}.$$

Therefore, using that $1 - \exp(-z) \leq z$ we obtain

$$\begin{aligned}\mathbb{E}_z\left[\mathbb{1}_{A_{ii}}|Z_{t_n}, \overline{Z}_{t_n}\right] &\leq \min\left\{\int_0^{\delta_{n+1}} \overline{\lambda}_i(\overline{Z}_{t_n}, s)\mathrm{d}s, \int_0^{\delta_{n+1}} \lambda_i(\varphi_s(Z_{t_n}))\mathrm{d}s\right\}\\ &\leq \delta_{n+1}L(t_{n+1},z),\end{aligned} \tag{3.47}$$

where $L(t_{n+1},z) < \infty$ was defined in Lemma 3.62. We can then separate the effects of the different approximations by the triangle inequality:

$$\begin{aligned}&\mathbb{E}_z\left[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|\mathbb{1}_{A_{ii}}\right] =\\ &= \mathbb{E}_z[\|\varphi_{\delta_{n+1}-\tau_{n+1}}(F_i(\varphi_{\tau_{n+1}}(Z_{t_n}), U_{n+1})) - \overline{\varphi}_{\delta_{n+1}-\overline{\tau}_{n+1}}(\overline{F}_i(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}), U_{n+1}))\|\mathbb{1}_{A_{ii}}]\\ &\leq \mathbb{E}_z\left[\|\varphi_{\delta_{n+1}-\tau_{n+1}}(F_i(\varphi_{\tau_{n+1}}(Z_{t_n}), U_{n+1})) - \varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}F_i(\varphi_{\overline{\tau}_{n+1}}(Z_{t_n}), U_{n+1})\|\mathbb{1}_{A_{ii}}\right]\end{aligned} \tag{*}$$

$$+\mathbb{E}_z\Big[\|\varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}(F_i(\varphi_{\overline{\tau}_{n+1}}(Z_{t_n}),U_{n+1}))-\varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}(F_i(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}),U_{n+1}))\|\mathbb{1}_{A_{ii}}\Big] \tag{**}$$

$$+\mathbb{E}_z\Big[\|\varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}(F_i(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}),U_{n+1}))-\varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}(\overline{F}_i(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}),U_{n+1}))\|\mathbb{1}_{A_{ii}}\Big] \tag{***}$$

$$+\mathbb{E}_z\Big[\|\varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}(\overline{F}_i(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}),U_{n+1}))-\overline{\varphi}_{\delta_{n+1}-\overline{\tau}_{n+1}}(\overline{F}_i(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}),U_{n+1}))\|\mathbb{1}_{A_{ii}}\Big]. \tag{****}$$

For term (*) we first compare both terms to $F_i(\varphi_{\delta_{n+1}}(Z_{t_n}),U_{n+1})$, and then we condition on all random variables apart from U_{n+1} in order to apply Assumption 3.8(c):

$$\begin{aligned}(*) &\le \mathbb{E}_z\big[\|\varphi_{\delta_{n+1}-\tau_{n+1}}(F_i(\varphi_{\tau_{n+1}}(Z_{t_n}),U_{n+1}))-F_i(\varphi_{\delta_{n+1}}(Z_{t_n}),U_{n+1})\|\mathbb{1}_{A_{ii}}\big]\\ &\quad+\mathbb{E}_z\left[\|F_i(\varphi_{\delta_{n+1}}(Z_{t_n}),U_{n+1})-\varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}F_i(\varphi_{\overline{\tau}_{n+1}}(Z_{t_n}),U_{n+1})\|\mathbb{1}_{A_{ii}}\right]\\ &\le 2D_3\delta_{n+1}\mathbb{P}_z(A_{ii})\\ &\le \delta_{n+1}^2 2D_3L(t_{n+1},z).\end{aligned}$$

In the last inequality we used the inequality derived in (3.47). Term (**) can be bounded applying inequality (3.46), then conditioning on $Z_{t_n},\overline{Z}_{t_n},\overline{\tau}_{n+1}$ and using Assumption 3.8(b), and finally applying Lemma 3.61 and (3.47):

$$\begin{aligned}(**) &\le C'\mathbb{E}_z\left[\|F_i(\varphi_{\overline{\tau}_{n+1}}(Z_{t_n}),U_{n+1})-F_i(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}),U_{n+1})\|\mathbb{1}_{A_{ii}}\right]\\ &\le C'D_2\mathbb{E}_z\left[\|\varphi_{\overline{\tau}_{n+1}}(Z_{t_n})-\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n})\|\mathbb{1}_{A_{ii}}\right]\\ &\le (C')^2D_2\mathbb{E}_z[\|Z_{t_n}-\overline{Z}_{t_n}\|\mathbb{1}_{A_{ii}}]+\delta_{n+1}^2C'D_2\tilde{C}\\ &\le (C')^2D_2L(t_{n+1},z)\delta_{n+1}\mathbb{E}_z[\|Z_{t_n}-\overline{Z}_{t_n}\|]+\delta_{n+1}^2C'D_2\tilde{C}.\end{aligned} \tag{3.48}$$

Term (***) is estimated by inequality (3.46), then again conditioning on $Z_{t_n},\overline{Z}_{t_n},\overline{\tau}_{n+1}$ and applying Assumption 3.11, and finally using (3.47):

$$\begin{aligned}(***) &\le C'\mathbb{E}_z\left[\|F_i(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}),U_{n+1})-\overline{F}_i(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}),U_{n+1})\|\mathbb{1}_{A_{ii}}\right]\\ &\le C'M_1\delta_{n+1}\mathbb{P}_z(A_{ii})\\ &\le \delta_{n+1}^2C'M_1L(t_{n+1},z).\end{aligned} \tag{3.49}$$

Term (****) is bounded using Assumption 3.10 and bounding by 1 the probability of A_{ii}:

$$(****)\le\tilde{C}\delta_{n+1}^2\mathbb{P}_z(A_{ii})\le\tilde{C}\delta_{n+1}^2.$$

Putting together terms (*), (**), (***), (****) we obtain the following bound on event A_{ii}:

$$\mathbb{E}_z\left[\|Z_{t_{n+1}}-\overline{Z}_{t_{n+1}}\|\mathbb{1}_{A_{ii}}\right]\le\delta_{n+1}^2K_2+\delta_{n+1}K_1\mathbb{E}_z\left[\|Z_{t_n}-\overline{Z}_{t_n}\|\right] \tag{3.50}$$

where K_1, K_2 are as in the statement of the lemma.

Now consider event A_{ij} for $i \neq j$. In this case we take advantage of independence of $\tilde{U}^i_{n+1}$ and $\tilde{U}^j_{n+1}$ to conclude that

$$\begin{aligned}
&\mathbb{E}_z[\mathbb{1}_{A_{ij}}|Z_{t_n}, \overline{Z}_{t_n}] \leq \\
&\quad \leq \left(1 - \exp\left(-\int_0^{\delta_{n+1}} \lambda_i(\varphi_s(Z_{t_n}))\mathrm{d}s\right)\right)\left(1 - \exp\left(-\int_0^{\delta_{n+1}} \overline{\lambda}_j(\overline{Z}_{t_n}, s)\mathrm{d}s\right)\right) \\
&\quad \leq \int_0^{\delta_{n+1}} \lambda_i(\varphi_s(Z_{t_n}))\mathrm{d}s \int_0^{\delta_{n+1}} \overline{\lambda}_j(\overline{Z}_{t_n}, s)\mathrm{d}s \\
&\quad \leq \delta^2_{n+1}(L(t_{n+1}, z))^2. \qquad (3.51)
\end{aligned}$$

Then we can use the decomposition

$$\begin{aligned}
&\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|\mathbb{1}_{A_{ij}}] = \\
&= \mathbb{E}_z[\|\varphi_{\delta_{n+1}-\tau_{n+1}}(F_i(\varphi_{\tau_{n+1}}(Z_{t_n}), U_{n+1})) \\
&\quad - \overline{\varphi}_{\delta_{n+1}-\overline{\tau}_{n+1}}(\overline{F}_j(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}), U_{n+1})))\|\mathbb{1}_{A_{ij}}] \\
&\leq \mathbb{E}_z[\|\varphi_{\delta_{n+1}-\tau_{n+1}}(F_i(\varphi_{\tau_{n+1}}(Z_{t_n}), U_{n+1}))) - F_i(\varphi_{\delta_{n+1}}(Z_{t_n}), U_{n+1})))\|\mathbb{1}_{A_{ij}}] \qquad (\dagger) \\
&+ \mathbb{E}_z[\|F_i(\varphi_{\delta_{n+1}}(Z_{t_n}), U_{n+1}))) - F_j(\varphi_{\delta_{n+1}}(Z_{t_n}), U_{n+1})))\|\mathbb{1}_{A_{ij}}] \qquad (\ddagger) \\
&+ \mathbb{E}_z[\|F_j(\varphi_{\delta_{n+1}}(Z_{t_n}), U_{n+1}))) - \varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}(F_j(\varphi_{\overline{\tau}_{n+1}}(Z_{t_n}), U_{n+1})))\|\mathbb{1}_{A_{ij}}] \qquad (\ddagger\dagger) \\
&+ \mathbb{E}_z[\|\varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}(F_j(\varphi_{\overline{\tau}_{n+1}}(Z_{t_n}), U_{n+1}))) \qquad (\ddagger\ddagger) \\
&\quad - \overline{\varphi}_{\delta_{n+1}-\overline{\tau}_{n+1}}(\overline{F}_j(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}), U_{n+1})))\|\mathbb{1}_{A_{ij}}]
\end{aligned}$$

To bound (†) and (‡†) we use Assumption 3.8(c), while for (‡) we add and subtract $\varphi_{\delta_{n+1}}(Z_{t_n})$ and use Assumption 3.8(a), and for (‡‡) we use a similar argument to the A_{ii} case. Combining this with the bound in (3.51) we obtain

$$\begin{aligned}
\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|\mathbb{1}_{A_{ij}}] \leq &((L(t_{n+1}, z))^2(2D_3 + D_1) + \tilde{K_2})\delta^2_{n+1} \\
&+ \delta_{n+1}\tilde{K}_1\mathbb{E}_z\left[\|Z_{t_n} - \overline{Z}_{t_n}\|\right]
\end{aligned}$$

where $\tilde{K}_1, \tilde{K}_2$ are as in the statement of the lemma.

Let us finally consider $\overline{E}^c$, i.e. the case in which Z_t has two or more jumps. The

probability this event is given by

$$\begin{aligned}\mathbb{P}_z(\overline{E}^c) &= \mathbb{E}_z\Bigg[\int_0^{\delta_{n+1}} \left(1-\exp\left(-\int_0^{\delta_{n+1}-t}\lambda(\varphi_s(F_{I_{n+1}}(\varphi_t(Z_{t_n}),U_{n+1})))\mathrm{d}s\right)\right)\\ &\qquad \lambda(\varphi_t(Z_{t_n}))\exp\left(-\int_0^t \lambda(\varphi_r(Z_{t_n}))\mathrm{d}r\right)\mathrm{d}t\Bigg]\\ &\leq \mathbb{E}_z\left[\int_0^{\delta_{n+1}}\left(\int_0^{\delta_{n+1}-t}\lambda(\varphi_s(F_{I_{n+1}}(\varphi_t(z),U_{n+1})))\mathrm{d}s\right)\lambda(\varphi_t(z))\mathrm{d}t\right]\\ &\leq \delta_{n+1}^2(L(t_{n+1},z))^2.\end{aligned} \tag{3.52}$$

Then we can bound the norms $\|Z_{t_{n+1}}\|$ and $\|\overline{Z}_{t_{n+1}}\|$ by Assumption 3.13 to obtain

$$\begin{aligned}\mathbb{E}_z\left[\|Z_{t_{n+1}}-\overline{Z}_{t_{n+1}}\|\mathbb{1}_{E_{11}\cap\overline{E}^c}\right] &\leq 2B(t_{n+1},z)\mathbb{E}_z\left[\mathbb{1}_{\overline{E}^c}\right]\\ &\leq 2\delta_{n+1}^2 B(t_{n+1},z)(L(t_{n+1},z))^2.\end{aligned} \tag{3.53}$$

Combining the bounds on $\overline{E}$ and $\overline{E}^c$ we obtain the statement of the lemma. □

Lemma 3.65. *Under Assumptions 3.7, 3.8(a), 3.9-3.13, it holds that*

$$\begin{aligned}\mathbb{E}_z\left[\|Z_{t_{n+1}}-\overline{Z}_{t_{n+1}}\|\mathbb{1}_{E_{10}}\right] \leq \delta_{n+1}\, m(C')^2(D_1D_4+2D_2L(t_{n+1},z))\,\mathbb{E}_z\left[\|Z_{t_n}-\overline{Z}_{t_n}\|\right]\\ +\delta_{n+1}^2(\tilde{C}+C'(D_2\tilde{C}+mD_1M_2(t_n,z)+2mM_1L(t_{n+1},z))).\end{aligned}$$

Proof. Recall that E_{10} is the event in which there are no switches for Z_s for $s\in(t_n,t_{n+1}]$ and there is one event for the approximation. Taking advantage of the coupling of the two processes as described in Coupling 3.54 we find that E_{10} takes place as long as for some i

$$\tilde{U}_{n+1}^i \in \left(1-\exp\left(-\int_0^{\delta_{n+1}}\lambda_i(\varphi_s(Z_{t_n}))\mathrm{d}s\right), 1-\exp\left(-\int_0^{\delta_{n+1}}\overline{\lambda}_i(\overline{Z}_{t_n},s)\mathrm{d}s\right)\right].$$

Then we can estimate the probability of this event as follows:

$$\begin{aligned}&\mathbb{E}_z[\mathbb{1}_{E_{10}}|Z_{t_n},\overline{Z}_{t_n}] \leq\\ &\leq \sum_{i=1}^m\left|\exp\left(-\int_0^{\delta_{n+1}}\lambda_i(\varphi_s(Z_{t_n}))\mathrm{d}s\right)-\exp\left(-\int_0^{\delta_{n+1}}\overline{\lambda}_i(\overline{Z}_{t_n},s)\mathrm{d}s\right)\right|\\ &\leq \sum_{i=1}^m\int_0^{\delta_{n+1}}\left|\lambda_i(\varphi_s(Z_{t_n}))-\overline{\lambda}_i(\overline{Z}_{t_n},s)\right|\mathrm{d}s\end{aligned}$$

where we used that $\exp(-z)$ is 1-Lipschitz for $z\geq 0$. Then we find bounds for $\mathbb{E}_z[\mathbb{1}_{E_{10}}]$ and $\mathbb{E}_z[\mathbb{1}_{E_{10}}|Z_{t_n},\overline{Z}_{t_n}]$ respectively. For the first case we use the triangle inequality,

followed by observing that λ and φ_s are Lipschitz by the inequality shown in (3.46) and then Assumption 3.12:

$$\begin{aligned}
\mathbb{E}_z[\mathbb{1}_{E_{10}}] &\leq \sum_{i=1}^{m} \mathbb{E}_z\Bigg[\int_0^{\delta_{n+1}} \left|\lambda_i(\varphi_s(Z_{t_n})) - \lambda_i(\varphi_s(\overline{Z}_{t_n}))\right| \mathrm{d}s \\
&\quad + \int_0^{\delta_{n+1}} \left|\lambda_i(\varphi_s(\overline{Z}_{t_n})) - \overline{\lambda}_i(\overline{Z}_{t_n}, s)\right| \mathrm{d}s\Bigg] \\
&\leq \sum_{i=1}^{m} \mathbb{E}_z\left[\int_0^{\delta_{n+1}} D_4 C' \|Z_{t_n} - \overline{Z}_{t_n}\| \mathrm{d}s + \int_0^{\delta_{n+1}} \delta_{n+1} \overline{M}_2(\overline{Z}_{t_n}) \mathrm{d}s\right] \\
&\leq \delta_{n+1} m D_4 C' \mathbb{E}_z[\|Z_{t_n} - \overline{Z}_{t_n}\|] + m\delta_{n+1}^2 \mathbb{E}_z[\overline{M}_2(\overline{Z}_{t_n})] \\
&\leq \delta_{n+1} m D_4 C' \mathbb{E}_z[\|Z_{t_n} - \overline{Z}_{t_n}\|] + \delta_{n+1}^2 m M_2(t_n, z),
\end{aligned} \tag{3.54}$$

where in the last inequality we used again Assumption 3.12. Alternatively, we can bound the switching rates by Lemma 3.62:

$$\begin{aligned}
\mathbb{E}_z[\mathbb{1}_{E_{10}} | Z_{t_n}, \overline{Z}_{t_n}] &\leq \sum_{i=1}^{m} \int_0^{\delta_{n+1}} \left(|\lambda_i(\varphi_s(Z_{t_n}))| + |\overline{\lambda}_i(\overline{Z}_{t_n}, s)|\right) \mathrm{d}s \\
&\leq 2m\delta_{n+1} L(t_{n+1}, z),
\end{aligned} \tag{3.55}$$

Let us now focus on bounding the distance between the two processes. On event E_{10} we have $Z_{t_{n+1}} = \varphi_{\delta_{n+1}}(Z_{t_n})$, while $\overline{Z}_{t_{n+1}} = \overline{\varphi}_{\delta_{n+1}-\overline{\tau}_{n+1}}(\overline{F}(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}), U_{n+1}))$ where $\overline{\tau}_{n+1}$ is the time of the event for the approximation. By triangle inequality we can decompose the distance in the following terms

$$\begin{aligned}
&\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\| \mathbb{1}_{E_{10}}] = \\
&= \mathbb{E}_z\left[\|\varphi_{\delta_{n+1}}(Z_{t_n}) - \overline{\varphi}_{\delta_{n+1}-\overline{\tau}_{n+1}}(\overline{F}_{\overline{I}_{n+1}}(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}), U_{n+1}))\| \mathbb{1}_{E_{10}}\right] \\
&\leq \mathbb{E}_z\left[\|\varphi_{\delta_{n+1}}(Z_{t_n}) - \varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}(F_{\overline{I}_{n+1}}(\varphi_{\overline{\tau}_{n+1}}(Z_{t_n}), U_{n+1}))\| \mathbb{1}_{E_{10}}\right] \qquad (*) \\
&+ \mathbb{E}_z\Big[\|\varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}(F_{\overline{I}_{n+1}}(\varphi_{\overline{\tau}_{n+1}}(Z_{t_n}), U_{n+1})) \qquad (**) \\
&\quad - \varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}(F_{\overline{I}_{n+1}}(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}), U_{n+1}))\| \mathbb{1}_{E_{10}}\Big] \\
&+ \mathbb{E}_z\Big[\|\varphi_{\delta_{n+1}-\overline{\tau}_{n+1}}(F_{\overline{I}_{n+1}}(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}), U_{n+1})) \qquad (***) \\
&\quad - \overline{\varphi}_{\delta_{n+1}-\overline{\tau}_{n+1}}(\overline{F}_{\overline{I}_{n+1}}(\overline{\varphi}_{\overline{\tau}_{n+1}}(\overline{Z}_{t_n}), U_{n+1}))\| \mathbb{1}_{E_{10}}\Big] \\
&= (*) + (**) + (***).
\end{aligned}$$

In order to estimate term (*) we apply inequality (3.46), then Assumption 3.8(a) by

conditioning on $Z_{t_n}, \overline{\tau}_{n+1}$, and then we apply (3.54):

$$\begin{aligned}(*) &\leq C'\mathbb{E}_z[\|\varphi_{\overline{\tau}_{n+1}}(Z_{t_n}) - F_{\overline{I}_{n+1}}(\varphi_{\overline{\tau}_{n+1}}(Z_{t_n}), U_{n+1})\|\mathbb{1}_{E_{10}}] \\ &\leq C'D_1\mathbb{E}_z[\mathbb{1}_{E_{10}}] \\ &\leq \delta_{n+1}mD_4(C')^2D_1\mathbb{E}_z[\|Z_{t_n} - \overline{Z}_{t_n}\|] + \delta_{n+1}^2 mC'D_1M_2(t_n, z).\end{aligned} \tag{3.56}$$

For term (**) we use the same reasoning of (3.48) together with the estimate (3.55):

$$(**) \leq \delta_{n+1}m2L(t_{n+1}, z)(C')^2D_2\mathbb{E}_z[\|Z_{t_n} - \overline{Z}_{t_n}\|] + C'D_2\tilde{C}\delta_{n+1}^2.$$

Then for term (***) we follow the reasoning in (3.49) and apply estimate (3.55) to obtain

$$(***) \leq \delta_{n+1}^2\big(2mC'M_1L(t_{n+1}, z) + \tilde{C}\big).$$

The statement of the lemma follows then by combining estimates (*), (**), (***). □

Lemma 3.66. *Under Assumptions 3.7, 3.8(a), 3.9, 3.10, 3.12, 3.13, it holds that*

$$\begin{aligned}\mathbb{E}_z\left[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|\mathbb{1}_{E_{01}}\right] \leq \delta_{n+1}\, mC'\big(L(t_{n+1}, z) + C'D_1D_4\big)\,\mathbb{E}_z\left[\|Z_{t_n} - \overline{Z}_{t_n}\|\right] \\ + \delta_{n+1}^2\big(2B(t_{n+1}, z)(L(t_{n+1}, z))^2 + mC'D_1M_2(t_n, z) + \tilde{C}\big).\end{aligned}$$

Proof. Recall that E_{01} is the event in which for $s \in [t_n, t_{n+1})$ there are no switches for $\overline{Z}_s$, while there is at least one event for Z_s. Similarly to the proof of Lemma 3.64, let us denote as $\overline{E}$ the event in which there is exactly one event for Z_s in the current time interval. On $\overline{E}^c$ we can use the bound (3.53). On $\overline{E}$

$$\begin{aligned}&\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|\mathbb{1}_{E_{01}\cap\overline{E}}] = \\ &= \mathbb{E}_z\left[\left\|\varphi_{\delta_{n+1}-\tau_{n+1}}(F_{I_{n+1}}(\varphi_{\tau_{n+1}}(Z_{t_n}), U_{n+1})) - \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n})\right\|\mathbb{1}_{E_{01}\cap\overline{E}}\right] \\ &\leq \mathbb{E}_z\left[\left\|\varphi_{\delta_{n+1}-\tau_{n+1}}(F_{I_{n+1}}(\varphi_{\tau_{n+1}}(Z_{t_n}), U_{n+1})) - \varphi_{\delta_{n+1}}(Z_{t_n})\right\|\mathbb{1}_{E_{01}\cap\overline{E}}\right] \qquad (*) \\ &\quad + \mathbb{E}_z\left[\|\varphi_{\delta_{n+1}}(Z_{t_n}) - \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n})\|\mathbb{1}_{E_{01}\cap\overline{E}}\right]. \qquad (**)\end{aligned}$$

In order to find an estimate for term (*) observe that the probability of E_{01} can be estimated similarly to what done for the probability of E_{10}. Then following the reasoning in (3.56) we obtain

$$(*) \leq \delta_{n+1}m(C')^2D_1D_4\mathbb{E}_z[\|Z_{t_n} - \overline{Z}_{t_n}\|] + \delta_{n+1}^2 mC'D_1M_2(t_n, z).$$

Similarly, for term (**) it is sufficient to apply Lemma 3.61 and then to bound $\mathbb{E}[\mathbb{1}_{E_{01}}]$ by the probability that the continuous process has a random event:

$$(**) \leq \delta_{n+1}mC'L(t_{n+1}, z)\mathbb{E}_z\left[\|Z_{t_n} - \overline{Z}_{t_n}\|\right] + \tilde{C}\delta_{n+1}^2.$$

□

3.A.2 Proof of Corollary 3.17

Proof. We only need to show that Assumptions 3.8 and 3.13 are verified in this setting under Assumption 3.16. Clearly the process moves with bounded velocity, and thus 3.13 holds. Then let us focus on verifying Assumption 3.8 and consider the ℓ^1-norm. For part (a) it is clear that for $z = (x, v)$

$$\mathbb{E}[\|z - F_i(z, \tilde{U})\|] = \mathbb{E}[\|v - F_i^v((x, v), \tilde{U})\|] \leq V_{max}.$$

Then consider part (b). For $z' = (y, w)$

$$\begin{aligned}\mathbb{E}[\|F_i(z, \tilde{U}) - F_i(z', \tilde{U})\|] &\leq \|x - y\| + \mathbb{E}[\|F_i^v((x, v), \tilde{U}) - F_i^v((y, w), \tilde{U})\|] \\ &\leq \|x - y\| + \frac{\|v - w\|}{V_{min}} V_{max} + \mathbb{E}[\|F_i^v((x, w), \tilde{U}) - F_i^v((y, w), \tilde{U})\|] \\ &\leq \max\left\{\frac{V_{max}}{V_{min}}, 1 + D\right\} \|z - z'\|.\end{aligned}$$

In the second inequality we used the triangle inequality and that $\|v - w\| \geq V_{min}$, while in the last inequality we bounded the rightmost term by Assumption 3.16. Let us focus on part (c). For the position part we have for $s \in [0, \delta]$ and $z = (x, v)$

$$\begin{aligned}&\mathbb{E}[\|\varphi_{\delta-s}(\varphi_s(z), F_i^v((\varphi_s(z), v), \tilde{U})) - \varphi_\delta(z)\|] = \\ &= \mathbb{E}\Big[\Big\|x + \int_0^s \Phi(v)\mathrm{d}s + \int_0^{\delta-s} \Phi(F_i^v((\varphi_s(z), v), U))\mathrm{d}r - x - \int_0^\delta \Phi(v)\mathrm{d}s\Big\|\Big] \\ &\leq \mathbb{E}[\|s\Phi(v) + (\delta - s)\Phi(F_i^v((\varphi_s(z), v), U)) - \delta\Phi(v)\|] \\ &\leq \mathbb{E}[\|(\delta - s)(\Phi(F_i^v((\varphi_s(z), v), U)) - \Phi(v))\|] \\ &\leq \delta C\mathbb{E}[\|F_i^v((\varphi_s(z), v), U) - v\|] \\ &\leq \delta C V_{max}.\end{aligned}$$

On the other hand, for the velocity part we obtain using Assumptions 3.16 and 3.7

$$\begin{aligned}\mathbb{E}[\|F_i^v((\varphi_s(z), v), \tilde{U}) - F_i^v((\varphi_\delta(z), v), \tilde{U})\|] &\leq D\|\varphi_s(z) - \varphi_\delta(z)\| \\ &\leq DC'\|z - \varphi_{\delta-s}(z)\| \\ &\leq \delta D(C')^2.\end{aligned}$$

Therefore part (c) of Assumption 3.8 holds with $D_3 = D(C')^2 + CV_{max}$. □

3.A.3 Proof of Proposition 3.19

Proof. In the proof of Theorem 3.15 boundedness of the PDMP is used only in the case $p = 1$ to deal with the event in which the PDMP has two or more jumps in the same time step (see Lemmas 3.64 and 3.65, in particular Equation (3.53)). Then it is sufficient to show that a similar bound holds also under Assumption 3.18 instead of Assumption 3.13. Let $p = 1$ and consider the case of Lemma 3.65, i.e. restricting

to event $E_{10} \cap \overline{E}$, which is the event in which the continuous time process has two or more jumps, while the approximation has zero jumps. Then we want to bound

$$\begin{aligned}
&\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\| \mathbb{1}_{E_{10}\cap\overline{E}}] = \\
&= \sum_{\ell\geq 2} \int_{A_\ell} \mathbb{E}_z\big[\|\varphi_{s_\ell} \circ F_{I^\ell_{n+1}}(\cdot, U_{\ell-1}) \circ \varphi_{s_{\ell-1}} \circ \ldots \circ F_{I^1_{n+1}}(\cdot, U_0) \circ \varphi_{s_0}(Z_{t_n}) \\
&\qquad - \overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n})\| p_{Z_{t_n}}(\mathrm{d}s)\big],
\end{aligned}$$

where $s_0, \ldots, s_{\ell-1}$ are the interarrival times of the random jumps of Z_t, I^ℓ_{n+1} denote the index of the ℓ-th jump to occur between time t_n and t_{n+1}, $s_\ell = \delta_{n+1} - \sum_{i=1}^{\ell-1} s_i$, $U_0, \ldots, U_{\ell-1} \overset{iid}{\sim} \nu_{\mathcal{U}}$,

$$A_\ell = \left\{ s = (s_0, \ldots, s_\ell) : \sum_{i=1}^{\ell} s_i = \delta_{n+1}, s_i > 0 \right\},$$

and $p_{Z_{t_n}}(\mathrm{d}s)$ is the law of the interarrival times,. Then we have

$$\begin{aligned}
\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\| \mathbb{1}_{E_{10}\cap\overline{E}}] \leq \mathbb{E}\Bigg[\sum_{\ell\geq 2} \int_{A_\ell} \mathbb{E}_z\Big[\Big(&\|\varphi_{s_\ell} \circ F_{I^\ell_{n+1}}(\cdot, U_{\ell-1}) \circ \cdots \circ \varphi_{s_0}(Z_{t_n})\| \\
&+ \|\overline{\varphi}_{\delta_{n+1}}(\overline{Z}_{t_n})\| \Big) \Big| Z_{t_n} \Big] p_{Z_{t_n}}(\mathrm{d}s) \Bigg]
\end{aligned}$$

Now we use (3.46) to conclude that φ_s has linear growth for some constant L, and so $\|\varphi_s(z)\| \leq L(\|z\| + 1)$. It follows that

$$\begin{aligned}
&\mathbb{E}_z\left[\|\varphi_{s_\ell} \circ F_{I^\ell_{n+1}}(\cdot, U_{\ell-1}) \circ \varphi_{s_{\ell-1}} \circ F_{I^{\ell-1}_{n+1}}(\cdot, U_{\ell-2}) \circ \cdots \circ \varphi_{s_0}(Z_{t_n})\| | Z_{t_n} \right] \leq \\
&\quad \leq L\mathbb{E}_z\left[\Big(\|F_{I^\ell_{n+1}}(\cdot, U_{\ell-1}) \circ \varphi_{s_{m-1}} \circ F_{I^{\ell-1}_{n+1}}(\cdot, U_{\ell-2}) \circ \cdots \circ \varphi_{s_0}(Z_{t_n})\| + 1 \Big) | Z_{t_n} \right] \\
&\quad \leq L(1 + D_1) + L\mathbb{E}_z\left[\Big(\|\varphi_{s_{\ell-1}} \circ F_{I^{\ell-1}_{n+1}}(\cdot, U_{\ell-2}) \circ \cdots \circ \varphi_{s_0}(Z_{t_n})\| \Big) | Z_{t_n} \right].
\end{aligned}$$

In the last inequality we used that $\mathbb{E}[\|F_i(z, U)\|] \leq \|z\| + D_1$ for any i, which is implied by Assumption 3.8(a). Therefore by recursion we have

$$\mathbb{E}_z\left[\|\varphi_{s_\ell} \circ F_{I^\ell_{n+1}}(\cdot, U_{\ell-1}) \circ \cdots \circ \varphi_{s_0}(Z_{t_n})\| | Z_{t_n} \right] \leq \sum_{i=1}^{\ell} (1 + D_1) L^i + L^{\ell+1}(1 + \|Z_{t_n}\|).$$

Moreover we also have that $\|\overline{\varphi}_s(z)\| \leq \delta^2_{n+1} + L(\|z\| + 1)$ by Assumption 3.10. It follows that

$$\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\| \mathbb{1}_{E_{10}\cap\overline{E}}] =$$

$$\leq \sum_{\ell\geq 2}\mathbb{E}_z\left[\int_{A_\ell}\left(\sum_{i=1}^{\ell}(1+D_1)L^i + L^{\ell+1}(1+\|Z_{t_n}\|)+\delta_{n+1}^2+L(\|\overline{Z}_{t_n}\|+1)\right)p_{Z_{t_n}}(\mathrm{d}s)\right]$$
$$\leq \mathbb{E}_z\left[\sum_{\ell\geq 2}\left(\sum_{i=1}^{\ell}(1+D_1)L^i + L^{\ell+1}(1+\|Z_{t_n}\|)+\delta_{n+1}^2+L(\|\overline{Z}_{t_n}\|+1)\right)p_{Z_{t_n}}(A_\ell)\right]$$
$$\leq \tilde{L}\mathbb{E}_z\left[(1+\|Z_{t_n}\|+\|\overline{Z}_{t_n}\|)\sum_{\ell\geq 2}\ell L^\ell p_{Z_{t_n}}(A_\ell)\right]$$

for some constant $\tilde{L}$ which depends only on D_1, L, δ_0. The function $f(\ell) = \ell L^\ell$ is increasing in the number of jumps and therefore because the switching rates have a global upper bound λ_{max} we obtain

$$\begin{aligned}
&\mathbb{E}_z[\|Z_{t_{n+1}} - \overline{Z}_{t_{n+1}}\|\mathbb{1}_{E_{10}\cap\overline{E}}] \leq \\
&\quad\leq \tilde{L}\mathbb{E}_z\left[(1+\|Z_{t_n}\|+\|\overline{Z}_{t_n}\|)\sum_{\ell\geq 2}\ell L^\ell e^{-\delta_{n+1}\lambda_{max}}\frac{(\delta_{n+1}\lambda_{max})^\ell}{\ell!}\right] \\
&\quad\leq \tilde{L}(1+2B(t_n,z))\sum_{\ell\geq 2}\ell L^\ell e^{-\delta_{n+1}\lambda_{max}}\frac{(\delta_{n+1}\lambda_{max})^\ell}{\ell!},
\end{aligned}$$

where in the last inequality we used Assumption 3.18. It remains to show that the sum is of order δ_{n+1}^2. This can be proved as follows

$$\begin{aligned}
&\sum_{\ell\geq 2}\ell L^\ell e^{-\delta_{n+1}\lambda_{max}}\frac{(\delta_{n+1}\lambda_{max})^\ell}{\ell!} = e^{(L-1)\delta_{n+1}\lambda_{max}}\sum_{\ell\geq 2}e^{-L\delta_{n+1}\lambda_{max}}\frac{(L\delta_{n+1}\lambda_{max})^\ell}{(\ell-1)!} \\
&\quad= e^{(L-1)\delta_{n+1}\lambda_{max}}L\delta_{n+1}\lambda_{max}\sum_{\ell\geq 1}e^{-L\delta_{n+1}\lambda_{max}}\frac{(L\delta_{n+1}\lambda_{max})^\ell}{\ell!} \\
&\quad= e^{(L-1)\delta_{n+1}\lambda_{max}}L\delta_{n+1}\lambda_{max}(1-e^{-L\delta_{n+1}\lambda_{max}}) \\
&\quad\leq \delta_{n+1}^2 e^{(L-1)\delta_0\lambda_{max}}L\lambda_{max}.
\end{aligned}$$

In particular we used that $\delta_n \leq \delta_0$ for all $n \in \mathbb{N}$.

The same proof holds on the event $E_{10}\cap\overline{E}$, and thus we have proved the wanted result. □

3.B Proofs of Section 3.4.2

3.B.1 Proof of Theorem 3.23: the case of $p = 1$

Proof of Lemma 3.58. Let us take advantage of the construction in Coupling 3.57. First consider the case in which $T_{i^*}(Z_{t_{n-1}}) > \delta_n$. In this case there are no random

events for either process in the time interval $(t_{n-1}, t_n]$ and therefore $Z_{t_n} = \overline{Z}_{t_n} = \varphi_{\delta_n}(Z_{t_{n-1}})$. Now, consider the case where $T_{i^*} \leq \delta_n$. In this scenario, there are three disjoint events:

- The proposed switching time is accepted by both processes. Denote this event as E_1.
- The proposed switching time is accepted by one process, and rejected by the other. Denote this event as E_2.
- The proposed switching time is rejected for both processes. Denote this event as E_3.

Therefore we have

$$\mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n} | Z_{t_{n-1}} = \overline{Z}_{t_{n-1}}) = \sum_{i=1}^{3} \mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n}, E_i | Z_{t_{n-1}} = \overline{Z}_{t_{n-1}}).$$

We start with event E_1. In this case we have that $\overline{Z}_{t_n} \neq Z_{t_n}$ if the continuous time process has at least one more jump in time interval $(t_{n-1} + T_{i^*}, t_n]$. Now let $\lambda(z) = \sum_{i=1}^m \lambda_i(z)$ and $\lambda_{tot}(z, t; \delta_n) = \sum_{i=1}^m \lambda^i_{tot}(z, t; \delta_n)$. Observe that, conditional on $Z_{t_{n-1}}$, the minimum of the m proposed random times is distributed as $\mathbb{P}(T_{i^*} \leq t) = 1 - \exp(-\int_0^t \lambda_{tot}(Z_{t_{n-1}}, s; \delta_n)\mathrm{d}s)$. Then bounding by 1 the probability that both proposals are accepted, and conditioning on $Z_{t_{n-1}}$ we obtain

$$\begin{aligned}
&\mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n}, E_1 | Z_{t_{n-1}} = \overline{Z}_{t_{n-1}}) \leq \\
&\leq \mathbb{E}_z\Bigg[\int_0^{\delta_n} \lambda_{tot}(Z_{t_{n-1}}, t; \delta_n) e^{-\int_0^t \lambda_{tot}(Z_{t_{n-1}}, s; \delta_n)\mathrm{d}s} \\
&\quad \times \left(1 - \exp\left(-\int_t^{\delta_n} \lambda(\varphi_s(F_{i^*}(\varphi_t(Z_{t_{n-1}}), U_n)))\mathrm{d}s\right)\right)\mathrm{d}t \Bigg].
\end{aligned}$$

Then using that $1 - \exp(-z) \leq z$, that $\exp(-z) \leq 1$ for $z \geq 0$ and by Fubini's theorem we obtain the following bound:

$$\begin{aligned}
&\mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n}, E_1 | Z_{t_{n-1}} = \overline{Z}_{t_{n-1}}) \leq \\
&\quad \leq \mathbb{E}_z\left[\int_0^{\delta_n} \int_t^{\delta_n} \lambda_{tot}(Z_{t_{n-1}}, t; \delta_n) \lambda(\varphi_s(F_{i^*}(\varphi_t(Z_{t_{n-1}}), U_n)))\mathrm{d}s\mathrm{d}t \right] \\
&\quad \leq \int_0^{\delta_n} \int_t^{\delta_n} \mathbb{E}_z\left[\lambda_{tot}(Z_{t_{n-1}}, t; \delta_n) \lambda(\varphi_s(F_{i^*}(\varphi_t(Z_{t_{n-1}}), U_n))\right] \mathrm{d}s\mathrm{d}t \\
&\quad \leq \delta_n^2 L_1(t_n, z)/2.
\end{aligned}$$

Note that in the last inequality the bound $L_1(t_n, z)$ follows from part (a) of Assumption 3.20.

Let us now consider event E_2. As the proposal $T_{i^*}(Z_{t_{n-1}})$ is accepted for one process only, it must be that

$$\overline{U} \in \left(\min\left\{ \frac{\lambda_{i^*}(\varphi_{T_{i^*}}(Z_{t_{n-1}}))}{\lambda^{i^*}_{tot}(Z_{t_{n-1}}, T_{i^*}(Z_{t_{n-1}}); \delta_n)}, \frac{\overline{\lambda}_{i^*}(Z_{t_{n-1}}, T_{i^*}(Z_{t_{n-1}}))}{\lambda^{i^*}_{tot}(Z_{t_{n-1}}, T_{i^*}(Z_{t_{n-1}}); \delta_n)} \right\}, \right.$$
$$\left. \max\left\{ \frac{\lambda_{i^*}(\varphi_{T_{i^*}}(Z_{t_{n-1}}))}{\lambda^{i^*}_{tot}(Z_{t_{n-1}}, T_{i^*}(Z_{t_{n-1}}); \delta_n)}, \frac{\overline{\lambda}_{i^*}(Z_{t_{n-1}}, T_{i^*}(Z_{t_{n-1}}); \delta_n)}{\lambda^{i^*}_{tot}(Z_{t_{n-1}}, T_{i^*}(Z_{t_{n-1}}); \delta_n)} \right\} \right].$$

Therefore using that $\overline{U}$ and T_{i^*} are independent we obtain

$$\mathbb{P}(Z_{t_n} \neq \overline{Z}_{t_n}, E_2 | Z_{t_{n-1}} = \overline{Z}_{t_{n-1}}) =$$
$$= \mathbb{E}\left[\mathbb{1}_{\{T_{i^*}(Z_{t_{n-1}}) \leq \delta_n\}} \left| \frac{\overline{\lambda}_{i^*}(Z_{t_{n-1}}, T_{i^*}(Z_{t_{n-1}}); \delta_n) - \lambda_{i^*}(\varphi_{T_{i^*}}(Z_{t_{n-1}}))}{\lambda^{i^*}_{tot}(Z_{t_{n-1}}, T_{i^*}(Z_{t_{n-1}}); \delta_n)} \right| \right].$$

By the definition given in Coupling 3.57 we have $\lambda^{i^*}_{tot}(z, t; \delta_n) \geq 1$. Using part (b) of Assumption 3.20 and Fubini's theorem:

$$\begin{aligned}
\mathbb{P}(Z_{t_n} \neq \overline{Z}_{t_n}, E_2 | Z_{t_{n-1}} = \overline{Z}_{t_{n-1}}) &\leq \delta_n \mathbb{E}_z \left[\overline{M}_2(Z_{t_{n-1}}) \mathbb{1}_{\{T_{i^*}(Z_{t_{n-1}}) \leq \delta_n\}} \right] \\
&\leq \delta_n \mathbb{E}_z \left[\overline{M}_2(Z_{t_{n-1}}) \int_0^{\delta_n} \lambda_{tot}(Z_{t_{n-1}}, t; \delta_n) e^{-\int_0^t \lambda_{tot}(Z_{t_{n-1}}, s; \delta_n) \mathrm{d}s} \mathrm{d}t \right] \\
&\leq \delta_n \int_0^{\delta_n} \mathbb{E}_z \left[\lambda_{tot}(Z_{t_{n-1}}, t; \delta_n) \overline{M}_2(Z_{t_{n-1}}) \right] \mathrm{d}t \\
&\leq \delta_n^2 L_2(t_n, z).
\end{aligned}$$

Finally, we focus on E_3. On this event, the processes remain equal unless there is (at least) a switch for either process for $t \in (t_{n-1} + T_{i^*}(Z_{t_n}), t_n)$. Recall $\overline{\lambda}(z, s; \delta_n) = \sum_{i=1}^m \overline{\lambda}_i(z, s; \delta_n)$. Using this observation together with Assumption 3.20 and the facts that on this event $T_{i^*}(Z_{t_n}) \leq \delta_n$ and that $1 - \exp(-z) \leq z$ we obtain

$$\begin{aligned}
\mathbb{P}(Z_{t_n} \neq \overline{Z}_{t_n}, E_3 | Z_{t_{n-1}} = \overline{Z}_{t_{n-1}}) \leq \mathbb{E}_z \Bigg[\int_0^{\delta_n} \Bigg(\Bigg(1 - \exp\left(- \int_t^{\delta_n} \lambda(\varphi_r(Z_{t_{n-1}})) \mathrm{d}r \right) \Bigg) \\
+ \left(1 - \exp\left(- \int_t^{\delta_n} \overline{\lambda}(Z_{t_{n-1}}, r; \delta_n) \mathrm{d}r \right) \right) \Bigg) \\
\lambda_{tot}(Z_{t_{n-1}}, t; \delta_n) \exp\left(- \int_0^t \lambda_{tot}(Z_{t_{n-1}}, s; \delta_n) \mathrm{d}s \right) \mathrm{d}t \Bigg] \\
\leq \mathbb{E}_z \left[\int_0^{\delta_n} \int_t^{\delta_n} \lambda_{tot}(Z_{t_{n-1}}, t; \delta_n) \left(\lambda(\varphi_r(Z_{t_{n-1}})) + \overline{\lambda}(Z_{t_{n-1}}, r; \delta_n) \right) \mathrm{d}r \mathrm{d}t \right]
\end{aligned}$$

$$= \int_0^{\delta_n} \int_t^{\delta_n} \mathbb{E}_z \Big[\lambda_{tot}(Z_{t_{n-1}}, t; \delta_n) \left(\lambda(\varphi_r(Z_{t_{n-1}})) + \overline{\lambda}(Z_{t_{n-1}}, r; \delta_n) \right) \Big] \mathrm{d}r \mathrm{d}t$$
$$\leq \delta_n^2 L_3(t_n, z)/2.$$

Combining the three bounds on events E_1, E_2, E_3 we obtain the statement. □

3.B.2 Proof of Theorem 3.23: the case of $p > 1$

Proof of Theorem 3.23. Observe that if $T_{i^*} > \delta$ the two processes are equal at time δ and thus the probability that $Z_{t_n} \neq \overline{Z}_{t_n}$ is 0. We analyse in turn the three events E_1, E_2, E_3 which were defined in Section 3.7.1 in the proof of Lemma 3.58. Define the event $E_= = \{Z_{t_{n-1}} = \overline{Z}_{t_{n-1}}\}$.

On event E_1, the proposal T_{i^*} is accepted by both processes. Then we reformulate $\mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n}, E_1 | E_=)$ in terms of the conditional probability

$$\mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n}, E_1 | E_=) = \mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n} | Z_{t_{n-1}+T_{i^*}} = \overline{Z}_{t_{n-1}+T_{i^*}}, T_{i^*} < \delta) \mathbb{P}_z(E_1 | E_=).$$

The first term on the right hand side can be bounded by applying Inductive Hypothesis 3.59. Moreover we can use the bound $\mathbb{P}_z(E_1|E_=) \leq \mathbb{P}_z(T_{i^*} < \delta | E_=)$ for the rightmost term to obtain

$$\mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n}, E_1 | E_=) \leq A\delta^{p+1} \mathbb{E}_z \left[\left(1 - \exp\left(- \int_0^{\delta} \lambda_{tot}(Z_{t_{n-1}}, t; \delta, p+1) \mathrm{d}t \right) \right) \right]$$
$$\leq A\delta^{p+2} \sup_{s \in [0,\delta]} \mathbb{E}_z \left[\lambda_{tot}(Z_{t_n}, s; \delta, p+1) \right] \leq A\delta^{p+2} L_4(t_n, z).$$

In the last inequality we took advantage of the bound $1 - \exp(-z) \leq x$ which is true for $z > 0$.

On event E_2 the proposal T_{i^*} is accepted for one process, and rejected for the other. This happens when

$$\overline{U} \in \left(\min \left\{ \frac{\lambda_{i^*}(\varphi_{T_{i^*}}(Z_{t_{n-1}}))}{\lambda^{i^*}_{tot}(Z_{t_{n-1}}, T_{i^*}; \delta, p+1)}, \frac{\overline{\lambda}_{i^*}(Z_{t_{n-1}}, T_{i^*}; \delta, p+1)}{\lambda^{i^*}_{tot}(Z_{t_{n-1}}, T_{i^*}; \delta, p+1)} \right\}, \right.$$
$$\left. \max \left\{ \frac{\lambda_{i^*}(\varphi_{T_{i^*}}(Z_{t_{n-1}})))}{\lambda^{i^*}_{tot}(Z_{t_{n-1}}, T_{i^*}; \delta, p+1)}, \frac{\overline{\lambda}_{i^*}(Z_{t_{n-1}}, T_{i^*}; \delta, p+1)}{\lambda^{i^*}_{tot}(Z_{t_{n-1}}, T_{i^*}; \delta, p+1)} \right\} \right],$$

and therefore with probability

$$\left| \frac{\overline{\lambda}_{i^*}(Z_{t_{n-1}}, T_{i^*}; \delta, p+1) - \lambda_{i^*}(\varphi_{T_{i^*}}(Z_{t_{n-1}}))}{\lambda^{i^*}_{tot}(z, T_{i^*}(z); \delta, p+1)} \right| \leq$$
$$\leq \left| \overline{\lambda}_{i^*}(Z_{t_{n-1}}, T_{i^*}; \delta, p+1) - \lambda_{i^*}(\varphi_{T_{i^*}}(Z_{t_{n-1}})) \right|$$
$$\leq \delta^{p+1} \overline{M}_2(Z_{t_{n-1}})$$

where we used that by definition $\lambda_{tot}^{i^*} \geq 1$ and then that $\overline{\lambda}_{i^*}(\cdot,\cdot;\delta,p+1)$ is an approximation of $p+1$ order. Thus we have

$$\begin{aligned}\mathbb{P}_z(Z_{t_n} \neq \overline{Z}_{t_n}, E_2|E_=) &\leq \delta^{p+1}\mathbb{E}_z\left[\overline{M}_2(Z_{t_{n-1}})\mathbb{P}_z(T_{i^*} < \delta|Z_{t_{n-1}}, E_=)\right] \\ &\leq \delta^{p+2}\sup_{s\in[0,\delta]}\mathbb{E}_z\left[\overline{M}_2(Z_{t_{n-1}})\lambda_{tot}(Z_{t_n},s;\delta,p+1)\right] \\ &\leq \delta^{p+2}L_4(t_n,z)\end{aligned}$$

Finally consider event E_3. Similarly to the proof of Theorem 3.15 it is sufficient to bound the event that $p+2$ proposal times occur before the end of the time interval, which is bounded by Assumption 3.20.

□

3.C Proofs of Section 3.4.3

3.C.1 Proofs of Theorem 3.30 and its corollaries

Lemma 3.67. *Suppose λ and $\overline{\lambda}$ satisfy Assumption 3.12 (a). We will consider the two algorithms separately. For Algorithm 7, let $p_{\overline{\tau}}^{z,\delta,7}$ be given by* (3.37) *then for any $h \in \mathcal{C}_b^1([0,\delta])$ we have*

$$\begin{aligned}\left|\int_0^\delta h_s p_{\overline{\tau}}^{z,\delta,7}(\mathrm{d}s) - h_s\lambda(\varphi_s(z))\mathrm{d}s\right| \leq \delta^2 \sup_{s,r\in[0,\delta]} \big(&|\partial_r h_r|\overline{\lambda}(z,s;\delta) \\ &+ |h_s|\big(\overline{\lambda}(z,s;\delta)\overline{\lambda}(z,r;\delta) + \overline{M}_2(z)\big)\big).\end{aligned}$$

For Algorithm 8, let $p_{\overline{\tau}}^{z,\delta,8}$ be given by (3.38) *then for any $h \in \mathcal{C}_b([0,\delta])$*

$$\left|\int_0^\delta h_s p_{\overline{\tau}}^{z,\delta,8}(\mathrm{d}s) - \int_0^\delta \lambda(\varphi_s(z))h_s\mathrm{d}s\right| \leq \delta^2 \sup_{s,r\in[0,\delta]} \big(|h_s|(\overline{\lambda}(z,s;\delta)\overline{\lambda}(z,r;\delta) + \overline{M}_2(z))\big)$$

Proof of Lemma 3.67. First consider the case where $p_{\overline{\tau}}^{z,\delta,7}$ is given by (3.37), and fix $h \in C_b^1([0,\delta])$. Then

$$\left|\int_0^\delta h_s p_{\overline{\tau}}^{z,\delta,7}(\mathrm{d}s) - h_s\lambda(\varphi_s(z))\mathrm{d}s\right| = \left|h_\delta\left(1 - e^{-\int_0^\delta \overline{\lambda}(z,s;\delta)\mathrm{d}s}\right) - \int_0^\delta h_s\lambda(\varphi_s(z))\mathrm{d}s\right|.$$

We can rewrite

$$1 - e^{-\int_0^\delta \overline{\lambda}(z,s;\delta)\mathrm{d}s} = \int_0^\delta \overline{\lambda}(z,s;\delta)e^{-\int_0^\delta \overline{\lambda}(z,r;\delta)\mathrm{d}r}\mathrm{d}s.$$

Therefore we have

$$\left|\int_0^\delta h_s p_{\overline{\tau}}^{z,\delta,7}(\mathrm{d}s) - h_s\lambda(\varphi_s(z))\mathrm{d}s\right| = \tag{3.57}$$

$$\begin{aligned}
&= \left|h_\delta \int_0^\delta \overline{\lambda}(z,s;\delta)e^{-\int_0^\delta \overline{\lambda}(z,r;\delta)\mathrm{d}r}\mathrm{d}s - \int_0^\delta h_s\lambda(\varphi_s(z))\mathrm{d}s\right| \\
&\leq \left|\int_0^\delta (h_\delta - h_s)\overline{\lambda}(z,s;\delta)e^{-\int_0^\delta \overline{\lambda}(z,r;\delta)\mathrm{d}r}\mathrm{d}s\right| \\
&\quad + \left|\int_0^\delta h_s\left(\overline{\lambda}(z,s;\delta)e^{-\int_0^\delta \overline{\lambda}(z,r;\delta)\mathrm{d}r} - \lambda(\varphi_s(z))\right)\mathrm{d}s\right|.
\end{aligned} \tag{3.58}$$

We can use Assumption 3.12 (a) and that that $1 - e^{-y} \leq y$ for $y > 0$ to bound the integrand of the second term on the right of (3.58),

$$\begin{aligned}
\left|\overline{\lambda}(z,s;\delta)e^{-\int_0^\delta \overline{\lambda}(z,r;\delta)\mathrm{d}r} - \lambda(\varphi_s(z))\right| &\leq \left|\overline{\lambda}(z,s;\delta)(1 - e^{-\int_0^\delta \overline{\lambda}(z,r;\delta)\mathrm{d}r})\right| \\
&\quad + \left|\overline{\lambda}(z,s;\delta) - \lambda(\varphi_s(z))\right| \\
&\leq \overline{\lambda}(z,s;\delta)\int_0^\delta \overline{\lambda}(z,r;\delta)\mathrm{d}r + \delta\overline{M}_2(z).
\end{aligned} \tag{3.59}$$

For the first term on the right hand side of (3.58) we use that

$$|h_\delta - h_s| \leq (\delta - s)\sup_{r\in[0,\delta]}|\partial_r h_r|. \tag{3.60}$$

Applying (3.59) and (3.60) to (3.58) we have

$$\begin{aligned}
&\left|\int_0^\delta h_s p_{\overline{\tau}}^{z,\delta,7}(\mathrm{d}s) - h_s\lambda(\varphi_s(z))\mathrm{d}s\right| \leq \\
&\leq \left|\sup_{r\in[0,\delta]}|\partial_r h_r|\int_0^\delta (\delta - s)\overline{\lambda}(z,s;\delta)e^{-\int_0^\delta \overline{\lambda}(z,r;\delta)\mathrm{d}r}\mathrm{d}s\right| \\
&\quad + \left|\int_0^\delta |h_s|\left(\overline{\lambda}(z,s;\delta)\int_0^\delta \overline{\lambda}(z,r;\delta)\mathrm{d}r + \delta\overline{M}_2(z)\right)\mathrm{d}s\right| \\
&\leq \delta^2 \sup_{s,r\in[0,\delta]}\left(|\partial_r h_r|\overline{\lambda}(z,s;\delta) + |h_s|\left(\overline{\lambda}(z,s;\delta)\overline{\lambda}(z,r) + \overline{M}_2(z)\right)\right).
\end{aligned}$$

Let us consider the case where $p_{\overline{\tau}}^{z,\delta,8}$ is given by (3.38). We use (3.59) to bound

$$\left|\int_0^\delta h_s p_{\overline{\tau}}^{z,\delta,8}(\mathrm{d}s) - h_s\lambda(\varphi_s(z))\mathrm{d}s\right| \leq$$

$$\begin{aligned}
&\leq \int_0^\delta |h_s| \left| \overline{\lambda}(z,s;\delta) \exp\left(- \int_0^s \overline{\lambda}(z,r;\delta) \mathrm{d}r \right) - \lambda(\varphi_s(z)) \right| \mathrm{d}s \\
&\leq \int_0^\delta |h_s| \left(\overline{\lambda}(z,s;\delta) \int_0^\delta \overline{\lambda}(z,r;\delta) \mathrm{d}r + \delta \overline{M}_2(z) \right) \mathrm{d}s \\
&\leq \delta^2 \sup_{s,r \in [0,\delta]} \left(|h_s| (\overline{\lambda}(z,s;\delta) \overline{\lambda}(z,r;\delta) + \overline{M}_2(z)) \right).
\end{aligned}$$

□

Proof of Corollary 3.33. First observe that (3.15) follows from (3.12). Then (3.16) is obtained by adding (3.10) and (3.12). To obtain (3.17) we use that

$$\begin{aligned}
\left| \frac{1}{N} \sum_{n=1}^N \mathbb{E}_z[g(\overline{Z}_{t_n})] - \mu(g) \right| \leq & \left| \frac{1}{N} \sum_{n=1}^N (\mathbb{E}_z[g(\overline{Z}_{t_n})] - \mathbb{E}_z[g(Z_{t_n})]) \right| \\
& + \left| \frac{1}{N} \sum_{n=1}^N \mathbb{E}_z[g(Z_{t_n})] - \mu(g) \right|.
\end{aligned}$$

We bound this using (3.10) and (3.12)

$$\begin{aligned}
\left| \frac{1}{N} \sum_{n=1}^N \mathbb{E}_z[g(\overline{Z}_{t_n})] - \mu(g) \right| &\leq C\delta \overline{G}_2(z) + C\overline{G}_2(z) \frac{1}{N} \sum_{n=1}^N e^{-\omega t_n} \\
&\leq C\overline{G}_2(z) \left(\delta + \frac{1}{t_N} \right).
\end{aligned}$$

□

Proof of Corollary 3.34. It is sufficient to show for S_n given by (3.13) that $S_n \to 0$ as $n \to \infty$. Fix $\eta > 0$. Then we have

$$S_n = \sum_{k=0}^{\eta-1} \delta_{k+1}^2 e^{-\omega(t_n - t_{k+1})} + \sum_{k=\eta}^{n-1} \delta_{k+1}^2 e^{-\omega(t_n - t_{k+1})}.$$

Consider the first term:

$$\sum_{k=0}^{\eta-1} \delta_{k+1}^2 e^{-\omega(t_n - t_{k+1})} \leq \sup_k \delta_k \int_0^{t_n - \eta} e^{-\omega(t_n - s)} \mathrm{d}s = \sup_k \delta_k \frac{e^{-\omega\eta} - e^{-\omega t_n}}{\omega}.$$

Consider the second term:

$$\sum_{k=\eta}^{n-1} \delta_{k+1}^2 e^{-\omega(t_n - t_{k+1})} \leq \left(\sup_{k \in \{\eta, \ldots, n\}} \delta_k \right) \int_{t_\eta}^{t_n} e^{-\omega(t_n - s)} \mathrm{d}s$$

$$= \frac{1 - e^{-\omega(t_n - t_\eta)}}{\omega} \sup_{k \in \{\eta, \dots, n\}} \delta_k.$$

Therefore

$$\limsup_{n \to \infty} S_n \leq \left(\sup_{k \geq 0} \delta_k \right) \frac{e^{-\omega\eta}}{\omega} + \frac{1}{\omega} \sup_{k \geq \eta} \delta_k.$$

Since η is arbitrary we let η tend to ∞ which gives that $S_n \to 0$ as $n \to \infty$. □

3.C.2 Proofs of Example 3.42

Proof of Proposition 3.43. Fix $f \in C_b^1(\mathbb{R}^d \times \mathbb{R}^d)$. Then by the chain rule

$$\|\nabla_{q,p} \mathcal{P}_t f(q,p)\| = \|\mathbb{E}[\nabla_{q,p}(Q_t, P_t)(\nabla_{q,p} f)(Q_t, P_t)]\| \leq \|f\|_{C_b^1} \mathbb{E}[\|\nabla_{q,p}(Q_t, P_t)\|].$$

Notice that there is a version of (Q_t, P_t) which is differentiable with respect to the initial conditions since we can write (Q_t, P_t) as the composition of smooth operators. Let T_i denote the i-th refreshment time and $\xi_i \sim N(0_d, I_d)$ the corresponding refreshed velocity. Set $T_0 = 0$. We shall track for which refreshment times we have that $\nu \leq T_i - T_{i-1} \leq K$. Let M_t denote the number of refreshment times before time t which have this property and let N_t denote the total number of refreshment times before time t. Note that conditional on N_t, M_t is distributed according to a Binomial distribution with N_t trials and success rate $e^{-\lambda\nu} - e^{-\lambda K_1}$.

To stress the dependence on the initial condition for the remainder of the proof we shall write $(Q_t^{q,p}, P_t^{q,p})$ to denote the process at time t with initial condition (q,p). Then by (3.22) we have

$$\begin{aligned} \|(Q_t^{q,p}, P_t^{q,p}) - (Q_t^{\overline{q},\overline{p}}, P_t^{\overline{q},\overline{p}})\| &= \|\varphi_{t - T_{N_t}}(Q_{T_{N_t}}^{q,p}, \xi_{N_t}) - \varphi_{t - T_{N_t}}(Q_{T_{N_t}}^{\overline{q},\overline{p}}, \xi_{N_t})\| \\ &\leq C \|Q_{T_{N_t}}^{q,p} - Q_{T_{N_t}}^{\overline{q},\overline{p}}\|. \end{aligned}$$

There are now three possible events either $N_t = 0$, $\nu \leq T_{N_t} - T_{N_t - 1} \leq K$ or $T_{N_t} - T_{N_t - 1} \geq K$. If $\nu \leq T_{N_t} - T_{N_t - 1} \leq K$ then we use (3.23), however if $T_{N_t} - T_{N_t - 1} \geq K$ then we use (3.22). By doing this for each refreshment we have

$$\|(Q_t^{q,p}, P_t^{q,p}) - (Q_t^{\overline{q},\overline{p}}, P_t^{\overline{q},\overline{p}})\| \leq C^{1 + N_t - M_t} \gamma^{M_t} \|Q_{T_1}^{q,p} - Q_{T_1}^{\overline{q},\overline{p}}\|.$$

Then by applying (3.22) once more we have

$$\|(Q_t^{q,p}, P_t^{q,p}) - (Q_t^{\overline{q},\overline{p}}, P_t^{\overline{q},\overline{p}})\| \leq C^{2 + N_t - M_t} \gamma^{M_t} \|(q,p) - (\overline{q},\overline{p})\|.$$

Dividing by $\|(q,p) - (\overline{q},\overline{p})\|$ and taking the limit as $\|(q,p) - (\overline{q},\overline{p})\| \to 0$ we have that

$$\|\nabla_{q,p}(Q_t^{q,p}, P_t^{q,p})\| \leq C^{2 + N_t - M_t} \gamma^{M_t}.$$

It remains to bound $\mathbb{E}[C^{N_t - M_t} \gamma^{M_t}]$. By conditioning on N_t we can use the moment generating function of a Binomial distribution to find

$$\mathbb{E}[C^{N_t - M_t} \gamma^{M_t} | N_t] = C^{N_t} (1 - (e^{-\lambda\nu} - e^{-\lambda K})(1 - \gamma C^{-1}))^{N_t}.$$

Now N_t is a Poisson process with rate λ so we have

$$\mathbb{E}[C^{N_t-M_t}\gamma^{M_t}] = \exp\left(\lambda\left(C(1-(e^{-\lambda\nu}-e^{-\lambda K})(1-\gamma C^{-1}))-1\right)\right).$$

This is decays exponentially provided

$$C(1-(e^{-\lambda\nu}-e^{-\lambda K_1})(1-\gamma C^{-1})) < 1.$$

□

3.C.3 Proofs of Example 3.45

Proof of Lemma 3.46. Note that for the ZZS

$$\begin{aligned}[\mathbf{\Phi}, Q]f(x,v) &= \sum_{i=1}^{d}\left\langle v, \nabla_x\left(\frac{\lambda_i(x,v)}{\lambda(x,v)}f(x,F_iv)\right)\right\rangle - \sum_{i=1}^{d}\frac{\lambda_i(x,v)}{\lambda(x,v)}\langle F_iv, \nabla_x(f(x,F_iv))\rangle \\ &= \sum_{i=1}^{d}\left\langle v, \nabla_x\left(\frac{\lambda_i(x,v)}{\lambda(x,v)}\right)\right\rangle f(x,F_iv) + 2\sum_{i=1}^{d}\left(\frac{\lambda_i(x,v)}{\lambda(x,v)}\right)v_i\partial_{x_i}f(x,F_iv).\end{aligned}$$

When we apply this with $f = \mathcal{P}_t g \circ \varphi_{\delta-s}$ and (x,v) replaced by $(x+vs,v)$ we have

$$\begin{aligned}[\mathbf{\Phi}, Q]f(x,v) &= \sum_{i=1}^{d}\left\langle v, \nabla_x\left(\frac{\lambda_i(x,v)}{\lambda(x,v)}\right)\right\rangle \mathcal{P}_t g(x+sv+(\delta-s)F_iv, F_iv) \\ &\quad + 2\sum_{i=1}^{d}\left(\frac{\lambda_i(x,v)}{\lambda(x,v)}\right)v_i\partial_{x_i}(\mathcal{P}_t g)(x+sv+(\delta-s)F_iv, F_iv) \\ &\leq \sum_{i=1}^{d}\left\langle v, \nabla_x\left(\frac{\lambda_i(x,v)}{\lambda(x,v)}\right)\right\rangle |\mathcal{P}_t g(x+sv+(\delta-s)F_iv, F_iv)| \\ &\quad + 2\|\nabla_x(\mathcal{P}_t g)(x+sv+(\delta-s)F_iv, F_iv)\|.\end{aligned}$$

Observe that

$$\partial_r\left(-\log\left(\phi(\exp(-r))\right)\right) = \frac{1}{1+e^{-r}}$$

then we have

$$\begin{aligned}\nabla_x\left(\frac{\lambda_i(x,v)}{\lambda(x,v)}\right) &= \left(\frac{\nabla_x\lambda_i(x,v)}{\lambda(x,v)}\right) - (\nabla_x\lambda(x,v))\left(\frac{\lambda_i(x,v)}{\lambda(x,v)^2}\right) \\ &= \left(\frac{v_i\nabla_x\partial_{x_i}\psi(x,v)}{(1+e^{-v_i\partial_{x_i}\psi(x)})\lambda(x,v)}\right) - \sum_{j=1}^{d}\left(\frac{v_j\nabla_x\partial_{x_j}\psi(x,v)\lambda_i(x,v)}{(1+e^{-v_i\partial_{x_i}\psi(x)})\lambda(x,v)^2}\right)\end{aligned}$$

Under our assumptions this is bounded. □

Proof of Theorem 3.47. This proof is omitted from this thesis, but can be found in the corresponding paper [16].

□

Lemma 3.68. *Let $\{(\overline{X}_{t_n}, \overline{X}_{t_n})\}_{n\in\mathbb{N}}$ denote the Euler Zig Zag algorithm in 1-d using Algorithm 7 or 8. Let $\overline{\lambda}(x,v;\delta) = \lambda(x,v)$ and $\lambda(x,v) = (\psi'(x)v)_+ + \gamma(x)$ for $\gamma : \mathbb{R} \to [0,\overline{\gamma}]$ with $\overline{\gamma} < \infty$. Assume that $\psi \in \mathcal{C}^2$ is such that (3.24) is satisfied. Let $\alpha \in (0,1)$, $\beta > 0$ be such that $\alpha < 2\beta$ and define*

$$\overline{G}_{\alpha,\beta}(x,v;\delta) = \begin{cases} \exp\left(\alpha\psi(x) + \beta\delta\psi'(x)v\right), & \text{if } v\psi'(x) \geq 0, \\ \exp\left(\alpha\psi(x) - \beta\delta\psi'(x)v\right), & \text{if } v\psi'(x) < 0. \end{cases} \tag{3.61}$$

Then there exists a compact set C and $\kappa \in (0,1)$ such that

$$\mathbb{E}_{x,v}\overline{G}_{\alpha,\beta}(\overline{X}_{t_n}, \overline{V}_{t_n}; \delta) \leq \kappa^n \overline{G}_{\alpha,\beta}(x,v;\delta) \qquad \text{for all } x \notin C.$$

Proof. Let $\{(\overline{X}^7_\delta, \overline{V}^7_\delta)\}$ ($\{(\overline{X}^8_\delta, \overline{V}^8_\delta)\}$ respectively) be given by Algorithm 7 (Algorithm 8 resp.). To simplify the notation in this proof suppress the δ dependence of $\overline{G}_{\alpha,\beta}$. Set $\beta_\pm = \beta$ if $v\psi'(x) \geq 0$ and $\beta_\pm = -\beta$ otherwise. Observe that

$$\begin{aligned}
&\mathbb{E}_{x,v}[\overline{G}_{\alpha,\beta}(\overline{X}^8_\delta, \overline{V}^8_\delta)] - \mathbb{E}_{x,v}[\overline{G}_{\alpha,\beta}(\overline{X}^7_\delta, \overline{V}^7_\delta)] = \\
&= \int_0^\delta \lambda(x,v) e^{-\lambda(x,v)s} \\
&\quad \times \left(e^{\alpha\psi(x+vs-(\delta-s)v) - \delta\beta_\pm v\psi'(x+sv-(\delta-s)v)} - e^{\alpha\psi(x+\delta v) - \delta\beta_\pm v\psi'(x+\delta v)} \right) ds \\
&= \int_0^\delta \lambda(x,v) \exp\left(-\lambda(x,v)s + \alpha\psi(x+\delta v) - \delta\beta_\pm v\psi'(x+\delta v)\right) \left(e^{I(x,v,s;\delta)} - 1 \right) \mathrm{d}s,
\end{aligned}$$

where

$$\begin{aligned}
I(x,v,s;\delta) \coloneqq\ & \alpha\psi(x+vs-(\delta-s)v) - \alpha\psi(x+\delta v) \\
& - \delta\beta_\pm v\psi'(x+sv-(\delta-s)v) + \delta\beta_\pm v\psi'(x+\delta v).
\end{aligned}$$

By Taylor's theorem we can find ξ_1, ξ_2

$$I(x,v,s;\delta) = \alpha 2(s-\delta)v\psi'(x+\delta v) + 2\alpha(\delta-s)^2\psi''(\xi_1) + 2\beta_\pm\delta(s-\delta)v\psi''(\xi_2).$$

By taking x sufficiently large we can ensure that the sign of $I(x,v,s;\delta)$ is equal to the sign of $-v\psi'(x)$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{x,v}[\overline{G}_{\alpha,\beta}(\overline{X}^8_\delta, \overline{V}^8_\delta)] \leq \mathbb{E}_{x,v}[\overline{G}_{\alpha,\beta}(\overline{X}^7_\delta, \overline{V}^7_\delta)] \quad \text{if } v\psi'(x) > 0, \\
&\mathbb{E}_{x,v}[\overline{G}_{\alpha,\beta}(\overline{X}^8_\delta, \overline{V}^8_\delta)] \geq \mathbb{E}_{x,v}[\overline{G}_{\alpha,\beta}(\overline{X}^7_\delta, \overline{V}^7_\delta)] \quad \text{if } v\psi'(x) < 0.
\end{aligned}$$

In the first case it is sufficient to consider Algorithm 7, while in the latter it is sufficient to consider Algorithm 8. We shall consider these two cases separately.

Case $v\psi'(x) > 0$**:** Note that it is sufficient to show that outside of a sufficiently large compact set

$$\frac{\mathbb{E}_{x,v}\overline{G}_{\alpha,\beta}(\overline{X}^7_\delta, \overline{V}^7_\delta)}{\overline{G}_{\alpha,\beta}(x,v)} < 1.$$

We can expand $\mathbb{E}_{x,v}\overline{G}_{\alpha,\beta}(\overline{X}^7_\delta, \overline{V}^7_\delta)$ as

$$\mathbb{E}_{x,v}\overline{G}_{\alpha,\beta}(\overline{X}^7_\delta, \overline{V}^7_\delta) = e^{-\delta\lambda(x,v)}\overline{G}_{\alpha,\beta}(x+v\delta, v) + (1-e^{-\delta\lambda(x,v)})\overline{G}_{\alpha,\beta}(x+v\delta, -v).$$

Using the definition of $\overline{G}_{\alpha,\beta}$ we can write

$$\begin{aligned}
\frac{\overline{G}_{\alpha,\beta}(x+v\delta, v)}{\overline{G}_{\alpha,\beta}(x,v)} &= \exp\left(\alpha(\psi(x+v\delta)-\psi(x)) + \beta\delta v(\psi'(x+v\delta)-\psi'(x))\right),\\
\frac{\overline{G}_{\alpha,\beta}(x+v\delta, -v)}{\overline{G}_{\alpha,\beta}(x,v)} &= \exp\left(\alpha(\psi(x+v\delta)-\psi(x)) - \beta\delta v(\psi'(x+v\delta)+\psi'(x))\right).
\end{aligned}$$

We can Taylor expand U to find some z_1, z_2, z_3 such that

$$\begin{aligned}
\frac{\overline{G}_{\alpha,\beta}(x+v\delta, v)}{\overline{G}_{\alpha,\beta}(x,v)} &= \exp\left(\alpha(\psi'(x)v\delta + \frac{1}{2}\psi''(z_1)\delta^2) + \beta\delta^2\psi''(z_2)\right),\\
\frac{\overline{G}_{\alpha,\beta}(x+v\delta, -v)}{\overline{G}_{\alpha,\beta}(x,v)} &= \exp\left(\alpha(\psi'(x)v\delta + \frac{1}{2}\psi''(z_1)\delta^2) - \beta\delta(2\psi'(x)v + \psi''(z_3)\delta)\right).
\end{aligned}$$

Thus we have

$$\begin{aligned}
\frac{\mathbb{E}_{x,v}\overline{G}_{\alpha,\beta}(\overline{X}^7_\delta, \overline{V}^7_\delta)}{\overline{G}_{\alpha,\beta}(x,v)} &= e^{-\delta\lambda(x,v)}\exp\left(\alpha(\psi'(x)v\delta + \frac{1}{2}\psi''(z_1)\delta^2) + \beta\delta^2\psi''(z_2)\right)\\
&\quad + (1-e^{-\delta\lambda(x,v)})\exp\left(\alpha(\psi'(x)v\delta + \frac{1}{2}\psi''(z_1)\delta^2) - \beta\delta(2\psi'(x)v + \psi''(z_3)\delta)\right).
\end{aligned}$$

Rearranging we can rewrite this as

$$\begin{aligned}
&\frac{\mathbb{E}_{x,v}\overline{G}_{\alpha,\beta}(\overline{X}^7_\delta, \overline{V}^7_\delta)}{\overline{G}_{\alpha,\beta}(x,v)} =\\
&= \exp\left(-\delta\lambda(x,v) + \alpha\psi'(x)v\delta + \frac{\alpha}{2}\psi''(z_1)\delta^2\right)\left(e^{\left(\beta\delta^2\psi''(z_2)\right)} - e^{\left(-\beta\delta(2\psi'(x)v+\psi''(z_3)\delta)\right)}\right) \qquad (3.62)\\
&\quad + \exp\left((\alpha-2\beta)\psi'(x)v\delta + \frac{1}{2}\alpha\psi''(z_1)\delta^2 - \beta\delta^2\psi''(z_3)\right). \qquad (3.63)
\end{aligned}$$

Recall that in this case $\lambda(x,v) \geq v\psi'(x) > 0$. Thus for the first term (3.62)

$$\exp\left(-\delta\lambda(x,v) + \alpha\psi'(x)v\delta + \frac{\alpha}{2}\psi''(z_1)\delta^2\right)\left(e^{\left(\beta\delta^2\psi''(z_2)\right)} - e^{\left(-\beta\delta(2\psi'(x)v+\psi''(z_3)\delta)\right)}\right) \leq$$

$$\leq \exp\left(-\delta\lambda(x,v) + \alpha\psi'(x)v\delta + \frac{\alpha}{2}\psi''(z_1)\delta^2\right)e^{\left(\beta\delta^2\psi''(z_2)\right)}$$
$$\leq \exp\left(-(1-\alpha)\delta v\psi'(x) + \frac{\alpha}{2}\psi''(z_1)\delta^2\right)e^{\left(\beta\delta^2\psi''(z_2)\right)}.$$

Now choose $0 < \alpha < \min\{1, 2\beta\}$ and recall that by assumption ψ' diverges to infinity faster than ψ''. It follows that, outside of a large enough compact set, both (3.62) and (3.63) can be made arbitrarily small.

Case $v\psi'(x) < 0$**:** In this case $\lambda(x,v) = \gamma(x)$. We expand $\mathbb{E}_{x,v}\overline{G}_{\alpha,\beta}(\overline{X}^8_\delta, \overline{V}^8_\delta)$ as

$$\frac{\mathbb{E}_{x,v}\overline{G}_{\alpha,\beta}(\overline{X}^8_\delta, \overline{V}^8_\delta)}{\overline{G}_{\alpha,\beta}(x,v)} = e^{-\delta\lambda(x,v)}\frac{\overline{G}_{\alpha,\beta}(x+v\delta, v)}{\overline{G}_{\alpha,\beta}(x,v)} + \int_0^\delta \lambda(x,v)e^{-s\lambda(x,v)}\frac{\overline{G}_{\alpha,\beta}(x+v(2s-\delta), -v)}{\overline{G}_{\alpha,\beta}(x,v)}\mathrm{d}s.$$

Similarly to above, we can use Taylor's theorem to find z_1, z_2, z_3, z_4 with

$$\begin{aligned}\frac{\mathbb{E}_{x,v}\overline{G}_{\alpha,\beta}(\overline{X}^8_\delta, \overline{V}^8_\delta)}{\overline{G}_{\alpha,\beta}(x,v)} &= e^{-\delta\lambda(x,v)}\exp\left(\alpha(\psi'(x)v\delta + \frac{1}{2}\psi''(z_1)\delta^2) - \beta\delta^2\psi''(z_2)\right)\\ &+ \int_0^\delta \lambda(x,v)e^{-s\lambda(x,v)}\exp\left(\alpha(\psi'(x)v(2s-\delta) + \frac{1}{2}\psi''(z_3)(2s-\delta)^2)\right.\\ &\left. + \beta\delta(2\psi'(x)v + \psi''(z_4)(2s-\delta))\right)\mathrm{d}s.\end{aligned} \tag{3.64}$$

Taking advantage of $-\delta \leq 2s - \delta \leq \delta$ we obtain the bound

$$\exp\left(\alpha(\psi'(x)v(2s-\delta) + \beta\delta(2\psi'(x)v)\right) \leq \exp\left((-\psi'(x)v)(\alpha - 2\beta)\delta\right).$$

Using this bound together with the assumption that $-v\psi'(x)$ diverges to $+\infty$ faster than ψ'', we obtain that for $0 < \alpha < 2\beta$ the right hand side of (3.64) can be made arbitrarily small for sufficiently large values of x.

Combining the two cases above we obtain the statement of the lemma. □

Lemma 3.69. *Assume that $\psi \in \mathcal{C}^2$ satisfies (3.24). Let $\overline{G}_{\alpha,\beta}(x,v;\delta)$ be given by (3.61) and $\overline{G}_{\alpha,\epsilon}(x,v)$ be given by (3.25). Then for any $0 < \alpha_1 < \overline{\alpha} < \alpha_2 < 1$ there exist positive constants $C, C' > 0$ with*

$$\overline{G}_{\alpha_1,\epsilon}(x,v) \leq C'e^{\overline{\alpha}\psi(x)} \leq C\overline{G}_{\alpha_2,\beta}(x,v;\delta). \tag{3.65}$$

Proof of Lemma 3.69. Let us first consider $\overline{G}_{\alpha_1,\epsilon}(x,v)$, since $|\phi_\epsilon(s)| \leq \epsilon|s|/2$ we have

$$\overline{G}_{\alpha_1,\epsilon}(x,v) \leq \exp\left(\alpha_1\psi(x) + \frac{\epsilon}{2}|\psi'(x)|\right).$$

By (3.24) there exists $R > 0$ such that for any $|x| > R$ we have $|\psi'(x)| \leq 2\epsilon^{-1}(\overline{\alpha} - \alpha_1)\psi(x)$. Therefore for $|x| > R$ we have

$$\overline{G}_{\alpha_1,\epsilon}(x, v) \leq \exp\left(\overline{\alpha}\psi(x)\right).$$

Setting $C' = \exp(\sup_{|x|\leq R}|\psi'(x)|)$ we have the left hand side of (3.65).

Similarly, we have

$$\overline{G}_{\alpha_2,\beta}(x, v; \delta) \geq \exp\left(\alpha_2\psi(x) - \beta\delta_0|\psi'(x)|\right).$$

Using (3.24) for x sufficiently large we have that $\beta\delta_0|\psi'(x)| \leq (\alpha_2 - \overline{\alpha})\psi(x)$ and hence the right hand side of (3.65) follows. □

Chapter 4

Splitting schemes for second order approximations of piecewise deterministic Markov processes

4.1 Introduction

Piecewise deterministic Markov processes (PDMPs) are non-diffusive Markov processes combining a deterministic motion and random jumps. They appear in a wide range of modelling problems [39, 94, 98] and, over the last decade, have gained considerable interest as Markov Chain Monte Carlo (MCMC) methods [125, 110, 21, 32, 64, 152]. Similarly to Chapter 3, this chapter addresses the question of the simulation of a PDMP with generator (2.31). The classical method is to use a Poisson thinning procedure [95, 93] to sample the jump times, and then to solve the ODE exactly if possible, or otherwise by a standard numerical scheme. Similar to rejection sampling which requires a good reference measure, an efficient Poisson thinning algorithm requires the knowledge of good bounds for the jump rate λ along the trajectory of the ODE. In this work, we focus on the case in which such bounds are not available, or are so crude that thinning would not be numerically efficient. In that case, the random event times have to be approximated even if the ODE can be solved exactly. This question has recently been addressed in [16, 121, 41] with three different schemes. Here, rather than designing an ad hoc numerical scheme, we work in the general framework of splitting schemes, which are widely used for e.g. Hamiltonian or underdamped Langevin processes [89, 91, 112]. One of the main interests is that, by design, such schemes have a numerical error which is of order 2 in

the step-size, without the need of an approximation of the jump rate along the ODE. Moreover, it is a flexible framework and thus such schemes can be easily combined with multi-time-step or factorization methods [90] or integrated in hybrid PDMP/diffusion schemes [113, 111]. Note that, by using a numerical approximation, we lose one of the interests of PDMPs for MCMC purposes, which is the exact simulation by thinning, while in our case the invariant measure of the scheme will have a deterministic bias with respect to the true target measure. However, we still benefit from the good long-time convergence properties of the ballistic non-reversible process and, contrary to Hamiltonian-based dynamics, it is still possible to factorize the target measure and define efficient schemes in terms of number of computations of forces (see [113, 111] and Section 4.5.3). We shall also show how the correct stationary distribution can be recovered by means of a non-reversible Metropolis-Hastings acceptance/rejection step (see Section 4.1.2). Moreover, for classical velocity jump processes used in MCMC, since the norm of the velocity is constant (between possible refreshments which are independent of the potential), these schemes are numerically stable (see the numerical experiments in Section 4.5 where the step-size of PDMP schemes can be taken larger than for the classical ULA), even for non-globally Lipschitz potentials.

The core idea of splitting schemes is first to split the generator in several parts such that a process associated to each part can be simulated exactly. For instance, when the ODE can be solved exactly, one can write $\mathcal{L} = \mathcal{L}_D + \mathcal{L}_J$ with

$$\mathcal{L}_D f(z) = \langle \Phi(z), \nabla_z f(z) \rangle,$$
$$\mathcal{L}_J f(z) = \lambda(z) \int_E (f(y) - f(z)) Q(z, \mathrm{d}y)\,,$$

in which case the process associated to $\mathcal{L}_D$ is simply the solution of the ODE, hence D stands for drift, while the process associated to $\mathcal{L}_J$ is a continuous-time Markov chain, for which the jump rate is constant between two jumps (so that the jump times are simple exponential random variables), hence J stands for jumps. Then, one approximates the semigroup of the true process by a Strang splitting

$$P_\delta = e^{\delta(\mathcal{L}_D + \mathcal{L}_J)} \approx e^{\frac{\delta}{2}\mathcal{L}_D} e^{\delta \mathcal{L}_J} e^{\frac{\delta}{2}\mathcal{L}_D} \tag{4.1}$$

for a small step size $\delta > 0$. Therefore, over one time step the approximation follows $\mathcal{L}_D$ for time $\delta/2$, then $\mathcal{L}_J$ for time δ and finally $\mathcal{L}_D$ again for time $\delta/2$. Given a step size δ, now we illustrate how the $(n+1)$-th iteration works. Starting at time $t_n = n\delta$ at state $\overline{Z}_{t_n}$ the process first moves deterministically for a half step:

$$\overline{Z}_{t_n + \delta/2} = \varphi_{\delta/2}(\overline{Z}_{t_n}).$$

Then we simulate the pure jump part of the process: we generate an event time $\tau_1 \sim \mathrm{Exp}(\lambda(\overline{Z}_{t_n+\delta/2}))$ and, if $\tau_1 < \delta$, we set $\overline{Z}_{t_n+\delta/2} \sim Q(\overline{Z}_{t_n+\delta/2}, \cdot)$. Then we repeat this step as long as $\sum_i \tau_i < \delta$, though, since we are interested in second order schemes, it is enough to limit ourselves to two jumps per time step. Note that the

rate is updated after every jump and is constant between jumps. We conclude the iteration by a final half step of deterministic motion:

$$\overline{Z}_{t_{n+1}} = \varphi_{\delta/2}(\overline{Z}_{t_n+\delta/2}).$$

We refer to this scheme as the splitting scheme **DJD**, where consistently with above **D** stands for drift and **J** for jumps. When the ODE cannot be solved exactly, any second-order numerical scheme can be used instead of φ_t. Moreover, in some cases (typically for the Hamiltonian dynamics) the generator $\mathcal{L}_D$ can be further divided in several ODEs. Similarly, for computational purpose, it can be interesting in some cases to split the jump part $\mathcal{L}_J$ in several operators. It is also possible to keep in $\mathcal{L}_D$ a combination of ODE and jump, simulated e.g. by thinning, while some parts of the jump are treated separately in $\mathcal{L}_J$ (it could make sense for instance in the context of [113]). When there are more than two parts in the splitting of $\mathcal{L}$, a scheme is obtained by starting from (4.1) and using e.g. $e^{\delta\mathcal{L}_J} \approx e^{\frac{\delta}{2}\mathcal{L}_A} e^{\delta\mathcal{L}_B} e^{\frac{\delta}{2}\mathcal{L}_A}$ if $\mathcal{L}_J = \mathcal{L}_A + \mathcal{L}_B$, etc.

Such splitting schemes can be used to simulate any PDMP. For some modelling problems, it can be interesting to have estimates on the trajectorial error between the approximated process and the two process, for instance when dynamical properties (like mean squared displacement or transition rates) are of interest. However, in this work, we have mainly in mind the PDMPs which are used for MCMC methods, in particular our recurrent examples will be the Zig-Zag sampler (ZZS) [21, 23] and the Bouncy Particle sampler (BPS) [125, 110, 32]. As a consequence, we will not discuss trajectorial errors but rather focus on what is relevant for MCMC purposes, namely the long-time convergence of the Markov chain (which should scale properly as the step size vanishes) and the numerical bias on the invariant measure and on empirical averages of the chain.

Organisation of the chapter

The chapter is organized as follows. We conclude this introduction by presenting the algorithms we focus on in this chapter. In Section 4.1.1 we discuss our two main examples and their approximation with splitting schemes. In Sections 4.1.2 and 4.1.3 we discuss respectively how we can Metropolis-adjust our schemes in a non-reversible fashion and how we can modify the algorithms to do subsampling. We conclude our introduction with Section 4.1.4, where we describe how boundaries can be treated with our splitting schemes. Section 4.2 is devoted to the analysis of the weak error for the finite-time empirical averages of the scheme **DJD**. The main result, Theorem 4.10, states that for this scheme the *weak error is of order* 2 in the step-size. The geometric ergodicity of splitting schemes based on our main examples is established in Section 4.3, with a consistent dependency of the estimates on the step-size. In Section 4.4, we provide a formal expansion (in terms of the step-size) of the invariant measure of the scheme based on the so-called Bouncy Particle Sampler depending on the choice of the splitting, in the spirit of [89], with a particular focus in Section 4.4.2 on three one-dimensional examples where everything can be made explicit. Numerical experiments are provided in Section 4.5. Finally, technical proofs are gathered in an Appendix.

Comparison to related works

The work in this chapter can be seen as a continuation of the work that two of the authors started with their coauthors in [16], in which a general framework to approximate PDMPs is introduced and studied. In this previous work, the focus is not a specific scheme and thus the results are mostly general and not tailored for particular processes or schemes, though the ZZS and BPS are considered as recurrent examples. In particular, the schemes introduced in [16] leave considerable freedom to the user in the choice of some crucial components of the algorithm, namely an approximation of the switching rates or a numerical integrator in place of the exact flow map. On the other hand, in this chapter we follow the philosophy of splitting schemes to describe a simple recipe to approximate PDMPs. Note that splitting schemes are not considered in [16]. The main advantage of splitting schemes is the second order of accuracy with one gradient evaluation per iteration, whereas second order algorithms considered in [16] relied on approximations of second order of the switching rates, which can be usually obtained with the expensive computation of the Hessian of the negative log-target. Moreover, in this work we describe how to remove the bias introduced by our approximation with a non-reversible Metropolis-Hastings step. Two other works ([121] and [41]) focus on approximate simulation of the Zig-Zag sampler, which is one of our two main examples. In [121] the authors suggest to approximate event times by using numerical approximations of the integral of the rates along the dynamics (3.4), as well as a root finding algorithm. In [41], the authors suggest using a numerical optimisation algorithm at each iteration to obtain a suitable bound that enables the use of Poisson thinning. The first difference is that we mainly consider our approximations as discrete time Markov chains, whereas the processes of [121] and [41] are interpreted in continuous time, although neither resulting process is a Markov process due to the nature of the numerical algorithms that are used. Naturally, one could interpret our algorithms as continuous time processes, which again would not be Markov processes. Secondly, without assuming any properties that we do not verify, we derive theoretical justifications of our proposed algorithms, such as bounds on the weak error and existence, uniqueness, and geometric convergence to a stationary distribution under simple conditions. Moreover, we introduce Metropolis adjusted algorithms to eliminate the error introduced by the numerical approximations, while this aspect is not studied in previous works and thus we introduce the first exact PDMP based samplers that can be simulated with only access to the gradient of the negative logarithm of the target distribution.

4.1.1 Main examples

Let us now introduce two examples from the computational statistics literature. In this setting we have a target probability measure with density $\pi(x) \propto \exp(-\psi(x))$ for $x \in \mathbb{R}^d$.

Example 4.1 (Zig-Zag sampler [23])**.** *As we have seen in the previous chapter, simulating the event times of ZZS is in general a very challenging problem. We can apply*

Algorithm 11: Splitting scheme **DBD** for ZZS

Input : Number of iterations N, initial condition (x, v), step size δ.
Output: Chain $(\overline{X}_{t_n}, \overline{V}_{t_n})_{n=0}^{N}$.
Set $n = 0$, $(\overline{X}_0, \overline{V}_0) = (x, v)$;
while $n < N$ **do**
  Set $\overline{X}_{t_{n+1}} = \overline{X}_{t_n} + \frac{\delta}{2}\overline{V}_{t_n}$;
  Set $\overline{V}_{t_{n+1}} = \overline{V}_{t_n}$;
  for $i = 1 \ldots, d$ **do**
    With probability $(1 - \exp(\delta\lambda_i(\overline{X}_{t_{n+1}}, \overline{V}_{t_{n+1}})))$ set $\overline{V}_{t_{n+1}} = R_i\overline{V}_{t_{n+1}}$;
  end
  Set $\overline{X}_{t_{n+1}} = \overline{X}_{t_{n+1}} + \frac{\delta}{2}\overline{V}_{t_{n+1}}$;
  Set $n = n + 1$;
end

to ZZS the splitting scheme above as follows. Assume the process has canonical rates, *i.e.* $\gamma_i = 0$ *for all* i. *Then we can split the generator* (2.32) *as*

$$\mathcal{L}_D f(x, v) = \langle v, \nabla_x f(x)\rangle,$$
$$\mathcal{L}_B f(x, v) = \sum_{i=1}^{d} \lambda_i(x, v)[f(x, R_i v) - f(x, v)].$$

Here we define the scheme ***DBD****, where* ***B*** *stands for* bounces. *Given* $(\overline{X}_{t_n}, \overline{V}_{t_n})$, *we start by a half step of deterministic motion:*

$$\overline{X}_{t_n+\frac{\delta}{2}} = \overline{X}_{t_n} + \frac{\delta}{2}\overline{V}_{t_n}.$$

Then for $i = 1, \ldots, d$ *we draw* $\tau_i \overset{iid}{\sim} \mathrm{Exp}(\lambda_i(\overline{X}_{t_n+\delta/2}, \overline{V}_{t_n}))$, *which are homogeneous exponential random variables. Then let* $\tau_{(1)} = \min \tau_i$ *and set*

$$\overline{V}_{t_{n+1}} = \begin{cases} \overline{V}_{t_n} & \text{if } \tau_{(1)} > \delta \\ R_I\overline{V}_{t_n} & \text{if } \tau_{(1)} \le \delta \end{cases}$$

where $R_I = \prod_{i\in I} R_i$ *and* I *is the set of indices* i *for which* $\tau_i \le \delta$. *Alternatively to have a second order scheme it is sufficient to flip only the two components with the smallest switching time* τ_i, *given that it is before time* δ. *Observe that for canonical rates flipping the sign of a component does not affect the other switching rates, and thus it is not possible to have two flips in the same component when* $\gamma_i = 0$. *Finally, set*

$$\overline{X}_{t_{n+1}} = \overline{X}_{t_n+\frac{\delta}{2}} + \frac{\delta}{2}\overline{V}_{t_{n+1}},$$

which concludes the iteration. The procedure is described in pseudo code form in Algorithm 11. An interesting feature of the algorithm is that the jump part of the

Algorithm 12: Splitting scheme **RDBDR** for BPS

Input : Number of iterations N, initial condition (x, v), step size δ.
Output: Chain $(\overline{X}_{t_n}, \overline{V}_{t_n})_{n=0}^{N}$.
Set $n = 0$, $(\overline{X}_0, \overline{V}_0) = (x, v)$;
while $n < N$ **do**
  Set $\overline{V}_{t_{n+1}} = \overline{V}_{t_n}$;
  With probability $(1 - \exp(-\lambda_r \frac{\delta}{2}))$ draw $\overline{V}_{t_{n+1}} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$;
  Set $\overline{X}_{t_{n+1}} = \overline{X}_{t_n} + \frac{\delta}{2}\overline{V}_{t_{n+1}}$;
  With probability $(1 - \exp(\delta\lambda_1(\overline{X}_{t_{n+1}}, \overline{V}_{t_{n+1}})))$ set $\overline{V}_{t_{n+1}} = R(\overline{X}_{t_{n+1}})\overline{V}_{t_{n+1}}$;
  Set $\overline{X}_{t_{n+1}} = \overline{X}_{t_{n+1}} + \frac{\delta}{2}\overline{V}_{t_{n+1}}$;
  With probability $(1 - \exp(-\lambda_r \frac{\delta}{2}))$ set $\overline{V}_{t_{n+1}} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$;
  Set $n = n + 1$;
end

chain can be computed in parallel*, since in that stage a velocity flip in one component does not affect the other components of the process.*

Example 4.2 (Bouncy Particle Sampler [32])**.** *Recall the BPS of Example 2.12. In this case we split the generator in three parts:* $\mathcal{L}f(x,v) = \mathcal{L}_D f(x,v) + \mathcal{L}_B f(x,v) + \mathcal{L}_R f(x,v)$ *where*

$$\begin{aligned}
\mathcal{L}_D f(x,v) &= \langle v, \nabla_x f(x)\rangle, \\
\mathcal{L}_B f(x,v) &= \lambda_1(x,v)[f(x, R(x)v) - f(x,v)], \\
\mathcal{L}_R f(x,v) &= \lambda_2 \int \big(f(x,w) - f(x,v)\big)\nu(\mathrm{d}y),
\end{aligned}$$

We then define the scheme ***RDBDR****, where* ***R*** *stands for* refreshments*. Starting at time* $t_n = n\delta$ *at state* $(\overline{X}_{t_n}, \overline{V}_{t_n})$ *we begin by drawing* $\tau_1 \sim \mathrm{Exp}(\lambda_r)$ *and setting*

$$\tilde{V}_{t_n + \frac{\delta}{2}} = \begin{cases} \overline{V}_{t_n} & \text{if } \tau_1 > \delta/2 \\ W_1 & \text{if } \tau_1 \le \delta/2 \end{cases}$$

for $W_1 \sim \nu$*. Then the process evolves deterministically for time* $\delta/2$*:*

$$\overline{X}_{t_n + \frac{\delta}{2}} = \overline{X}_{t_n} + \frac{\delta}{2}\tilde{V}_{t_n + \frac{\delta}{2}}.$$

At this point, we check if a reflection takes place by drawing

$$\tau_2 \sim \mathrm{Exp}(\lambda_1(\overline{X}_{t_n + \frac{\delta}{2}}, \tilde{V}_{t_n + \frac{\delta}{2}}))$$

and set

$$\overline{V}_{t_n + \frac{\delta}{2}} = \begin{cases} \tilde{V}_{t_n + \frac{\delta}{2}} & \text{if } \tau_2 > \delta \\ R(\overline{X}_{t_n + \frac{\delta}{2}})\tilde{V}_{t_n + \frac{\delta}{2}} & \text{if } \tau_2 \le \delta \end{cases}$$

Importantly, $\lambda_1(\overline{X}_{t_n+\frac{\delta}{2}}, \overline{V}_{t_n+\frac{\delta}{2}}) = 0$ *if a reflection takes place and thus at most one reflection can happen. This is a consequence of the fact that* $\langle R(x)v, \nabla\psi(x)\rangle = -\langle v, \nabla\psi(x)\rangle$ *by definition of the reflection operator. After this we set*

$$\overline{X}_{t_{n+1}} = \overline{X}_{t_n+\frac{\delta}{2}} + \frac{\delta}{2},$$

and finally conclude the iteration drawing $\tau_3 \sim \text{Exp}(\lambda_r)$ *and letting*

$$\tilde{V}_{t_{n+1}} = \begin{cases} \overline{V}_{t_n+\frac{\delta}{2}} & \text{if } \tau_3 > \delta/2 \\ W_2 & \text{if } \tau_3 \leq \delta/2 \end{cases}$$

where $W_2 \sim \nu$*. The pseudo code can be found in Algorithm 12.*

4.1.2 Metropolis adjusted algorithms

Naturally, the use of splitting schemes to approximate a PDMP introduces a discretisation error. In the context of Bayesian statistics, this means that a bias term is introduced in the estimators for statistics of interest. In this section we discuss how to eliminate this bias with the addition of a Metropolis-Hastings (MH) acceptance-rejection step. We shall define *skew-reversible* MH algorithms based on our splitting schemes of ZZS and BPS relying on the general framework we introduced in Chapter 2 (see Section 2.2.4.1 for the details). In both cases, the splitting schemes give a proposal mechanism which fits in one of our examples in Section 2.2.4.1, in particular giving an acceptance probability of the form (2.19), where fortunately the determinant term equals 1. This is due to the fact that all proposals are formed by composing volume preserving transformations, and hence the determinant of the Jacobian equals one in absolute value. We observe that the ideas below can be applied to other kinetic PDMPs used in MCMC in a similar fashion.

4.1.2.1 Skew-reversible Metropolis adjusted ZZS

Consider the splitting **DBD** of ZZS with initial condition (x, v). Let $\delta > 0$ be the step size and $x_{1/2}(x, v) = x + v\delta/2$ (we will drop the dependence on (x, v) when clear). As explained in Example 4.1, after one iteration the algorithm has state

$$(\tilde{X}, \tilde{V}) = (x_{1/2} + \frac{\delta}{2} R_I v, R_I v)$$

with corresponding probability

$$\exp\left(-\delta \sum_{i \notin I} \lambda_i(x_{1/2}, v)\right) \prod_{i \in I} (1 - \exp(-\delta\lambda_i(x_{1/2}, v)). \tag{4.2}$$

We now want to accept or reject the proposed state with suitable probability to ensure μ-stationarity. Note the classical MH scheme of Section 2.2.1 is not directly

Algorithm 13: Skew-reversible Metropolis adjusted ZZS

Input : Number of iterations N, initial condition (x, v), step size δ.
Output: Chain $(\overline{X}_{t_n}, \overline{V}_{t_n})_{n=0}^N$.
Set $n = 0$, $(\overline{X}_0, \overline{V}_0) = (x, v)$;
while $n < N$ **do**
  Set $\overline{X}_{t_n+\delta/2} = \overline{X}_{t_n} + \frac{\delta}{2}\overline{V}_{t_n}$;
  Set $\tilde{V} = \overline{V}_{t_n}$;
  for $i = 1 \dots, d$ **do**
    With probability $(1 - \exp(\delta\lambda_i(\overline{X}_{t_n+\delta/2}, \tilde{V})))$ set $\tilde{V} = R_i\tilde{V}$;
  end
  Set $\tilde{X} = \overline{X}_{t_n+\delta/2} + \frac{\delta}{2}\tilde{V}$;
  Set $(\overline{X}_{t_{n+1}}, \overline{V}_{t_{n+1}}) = (\tilde{X}, \tilde{V})$ with probability

$$1 \wedge \frac{\pi(\tilde{X})}{\pi(\overline{X}_{t_n})} \exp\left(\delta \sum_{j=1}^{d} \left(\lambda_j(\overline{X}_{t_n+\delta/2}, \overline{V}_{t_n}) - \lambda_j(\overline{X}_{t_n+\delta/2}, -\tilde{V})\right)\right)$$

  else set $(\overline{X}_{t_{n+1}}, \overline{V}_{t_{n+1}}) = (\overline{X}_{t_n}, -\overline{V}_{t_n})$;
  Set $n = n + 1$;
end

applicable, as typically there is a 0 probability that the process goes from $(\tilde{X}, \tilde{V})$ to (x, v). Hence we use the non-reversible MH acceptance probability (2.19). For this we need to compute the probability of going from $(\tilde{X}, -\tilde{V})$ to $(x, -v)$ according the transition kernel of **DBD**. This can only be achieved by following the same path of $(x, v) \to (\tilde{X}, \tilde{V})$ with reversed time. Hence the sign of the velocity of the components in α needs to be flipped. Noticing that $x_{1/2}(x, v) = x_{1/2}(\tilde{X}, -\tilde{V})$, we find that the probability of this path is

$$\exp\left(-\delta \sum_{i \notin I} \lambda_i(x_{1/2}, -\tilde{V})\right) \prod_{i \in I} (1 - \exp(-\delta\lambda_i(x_{1/2}, -\tilde{V})),$$

where I is the same set of indices of (4.2). Observe that for $i \in I$ it holds that $\tilde{V}_i = -v_i$ and thus $\lambda_i(x_{1/2}, v) = \lambda_i(x_{1/2}, -\tilde{V})$, while for $i \notin I$ we have $\tilde{V}_i = v_i$ and hence $\lambda_i(x_{1/2}, v) - \lambda_i(x_{1/2}, -\tilde{V}) = v_i\partial\psi(x_{1/2})$. Therefore the acceptance probability (2.19) simplifies to

$$1 \wedge \frac{\pi(\tilde{X}) \exp(-\delta \sum_{i \notin I} \lambda_i(x_{1/2}, -\tilde{V}))}{\pi(x) \exp(-\delta \sum_{i \notin I} \lambda_i(x_{1/2}, v))} = 1 \wedge \exp\left(\psi(x) - \psi(\tilde{X}) + \delta \sum_{i \notin I} v_i \partial_i \psi(x_{1/2})\right). \tag{4.3}$$

In case of rejection, the state is set to $(x, -v)$. The procedure is written as pseudo-code in Algorithm 13.

Algorithm 14: Non-reversible Metropolis adjusted BPS

Input : Number of iterations N, initial condition (x, v), step size δ.
Output: Chain $(\overline{X}_{t_n}, \overline{V}_{t_n})_{n=0}^{N}$.
Set $n = 0$, $(\overline{X}_0, \overline{V}_0) = (x, v)$;
while $n < N$ **do**
  Set $\overline{V}_{t_n+\delta/2} = \overline{V}_{t_n}$;
  With probability $(1 - \exp(-\lambda_r \delta/2))$ draw $\overline{V}_{t_n+\delta/2} \sim \text{Unif}(\mathbb{S}^{d-1})$;
  Set $\overline{X}_{t_n+\delta/2} = \overline{X}_{t_n} + \frac{\delta}{2}\overline{V}_{t_n+\delta/2}$;
  Set $\tilde{V} = \overline{V}_{t_n+\delta/2}$;
  With probability $(1 - \exp(\delta\lambda_1(\overline{X}_{t_n+\delta/2}, \tilde{V})))$ set $\tilde{V} = R(\overline{X}_{t_n+\delta/2})\tilde{V}$;
  Set $\tilde{X} = \tilde{X} + \frac{\delta}{2}\tilde{V}$;
  Set $(\overline{X}_{t_{n+1}}, \overline{V}_{t_{n+1}}) = (\tilde{X}, \tilde{V})$ with probability

$$1 \wedge \frac{\pi(\tilde{X}) \times \exp(-\delta\lambda(\overline{X}_{t_n+\delta/2}, -\tilde{V}))}{\pi(\overline{X}_{t_n}) \times \exp(-\delta\lambda(\overline{X}_{t_n+\delta/2}, \overline{V}_{t_n}))}$$

  else set $(\overline{X}_{t_{n+1}}, \overline{V}_{t_{n+1}}) = (\overline{X}_{t_n}, -\overline{V}_{t_n+\delta/2})$;
  With probability $(1 - \exp(-\lambda_r \delta/2))$ set $\overline{V}_{t_{n+1}} \sim \text{Unif}(\mathbb{S}^{d-1})$;
  Set $n = n + 1$;
end

Remark 4.3. Assuming $\psi \in \mathcal{C}^2$, the terms $\psi(x)$ and $\psi(\tilde{X}) = \psi(x_{1/2} + R_I v\, \delta/2)$ can be expanded by Taylor's theorem around $x_{1/2}$. This gives that the acceptance probability (4.3) is

$$1 \wedge \exp\left(\frac{\delta^2}{8}\left(\langle v, \nabla^2\psi(\overline{x}_1)v\rangle - \langle R_I v, \nabla^2\psi(\overline{x}_2)R_I v\rangle\right)\right), \tag{4.4}$$

where $\overline{x}_1 \in (x, x_{1/2})$ and $\overline{x}_2 \in (x_{1/2}, x_{1/2} + R_I v\delta/2)$. This result shows that the probability of rejecting the proposed state is of order δ^2 and gives first evidence that the splitting scheme introduces an error of second order in the invariant measure. Moreover, it is clear from (4.4) that the acceptance probability equals 1 for instance when $\nabla^2\psi$ is a constant diagonal matrix, as is the case in a d-dimensional independent Gaussian vector. In this setting the splitting scheme **DBD** has the correct stationary distribution μ and does not need a Metropolis correction. Finally, observe that *the rejection probability is of order* δ^3 if $\nabla^2\psi$ is diagonal but not constant.

4.1.2.2 Non-reversible Metropolis adjusted BPS

Here we consider scheme **RDBDR** of BPS. The refreshment does not alter the stationary distribution of the process, thus we focus first on the **DBD** part. Denote $x_{1/2}(x, v) = x + \delta v/2$. According to **DBD**, the process moves from an initial condition

(x, v) to

$$(\tilde{X}, \tilde{V}) = \begin{cases} (x_{1/2} + \frac{\delta}{2} R(x_{1/2})v, R(x_{1/2})v) & \text{with probability } 1 - \exp(-\delta\lambda(x_{1/2}, v)), \\ (x + \delta v, v) & \text{with probability } \exp(-\delta\lambda(x_{1/2}, v)). \end{cases} \tag{4.5}$$

Observe that for both states in (4.5) it holds $x_{1/2}(x, v) = x_{1/2}(\tilde{X}, -\tilde{V}) = x_{1/2}$. We now focus on computing the acceptance probability (2.19) in the two cases in (4.5).

Consider first the case in which a reflection took place, which corresponds to the first line of (4.5). Then we need to compute the probability that the process goes from $(\tilde{X}, -\tilde{V})$ back to $(x, -v)$ using scheme **DBD**, which is equal to the probability that the process has a reflection at $x_{1/2}$. By definition of the reflection rate λ, it holds that $\lambda(x_{1/2}, v) = \lambda(x_{1/2}, -R(x_{1/2})v)$. Therefore in this case the probability that the process goes from $(\tilde{X}, -\tilde{V})$ to $(x, -v)$ is the same as that of going from (x, v) to $(\tilde{X}, \tilde{V})$ and thus the acceptance probability (2.19) is

$$1 \wedge \frac{\pi(x_{1/2} + \frac{\delta}{2} R(x_{1/2})v)}{\pi(x)} = 1 \wedge \exp\Big(\psi(x) - \psi\left(x_{1/2} + \delta R(x_{1/2})v/2\right)\Big). \tag{4.6}$$

Observe that moves that decrease ψ are accepted with probability 1.

Consider now the second case in (4.5). The probability that the process goes from $(x + \delta v, -v)$ to $(x, -v)$ is $\exp(-\delta\lambda(x_{1/2}, -v))$, while the probability of going from (x, v) to $(x + \delta v, v)$ is $\exp(-\delta\lambda(x_{1/2}, v))$. Observing that $\lambda(x_{1/2}, v) - \lambda(x_{1/2}, -v) = \langle v, \nabla\psi(x_{1/2})\rangle$ we find that in this case the MH acceptance probability is

$$1 \wedge \frac{\pi(x + \delta v)\, \exp(-\delta\lambda(x_{1/2}, -v))}{\pi(x)\, \exp(-\delta\lambda(x_{1/2}, v))} = 1 \wedge \exp\Big(\psi(x) - \psi(x + v\delta) + \delta\langle v, \nabla\psi(x_{1/2})\rangle\Big). \tag{4.7}$$

Hence we have shown that the acceptance probability can in general be written as

$$1 \wedge \frac{\pi(\tilde{X})\, \exp(-\delta\lambda(x_{1/2}, -\tilde{V}))}{\pi(x)\, \exp(-\delta\lambda(x_{1/2}, v))} = 1 \wedge \exp\Big(\psi(x) - \psi(\tilde{X}) + \delta(\lambda(x_{1/2}, v) - \lambda(x_{1/2}, -\tilde{V}))\Big).$$

In case of rejection, the state is set to $(x, -v)$. Two refreshments half-steps, to be executed before and after the scheme **DBD**, are necessary to ensure irreducibility of the Markov chain.

The described procedure is written in pseudo code form in Algorithm 14.

Remark 4.4. Let us Taylor expand the acceptance probabilities similarly to Note 4.3. Indeed for (4.6) we expand both terms around $x_{1/2}$ and for $\overline{x}_1 \in (x, x_{1/2})$ and $\overline{x}_2 \in (x_{1/2}, x + v\delta)$ we obtain

$$\psi(x) - \psi\left(x_{1/2} + \frac{\delta}{2} R(x_{1/2})v\right) = -\frac{\delta}{2}\left(\langle v, \nabla\psi(x_{1/2})\rangle + \langle R(x_{1/2})v, \nabla\psi(x_{1/2})\rangle\right)$$

$$+\frac{\delta^2}{8}\left(\langle v,\nabla^2\psi(\overline{x}_1)v\rangle-\langle R(x_{1/2})v,\nabla^2\psi(\overline{x}_2)R(x_{1/2})v\rangle\right)$$
$$=\frac{\delta^2}{8}\left(\langle v,\nabla^2\psi(\overline{x}_1)v\rangle-\langle R(x_{1/2})v,\nabla^2\psi(\overline{x}_2)R(x_{1/2})v\rangle\right).$$

In the last line we used that $\langle R(x_{1/2})v,\nabla\psi(x_{1/2})\rangle=-\langle v,\nabla\psi(x_{1/2})\rangle$. Similarly, in (4.7) we expand terms $\psi(x)$ and $\psi(x+v\delta)$ around $x_{1/2}$ to find

$$\psi(x)-\psi(x+v\delta)+\delta\langle v,\nabla\psi(x_{1/2})\rangle=\frac{\delta^2}{8}\left(\langle v,\nabla^2\psi(\overline{x}_1)v\rangle-\langle v,\nabla^2\psi(\overline{x}_2)v\rangle\right)$$

for $\overline{x}_1\in(x,x_{1/2})$ and $\overline{x}_2\in(x_{1/2},x+v\delta)$. If the Hessian of ψ is constant, as for instance in the Gaussian case, then proposals of this type are accepted with probability 1.

Overall, this means the acceptance probability has the form

$$1\wedge\exp\left(\frac{\delta^2}{8}\left(\langle v,\nabla^2\psi(\overline{x}_1)v\rangle-\langle\tilde{V},\nabla^2\psi(\overline{x}_2)\tilde{V}\rangle\right)\right),$$

which means that the probability of rejecting the proposed state is of order δ^2. In particular if π is a d-dimensional Gaussian with covariance $\Sigma=cI_d$, then the probability of accepting the proposed state in the MH step is equal to 1, as $\|\tilde{V}\|=\|v\|=1$. Hence in this case the splitting scheme **RDBDR** has the correct stationary distribution.

4.1.3 Algorithms with subsampling

One of the attractive features of ZZS and BPS is *exact subsampling*, i.e. the possibility when the potential is of the form $\psi(x)=\frac{1}{N}\sum_{j=1}^N\psi_j(x)$ of using only a randomly chosen ψ_j to simulate the next event time. The clearest application of this technique is Bayesian statistics, where $\psi(x)$ is the *posterior distribution*, x is the parameter of the chosen statistical model and, when the data points are independent realisations, ψ_j can be chosen to depend only on the j-th batch of data points and not on the rest of the dataset. Therefore, this technique can greatly reduce the computational cost per event time. Naturally, Bayesian statistics is not the only area where this structure of ψ arises. An example from molecular dynamics whit this type of potential is considered in Section 4.5.3. Here we define a splitting scheme of ZZS with this feature. With the same ideas it is possible to define a splitting scheme with subsampling based on BPS, but we do not give the details here for the sake of brevity.

Let us briefly explain the basic idea in the case of ZZS, as given in [23]. Assume the target distribution is of the form $\psi(x)=\frac{1}{N}\sum_{j=1}^N\psi_j(x)$ and define the switching rates $\lambda_i^j(x,v)=(v_i\partial_i\psi_j(x))_+$ for $i=1,\dots,d$ and $j=1,\dots,N$. Assuming we have a tractable M such that $\lambda_i^j(x+vt,v)\le M(t)$ for all $j=1,\dots,N$, one can use Poisson thinning to obtain a proposal τ for the next event time distributed as $\mathrm{Exp}(M(t))$. This proposal is then accepted with probability $\lambda_i^J(x+v\tau,v)/M(\tau)$, where $J\sim\mathrm{Unif}(\{1,\dots,N\})$ independently of the rest. This procedure defines a ZZS

Algorithm 15: Splitting scheme **DBD** for ZZS with subsampling

Input : Number of iterations N, initial condition (x, v), step size δ.
Output: Chain $(\overline{X}_{t_n}, \overline{V}_{t_n})_{n=0}^N$.
Set $n = 0$, $(\overline{X}_0, \overline{V}_0) = (x, v)$;
while $n < N$ **do**
  Set $\overline{X}_{t_{n+1}} = \overline{X}_{t_n} + \frac{\delta}{2}\overline{V}_{t_n}$;
  Draw $J \sim \text{Unif}(\{1, \dots, N\})$;
  for $i = 1 \dots, d$ **do**
    Obtain $(\overline{V}_{t_{n+1}})_i$ by simulating a pure jump process with kernel R_i and rate $v \mapsto (v\partial_i\psi_j(\overline{X}_{t_{n+1}}))_+$ with initial velocity $(\overline{V}_{t_n})_i$ and time horizon δ ;
  end
  Set $\overline{X}_{t_{n+1}} = \overline{X}_{t_{n+1}} + \frac{\delta}{2}\overline{V}_{t_{n+1}}$;
  Set $n = n + 1$;
end

with switching rates $\lambda_i(x, v) = \frac{1}{N}\sum_{j=1}^N \lambda_i^j(x, v)$, which are larger than the canonical rates, but keep π stationary.

Clearly the bottleneck of this procedure is that a sharp bound M needs to be available. Algorithm 15 defines an approximation of this process with a similar idea as [16]. At each iteration the algorithm draws $J \sim \text{Unif}(\{1, \dots, N\})$ independently of the rest and uses the corresponding ψ_J to update the process. Since the rates are now larger than the canonical rates, that is $\lambda_i(x, v) > (v_i\partial_i\psi(x))_+$, there can be more than one jump per component at each iteration. Nonetheless, the algorithm requires only one gradient computation per iteration since the position is not updated during the jump part. Moreover, in this case obtaining the gradient $\nabla\psi_J$ is an order 1 computation as opposed to the usual order N needed to compute the full gradient $\nabla\psi$.

4.1.4 PDMPs with boundaries

Another interesting feature of PDMPs such as BPS and ZZS is that, thanks to the simple deterministic dynamics, boundary conditions can be included and hitting times of the boundary can be easily computed (see [48] or [36] for a discussion of PDMPs with boundaries). Here we illustrate how to simply adapt splitting schemes to these settings by adding the boundary behaviour to the **D** part of the scheme.

Boundary terms appear for instance when the target distribution π is defined on a restricted domain, in which case a boundary jump kernel can be introduced as considered in [22]. In this case, Algorithms 11 and 12 can be easily modified by incorporating the boundary term in part **D** of the splitting scheme, as typically the boundary can be hit only if there is deterministic motion. Hence, the continuous

deterministic dynamics are applied as in the exact process, while other jumps are performed in the **B** steps.

Another example of this setting is when π is a mixture of a continuous density and a discrete distribution on finitely many states, as in Bayesian variable selection when a spike and slab prior is chosen. Sticky PDMPs were introduced in [26] to target a distribution of the form

$$\mu(\mathrm{d}x) \propto \exp(-\psi(x)) \prod_{i=1}^{d} (\mathrm{d}x_i + \frac{1}{c_i}\delta_0(\mathrm{d}x_i)),$$

which assigns strictly positive mass to events $\{x_i = 0\}$. The sticky ZZS of [26] is obtained following the usual dynamics of the standard ZZS and in addition freezing the i-th component for a time $\tau \sim \mathrm{Exp}(c_i)$ when x_i hits zero. The simulation of this process is challenging for the same reasons of the standard ZZS, since the two processes have the same switching rates λ_i for $i = 1, \dots, d$. The i-th component is either frozen, which is denoted by $(x_i, v_i) \in A_i$, or it evolves as given by the usual dynamics of ZZS. The generator can then be decomposed as $\mathcal{L} = \mathcal{L}_D + \mathcal{L}_B$ where $\mathcal{L}_D = \sum_{i=1}^{d} \mathcal{L}_{D,i}$ and $\mathcal{L}_B = \sum_{i=1}^{d} \mathcal{L}_{B,i}$,

$$\mathcal{L}_{D,i} f(x,v) = v_i \frac{\partial}{\partial x_i} f(x,v) \mathbb{1}_{A_i^C}(x_i, v_i) + c_i (f(T_i(x,v)) - f(x,v)) \mathbb{1}_{A_i}(x_i, v_i),$$
$$\mathcal{L}_{B,i} f(x,v) = \lambda_i(x,v)[f(x, R_i v) - f(x,v)] \mathbb{1}_{A_i^C}(x_i, v_i),$$

and $T_i(x,v)$ corresponds to unfreezing the i-th component (we refer to [26] for a detailed description). An iteration of the scheme **DBD** in this case proceeds by a first half step of **D**, which is identical to the continuous sticky ZZS but with λ_i temporarily set to 0. Hence frozen components are unfrozen with rate c_i and then start moving again, or unfrozen components move with their corresponding velocity v_i and become frozen for a random time with rate c_i if they hit $x_i = 0$. Then a full step of the usual bounce kernel **B** is done for the components which are not frozen, while for the frozen components, that is $(x_i, v_i) \in A_i$, the generator $\mathcal{L}_{B,i}$ does nothing and so the velocity cannot be flipped. So unfreezing is not possible in this step. The iteration ends with another half step of **D** in a similar fashion to the previous one.

These ideas are more general than the two specific examples we considered and do not introduce further difficulties for our schemes. We observe that in these cases it might be useful to consider the process obtained with the splitting schemes as continuous time processes. Finally, notice that a Metropolis correction can be added following Section 4.1.2, and subsampling is possible following Section 4.1.3.

4.2 Convergence of the splitting scheme

In this section we prove that under suitable conditions the splitting scheme **DJD** described in Section 4.1 is indeed a second order approximation of the original PDMP (2.31).

Note that in this section we have a PDMP defined on some arbitrary space E therefore it is not clear what it means to have a derivative, indeed we will typically be interested in the setting $E = \mathbb{R}^d \times \mathcal{V}$ for some set $\mathcal{V}$ which may be a discrete set. Instead of working with a full derivative we will define the directional derivative, D_Φ, in the direction Φ as

$$D_\Phi g(z) = \lim_{t\to 0} \frac{\mathrm{d}}{\mathrm{d}t} g(\varphi_t(z))$$

for any $g \in C(E)$ for which $t \mapsto g(\varphi_t(z))$ is continuously differentiable in t for every z. Note if E is a subset of $\mathbb{R}^d$ for some d and g is continuously differentiable then

$$D_\Phi g(z) = \Phi(z)^T \nabla g(z).$$

We extend this definition to multi-dimensional valued functions $G : E \to \mathbb{R}^m$ by defining $D_\Phi G(z) = (D_\Phi G^i(z))_{i=1}^m$. We define the space $\mathcal{C}_\Phi^{k,m}$ to be the set of all functions $g : E \to \mathbb{R}$ which are k times continuously differentiable in the direction Φ with all derivatives $D_\Phi^\ell g(z)$ up to order k bounded by a polynomial of order m. We endow this space with the norm

$$\|g\|_{\mathcal{C}_\Phi^{k,m}} := \sup_{z\in E} \frac{|g(z)| + \sum_{\ell=1}^k |D_\Phi^\ell g(z)|}{1+|z|^m}.$$

Let us make the following assumptions.

Assumption 4.5. *Let Φ be a globally Lipschitz vector field defined on E. We assume that the directional derivative in the direction Φ is well-defined and that Φ be continuous.*

Assumption 4.6. *The switching rate $\lambda : E \to [0,\infty)$ is twice continuously differentiable in the direction Φ and $\lambda, D_\Phi\lambda, D_\Phi^2\lambda$ grow at most polynomially. We denote by m_λ a constant such that $\|\lambda\|_{\mathcal{C}_\Phi^{2,m_\lambda}} < \infty$.*

Assumption 4.7. *Let Q be a probability kernel defined on E. We shall consider the operator $Q : \mathcal{C}_b(E) \to \mathcal{C}_b(E)$ defined by*

$$Qg(z) = \int g(\tilde{z}) Q(z, \mathrm{d}\tilde{z}), \qquad \text{for any } g \in \mathcal{C}_b(E). \tag{4.8}$$

Moreover we assume that Q has moments of all orders and Qg has at most polynomial growth of order m whenever g has at most polynomial growth of order m. For any $m \in \mathbb{N}$, and $g \in \mathcal{C}_\Phi^{1,m}$ we assume the following distribution is well-defined:

$$(D_\Phi Q)g(z) = D_\Phi(Qg)(z). \tag{4.9}$$

As an abuse of notation we shall write $D_\Phi Q$ also as a kernel. We assume for any $m \in \mathbb{N}$, and $g \in \mathcal{C}_\Phi^{1,m}$

$$|Qg(\varphi_s(z)) - Qg(z)| \le Cs(1+|z|^m)\|g\|_{\mathcal{C}_\Phi^{1,m}}, \tag{4.10}$$

and also that there exists a constant C such that for any $g \in \mathcal{C}_\Phi^{2,m}$

$$|Qg(\varphi_s(z)) - Qg(z) - sD_\Phi Qg(z)| \leq Cs^2(1+|z|^m)\|g\|_{\mathcal{C}_\Phi^{2,m}}. \tag{4.11}$$

Assumption 4.8. *The closure $(\mathcal{L}, D(\mathcal{L}))$ of the operator $(\mathcal{L}, \mathcal{C}_c^1(E))$ in L_μ^2 generates a C_0-semigroup $\mathcal{P}_t$. If $g \in \mathcal{C}_\Phi^{2,0}$ then we assume that $\mathcal{P}_t g$ is also twice continuously differentiable in the direction Φ and $\mathcal{L}\mathcal{P}_t g$ is continuously differentiable in the direction Φ. Moreover we assume $D_\Phi^2 \mathcal{P}_t g$ and $D_\Phi \mathcal{L}\mathcal{P}_t g$ are both polynomially bounded for finite t and for some $C > 0, R \in \mathbb{R}, m_{\mathcal{P}} \in \mathbb{N}$*

$$|D_\Phi \mathcal{P}_t g(z)| + |D_\Phi^2 \mathcal{P}_t g(z)| + |D_\Phi \mathcal{L}\mathcal{P}_t g(z)| \leq C(1+|z|^{m_{\mathcal{P}}})e^{Rt}\|g\|_{\mathcal{C}_\Phi^{2,0}}.$$

Assumption 4.9. *Let $\overline{Z}_{t_k}$ denote the approximation obtained by the splitting scheme **DJD**. Assume that for each k, $\overline{Z}_{t_k}$ has moments of all orders and moreover for every $M \in \mathbb{N}$ there exists some $\overline{G}_M$ such that*

$$\sup_{m \leq M} \mathbb{E}_z[|\overline{Z}_{t_k}|^m] \leq \overline{G}_M(z).$$

Theorem 4.10. *Let Z_t be a PDMP corresponding to the generator (2.31). Assume that Assumption 4.5 to Assumption 4.9 hold. Then there exist constants C, R such that for any $g \in \mathcal{C}_\Phi^{2,0} \cap D(\mathcal{L})$ we have for some $M \in \mathbb{N}$*

$$\sup_{k \leq n} |\mathbb{E}[g(Z_{t_k})] - \mathbb{E}[g(\overline{Z}_{t_k})]| \leq Ce^{Rt_n}\overline{G}_M(z)\delta^3 n\|g\|_{\mathcal{C}_\Phi^{2,0}}.$$

Proof. The proof is adapted from [16, Theorem 4.24] and can be found in Appendix 4.A.1. □

Example 4.11 (ZZS continued)**.** *Let us verify Assumptions 4.5 to 4.9 for the splitting scheme of ZZS defined in 4.1. In order to have a smooth switching rate we replace $\lambda_i(x,v)$ by*

$$\lambda_i(x,v) = \log\left(1 + \exp(v_i \partial_i \psi(x))\right).$$

This is shown to be a valid switching rate in [2]. We will assume that $\psi \in \mathcal{C}^2$ with bounded second and third derivatives. Let us now consider each assumption in turn.

Assumption 4.5: In this case $\Phi(x,v) = (v,0)^T$ which is clearly smooth and globally Lipschitz.

Assumption 4.6: Since λ_i is the composition of smooth maps and ψ we have that λ_i has the same smoothness in x as ψ and hence $x \mapsto \lambda_i(x,v)$ is $\mathcal{C}^2$. As $s \mapsto \log(1+e^s)$ grows at most linearly, has first and second derivatives bounded by 1 we have that $\lambda_i, \nabla_x \lambda_i$ and $\nabla_x^2 \lambda_i$ are all polynomially bounded.

Assumption 4.7: The proof of this can be found in Section 4.A.2.

Assumption 4.8: By [2] we have that $\mathcal{P}_t$ is a strongly continuous semigroup on L^2_μ with generator $(\mathcal{L}, D(\mathcal{L}))$ given as the closure of $(\mathcal{L}, C^1_c(E))$. Moreover we have that the assumptions of [65, Theorem 17] are satisfied and hence $\mathcal{P}_t g(x,v)$ is differentiable in x. Following the proof of [65, Theorem 17] one also has

$$|\nabla_x \mathcal{P}_t g| \leq C(1+|x|^m)e^{Rt}\|g\|_{\mathcal{C}^{1,0}_\Phi}.$$

Note here since $D_\Phi g(x,v) = v^T \nabla_x g$ we have that $\mathcal{C}^{k,0}_\Phi$ coincides with the space of continuous functions which are k-times continuously differentiable in the variable x. By the same arguments one can also obtain

$$|\nabla^2_x \mathcal{P}_t g| \leq C(1+|x|^m)e^{Rt}\|g\|_{\mathcal{C}^{2,0}_\Phi}.$$

Assumption 4.9: This will be established in Theorem 4.17.

4.3 Ergodicity of splitting schemes of BPS and ZZS

We shall now focus on results on ergodicity of splitting schemes of BPS and ZZS. In particular we show existence of an invariant distribution, characterise the set of all invariant distributions, and establish convergence of the law of the process to such distributions with geometric rate. In order to prove this we rely on the following classical result, due to Meyn and Tweedie [107] (here the specific statement is based on [79, Theorem 1.2], see also [64, Theorem S.7] for the explicit constants). Recall the definition of V-norm: $\|\mu\|_V := \sup_{|g|\leq V}|\mu(g)|$.

Theorem 4.12. *Consider a Markov chain with transition kernel P on a set E. Suppose that there exist constants $\rho \in [0,1)$, $C, \alpha > 0$, a function $V : E \to [1, +\infty)$ and a probability measure ν on E such that the two following conditions are verified:*

1. *Drift condition: for all $x \in E$,*

$$PV(x) \leq \rho V(x) + C. \tag{4.12}$$

2. *Local Dobelin condition: for all $x \in E$ with $V(x) \leqslant 4C/(1-\rho)$,*

$$\delta_x P \geqslant \alpha \nu\,.$$

Then, for all probability measures μ, μ' on E and all $n \in \mathbb{N}$,

$$\|\mu P^n - \mu' P^n\|_V \leqslant \frac{C}{\alpha}\kappa^n \|\mu - \mu'\|_V \tag{4.13}$$

where $\kappa = \max(1-\alpha/2, (3+\rho)/4)$. Moreover P admits a unique stationary distribution μ_* satisfying $\mu_*(V) < \infty$.

Remark 4.13. Under the Drift condition (4.12) alone, following the proof of [79, Theorem 1.2] in the case $\alpha = 0$ we get that for all probability measures μ, μ' on E,

$$\|\mu P - \mu' P\|_V \leqslant (\rho + 2C)\|\mu - \mu'\|_V \,.$$

We shall now consider our splitting schemes and prove geometric ergodicity under suitable conditions by showing that the assumptions of Theorem 4.12 are satisfied. The splitting schemes of BPS and ZZS are respectively addressed in Theorems 4.15 and 4.17 below and, in both cases, the dependence of all constants in (4.13) on the step size is made explicit (statements with more details are postponed to Appendixes 4.B and 4.C). More precisely, in both cases, we obtain a local Doeblin (or minorisation) condition with constant α after $n_* = \lceil t_*/\delta \rceil$ steps, where $t_* > 0$ plays the role of physical time and n_* is the number of steps needed to travel for an equivalent time. Here t_*, α are independent of δ. On the other hand, we show that the drift condition holds for one step of the kernel with constants $\rho = 1 - b\delta$ and $C = D\delta$, where b, D and the Lyapunov function V are independent of δ. This implies that for any $s > 0$ and any $\delta \in (0, \delta_0]$

$$\begin{aligned}(P_\delta)^{\lceil s/\delta \rceil} V &\leqslant (1 - b\delta)^{\lceil s/\delta \rceil}\, V + D\delta \sum_{k=0}^{\lceil s/\delta \rceil - 1} (1 - b\delta)^k \\ &\leqslant e^{-bs} V + \frac{D}{b}\,.\end{aligned}$$

Applying Theorem 4.12, we get for $P_\delta^{n_*}$ a long-time convergence estimate which is uniform over $\delta \in (0, \delta_0]$, that is for all $\delta \in (0, \delta_0]$ and $n \geq 1$ we find

$$\|\mu (P_\delta^{n_*})^n - \mu' (P_\delta^{n_*})^n\|_V \leqslant \frac{C'}{\alpha} \kappa^n \, \|\mu - \mu'\|_V,$$

where $C' = D/b$ and $\kappa = \max(1 - \alpha/2, (3 + e^{-bt_*})/4)$. Observe that the rhs does not depend on δ. Using the observation in Note 4.13, we can get convergence in V-norm for P^n. Indeed for $n = mn_* + r$ with $r < n_*$ we have

$$\begin{aligned}\|\mu P_\delta^n - \mu' P_\delta^n\|_V &= \|\mu P_\delta^{mn_*+r} - \mu' P_\delta^{mn_*+r}\|_V \\ &\leqslant \frac{C'}{\alpha} \kappa^m \|\mu P_\delta^r - \mu' P_\delta^r\|_V \\ &\leqslant \frac{C'}{\alpha} \left(1 + 2C'\right) \kappa^m \|\mu - \mu'\|_V \\ &\leqslant C'' \tilde{\kappa}^{n\delta} \|\mu - \mu'\|_V, \end{aligned} \tag{4.14}$$

where $\tilde{\kappa} = \kappa^{1/(t_* + \delta_0)} \in (0, 1)$ and $C'' = C' \left(1 + 2C'\right) / (\alpha\kappa)$ are independent from δ. Here we used that with computations identical to above we get the drift condition $P_\delta^r V \leq (1 - b\delta)V + D(1 - (1 - b\delta)^r)/b \leq V + C'$, which is enough for the current purpose. As a conclusion, the estimates given in Theorems 4.15 and 4.17 below (or

in Appendixes 4.B and 4.C for more details) give the expected dependency in δ for the convergence rate of the process toward equilibrium.

For splitting schemes of the BPS, we work under the following condition.

Assumption 4.14. *The dimension is $d \geq 2$, the velocity equilibrium ν is the uniform measure on $\mathbb{S}^{d-1}$. There exists $C > 0$ such that*

$$\frac{1}{C}|x|^2 - C \leqslant \psi(x) \leqslant C|x|^2 + C\,, \qquad \frac{1}{C}|x| - C \leqslant |\nabla\psi(x)| \leqslant C|x| + C$$

for all $x \in \mathbb{R}^d$. Moreover, $\|\nabla^2\psi\|_\infty < \infty$ and, without loss of generality, $\inf\psi = 1$.

Notice that, when $d = 1$, the BPS and the ZZS coincide, in which case we refer to Theorem 4.17 below. Our result of ergodicity for splitting schemes of the BPS is the following.

Theorem 4.15. *Consider any scheme of the BPS based on the decomposition **D**,**R**,**B**. Under Assumption 4.14, there exist $\delta_0, a, C'' > 0, \tilde{\kappa} \in (0,1)$ and $V : \mathbb{R}^d \times \mathbb{S}^{d-1} \to [1, +\infty)$ satisfying*

$$\text{for all } x \in \mathbb{R}^d, v \in \mathbb{S}^{d-1}, \qquad e^{|x|/a}/a \leqslant V(x,v) \leqslant ae^{a|x|}$$

such that, for all $\delta \in (0, \delta_0]$, Theorem 4.12 is applicable and (4.14) *holds with these $C'', \tilde{\kappa}, V$.*

Proof. The proof can be found in Appendix 4.B. □

More care is required for the **DBD** scheme of the ZZS since this Markov chain has periodicity and is not irreducible, which is reminiscent of the discrete-space Zig-Zag chain studied in [111]. Let us illustrate this behaviour by considering the one dimensional setting. Let (x, v) be the initial condition of the process. Since v has magnitude 1, the position component x can only vary by multiples of the step size δ. Thus for a fixed initial condition (x, v) the process remains on a grid $(x+\delta\mathbb{Z})\times\{-1, 1\}$. Moreover, after a single step of the scheme there are two possible outcomes: either the velocity does not change, in which case x moves to $x+\delta v$, or the velocity is flipped and the position remains the same. This means that the change in the position (by amounts of δ) plus half the difference in the velocity always changes by ± 1 each step and hence is equal to the number of steps in the scheme up to multiples of two, i.e.

$$\frac{\overline{X}_{n\delta} - x}{\delta} + \frac{1}{2}(\overline{V}_{n\delta} - v) \in n + 2\mathbb{Z}.$$

As a consequence, the chain lives on two disjoint sets depending on whether n is even or odd, which means that it is periodic. To overcome this issue, we consider the chain with one step transition kernel given by $P_\delta^2 = P_\delta P_\delta$, i.e. we restrict to the case of an even number of steps. The Markov chain with kernel P_δ^2 is aperiodic, but it is

not irreducible on $\mathbb{R}^d$ and hence has (infinitely) many invariant measures. In order to characterise the invariant measures we restrict to the set in which the Markov kernel P_δ^2 is irreducible. For fixed $(x, v) \in \mathbb{R}^d \times \{-1, 1\}^d$ we construct the grid which contains (x, v) as follows:

$$D(x, v) := \{(y, w) \in C \times \{\pm 1\}^d : (y_i, w_i) \in D_1(x_i, v_i) \text{ for all } i = 1, \dots, d\}, \quad (4.15)$$

where $D_1(x_i, v_i) := D_+(x_i, v_i) \cup D_-(x_i, v_i)$, with

$$D_+(x_i, v_i) := \{(y_i, w_i) : w_i = v_i,\ y_i = x_i + m\delta,\ m \in 2\mathbb{Z}\},$$
$$D_-(x_i, v_i) := \{(y_i, w_i) : w_i = -v_i,\ y = x_i + m\delta,\ m \in 2\mathbb{Z} + 1\}.$$

In this case we show in Theorem 4.17 that the Markov chain with transition kernel P_δ^2 is irreducible on $D(x, v)$, has a unique invariant measure, $\pi_\delta^{x,v}$, and is geometrically ergodic. Now we can characterise all the invariant measures of the Markov chain with transition kernel P_δ^2 defined on $\mathbb{R}^d \times \{-1, 1\}^d$ as the closed convex hull of the set $\{\pi_\delta^{x,v} : x \in \mathbb{R}^d, v \in \{-1, 1\}^d\}$. Now consider the Markov chain with transition kernel P_δ^2 on $\mathbb{R}^d \times \{-1, 1\}^d$. For any initial distribution μ we have convergence of μP_δ^{2n} to some measure π_δ^μ as n tends to ∞ and π_δ^μ is given by

$$\pi_\delta^\mu(\varphi) = (\mu\pi_\delta^{x,v})(\varphi) := \int_{\mathbb{R}^d \times \{-1,1\}^d} \int_{\mathbb{R}^d \times \{-1,1\}^d} \varphi(y, w)\pi_\delta^{x,v}(\mathrm{d}y, \mathrm{d}w)\mu(\mathrm{d}x, \mathrm{d}v). \quad (4.16)$$

We use the next assumption to verify that Theorem 4.12 applies for initial conditions drawn from probability distributions with support on $D(x, v)$.

Assumption 4.16. *Consider switching rates $\lambda_i(x, v) = (v_i\partial_i\psi(x))_+ + \gamma_i(x)$ for $i = 1, \dots, d$. $\psi \in \mathcal{C}^2(\mathbb{R}^d)$ and the following conditions hold:*

(a) The switching rates $\lambda_i(x, v)$ are such that there exist $x_0 \geq 0$ such that for all $x_1 > x_0$

$$\underline{\lambda}(x_1) := \min_{i=1,\dots,d} \quad \min_{(x,v):\, x_i v_i \in [x_0, x_1],\, |x_j| \in [x_0, x_1] \text{ for all } j \neq i} \lambda_i(x, v) > 0.$$

(b) For $|x| \geq R$ for some $R > 0$

$$\sup e^{\left(t^2(|(v+w)^T\nabla^2\psi(y_1))_i| + 2t|(w\nabla^2\psi(y_2))_i|\right)}\gamma_i(x + vt)e^{tv_i\partial_i\psi(x)} \leq \gamma_0 < 1, \quad (4.17)$$

where the supremum is over $t \in (0, 1), y_1, y_2 \in B(x, t\sqrt{d}), v, w \in \{-1, 1\}^d$.

(c) Denote as $B(x, \delta\sqrt{d})$ the ball with centre at x and radius $\delta\sqrt{d}$. Then

$$\lim_{\|x\| \to \infty} \sup_{y_1, y_2 \in B(x, \delta\sqrt{d})} \frac{\max\{1, \|\nabla^2\psi(y_1)\|\}}{|\partial_i\psi(y_2)|} = 0$$

for all $0 \leq \delta \leq \delta_0$, $i = 1, \dots, d$, where $\delta_0 = 2(1 + \gamma_0)^{-1}$, for γ_0 as in part (b).

Part (a) in Assumption 4.16 is inspired by [21, Assumption 3] and is used to show that a minorisation condition holds. This condition is either a consequence of properties of the target, or else can be enforced by taking a non-negative excess switching rate, in which case $\gamma_i(x)$ can be chosen to be a continuous function $\gamma_i : \mathbb{R}^d \to (0, \infty)$. In principle one could prove a minorisation condition using the techniques of [24], but this is beyond the scope of this paper. Part (b) is a condition on the decay of the refreshment rate, while Part (c) is similar to Growth Condition 3 in [24] and is satisfied for instance if ψ is strongly convex with globally Lipschitz gradient. These two conditions are used to show that a drift condition holds.

Theorem 4.17. *Consider the splitting scheme **DBD** for ZZS. Suppose Assumption 4.16 holds. Then there exist $C'', \delta_0 > 0, \tilde{\kappa} \in (0,1)$ and $V : \mathbb{R}^d \times \{-1,1\}^d \to [1, \infty)$ satisfying for all $(x, v) \in \mathbb{R}^d \times \{-1,1\}^d$*

$$\prod_{i=1}^{d} (1 + 2|\partial_i \psi(x)|)^{-\frac{1}{2}} \leq \frac{V(x,v)}{\exp(\beta\psi(x))} \leq \prod_{i=1}^{d} (1 + 2|\partial_i \psi(x)|)^{\frac{1}{2}}$$

for all $\beta \in (0, 1/2)$ such that, for all $\delta \in (0, \delta_0]$, the following holds:

1. *Fix $(x, v) \in \mathbb{R}^d \times \{-1,1\}^d$. Theorem 4.12 is applicable to $P_\delta^2 = P_\delta P_\delta$ seen as a transition kernel on $D(x, v)$, and the inequality (4.14) holds (with P_δ replaced by P_δ^2) with these $C'', \tilde{\kappa}, V$ for any μ, μ' having support on $D(x, v)$.*
2. *For any probability measure μ on $\mathbb{R}^d \times \{-1,1\}^d$ with $\mu(V) < \infty$, we have that μP_δ^{2n} converges as $n \to \infty$ to the measure π_δ^μ given by (4.16) where $\pi_\delta^{x,v}$ is the unique invariant measure of P_δ^2 on $D(x, v)$ and we have*

$$\|\mu P_\delta^{2n} - \mu \pi_\delta^{x,v}\|_V \leq C'' \tilde{\kappa}^{n\delta} \int \|\delta_{(x,v)} - \pi_\delta^{x,v}\|_V \mu(\mathrm{d}x, \mathrm{d}v)\,. \tag{4.18}$$

Proof. The proof can be found in Appendix 4.C.1. □

Under similar assumptions we establish geometric ergodicity of schemes **DRBRD**, **RDBDR** of ZZS, where the switching rates in the **B** part are $\lambda_i(x, v) = (v_i \partial_i \psi(x))_+$, i.e. the canonical rates, while refreshments in the **R** part are independent draws from $\mathrm{Unif}(\{\pm 1\}^d)$ with rate $\gamma(x) : \mathbb{R}^d \to [0, \infty)$. The rigorous statement of this result, Theorem 4.31, and its proof can be found in Appendix 4.C.2.

4.4 Expansion of the invariant measure of splitting schemes for BPS

In this section we investigate the bias in the invariant measure of different splittings of BPS and draw conclusions on which schemes perform best. Motivated by Theorems

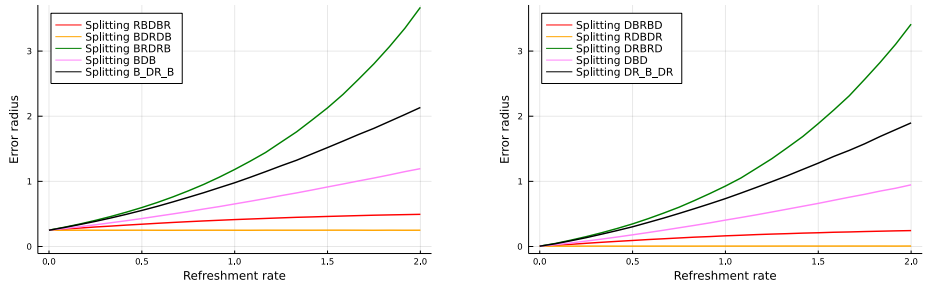


Figure 4.1: Empirical error for the radius statistic $t(x) = x^2$ with a one-dimensional standard Gaussian target. The step size is set to $\delta = 1.0$, the number of iterations is $N = 10^5$, and the experiment is repeated 250 times. The schemes **BDB** (left) and **DBD** (right) correspond to including the refreshment part in **B**. In schemes **B_DR_B** (left) and **DR_B_DR** (right) we denote by **B** the standard bounce part, by **DR** the transition kernel which corresponds to having refreshments and deterministic motion together, and we use underscores to divide these two kernels. Here ν is the uniform distribution on $\{\pm 1\}$.

4.10 and 4.15, we assume that the processes corresponding to our splitting schemes have an invariant distribution with density

$$\mu_\delta(x, v) = \mu(x, v)(1 - \delta^2 f_2(x, v) + \mathcal{O}(\delta^4)), \tag{4.19}$$

where $\mu(x, v) = \nu(v)\pi(x)$, π is the target and ν is a distribution satisfying Assumption 4.19 below, e.g. the uniform distribution on the unit sphere or the standard Gaussian. It is then our goal to compute and compare f_2 for different schemes.

There are several splitting schemes that could be compared, and thus we make a selection of the ones it is worth focusing on. The numerical simulations shown in Figure 4.1 give an idea of the relative performance of the schemes. The plots show that the schemes that have **DBD** as their limit as the refreshment rate goes to zero have a smaller bias in the x component compared to those that converge to **BDB**. Naturally the difference between the two schemes is expected to vanish as $\delta \to 0$ and also appears to be diminishing as the dimension increases (see Figure 4.4). Based on this result we decide to concentrate on schemes **RDBDR**, **DBRBD**, **DRBRD**, as well as **BDRDB**. Note that all these schemes have the same cost of one gradient computation per iteration (in **BDRDB** it is sufficient to keep track of the gradient at the previous iteration).

Following the approach of [89] , we will show in Section 4.4.1 (more precisely in Proposition 4.22) that the second order of the bias f_2 can be computed analytically for one-dimensional targets. We then focus on the dependence of f_2 on the refreshment rate, which is the only parameter of the algorithm (outside of δ). As we will see, and as already hinted by Figure 4.1, some splittings like **RDBDR** and **BDRDB** are robust

to poor choices of the refreshment rate, while others like **DBRBD** and **DRBRD** have linear or quadratic dependence on λ_r. The numerical experiments of Section 4.4.2 confirm the theoretical results of Proposition 4.22 and suggest that the bias behaves similarly in higher dimensions, where obtaining f_2 analytically is very challenging. In particular, in Figure 4.3 we show that, in the cases we consider, splitting **RDBDR** is the scheme that shows the best overall behaviour. This scheme was shown to be unbiased for standard Gaussian targets in Section 4.1.2, and is confirmed to have $f_2 = 0$ in such cases in Section 4.4.1. Moreover, we fully characterise the invariant distribution of **RDBDR** in one dimension in Proposition 4.24.

Remark 4.18. In Section 4.3 we will see cases where a splitting scheme may admit more than one invariant measure. In such cases it is not immediately clear what the expansion (4.19) means. In order to make (4.19) consistent as $\delta \to 0$, in those cases we consider μ_δ as the limit of the law of the splitting scheme as the number of steps tends to infinity when the process is started according to μ.

4.4.1 Computing f_2

Let us discuss briefly how to find f_2 with the approach of [89]. Using the Baker-Campbell-Hausdorff (BCH) formula (see e.g. [28]) we can find $\mathcal{L}_2$ such that

$$\mathbb{E}_{x,v}[f(\overline{X}_\delta, \overline{V}_\delta)] = f(x,v) + \delta \mathcal{L} f(x,v) + \delta^3 \mathcal{L}_2 f(x,v) + \mathcal{O}(\delta^4).$$

Here $\mathcal{L}$ is the infinitesimal generator of the continuous time process. Integrating both sides with respect to μ_δ and using that μ_δ is an invariant measure for the splitting scheme we obtain

$$\begin{aligned}\int f(x,v)\mu_\delta(x,v)\mathrm{d}x\mathrm{d}v = \int f(x,v)\mu_\delta(x,v)\mathrm{d}x\mathrm{d}v + \delta \int \mathcal{L} f(x,v)\mu_\delta(x,v)\mathrm{d}x\mathrm{d}v \\ + \delta^3 \int \mathcal{L}_2 f(x,v)\mu_\delta(x,v)\mathrm{d}x\mathrm{d}v + \mathcal{O}(\delta^4).\end{aligned}$$

Substituting for μ_δ with the expansion (4.19) we have

$$\begin{aligned}0 = \delta \int \mathcal{L} f(x,v)\mu(x,v)\mathrm{d}x\mathrm{d}v - \delta^3 \int \mathcal{L} f(x,v)\mu(x,v) f_2(x,v)\mathrm{d}x\mathrm{d}v \\ + \delta^3 \int \mathcal{L}_2 f(x,v)\mu(x,v)\mathrm{d}x\mathrm{d}v + \mathcal{O}(\delta^4).\end{aligned}$$

Since μ is an invariant measure for BPS we have $\int \mathcal{L} f p \,\mathrm{d}x\mathrm{d}v = 0$ which gives the equation

$$\mathcal{L}^*(\mu f_2) = \mathcal{L}_2^* \mu. \tag{4.20}$$

Here $\mathcal{L}^*$ and $\mathcal{L}_2^*$ are the adjoints on $\mathcal{L}$ and $\mathcal{L}_2$ in L^2 with respect to Lebesgue measure. Since there is not a unique solution to the equation (4.20) we need to impose a

compatibility condition. Since both $\hat{\mu}$ and μ are probability densities, integrating (4.19) gives the requirement

$$\int f_2(x,v)\mu(x,v)\mathrm{d}x\mathrm{d}v = 0. \tag{4.21}$$

It is then the goal of this section to solve (4.20) and compare the solutions corresponding to the different splitting schemes. We start by computing the term $\mathcal{L}_2^*$ using the BCH formula. Recall that the adjoint of the generator of BPS is given by

$$\begin{aligned}\mathcal{L}^* g(x,v) = &-\langle v, \nabla_x g(x,v)\rangle + ((g\lambda_1)(x,R(x)v) - (g\lambda_1)(x,v)) \\ &+ \lambda_r\left(\nu(v)\int g(x,y)\mathrm{d}y - g(x,v)\right).\end{aligned}$$

We now compute $\mathcal{L}_2^*$ for the splitting schemes **DBRBD**, **RDBDR**, **DRBRD**, **BDRDB**. Let us start with an assumption on the invariant distribution of the velocity vector.

Assumption 4.19. *The invariant measure for the velocity component ν satisfies the following conditions:*

1. Invariance under rotations*: $\nu(w) = \nu(v)$ for any v, w such that $|v| = |w|$;*
2. Mean zero*: $\mathbb{E}_\nu[V] = 0$;*
3. Isotropic*: for some $b > 0$ it holds that $\mathrm{Cov}_\nu(V) = bI$.*

These properties hold for instance if ν is the standard Gaussian distribution, as well as if ν is the uniform on the unit sphere (in that case $b = 1/d$).

Proposition 4.20. *Let Assumption 4.19 hold and define*

$$\begin{aligned}A(x,v) &= \frac{3}{2}\lambda_r\Big(b\,\mathrm{tr}\big(\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x)\big) \\ &\quad + 2\langle v, \nabla\psi(x)\rangle\lambda_1(x,R(x)v) + \langle v, \nabla^2\psi(x)v\rangle\Big), \\ B(x,v) &= \frac{3}{2}\lambda_1(x,R(x)v)\Big(\langle v, \nabla^2\psi(x)v\rangle - \langle R(x)v, \nabla^2\psi(x)R(x)v\rangle\Big) \\ &\quad + \frac{1}{2}\langle v, \nabla_x(\langle v, \nabla^2\psi(x)v\rangle)\rangle, \\ C(x,v) &= 3\lambda_1(x,R(x)v)\big(-2\langle v, \nabla^2\psi(x)v\rangle + \langle v, \nabla\psi(x)\rangle^2\big) \\ &\quad - \langle v, \nabla(\langle v, \nabla^2\psi(x)v\rangle)\rangle, \\ D(x,v) &= \frac{3}{2}\lambda_r\Big(b\,\mathrm{tr}\big(\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x)\big) + \langle v, \nabla^2\psi(x)v\rangle \\ &\quad + \langle v, \nabla\psi(x)\rangle\big(3\lambda_1(x,R(x)v) + \lambda_1(x,v)\big)\Big).\end{aligned}$$

The splitting scheme ***DBRBD*** *satisfies*

$$\mathcal{L}_2^*\mu(x,v) = \frac{\mu(x,v)}{12}\Big(A(x,v) + B(x,v)\Big).$$

The splitting scheme ***RDBDR*** *satisfies*

$$\mathcal{L}_2^*\mu(x,v) = \frac{\mu(x,v)}{12}B(x,v).$$

The splitting scheme ***DRBRD*** *satisfies*

$$\mathcal{L}_2^*\mu(x,v) = \frac{\mu(x,v)}{12}\left(D(x,v) + B(x,v) + \frac{3}{2}\lambda_r^2\langle v, \nabla\psi(x)\rangle\right).$$

The splitting scheme ***BDRDB*** *satisfies*

$$\mathcal{L}_2^*\mu(x,v) = \frac{\mu(x,v)}{12}\Big(-A(x,v) + C(x,v)\Big).$$

Proof. The proof can be found in Appendix 4.E. □

Remark 4.21. Clearly, if $\mathcal{L}_2^*\mu = 0$ then f_2 must be a constant that satisfies (4.21) and hence it must be that $f_2 = 0$, i.e. the second order term in μ_δ is zero. This is the case for instance for scheme **RDBDR** when the target is a multidimensional standard Gaussian. Indeed in Section 4.1.2 we proved that **RDBDR** is unbiased for standard Gaussian targets, thus this is a consistent result. In the same setting, we observe that $f_2 = 0$ for schemes **DBRBD** and **DRBRD** when the refreshment rate is $\lambda_r = 0$. This is an expected result, as when $\lambda_r = 0$ these schemes coincide with **RDBDR**. In Figure 4.2 we confirm that, for a one-dimensional standard Gaussian and when $\lambda_r = 0$, the scheme **DBD** is unbiased, while the scheme **BDB** is of second order.

Equation (4.20) is in general hard to solve, as the adjoint of BPS contains both derivatives and integrals. Nonetheless, we are able to solve (4.20) and find f_2 in the one-dimensional case, as stated in the next Proposition.

Proposition 4.22. *Consider the one-dimensional setting with state space* $\mathbb{R}\times\{\pm 1\}$ *and target distribution* $\mu(x,v) = \pi(x)\nu(v)$ *with* $\pi \propto \exp(-\psi)$ *and* $\nu = \mathrm{Unif}(\{\pm 1\})$. *Let* $\lambda_r \geq 0$ *be the refreshment rate. Then the function* f_2 *that solves* (4.20) *is*

$$f_2(x,+1) = f_2^+(0) + \int_0^x \left(\left(\frac{\lambda_r}{2} + (-\psi'(y))_+\right)g(y) - \frac{\mathcal{L}_2^*\mu(y,+1)}{\mu(y,+1)}\right)\mathrm{d}y,$$
$$f_2(x,-1) = f_2(x,+1) + g(x),$$

where

$$g(x) = \exp\left(\psi(x)\right)\int_{-\infty}^x \left(\frac{\mathcal{L}_2^*\mu(y,+1)}{\mu(y,+1)} + \frac{\mathcal{L}_2^*\mu(y,-1)}{\mu(y,-1)}\right)\exp(-\psi(y))\mathrm{d}y,$$

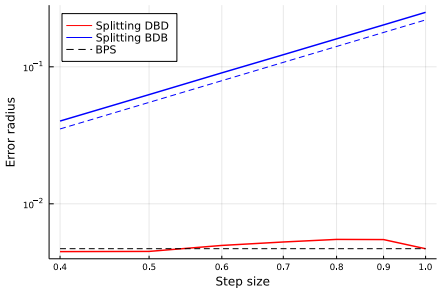


Figure 4.2: Error for the radius statistic for a one-dimensional standard Gaussian target. Here $\lambda_r = 0$ for both schemes **DBD** and **BDB**. The *dashed, blue line* corresponds to second order convergence. The time horizon is fixed to $T = 10^5$ and the number of iterations is $N = T/\delta$. The experiment is repeated 250 times.

$$f_2^+(0) = -\int_{-\infty}^{\infty} \left(\frac{g(x)}{2} + \int_0^x \left(\left(\frac{\lambda_r}{2} + (-\partial\psi(y))_+ \right) g(y) - \frac{\mathcal{L}_2^* \mu(y,+1)}{\mu(y,+1)} \right) \mathrm{d}y \right) \pi(x) \mathrm{d}x.$$

Proof. The proof can be found in Appendix 4.D.1. □

Remark 4.23. An immediate consequence of Propositions 4.20 and 4.22 is that in the one dimensional case the second order term of the bias of scheme **RDBDR** is always independent of the refreshment rate and of v. Indeed applying the propositions we find

$$f_2(x,v) = f_2^+(0) - \frac{1}{24} \int_0^x \psi^{(3)}(y) \mathrm{d}y \tag{4.22}$$

with $f_2^+(0) = \frac{1}{24} \int_{-\infty}^{\infty} \int_0^x \psi^{(3)}(y) \mathrm{d}y \pi(\mathrm{d}x)$.

In fact, in 1D, for the scheme **RDBDR**, we can get an explicit expression for the invariant measure.

Proposition 4.24. *Consider the scheme **RDBDR** for BPS in one dimension, where the velocity is refreshed from $\nu = \mathrm{Unif}(\{\pm 1\})$. Then for a fixed initial condition $x \in \mathbb{R}$ and step size δ the distribution with support on $\{y \in \mathbb{R} : y = x + n\delta,\ n \in \mathbb{Z}\} \times \{\pm 1\}$ given by*

$$\mu_\delta(y,v) \propto e^{-\psi_\delta(y)}$$

where $\psi_\delta(x) = \psi(x)$ and for $y = x + nv\delta$, $n \in \mathbb{N}$

$$\psi_\delta(y) = \psi(x) + \delta \sum_{\ell=1}^{n} \psi'(x + (\ell - 1/2)v\delta)$$

is stationary for the process. Moreover, under the conditions of Theorem 4.17 we obtain that μ_δ is ergodic, in the sense that for all bounded functions

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(\overline{X}_{t_n}, \overline{V}_{t_n}) = \mu_\delta(f) \quad \mathbb{P}_{x,v} - a.s.$$

Proof. The proof can be found in Appendix 4.D.2. □

Once again it is clear that the scheme is unbiased in the Gaussian case $\psi(x) = x^2/(2\sigma^2)$ (in the sense that $\psi_\delta(y) = \psi(y)$ for all $y = x + n\delta v$, i.e. the BPS is ergodic with respect to the restriction of the true Gaussian target to the grid, and moreover the target measure is invariant for the scheme). More generally, for $y = x + vn\delta$ with $n \in \mathbb{N}$ we get

$$\begin{aligned}
\psi(y) &= \psi(x) + \sum_{\ell=1}^{n} \int_{-\delta/2}^{\delta/2} \psi'(x + v(\ell - 1/2)\delta + u)\mathrm{d}u \\
&= \psi_\delta(y) + \frac{1}{2} \sum_{\ell=1}^{n} \int_{-\delta/2}^{\delta/2} u^2 \psi^{(3)}(x + v(\ell - 1/2)\delta)\mathrm{d}u + O(n\delta^5) \\
&= \psi_\delta(y) + \frac{\delta^2}{24} \int_x^y \psi^{(3)}(u)\mathrm{d}u + O(\delta^4|x - y|)\,.
\end{aligned}$$

Setting $x = 0$ this gives $\psi_\delta = \psi + \delta^2 f_2 + O(\delta^4)$ with $f_2(y) = \int_0^y \psi^{(3)}(u)du$, which is the same of Equation (4.22). Indeed the term $f_2^+(0)$ in (4.22) was introduced to make $\exp(-\psi)(1 + \delta^2 f_2)$ a probability distribution and would appear also in the present context. Hence Propositions 4.22 and 4.24 agree.

4.4.2 Application to three one-dimensional target distributions

In this section we compare the splitting schemes by applying Proposition 4.22 to three one-dimensional target distributions: a centred Gaussian distribution, a distribution with non-Lipschitz potential $\psi(x) = x^4$, and a Cauchy distribution. The formal statements can be found in the Appendix 4.D.3, correspondingly in Propositions 4.35, 4.36, 4.37. Here instead of giving the complicated analytic expressions for f_2 in all cases, we give plots of the TV distance between μ and μ_δ as a function of λ_r as given by Propositions 4.35, 4.36, 4.37. The results, both according to the theory and numerical simulations, are shown in Figure 4.3.

Let us briefly explain how the TV distance is derived from the analytic expression of f_2. We shall focus on the position part of μ_δ, which we denote as π_δ. By marginalising and recalling in this context $\nu = \mathrm{Unif}(\{\pm 1\})$ we obtain

$$\pi_\delta(x) = \pi(x)\left(1 - \frac{\delta^2}{2}(f_2(x, +1) + f_2(x, -1))\right) + \mathcal{O}(\delta^4), \tag{4.23}$$

Using (4.23) we can express the TV distance between π and π_δ as

$$\|\pi - \pi_\delta\|_{TV} = \frac{\delta^2}{2} \sup_A \left| \int_A (f_2(x,+1) + f_2(x,-1))\pi(x)\mathrm{d}x \right| + \mathcal{O}(\delta^4). \tag{4.24}$$

The δ^2 contribution of the rhs can be computed by plugging in the expressions for f_2 found in Propositions 4.35, 4.36, and 4.37. We neglect higher order terms.

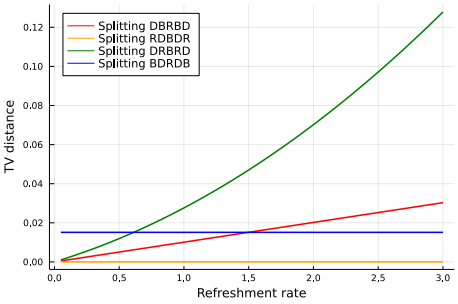
Let us comment on these results. First of all, the theoretical results are consistent with the numerical experiments of Figure 4.3. Indeed, it is clear that the schemes **RDBDR** and **BDRDB** have a bias that is independent of the refreshment rate, while **DBRBD** and **DRBRD** have respectively linear and quadratic dependence. In the one-dimensional case, the plots show that it is best to choose $\lambda_r = 0$, which is possible as in this case BPS is irreducible. However, in higher dimensional settings it is necessary to take $\lambda_r > 0$ as shown in Figure 4.4. Since choosing a good value of λ_r is difficult and depends on the target distribution, it is desirable to use schemes that have good performance for most values of λ_r. Moreover, it is clear from Figure 4.3 that **RDBDR** is indeed unbiased in the Gaussian case, and also has the smallest bias out of all the considered splittings with the exception of the Cauchy target, where the difference in performance between **RDBDR** and **BDRDB** is almost negligible and seems to slightly favour the latter in experiments. In this case, we also see a small dependence on λ_r for **RDBDR** and **BDRDB**, which could be due to higher order terms.

The experiments in Figure 4.4 suggest that the findings of the one-dimensional case extend to multi-dimensional targets. In particular, **RDBDR** has either a better performance than other splittings or behaves very similarly to **BDRDB** both on an independent as well as a correlated Gaussian. Moreover, the independence on λ_r of the bias of schemes **DBRBD** and **DRBRD** is confirmed also when $d > 1$.

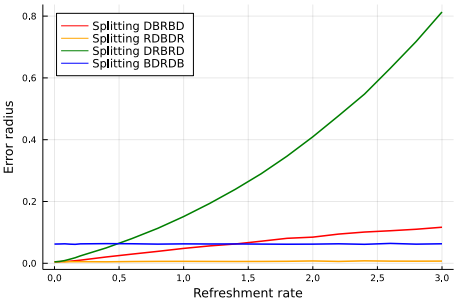
In conclusion, we have conducted a detailed analysis of the bias in the invariant measure, both theoretical in Propositions 4.35, 4.36, 4.37, and empirical in Figures 4.1, 4.2, 4.3, 4.4, and the evidence suggests that **RDBDR** is the best candidate out of the pool of splitting schemes that are available. The closest competitor **BDRDB** shows similar performance in some settings, but a larger bias in others in which **RDBDR** enjoys desirable properties.

4.5 Numerical experiments

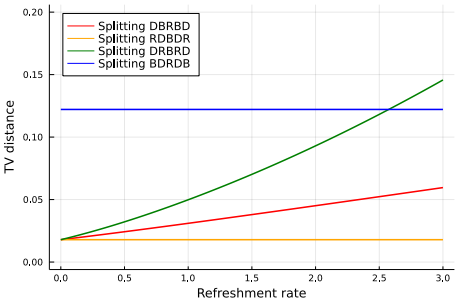
In this section we discuss some numerical simulations for the proposed samplers. The codes for all these experiments can be found at `https://github.com/andreabertazzi/splittingschemes_PDMP`.



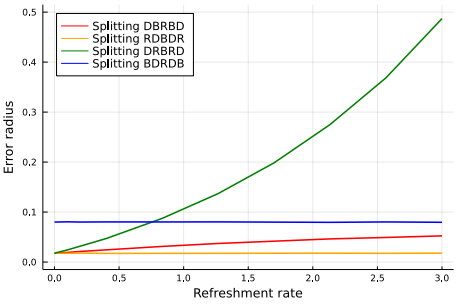
(a) TV distance according to Proposition 4.35



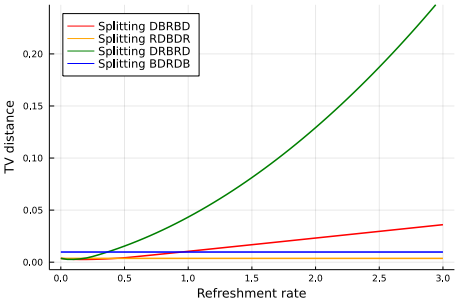
(b) Absolute value of the error for the radius statistic.



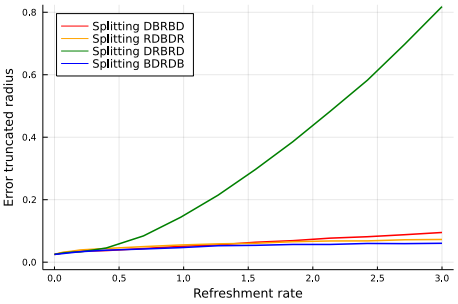
(c) TV distance according to Proposition 4.36.



(d) Absolute value of the error for the radius statistic.

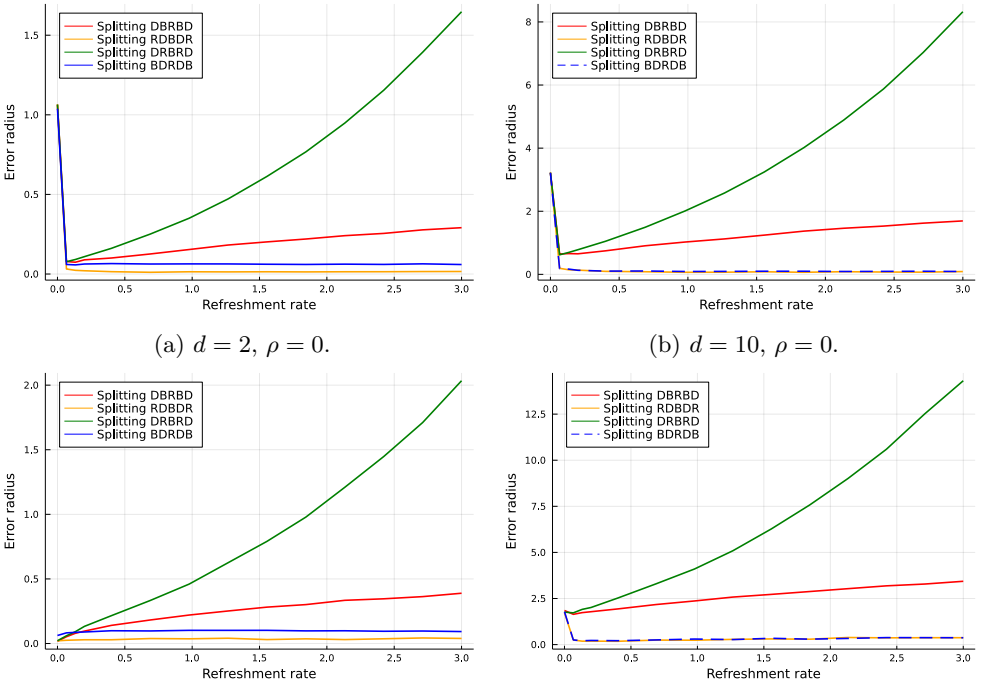


(e) TV distance according to Proposition 4.37.



(f) Absolute value of the error for $\min\{4, x^2\}$.

Figure 4.3: Total variation distance to the true target as given by the second order term in (4.24) (*left*) and numerical simulations (*right*) for the various splittings. The *top row* is obtained with standard Gaussian target, the *middle row* with $\psi(x) = x^4$, and the *bottom row* with a one dimensional Cauchy target with $\gamma = 1$. In all plots $\delta = 0.5$ and the number of iterations is $N = 2 \cdot 10^5$. In the Gaussian and Cauchy cases we initialise the processes at μ.



(a) $d = 2$, $\rho = 0$.

(b) $d = 10$, $\rho = 0$.

(c) $d = 2$, $\rho = 0.7$.

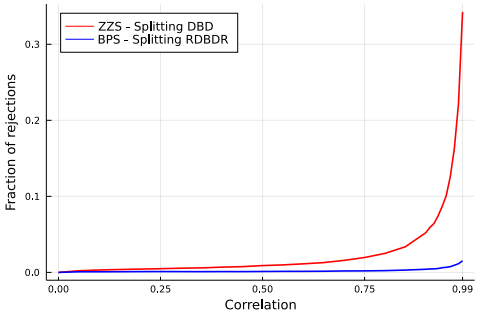
(d) $d = 10$, $\rho = 0.7$.

Figure 4.4: Error of estimators of the radius with splitting schemes for BPS with a Gaussian target with covariance $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho$ for $i \neq j$. The step size is $\delta = 0.5$ and the number of iterations is $N = 2 \cdot 10^5$. The processes are initialised with a draw from μ.

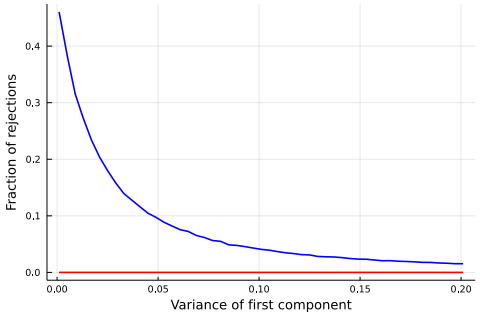
4.5.1 Gaussian target

Here we study the behaviour of the proposed algorithms on two types of Gaussian targets. The first type is a correlated Gaussian, for which the covariance matrix has unitary variances and correlation ρ between all components. The second type is an independent Gaussian, where components $i \geq 2$ have unitary variance, while the first component has (small) variance σ^2. We study the performance of our algorithms as a function of ρ and σ^2, as well as of the step size δ and the dimension of the target. In particular we first focus on the number of rejections in the Metropolised algorithms, that is Algorithms 13 and 14, and then we focus on the error in the estimation of the expected radius for all our algorithms.

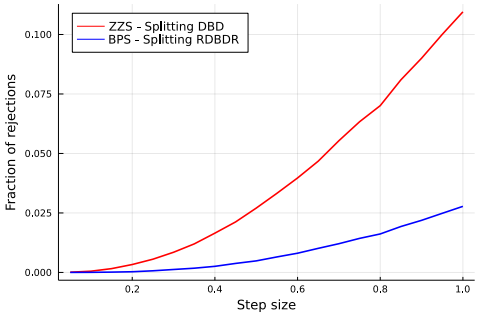
Figure 4.5 shows the number of rejections in adjusted algorithms, with the left part of the plot showing the first type of Gaussian target and the right part showing the second. This experiment allows us to understand the efficiency of the Metropolis



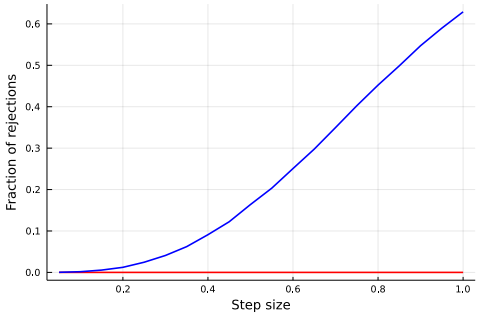
(a) Here $\delta = 0.3$ and $d = 20$.



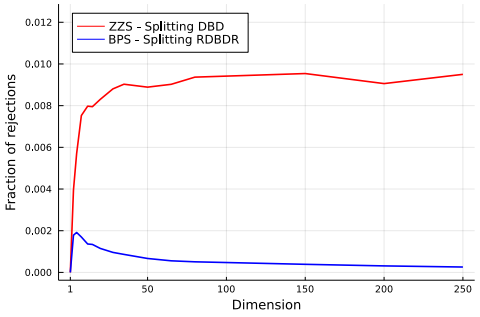
(b) Here $\delta = 0.3$ and $d = 20$.



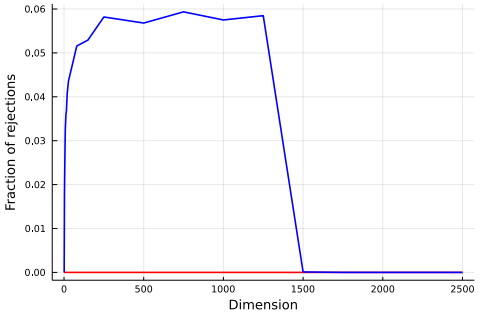
(c) Here $\rho = 0.5$ and $d = 20$.



(d) Here $\sigma^2 = 0.1$ and $d = 20$.

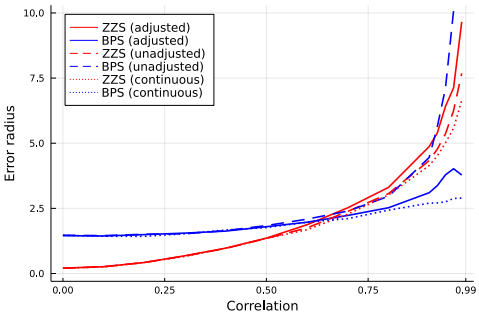


(e) Here $\delta = 0.3$ and $\rho = 0.5$.

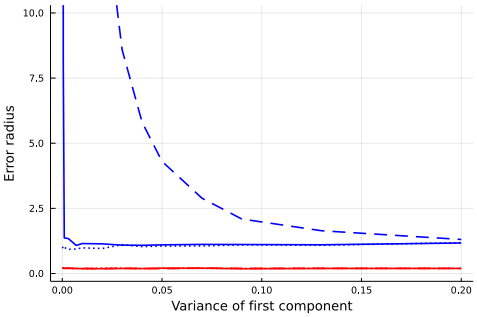


(f) Here $\delta = 0.3$ and $\sigma^2 = 0.1$.

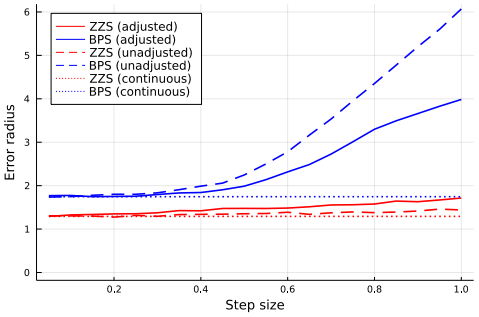
Figure 4.5: Fraction of rejected proposals in the Metropolis step for Algorithms 13 and 14. The plots on the *left* are obtained running the algorithms with a Gaussian target with covariance $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho$ for $j \neq i$, while the plots on the *right* with diagonal covariance $\Sigma_{11} = \sigma^2$, $\Sigma_{ii} = 1$ for $i \neq 1$. In all experiments the refreshment rate for BPS is $\lambda_r = 0.5$.



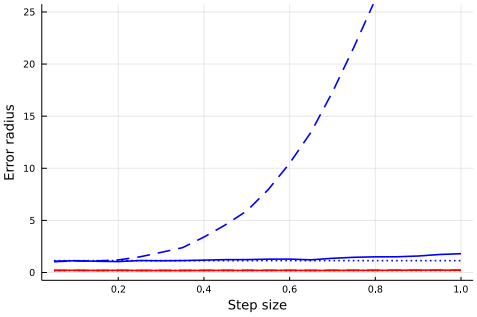
(a) Here $\delta = 0.3$ and $d = 20$.



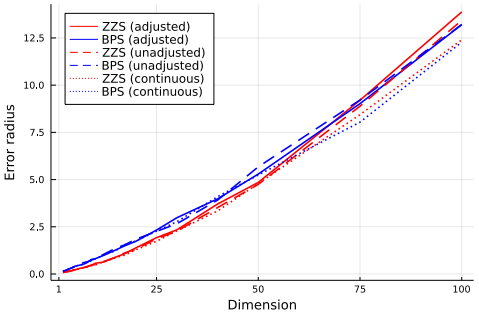
(b) Here $\delta = 0.3$ and $d = 20$.



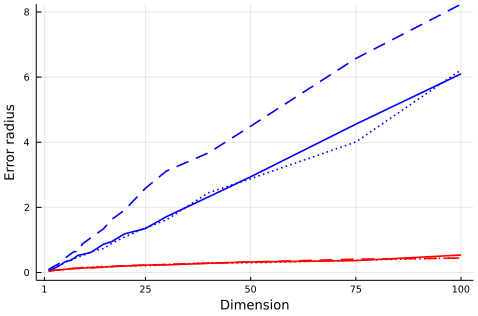
(c) Here $\rho = 0.5$ and $d = 20$.



(d) Here $\sigma^2 = 0.1$ and $d = 20$.



(e) Here $\delta = 0.3$ and $\rho = 0.5$.



(f) Here $\delta = 0.3$ and $\sigma^2 = 0.1$.

Figure 4.6: Error for the radius statistic for Algorithms 11, 12, 13, 14, as well as the continuous ZZS and BPS. The plots on the *left* are obtained running the algorithms with a Gaussian target with covariance $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho$ for $j \neq i$, while the plots on the *right* with diagonal covariance $\Sigma_{11} = \sigma^2$, $\Sigma_{ii} = 1$ for $i \neq 1$. In all experiments the refreshment rate for BPS is $\lambda_r = 0.5$. The processes are started from a draw of the target. The time horizon is $T = 10^3$ and the number of iterations is $N = \lceil T/\delta \rceil$. The radius is estimated with the usual Monte Carlo averages.

adjusted algorithms, as a larger fraction of rejections corresponds to more computations required to obtain an accepted state. What we observe is that the adjusted ZZS defined in Algorithm 13 is exact for targets with diagonal covariance as expected, but the number of rejections increases with the correlation between components of the target. It is well known that the continuous time ZZS has lower efficiency for correlated targets (see [15]), and in the case of Algorithm 13 this is seen as a large number of reflections. On the other hand, the adjusted BPS given by Algorithm 14 appears to suffer when σ^2 is small, while the number of rejections remains controlled for large correlation ρ.

Figure 4.6 shows the error in the estimation of the expected radius for the adjusted and unadjusted algorithms, as well as for the continuous BPS and ZZS. As expected, ZZS is sensitive to high correlation between components and its error increases with ρ. It is possible to improve in these cases by applying the adaptive schemes proposed in [15], which learn the covariance structure of the target and use this information to tune the set of velocities of the ZZS suitably. It seems also clear that the schemes based on ZZS are more robust when the target is very narrow in some components. This is a reasonable behaviour, as **DBD** schemes for ZZS essentially decompose the target in one dimensional problems, hence the chain can explore efficiently some components while being stuck in others. As a consequence, in the second type of Gaussian target the chain will rarely move in the component with small variance, but it can freely move in the other components. On the other hand, in BPS the switching rate and reflection operator are dominated by the component with small variance, thus the whole chain is affected by settings with e.g. small variances of some components. We observe that the adjusted BPS given by Algorithm 14 is more robust than its unadjusted counterpart. For these reasons, Algorithms 11, 13, and 14 are to be preferred in case of stiff targets.

4.5.2 Image deconvolution using a total variation prior

In this section we test Algorithm 13 on an imaging inverse problem, which we solve with a Bayesian approach. In the following we shall refer to an image either as a $N \times N$ matrix or as a vector of length N^2, which is obtained by placing each column of the matrix below the previous one. We set $d = N^2$. In both cases each entry corresponds to a pixel. Now denote as $x \in \mathbb{R}^d$ the image we are interested in estimating and $y \in \mathbb{R}^d$ the observation. The observation is related to x via the statistical model

$$y = Ax + \xi,$$

where A is a $d \times d$-dimensional matrix which may be degenerate and ill-conditioned and ξ a d-dimensional Gaussian random variable with mean zero and variance $\sigma^2 I_d$. The forward problem we consider is given by a blurring operator, i.e. A acts by a discrete convolution with a kernel h. In our examples h will be a uniform blur operator

with blur length either 9 or 25. The likelihood of y given x is given by

$$p(y|x) \propto e^{-f_y(x)},$$
$$f_y(x) = \frac{1}{2\sigma^2}\|Ax - y\|^2.$$

In the Bayesian approach one then has to place a prior distribution on x. Here we choose the total variation prior:

$$p(x) \propto e^{-g(x)},$$

where $\theta > 0$ and $g(x) = \theta\|x\|_{TV}$ is the total variation of the image x (see [142]) and is given by

$$\|x\|_{TV} := \sum_{i,j=1}^{N} (|x_{i+1,j} - x_{i,j}| + |x_{i,j+1} - x_{i,j}|).$$

The total variation prior corresponds to the ℓ^1-norm of the discrete gradient of the image and therefore promotes piecewise constant reconstructions. Note that this prior is not smooth and hence we cannot directly apply the gradient based algorithms such as the unadjusted Langevin algorithm (ULA), BPS or ZZS. Therefore we approximate g with a Moreau-Yosida envelope

$$g^\lambda(x) = \min_{z\in\mathbb{R}^d} \left\{ g(z) + \frac{1}{2\lambda}\|x - z\|^2 \right\}.$$

By [140, Proposition 12.19] we have that g^λ is Lipschitz differentiable with Lipschitz constant λ^{-1} and

$$\nabla g^\lambda(x) = \frac{1}{\lambda}(x - \text{prox}_g^\lambda(x)),$$
$$\text{prox}_g^\lambda(x) = \arg\min_{z\in\mathbb{R}^d} \left\{ g(z) + \frac{1}{2\lambda}\|x - z\|^2 \right\}.$$

Using Bayes theorem, we have the posterior distribution

$$\pi(x) := p(x|y) \propto e^{-f_y(x) - \theta g^\lambda(x)}. \tag{4.25}$$

We select the optimal θ by using the SAPG algorithm [156, 50] and we choose λ based on the guidelines given in [63], which set $\lambda = 1/L_f$ where L_f is the Lipschitz constant of f_y. Sampling from this model using MCMC schemes is difficult because x is usually very high dimensional and the problem is ill-conditioned. In this case the unadjusted Langevin algorithm can be very expensive to run since the step size is limited by $2/L$, where $L = L_f + \lambda^{-1}$ is the Lipschitz constant of $\nabla \log \pi$. Note that we do not consider an unadjusted underdamped Langevin algorithm since this algorithm scales poorly (see [35, 60]) with the conditioning number which is very large in these examples.

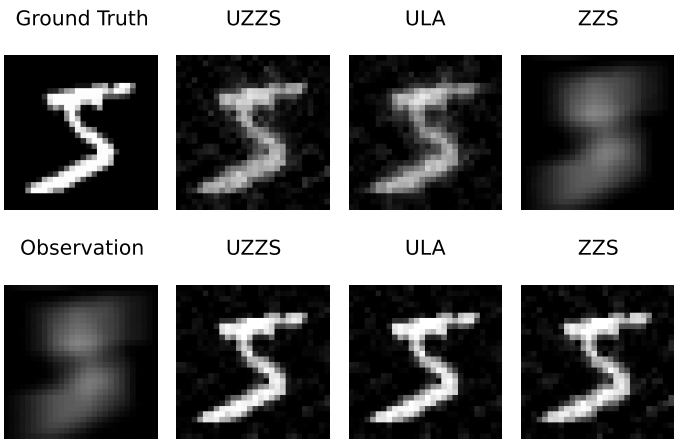


Figure 4.7: Results for the reconstruction of one of the MNIST handwritten digits using a TV prior. The observed image is obtained with a blur length of 9 pixels and then adding Gaussian noise with standard deviation $\sigma = 0.0014$. The step size for ULA is $1.98/L$, for ZZS is $3000/L$, where $L \approx 1042855.81$. Mean after 2×10^3 iterations (*second column*) and after 10^6 iterations (*third column*) of the samplers based on the states at iterations $n \times 10^3$ for $n = 0, \ldots, 100$.

We are now interested in drawing samples from the posterior (4.25), and in particular we compare the unadjusted ZZS (Algorithm 11, abbreviated as UZZS in the plots)), the unadjusted Langevin algorithm (ULA), as well as the continuous ZZS. Indeed, we can compute the Lipschitz constant of the gradient of the negative log-posterior, L, and thus we can implement the exact ZZS using the Poisson thinning technique based on the simple bound

$$\lambda_i(x + vt, v) \leq tL\sqrt{d} + \lambda_i(x, v). \tag{4.26}$$

In order to compare the computational cost of the continuous ZZS to the unadjusted ZZS and to ULA we count each proposal for an event time obtained by Poisson thinning as a gradient evaluation and thus as an iteration. Indeed, an update of the computational bounds requires the evaluation of $\lambda_i(x, v)$ for all $i = 1, \ldots, d$ and thus the full gradient has to be computed. To estimate the posterior mean for the continuous ZZS we compute the time average $T^{-1} \int_0^T X_t dt$.

In Figures 4.7 and 4.9 we show the original images, the observed images after blurring and adding noise, and the estimated posterior mean using the different samplers. Figure 4.8 shows the mean square error (MSE) between the true image and the estimated posterior mean as a function of the number of iterations. The MSE is computed

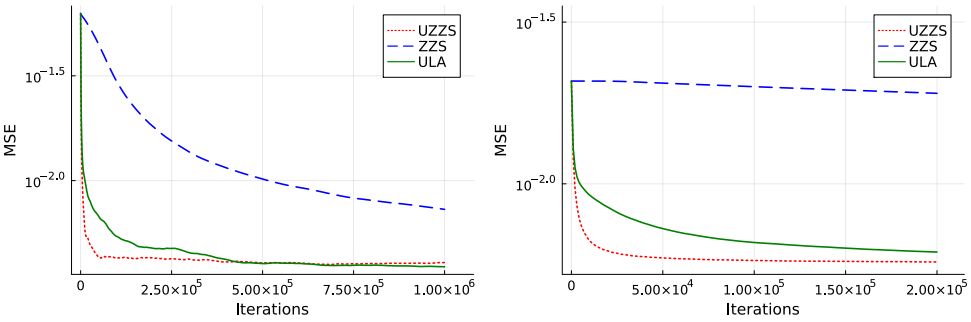


Figure 4.8: Mean square errors as defined in (4.27) for the setting of Figure 4.7 (*left*) and of Figure 4.9 (*right*).

for two images $x, y \in [0,1]^{N\times N}$ as

$$\mathrm{MSE}(x,y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_{ij}-y_{ij})^2. \tag{4.27}$$

It is clear in both Figures 4.7 and 4.9 that the unadjusted ZZS shows the fastest convergence to the posterior mean. This is clear from visual inspection of the reconstructed images, as the reconstruction of ZZS are less noisy after just a few thousand iterations, and even more evident from the MSE shown in Figure 4.8. In the case of Figure 4.7 it appears ZZS has essentially converged after 10^5 iterations, while it takes around 4×10^5 iterations for ULA to obtain a comparable approximation of the posterior mean. This is likely due to the fact that the step size of ULA must be very small or otherwise the process tends to infinity, while for ZZS larger step sizes can be selected. For instance in the context of Figure 4.7 the step size for ULA is approximately 1.9×10^{-6}, which is the largest available value without going over the stability barrier, while for ZZS the step size is 2.9×10^{-3}. This constitutes a major difference because every iteration is very computationally intensive, as the target distribution e.g. for the context of Figure 4.9 is of dimension 65536. Notably each iteration involves solving an optimisation problem, which is solved by the SAPG algorithm. A similar behaviour is observed for the cameraman image shown in Figure 4.9. In this case the unadjusted ZZS needs around 6×10^4 gradient evaluations to converge, while ULA still has not achieved the same accuracy after 2×10^5 iterations. This difference can also be seen from the reconstructions after 2×10^3 iterations shown in Figure 4.9, as indeed the unadjusted ZZS gives a clearly better estimate for the posterior mean. Finally, let us compare the unadjusted ZZS with the continuous time ZZS. It is clear from our experiments that ZZS performs poorly compared to its discretisation. The reason is twofold. First, the major drawback of Poisson thinning using the bounds (4.26) is that a considerable proportion of the proposed event times are rejected (in our examples the rejection rate is around $70-80\%$). Moreover, the rates λ_i are very

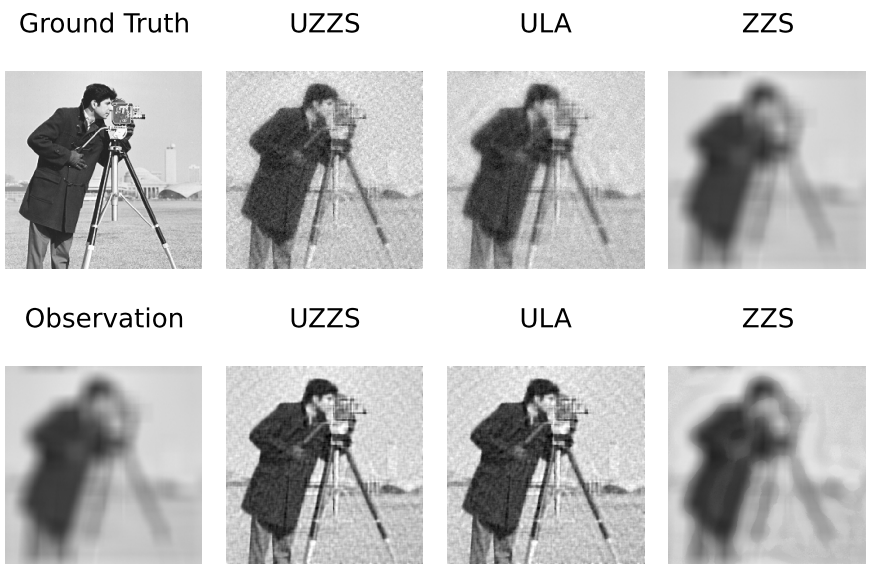


Figure 4.9: Results for the reconstruction of full 256×256 pixels cameraman image using a TV prior. The observed image is obtained with a blur length of 25 pixels and then adding Gaussian noise with standard deviation $\sigma = 0.0021$. The reconstructions show the estimated mean after 2×10^3 iterations (*top row*) and after 2×10^5 iterations (*bottom row*) of the samplers based on the states at iterations $n \times 10^3$ for $n = 0, \dots, 200$. The step size for ULA is $1.98/L$, for ZZS is $1000/L$, where $L \approx 518349.52$.

large in the current framework and the process can have even 10^9 switches per continuous time unit. This means that many gradient computations are required to travel a decent distance and thus the process itself is expensive to run. The combination of these two phenomena implies an important loss of efficiency, which explains the results of our simulations.

4.5.3 Chain of interacting particles

Finally, let us consider a problem which will serve as an illustration of a typical context where ZZS is favored with respect to other samplers. This is a toy model that presents in a simpler form features which are similar to the molecular system considered in [113], where splitting schemes involving velocity bounces have proven efficient. We consider a chain of N particles in 1D, labeled from 1 to N. The particles interact through two potentials: a chain interaction, where the particle i interacts with the particles $i-1$ and $i+1$; and a mean-field interaction, where each particle interacts

with all the others. For $x \in \mathbb{R}^N$, the potential is thus of the form

$$\psi(x) = \sum_{i=1}^{N-1} V(x_i - x_{i+1}) + \frac{a}{N} \sum_{i,j=1}^{N} W(x_i - x_j)\,,$$

where $a > 0$ measures the strength of the mean-field interaction, V is the chain potential and W is the mean-field potential. In the following we take

$$V(s) = s^4, \qquad W(s) = -\sqrt{1+s^2}\,,$$

for $s \in \mathbb{R}$, i.e. the chain interaction is an anharmonic quartic potential which constrains two consecutive particles in the chain to stay close, while the mean-field interaction induces a repulsion from the rest of the system. Although this specific ψ is an academic example meant for illustration purpose, its general form is classical in statistical physics.

Notice that ψ is invariant by translation of the whole system, so that $e^{-\psi}$ is not integrable on $\mathbb{R}^N$. However we are not interested in the behavior of the barycentre $\overline{x} = \frac{1}{N}\sum_{i=1}^N x_i$, so we consider $e^{-\psi}$ as a probability density on the subspace $\{x \in \mathbb{R}^N, \overline{x} = 0\}$, which amounts to looking at the system of particles from its center of mass. Anyway, in practice, we run particles in $\mathbb{R}^N$ without constraining their barycentre to zero, which does not change the output as long as we estimate the expectations of translation-invariant functions. Specifically, here, we consider the empirical variance of the system

$$v(x) = \frac{1}{N^2} \sum_{i,j=1}^{N} (x_i - x_j)^2\ .$$

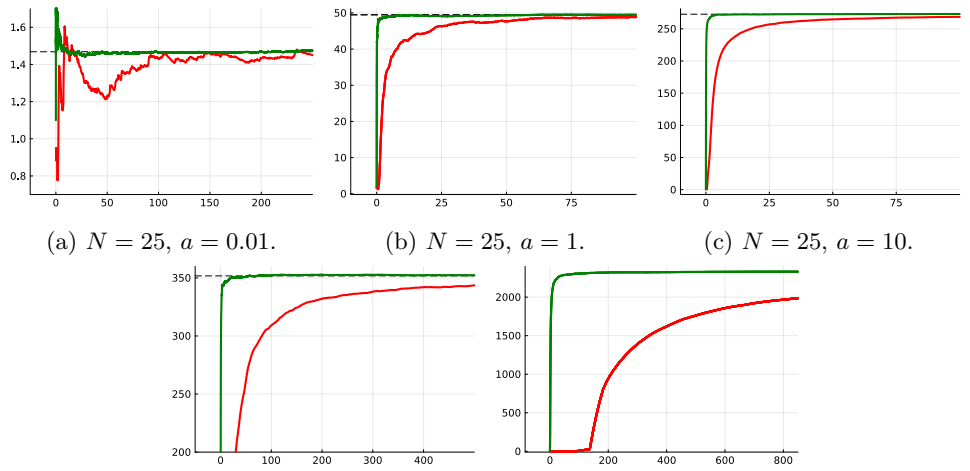
The important points concerning this model (which are typically met in real molecular dynamics as in [113]) are the following. The forces $\nabla\psi$ can be decomposed in two parts, one of which (the chain interaction) is unbounded and not globally Lipschitz but is relatively cheap to compute (with a complexity $O(N)$), while the second part (the mean-field interaction) is bounded but numerically expensive (with a complexity $O(N^2)$). If this decomposition is not taken into account, so that we simply run a classical MCMC sampler based on the computation of $\nabla\psi$, then the step size has to be very small because of the non-Lipschitz part of the forces, and then each step is very costly because of the mean-field force. Besides, due to the non-Lipschitz part, sampling a continuous-time PDMP via thinning would not be very efficient (in fact in this specific simple case it could be possible to design a suitable thinning procedure with some effort, but this would be more difficult with 3D particles and singular potentials such as the Lennard-Jones one [113]).

Now, as was already discussed in Section 4.1.3 for subsampling, PDMPs and their splitting schemes can be used with a splitting of the forces. In the present case, we

consider a ZZS where the switching rate of the i-th velocity is given by

$$\lambda_i(x,v) = (v_i(V'(x_i - x_{i+1}) - V'(x_{i-1} - x_i)))_+ + \frac{a}{N}\sum_{j\neq i}(v_i W'(x_i - x_j))_+ ,$$

where, for the particles 1 and N, we set $x_0 = x_1$ and $x_{N+1} = x_N$ to cancel out the corresponding terms. The corresponding continuous-time ZZS has the correct invariant measure (once centered). We consider the **DBD** splitting to approximate this ZZS (although several other choices are possible, e.g. including the Poisson thinning part in the **D** step and having only jumps according to the potential V in the **B** part). To sample the jump times of the i-th velocity, using that $|W'(s)| \leqslant 1$ for all $s \in \mathbb{R}$, we sample two jump times with rates respectively $(v_i(V'(x_i - x_{i+1}) - V'(x_{i-1} - x_i)))_+$ and a. If both times are larger than the step size δ, then the velocity is not flipped. Else, if the time corresponding to the first rate is smaller than δ and than that corresponding to the second, then we flip the i-th velocity. Alternatively, if the second time is smaller than δ and than the first, we draw $J \sim \mathrm{Unif}(\{1,\dots,N\})$ and we flip the sign of the i-th velocity with probability $(v_i W'(x_i - x_J))_+ / a$ (note that if $J = i$ then this probability is indeed 0). Since in this case the rates are not canonical due to the splitting of forces, this procedure is repeated until there are no events before the end of the time step. This results in $O(1)$ computations per particle on average, hence $O(N)$ for the whole system.



(a) $N = 25$, $a = 0.01$. (b) $N = 25$, $a = 1$. (c) $N = 25$, $a = 10$.

(d) $N = 50$, $a = 1$. (e) $N = 100$, $a = 1$.

Figure 4.10: Empirical variance (on the y-axis) and runtime in seconds (on the x-axis) for various values of N and a in the setting of Section 4.5.3. The *green* line corresponds to the unadjusted ZZS, the *red* line corresponds to HMC, and the *dashed black* line corresponds to the estimated empirical variance with a long run of HMC (when computationally feasible).

We compare this scheme with an HMC sampler implemented in the Julia package [160]. This contains state of the art techniques and implementation of HMC. The results of our simulations are shown in Figure 4.10. Initially, particles are i.i.d. standard Gaussian variables. This gives a configuration which is far from the modes of the target distribution, where particles are organized so that $i \mapsto x_i$ is close to the monotonous profile which minimizes ψ (by the symmetry $i \longleftrightarrow N+1-i$ there are two such modes, but the empirical variance is unchanged by this symmetry so that it is sufficient to see one of them). For both ZZS and HMC, the convergence of the estimator is thus essentially driven by a deterministic motion from the biased initial condition towards a mode. It is clear that our algorithm gives considerably cheaper yet accurate estimates of the empirical variance for all values of a and N considered. This is the result of the subsampling procedure, which reduces the cost per iteration from $O(N^2)$ to $O(N)$, whereas in each iteration of HMC the full mean-field interaction needs to be computed. Notably, as expected the gain in performance increases with the number of particles N, which makes the required runtime of HMC prohibitive for large values of N.

4.A Proofs of Section 4.2

4.A.1 Proof of Theorem 4.10

Proof. Fix $g \in \mathcal{C}^{2,0}_\Phi \cap D(\mathcal{L})$. By a telescoping sum we have

$$\mathbb{E}_z[g(\overline{Z}_{t_n})] - \mathbb{E}_z[g(Z_{t_n})] = \sum_{k=0}^{n-1} (\mathbb{E}_z[\mathcal{P}_{t_n - t_{k+1}} g(\overline{Z}_{t_{k+1}})] - \mathbb{E}_z[\mathcal{P}_{t_n - t_k} g(\overline{Z}_{t_k})]).$$

For each $k \in \{0, \dots, n-1\}$, set $f_k(y,s) = \mathcal{P}_{t_n - t_k - s} g(y)$ then we have

$$\mathbb{E}_z[g(\overline{Z}_{t_n})] - \mathbb{E}_z[g(Z_{t_n})] = \sum_{k=0}^{n-1} \mathbb{E}_z[f_k(\overline{Z}_{t_{k+1}}, \delta) - f_k(\overline{Z}_{t_k}, 0)].$$

By conditioning on $\overline{Z}_{t_k}$ it is sufficient to prove that

$$|\mathbb{E}_z[f_k(\overline{Z}_\delta, \delta)] - f_k(z,0)| \leq R(1+|z|^M) \|g\|_{\mathcal{C}^{2,0}_\Phi} \delta^3. \tag{4.28}$$

Indeed if we have that (4.28) holds then by Assumption 4.9 we have

$$\begin{aligned} |\mathbb{E}_z[g(Z_{t_n})] - \mathbb{E}_z[g(\overline{Z}_{t_n})]| &\leq C\delta^3 \sum_{k=0}^{n-1} e^{R(t_n - t_k)} \mathbb{E}_z[\overline{G}(\overline{Z}_{t_k})] \\ &\leq C \|g\|_{\mathcal{C}^{2,0}_\Phi} e^{R t_n} \delta^3 n \overline{G}_M(z), \end{aligned}$$

which gives the desired result. It remains to show that (4.28) holds.

As done in [16] we rewrite the lhs as

$$\mathbb{E}_z[f_k(\overline{Z}_\delta,\delta)] - f_k(z,0) = \mathbb{E}_z[f_k(\overline{Z}_\delta,\delta)] - f_k(\varphi_\delta(z),\delta) + f_k(\varphi_\delta(z),\delta) - f_k(z,0). \tag{4.29}$$

In particular, with identical steps to [16] we can rewrite the last two terms on the left hand side of (4.29) using the fundamental theorem of calculus and that $\partial_s f_k(z,s) = -\mathcal{L}f_k(z,s)$:

$$f_k(\varphi_\delta(z),\delta) - f_k(z,0) = -\int_0^\delta \lambda(\varphi_r(z))[Q(f_k(\cdot,r))(\varphi_r(z)) - f_k(\varphi_r(z),r)]\mathrm{d}r.$$

Then we compute the expectation in the right hand side of (4.29), collecting a term for the case of no jumps, a single jump and the case of multiple jumps

$$\begin{aligned}
&\mathbb{E}_z[f_k(\overline{Z}_\delta,\delta)] - f_k(z,\delta) = \\
&= \iint_0^\delta Q(\varphi_{\delta/2}(z),\mathrm{d}\tilde{z})(f_k(\varphi_{\delta/2}(\tilde{z}),\delta) - f_k(\varphi_\delta(z),\delta))\lambda(\varphi_{\delta/2}(z))e^{-s\lambda(\varphi_{\delta/2}(z))-(\delta-s)\lambda(\tilde{z})}\mathrm{d}s \qquad (\dagger)\\
&\quad + \sum_{\ell=2}^\infty \mathbb{E}_z[(f_k(\overline{Z}_\delta,\delta) - f_k(\varphi_\delta(z),\delta))\mathbb{1}_{\{\ell \text{ events}\}}] \qquad (\ddagger)\\
&\quad - \int_0^\delta \lambda(\varphi_r(z))[Q(f_k(\cdot,r))(\varphi_r(z)) - f_k(\varphi_r(z),r)]\mathrm{d}r. \qquad (\ddagger\dagger)
\end{aligned}$$

Observe that the sum in the second term ($\ddagger$) can be truncated from $\ell = 3$ onward as we only wish to get an order δ^3 local error. Indeed, we have $|f_k| \le \|g\|_\infty$ and hence

$$\begin{aligned}
&\left|\sum_{\ell=3}^\infty \mathbb{E}_z[(f_k(\overline{Z}_\delta,\delta) - f_k(\varphi_\delta(z),\delta))\mathbb{1}_{\{\ell \text{ events}\}}]\right| \le 2\|g\|_\infty \mathbb{P}_z(\ell \ge 3 \text{ events})\\
&\le 2\|g\|_\infty \int\int_0^\delta \int_0^{\delta-s_1}\int_0^{\delta-s_1-s_2} \lambda(\varphi_{\delta/2}(z))e^{-s_1\lambda(\varphi_{\delta/2}(z))}\lambda(z_1)e^{-s_2\lambda(z_1)}\lambda(z_2)e^{-s_3\lambda(z_2)}\\
&\quad Q(\varphi_{\delta/2}(z),\mathrm{d}z_1)Q(z_1,\mathrm{d}z_2)Q(z_2,\mathrm{d}z_3)\,\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}s_3\\
&\le 2\|g\|_\infty \int (1-e^{-\delta\lambda(\varphi_{\delta/2}(z))})(1-e^{-\delta\lambda(z_1)})(1-e^{-\delta\lambda(z_2)})Q(\varphi_{\delta/2}(z),\mathrm{d}z_1)Q(z_1,\mathrm{d}z_2)\\
&\le 2\delta^3\|g\|_\infty \int \lambda(\varphi_{\delta/2}(z))\lambda(z_1)\lambda(z_2)Q(\varphi_{\delta/2}(z),\mathrm{d}z_1)Q(z_1,\mathrm{d}z_2)\\
&\le 2\delta^3\|g\|_\infty \int \lambda(\varphi_{\delta/2}(z))\lambda(z_1)Q(\varphi_{\delta/2}(z),\mathrm{d}z_1)Q\lambda(z_1)\\
&\le 2\delta^3\|g\|_\infty \lambda(\varphi_{\delta/2}(z))Q(\lambda(\cdot)Q\lambda(\cdot))(\varphi_{\delta/2}(z))
\end{aligned}$$

where we used that $1-\exp(-z) \le z$ for $z \ge 0$. By Assumption 4.7 we have that $\lambda Q(\lambda Q\lambda)$ is polynomially bounded and therefore we can bound ($\ddagger$) by

$$(\ddagger) = \int Q(\varphi_{\delta/2}(z),\mathrm{d}z_1)Q(z_1,\mathrm{d}z_2)\Big(f_k(\varphi_{\delta/2}(z_2),\delta) - f_k(\varphi_\delta(z),\delta)\Big).$$

$$\cdot \int_0^\delta \int_0^{\delta - s_1} \lambda(\varphi_{\delta/2}(z)) e^{-s_1 \lambda(\varphi_{\delta/2}(z))} \lambda(z_1) e^{-s_2 \lambda(z_1)} e^{-(\delta - s_1 - s_2)\lambda(z_2)} \mathrm{d}s_2 \, \mathrm{d}s_1$$
$$+ \mathcal{O}(\|g\|_\infty (1 + |z|^{3m_\lambda}) \delta^3).$$

Here and throughout we understand $F(z, \delta, g) = \mathcal{O}(\|g\|_{\mathcal{C}^2_\Phi}(1 + |z|^m)\delta^n)$ to mean that

$$\limsup_{\delta \to 0} \sup_{z \in E} \sup_g \frac{|F(z, \delta, g)|}{\|g\|_{\mathcal{C}^2_\Phi} \delta^n (1 + |z|^m)} \leq C.$$

We Taylor expand several terms in order to verify that the local error is of order δ^3. We use repeatedly the following expansions:

$$\lambda(\varphi_s(z)) = \lambda(z) + s D_\Phi \lambda(z) + s^2 R(z, \tilde{s}; \lambda),$$
$$f_k(\varphi_s(z), \delta) = f_k(z, \delta) + s D_\Phi f_k(z, \delta) + s^2 R(z, \tilde{s}; f_k),$$
$$R(z, \tilde{s}; g) = D^2_\Phi g(\varphi_{\tilde{s}}(z))/2,$$
$$f_k(z, s) = f_k(z, 0) - s \mathcal{L} f_k(z, 0) + s^2 \mathcal{L}^2 f_k(z, \tilde{s})/2,$$

for some $\tilde{s} \in [0, s]$ (note that $\tilde{s}$ may vary with each term so when we use these expansions we include an index to distinguish different incidents of $\tilde{s}$). Note that by Assumption 4.8 we have

$$\|f_k\|_{\mathcal{C}^2_\Phi} \leq C e^{R(t_n - t_k)} (1 + |z|^{m_\mathcal{P}}) \|g\|_{\mathcal{C}^{2,0}_\Phi}$$

which gives us a bound on the remainder terms. Applying the expansions above to (†) we obtain

$$\begin{aligned}
(\dagger) = &\int Q(\varphi_{\delta/2}(z), \mathrm{d}\tilde{z}) \Big(f_k(\tilde{z}, 0) - \delta \mathcal{L} f_k(\tilde{z}, 0) + \frac{\delta^2}{2} \mathcal{L}^2 f_k(\tilde{z}, \tilde{s}_2) + \frac{\delta}{2} D_\Phi f_k(\tilde{z}, \delta) \\
&+ \frac{1}{8} \delta^2 R(\tilde{z}, \tilde{s}_2; f_k) - f_k(z, 0) + \delta \mathcal{L} f_k(z, 0) - \frac{1}{2} \delta^2 \mathcal{L}^2 f_k(z, \tilde{s}_3) - \delta D_\Phi f_k(z, \delta) \\
&- \frac{1}{2} \delta^2 R(z, \tilde{s}_3; f_k) \Big) \int_0^\delta \Big(\lambda(z) + \frac{\delta}{2} D_\Phi \lambda(z) + \frac{1}{8} \delta^2 R(z, \tilde{s}_4; \lambda) \Big) \\
&\Big(1 - s\lambda(\varphi_{\delta/2}(z)) - (\delta - s)\lambda(\tilde{z}) + \frac{1}{2} (s\lambda(\varphi_{\delta/2}(z)) + (\delta - s)\lambda(\tilde{z}))^2 e^{-\eta} \Big) \mathrm{d}s \\
= &\int Q(\varphi_{\delta/2}(z), \mathrm{d}\tilde{z}) \Big(f_k(\tilde{z}, 0) - f_k(z, 0) \Big) \int_0^\delta \Big(\lambda(z) + \frac{\delta}{2} D_\Phi \lambda(z) \Big) \times \\
&\quad \times \Big(1 - s\lambda(z) - (\delta - s)\lambda(\tilde{z}) \Big) \mathrm{d}s \\
&+ \delta \int Q(\varphi_{\delta/2}(z), \mathrm{d}\tilde{z}) \Big(- \mathcal{L} f_k(\tilde{z}, 0) + \frac{1}{2} D_\Phi f_k(\tilde{z}, 0) + \mathcal{L} f_k(z, 0) - D_\Phi f_k(z, 0) \Big) \\
&\int_0^\delta \Big(\lambda(z) + \frac{\delta}{2} D_\Phi \lambda(z) \Big) \Big(1 - s\lambda(z) - (\delta - s)\lambda(\tilde{z}) \Big) \mathrm{d}s
\end{aligned}$$

$$+ e^{R(t_n - t_k)} \mathcal{O}(\|g\|_{\mathcal{C}^{2,0}_\Phi}(1 + |z|^M)\delta^3)$$

where we used

$$\eta \in [0, s\lambda(\varphi_{\delta/2}(z)) + (\delta - s)\lambda(\tilde{z})]$$

in the first equality and further Taylor expansions to obtain the second equality. Here $M = 3m_\lambda + m_\mathcal{P}$. Now using Assumption 4.7 we can expand the Q term

$$\begin{aligned}
(\dagger) = &\int \Big(Q(z, \mathrm{d}\tilde{z}) + \frac{\delta}{2} D_\Phi Q(z, \mathrm{d}\tilde{z})\Big)\Big(f_k(\tilde{z}, 0) - f_k(z, 0)\Big) \\
&\int_0^\delta \Big(\lambda(z) + \frac{\delta}{2} D_\Phi \lambda(z)\Big)\Big(1 - s\lambda(z) - (\delta - s)\lambda(\tilde{z})\Big)\mathrm{d}s \\
&+ \delta \int Q(z, \mathrm{d}\tilde{z})\Big(- \mathcal{L} f_k(\tilde{z}, 0) + \frac{1}{2} D_\Phi f_k(\tilde{z}, 0) + \mathcal{L} f_k(z, 0) - D_\Phi f_k(z, 0)\Big) \\
&\int_0^\delta \Big(\lambda(z) + \frac{\delta}{2} D_\Phi \lambda(z)\Big)\Big(1 - s\lambda(z) - (\delta - s)\lambda(\tilde{z})\Big)\mathrm{d}s \\
&+ e^{R(t_n - t_k)} \mathcal{O}(\|g\|_{\mathcal{C}^{2,0}_\Phi}(1 + |z|^M)\delta^3)
\end{aligned}$$

Term (‡) can be expanded as

$$\begin{aligned}
(\ddagger) = &\int Q(\varphi_{\delta/2}(z), \mathrm{d}\tilde{z}) Q(z_1, \mathrm{d}z_2)\Big(f_k(z_2, 0) - f_k(z, 0)\Big) \\
&\int_0^\delta \int_0^{\delta - s_1} \big(\lambda(z) + \delta D_\Phi(\lambda)(\varphi_{\tilde{s}_4}(z))\big)\lambda(z_1) \\
&\Big(1 + (-s_1\lambda(\varphi_{\delta/2}(z)) - s_2\lambda(z_1) - (\delta - s_1 - s_2)\lambda(z_2))e^{-\xi}\Big)\mathrm{d}s_2\,\mathrm{d}s_1 \\
&+ e^{R(t_n - t_k)} \mathcal{O}(\|g\|_{\mathcal{C}^{1,0}_\Phi}(1 + |z|^{m_\mathcal{P} + 3m_\lambda})\delta^3) \\
= &\frac{\delta^2}{2} \int Q(\varphi_{\delta/2}(z), \mathrm{d}\tilde{z}) Q(z_1, \mathrm{d}z_2)\Big(f_k(z_2, 0) - f_k(z, 0)\Big)\lambda(z)\lambda(z_1) \\
&+ e^{R(t_n - t_k)} \mathcal{O}(\|g\|_{\mathcal{C}^{1,0}_\Phi}(1 + |z|^{m_\mathcal{P} + 3m_\lambda})\delta^3) \\
= &\frac{\delta^2}{2} \int Q(z, \mathrm{d}z_1) Q(z_1, \mathrm{d}z_2)\Big(f_k(z_2, 0) - f_k(z, 0)\Big)\lambda(z)\lambda(z_1) \\
&+ e^{R(t_n - t_k)} \mathcal{O}(\|g\|_{\mathcal{C}^{1,0}_\Phi}(1 + |z|^{m_\mathcal{P} + 3m_\lambda})\delta^3),
\end{aligned}$$

where

$$\xi \in [0, s_1\lambda(\varphi_{\delta/2}(z)) + s_2\lambda(z_1) + (\delta - s_1 - s_2)\lambda(z_2)].$$

By Assumption 4.7 we can expand the term (‡†) as follows:

$$(\ddagger\dagger) = -\int_0^\delta \lambda(\varphi_r(z))\left(Q f_k(\cdot, r)(z) + r \int D_\Phi Q(z, \mathrm{d}\tilde{z}) f_k(\tilde{z}, r) - f_k(\varphi_r(z), r)\right)\mathrm{d}r$$

$$+ e^{R(t_n-t_k)}\mathcal{O}(\delta^3\|g\|_{C^{2,0}_\Phi}(1+|z|^{m_\lambda})).$$

By Taylor's theorem

$$\begin{aligned}(\ddagger\dagger) = -\int_0^\delta \lambda(\varphi_r(z))\Big(Qf_k(\cdot,0)(z) - r\int Q(z,\mathrm{d}\tilde{z})\mathcal{L}f_k(\tilde{z},0) + r\int D_\Phi Q(z,\mathrm{d}\tilde{z})(f_k(\tilde{z},0)\\ - r\mathcal{L}f_k(\tilde{z},\tilde{r})) - (f_k(z,0) + rD_\Phi f_k(z,0) - r\mathcal{L}f_k(z,0))\Big)\mathrm{d}r\\ + e^{R(t_n-t_k)}\mathcal{O}(\delta^3\|g\|_{C^{2,0}_\Phi}(1+|z|^{m_\lambda+m_\mathcal{P}}).\end{aligned}$$

Note that

$$\int D_\Phi Q(z,\mathrm{d}\tilde{z})\mathcal{L}f_k(\tilde{z},\tilde{r}) = Q(D_\Phi\mathcal{L}f_k(\cdot,\tilde{r}))(z) = e^{R(t_n-t_k)}\mathcal{O}((1+|z|^{m_\mathcal{P}})\|g\|_{\mathcal{C}^{2,0}_\Phi}).$$

Using this and Taylor expanding $\lambda(\varphi_r(z))$ we have

$$\begin{aligned}(\ddagger\dagger) = -\int_0^\delta \lambda(z)\Big(Qf_k(\cdot,0)(z) - r\int Q(z,\mathrm{d}\tilde{z})\mathcal{L}f_k(\tilde{z},0) + r\int D_\Phi Q(z,\mathrm{d}\tilde{z})f_k(\tilde{z},0)\\ - (f_k(z,0) + rD_\Phi f_k(z,0) - r\mathcal{L}f_k(z,0))\Big)\mathrm{d}r - \int_0^\delta rD_\Phi\lambda(z)\left(Qf_k(\cdot,0)(z) - f_k(z,0)\right)\mathrm{d}r\\ + e^{R(t_n-t_k)}\mathcal{O}(\delta^3\|g\|_{C^{2,0}_\Phi}(1+|z|^{m_\lambda+m_\mathcal{P}}).\end{aligned}$$

Evaluating the integral over r

$$\begin{aligned}(\ddagger\dagger) = -\lambda(z)\Bigg(Qf_k(\cdot,0)(z)\delta - \frac{1}{2}\delta^2\int Q(z,\mathrm{d}\tilde{z})\mathcal{L}f_k(\tilde{z},0) + \frac{1}{2}\delta^2\int D_\Phi Q(z,\mathrm{d}\tilde{z})f_k(\tilde{z},0)\\ - \Big(f_k(z,0)\delta + \frac{1}{2}\delta^2 D_\Phi f_k(z,0) - \frac{1}{2}\delta^2\mathcal{L}f_k(z,0)\Big)\Bigg)\\ - \frac{1}{2}\delta^2 D_\Phi\lambda(z)\left(Qf_k(\cdot,0)(z) - f_k(z,0)\right) + e^{R(t_n-t_k)}\mathcal{O}(\delta^3\|g\|_{C^{2,0}_\Phi}(1+|z|^{m_\lambda+m_\mathcal{P}}).\end{aligned}$$

First order terms. Terms of order δ appear only in (†) and (‡†) and clearly they cancel out.

Second order terms. In (†) we can further expand terms of the form $D_\Phi(f)(\tilde{z},\delta)$ and rearrange as

$$\begin{aligned}\text{Order } \delta^2 \text{ of } (\dagger) = \delta^2\Bigg[\int \lambda(z)\Big(-Q(z,\mathrm{d}\tilde{z})\mathcal{L}f_k(\tilde{z},0) + Q(z,\mathrm{d}\tilde{z})\frac{1}{2}D_\Phi f_k(\tilde{z},0)\\ + \frac{1}{2}D_\Phi Q(z,\mathrm{d}\tilde{z})(f_k(\tilde{z},0) - f_k(z,0))\Big)\\ + \lambda(z)(\mathcal{L}f_k(z,0) - D_\Phi f_k(z,0))\Big)\end{aligned}$$

$$
\begin{aligned}
&+\int Q(z,\mathrm{d}\tilde{z})(f_k(\tilde{z},0)-f_k(z,0))\Big(-\frac{1}{2}\lambda(z)(\lambda(\tilde{z})+\lambda(z))+\frac{1}{2}D_\Phi\lambda(z)\Big)\Big]\\
=&\delta^2\Big[\int\lambda(z)\Big(-Q(z,\mathrm{d}\tilde{z})\mathcal{L}f_k(\tilde{z},0)+Q(z,\mathrm{d}\tilde{z})\frac{1}{2}D_\Phi f_k(\tilde{z},0)\rangle\\
&+\underbrace{\frac{1}{2}D_\Phi Q(z,\mathrm{d}\tilde{z})(f_k(\tilde{z},0)-f_k(z,0))}_{\text{Term }A}\Big)\\
&+\underbrace{\lambda(z)\Big(\mathcal{L}f_k(z,0)-D_\Phi f_k(z,0))-\frac{1}{2}\lambda(z)\int Q(z,\mathrm{d}\tilde{z})(f_k(\tilde{z},0)-f_k(z,0))\Big)}_{\text{Term }B}\\
&+\underbrace{\frac{1}{2}D_\Phi\lambda(z)\int Q(z,\mathrm{d}\tilde{z})(f_k(\tilde{z},0)-f_k(z,0))}_{\text{Term }C}\\
&-\frac{1}{2}\int Q(z,\mathrm{d}\tilde{z})(f_k(\tilde{z},0)-\underbrace{f_k(z,0)}_{\text{Term }D})\lambda(z)\lambda(\tilde{z})\Big].
\end{aligned}
$$

For term (‡) we have

$$
\text{Order }\delta^2\text{ of }(\ddagger)=\frac{1}{2}\delta^2\int Q(z,\mathrm{d}z_1)Q(z_1,\mathrm{d}z_2)(f_k(z_2,0)-\underbrace{f_k(z,0)}_{\text{Term }D})\lambda(z)\lambda(z_1).
$$

Similarly for (‡†) we have

$$
\begin{aligned}
\text{Order }\delta^2\text{ of }(\ddagger\dagger)=-\frac{\delta^2}{2}\Big(&\underbrace{\int D_\Phi(Q)(z)\lambda(z)(f_k(\tilde{z},0)-f_k(z,0))}_{\text{Term }A}\\
&+\underbrace{\int Q(z,\mathrm{d}\tilde{z})D_\Phi(\lambda)(z)(f_k(\tilde{z},0)-f_k(z,0))}_{\text{Term }C}\\
&+\int Q(z,\mathrm{d}\tilde{z})\lambda(z)(-\mathcal{L}f_k(\tilde{z},0)+\underbrace{\mathcal{L}f_k(z,0)-D\Phi(f_k)(z,0)}_{\text{Term }B})\Big).
\end{aligned}
$$

After cancellations we obtain

$$
\begin{aligned}
&\text{Order }\delta^2\text{ of }(\dagger)+(\ddagger)+(\ddagger\dagger)=\\
&=\delta^2\lambda(z)\Big(\int\Big(-Q(z,\mathrm{d}\tilde{z})\mathcal{L}f_k(\tilde{z},0)+Q(z,\mathrm{d}\tilde{z})\frac{1}{2}D_\Phi(f_k)(\tilde{z},0)\Big)\\
&\quad-\frac{1}{2}\int Q(z,\mathrm{d}\tilde{z})f_k(\tilde{z},0)\lambda(\tilde{z})+\frac{1}{2}\int Q(z,\mathrm{d}\tilde{z})Q(\tilde{z},\mathrm{d}z_2)f_k(z_2,0)\lambda(\tilde{z})
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{2}\int Q(z,\mathrm{d}\tilde{z})\mathcal{L}f_k(\tilde{z},0)\Big)\\
&= \frac{1}{2}\delta^2\lambda(z)\Bigg(\int Q(z,\mathrm{d}\tilde{z})\Big(-\mathcal{L}f_k(\tilde{z},0) + D_\Phi(f_k)(\tilde{z},0)\\
&\quad + \lambda(\tilde{z})\int Q(\tilde{z},\mathrm{d}z_2)[f_k(z_2,0) - f_k(\tilde{z},0)]\Big)\Bigg)\\
&= 0,
\end{aligned}
$$

where the last equality follows by the definition of the generator of the PDMP. Therefore we have shown that second order terms cancel out.

□

4.A.2 Proofs for Example 4.11

Let us verify that Assumption 4.7 holds. Note that

$$(D_\Phi Q)(g)(x,v) = v^T \sum_{i=1}^{d} \nabla_x \left(\frac{\lambda_i(x,v)}{\lambda(x,v)} g(x, F_i v)\right).$$

Therefore by Taylor's theorem we have for some η between x and $x+sv$

$$Qg(x+sv,v) - Qg(x,v) - (D_\Phi Q)(g)(x,v) = \frac{1}{2}s^2 \sum_{i=1}^{d} v^T \nabla_x^2 \left(\frac{\lambda_i(x,v)}{\lambda(x,v)} g(x,F_i v)\right)\Big|_{x=\eta} v.$$

It remains to show that $\frac{\lambda_i}{\lambda}, \nabla_x\left(\frac{\lambda_i(x,v)}{\lambda(x,v)}\right), \nabla_x^2\left(\frac{\lambda_i(x,v)}{\lambda(x,v)}\right)$ are bounded. It is clear that $0 < \lambda_i/\lambda \le 1$. Let us consider the first derivative. Set $\Xi(s) = \log(1+e^s)$ so that $\lambda_i(x,v) = \Xi(v_i\partial_i\psi(x))$ and note that

$$0 \le \frac{\Xi'(s)}{\Xi(s)} \le 1, \quad 0 \le \frac{\Xi''(s)}{\Xi(s)} \le 1.$$

Now

$$
\begin{aligned}
\left|\nabla_x\left(\frac{\lambda_i(x,v)}{\lambda(x,v)}\right)\right| &= \left|\frac{\nabla_x\lambda_i(x,v)}{\lambda(x,v)} - \frac{\lambda_i(x,v)\nabla_x\lambda(x,v)}{\lambda(x,v)^2}\right|\\
&\le \left|\frac{\nabla_x\lambda_i(x,v)}{\lambda(x,v)}\right| + \sum_{j=1}^{d}\left|\frac{\nabla_x\lambda_j(x,v)}{\lambda(x,v)}\right|.
\end{aligned}
$$

So it remains to show that $\nabla_x\lambda_i/\lambda$ is bounded. Using the bounds on Ξ we have

$$\left|\frac{\nabla_x\lambda_i(x,v)}{\lambda(x,v)}\right| \le \left|\frac{\nabla_x\lambda_i(x,v)}{\lambda_i(x,v)}\right| = \left|\frac{\Xi'(v_i\partial_i\psi(x))}{\Xi(v_i\partial_i\psi(x))} v_i \nabla_x\partial_i\psi(x)\right| \le |\nabla_x\partial_i\psi(x)|.$$

This is bounded by our assumptions on ψ. Let us now consider $\nabla_x^2 \left(\frac{\lambda_i(x,v)}{\lambda(x,v)} \right)$:

$$\left| \nabla_x^2 \left(\frac{\lambda_i(x,v)}{\lambda(x,v)} \right) \right| = \left| \frac{\nabla_x^2 \lambda_i(x,v)}{\lambda(x,v)} - \frac{2\nabla_x \lambda_i(x,v) \nabla_x \lambda(x,v)^T}{\lambda(x,v)^2} + \frac{2\lambda_i(x,v) \nabla_x \lambda(x,v) \nabla_x \lambda(x,v)^T}{\lambda(x,v)^3} - \frac{\lambda_i(x,v) \nabla_x^2 \lambda(x,v)}{\lambda(x,v)^2} \right|.$$

Using the bound on $\nabla_x \lambda_i / \lambda$ we can bound all the terms aside from the one involving the second derivative of λ so it suffices to consider

$$\begin{aligned} &\left| \frac{\nabla_x^2 \lambda_i(x,v)}{\lambda(x,v)} \right| = \\ &= \left| \frac{\Xi''(v_i \partial_i \psi(x))}{\sum_j \Xi(v_j \partial_j \psi(x))} \nabla_x \partial_i \psi(x) \nabla_x \partial_i \psi(x)^T + \frac{\Xi'(v_i \partial_i \psi(x))}{\sum_j \Xi(v_j \partial_j \psi(x))} v_i \nabla_x^2 \partial_i \psi(x) \right| \\ &\leq \left| \nabla_x \partial_i \psi(x) \nabla_x \partial_i \psi(x)^T \right| + \left| \nabla_x^2 \partial_i \psi(x) \right| \end{aligned}$$

This is bounded since ψ has bound second and third order derivatives.

4.B Ergodicity for splitting schemes of the BPS

We have the following theorem, which implies Theorem 4.15.

Theorem 4.25. *Consider any scheme of the BPS based on the decomposition* ***D,R,B****. Under Assumption 4.14, the following hold:*

1. *There exist $a, b, C, \delta_0 > 0$ and a function V (with a, b, C, δ_0, V depending only on ψ and λ_r, but not on δ) such that, for all x, v,*

$$e^{|x|/a}/a \leqslant V(x,v) \leqslant a e^{a|x|}$$

and for all $\delta \in (0, \delta_0]$ and all x, v,

$$P_\delta V(x,v) \leqslant (1 - b\delta) V(x,v) + C\delta .$$

2. *For all $R > 0$, there exist $c, \delta_0 > 0$ and a probability measure ν on E such that for all x, v with $|x| \leqslant R$ and all $\delta \in (0, \delta_0]$, setting $n_* = \lceil 4R/\delta \rceil$,*

$$\delta_{x,v} P_\delta^{n_*} \geqslant c\nu .$$

The proof is omitted from the thesis, but can be found in the paper on which this chapter is based [17].

4.C Ergodicity for splitting schemes of ZZS

4.C.1 Splitting DBD

In this Section we focus on splitting scheme DBD for ZZS. We shall prove the following, which implies Theorem 4.17.

Theorem 4.26. *Consider the splitting scheme **DBD** for ZZS. Suppose Assumption 4.16 holds. Then the following hold:*

1. *There exists a function $V : \mathbb{R}^d \times \{\pm 1\}^d \to \mathbb{R}$, and constants $b \in (0,1), C < \infty$ such that for all (x,v) and all $\delta \in (0, \delta_0)$ with $\delta_0 = 2(1+\gamma_0)^{-1}$, where γ_0 is as in Assumption 4.16(b), it holds that*

$$P_\delta V(x,v) \leq (1 - b\delta)V(x,v) + C\delta. \tag{4.30}$$

2. *For any $R > 0$ consider a set $C = [-R, R]^d$. For some $L > 0$ let*

$$n^* = 2 + \frac{4x_0 + 2R}{\delta} + 2\left\lceil \frac{L}{\delta} \right\rceil \in 2\mathbb{N}.$$

For $(x,v) \in C \times \{\pm 1\}^d$ define the set $D(x,v)$ given by (4.15). Then for any $(y,w) \in D(x,v) \cap (C \times \{\pm 1\}^d)$ and $\delta \in (0, \delta_0]$ for $\delta_0 > 0$ it holds that

$$\delta_{y,w} P_\delta^{n^*} \geqslant c\nu\,.$$

where c is independent of δ and ν is uniform over $D(x,v) \cap (C \times \{\pm 1\}^d)$.

In Section 4.C.1.1 we prove the minorisation condition, in Section 4.C.1.2 we prove the drift condition, while in Section 4.C.1.3 we prove Equation (4.18).

4.C.1.1 Minorisation condition

We now prove a minorisation condition for splitting scheme DBD of ZZS. In the following Lemma we consider the one-dimensional setting, for which the reasoning is similar to that of the proof of a minorisation condition for the continuous ZZS done in [15, Lemma B.2].

Lemma 4.27. *Consider the splitting scheme DBD of ZZS with step size $\delta \leq \delta_0$. Suppose Assumption 4.16(a) holds for some $x_0 \geq 0$ and consider a set $C = [-R, R]$ for $R > 0$. For $L > 0$ let*

$$N = 2 + \frac{4x_0 + 2R}{\delta} + 2\left\lceil \frac{L}{\delta} \right\rceil \in 2\mathbb{N}. \tag{4.31}$$

For $(x,v) \in C \times \{\pm 1\}$ define the set $D(x,v) := D_+(x,v) \cup D_-(x,v)$, where

$$\begin{aligned} D_+(x,v) &:= \{(y,w) : w = v,\ y = x + m\delta,\ m \in 2\mathbb{Z}\}, \\ D_-(x,v) &:= \{(y,w) : w = -v,\ y = x + m\delta,\ m \in 2\mathbb{Z} + 1\}. \end{aligned} \tag{4.32}$$

Then for any $(y,w) \in D(x,v) \cap (C \times \{\pm 1\})$ *it holds that*

$$\mathbb{P}_{(y,w)}((X_N, V_N) \in \cdot) \geq b\nu(\cdot)$$

where b *is independent of* δ *and* ν *is uniform over* $D(x,v) \cap (C \times \{\pm 1\})$.

Proof. Let $C = [-R, R]$ for a fixed $R > 0$ and let $x \in C$. We shall consider only the case of $v = +1$, as the same arguments extend to the symmetric case $v = -1$. In particular observe that if the process is started in set $D(x,+1)$ (respectively $D(x,-1)$), then after an even number of iterations it will again be in $D(x,+1)$ ($D(x,-1)$). This means that the process lives on $D(x,+1)$ (respectively on $D(x,-1)$). To shorten the notation we denote by D_+, D_- the sets $D_+(x,+1), D_-(x,+1)$ as defined in (4.32). Below we focus on the case of an initial condition in D_+, while the case of D_- follows with an identical reasoning and obvious changes.

Fix $N \in 2\mathbb{N}$ and define

$$\overline{\lambda} := \max_{x \in C} \max_{y : |y-x| \leq N\delta, v = \pm 1} \lambda(y,v)$$

which is the largest switching rate that can be reached within N iterations starting in C. Note that taking N as in (4.31) implies that $N\delta$ is upper bounded by a constant as $\delta \leq \delta_0$ and thus $\overline{\lambda}$ can be chosen independently of the step size δ. Recall $\underline{\lambda} > 0$.

From here on we shall denote the initial condition as $(y,w) \in D_+$, and without loss of generality we shall assume $x_0 = x + \ell\delta$ for some $\ell \in \mathbb{N}$, where x_0 is as in Assumption 4.16(a). We want to lower bound the probability that after N iterations the process is in measurable sets $\overline{B} \subset D$. We consider two cases: in the first one the final state of the process is of the form $(X_N, V_N) = (z, -1) \in D_- \cap (C \times \{\pm 1\})$, while in the second case $(X_N, V_N) = (z, +1) \in D_+ \cap (C \times \{\pm 1\})$.

First case

Consider the case in which the final state has negative velocity, i.e. $V_N = -1$. To lower bound the probability of reaching this state, we consider the case in which only one switching event takes place. Let $z = y + m\delta$ with $m \in \mathbb{N}$ odd. Then in order to have $(X_N, V_N) = (z, -1)$ with exactly one event taking place at time N_1 it must be that

$$y + (N_1 - 1)\delta - (N - N_1)\delta = z.$$

Thus we find that the event should take place at

$$N_1 = \frac{z - y}{2\delta} + \frac{N+1}{2}.$$

In order to guarantee the switching rate is strictly positive it must also be that $X_{N_1} \geq x_0$, i.e. $y + (N_1 - 1)\delta \geq x_0$ and thus $N_1 \geq 1 + (x_0 - y)/\delta$. Note $N_1 < N$, where N is as in (4.31). Denote the position at the time of the switching event by

$\tilde{x} = y + \delta(N_1 - 1/2)$. Then probability of exactly one event taking place at iteration N_1 is given by

$$\int_0^\delta \lambda(\tilde{x}, 1) \exp(-s\lambda(\tilde{x}, 1)) \exp(-(\delta - s)\lambda(\tilde{x}, -1))ds \geq \delta \underline{\lambda} \exp(-\delta\overline{\lambda}).$$

The probability of this path is simple to lower bound, since upper bounding the switching rates gives a smaller probability:

$$\begin{aligned}
\mathbb{P}_{(y,+1)}((X_N, V_N) = (z, -1)) &\geq \underbrace{\prod_{n=0}^{N_1-1} \exp(-\delta\lambda(y + (n + 1/2)\delta))}_{\text{no jumps before } N_1} \times \underbrace{\delta\underline{\lambda}\exp(-\delta\overline{\lambda})}_{\text{a jump at } N_1} \times \\
&\quad \times \underbrace{\prod_{n=0}^{N-N_1} \exp(-\delta\lambda(y + (N_1 - 1 - n)\delta))}_{\text{no jumps after } N_1} \\
&\geq \exp(-(N_1 - 1)\delta\overline{\lambda})\ \delta\underline{\lambda}\exp(-\delta\overline{\lambda})\ \exp(-(N - N_1)\delta\overline{\lambda}) \\
&\geq 2\exp(-(N - 1)\delta\overline{\lambda})\ \underline{\lambda}\exp(-\delta_0\overline{\lambda}) \times \left(\frac{1}{2}\frac{\delta M}{M}\right) \\
&\geq 2\exp(-(N - 1)\delta\overline{\lambda})\ \underline{\lambda}\exp(-\delta_0\overline{\lambda})(2R - \delta)\frac{\nu(-1)}{M},
\end{aligned}$$

where $M \in \mathbb{N}$ is the number of points in $D_+ \cap (C \times \{\pm 1\})$. In the last line we used that $\delta M \geq 2R - \delta$. Recall that $\delta \leq \delta_0$ and that N is given by (4.31). This concludes as $(N - 1)\delta \leq 4x_0 + R + 2L + 3\delta_0$ and $2R - \delta \geq 2R - \delta_0$.

Second case

We now focus on the case in which $V_N = +1$. We shall find an appropriate lower bound by restricting to the case in which exactly two switching events take place. Denoting the times of the two events as N_1, N_2, if the final position is z it must be

$$y + (N_1 - 1)\delta - (N_2 - 1)\delta + (N - N_1 - N_2)\delta = z$$

which implies

$$N_2 = \frac{y - z}{2\delta} + \frac{N}{2}. \tag{4.33}$$

Moreover, at event times the process should be in regions with strictly positive switching rate:

$$\begin{aligned}
&y + (N_1 - 1/2)\delta \geq x_0, \\
&y + (N_1 - 1)\delta - (N_2 - 1/2)\delta \leq -x_0.
\end{aligned}$$

These imply respectively

$$N_1 \geq \frac{x_0 - y}{\delta} + 1 =: \underline{N}_1,$$

$$N_2 \geq \frac{y + x_0}{\delta} + N_1.$$

Since N_2 is determined by (4.33), we enforce that the second inequality holds:

$$\frac{y - z}{2\delta} + \frac{N}{2} \geq \frac{y + x_0}{\delta} + N_1$$

which implies

$$N_1 \leq \frac{N}{2} - \frac{y + 2x_0 + z}{2\delta} =: \overline{N}_1.$$

Now to obtain the right dependence on δ, we shall take N such that $\overline{N}_1 - \underline{N}_1$ is increasing as $1/\delta$. It holds

$$\overline{N}_1 - \underline{N}_1 = \frac{N - 2}{2} - \frac{4x_0 - y + z}{2\delta}$$

and thus it is sufficient to take

$$N = 2 + \frac{4x_0 - y + z}{\delta} + 2\left\lceil \frac{L}{\delta} \right\rceil$$

for some constant $L > 0$, as with this choice $\overline{N}_1 - \underline{N}_1 = \lceil L/\delta \rceil$.

Using the results above we find

$$\begin{aligned}
&\mathbb{P}_{(y,+1)}((X_N, V_N) = (z, +1)) \geq \\
&\geq \sum_{N_1=\underline{N}_1}^{\overline{N}_1} \Bigg[\underbrace{\prod_{n=0}^{N_1-1} \exp(-\delta\lambda(y + (n + 1/2)\delta))}_{\text{no jumps before } N_1} \times \underbrace{\delta\underline{\lambda}\exp(-\delta\overline{\lambda})}_{\text{a jump at } N_1} \times \\
&\quad \times \underbrace{\prod_{m=0}^{N_2-1} \exp(-\delta\lambda(y + (N_1 - 1 - (m + 1/2))\delta))}_{\text{no jumps until } N_2} \times \underbrace{\delta\underline{\lambda}\exp(-\delta\overline{\lambda})}_{\text{a jump at } N_2} \\
&\quad \times \underbrace{\prod_{\ell=0}^{N-N_1-N_2} \exp(-\delta\lambda(y + (N_1 - 1 - (N_2 - 1) + (\ell + 1/2))\delta))}_{\text{no jumps after } N_2} \Bigg] \\
&\geq \sum_{N_1=\underline{N}_1}^{\overline{N}_1} \exp(-\delta\overline{\lambda}N\delta))\delta^2\underline{\lambda}^2\exp(-2\delta\overline{\lambda}) \\
&= \left\lceil \frac{L}{\delta} \right\rceil \exp(-\overline{\lambda}N\delta))\delta^2\underline{\lambda}^2\exp(-2\delta\overline{\lambda}) \\
&\geq L\exp(-\delta\overline{\lambda}N))\underline{\lambda}^2\exp(-2\delta_0\overline{\lambda})\delta
\end{aligned}$$

$$\geq 2L\exp(-\delta\overline{\lambda}N))\underline{\lambda}^2\exp(-2\delta_0\overline{\lambda})(2R-\delta_0)\left(\nu(+1)\times\frac{1}{M}\right).$$

Similarly to above it is now sufficient to note that $N\delta \leq 4\delta_0 + 4x_0 + 2R + 2L$.

Conclusion

To conclude it is sufficient to observe that the conditions above hold for any choice of $x, y, z \in C$ since N is as in (4.31). □

Multidimensional case

To extend to the higher dimensional setting, first observe that it is possible to apply the same ideas in the proof of Lemma 4.27 to each component, in particular requiring that the events happen when all components of the process are outside of the rectangle $[-x_0, +x_0]$. This implies that Assumption 4.16(a) can be used to lower bound the probability of flipping each component of the velocity vector. Hence each coordinate can be controlled independently of the others. It is clear that the following minorisation condition is implied: let $C = [-R, R]^d$ for $R > 0$, $(x, v) \in C \times \{\pm 1\}^d$, and let $D(x, v)$ as in (4.15); then for all $(y, w) \in (x, v)$ it holds that

$$\mathbb{P}_{(y,w)}((X_N, V_N) \in \cdot) \geq b^d\ \nu_d(\cdot),$$

where N, b are as in Lemma 4.27 and ν_d is the uniform distribution over states in the grid $D(x, v) \cap (C \times \{\pm 1\}^d)$.

4.C.1.2 Drift condition

Let us first characterise in the following Lemma the law of the jump part of the process. This result is then used to prove the wanted drift condition in Lemma 4.29 below.

Lemma 4.28. *Let $\tilde{V}_t^x$ denote the PDMP corresponding to the generator $\mathcal{L}_2$ (for this process x acts as a parameter). Suppose that $\lambda_i(x, v)$ is independent of v_j for $j \neq i$. Then for any $w \in \{\pm 1\}^d$ we have*

$$\mathbb{P}_v(\tilde{V}_t^x = w) = \prod_{i=1}^{d} \frac{\lambda_i(x, F_i w) + \frac{w_i}{v_i}\lambda_i(x, v)e^{-(\lambda_i(x,v)+\lambda_i(x,F_i v))t}}{\lambda_i(x, v) + \lambda_i(x, F_i v)}.$$

Proof of Lemma 4.28. To simplify notation we will suppress the dependence on x and set $\Lambda_i(v) = \lambda_i(v) + \lambda_i(-v) = \lambda_i(x, v) + \lambda_i(x, F_i v)$. Since $\tilde{V}_t^x$ jumps according to λ_i which does not depend on v_j we have that the coordinates of $\tilde{V}_t^x$ are all independent. Hence it is sufficient to show

$$\mathbb{P}_{v_i}((\tilde{V}_t^x)^i = w_i) = \frac{\lambda_i(-w_i) + \frac{w_i}{v_i}\lambda_i(v_i)e^{-\Lambda_i(v_i)t}}{\Lambda_i(v_i)}.$$

Therefore it is sufficient to consider the setting $d = 1$. Define for any $t \geq 0, v, w \in \{1, -1\}$

$$\varphi_t(v; w) := \frac{\lambda(-w) + \frac{w}{v}\lambda(v)e^{-\Lambda(v)t}}{\Lambda(v)}.$$

If we show that for all $t \geq 0, v, w \in \{1, -1\}$

$$\partial_t \varphi_t(v; w) = \mathcal{L}_2 \varphi_t(v; w), \quad \varphi_0(v; w) = \mathbb{1}_w(v), \tag{4.34}$$

then φ_t coincides with the semigroup applied to $\mathbb{1}_w$ and we have the desired result

$$\varphi_t(v; w) = \mathbb{E}_v[\varphi_0(\tilde{V}_t; w)] = \mathbb{P}_v[\tilde{V}_t = w].$$

It is straightforward to confirm the initial condition $\varphi_0(v; w) = \mathbb{1}_w(v)$ holds. So it remains to show that φ_t satisfies the PDE (4.34). Note that

$$\begin{aligned}
\partial_t \varphi_t(v; w) &= -\frac{w}{v}\lambda(v)e^{-\Lambda(v)t} \\
\mathcal{L}_2 \varphi_t(v; w) &= \lambda(v)\left(\varphi_t(-v; w) - \varphi_t(v; w)\right) \\
&= \lambda(v)\left(\frac{\lambda(-w) - \frac{w}{v}\lambda(-v)e^{-\Lambda(v)t}}{\Lambda(v)} - \frac{\lambda(-w) + \frac{w}{v}\lambda(v)e^{-\Lambda(v)t}}{\Lambda(v)}\right) \\
&= -\lambda(v)\left(\frac{\frac{w}{v}\lambda(-v)e^{-\Lambda(v)t}}{\Lambda(v)} + \frac{\frac{w}{v}\lambda(v)e^{-\Lambda(v)t}}{\Lambda(v)}\right) = -\lambda(v)\frac{w}{v}e^{-\Lambda(v)t}.
\end{aligned}$$

Therefore we have that (4.34) holds. □

Lemma 4.29. *Consider the splitting scheme **DBD** of ZZS. Let $\lambda_i(x, v) = (v_i \partial_i \psi(x))_+ + \gamma_i(x)$ and let Assumption 4.16 be verified. Then there exists a function $V : \mathbb{R} \times \{\pm 1\}^d \to \mathbb{R}$, and constants $\rho \in (0, 1), C < \infty$ such that for all (x, v) and all $t \in (0, t_0)$ with $t_0 < (1 + \gamma_0)^{-1}$*

$$\overline{P}_t V(x, v) = P_t^D P_{2t}^B P_t^D V(x, v) \leq (1 - \rho t)V(x, v) + Ct. \tag{4.35}$$

Proof. For a function $g(x, v)$ conditioning on the event $v = w$ and using Lemma 4.28 we have

$$\begin{aligned}
\overline{P}_t g(x, v) = &\sum_{w \in \{\pm 1\}^d} g(x + vt + wt, w) \\
&\prod_{i=1}^{d} \left[\frac{\lambda_i(x + vt, F_i w) + \frac{w_i}{v_i}\lambda_i(x + vt, v)e^{-(\lambda_i(x+vt,v)+\lambda_i(x+vt,F_i v))t}}{\lambda_i(x + vt, v) + \lambda_i(x + vt, F_i v)}\right].
\end{aligned}$$

We now construct our Lyapunov function V. Let $\beta \in (0, 1/2)$, define

$$\phi(s) = \frac{1}{2}\text{sign}(s)\ln(1 + 2|s|)$$

and

$$V(x,v) = \exp\left(\beta\psi(x) + \sum_{i=1}^{d}\phi(v_i\partial_i\psi(x))\right).$$

This is the same Lyapunov function defined in [24]. For this function we have

$$\frac{\overline{P}_t V(x,v)}{V(x,v)} = \sum_{w\in\{\pm1\}^d}\frac{V(x+vt+wt,w)}{V(x,v)} \prod_{i=1}^{d}\left[\frac{\lambda_i(x+vt,F_iw)+\frac{w_i}{v_i}\lambda_i(x+vt,v)e^{-(\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv))t}}{\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv)}\right]. \tag{4.36}$$

By Taylor's theorem there exists $\overline{x}_1 = \overline{x}_1(x,v,w,t) \in B(x,t\sqrt{d})$ such that

$$\psi(x+vt+wt) = \psi(x) + t\langle v+w, \nabla\psi(x)\rangle + \frac{t^2}{2}(v+w)^T\nabla^2\psi(\overline{x}_1)(v+w).$$

Therefore we can rewrite (4.36) as

$$\frac{\overline{P}_t V(x,v)}{V(x,v)} = \sum_{w\in\{\pm1\}^d} e^{\frac{t^2}{2}(v+w)^T\nabla^2\psi(\overline{x}_1)(v+w)} \prod_{i=1}^{d} e^{t(v_i+w_i)\beta\partial_i\psi(x)+\phi(w_i\partial_i\psi(x+vt+wt))-\phi(v_i\partial_i\psi(x))} \times\left[\frac{\lambda_i(x+vt,F_iw)+\frac{w_i}{v_i}\lambda_i(x+vt,v)e^{-(\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv))t}}{\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv)}\right]. \tag{4.37}$$

Since $|\phi'(s)| \le 1$ for all s, by Taylor's Theorem we have

$$\phi(w_i\partial_i\psi(x+vt+wt)) - \phi(v_i\partial_i\psi(x)) \le |w_i\partial_i\psi(x+vt+wt) - w_i\partial_i\psi(x)| + \phi(w_i\partial_i\psi(x)) - \phi(v_i\partial_i\psi(x)).$$

Then we can write

$$\frac{\overline{P}_t V(x,v)}{V(x,v)} \le \sum_{w\in\{\pm1\}^d} K_1\prod_{i=1}^{d} I(i) \tag{4.38}$$

with

$$K_1 = e^{\frac{t^2}{2}(v+w)^T\nabla^2\psi(\overline{x}_1)(v+w)}e^{\sum_{i=1}^{d}|w_i\partial_i\psi(x+vt+wt)-w_i\partial_i\psi(x)|}$$

$$I(i) = e^{t(v_i+w_i)\beta\partial_i\psi(x)+\phi(w_i\partial_i\psi(x))-\phi(v_i\partial_i\psi(x))}\times \times\frac{\lambda_i(x+vt,F_iw)+\frac{w_i}{v_i}\lambda_i(x+vt,v)e^{-(\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv))t}}{\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv)}.$$

Bound outside of a compact set

We split the product in (4.38) into four cases: (i) $w_i = v_i$ and $v_i\partial_i\psi(x+vt) > 0$; (ii) $w_i = v_i$ and $v_i\partial_i\psi(x+vt) < 0$; (iii) $w_i = -v_i$ and $v_i\partial_i\psi(x+vt) > 0$; (iv) $w_i = -v_i$ and $v_i\partial_i\psi(x+vt) < 0$.

Consider first case (i). Let i be such that $w_i = v_i$ and $v_i\partial_i\psi(x+vt) > 0$. Then

$$I(i) = e^{2tv_i\beta\partial_i\psi(x)}\frac{\lambda_i(x+vt,F_iv)+\lambda_i(x+vt,v)e^{-(\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv))t}}{\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv)}.$$

Using the form of λ_i we can write this as

$$\begin{aligned} I(i) = {} & \frac{\gamma_i(x+vt)e^{2\beta tv_i\partial_i\psi(x)}(1-e^{-(|\partial_i\psi(x+vt)|+2\gamma_i(x+vt))t})}{|\partial_i\psi(x+vt)|+2\gamma_i(x+vt)} \\ & + e^{-(|\partial_i\psi(x+vt)|+2\gamma_i(x+vt))t+2tv_i\beta\partial_i\psi(x)}. \end{aligned}$$

Using that $1-e^{-z} \le z$ for all $z > 0$ (note we make use of this inequality several times in the following computations) we find

$$I(i) \le \gamma_i(x+vt)e^{2\beta tv_i\partial_i\psi(x)}t + e^{-(|\partial_i\psi(x+vt)|+2\gamma_i(x+vt))t+2tv_i\beta\partial_i\psi(x)}.$$

By Assumption 4.16(b) we can bound γ_i for $|x| \ge R$ with R sufficiently large and we have $v_i\partial_i\psi(x+vt) \ge 1+\gamma_i(x+vt)$ which gives

$$\begin{aligned} I(i) &\le \frac{1+(v_i\partial_i\psi(x+vt)+\gamma_i(x+vt))}{|\partial_i\psi(x+vt)|+2\gamma_i(x+vt)}e^{-|\partial_i\psi(x+vt)|t+2tv_i\beta\partial_i\psi(x)} \\ &\le 2e^{-|\partial_i\psi(x+vt)|t+2tv_i\beta\partial_i\psi(x)}. \end{aligned}$$

For case (ii), let i be such that $w_i = v_i$ and $v_i\partial_i\psi(x+vt) < 0$. Then

$$\begin{aligned} I(i) &= e^{2\beta tv_i\partial_i\psi(x)}\frac{|\partial_i\psi(x+vt)|+\gamma_i(x+vt)+\gamma_i(x+vt)e^{-(|\partial_i\psi(x+vt)|+2\gamma_i(x+vt))t}}{|\partial_i\psi(x+vt)|+2\gamma_i(x+vt)} \\ &\le e^{2\beta tv_i\partial_i\psi(x)}. \end{aligned}$$

For case (iii), let i be such that $w_i = -v_i$ and $v_i\partial_i\psi(x+vt) > 0$. Then

$$I(i) = e^{\phi(-v_i\partial_i\psi(x))-\phi(v_i\partial_i\psi(x))}\frac{\lambda_i(x+vt,v)-\lambda_i(x+vt,v)e^{-(\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv))t}}{\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv)}.$$

For $s > 0$ it holds that $\phi(-s)-\phi(s) = -\ln(1+2s)$ and hence

$$I(i) = \frac{\lambda_i(x+vt,v)}{1+2v_i\partial_i\psi(x)}\frac{1-e^{-(\lambda_i(x+vt,v)+\lambda_i(x+vt,-v))t}}{\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv)} \le \frac{\lambda_i(x+vt,v)}{1+2v_i\partial_i\psi(x)}t.$$

For case (iv), let i be such that $w_i = -v_i$ and $v_i\partial_i\psi(x+vt) < 0$. Then

$$I(i) = e^{\phi(-v_i\partial_i\psi(x))-\phi(v_i\partial_i\psi(x))}\frac{\gamma_i(x+vt) - \gamma_i(x+vt)e^{-(\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv))t}}{\lambda_i(x+vt,v)+\lambda_i(x+vt,F_iv)}.$$

For $s<0$ we have $\phi(-s)-\phi(s) = \ln(1+2|s|)$, and thus we obtain

$$I(i) \le \gamma_i(x+vt)(1+2|\partial_i\psi(x)|)t.$$

Combining these estimates we have for $|x| \ge R$ with R sufficiently large

$$\begin{aligned}\frac{\overline{P}_tV(x,v)}{V(x,v)} &\le \sum_{w\in\{\pm1\}^d} K_1 \prod_{i:w_i=v_i,\, v_i\partial_i\psi>0} (\gamma_i(x+vt)e^{2\beta t v_i\partial_i\psi(x)}t \\ &+ e^{-(|\partial_i\psi(x+vt)|+2\gamma_i(x+vt))t+2tv_i\beta\partial_i\psi(x)}) \prod_{i:w_i=v_i,\, v_i\partial_i\psi<0} e^{2\beta t v_i\partial_i\psi(x)} \\ &\times \prod_{i:w_i=-v_i,\, v_i\partial_i\psi>0} \frac{\lambda_i(x+vt,v)}{1+2v_i\partial_i\psi(x)}t \prod_{i:w_i=-v_i,\, v_i\partial_i\psi<0} \gamma_i(x+vt)(1+2|\partial_i\psi(x)|)t.\end{aligned}$$

Now consider K_1. By Taylor's theorem there exists $\overline{x}_2 \in B(x, 2\sqrt{dt})$ such that

$$K_1 \le \exp\left(\sum_{i=1}^d \left(\frac{t^2}{2}|((v+w)^T\nabla^2\psi(\overline{x}_1))_i| + t|(w\nabla^2\psi(\overline{x}_2))_i|\right)|v_i+w_i|\right).$$

Using this bound and the four cases above we now obtain

$$\begin{aligned}\frac{\overline{P}_tV(x,v)}{V(x,v)} &\le \prod_{i:v_i\partial_i\psi>0} e^{2t^2|(v^T\nabla^2\psi(\overline{x}_1))_i|+\frac{2t}{2}|(v\nabla^2\psi(\overline{x}_2))_i|} \\ &\prod_{i:v_i\partial_i\psi<0} e^{\left(\frac{t^2}{2}(2|v^T\nabla^2\psi(\overline{x}_1))_i|+2t|(v\nabla^2\psi(\overline{x}_2))_i|\right)+2\beta t v_i\partial_i\psi(x)} \\ &\times (\gamma_i(x+vt)e^{2\beta t v_i\partial_i\psi(x)}t + e^{-(1-2\beta)|\partial_i\psi(x+vt)|t-2\gamma_i(x+vt)t}) \\ &+ \sum_{w\in\{\pm1\}^d\setminus\{v\}} t^{|\{i:w_i\ne v_i\}|} \prod_{i:w_i=v_i,v_i\partial_i\psi>0} e^{\left(t^2(|(v+w)^T\nabla^2\psi(\overline{x}_1))_i|+2t|(w\nabla^2\psi(\overline{x}_2))_i|\right)} \\ &\times (\gamma_i(x+vt)e^{2\beta t v_i\partial_i\psi(x)}t + e^{-(1-2\beta)|\partial_i\psi(x+vt)|t-2\gamma_i(x+vt)t}) \\ &\times \prod_{i:w_i=v_i,v_i\partial_i\psi<0} e^{\left(t^2(|(v+w)^T\nabla^2\psi(\overline{x}_1))_i|+2t|(w\nabla^2\psi(\overline{x}_2))_i|\right)}e^{2\beta t v_i\partial_i\psi(x)} \\ &\times \prod_{i:w_i=-v_i,v_i\partial_i\psi>0} \frac{\lambda_i(x+vt,v)}{1+2v_i\partial_i\psi(x)} \prod_{i:w_i=-v_i,v_i\partial_i\psi<0} e^{-t_0|\partial_i\psi(x+vt)|}(1+2|\partial_i\psi(x)|).\end{aligned}$$

By (4.17) we have

$$\frac{\overline{P}_tV(x,v)}{V(x,v)} \le$$

$$\leq \prod_{i:v_i\partial_i\psi>0} (\gamma_0 t + e^{2t^2|(v^T\nabla^2\psi(\overline{x}_1))_i|+\frac{2t}{2}|(v\nabla^2\psi(\overline{x}_2))_i|} e^{-(1-2\beta)|\partial_i\psi(x+vt)|t-2\gamma_i(x+vt)t})$$
$$\times \prod_{i:v_i\partial_i\psi<0} e^{\left(\frac{t^2}{2}(2|v^T\nabla^2\psi(\overline{x}_1))_i|+2t|(v\nabla^2\psi(\overline{x}_2))_i|\right)+2\beta t v_i\partial_i\psi(x)}$$
$$+ \sum_{w\in\{\pm1\}^d\setminus\{v\}} t^{|\{i:w_i\neq v_i\}|}$$
$$\times \prod_{i:w_i=v_i,v_i\partial_i\psi>0} (\gamma_0 t + e^{\left(t^2(|(v+w)^T\nabla^2\psi(\overline{x}_1))_i|+2t|(w\nabla^2\psi(\overline{x}_2))_i|\right)} e^{-(1-2\beta)|\partial_i\psi(x+vt)|t-2\gamma_i(x+vt)t})$$
$$\times \prod_{i:w_i=v_i,v_i\partial_i\psi<0} e^{\left(t^2(|(v+w)^T\nabla^2\psi(\overline{x}_1))_i|+2t|(w\nabla^2\psi(\overline{x}_2))_i|\right)} e^{2\beta t v_i\partial_i\psi(x)}$$
$$\times \prod_{i:w_i=-v_i,v_i\partial_i\psi>0} \frac{\lambda_i(x+vt,v)}{1+2v_i\partial_i\psi(x)} \prod_{i:w_i=-v_i,v_i\partial_i\psi<0} e^{-t_0|\partial_i\psi(x+vt)|}(1+2|\partial_i\psi(x)|).$$

Since $\beta < 1/2$, by Assumption 4.16(c) there exists β_1 such that for $|x| \geq R$ with R sufficiently large

$$\frac{\overline{P}_t V(x,v)}{V(x,v)} \leq \prod_{i:v_i\partial_i\psi>0} (\gamma_0 t + e^{-\beta_1|\partial_i\psi(x+vt)|t}) \prod_{i:v_i\partial_i\psi<0} e^{-\beta_1|\partial_i\psi(x)|t}$$
$$+ \sum_{w\in\{\pm1\}^d\setminus\{v\}} t^{|\{i:w_i\neq v_i\}|} \prod_{i:w_i=v_i,v_i\partial_i\psi>0} (\gamma_0 t + e^{-\beta_1|\partial_i\psi(x+vt)|t}) \prod_{i:w_i=v_i,v_i\partial_i\psi<0} e^{-\beta_1 t|\partial_i\psi(x)|}$$
$$\times \prod_{i:w_i=-v_i,v_i\partial_i\psi>0} \frac{\lambda_i(x+vt,v)}{1+2v_i\partial_i\psi(x)} \prod_{i:w_i=-v_i,v_i\partial_i\psi<0} e^{-t_0|\partial_i\psi(x+vt)|}(1+2|\partial_i\psi(x)|).$$

For $|x| \geq R$ with R sufficiently large $\lambda_i(x+vt,v)/(1+2v_i\partial_i\psi(x)) \leq 1$ and by (4.17) we have $\gamma_i(x)(1+2|\partial_i\psi(x)|) \leq 1$. We also have that $|\nabla\psi(x+vt)| \geq M$ for any $M > 0$ for $|x| \geq R$ with R sufficiently large. Then we have

$$\frac{\overline{P}_t V(x,v)}{V(x,v)} \leq (\gamma_0 t + e^{-\beta_1 Mt})^{|\{i:v_i\partial_i\psi(x+vt)>0\}|} e^{-\beta_1 Mt|\{i:v_i\partial_i\psi(x+vt)<0\}|}$$
$$+ \sum_{w\in\{\pm1\}^d\setminus\{v\}} t^{|\{i:w_i\neq v_i\}|} \prod_{i:w_i=v_i,v_i\partial_i\psi>0} (\gamma_0 t + e^{-\beta_1 Mt}) \prod_{i:w_i=v_i,v_i\partial_i\psi<0} e^{-\beta_1 tM}.$$

Since $e^{-\beta_1 Mt} \leq \gamma_0 t + e^{-\beta_1 Mt}$ we obtain

$$\frac{\overline{P}_t V(x,v)}{V(x,v)} \leq (\gamma_0 t + e^{-\beta_1 Mt})^{|\{i:v_i\partial_i\psi(x+vt)>0\}|} (\gamma_0 t + e^{-\beta_1 Mt})^{|\{i:v_i\partial_i\psi(x+vt)<0\}|}$$
$$+ \sum_{w\in\{\pm1\}^d\setminus\{v\}} t^{|\{i:w_i\neq v_i\}|} (\gamma_0 t + e^{-\beta_1 Mt})^{|\{i:w_i=v_i,v_i\partial_i\psi(x+vt)>0\}|} \times$$
$$\times (\gamma_0 t + e^{-\beta_1 Mt})^{|\{i:w_i=v_i,v_i\partial_i\psi(x+vt)<0\}|}.$$

Hence

$$\begin{aligned}\frac{\overline{P}_t V(x,v)}{V(x,v)} &\le (\gamma_0 t + e^{-\beta_1 M t})^d + \sum_{w\in\{\pm1\}^d:\, w\neq v} t^{|\{i:w_i\neq v_i\}|}(\gamma_0 t + e^{-\beta_1 M t})^{d-|\{i:w_i\neq v_i\}|} \\ &= \sum_{w\in\{\pm1\}^d} t^{|\{i:w_i\neq v_i\}|}(\gamma_0 t + e^{-\beta_1 M t})^{d-|\{i:w_i\neq v_i\}|} \\ &= \sum_{k=0}^{d}\binom{d}{k} t^k\,(\gamma_0 t + e^{-\beta_1 M t})^{d-k} \\ &= \big((1+\gamma_0)t + e^{-\beta_1 M t}\big)^d.\end{aligned}$$

To show that (4.35) holds for $|x| \ge R$ it is sufficient to show that $(1+\gamma_0)t + e^{-\beta_1 M t} < 1-\rho t$ for some $\rho > 0$. Indeed in that case $1-\rho t < 1$ and thus

$$((1+\gamma_0)t + e^{-\beta_1 M t})^d < (1-\rho t)^d < 1-\rho t.$$

Note that for $t \le t_0$, with $t_0 \in [0,1]$, it holds that $e^{-\beta_1 M t} \le 1 - ct$ for $c = \frac{1-e^{-\beta_1 M t_0}}{t_0}$. Then for $t \le t_0$ we have

$$(1+\gamma_0)t + e^{-\beta_1 M t} \le 1 - t(c - 1 - \gamma_0).$$

Then it is needed that $c > 1+\gamma_0$, that is t_0 should be such that

$$\frac{1-e^{-\beta_1 M t_0}}{t_0} > 1+\gamma_0. \tag{4.39}$$

Note we can always increase M by taking R larger. Choose M such that $e^{-\beta_1 M t_0} < 1 - t_0(1+\gamma_0)$, which is possible since $t_0 < (1+\gamma_0)^{-1}$, then (4.39) holds. Hence (4.35) holds for $|x| \ge R$ with $\rho = (1-e^{-\beta_1 M t_0})t_0^{-1} - 1 - \gamma_0$.

Bound inside of a compact set

It remains to show that (4.35) holds for $|x| \le R$. Let $C = \{x : |x| \le R\} \times \{\pm1\}^d$. Recall $t < 1$ and $\psi \in \mathcal{C}^2$. We shall use the inequality $e^{tr} \leqslant 1 + tr + t^2 r^2 e^r/2 \le 1 + t(r + e^{3r}/2)$, which holds for for $t \leqslant 1, r > 0$. First of all we consider the term in the sum corresponding to the case $w = v$. Bounding the probability of this event by 1 we find

$$\begin{aligned}K_1 \prod_{i=1}^d I(i) &\le e^{t^2 2v^T\nabla^2\psi(\overline{x}_1)v + \frac{2}{2}\sum_{i=1}^d |\partial_i\psi(x+2vt) - \partial_i\psi(x)| + 2t\beta\langle v, \nabla\psi(x)\rangle} \\ &\le 1 + t(A(x,v,t) + e^{3A(x,v,t)}/2),\end{aligned}$$

where $A(x,v,t) = 2v^T\nabla^2\psi(\overline{x}_1)v + (2/2)\sum_{i=1}^d |\langle v, \nabla\partial_i\psi(\overline{x}_2)\rangle| + 2\beta\langle v, \nabla\psi(x)\rangle$. Taking the maximum of A over $(x,v,t) \in C \times \{\pm1\}^d \times (0,1)$ we find

$$K_1 \prod_{i=1}^d I(i) \le 1 + t\overline{A}.$$

Let us now consider the remaining elements in the sum. Here we take advantage that a velocity flip is an order t event. Consider for the moment only the i-th component of the velocity vector. The probability that this is flipped (i.e. $w_i = -v_i$) satisfies

$$\begin{aligned}
&\frac{\lambda_i(x+vt, F_i w) - \lambda_i(x+vt, v) e^{-(\lambda_i(x+vt,v)+\lambda_i(x+vt,F_i v))t}}{\lambda_i(x+vt,v) + \lambda_i(x+vt, F_i v)} \leq \\
&\quad \leq \frac{\lambda_i(x+vt,v)(1 - e^{-t(|\partial_i \psi(x+tv)|+2\gamma_i(x+tv))})}{|\partial_i \psi(x+tv)| + 2\gamma_i(x+tv)} \\
&\quad \leq t\lambda_i(x+vt,v) \\
&\quad \leq t \max_{i=1,\dots,d,\,(x,v)\in C\times\{\pm 1\}^d,\,t\in(0,1)} \lambda_i(x+vt,v).
\end{aligned}$$

Here we used that $1 - \exp(-z) \leq z$ for $z \geq 0$. All other probabilities can be bounded by 1 and hence

$$\sum_{w\neq v} K_1 \prod_{i=1}^{d} I(i) \leq t \max_{i=1,\dots,d,\,(x,v)\in C\times\{\pm 1\}^d,\,t\in(0,1)} \lambda_i(x+vt,v) \sum_{w\neq v} \frac{V(x+vt+wt,w)}{V(x,v)}.$$

Since V is continuous under our assumptions we proved that for every compact set $C \times \{\pm 1\}^d$ there exists a constant $B > 0$ such that for all $(x,v) \in C \times \{\pm 1\}^d$ it holds

$$\overline{P}_t V(x,v) \leq (1+tB)V(x,v). \tag{4.40}$$

Therefore we have (4.35) holds for all $x \in \mathbb{R}^d, v \in \{\pm 1\}^d$. □

4.C.1.3 Proof of Equation (4.18)

Let us prove (4.18). Fix a probability measure μ on $\mathbb{R}^d \times \{\pm 1\}^d$, and let (x,v) be a point in the support of μ. Then we can construct the set $D(x,v)$ corresponding to (x,v) and given by (4.15). By Theorem 4.17 there is a unique invariant measure $\pi_\delta^{x,v}$ for the Markov process with kernel P_δ^2, and by (4.14) for any probability measures ν, ν' on $D(x,v)$

$$\|\nu P_\delta^{2n} - \nu' P_\delta^{2n}\|_V \leq \frac{C}{\alpha} \tilde{\kappa}^{n\delta} \|\nu - \nu'\|_V.$$

Setting $\nu = \delta_{x,v}$, $\nu' = \pi_\delta^{x,v}$ and using that $\pi_\delta^{x,v}$ is an invariant measure for the kernel P_δ^2 we have

$$\|\delta_{x,v} P_\delta^{2n} - \pi_\delta^{x,v}\|_V \leq \frac{C}{\alpha} \tilde{\kappa}^{n\delta} \|\nu - \nu'\|_V.$$

Then integrating with respect to the probability measure μ we obtain

$$\begin{aligned}
\|\mu P_\delta^{2n} - \mu \pi_\delta^{x,v}\|_V &= \sup_{|g|\leq V} \left| \int [P_\delta^{2n} g(x,v) - \pi_\delta^{x,v}(g)] \mu(\mathrm{d}x, \mathrm{d}v) \right| \\
&\leq \int \sup_{|g|\leq V} \left| P_\delta^{2n} g(x,v) - \pi_\delta^{x,v}(g) \right| \mu(\mathrm{d}x, \mathrm{d}v)
\end{aligned}$$

$$\leq \int \left\|\delta_{x,v}P_\delta^{2n} - \pi_\delta^{x,v}\right\|_V \mu(\mathrm{d}x,\mathrm{d}v)$$
$$\leq \frac{C}{\alpha}\tilde{\kappa}^{n\delta} \int \|\delta_{(x,v)} - \pi_\delta^{x,v}\|_V \mu(\mathrm{d}x,\mathrm{d}v).$$

4.C.2 Other splitting schemes

In this Section we consider splitting schemes **DRBRD** and **RDBDR** of ZZS and prove geometric ergodicity in Theorem 4.31. The minorisation and drift conditions are proved in Sections 4.C.2.1 and 4.C.2.2 respectively. We shall work under the following assumption.

Assumption 4.30. *There exists $\gamma_0 \in (0,\infty)$ such that the following conditions for the refreshment rate hold:*

(a) there exists $R > 0$ for which for any $|x| \geq R$ it holds that

$$\gamma(x)\prod_{j=1}^{d}(1+|\partial_j\psi(x)|) \leq \gamma_0.$$

(b) For $|x| > R$ for some $R > 0$ it holds that

$$\sup \gamma(x+vt)e^{t|\nabla\psi(x)|+\frac{t^2}{2}(v+w)^T\nabla^2\psi(y_1)(v+w)+|\nabla\psi(y_2)|}\prod_{i=1}^{d}(1+2|\partial_i\psi(x)|) \leq \gamma_0,$$

where the supremum is over $t \in (0,1)$, $y_1, y_2 \in B(x, t\sqrt{d})$, $v, w \in \{-1,1\}^d$.

Theorem 4.31. *Consider splitting schemes **DRBRD** and **RDBDR** of ZZS. Suppose Assumption 4.16(a), (c) holds for switching rates $\lambda_i(x,v) = (v_i\partial_i\psi(x))_+$. Suppose moreover that the refreshment rate γ satisfies Assumption 4.30(a) for scheme **RDBDR** and Assumption 4.30(b) for scheme **DRBRD**. Then statements (1) and (2), as well as Equation* (4.18)*, hold. In particular (2) holds with $\delta_0 < 2(1+2\gamma_0+\gamma_0^2)^{-1}$ for **RDBDR** and with $\delta_0 < 2(1+2\gamma_0)^{-1}$ for **DRBRD**.*

4.C.2.1 Minorisation condition

Splitting DRBRD

The chain obtained by **DRBRD** has the same periodic behaviour of DBD. Hence this case can be treated in the same way and a minorisation condition follows by the same reasoning used in Section 4.C.1.1 for splitting DBD.

Splitting RDBDR

In this case we give a sketch of the proof. The chain obtained by **RDBDR** breaks the grid behaviour exhibited by DBD because of the two refreshment steps at the beginning and end of each step. Indeed, consider the one-dimensional case and recall

the definition of the grid $D(x,v)$ as in Lemma 4.27. Since $v = \pm 1$, there are two disjoint grids: $D(x,+1)$ and $D(x,-1)$, with the idea being that after even steps of DBD the process lives on the same grid it started from. However, the process **RDBDR** can swap between one grid and the other by having a velocity refreshment. Indeed, starting the process at $(x,+1)$ and having a velocity flip due to a refreshment at the end of the first step and having no other jumps, we find the state of the process is $(X_2,V_2) = (x,-1)$. Therefore after even steps this process lives on the grid $D(x) = \{y : y = x + m\delta,\, m \in \mathbb{Z}\}$. If the initial and final condition are on the same grid $D(x,v)$, then no refreshment is required and one can simply use the proof of the scheme DBD. On the other hand, if the two states are on different grids, i.e. one is on $D(x,+1)$ and the other on $D(x,-1)$, then a refreshment is required to choose the right grid.

In order to maintain the right dependence on the step size δ it is required to give the process additional $\lceil \frac{M}{\delta} \rceil$ iterations, for a constant $M > 0$. Indeed with this modification the probability of having a refreshment in the first $\lceil \frac{M}{\delta} \rceil$ is constant has a lower bound which is independent of δ, assuming $\delta \leq \delta_0$ for some $\delta_0 > 0$ (see for instance the second case in the proof of Lemma 4.27). After the first $\lceil \frac{M}{\delta} \rceil$ iterations the process is on the right grid and Lemma 4.27 can be applied with the further constraint that no (more) refreshments take place. Note that this event again has a lower bounded probability independent of δ. Since in the first $\lceil \frac{M}{\delta} \rceil$ iterations the process can go out of the initial compact set $C = [-R,R]$, it follows that the Lemma should be applied with set $\tilde{C} = [-R-M, R+M]$.

The extension to the multidimensional case follows by applying this same intuition to every component.

4.C.2.2 Drift condition

Let us start with an auxiliary result.

Lemma 4.32. *Suppose the refreshment rate γ satisfies Assumption 4.30(a). Then $P_t^R V \leq (1+\gamma_0 t)V + Mt$, where γ_0 is as in Assumption 4.30 and M independent of t.*

Proof. Let V be as in Lemma 4.29. Applying the transition kernel P_t^R to V we find

$$\begin{aligned}
P_t^R V(x,v) &= V(x,v)e^{-t\gamma(x)} + \frac{1}{2^d}(1-e^{-t\gamma(x)})\sum_{w\neq v} V(x,w) \\
&= V(x,v)e^{-t\gamma(x)} + (1-e^{-t\gamma(x)})V(x,v)\frac{1}{2^d}\sum_{w\neq v}\frac{V(x,w)}{V(x,v)} \\
&= V(x,v)\Big(e^{-t\gamma(x)} + (1-e^{-t\gamma(x)})\frac{1}{2^d}\sum_{w\neq v}\prod_{j:v_j\neq w_j}(1+|\partial_j\psi(x)|)\Big) \\
&\leq V(x,v)\left(e^{-t\gamma(x)} + t\gamma(x)\prod_{j=1}^{d}(1+|\partial_j\psi(x)|)\right).
\end{aligned}$$

Clearly for x inside of a compact set this implies $P_t^R V(x,v) \leq (1+Bt)V(x,v)$ by taking maximum over x. Outside of a compact set we use Assumption 4.30 to obtain

$$P_t^R V(x,v) \leq V(x,v)(1+t\gamma_0).$$

□

Lemma 4.33. *Consider the splitting scheme **RDBDR** of ZZS. Suppose Assumptions 4.16(c) and 4.30(a) hold. Then there exist a function V and constants $\rho \in (0,1)$, $C>0$ such that for any $t \in (0,t_0)$ with $t_0 < (1+2\gamma_0+\gamma_0^2)^{-1}$ it holds that*

$$P_t^R P_t^D P_{2t}^B P_t^D P_t^R V(x,v) \leq (1-\rho t)V(x,v) + Ct.$$

Proof. Let V be as in Lemma 4.29. In the current context the result of the Lemma is that for all $t \in (0,t_0)$ with $t_0 < 1$ it holds that $P_t^D P_{2t}^B P_t^D V(x,v) \leq (1-\rho t)V(x,v) + Bt$ where $\rho = (1-e^{-Rt_0})t_0^{-1} - 1$ for R sufficiently large such that $\rho > 0$. Applying Lemmas 4.29 and 4.32 we obtain

$$\begin{aligned} P_t^R P_t^D P_{2t}^B P_t^D P_t^R V(x,v) &\leq (1+t\gamma_0) P_t^R P_t^D P_{2t}^B P_t^D V(x,v) + Mt \\ &\leq (1+t\gamma_0)(1-\rho t) P_t^R V(x,v) + t(M + (1+\gamma_0)B) \\ &\leq (1+t\gamma_0)^2(1-\rho t)V(x,v) + t(M(2+\gamma_0) + (1+\gamma_0)B). \end{aligned}$$

It is left to ensure that $(1+t\gamma_0)^2(1-\rho t) \leq (1-\tilde{\rho}t)$ for $\tilde{\rho} > 0$. We have

$$(1+t\gamma_0)^2(1-\rho t) \leq (1 - t(\rho - 2\gamma_0 - \gamma_0^2)).$$

Hence it is needed that

$$\frac{(1-e^{-Rt_0})}{t_0} - 1 > 2\gamma_0 + \gamma_0^2$$

and thus that $e^{-Rt_0} < 1 - t_0(1+2\gamma_0+\gamma_0^2)$, which is valid as R can be taken as large as needed and $t_0 < (1+2\gamma_0+\gamma_0^2)^{-1}$.

□

Lemma 4.34. *Consider the splitting scheme **DRBRD** of ZZS. Suppose Assumptions 4.16(c) and 4.30 hold. Then there exist a function V and constants $\rho \in (0,1)$, $C>0$ such that for any $t \in (0,t_0)$ with $t_0 < (1+2\gamma_0)^{-1}$ it holds that*

$$P_t^D P_t^R P_{2t}^B P_t^R P_t^D V(x,v) \leq (1-\rho t)V(x,v) + Ct.$$

Proof. Let V be as in Lemma 4.29. Observe that by Lemma 4.32 we have that

$$P_t^R P_t^D V(x,v) = P_t^R V(x+vt,v) \leq (1+\gamma_0 t)V(x+vt,v) + Mt$$

and thus $P_t^R P_t^D V(x,v) \leq (1+\gamma_0 t) P_t^D V(x,v) + Mt$. Then

$$P_t^D P_t^R P_{2t}^B P_t^R P_t^D V(x,v) \leq (1+\gamma_0 t) P_t^D P_t^R P_{2t}^B P_t^D V(x,v) + Mt$$

and

$$P_t^D P_t^R P_{2t}^B P_t^D V(x,v) = e^{-t\gamma(x+vt)} P_t^D P_{2t}^B P_t^D V(x,v) + V(x,v)(1-e^{-t\gamma(x+vt)}) \sum_{w\in\{\pm 1\}^d} \frac{1}{2^d} \frac{V(x+vt+wt,w)}{V(x,v)}.$$

The first term corresponds to the case of no refreshments, while in the second term a refreshment takes place. For the first term we can directly apply Lemma 4.29, which in the current context shows that for $t < t_0 < 1$ it holds $P_t^D P_{2t}^B P_t^D V(x,v) \leq (1-\rho t)V(x,v) + Mt$ for $\rho = (1-e^{-\beta_1 M t_0})t_0^{-1} - 1$. The second term can be rewritten as in (4.37), that is for $\overline{x}_1 = \overline{x}_1(x,v,w,t) \in B(x, t\sqrt{d})$

$$\begin{aligned}
&\frac{V(x+vt+wt,w)}{V(x,v)} = \\
&= \exp\left(\beta(\psi(x+vt+wt)-\psi(x)) + \sum_{i=1}^d (\phi(w_i\partial_i\psi(x+vt+wt)) - \phi(v_i\partial_i\psi(x)))\right) \\
&= e^{t|\nabla\psi(x)| + \frac{t^2}{2}(v+w)^T\nabla^2\psi(\overline{x}_1)(v+w) + |\nabla\psi(\overline{x}_2)|} \prod_{i=1}^d (1+2|\partial_i\psi(x)|)
\end{aligned}$$

Using Assumption 4.30(b) we find

$$\begin{aligned}
(1-e^{-t\gamma(x+vt)}) \sum_{w\in\{\pm 1\}^d} \frac{1}{2^d} \frac{V(x+vt+wt,w)}{V(x,v)} &\leq \\
&\leq t\gamma(x+vt) \sum_{w\in\{\pm 1\}^d} \frac{1}{2^d} \frac{V(x+vt+wt,w)}{V(x,v)} \\
&\leq t\gamma_0.
\end{aligned}$$

Therefore we have shown

$$P_t^D P_t^R P_{2t}^B P_t^R P_t^D V \leq (1+\gamma_0 t)(1-\rho t + t\gamma_0)V + \tilde{M}t \leq (1-t(\rho-2\gamma_0))V + \tilde{M}t.$$

Hence it is sufficient to ensure that $\rho > 2\gamma_0$, which can be done similarly to the proof of Lemma 4.33. □

4.D Proof of Proposition 4.22 and related results

In this section we collect statements and proofs that are not included in Section 4.4.

4.D.1 Proof of Proposition 4.22

In this section we prove Proposition 4.22. We start by focusing on the left hand side of (4.20), i.e. $\mathcal{L}^*_{BPS}(\mu f_2)$. We find since μ is rotationally invariant in v

$$\mathcal{L}^*_{BPS}(\mu f_2)(x,v) =$$

$$= \mu(x,v)\Big\{\langle v, \nabla\psi(x)\rangle f_2(x,v) - \langle v, \nabla_x f_2(x,v)\rangle + (-\langle v, \nabla\psi(x)\rangle)_+ f_2(x, R(x)v)$$
$$- \langle v, \nabla\psi(x)\rangle_+ f_2(x,v) + \lambda_r \int f_2(x,y)\nu(\mathrm{d}y) - \lambda_r f_2(x,v)\Big\}.$$

We shall consider the case of $v = \pm 1$, hence $\nu = (1/2)\delta_{+1} + (1/2)\delta_{-1}$. In particular this choice satisfies Assumption 4.19 below. Introduce the notation $f_2^+(x) = f_2(x,1)$, $f_2^-(x) = f_2(x,-1)$. We have in the 1-dimensional setting

$$\mathcal{L}^*_{BPS}(\mu f_2)(x,+1) = -\mu(x,+1)\Big\{(f_2^+)'(x) + ((-\psi'(x))_+ + \lambda_r/2)f_2^+(x)$$
$$- (\lambda_r/2 + (-\psi'(x))_+)f_2^-(x)\Big\},$$
$$\mathcal{L}^*_{BPS}(\mu f_2)(x,-1) = +\mu(x,-1)\Big\{(f_2^-)'(x) + ((+\psi'(x))_+ + \lambda_r/2)f_2^+(x)$$
$$- (\lambda_r/2 + (+\psi'(x))_+)f_2^-(x)\Big\}.$$

Define function h such that $h\mu = \mathcal{L}_2^*\mu$, and also $h^+(x) = h(x,+1)$ and $h^-(x) = h(x,-1)$. Therefore we wish to solve the following system of ODEs

$$\begin{cases} (f_2^+)'(x) = -(\lambda_r/2 + (-\psi'(x))_+)f_2^+(x) + (\lambda_r/2 + (-\psi'(x))_+)f_2^-(x) - h^+(x), \\ (f_2^-)'(x) = -(\lambda_r/2 + (+\psi'(x))_+)f_2^+(x) + (\lambda_r/2 + (+\psi'(x))_+)f_2^-(x) + h^-(x), \end{cases} \tag{4.41}$$

with compatibility condition (4.21), which in this case can be written as

$$\int_{-\infty}^{\infty} (f_2^+(x) + f_2^-(x))\pi(x)\mathrm{d}x = 0 \tag{4.42}$$

with $\pi(x) = \mu(x,1) + \mu(x,-1)$. Let us find a solution to (4.41) for a generic (continuous and locally lipschitz) function h. Start by subtracting the first line to the second line in (4.41):

$$(f_2^-)'(x) - (f_2^+)'(x) = ((\psi'(x))_+ - (-\psi'(x))_+)(f_2^-(x) - f_2^+(x)) + h_s(x), \tag{4.43}$$

where $h_s(x) = h^+(x) + h^-(x)$. Define $g = f_2^- - f_2^+$ and notice that $(\psi'(x))_+ - (-\psi'(x))_+ = \psi'(x)$. Then we can rewrite (4.43) as

$$g'(x) = \psi'(x)g(x) + h_s(x).$$

Solving this ODE using an integrating factor we find

$$g(x) = \exp(\psi(x)) \lim_{y\to-\infty} [\exp(-\psi(y))\, g(y)] + \exp(\psi(x)) \int_{-\infty}^{x} h_s(y)\exp(-\psi(y))\mathrm{d}y.$$

Recall that $g = f^- - f^+$ and f^+, f^- satisfy (4.42). In order for f_2 to define a proper density we require

$$\int_{-\infty}^{\infty} g(x)\pi(x)\mathrm{d}x < \infty.$$

For this to hold it must be that $\lim_{y\to-\infty} \exp(-\psi(y))\, g(y) = 0$ and thus

$$g(x) = \exp(\psi(x)) \int_{-\infty}^{x} h_s(y) \exp(-\psi(y)) \mathrm{d}y. \tag{4.44}$$

Since $f_2^-(x) = f_2^+(x) + g(x)$ and plugging this in the first equation of (4.41) we obtain the ODE

$$(f_2^+)'(x) = (\lambda_r/2 + (-\psi'(x))_+) g(x) - h^+(x)$$

which can be integrated as

$$f_2^+(x) = f_2^+(0) + \int_0^x \left((\lambda_r/2 + (-\psi'(y))_+) g(y) - h^+(y)\right) \mathrm{d}y. \tag{4.45}$$

It follows that

$$\begin{aligned} f_2^-(x) = f_2^+(0) &+ \int_0^x \left((\lambda_r/2 + (-\psi'(y))_+) g(y) - h^+(y)\right) \mathrm{d}y \\ &+ \exp(\psi(x)) \int_{-\infty}^{x} h_s(y) \exp(-\psi(y)) \mathrm{d}y. \end{aligned} \tag{4.46}$$

Finally we compute $f_2^+(0)$ enforcing the compatibility condition (4.42). Plugging (4.45) and (4.46) in (4.42) we find the condition

$$f_2^+(0) = -\int_{-\infty}^{\infty} \Big(g(x)/2 + \int_0^x \left((\lambda_r/2 + (-\psi'(y))_+) g(y) - h^+(y)\right) \mathrm{d}y \Big) \pi(x) \mathrm{d}x. \tag{4.47}$$

4.D.2 Proof of Proposition 4.24

Fix $x \in \mathbb{R}$, $\delta > 0$ and let $G(x,\delta) := \{(z,v) \in \mathbb{R} \times \{\pm 1\} : (z-x)/\delta \in \mathbb{Z}\}$ be the state space of the chain with initial position x. For now, let μ_δ be any probability measure on $G(x,\delta)$ such that $\mu_\delta(y,w) = \mu_\delta(y,-w)$ for all $(y,w) \in G(x,\delta)$, and let us give a sufficient and necessary condition for it to be invariant by the chain. Since such a μ_δ is invariant by the refreshment step, it is invariant for the scheme **RDBDR** if and only if it is invariant for the scheme **R'DBD**, where **R'** is a deterministic flip of the velocity (which, as **R**, preserves μ_δ). Besides, from a state $(y,w) \in G(x,\delta)$, one transition of **R'DBD** can only lead to (y,w) or $(y+\delta w, -w)$, from which it can only stay or come back to the initial (y,w). In other words the pair $\{(y,w),(y+\delta w,-w)\}$ is irreducible for this chain, and thus μ_δ is invariant for **R'DBD** if and only if its restrictions on all these sets for $(y,w) \in G(x,\delta)$ are invariant by this scheme, which by definition reads

$$\forall (y,w) \in G(x,\delta)\,, \qquad \mu_\delta(y,w) e^{-\delta\lambda(y+w\delta/2,w)} = \mu_\delta(y+\delta w,-w) e^{-\delta\lambda(y+w\delta/2,-w)}.$$

It turns out that this is exactly the skew detailed balance condition (2.17) for the scheme **DBD**. Writing that $\mu_\delta(y,w) \propto \exp(-\psi_\delta(y))$ for some ψ_δ and recalling that $\lambda(y,w) - \lambda(y,-w) = \psi'(y)$ for all y, w, this is equivalent to

$$\forall (y,w) \in G(x,\delta), \qquad \psi_\delta(y+\delta w) - \psi_\delta(y) = \delta \psi'\left(y + \delta w/2\right) .$$

Up to an additive constant, the only function ψ_δ which satisfies this is the one given in the statement of Proposition 4.24. As a conclusion, we have proven that a probability measure on $G(x,\delta)$ which is independent from the velocity is invariant for the scheme **RDBDR** if and only if it is the one given in the proposition, which concludes the proof of the first statement.

Now we focus on the convergence of empirical means, assuming that the conditions of Theorem 4.17 are met. The reference position $x \in \mathbb{R}$ is still fixed. The long-time convergence established in Theorem 4.17 (for P_δ^2 where P_δ is one step of the scheme) is well-known to imply an ergodic Law of Large Numbers. In particular, for all initial conditions in $G(x,\delta)$ and all bounded f, distinguishing odd and even indexes, we see that $\frac{1}{N}\sum_{k=1}^N f(\overline{Z}_{t_k})$ (where $(\overline{Z}_{t_k})_{k\in\mathbb{N}}$ is a trajectory of the scheme) converges almost surely as $N \to \infty$ to $\tilde{\mu}_\delta(f) := (\mu'_\delta(f) + \mu''_\delta(f))/2$, where μ'_δ and μ''_δ are the unique invariant measures of P_δ^2 on each periodic component of the state space. In particular, $\tilde{\mu}_\delta$ is an invariant measure for P_δ. In dimension 1, the scheme **DBD** is such that for all y, for all times, the number of visits of the points $(y,1)$ and $(y,-1)$ differ at most by 1, which implies by ergodicity that $\tilde{\mu}_\delta(y,w) = \mu_\delta(y,-w)$ for all $(y,w) \in G(x,\delta)$, and we conclude thanks to the first part of the proof.

4.D.3 Application of Proposition 4.22 to three one-dimensional targets

In this section we give the function f_2 corresponding to the three cases considered in Figure 4.3.

4.D.3.1 Gaussian target

Let us start with a one-dimensional Gaussian target with mean zero and variance $\sigma^2 > 0$.

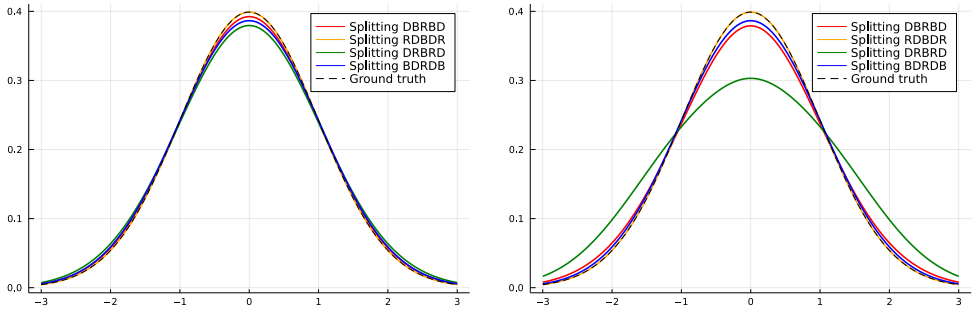
Proposition 4.35. *Let $\psi(x) = x^2/(2\sigma^2)$ for $\sigma^2 > 0$. Then:*

- *For the splitting scheme **DBRBD** it holds that*

$$f_2(x,+1) = f_2(x,-1) = \frac{\lambda_r}{24}\left(\frac{2\sqrt{2}}{\sigma\sqrt{\pi}} - \frac{x^3}{\sigma^4}\mathrm{sign}(x)\right).$$

- *For the splitting scheme **BDRDB** it holds that*

$$f_2(x,+1) = \frac{1}{8\sigma^2} - \frac{1}{4\sigma^4}x^2 \mathbb{1}_{x<0},$$



(a) Refreshment rate $\lambda_r = 1.0$. (b) Refreshment rate $\lambda_r = 3.0$.

Figure 4.11: Plots of the theoretical invariant measure up to second order for a standard Gaussian target, as given by Proposition 4.35. The step size is $\delta = 0.5$.

$$f_2(x,-1) = \frac{1}{8\sigma^2} - \frac{1}{4\sigma^4}x^2 \mathbb{1}_{x>0}.$$

- *For the splitting scheme* ***RDBDR*** *it holds that*

$$f_2(x,+1) = f_2(x,-1) = 0.$$

- *For the splitting scheme* ***DRBRD*** *it holds that*

$$f_2(x,+1) = f_2(x,-1) = \frac{\lambda_r}{12}\left(\frac{2\sqrt{2}}{\sigma\sqrt{\pi}} - \frac{|x|^3}{\sigma^4}\right) + \frac{\lambda_r^2}{16}\left(1 - \frac{x^2}{\sigma^2}\right).$$

Proof of Proposition 4.35. Recalling that $v \in \{-1,+1\}$, for all splitting schemes we can compute the functions $h = (\mathcal{L}_2^*\mu)/\mu$:

$$\begin{aligned}
h_{DBRBD}(x,v) &= \frac{\lambda_r}{8\sigma^4}(x^2 + 2vx(-vx)_+), \\
h_{BDRDB}(x,v) &= \frac{1}{8\sigma^6}\left(-\lambda_r\sigma^2(x^2 + 2vx(-vx)_+) + 2(-vx)_+(x^2 - 2\sigma^2)\right), \\
h_{RDBDR}(x,v) &= 0, \\
h_{DRBRD}(x,v) &= \frac{1}{8\sigma^4}\lambda_r\left(x^2 + vx(3(-vx)_+ + (vx)_+) + \lambda_r vx\sigma^2\right).
\end{aligned}$$

Splitting DBRBD

Observe that $h_s(x) = \frac{\lambda_r}{4\sigma^4}(x^2 + x((-x)_+ - (x)_+)) = 0$. Hence by (4.44) it holds $g(x) = 0$. Then by (4.45)

$$f^+_{2,DBRBD}(x) = f_2^+(0) - \int_0^x \frac{\lambda_r}{8\sigma^4}(y^2 + 2y(-y)_+)\mathrm{d}y$$

$$= f_2^+(0) - \frac{\lambda_r}{24\sigma^4} x^3(1 - 2\mathbb{1}_{x<0}).$$

Since $g = 0$ we have $f^+_{2,DBRBD} = f^-_{2,DBRBD}$. To find $f_2^+(0)$ we enforce (4.42):

$$\begin{aligned}\int_{-\infty}^{+\infty} f_2^+(x)\pi(x)\mathrm{d}x &= f_2^+(0) - \frac{\lambda_r}{24\sigma^4}\int_{-\infty}^{+\infty} x^3(1 - 2\mathbb{1}_{x<0})\pi(x)\mathrm{d}x \\ &= f_2^+(0) - \frac{\lambda_r}{12\sigma^4}\int_0^{+\infty} x^3\pi(x)\mathrm{d}x \\ &= f_2^+(0) - \frac{\lambda_r}{12\sigma^4}\sigma^3\sqrt{\frac{2}{\pi}} = 0.\end{aligned}$$

Clearly this is satisfied for $f_2^+(0) = \lambda_r/(6\sigma\sqrt{2\pi})$.

Splitting RDBDR

Clearly in this case $h_s(x) = 0$, hence $g(x) = 0$ and $f^+_{2,RDBDR} = f^-_{2,RDBDR} = 0$.

Splitting BDRDB

We have $h_s(x) = \frac{1}{4\sigma^6}|x|(x^2 - 2\sigma^2)$. Now inserting this into the expression for g we get

$$\begin{aligned}g(x) &= \frac{1}{4\sigma^6}\exp(x^2/(2\sigma^2))\int_{-\infty}^{x} |y|(y^2 - 2\sigma^2)\exp(-x^2/(2\sigma^2))\mathrm{d}y \\ &= \frac{1}{4\sigma^6}\exp(x^2/(2\sigma^2))(-\sigma^2\mathrm{sign}(x)\exp(-x^2/(2\sigma^2))x^2) \\ &= -\frac{1}{4\sigma^4}x^2\mathrm{sign}(x).\end{aligned}$$

We compute f_2^+ by applying (4.45). First observe that

$$\int_0^x h^+(y)\mathrm{d}y = -\frac{\lambda_r}{8\sigma^4}\int_0^x y^2\mathrm{sign}(y)\mathrm{d}y + \frac{1}{4\sigma^6}\int_0^x (-y)_+(y^2 - 2\sigma^2)\mathrm{d}y$$

and

$$\int_0^x (\lambda_r/2 + (-y/\sigma^2)_+)g(y)\mathrm{d}y = -\frac{\lambda_r}{8\sigma^4}\int_0^x y^2\mathrm{sign}(y)\mathrm{d}y - \frac{1}{4\sigma^6}\int_0^x (-y)_+y^2\mathrm{sign}(y)\mathrm{d}y.$$

Therefore we obtain

$$\begin{aligned}f^+_{2,BDRDB}(x) &= f_2^+(0) + \frac{1}{2\sigma^4}\int_0^x (-y)_+\mathrm{d}y \\ &= f_2^+(0) + \frac{1}{4\sigma^4}x^2\mathbb{1}_{x<0}.\end{aligned}$$

Enforcing the compatibility condition (4.47) we obtain $f_2^+(0) = 1/(8\sigma^2)$.

Splitting DRBRD

Similarly to the case of **DBRBD** observe that $h_s = 0$ and thus $g(x) = 0$. Observe that $h^+_{DRBRD}(x) = \frac{\lambda_r}{8\sigma^4}(2x^2\text{sign}(x) + \lambda_r x\sigma^2)$. Then by (4.45)

$$\begin{aligned} f^+_{2,DRBRD}(x) &= f^+_2(0) - \frac{\lambda_r}{8\sigma^4}\int_0^x (2y^2\text{sign}(y) + \sigma^2\lambda_r y)\text{d}y \\ &= f^+_2(0) - \frac{\lambda_r}{12\sigma^4}x^3\text{sign}(x) - \frac{\lambda_r^2}{16\sigma^2}x^2. \end{aligned}$$

To find $f^+_2(0)$ we enforce (4.47):

$$f^+_2(0) = \frac{\lambda_r}{6\sigma}\sqrt{\frac{2}{\pi}} + \frac{\lambda_r^2}{16}.$$

□

4.D.3.2 Non-Lipschitz potential

Now we focus on a target distribution with non-Lipschitz potential.

Proposition 4.36. *Let $\psi(x) = x^4$. Then:*

- *For the splitting scheme **DBRBD** it holds that*

$$f_2(x,+1) = f_2(x,-1) = \frac{\lambda_r}{7}\left(\frac{1}{2\Gamma(5/4)} - 2x^7\text{sign}(x)\right) + \frac{1}{2}\left(\frac{\Gamma(3/4)}{\Gamma(1/4)} - x^2\right).$$

- *For the splitting scheme **BDRDB** it holds that*

$$\begin{aligned} f_2(x,+1) &= \frac{5\Gamma(3/4)}{2\Gamma(1/4)} - x^2 - 4x^6\mathbb{1}_{x<0}\,, \\ f_2(x,-1) &= \frac{5\Gamma(3/4)}{2\Gamma(1/4)} - x^2 - 4x^6\mathbb{1}_{x\geq 0}\,. \end{aligned}$$

- *For the splitting scheme **RDBDR** it holds that*

$$f_2(x,+1) = f_2(x,-1) = \frac{\Gamma(3/4)}{2\Gamma(1/4)} - \frac{1}{2}x^2.$$

- *For the splitting scheme **DRBRD** it holds that*

$$\begin{aligned} f_2(x,+1) = f_2(x,-1) = \frac{\lambda_r}{7}\left(\frac{1}{\Gamma(5/4)} - 4x^7\text{sign}(x)\right) + \frac{1}{2}\left(\frac{\Gamma(3/4)}{\Gamma(1/4)} - x^2\right) \\ + \frac{\lambda_r^2}{8}\left(\frac{1}{4} - x^4\right). \end{aligned}$$

Proof of Proposition 4.36. By Proposition 4.20 we obtain

$$\begin{aligned}
h_{DBRBD}(x,v) &= +2\lambda_r(x^6 + 2vx^3(-vx^3)_+) + vx,\\
h_{BDRDB}(x,v) &= -2\lambda_r(x^6 + 2vx^3(-vx^3)_+) + 8(-vx^3)_+(-3x^2 + 2x^6) - 2vx,\\
h_{RDBDR}(x,v) &= vx,\\
h_{DRBRD}(x,v) &= +2\lambda_r(x^6 + vx^3(3(-vx^3)_+ + (vx^3)_+)) + vx + (\lambda_r^2 vx^3)/2.
\end{aligned}$$

Denote the normalisation constant of the target $\pi(x)$ by $Z = 2\Gamma(5/4)$.

Splitting DBRBD
Since $h_s(x) = 0$ we have

$$f^+_{2,DBRBD}(x) = f^-_{2,DBRBD}(x) = f_2^+(0) - \frac{2}{7}\lambda_r x^7 \mathrm{sign}(x) - \frac{1}{2}x^2,$$

with

$$f_2^+(0) = \frac{4\lambda_r}{7}\int_0^\infty x^7\pi(x)\mathrm{d}x + \int_0^\infty x^2\pi(x)\mathrm{d}x = \frac{\lambda_r}{14\Gamma(5/4)} + \frac{\Gamma(3/4)}{2\Gamma(1/4)}.$$

Splitting BDRDB
In this case $h_s(x) = 8x^2|x^3|(2x^4 - 3)$ and thus we find $g(x) = -4x^6\mathrm{sign}(x)$. It follows that

$$\begin{aligned}
f^+_{2,BDRDB}(x) &= f_2^+(0) - 4x^6\mathbb{1}_{x<0} - x^2,\\
f^-_{2,BDRDB}(x) &= f_2^+(0) - 4x^6\mathbb{1}_{x\geq 0} - x^2,
\end{aligned}$$

where

$$f_2^+(0) = \frac{\Gamma(7/4)}{2\Gamma(5/4)} + \frac{\Gamma(3/4)}{\Gamma(1/4)}.$$

Splitting RDBDR
Since $h_s(x) = 0$ we have $f^+_{2,RDBDR}(x) = f^-_{2,RDBDR}(x) = f_2^+(0) - x^2/2$ with $f_2^+(0) = \frac{\Gamma(3/4)}{2\Gamma(1/4)}$.

Splitting DRBRD
Since $h_s(x) = 0$ we have

$$f^+_{2,DRBRD}(x) = f^-_{2,DRBRD}(x) = f_2^+(0) - \frac{4}{7}\lambda_r x^7\mathrm{sign}(x) - \frac{1}{2}x^2 - \frac{1}{8}\lambda_r^2 x^4,$$

with

$$f_2^+(0) = \frac{\lambda_r}{7\Gamma(5/4)} + \frac{\Gamma(3/4)}{2\Gamma(1/4)} + \frac{1}{32}\lambda_r^2.$$

□

4.D.3.3 Heavy tailed target

Finally we consider a Cauchy distribution $\pi(x) = \gamma/(\pi(\gamma^2 + x^2))$ for $\gamma > 0$.

Proposition 4.37. *Let $\psi(x) = \ln(\gamma^2 + x^2)$. Then:*

- *For the splitting scheme **DBRBD** it holds that*

$$f_2(x,+1) = f_2(x,-1) = \frac{\lambda_r}{4\gamma}\left(\frac{\pi}{4} - |\arctan(x/\gamma)| + \frac{\gamma|x|}{\gamma^2+x^2} - \frac{1}{\pi}\right) + \frac{1}{12}\left(\frac{1}{4\gamma^2} + \frac{x^2-\gamma^2}{(\gamma^2+x^2)^2}\right).$$

- *For the splitting scheme **BDRDB** it holds that*

$$f_2(x,v) = \left(\frac{(x^2-3\gamma^2)^2}{48\gamma^2(x^2+\gamma^2)^2}\right)\mathbb{1}_{xv<0} + \left(\frac{x^4-54x^2\gamma^2+9\gamma^4}{48\gamma^2(x^2+\gamma^2)^2}\right)\mathbb{1}_{xv\geq 0}.$$

- *For the splitting scheme **RDBDR** it holds that*

$$f_2(x,+1) = f_2(x,-1) = \frac{1}{12}\left(\frac{1}{4\gamma^2} + \frac{x^2-\gamma^2}{(\gamma^2+x^2)^2}\right).$$

- *For the splitting scheme **DRBRD** it holds that*

$$f_2(x,+1) = f_2(x,-1) = \frac{\lambda_r}{2\gamma}\left(\frac{\pi}{4} - |\arctan(x/\gamma)| + \frac{\gamma|x|}{\gamma^2+x^2} - \frac{1}{\pi}\right) + \frac{1}{12}\left(\frac{1}{4\gamma^2} + \frac{x^2-\gamma^2}{(\gamma^2+x^2)^2}\right) + \frac{\lambda_r^2}{8}\left(\ln 4 - \ln\left(1+\frac{x^2}{\gamma^2}\right)\right).$$

Proof of Proposition 4.37. By Proposition 4.20 we obtain

$$\begin{aligned}
h_{DBRBD}(x,v) &= \frac{\lambda_r}{2(\gamma^2+x^2)^2}(x^2 + 2vx(-vx)_+) + \frac{1}{24}v\psi^{(3)}(x),\\
h_{BDRDB}(x,v) &= -\frac{\lambda_r}{2(\gamma^2+x^2)^2}(x^2 + 2vx(-vx)_+)\\
&\quad + \frac{2}{(\gamma^2+x^2)^3}(-vx)_+(-\gamma^2+2x^2) - \frac{1}{12}v\psi^{(3)}(x),\\
h_{RDBDR}(x,v) &= \frac{1}{24}v\psi^{(3)}(x),\\
h_{DRBRD}(x,v) &= \frac{\lambda_r}{2(\gamma^2+x^2)^2}\left(x^2 + vx(3(-vx)_+ + (vx)_+)\right)\\
&\quad + \frac{1}{24}v\psi^{(3)}(x) + \lambda_r^2\frac{xv}{4(\gamma^2+x^2)}.
\end{aligned}$$

Denote the normalisation constant of the target $\pi(x)$ by $Z = \pi/\gamma$.

Splitting DBRBD
Since $h_s(x) = 0$ we have

$$\begin{aligned} f^+_{2,DBRBD}(x) &= f^-_{2,DBRBD}(x) \\ &= f^+_2(0) - \frac{\lambda_r}{4}\mathrm{sign}(x)\left(\frac{\arctan(x/\gamma)}{\gamma} - \frac{x}{\gamma^2+x^2}\right) \\ &\quad - \frac{1}{12}\left(\frac{\gamma^2-x^2}{(\gamma^2+x^2)^2} - \frac{1}{\gamma^2}\right), \end{aligned}$$

with $f^+_2(0) = \frac{\lambda_r}{4\gamma}\left(\frac{\pi}{4} - \frac{1}{\pi}\right) - \frac{1}{16\gamma^2}$.

Splitting BDRDB
In this case $h_s(x) = 2(2x^2-\gamma^2)|x|/(\gamma^2+x^2)^3$ and thus we find

$$g(x) = -x^2\mathrm{sign}(x)/(\gamma^2+x^2)^2.$$

It follows that $f^+_2(0) = \frac{3}{16\gamma^2}$ and

$$\begin{aligned} f^+_{2,BDRDB}(x) &= \left(\frac{(x^2-3\gamma^2)^2}{48\gamma^2(x^2+\gamma^2)^2}\right)\mathbb{1}_{x<0} + \left(\frac{x^4-54x^2\gamma^2+9\gamma^4}{48\gamma^2(x^2+\gamma^2)^2}\right)\mathbb{1}_{x\geq 0}, \\ f^-_{2,BDRDB}(x) &= \left(\frac{(x^2-3\gamma^2)^2}{48\gamma^2(x^2+\gamma^2)^2}\right)\mathbb{1}_{x>0} + \left(\frac{x^4-54x^2\gamma^2+9\gamma^4}{48\gamma^2(x^2+\gamma^2)^2}\right)\mathbb{1}_{x<0}. \end{aligned}$$

Splitting RDBDR
Since $h_s(x) = 0$ we have

$$f^+_{2,RR}(x) = f^-_{2,RDBDR}(x) = \frac{1}{12\gamma^2}\left(\frac{1}{4} - 1\right) - \frac{1}{12}\left(\frac{\gamma^2-x^2}{(\gamma^2+x^2)^2} - \frac{1}{\gamma^2}\right).$$

Splitting DRBRD
Since $h_s(x) = 0$ we have

$$\begin{aligned} f^+_{2,DRBRD}(x) = f^-_{2,DRBRD}(x) &= f^+_2(0) - \frac{\lambda_r}{2}\mathrm{sign}(x)\left(\frac{\arctan(x/\gamma)}{\gamma} - \frac{x}{\gamma^2+x^2}\right) \\ &\quad - \frac{1}{12}\left(\frac{\gamma^2-x^2}{(\gamma^2+x^2)^2} - \frac{1}{\gamma^2}\right) - \frac{\lambda_r^2}{8}\ln\left(1+\frac{x^2}{\gamma^2}\right), \end{aligned}$$

with $f^+_2(0) = \frac{\lambda_r}{2\gamma}\left(\frac{\pi}{4} - \frac{1}{\pi}\right) + \frac{\lambda_r^2}{8}\ln 4 - \frac{1}{16\gamma^2}$. □

4.E Proof of Proposition 4.20

In Section 4.E.1 we obtain the first and second order commutators of BPS, while in Section 4.E.2 we use the BCH formula and the obtained results to prove Proposition 4.20.

4.E.1 Computing the commutators of BPS

In this section we compute the first and second order commutators for the various components of the adjoint of the BPS. In Section 4.E.1.1 we write down the commutator of the BPS and its decomposition in the three terms that represent the free transport, reflection mechanism, and velocity refreshments. In Section 4.E.1.2 we start with first order commutators, which are essential to compute second order commutators. The latter are computed in Section 4.E.1.3.

Now let us write the following identities, which form a lemma for convenience. These will be used countless times in the computation of the commutators below.

Lemma 4.38. *For $\lambda(x,v) = \langle v, \nabla\psi(x)\rangle_+$ it holds that*

$$\lambda_1(x, R(x)v) - \lambda_1(x,v) = -\langle v, \nabla\psi(x)\rangle, \tag{4.48}$$
$$\lambda_1(x, R(x)v) + \lambda_1(x,v) = +|\langle v, \nabla\psi(x)\rangle|. \tag{4.49}$$

The proof is trivial.

4.E.1.1 The adjoint of BPS

Consider the generator

$$\begin{aligned}\mathcal{L}f(x,v) = \langle v, \nabla_x f(x,v)\rangle + \lambda_1(x,v)[f(x,R(x)v) - f(x,v)] \\ + \lambda_2 \int \big(f(x,w) - f(x,v)\big)\nu(\mathrm{d}w)\end{aligned}$$

Then one obtains that the adjoint is given by

$$\begin{aligned}\mathcal{L}^*_{BPS} g(x,v) &= -\langle v, \nabla_x g(x,v)\rangle + ((g\lambda_1)(x,R(x)v) - (g\lambda_1)(x,v)) \\ &\quad + \lambda_r \left(\nu(v)\int g(x,y)\mathrm{d}y - g(x,v)\right) \\ &= (\mathcal{L}^*_D + \mathcal{L}^*_B + \mathcal{L}^*_R) g(x,v),\end{aligned}$$

where

$$\begin{aligned}\mathcal{L}^*_D g(x,v) &= -\langle v, \nabla_x g(x,v)\rangle, \\ \mathcal{L}^*_B g(x,v) &= g(x,R(x)v)\lambda_1(x,R(x)v) - g(x,v)\lambda_1(x,v), \\ \mathcal{L}^*_R g(x,v) &= \lambda_r \left(\nu(v)\int g(x,y)\mathrm{d}y - g(x,v)\right)\end{aligned}$$

Here the letters D, B, and R stand for drift, bounce, refreshment. If we take g to be the invariant measure of BPS, μ, then

$$\begin{aligned}\mathcal{L}^*_D \mu(x,v) &= \langle v, \nabla\psi(x)\rangle \mu(x,v), \\ \mathcal{L}^*_B \mu(x,v) &= -\langle v, \nabla\psi(x)\rangle \mu(x,v),\end{aligned}$$

$$\mathcal{L}_R^*\mu(x,v) = 0.$$

To obtain $\mathcal{L}_D^*\mu(x,v)$ we used the trivial, but useful, identity

$$\nabla_x\mu(x,v) = -\nabla\psi(x)\mu(x,v).$$

4.E.1.2 First order commutators

Let us start computing the three first order commutators $[\mathcal{L}_B^*, \mathcal{L}_D^*]$, $[\mathcal{L}_R^*, \mathcal{L}_D^*]$, and $[\mathcal{L}_R^*, \mathcal{L}_B^*]$, which are essential to compute higher order commutators. This is done below respectively in Lemmas 4.39, 4.41, 4.42.

Lemma 4.39. *Let g be a suitable function. It holds that*

$$\begin{aligned}[\mathcal{L}_B^*, \mathcal{L}_D^*]g(x,v) = &-\langle R(x)v, (\nabla_x g)(x, R(x)v)\rangle\lambda_1(x, R(x)v) + \langle v, \nabla_x g(x,v)\rangle\lambda_1(x,v)\\ &+ \langle v, \nabla_x\Big(g(x, R(x)v)\lambda_1(x, R(x)v) - g(x,v)\lambda_1(x,v)\Big)\rangle.\end{aligned}$$

In particular if $g = \mu$

$$[\mathcal{L}_B^*, \mathcal{L}_D^*]\mu(x,v) = \mu(x,v)\Big(\langle v, \nabla\psi(x)\rangle\big(\langle v, \nabla\psi(x)\rangle - |\langle v, \nabla\psi(x)\rangle|\big) - \langle v, \nabla^2\psi(x))v\rangle\Big).$$

Remark 4.40. Alternative ways to write $[\mathcal{L}_B^*, \mathcal{L}_D^*]\mu(x,v)$ can be found using the identities in Lemma 4.38. We find

$$\begin{aligned}[\mathcal{L}_B^*, \mathcal{L}_D^*]\mu(x,v) &= \mu(x,v)\Big(\lambda_1^2(x, R(x)v) - \lambda_1^2(x,v) + \langle v, \nabla\psi(x)\rangle^2 - \langle v, \nabla^2\psi(x)v\rangle\Big)\\ &= \mu(x,v)\Big(\lambda_1^2(x, R(x)v) - \lambda_1^2(x,v) + \langle v, (\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x))v\rangle\Big).\end{aligned}$$

Proof. We have

$$\begin{aligned}&[\mathcal{L}_B^*, \mathcal{L}_D^*]g(x,v) =\\ &= \mathcal{L}_B^*(-\langle v, \nabla_x g(x,v)\rangle) - \mathcal{L}_D^*(g(x, R(x)v)\lambda_1(x, R(x)v) - g(x,v)\lambda_1(x,v))\\ &= -\langle R(x)v, (\nabla_x g)(x, R(x)v)\rangle\lambda_1(x, R(x)v) + \langle v, \nabla_x g(x,v)\rangle\lambda_1(x,v)\\ &\quad + \langle v, \nabla_x\left(g(x, R(x)v)\lambda_1(x, R(x)v) - g(x,v)\lambda_1(x,v)\right)\rangle\end{aligned}$$

and hence

$$\begin{aligned}[\mathcal{L}_B^*, \mathcal{L}_D^*]\mu(x,v) &= -\mu(x,v)\langle v, \nabla\psi(x)\rangle(\lambda_1(x, R(x)v) + \lambda_1(x,v))\\ &\quad + \langle v, \nabla_x\Big(\mu(x,v)(\lambda_1(x, R(x)v) - \lambda_1(x,v))\Big)\rangle\\ &= \mu(x,v)(\lambda_1^2(x, R(x)v) - \lambda_1^2(x,v))\\ &\quad - \langle v, \nabla_x\Big(\mu(x,v)\langle v, \nabla\psi(x)\rangle\Big)\rangle.\end{aligned}$$

Then note that

$$\langle v, \nabla_x \Big(\mu(x,v) \langle v, \nabla \psi(x) \rangle \Big) \rangle = \langle v, \nabla^2 \psi(x) v - \nabla \psi(x) \langle v, \nabla \psi(x) \rangle \, \rangle \, \mu(x,v).$$

and hence

$$[\mathcal{L}_B^*, \mathcal{L}_D^*] \mu(x,v) = \mu(x,v) \Big(\langle v, \nabla \psi(x) \big(\langle v, \nabla \psi(x) \rangle - |\langle v, \nabla \psi(x) \rangle| \big) \rangle - \langle v, \nabla^2 \psi(x)) v \rangle \Big).$$

□

Lemma 4.41. *Let g be a suitable function. It holds that*

$$[\mathcal{L}_R^*, \mathcal{L}_D^*] g(x,v) = \lambda_r \nu(v) \Big(\langle v, \int \nabla_x g(x,y) \mathrm{d}y \rangle - \int \langle y, \nabla_x g(x,y) \rangle \mathrm{d}y \Big).$$

In particular if $g = \mu$

$$[\mathcal{L}_R^*, \mathcal{L}_D^*] \mu(x,v) = -\lambda_r \langle v, \nabla \psi(x) \rangle \mu(x,v).$$

Proof. We find

$$\begin{aligned}
[\mathcal{L}_R^*, \mathcal{L}_D^*] g(x,v) &= -\mathcal{L}_R^* (\langle v, \nabla_x g(x,v) \rangle) - \mathcal{L}_D^* \left(\lambda_r \left(\nu(v) \int g(x,y) \mathrm{d}y - g(x,v) \right) \right) \\
&= -\lambda_r \left(\nu(v) \int \langle y, \nabla_x g(x,y) \rangle \mathrm{d}y - \langle v, \nabla_x g(x,v) \rangle \right) \\
&\quad + \lambda_r \langle v, \nabla_x (\nu(v) \int g(x,y) \mathrm{d}y - g(x,v)) \rangle \\
&= \lambda_r \nu(v) \Big(\langle v, \int \nabla_x g(x,y) \mathrm{d}y \rangle - \int \langle y, \nabla_x g(x,y) \rangle \mathrm{d}y \Big)
\end{aligned}$$

and thus

$$\begin{aligned}
[\mathcal{L}_R^*, \mathcal{L}_D^*] \mu(x,v) &= \mathcal{L}_R^* (\langle v, \nabla \psi(x) \rangle \mu(x,v)) \\
&= -\lambda_r \langle v, \nabla \psi(x) \rangle \mu(x,v)
\end{aligned}$$

□

Lemma 4.42. *Let g be a suitable function. It holds that*

$$\begin{aligned}
[\mathcal{L}_R^*, \mathcal{L}_B^*] g(x,v) = \lambda_r \Bigg(& \nu(v) \int (g(x, R(x)y) \lambda_1(x, R(x)y) - g(x,y) \lambda_1(x,y)) \mathrm{d}y \\
& + \left(\nu(v) \int g(x,y) \mathrm{d}y \right) \langle v, \nabla \psi(x) \rangle \Bigg).
\end{aligned}$$

In particular if $g = \mu$

$$[\mathcal{L}_R^*, \mathcal{L}_B^*] \mu(x,v) = \lambda_r \langle v, \nabla \psi(x) \rangle \mu(x,v).$$

Proof. Compute

$$\begin{aligned}
[\mathcal{L}_R^*, \mathcal{L}_B^*]g(x,v) &= \mathcal{L}_R^*(g(x,R(x)v)\lambda_1(x,R(x)v) - g(x,v)\lambda_1(x,v)) \\
&\quad - \mathcal{L}_B^*\left(\lambda_r\left(\nu(v)\int g(x,y)\mathrm{d}y - g(x,v)\right)\right) \\
&= \lambda_r\Big(\nu(v)\int (g(x,R(x)y)\lambda_1(x,R(x)y) - g(x,y)\lambda_1(x,y))\mathrm{d}y \\
&\quad - \underbrace{g(x,R(x)v)\lambda_1(x,R(x)v)}_{A} + \underbrace{g(x,v)\lambda_1(x,v)}_{B}\Big) \\
&\quad - \lambda_r\Bigg(\left(\nu(R(x)v)\int g(x,y)\mathrm{d}y - \underbrace{g(x,R(x)v)}_{A}\right)\lambda_1(x,R(x)v) \\
&\quad - \left(\nu(v)\int g(x,y)\mathrm{d}y - \underbrace{g(x,v)}_{B}\right)\lambda_1(x,v)\Bigg).
\end{aligned}$$

It is now sufficient to cancel out the terms denoted by A and B that appear twice with opposite signs to obtain the final statement. For $g = \mu$

$$\begin{aligned}
[\mathcal{L}_R^*, \mathcal{L}_B^*]\mu(x,v) &= \mathcal{L}_R^*(-\langle v, \nabla\psi(x)\mu(x,v)) \\
&= -\lambda_r\left(\nu(v)\int \langle y, \nabla\psi(x)\rangle p(x,y)\mathrm{d}y - \langle v, \nabla\psi(x)\rangle\mu(x,v)\right)
\end{aligned}$$

which concludes by Assumption 4.19. □

Remark 4.43. It follows from Lemmas 4.41 and 4.42 that

$$[\mathcal{L}_R^*, \mathcal{L}_B^* + \mathcal{L}_D^*]\mu(x,v) = 0. \tag{4.50}$$

4.E.1.3 Higher order commutators

Let us now compute higher order commutators.

Lemma 4.44. *It holds that*

$$\begin{aligned}
[\mathcal{L}_B^*, [\mathcal{L}_R^*, \mathcal{L}_D^*]]\mu(x,v) = \lambda_r\mu(x,v)\Big(&\langle v, \nabla\psi(x)\rangle\Big(\lambda_1(x,R(x)v) + \lambda_1(x,v)\Big) \\
&+ b\,\mathrm{tr}\big(\nabla\psi(x)(\nabla\psi(x))^T - \nabla^2\psi(x)\big)\Big).
\end{aligned}$$

Proof. Applying Lemma 4.41 we obtain

$$\begin{aligned}
&[\mathcal{L}_B^*, [\mathcal{L}_R^*, \mathcal{L}_D^*]]\mu(x,v) = -\lambda_r\mathcal{L}_B^*\Big(\langle v, \nabla\psi(x)\rangle\mu(x,v)\Big) + [\mathcal{L}_R^*, \mathcal{L}_D^*]\Big(\langle v, \nabla\psi(x)\rangle\mu(x,v)\Big) \\
&= -\lambda_r\Big(\langle R(x)v, \nabla\psi(x)\rangle\mu(x,v)\lambda_1(x,R(x)v) - \langle v, \nabla\psi(x)\rangle\mu(x,v)\lambda_1(x,v)\Big)
\end{aligned}$$

$$
\begin{aligned}
&+ \lambda_r \nu(v)\Big(\langle v, \nabla_x \int (\langle y, \nabla\psi(x)\rangle \mu(x,y))\mathrm{d}y\rangle - \int \langle y, \nabla_x(\langle y, \nabla\psi(x)\rangle\mu(x,y))\rangle \mathrm{d}y\Big) \\
&= \lambda_r \mu(x,v)\langle v, \nabla\psi(x)\rangle\Big(\lambda_1(x, R(x)v) + \lambda_1(x,v)\Big) \\
&\quad - \lambda_r \mu(x,v)\Big(\int (\langle y, \nabla^2\psi(x)y\rangle - \langle y, \nabla\psi(x)\rangle^2)\nu(\mathrm{d}y)\Big) \\
&= \lambda_r \mu(x,v)\Big(\langle v, \nabla\psi(x)\rangle\Big(\lambda_1(x,R(x)v) + \lambda_1(x,v)\Big) \\
&\quad + b\,\mathrm{tr}\big(\nabla\psi(x)(\nabla\psi(x))^T - \nabla^2\psi(x)\big)\Big).
\end{aligned}
$$

Note that in the last line we used that $\langle a, b\rangle^2 = \langle a, bb^T a\rangle$ and that

$$
\begin{aligned}
&\int \langle y, (\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x))y\rangle \nu(\mathrm{d}y) = \\
&= \sum_{j=1}^{d}\sum_{\ell=1}^{d} (\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x))_{j\ell} \int (y_j y_\ell)\nu(\mathrm{d}y) \\
&= b\,\mathrm{tr}\Big(\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x)\Big)
\end{aligned}
\tag{4.51}
$$

which is a consequence of Assumption 4.19. □

Lemma 4.45. *It holds that*

$$
[\mathcal{L}_R^*, [\mathcal{L}_R^*, \mathcal{L}_B^*]]\mu(x,v) = -\lambda_r^2 \mu(x,v)\langle v, \nabla\psi(x)\rangle.
$$

Proof. Next we compute $[\mathcal{L}_R^*, [\mathcal{L}_R^*, \mathcal{L}_B^*]]$. Since $\mathcal{L}_R^*\mu(x,v) = 0$ we easily find

$$
\begin{aligned}
[\mathcal{L}_R^*, [\mathcal{L}_R^*, \mathcal{L}_B^*]]\mu(x,v) &= \mathcal{L}_R^*(\lambda_r\langle v, \nabla\psi(x)\rangle\mu(x,v)) \\
&= -\lambda_r^2 \mu(x,v)\langle v, \nabla\psi(x)\rangle.
\end{aligned}
$$

□

Lemma 4.46. *It holds that*

$$
[\mathcal{L}_R^*, [\mathcal{L}_R^*, \mathcal{L}_D^*]]\mu(x,v) = \lambda_r^2 \mu(x,v)\langle v, \nabla\psi(x)\rangle.
$$

Proof. The result follows from Lemma 4.45 and (4.50). □

Lemma 4.47. *It holds that*

$$
\begin{aligned}
[\mathcal{L}_R^*, [\mathcal{L}_B^*, \mathcal{L}_D^*]]\mu(x,v) = \lambda_r \mu(x,v)\Big(&b\,\mathrm{tr}\Big(\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x)\Big) \\
&- \lambda_1^2(x, R(x)v) + \lambda_1^2(x,v) - \langle v, (\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x))v\rangle\Big).
\end{aligned}
$$

Proof. Taking advantage of Lemma 4.39

$$
\begin{aligned}
&[\mathcal{L}_R^*,[\mathcal{L}_B^*,\mathcal{L}_D^*]]\mu(x,v) = \\
&= \mathcal{L}_R^*\Big(\mu(x,v)\Big(\lambda_1^2(x,R(x)v) - \lambda_1^2(x,v) + \langle v,(\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x))v\rangle\Big)\Big) \\
&= \lambda_r\mu(x,v)\Bigg(\int\Big(\lambda_1^2(x,R(x)y) - \lambda_1^2(x,y) + \langle y,(\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x))y\rangle\Big)\nu(\mathrm{d}y) \\
&\quad - \Big(\lambda_1^2(x,R(x)v) - \lambda_1^2(x,v) + \langle v,(\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x))v\rangle\Big)\Bigg).
\end{aligned}
$$

Observe that for $A = \{y : \langle y, \nabla\psi(x)\rangle \geq 0\}$ we have

$$
\begin{aligned}
&\int(\lambda_1^2(x,R(x)y) - \lambda_1^2(x,y))\nu(\mathrm{d}y) = \\
&\quad = \int_{A^C}\langle y,\nabla\psi(x)\rangle^2\nu(y)\mathrm{d}y - \int_A\langle y,\nabla\psi(x)\rangle^2\nu(y)\mathrm{d}y \\
&\quad = 0.
\end{aligned}
$$

This can be seen by the change of variables $y' = R(x)y$ in the first integral. The result then follows by using (4.51). □

Lemma 4.48. *It holds that*

$$
[\mathcal{L}_B^*,[\mathcal{L}_R^*,\mathcal{L}_B^*]]\mu(x,v) = -\lambda_r\mu(x,v)\langle v,\nabla\psi(x)\rangle(\lambda_1(x,R(x)v) + \lambda_1(x,v)).
$$

Proof. Consider now

$$
\begin{aligned}
[\mathcal{L}_B^*,[\mathcal{L}_R^*,\mathcal{L}_B^*]]\mu(x,v) &= \mathcal{L}_B^*(\lambda_r\langle v,\nabla\psi(x)\rangle\mu(x,v)) + [\mathcal{L}_R^*,\mathcal{L}_B^*](\langle v,\nabla\psi(x)\rangle\mu(x,v)) \\
&= \lambda_r\mu(x,v)\Bigg(\langle R(x)v,\nabla\psi(x)\rangle\lambda_1(x,R(x)v) - \langle v,\nabla\psi(x)\rangle\lambda_1(x,v) \\
&\quad + \int\Big(\langle R(x)y,\nabla\psi(x)\rangle\lambda_1(x,R(x)y\rangle - \langle y,\nabla\psi(x)\rangle\lambda_1(x,y)\Big)\nu(\mathrm{d}y) \\
&\quad + \int(\langle y,\nabla\psi(x)\rangle\nu(\mathrm{d}y)\langle v,\nabla\psi(x)\rangle\Bigg).
\end{aligned}
$$

The last term equals zero as ν has mean zero. Then observe that by Identity (4.48)

$$
\begin{aligned}
&\int\Big(\langle R(x)y,\nabla\psi(x)\rangle\lambda_1(x,R(x)y) - \langle y,\nabla\psi(x)\rangle\lambda_1(x,y)\Big)\nu(\mathrm{d}y) = \\
&\quad = -\int\langle y,\nabla\psi(x)\rangle(\lambda_1(x,R(x)y) + \lambda_1(x,y))\nu(\mathrm{d}y) \\
&\quad = -\Big(\int\lambda_1(x,R(x)y)^2\nu(\mathrm{d}y) - \int\lambda_1(x,y)^2\nu(\mathrm{d}y)\Big) \\
&\quad = 0,
\end{aligned}
\tag{4.52}
$$

where the last equality follows by invariance under rotation of ν as required in Assumption 4.19. Hence we have obtained the statement. □

Lemma 4.49. *It holds that*

$$[\mathcal{L}_B^*,[\mathcal{L}_B^*,\mathcal{L}_D^*]]\mu(x,v) = 2\mu(x,v)\lambda_1(x,R(x)v)\Big(\langle v,\nabla\psi(x)\rangle^2 - \langle v,\nabla^2\psi(x)v\rangle \\ - \langle R(x)v,\nabla\psi^2(x)R(x)v\rangle\Big).$$

Proof. By Lemma 4.39 we find

$$\begin{aligned}&[\mathcal{L}_B^*,[\mathcal{L}_B^*,\mathcal{L}_D^*]]\mu(x,v) = \\ &= \mathcal{L}_B^*\Big(\mu(x,v)(\lambda_1^2(x,R(x)v) - \lambda_1^2(x,v) + \langle v,\nabla\psi(x)\rangle^2 - \langle v,\nabla^2\psi(x)v\rangle)\Big) \qquad (*)\\ &\quad + [\mathcal{L}_B^*,\mathcal{L}_D^*]\Big(\langle v,\nabla\psi(x)\rangle\mu(x,v)\Big). \qquad (**)\end{aligned}$$

Let us treat the two terms separately, starting with (*). After applying $\mathcal{L}_B^*$ and using that $R(x)(R(x)v) = v$ the first term becomes

$$\begin{aligned}(*) &= \mu(x,v)\Big[-\Big(\lambda_1^2(x,R(x)v) - \lambda_1^2(x,v) + \langle v,\nabla\psi(x)\rangle^2 - \langle v,\nabla^2\psi(x)v\rangle\Big)\lambda_1(x,v)\\ &+ \big((\lambda_1^2(x,v) - \lambda_1^2(x,R(x)v) + \langle R(x)v,\nabla\psi(x)\rangle^2 - \langle R(x)v,\nabla^2\psi(x)R(x)v\rangle\big)\lambda_1(x,R(x)v)\Big]\\ &= \mu(x,v)\Big[(\lambda_1^2(x,v) - \lambda_1^2(x,R(x)v))(\lambda_1(x,R(x)v) + \lambda_1(x,v))\\ &\quad + \langle v,\nabla\psi(x)\rangle^2(\lambda_1(x,R(x)v) - \lambda_1(x,v))\\ &\quad - \langle R(x)v,\nabla^2\psi(x)R(x)v\rangle\lambda_1(x,R(x)v) + \langle v,\nabla^2\psi(x)v\rangle\lambda_1(x,v)\Big].\end{aligned}$$

Using Identity (4.48) we obtain that

$$\langle v,\nabla\psi(x)\rangle^2(\lambda_1(x,R(x)v) - \lambda_1(x,v)) = (\lambda_1^2(x,v) - \lambda_1^2(x,R(x)v)) \\ \times(\lambda_1(x,R(x)v) + \lambda_1(x,v))$$

and thus cancelling out the corresponding terms in (*) it follows that

$$(*) = \mu(x,v)\Big(\langle v,\nabla^2\psi(x)v\rangle\lambda_1(x,v) - \langle R(x)v,\nabla^2\psi(x)R(x)v\rangle\lambda_1(x,R(x)v)\Big).$$

Focusing now on (**), we apply Lemma 4.39 to find

$$\begin{aligned}(**) &= -\langle R(x)v,\nabla_x\big(\langle v,\nabla\psi(x)\rangle\mu(x,v)\big)(x,R(x)v)\rangle\lambda_1(x,R(x)v)\\ &\quad + \langle v,\nabla_x\big(\langle v,\nabla\psi(x)\rangle\mu(x,v)\big)\rangle\lambda_1(x,v)\\ &\quad + \langle v,\nabla_x\Big(\langle R(x)v,\nabla\psi(x)\rangle\mu(x,v)\lambda_1(x,R(x)v) - \langle v,\nabla\psi(x)\rangle\mu(x,v)\lambda_1(x,v)\Big)\rangle.\end{aligned}$$

Recalling that

$$\nabla_x(\langle v, \nabla\psi(x)\rangle\mu(x,v)) = \mu(x,v)(\nabla^2\psi(x)v - \nabla\psi(x)\langle v, \nabla\psi(x)\rangle),$$

we find

$$\begin{aligned}(**) = \mu(x,v)\Big[&\left(-\langle R(x)v, \nabla^2\psi(x)R(x)v\rangle + \langle v, \nabla\psi(x)\rangle^2\right)\lambda_1(x, R(x)v)\\ &+ \left(\langle v, \nabla^2\psi(x)v\rangle - \langle v, \nabla\psi(x)\rangle^2\right)\lambda_1(x,v)\Big]\\ &- \langle v, \nabla_x\Big(\langle v, \nabla\psi(x)\rangle\mu(x,v)|\langle v, \nabla\psi(x)\rangle|\Big)\rangle.\end{aligned}$$

In particular we used Lemma 4.38 to write the last term more compactly. The derivative in the last term can be computed as follows

$$\begin{aligned}-\langle v, \nabla_x\Big(\langle v, \nabla\psi(x)\rangle\mu(x,v)|\langle v, \nabla\psi(x)\rangle|\Big)\rangle &=\\ &= -\mu(x,v)\langle v, \nabla^2\psi(x)v|\langle v, \nabla\psi(x)\rangle| - \nabla\psi(x)\langle v, \nabla\psi(x)\rangle|\langle v, \nabla\psi(x)\rangle|\\ &\quad + \langle v, \nabla\psi(x)\rangle\mathrm{sign}(\langle v, \nabla\psi(x)\rangle)\nabla^2\psi(x)v\rangle\\ &= -\mu(x,v)\Big(- \langle v, \nabla\psi(x)\rangle^2|\langle v, \nabla\psi(x)\rangle| + \langle v, \nabla^2\psi(x)v\rangle\Big(|\langle v, \nabla\psi(x)\rangle|\\ &\quad + \langle v, \nabla\psi(x)\rangle\mathrm{sign}(\langle v, \nabla\psi(x)\rangle)\Big)\Big)\\ &= -\mu(x,v)\Big(- \langle v, \nabla\psi(x)\rangle^2|\langle v, \nabla\psi(x)\rangle| + \langle v, \nabla^2\psi(x)v\rangle 2|\langle v, \nabla\psi(x)\rangle|\Big)\\ &= \mu(x,v)|\langle v, \nabla\psi(x)\rangle|\big(\langle v, \nabla\psi(x)\rangle^2 - 2\langle v, \nabla^2\psi(x)v\rangle\big).\end{aligned}$$

Hence re-applying Lemma 4.38 we find

$$\begin{aligned}(**) = \mu(x,v)\Big[&- \langle R(x)v, \nabla^2\psi(x)R(x)v\rangle\lambda_1(x, R(x)v)\\ &+ 2\langle v, \nabla\psi(x)\rangle^2\lambda_1(x, R(x)v) - (2\lambda_1(x, R(x)v) + \lambda_1(x,v))\langle v, \nabla^2\psi(x)v\rangle\Big].\end{aligned}$$

The proof is now concluded by summing (*) and (**). □

Lemma 4.50. *It holds that*

$$[\mathcal{L}_D^*, [\mathcal{L}_R^*, \mathcal{L}_B^*]]\mu(x,v) = \lambda_r\mu(x,v)\left(\langle v, \nabla\psi(x)\rangle^2 - \langle v, \nabla^2\psi(x)v\rangle\right).$$

Proof. Consider now $[\mathcal{L}_D^*, [\mathcal{L}_R^*, \mathcal{L}_B^*]]$:

$$\begin{aligned}[\mathcal{L}_D^*, [\mathcal{L}_R^*, \mathcal{L}_B^*]]\mu(x,v) &= \mathcal{L}_D^*(\lambda_r\langle v, \nabla\psi(x)\rangle\mu(x,v)) - [\mathcal{L}_R^*, \mathcal{L}_B^*](\langle v, \nabla\psi(x)\rangle\mu(x,v))\\ &= -\langle v, \nabla_x(\lambda_r\langle v, \nabla\psi(x)\rangle\mu(x,v))\rangle\\ &\quad - \lambda_r\Big(\mu(x,v)\int(-\langle y, \nabla\psi(x)\rangle)(\lambda_1(x, R(x)y) + \lambda_1(x,y))\nu(\mathrm{d}y)\Big)\end{aligned}$$

$$= \lambda_r \mu(x,v) \left(\langle v, \nabla\psi(x)\rangle^2 - \langle v, \nabla^2\psi(x) v\rangle \right).$$

In particular we used that

$$\begin{aligned}\int (\langle y, \nabla\psi(x)\rangle)(\lambda_1(x, R(x)y) + \lambda_1(x,y))\nu(\mathrm{d}y) &= \int \lambda_1(x,y)^2 \nu(\mathrm{d}y) \\ &\quad - \int \lambda_1(x, R(x)y)^2 \nu(\mathrm{d}y) \\ &= 0\end{aligned}$$

which was shown in (4.52). □

Lemma 4.51. *It holds that*

$$\begin{aligned}[\mathcal{L}_D^*, [\mathcal{L}_R^*, \mathcal{L}_D^*]]\mu(x,v) = \lambda_r \mu(x,v) \Big(&\langle v, \nabla^2\psi(x) v\rangle - \langle v, \nabla\psi(x)\rangle^2 \\ &+ b \operatorname{tr}\left(\nabla^2\psi(x) - \nabla\psi(x)\nabla\psi(x)^T\right) \Big)\end{aligned}$$

Proof. By Lemma 4.41

$$\begin{aligned}[\mathcal{L}_D^*, [\mathcal{L}_R^*, \mathcal{L}_D^*]]\mu(x,v) &= -\lambda_r \mathcal{L}_D^*(\langle v, \nabla\psi(x)\rangle \mu(x,v)) - [\mathcal{L}_R^*, \mathcal{L}_D^*](\langle v, \nabla\psi(x)\rangle\mu(x,v)) \\ &= \lambda_r \mu(x,v) \Bigg(\langle v, \nabla^2\psi(x) v\rangle - \langle v, \nabla\psi(x)\rangle^2 \\ &\quad + \int (\langle y, \nabla^2\psi(x) y\rangle - \langle y, \nabla\psi(x)\rangle^2)\nu(\mathrm{d}y) \Bigg).\end{aligned}$$

The statement follows by Equation (4.51). □

Lemma 4.52. *It holds that*

$$\begin{aligned}[\mathcal{L}_D^*, [\mathcal{L}_B^*, \mathcal{L}_D^*]] = \mu(x,v) \Big(&- 4\langle v, \nabla\psi(x)\rangle^2 \lambda_1(x, R(x)v) + 7\langle v, \nabla^2\psi(x)v\rangle \lambda_1(x, R(x)v) \\ &+ \langle v, \nabla_x(\langle v, \nabla^2\psi(x))v\rangle)\rangle + \langle R(x)v, \nabla^2\psi(x) R(x)v\rangle \lambda_1(x, R(x)v) \Big).\end{aligned}$$

Proof. By Lemma 4.39 together with Lemma 4.38

$$\begin{aligned}&[\mathcal{L}_D^*, [\mathcal{L}_B^*, \mathcal{L}_D^*]]\mu(x,v) = \\ &= \mathcal{L}_D^* \Big(\mu(x,v) \Big(\langle v, \nabla\psi(x)\rangle \big(\langle v, \nabla\psi(x)\rangle - |\langle v, \nabla\psi(x)\rangle| \big) - \langle v, \nabla^2\psi(x)v\rangle \Big) \Big) \qquad (\dagger) \\ &\quad - [\mathcal{L}_B^*, \mathcal{L}_D^*](\langle v, \nabla\psi(x)\rangle \mu(x,v)). \qquad (\dagger\dagger) \\ &= (\dagger) - (\dagger\dagger).\end{aligned}$$

Consider the two terms separately, starting from the first one:

$$(\dagger) = \mu(x,v)\langle v, \nabla\psi(x)\rangle \Big(\langle v, \nabla\psi(x)\rangle \big(\langle v, \nabla\psi(x)\rangle - |\langle v, \nabla\psi(x)\rangle| \big) - \langle v, \nabla^2\psi(x))v\rangle \Big)$$

$$
\begin{aligned}
&- \mu(x,v)\langle v, 2\langle v, \nabla\psi(x)\rangle \nabla^2\psi(x)v - 2\nabla^2\psi(x)v|\langle v, \nabla\psi(x)\rangle| - \nabla_x(\langle v, \nabla^2\psi(x))v\rangle)\rangle \\
&= \mu(x,v)\Big(-2\langle v, \nabla\psi(x)\rangle^2\lambda_1(x,R(x)v) + \langle v, \nabla^2\psi(x)\rangle(-3\langle v, \nabla\psi(x)\rangle + 2|\langle v, \nabla\psi(x)\rangle|) \\
&\quad + \langle v, \nabla_x(\langle v, \nabla^2\psi(x))v\rangle)\rangle\Big).
\end{aligned}
$$

The second term (††) is the same as term (**) in the proof of Lemma 4.49. The statement follows taking the difference of the two terms (†) and (††) and using Lemma 4.38. □

4.E.2 Proof of Proposition 4.20

Consider symmetric splitting schemes of the form

$$
e^{\delta\mathcal{L}_S} = e^{\frac{\delta}{2}\mathcal{L}_A}e^{\frac{\delta}{2}\mathcal{L}_B}e^{\delta\mathcal{L}_C}e^{\frac{\delta}{2}\mathcal{L}_B}e^{\frac{\delta}{2}\mathcal{L}_A}.
$$

We have by the Baker-Campbell-Haussdorff formula

$$
\begin{aligned}
\mathcal{L}_S^* &= \mathcal{L}^* + \frac{\delta^2}{12}\Big([\mathcal{L}_C^*,[\mathcal{L}_C^*,\mathcal{L}_A^*+\mathcal{L}_B^*]] + [\mathcal{L}_B^*,[\mathcal{L}_B^*,\mathcal{L}_A^*]] + [\mathcal{L}_C^*,[\mathcal{L}_B^*,\mathcal{L}_A^*]] \\
&\quad + [\mathcal{L}_B^*,[\mathcal{L}_C^*,\mathcal{L}_A^*]] - \frac{1}{2}[\mathcal{L}_B^*,[\mathcal{L}_B^*,\mathcal{L}_C^*]] - \frac{1}{2}[\mathcal{L}_A^*,[\mathcal{L}_A^*,\mathcal{L}_C^*]] - \frac{1}{2}[\mathcal{L}_A^*,[\mathcal{L}_A^*,\mathcal{L}_B^*]]\Big) + \mathcal{O}(\delta^4) \\
&= \mathcal{L}^* + \delta^2\mathcal{L}_2^* + \mathcal{O}(\delta^4).
\end{aligned}
$$

where $\mathcal{L}^* = \mathcal{L}_A^* + \mathcal{L}_B^* + \mathcal{L}_C^*$. Therefore it is sufficient to use the commutators of Section 4.E.1. Observe that $\mathcal{L}_{BPS}^*\mu(x,v) = 0$. Let us start with $\mathcal{L}_{DBRBD}^*$:

$$
\begin{aligned}
&\mathcal{L}_{DBRBD}^*\mu(x,v) = \frac{\delta^2}{12}\Big([\mathcal{L}_R^*,[\mathcal{L}_R^*,\mathcal{L}_D^*+\mathcal{L}_B^*]] + [\mathcal{L}_B^*,[\mathcal{L}_B^*,\mathcal{L}_D^*]] + [\mathcal{L}_R^*,[\mathcal{L}_B^*,\mathcal{L}_D^*]] \\
&+ [\mathcal{L}_B^*,[\mathcal{L}_R^*,\mathcal{L}_D^*]] - \frac{1}{2}[\mathcal{L}_B^*,[\mathcal{L}_B^*,\mathcal{L}_R^*]] - \frac{1}{2}[\mathcal{L}_D^*,[\mathcal{L}_D^*,\mathcal{L}_R^*]] - \frac{1}{2}[\mathcal{L}_D^*,[\mathcal{L}_D^*,\mathcal{L}_B^*]]\Big) + \mathcal{O}(\delta^4) \\
&= \frac{\delta^2}{12}\ \mu(x,v)\times \\
&\times \Big(\frac{3}{2}\lambda_r\Big(b\,\mathrm{tr}\big(\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x)\big) + 2\langle v, \nabla\psi(x)\rangle\lambda_1(x,R(x)v) + \langle v, \nabla^2\psi(x)v\rangle\Big) \\
&+ \frac{3}{2}\lambda_1(x,R(x)v)\Big(\langle v, \nabla^2\psi(x)v\rangle - \langle R(x)v, \nabla^2\psi(x)R(x)v\rangle\Big) + \frac{1}{2}\langle v, \nabla_x(\langle v, \nabla^2\psi(x)v\rangle)\rangle\Big) \\
&+ \mathcal{O}(\delta^4).
\end{aligned}
$$

Then focus on $\mathcal{L}_{BDRDB}^*$:

$$
\begin{aligned}
\mathcal{L}_{BDRDB}^*\mu(x,v) &= \frac{\delta^2}{12}\Big([\mathcal{L}_R^*,[\mathcal{L}_R^*,\mathcal{L}_D^*+\mathcal{L}_B^*]] + [\mathcal{L}_D^*,[\mathcal{L}_D^*,\mathcal{L}_B^*]] + [\mathcal{L}_R^*,[\mathcal{L}_D^*,\mathcal{L}_B^*]] \\
&\quad + [\mathcal{L}_D^*,[\mathcal{L}_R^*,\mathcal{L}_B^*]] - \frac{1}{2}[\mathcal{L}_D^*,[\mathcal{L}_D^*,\mathcal{L}_R^*]] - \frac{1}{2}[\mathcal{L}_B^*,[\mathcal{L}_B^*,\mathcal{L}_R^*]] - \frac{1}{2}[\mathcal{L}_B^*,[\mathcal{L}_B^*,\mathcal{L}_D^*]]\Big) \\
&\quad + \mathcal{O}(\delta^4)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\delta^2}{12}\ \mu(x,v)\times \\
&\times \Bigg(-\frac{3}{2}\lambda_r\Big(b\,\mathrm{tr}\big(\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x)\big) + 2\langle v, \nabla\psi(x)\rangle\lambda_1(x,R(x)v) + \langle v, \nabla^2\psi(x)v\rangle\Big) \\
&+ 3\lambda_1(x,R(x)v)\Big(-2\langle v,\nabla^2\psi(x)v\rangle + \langle v,\nabla\psi(x)\rangle^2\Big) - \langle v, \nabla(\langle v,\nabla^2\psi(x)v\rangle)\rangle\Bigg) \\
&\quad + \mathcal{O}(\delta^4).
\end{aligned}
$$

Consider $\mathcal{L}^*_{RDBDR}$:

$$
\begin{aligned}
\mathcal{L}^*_{RDBDR}\mu(x,v) &= \frac{\delta^2}{12}\Big([\mathcal{L}^*_B,[\mathcal{L}^*_B,\mathcal{L}^*_R+\mathcal{L}^*_D]] + [\mathcal{L}^*_D,[\mathcal{L}^*_D,\mathcal{L}^*_R]] + [\mathcal{L}^*_B,[\mathcal{L}^*_D,\mathcal{L}^*_R]] \\
&\quad + [\mathcal{L}^*_D,[\mathcal{L}^*_B,\mathcal{L}^*_R]] - \frac{1}{2}[\mathcal{L}^*_D,[\mathcal{L}^*_D,\mathcal{L}^*_B]] - \frac{1}{2}[\mathcal{L}^*_R,[\mathcal{L}^*_R,\mathcal{L}^*_B]] - \frac{1}{2}[\mathcal{L}^*_R,[\mathcal{L}^*_R,\mathcal{L}^*_D]]\Big) + \mathcal{O}(\delta^4) \\
&= \frac{\delta^2}{12}\ \mu(x,v)\times \\
&\times \Big(\frac{3}{2}\lambda_1(x,R(x)v)\Big(\langle v,\nabla^2\psi(x)v\rangle - \langle R(x)v,\nabla^2\psi(x)R(x)v\rangle\Big) + \frac{1}{2}\langle v,\nabla_x(\langle v,\nabla^2\psi(x)v\rangle)\rangle\Big) \\
&+ \mathcal{O}(\delta^4).
\end{aligned}
$$

Finally focus on $\mathcal{L}^*_{DRBRD}$:

$$
\begin{aligned}
\mathcal{L}^*_{DRBRD}\mu(x,v) &= \frac{\delta^2}{12}\Big([\mathcal{L}^*_B,[\mathcal{L}^*_B,\mathcal{L}^*_D+\mathcal{L}^*_R]] + [\mathcal{L}^*_R,[\mathcal{L}^*_R,\mathcal{L}^*_D]] + [\mathcal{L}^*_B,[\mathcal{L}^*_R,\mathcal{L}^*_D]] \\
&\quad + [\mathcal{L}^*_R,[\mathcal{L}^*_B,\mathcal{L}^*_D]] - \frac{1}{2}[\mathcal{L}^*_R,[\mathcal{L}^*_R,\mathcal{L}^*_B]] - \frac{1}{2}[\mathcal{L}^*_D,[\mathcal{L}^*_D,\mathcal{L}^*_B]] - \frac{1}{2}[\mathcal{L}^*_D,[\mathcal{L}^*_D,\mathcal{L}^*_R]]\Big) \\
&\quad + \mathcal{O}(\delta^4) \\
&= \frac{\delta^2}{12}\ \mu(x,v)\times \\
&\times \Bigg(\frac{3}{2}\lambda_r\Big(b\,\mathrm{tr}\big(\nabla\psi(x)\nabla\psi(x)^T - \nabla^2\psi(x)\big) + \langle v,\nabla\psi(x)\rangle\big(3\lambda_1(x,R(x)v) + \lambda_1(x,v)\big) \\
&\quad + \langle v,\nabla^2\psi(x)v\rangle\Big) + \frac{3}{2}\lambda_1(x,R(x)v)\Big(\langle v,\nabla^2\psi(x)v\rangle - \langle R(x)v,\nabla^2\psi(x)R(x)v\rangle\Big) \\
&\quad + \frac{1}{2}\langle v,\nabla(\langle v,\nabla^2\psi(x)v\rangle)\rangle + \frac{3}{2}\lambda_r^2\langle v,\nabla\psi(x)\rangle\Bigg) + \mathcal{O}(\delta^4).
\end{aligned}
$$

Part III

Transformations of piecewise deterministic Markov processes

Chapter 5

Adaptive schemes for piecewise deterministic Monte Carlo algorithms

5.1 Introduction

Although PDMC sampling methods offer some important benefits, computation remains expensive, which requires us to investigate possible performance improvements. In particular, a strong performance degradation is observed when the target distribution π is anisotropic. Figure 5.1 illustrates this phenomenon in the case of Gaussian targets as a function of the correlation between all components. The performance drop occurs due to a combination of decreasing accuracy of the estimates and increasing computational complexity of the algorithms, which is implied by the growing number of velocity change events. Our idea to improve this issue is to let the process learn (part of) the covariance matrix Σ_π and take advantage of it to enhance the mixing properties. The covariance estimate is used to linearly transform the target in such a way that it becomes more isotropic, i.e. with unitary covariance matrix. The standard samplers are then run targeting the transformed version of π, and the obtained sample is finally re-transformed to be approximately from π. The procedure is applied iteratively, and once a new estimate of Σ_π is computed, it is used to define the linear transformation of π. The estimate will eventually be close to the true covariance matrix and the process targets an isotropic version of π. This scheme can also be interpreted as an application of a linear transformation directly to the standard ZZS and BPS. The natural applications of this procedure are then targets with elliptical level curves, although performance improvements can be observed also for distributions that deviate from this class.

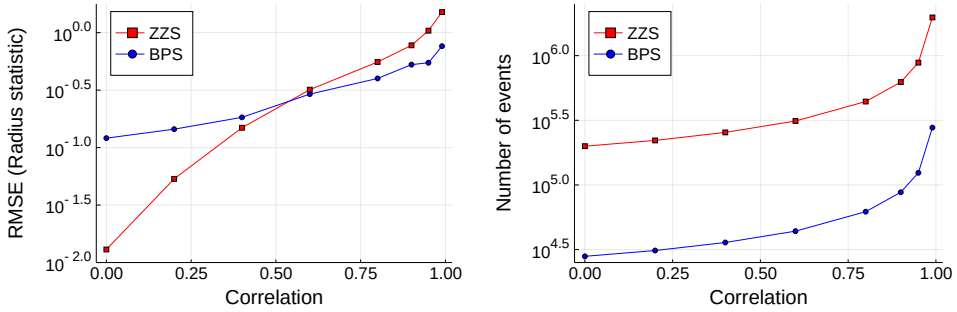


Figure 5.1: Root mean squared error (RMSE) for the radius statistic (left) and number of events (right) for 50-dimensional Gaussian targets for several values of the correlation between all components. The continuous time horizon is $T = 10^4$.

Furthermore, we address the problem of automatically choosing the rate of velocity refreshments. In [27] the authors consider a BPS with a specific target and derive that in the limit it is optimal to have a ratio of number of refreshments over number of total events of 0.7812. In this paper we use this criterion as a basis to define an adaptive algorithm that iteratively adjusts the refreshment rate to obtain the right ratio. This same adaptive scheme can be applied to the ZZS. Indeed it could be the case that adding velocity refreshments to the ZZ process leads to a faster convergence to the invariant measure, although it comes with a larger asymptotic variance. For an analysis of these two results we refer to [155] and [19] respectively.

Both the schemes we discussed take advantage of what the process has learned up until the current time to tune a parameter or to improve the performance. This idea is at the core of adaptive Markov chain Monte Carlo algorithms. For an introduction to this area we refer to [3, 136], while standard results on convergence of these methods can be found among others in [8, 71, 72, 135]. It is well known that adaptive MCMC algorithms can lose the right invariant measure if not applied with care (see for instance [135] and the examples therein). Therefore we study in depth the convergence properties of the proposed algorithms. To fit into the existing adaptive MCMC literature we let the adaptation happen at fixed points in time. The main challenge consists of establishing a simultaneous geometric drift condition for a family of BPS's (see Lemma 5.24) and a simultaneous small set condition for a family of ZZ processes (see Lemma 5.20). The former result is obtained taking advantage of the Lyapunov function found in [64], while the latter is proved by extending on the one-dimensional case. The ergodicity and a law of large numbers for the proposed adaptive PDMC algorithms are then established in Theorem 5.14 and Theorem 5.18.

In Section 5.2 we introduce the adaptive schemes, while in Section 5.3 the theoretical aspects of the algorithms are studied. The skeleton of the proofs of the two main theorems can be found in Section 5.4, while all other proofs can be found in

Appendix 5.B. In Section 5.5 the adaptive BPS and ZZS are tested empirically on various Gaussian targets, on a Bayesian logistic regression problem with correlated data, and on a mixture of two Gaussian distributions. The details on the implementation of adaptive PDMC algorithms (with and without subsampling), as well as an alternative adaptive scheme for the refreshment rate, can be found in Appendix 5.A.

5.2 The adaptive schemes

We are interested in building adaptive strategies to make the ZZS and the BPS choose the refreshment rate themselves and/or converge faster to the target density. We begin with an introduction of the standard versions of both samplers, followed by a characterisation of the preconditioned processes in Section 5.2.2 and a discussion on the choice of the transformation matrix in Section 5.2.3. Finally, the adaptive algorithms are defined in Section 5.2.4.

5.2.1 The standard ZZS and BPS

Let the target density π be defined on $\mathsf{X} \subset \mathbb{R}^d$ as

$$\pi(\xi) = \frac{1}{Z} \exp\left(-\psi(\xi)\right),$$

where $\psi(\xi)$ is called potential or energy function, and $\xi \in \mathsf{X}$. Recall the ZZS and BPS introduced in Examples 2.11 and 2.12. Throughout this chapter, the state at time t of both the standard ZZS and BPS is denoted as (Ξ_t, Θ_t) in order to distinguish them from preconditioned and adaptive algorithms. We shall distinguish between the Zig-Zag Sampler (i.e. a PDMC algorithm) and the Zig-Zag Process (i.e. the Markov process on which the algorithm is based).

5.2.2 Applying a linear transformation to the ZZS and BPS

In this section we suppose a matrix $M \in \mathbb{R}^{d\times d}$ is given. We then wish to define a transformation scheme encoded by M, which we should think of being such that, for a suitable choice of M, it gives a "more isotropic" version of the target, and analyse its effects on the PDMC samplers.

The transformation scheme encoded by M consists of a linear transformation of the state space, which defines a new target distribution $\tilde{\pi}_M$ given by

$$\tilde{\pi}_M(\xi) \coloneqq \frac{1}{\tilde{Z}_M} \exp(-\tilde{\psi}_M(\xi)),$$

with $\tilde{\psi}_M(\xi) = \psi(M\xi)$ and $\tilde{Z}_M = Z/|\det M|$. The idea is to apply the transformation to the target distribution π and simulate the standard PDMC sampler $(\Xi_t, \Theta_t)_{t\geq 0}$ with the resulting target $\tilde{\pi}_M$. Then the last thing to do is transform the obtained

sample, which is approximately from $\tilde{\pi}_M$, by applying the inverse transformation. For this reason it is important that the matrix M is invertible, and thus that we can go from one state space to the other. This procedure is illustrated in Figure 5.2. An equivalent option is to simulate directly the process $(X_t, \Theta_t)_{t\geq 0}$ that results from the scheme in Figure 5.2. We will conveniently alternate between these two formulations when studying the ergodic properties of the samplers, while we will use the latter formulation for our experiments. The dynamics of process $(X_t, \Theta_t)_{t\geq 0}$ are studied in the remainder of this section.

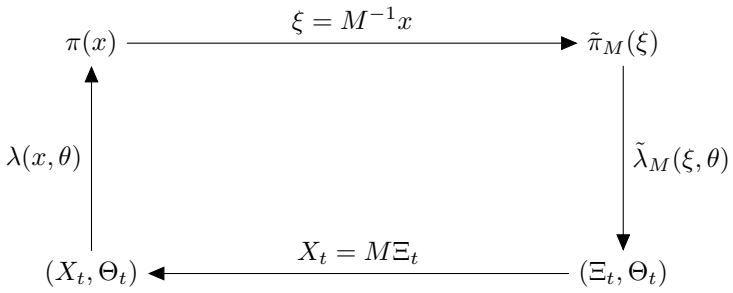


Figure 5.2: Transformation scheme.

Let us first focus on the case in which $(\Xi_t, \Theta_t)_{t\geq 0}$ is a standard ZZS with excess switching rate γ. In this case the switching rates are $\tilde{\lambda}_{M,i}(\xi,\theta) = (\theta_i \partial_i \tilde{\psi}_M(\xi))_+ + \gamma_i(\xi,\theta)$ for $i = 1,\dots,d$. Note that, unless stated otherwise, we will always use a tilde to indicate quantities related to the standard PDMC samplers with transformed target. In the next proposition we find the generator of the preconditioned ZZS. For a characterisation of the domain of the extended generator we refer to [49, Theorem 26.14].

Proposition 5.1. *Let $M \in \mathbb{R}^{d\times d}$ be an invertible matrix. Let $(\Xi_t, \Theta_t)_{t\geq 0}$ be a standard ZZS with target $\tilde{\pi}_M$ and excess switching rates $\gamma : E \to \mathbb{R}^d_+$. The process $(X_t, \Theta_t)_{t\geq 0} = (M\Xi_t, \Theta_t)_{t\geq 0}$ has extended generator $(\mathcal{L}_M, \mathcal{D}(\mathcal{L}_M))$ where for any $h \in \mathcal{D}(\mathcal{L}_M)$*

$$\mathcal{L}_M\, h(x,\theta) = \langle M\theta, \nabla_x h(x,\theta)\rangle + \sum_{i=1}^{d} \lambda_i^M(x,\theta)(h(x,R_i\theta) - h(x,\theta)), \tag{5.1}$$

in which for $i = 1,\dots,d$

$$\lambda_i^M(x,\theta) = \tilde{\lambda}_{M,i}(M^{-1}x,\theta) = (\theta_i \langle M_i, \nabla\psi(x)\rangle)_+ + \gamma_i(M^{-1}x,\theta), \tag{5.2}$$

where M_i denotes the i-th column of M.

Proof. All the proofs of this section can be found in Appendix 5.B. □

Proposition 5.1 shows that the transformed process is again a PDMP with linear trajectories between jumps. The transformation affects the velocity of the process, which is now $v = M\theta$, and the switching intensities, which are as defined in (5.2). In particular, the available velocities for a fixed transformation M are in the following set:

$$V \coloneqq \big\{v : v = M\theta,\, \theta \in \{-1, +1\}^d\big\}.$$

When a switch of the i-th velocity of the underlying standard ZZ process happens, the velocities of the transformed process change according to the operator $\overline{R}_i v = MR_i\theta = MR_i(M^{-1}v)$ for $i = 1, \dots, d$. Therefore, all components of the velocity are possibly affected by any single event. If M is diagonal the behaviour is more similar to the standard ZZ process and in particular we have that $\overline{R}_i \equiv R_i$. In the proposition below we check that $(X_t, \Theta_t)_{t\geq 0}$ targets the correct density function.

Proposition 5.2. *Consider the same setting of Proposition 5.1. Then, for any invertible $M \in \mathbb{R}^{d\times d}$, the modified ZZ process $(X_t, \Theta_t)_{t\geq 0}$ has invariant distribution $\mu = \pi \times \mathrm{Unif}(\{-1, +1\}^d)$.*

Let us now apply the same transformation scheme shown in Figure 5.2 to the BPS. In this case the switching rate of the standard BPS with target $\tilde{\pi}_M$ is $\tilde{\lambda}_M(\xi, \theta) = \langle \theta, \nabla\tilde{\psi}_M(\xi)\rangle_+$, while reflections on the level curves of $\tilde{\psi}_M$ are obtained by applying operator $\tilde{R}_M$. The following result, analogous to Proposition 5.1, studies the transformed process.

Proposition 5.3. *Let $M \in \mathbb{R}^{d\times d}$ be an invertible matrix. Let $(\Xi_t, \Theta_t)_{t\geq 0}$ be a standard BPS with target $\tilde{\pi}_M$ and refreshment rate $\lambda_r : E \to \mathbb{R}_+$. The process $(X_t, \Theta_t)_{t\geq 0} = (M\Xi_t, \Theta_t)_{t\geq 0}$ has extended generator $(\mathcal{L}_M, \mathcal{D}(\mathcal{L}_M))$ where for any $h \in \mathcal{D}(\mathcal{L}_M)$*

$$\begin{aligned}\mathcal{L}_M h(x,\theta) = \langle M\theta, \nabla_x h(x,\theta)\rangle + \lambda_M(x,\theta)\big(h(x, R_M(x)\theta) - h(x,\theta)\big) \\ + \lambda_r(M^{-1}x, \theta)\int (h(x,\theta') - h(x,\theta))\nu(\mathrm{d}\theta'),\end{aligned} \tag{5.3}$$

where we defined

$$\lambda_M(x,\theta) = \tilde{\lambda}(M^{-1}x, \theta) = \langle M\theta, \nabla\psi(x)\rangle_+ \tag{5.4}$$

and

$$R_M(x)\theta = \tilde{R}_M(M^{-1}x)\theta = \theta - 2\frac{\langle M^T\nabla\psi(x), \theta\rangle}{||M^T\nabla\psi(x)||^2}M^T\nabla\psi(x). \tag{5.5}$$

Once again the true velocity of the process is $v = M\theta$. When a velocity refreshment takes place the new θ is sampled from $\nu = \mathcal{N}(0_d, \mathbb{1}_d)$, while the new velocity v is $v = M\theta \sim \mathcal{N}(0_d, MM^T)$. Observe also that the reflection rate is $\lambda_M(x,\theta) = \langle v, \nabla\psi(x)\rangle_+$ and thus preserves the same structure as in the standard BPS. It follows that the complexity of the simulation of event times remains unchanged. Finally consider the reflection operator in (5.5). This corresponds to a reflection in the opposite direction

to the gradient in the transformed space, i.e. $\nabla_\xi \tilde{\psi}(\xi) = M^T \nabla \psi(x)$. After the bounce the process moves with velocity

$$v = M(R_M(x)\theta) = v - 2\frac{\langle \nabla\psi(x), v\rangle}{||M^T\nabla\psi(x)||^2} MM^T\nabla\psi(x).$$

This implies that

$$\langle v, \nabla\psi(x)\rangle = \langle M(R_M(x)\theta), \nabla\psi(x)\rangle = -\langle M\theta, \nabla\psi(x)\rangle = -\langle v, \nabla\psi(x)\rangle.$$

Proposition 5.4. *Consider the same setting of Proposition 5.3. Then, for any invertible $M \in \mathbb{R}^{d\times d}$, the transformed BPS $(X_t, \Theta_t)_{t\geq 0}$ has invariant distribution $\mu = \pi \times \nu$.*

5.2.3 Choosing the transformation matrix

As explained above, we wish to transform the target to mitigate its anisotropies. To this end, some alternative choices of the transformation matrix M are the following:

a) $M = \sqrt{\mathrm{Cov}_\pi(X)}$: this transformation is such that the target $\tilde{\pi}_M$ has unitary covariance matrix. The downside of this choice is the additional $\mathcal{O}(d^3)$ computations that are introduced by the calculation of the square root of the covariance;

b) M is a rotation matrix ($\det M = 1$) such that the transformed density has a certain angle. Although an interesting case, it is not clear whether there is an optimal angle that speeds up the convergence;

c) M is the diagonal matrix with $M_{ii} = \sqrt{\mathrm{Var}_\pi(X_i)}$ for any $1 \leq i \leq d$. This choice introduces $\mathcal{O}(d)$ computations due to the square root of the variances, which is a negligible additional computational burden. However, correlations in the target are not picked up and only a rescaling of the axes is performed. The main advantage of this choice is that the scenario in which some components are explored quickly and others slowly is avoided.

Both the first and the third option can potentially change the expected number of switching events, and this could be an inconvenience in certain settings. It is not difficult to modify these transformations in such a way that the expected switching rate is enforced to be close to that of the original standard PDMC algorithm. For example, consider the transformed BPS with generator as in Proposition 5.3. Then in stationarity we have

$$\mathbb{E}_\mu(\langle M\Theta, \nabla\psi(X)\rangle_+) = \mathbb{E}_\pi\|M^T\nabla\psi(X)\|_2 \leq \|M\|_2\, \mathbb{E}_\pi\|\nabla\psi(X)\|_2.$$

Since the standard case corresponds to $M = \mathbb{I}_d$, we can normalise any M by dividing it by its Frobenius norm. Then the upper bound is the same for all such choices of M and the expected switching intensity will be close to the standard case. This does not make a difference from a computational point of view as it just amounts to reparametrization of the time parameter, but prevents unpredictable behaviour of the algorithm.

Naturally the options above are not available in practice as the covariance matrix is unknown. It is the goal of the next section to propose an adaptive scheme that overcomes this issue.

5.2.4 Adaptive PDMC algorithms

In the previous sections we defined the transformation scheme and we discussed the effect it has on the underlying process, together with different choices of the preconditioning matrix. We now describe how this idea can be applied in practice by designing an adaptive PDMC algorithm. Our general strategy is to simulate the process in continuous time and store the states of the process at discrete times. Then at predefined times the stored states are used to update the adaptation parameters. In addition to the adaptive preconditioner, we incorporate an adaptation of the refreshment rate, which makes its choice automatic.

Let us then define a family of Markov semigroups by $\mathcal{P} := \{(P_\Gamma^t)_{t\geq 0} : \Gamma \in \mathcal{Y}\}$, in which Γ is the adaptation parameter, $\mathcal{Y}$ is a compact space, and $(P_\Gamma^t)_{t\geq 0}$ is the semigroup of a modified ZZS or BPS. The modification is given by the adaptation parameter, which is then $\Gamma = (M, \lambda_r)$ for BPS and $\Gamma = (M, \gamma)$ for ZZS. Thus $\mathcal{Y}$ is a suitable compact space of preconditioners and refreshment rates/excess switching rates. Naturally, it is also possible to choose $\Gamma = M$ or $\Gamma = \lambda_r$ only. Now that we have defined a family of Markov processes, we define a rule that establishes how to choose a $P_\Gamma \in \mathcal{P}$ at every iteration. Let us begin by introducing a discretisation step Δt, which defines a discretisation of the time variable. At each time step $n \in \mathbb{N}$, which corresponds to continuous time $t = n\Delta t$, the adaptive scheme can update the parameter Γ_n based on the new information available, that is the new state of the process (X_n, Θ_n). This defines a random sequence $\{\Gamma_n\}_{n\geq 0}$. Once Γ_n is computed, the next state of the process is given by $(X_{n+1}, \Theta_{n+1}) \sim P_{\Gamma_n}^{\Delta t}((X_n, \Theta_n), \cdot)$. Then one updates the parameter, obtains the next state of the process, and so on.

The definition above defines the core ideas, which are written in pseudo-code form in Algorithm 16. A few issues remain to be clarified. A first question is how to simulate the PDMP semigroup of either ZZS or BPS. Details on how the processes can be simulated in the case of a target with dominated Hessian can be found in Appendix 5.A.1. In a big data setting (large number of observations, moderate dimensionality of the problem) it can be beneficial to take advantage of subsampling techniques that can be implemented with PDMC algorithms. In Appendix 5.A.2 details can be found on how to make use of subsampling in the context of the adaptive schemes here discussed. For further information on the general implementation of ZZS and BPS we refer to [23, 32]. A second aspect of Algorithm 16 we focus on is the introduction of set B, and thus of the auxiliary sequence of random variables $(Q_n)_{n\geq 0}$. The idea is to update the adaptation parameter Γ_n only if $(X_n, \Theta_n) \in B$. This is useful from a theoretical point of view as it ensures that the process remains bounded in probability. Note that set B is defined by the user and can be chosen large. The auxiliary variable Q_n is updated even if the process is outside of B and then, as soon as the process enters B,

Algorithm 16: Adaptive PDMC sampler

Input : family of kernels $\mathcal{P} = \{P_\Gamma : \Gamma \in \mathcal{Y}\}$, initial condition $(x, \theta) \in E$, $\Gamma_0 \in \mathcal{Y}$, set B, Δt, $\{p_n\}_{n\geq 0}$, number of steps N.
Output: Chain $\{X_n, \Theta_n\}_{n=0}^N$.
Initialise $n = 0$, $(X_0, \Theta_0) = (x, \theta)$, $Q_0 = \Gamma_0$;
while $n \leq N$ **do**
  $(X_{n+1}, \Theta_{n+1}) \sim P_{\Gamma_n}^{\Delta t}((X_n, \Theta_n), \cdot)$;
  $Q_{n+1} = \textbf{update}(Q_n, (X_{n+1}, \Theta_{n+1}))$;
  if $(X_{n+1}, \Theta_{n+1}) \in B$ **then**
    With probability p_n, set $\Gamma_{n+1} = Q_{n+1}$, else set $\Gamma_{n+1} = \Gamma_n$;
  end
  Set $n = n + 1$;
end

or if it was already in B, we let $\Gamma_n = Q_n$. A third characteristic of Algorithm 16 is the sequence $\{p_n\}_{n\geq 0}$. This is a sequence for which $p_n \in [0, 1]$ for all $n \in \mathbb{N}$ and $p_n \to 0$ as $n \to \infty$. The meaning is that at time step n we update the parameter Γ_n with probability p_n (assuming $(X_n, \Theta_n) \in B$), and with remaining probability $(1 - p_n)$ we set $\Gamma_n = \Gamma_{n-1}$. This choice is helpful when proving ergodicity of the adaptive scheme, as it enforces that the quantity of adaptation diminishes and eventually vanishes.

The function $\textbf{update}(Q_n, (X_{n+1}, \Theta_{n+1}))$ outputs the updated parameter given the new observation (X_{n+1}, Θ_{n+1}). As suggested in [3], the estimation of the covariance matrix can be done sequentially, or online, by applying

$$\begin{aligned} \hat{\mu}_{n+1} &= \hat{\mu}_n + r_{n+1}(X_{n+1} - \hat{\mu}_n), \\ \hat{\Sigma}_{n+1} &= \hat{\Sigma}_n + r_{n+1}((X_{n+1} - \hat{\mu}_n)(X_{n+1} - \hat{\mu}_n)^T - \hat{\Sigma}_n). \end{aligned} \tag{5.6}$$

Here $\{r_n\}_{n\geq 0}$ is a positive, decreasing sequence such that $r_n \to 0$ as $n \to \infty$. In our simulations we choose $r_n = 1/n$. Equation (5.6) is then used to define M_{n+1} such that $\hat{\Sigma}_{n+1} = M_{n+1}^T M_{n+1}$. The same principle can be used if one is not interested in estimating the full covariance matrix, but only the diagonal, or more generally only a subset of it. More advanced estimation techniques can be employed to preserve any existing conditional independence structure in the target, as discussed in [158]. We remark that $\hat{\Sigma}_{n+1}$ needs to be positive definite in order for M_{n+1} to be invertible as required. This property is achieved by choosing $\hat{\Sigma}_0 = \mathbf{1}_{d\times d}$, i.e. the identity matrix, and then observing that the second equation in (5.6) can be reformulated as

$$\hat{\Sigma}_{n+1} = (1 - r_{n+1})\hat{\Sigma}_n + r_{n+1}(X_{n+1} - \hat{\mu}_n)(X_{n+1} - \hat{\mu}_n)^T.$$

Indeed $(X_{n+1} - \hat{\mu}_n)(X_{n+1} - \hat{\mu}_n)^T$ is non-negative definite and by induction $\hat{\Sigma}_n$ is positive definite, and therefore $\hat{\Sigma}_{n+1}$ is itself positive definite. Moreover, in Section 5.3 we will see that to show ergodicity of the adaptive algorithms it is required that M_n

lies in a compact space of positive definite matrices. Observe that positive definiteness follows from the fact that M_n is the square root of a positive definite matrix, while it is sufficient to set bounds on the norm of M_n in order to force it to be in a compact space. In particular we can impose that M_n is not updated if the norm of the new estimate is outside of a user chosen interval $[M_{\min}, M_{\max}]$. As $M_{\min}$ and $M_{\max}$ can be chosen arbitrarily small and large respectively, this condition is not restrictive in practice, although the choice of the cut-off value may influence convergence properties of the algorithm. Then refreshment rate of the BPS is assumed to be constant and is updated iteratively as follows. At time step n, n_{refl} reflections took place and thus we estimate the average reflection rate as $\overline{\lambda}_{\text{refl}}(n) = n_{\text{refl}}/(n\Delta t)$. Therefore, using the optimality criterion in [27] we have

$$\frac{\lambda_r^n}{\lambda_r^n + \overline{\lambda}_{\text{refl}}(n)} = 0.7812 \implies \lambda_r^n = \frac{0.7812}{0.2188}\, \overline{\lambda}_{\text{refl}}(n). \tag{5.7}$$

An alternative adaptive strategy for the refreshment can be found in Appendix 5.B.2. The scheme above can be applied to the excess switching rate of the ZZS. Although the analysis in [19] suggests that the best choice in terms of asymptotic variance is $\gamma \equiv 0$, adding some diffusivity could speed up the convergence to the target measure. In practice the user can select the wanted ratio of velocity switches over total events and proceed as above. However, a criterion to choose this ratio is currently unavailable for ZZS, and thus in this paper we limit ourselves to a theoretical study of this option.

Finally, we remark that in practice it is not reasonable to update the parameters at every iteration. The main reason for this is the computational cost of such an operation. In the most general case, the task of learning all components of Σ takes $\mathcal{O}(d^2)$ operations, while the computation of its square root, which is needed to obtain the transformation matrix M, is an $\mathcal{O}(d^3)$ operation. Therefore it is rather inconvenient to perform this at every time step. To avoid this issue it is sufficient to define the adaptive scheme such that adaptations happen every n_{adap} time steps, where n_{adap} is a user-defined integer. A possible choice is for instance $n_{\text{adap}} = 1000$. This modification is beneficial also because it allows the process to explore the target distribution before updating the parameters. Similarly, it is reasonable to update the refreshment rate based on the previous n_{adap} time steps, as in the long term this allows to stabilise around the wanted ratio. The covariance matrix can be updated as in (5.6) by simply processing the entire batch of n_{adap} data points one at a time.

5.3 Theoretical results

In the context of adaptive MCMC algorithms, convergence to the target density is usually proved with simultaneous drift conditions and small set conditions. In Section 5.3.1 we introduce the notation and the main existing theorems we make use of, and we extend these results to more general conditions in Theorem 5.9. In Section 5.3.2 we state Theorem 5.11, which shows ergodicity for an adaptive MCMC algorithm

based on a continuous time process. In this result, the assumptions are formulated directly in continuous time. Finally, Theorems 5.14 and 5.18 in Section 5.3.3 show that the adaptive ZZS and the adaptive BPS discussed in Section 5.2.4 are ergodic and satisfy a weak law of large numbers under reasonable growth conditions on the potential.

5.3.1 Theory of adaptive MCMC

We denote the parameter that specifies the kernel as $\Gamma \in \mathcal{Y}$. At time step n a $\mathcal{Y}$-valued random variable Γ_n determines which transition kernel will be used to move to the next step. From here on each Markov transition kernel P_Γ is assumed to define a Markov chain that has μ as stationary measure, and moreover it is aperiodic and irreducible. An adaptive MCMC algorithm is then said to be *ergodic* if

$$\lim_{n\to\infty} \|P(Z_n \in \cdot \,|z_0, \Gamma_0) - \mu(\cdot)\|_{\mathrm{TV}} = 0 \qquad \text{for all } z_0 \in E, \Gamma_0 \in \mathcal{Y}, \tag{5.8}$$

where $||\cdot||_{\mathrm{TV}}$ is the total variation distance, i.e. $\|\mu - \nu\|_{\mathrm{TV}} = \sup_{A\subset E}|\mu(A) - \nu(A)|$. A crucial quantity turns out to be the ε-convergence time function $M_\varepsilon : E \times \mathcal{Y} \to \mathbb{N}$, defined as

$$M_\varepsilon(z, \Gamma) = \inf\{n \geq 1 : \|P_\Gamma^n(z, \cdot) - \mu(\cdot)\|_{\mathrm{TV}} \leq \varepsilon\}.$$

The next theorem, proved in [135], is arguably the most important result for establishing ergodicity of adaptive MCMC methods.

Theorem 5.5 (Theorem 2 in [135])**.** *Consider an adaptive MCMC algorithm on a state space E with adaption parameter in a space $\mathcal{Y}$. Let μ be stationary for P_Γ for each $\Gamma \in \mathcal{Y}$. The adaptive algorithm is ergodic if the two following conditions hold:*

(a) (Containment condition) For all $z_0 \in E, \Gamma_0 \in \mathcal{Y}$, and$\varepsilon > 0$ the sequence $\{M_\varepsilon(Z_n, \Gamma_n)\}_{n=0}^\infty$ is bounded in probability given z_0, Γ_0;

(b) (Diminishing adaptation) The following limit holds in probability:

$$\lim_{n\to\infty} \left(\sup_{z\in E} \|P_{\Gamma_{n+1}}(z, \cdot) - P_{\Gamma_n}(z, \cdot)\|_{\mathrm{TV}} \right) = 0. \tag{5.9}$$

The boundedness of $\{M_\varepsilon(Z_n, \Gamma_n)\}_{n=0}^\infty$ can be rephrased as for all $z_0 \in E, \Gamma_0 \in \mathcal{Y}$, $\delta > 0$, there exists $N \in \mathbb{N}$ such that $P(M_\varepsilon(Z_n, \Gamma_n) \leq N | z_0, \Gamma_0) \geq 1 - \delta$, for all $n \in \mathbb{N}$.

We are then interested in sufficient conditions that imply containment. A first case is the following, and was studied in [8].

Assumption 5.6 ([8])**.** *The family $\{P_\Gamma : \Gamma \in \mathcal{Y}\}$ is simultaneously geometrically ergodic (SGE), i.e. there are $C \in \mathcal{B}(E)$, some integer $n_0 \geq 1$, a function $V : E \to [1, \infty)$, $\delta > 0$, $0 < \lambda < 1$, and $b < \infty$, such that $\sup_{z\in C} V(z) < \infty$, $\mu(V) < \infty$, and*

(a) C is a uniform $(\nu_\Gamma, \delta, n_0)$-small set, i.e. for each Γ, there exists a probability measure $\nu_\Gamma(\cdot)$ on C such that $P_\Gamma^{n_0}(z, \cdot) \geq \delta\nu_\Gamma(\cdot)$ for all $z \in C$;

(b) (simultaneous geometric drift condition) $P_\Gamma V \leq \lambda V + b\mathbb{1}_C$ *for all* $\Gamma \in \mathcal{Y}$.

Then [8, Theorem 3] establishes that an SGE family satisfies the containment condition. In Section 5.3.3 we use this result to show that containment holds for the adaptive ZZS when the class of preconditioners is restricted to diagonal matrices.

In practice it is often hard to show that the family of Markov kernels is SGE, as it is not trivial to find a Lyapunov function that satisfies the simultaneous geometric drift condition. In [44] the authors introduced a way around this problem, although in a different context, and in [38] this was applied to adaptive MCMC. The fundamental idea is that it is possible to weaken the simultaneous drift condition by allowing adaptations only at time steps n at which the process Z_n is inside of a compact set B. This means that, defining an auxiliary random process $\{Q_n\}_{n\geq 1}$ that contains the current adaptation parameter independently of the position of Z_n, Γ_n is updated as

$$\Gamma_{n+1} = \begin{cases} \Gamma_n & \text{if } Z_{n+1} \notin B, \\ Q_{n+1} & \text{if } Z_{n+1} \in B. \end{cases} \tag{5.10}$$

This modification avoids unbounded detours of the process by sticking to the same ergodic kernel once the process exits a fixed compact set. The compact set can be chosen arbitrarily large, and therefore in most applications the process will not exit from it.

With this is mind we introduce the following sets of assumptions, which we show in Theorem 5.9 to be sufficient to enforce the containment condition.

Assumption 5.7. *Let* $\{P_\Gamma : \Gamma \in \mathcal{Y}\}$ *be a family of discrete time Markov chains with state space* E. *There are* $C \in \mathcal{B}(E)$, *an integer* $n_0 \geq 1$, *a class of functions* $\{V_\Gamma : E \to [1,\infty) : \Gamma \in \mathcal{Y}\}$, $\delta > 0$, $0 < \lambda < 1$, *and* $b < \infty$, *such that* $\sup_{z\in C,\, \Gamma\in\mathcal{Y}} V_\Gamma(z) < \infty$, $\mu(V_\Gamma) < \infty$, *and*

(a) C *is a uniform* $(\nu_\Gamma, \delta, n_0)$*-small set, i.e. for each* $\Gamma \in \mathcal{Y}$, *there exists a probability measure* $\nu_\Gamma(\cdot)$ *on* C *such that* $P_\Gamma^{n_0}(z,\cdot) \geq \delta\nu_\Gamma(\cdot)$ *for all* $z \in C$;

(b) for each $\Gamma \in \mathcal{Y}$, $z \in E$, $P_\Gamma V_\Gamma(z) \leq \lambda V_\Gamma(z) + b\mathbb{1}_C(z)$;

(c) the adaptation parameter is allowed to be updated only if the process is inside of a compact set B, *as defined in* (5.10).

Assumption 5.8. *Let* $\{P_\Gamma : \Gamma \in \mathcal{Y}\}$ *be a family of discrete time Markov chains with state space* E. *There exist* $\alpha, \lambda \in (0,1)$, $C_1 > 0$, $C_2 > 2C_1$, *a class of functions* $\{V_\Gamma : E \to [1,\infty) : \Gamma \in \mathcal{Y}\}$ *with* $\mu(V_\Gamma) < +\infty$, *such that*

(a) for each $\Gamma \in \mathcal{Y}$, *for all* $x, y \in E$ *such that* $V_\Gamma(x) + V_\Gamma(y) \leq C_2$ *it holds that*

$$\|P_\Gamma(x,\cdot) - P_\Gamma(y,\cdot)\|_{\mathrm{TV}} \leq 2(1-\alpha);$$

(b) for each $\Gamma \in \mathcal{Y}$ *and for any* $z \in E$, $P_\Gamma V_\Gamma(z) \leq \lambda V_\Gamma(z) + C_1(1-\lambda)$;

(c) the adaptation parameter is allowed to be updated only if the process is inside of a compact set B, as defined in (5.10).

Theorem 5.9. *Consider a family of discrete time Markov transition kernels $\{P_\Gamma : \Gamma \in \mathcal{Y}\}$. Assume that all kernels P_Γ are aperiodic, irreducible, and have stationary measure μ. Suppose the adaptive algorithm satisfies the diminishing adaptation condition, i.e. assumption (b) in Theorem 5.5, and let either Assumption 5.7 or Assumption 5.8 hold. Then the containment condition holds and the adaptive MCMC algorithm is ergodic.*

Proof. The proof can be found in Appendix 5.B.2. □

Remark 5.10. A weak law of large numbers (WLLN) for bounded and measurable functions follows immediately from containment and diminishing adaptation by Theorem 3.4 in [126]. Therefore under the conditions of Theorem 5.9 a WLLN holds.

5.3.2 Convergence properties of adaptive MCMC algorithms based on continuous time Markov processes

It could be the case, as it is in the present work, that one is interested in defining an adaptive scheme based on a family of continuous time Markov processes in continuous time. In this case a grid for the time variable needs to be introduced in order to indicate the times at which the adaptation occurs. In fact the adaptive chain only sees the process at times $m\Delta t$, where $\Delta t > 0$ is the step size and $m \in \mathbb{N}$. Although the resulting chain is in discrete time, it is in most cases easier to work directly with the continuous time process. The following result, which is analogous to Theorem 5.9, is helpful in this sense.

Theorem 5.11. *Consider a family of Markov processes with generators $\{\mathcal{L}_\Gamma : \Gamma \in \mathcal{Y}\}$, each being irreducible, aperiodic, and having μ as invariant measure. Consider a grid for the time variable with step Δt. Consider an adaptive scheme that at times $m\Delta t$, with $m \in \mathbb{N}$, chooses a process from the aforementioned family. Furthermore, suppose that the adaptive scheme satisfies the diminishing adaptation condition* (5.9) *for $P := P^{\Delta t}$. Finally assume one of the following two sets of conditions holds:*

1. *There exist a set $C \in \mathcal{B}(E)$, $t_0 > 0$, a class of functions $\{V_\Gamma : E \to [1, \infty) : \Gamma \in \mathcal{Y}\}$, $\delta > 0$, $A_1, A_2 > 0$, such that for each Γ, $\sup_{z \in C, \Gamma \in \mathcal{Y}} V_\Gamma(z) < \infty$, $\pi(V_\Gamma) < \infty$, and*

   (a) for each $\Gamma \in \mathcal{Y}$ there exists a probability measure ν_Γ such that $P_\Gamma^{t_0}(z, \cdot) \geq \delta \nu_\Gamma(\cdot)$ for all $z \in C$;

   (b) for each $\Gamma \in \mathcal{Y}$ and $z \in E$ it holds that $\mathcal{L}_\Gamma V_\Gamma(z) \leq -A_1 V_\Gamma(z) + A_2 \mathbb{1}_C(z)$;

   (c) it holds that $\Delta t = m t_0$, for some $m \in \mathbb{N}$.

2. *There exist* $A_1, A_2 > 0$, $C_2 > 2A_2/A_1$, *a class of functions* $\{V_\Gamma : E \to [1,\infty) : \Gamma \in \mathcal{Y}\}$ *with* $\pi(V_\Gamma) < +\infty$, *such that*

   (a) *for each* $\Gamma \in \mathcal{Y}$, *for all* $x, y \in E$ *such that* $V_\Gamma(x) + V_\Gamma(y) \leq C_2$, *there exists* $\alpha, t_0 > 0$ *such that*

   $$\|P_\Gamma^{t_0}(x,\cdot) - P_\Gamma^{t_0}(y,\cdot)\|_{\mathrm{TV}} \leq 2(1-\alpha);$$

   (b) *for each* $\Gamma \in \mathcal{Y}$ *and for any* $z \in E$, $\mathcal{L}_\Gamma V_\Gamma(z) \leq -A_1 V_\Gamma(z) + A_2$;

   (c) *it holds that* $\Delta t = t_0$.

If the adaptation parameter is allowed to be updated only if the process is inside of a compact set B, *as defined in* (5.10), *then the adaptive algorithm satisfies the containment condition and is thus ergodic.*

Proof. The proof of this theorem can be found in Appendix 5.B.3. □

Remark 5.12. The restrictions on the step size can be milder than as stated in Theorem 5.11. For instance, if the minorisation condition (1a) of Theorem 5.11 holds for all $t \geq t_0$, then one is free to choose any step size $\Delta t > 0$. Furthermore, in both cases the assumption that the parameter can be updated only if the process is inside of a compact set at the adaptation time can be dropped when a simultaneous geometric drift condition holds (Assumption 5.6(b)).

5.3.3 Convergence properties of adaptive PDMC algorithms

Relying on Theorem 5.11, in this section we turn our attention to the ergodicity of adaptive PDMC algorithms. The proofs of the two theorems are postponed to Section 5.4. First, let us consider the adaptive ZZS. We assume the following conditions on the potential.

Assumption 5.13 (Growth Condition 3 in [24])**.** $\psi \in \mathcal{C}^2(\mathbb{R}^d)$ *and*

$$\lim_{\|x\|\to\infty} \frac{\max(1, \|\nabla^2\psi(x)\|)}{\|\nabla\psi(x)\|} = 0, \qquad \lim_{\|x\|\to\infty} \frac{\|\nabla\psi(x)\|}{\psi(x)} = 0.$$

Let us now state the ergodicity result for the adaptive ZZS.

Theorem 5.14. *Let* $\mathcal{M}$ *be a compact set of positive definite matrices and let* Λ *be a set of excess switching rates* $\gamma : E \to \mathbb{R}^d_+$ *for which there are* $0 < \gamma_{\min} \leq \gamma_{\max} < \infty$ *such that for all* $\gamma \in \Lambda$

$$\gamma_{\min} \leq \gamma(x,\theta) \leq \gamma_{\max} \qquad \text{for all } (x,\theta) \in E.$$

Let $\mathcal{P} = \{P_{M,\gamma} : M \in \mathcal{M}, \gamma \in \Lambda\}$ *be a family of preconditioned Zig-Zag processes with generators defined by Equation* (5.1). *Suppose Assumption 5.13 holds and assume either of the following conditions holds:*

a) $\mathcal{M} = \{M \in \mathbb{R}^{d\times d} : M_{ii} \in [a,b], M_{jk} = 0 \quad \textit{for all } j \neq k\}$ *with* $b > a > 0$ *;*

b) $\mathcal{M}$ *has no additional restrictions but adaptations are allowed only inside of a compact set* B*,*

and let Δt *be the discretisation step. Then the containment condition holds. Moreover, if the adaptive strategy is as described in Section 5.2.4 and is such that* $p_n \to 0$ *as* $n \to \infty$*, then the diminishing adaptation condition holds and thus for* $\mu = \pi \times \mathrm{Unif}(\Theta)$*:*

$$\lim_{n\to\infty} ||\mathbb{P}((X_n, \Theta_n) \in \cdot \,|x_0, \theta_0, \gamma_0) - \mu(\cdot)||_{TV} = 0 \qquad \textit{for all } (x_0,\theta_0) \in E, \gamma_0 \in \Lambda. \tag{5.11}$$

Finally, for any bounded and measurable $f : \mathbb{R}^d \to \mathbb{R}$ *a weak law of large numbers holds, i.e.*

$$\frac{\sum_{n=1}^N f(X_n)}{N} \to \pi(f) \qquad \textit{in probability.} \tag{5.12}$$

Remark 5.15. The time discretisation step Δt can be chosen freely and is not subject to constraints. Moreover, we remark that under condition (a) on $\mathcal{M}$ the adaptive algorithm is SGE, i.e. satisfies Assumption 5.6, while under condition (b) it satisfies the first set of conditions in Theorem 5.11. Thus if one is interested in learning only the diagonal elements of the covariance, then it is possible to take $B = \mathbb{R}^d$ and allow adaptations independently of the state of the process.

Remark 5.16. It was shown in [24] that the ZZS is geometrically ergodic under Assumption 5.13 also in the case $\gamma = 0$, whereas in Theorem 5.14 we require $\gamma(x,\theta) \geq \gamma_{\min} > 0$. This extra assumption is convenient when proving a simultaneous small set condition (see Lemma 5.20). Based on similar arguments as in [24], we expect the statement of Theorem 5.14 to remain valid even in the case $\gamma_{\min} = 0$. In practice, one is free to choose $\gamma_{\min}$ very small and thus this assumption does not represent a severe limitation.

Below we introduce a set of assumptions that is used to show ergodicity of the adaptive BPS. Here we limit our attention to the case of $\nu = \mathcal{N}(0, \mathbb{1}_d)$.

Assumption 5.17 (Assumptions A1, A2, and A7 in [64])**.** *Let* $\psi : \mathbb{R}^d \to [0,\infty)$ *satisfy*

(a) $\psi \in \mathcal{C}^2(\mathbb{R}^d)$*, and* $x \to \|\nabla\psi(x)\|$ *is integrable w.r.t.* π*;*

(b) $\int_{\mathbb{R}^d} e^{-\psi(x)/2}\mathrm{d}x < +\infty$ *and* $\lim_{\|x\|\to\infty} \psi(x) = +\infty$*;*

(c) There exists $\zeta \in (0,1)$ *such that*

$$\liminf_{\|x\|\to\infty} \frac{\|\nabla\psi(x)\|}{\psi^{1-\zeta}(x)} > 0, \qquad \limsup_{\|x\|\to\infty} \frac{\|\nabla\psi(x)\|}{\psi^{1-\zeta}(x)} < \infty,$$

and

$$\limsup_{\|x\|\to\infty} \frac{\|\nabla^2\psi(x)\|}{\psi^{1-2\zeta}(x)} < \infty.$$

Theorem 5.18. *Let $\mathcal{M}$ be a compact set of positive definite matrices and $\Lambda_r = [\lambda_{\min}, \lambda_{\max}]$ for $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$. Let $\mathcal{P} = \{P_{M,\lambda_r} : M \in \mathcal{M}, \lambda_r \in \Lambda_r\}$ be a family of preconditioned BPS's as defined in Equation (5.3), where λ_r is the refreshment rate. Suppose Assumption 5.17 holds and let Δt be the discretisation step. Assume that $\nu = \mathcal{N}(0, \mathbb{1}_d)$. If adaptations are allowed only inside of a compact set as explained in Equation (5.10), then the containment condition holds. Furthermore, for the strategy discussed in Section 5.2.4 the diminishing adaptation holds as long as $p_n \to 0$ as $n \to \infty$. Thus the ABPS is ergodic in the sense of Equation (5.11) and satisfies a WLLN of the form (5.12) for any bounded and measurable $f : \mathbb{R}^d \to \mathbb{R}$.*

Remark 5.19. Proving ergodicity of the adaptive BPS with refreshments from $\mathbb{S}^{d-1}$, i.e. the unit sphere centred at the origin, is more challenging due to the more involved drift condition proved in [51]. In particular, it is not straightforward to convert it into a simultaneous drift condition as required in assumption 2(b) of Theorem 5.11.

5.4 Proofs of the main theorems

5.4.1 Proof of Theorem 5.14

In order to prove the theorem we show that, for suitable families of preconditioners, either Assumption 5.6 holds for the family of discretised ZZ processes, or condition (1) in Theorem 5.11 holds for the family of continuous time ZZ processes. In the next two sections we state auxiliary results, while in Section 5.4.1.3 we assemble them to show the theorem. Proofs of the auxiliary results can be found in Appendix 5.B.4.

5.4.1.1 Minorisation condition for the ZZS

The following lemma shows that a simultaneous small set condition holds for the family of ZZ processes. The strategy of the proof is to reduce the d-dimensional minorisation condition to 1-dimensional conditions for every component of the process. Then we can take advantage of Lemma 5.32, which establishes that a simultaneous minorisation condition holds for a 1-dimensional ZZ process as long as lower and upper bounds for the switching rates are available.

Lemma 5.20. *Let $\psi \in \mathcal{C}^1$. Consider the family of d-dimensional Zig-Zag processes with generators $\{\mathcal{L}_{M,\gamma} : M \in \mathcal{M}, \gamma \in \Lambda\}$, in which $\mathcal{M}$ is a compact set of positive definite matrices and Λ is a set of switching rates $\gamma : E \to \mathbb{R}^d_+$. Assume that there are $\gamma_{\min}, \gamma_{\max}$ such that for all $\gamma \in \Lambda$*

$$0 < \gamma_{\min} \leq \gamma(x, \theta) \leq \gamma_{\max} < \infty \qquad \text{for all } (x, \theta) \in E.$$

Then for any set of the form $C = D \times V$, where $D \subset \mathbb{R}^d$ is a compact set and $V \subseteq \Theta$, there exists $t_0 > 0$ such that for any $t \geq t_0$ there are $\delta > 0$, and probability measures $\{\nu_M\}_{M \in \mathcal{M}}$ on E such that

$$P^t_{M,\gamma}((x, \theta), \cdot) \geq \delta \nu_M(\cdot) \qquad \text{for all } (x, \theta) \in C.$$

In particular, t_0 and δ do not depend neither on M nor on γ.

Proof. The proof can be found in Appendix 5.B.4.1. □

5.4.1.2 Drift conditions for the Zig-Zag process

If we restrict our attention to the class of diagonal matrices with positive, bounded entries, then the Lyapunov function in [24, Lemma 11] satisfies also a simultaneous drift condition. This is shown in the following lemma.

Lemma 5.21. *Let Assumption 5.13 hold. Consider the family of linearly transformed Zig-Zag processes with generators $\{\mathcal{L}_{M,\gamma} : M \in \mathcal{M}, \gamma \in \Lambda\}$, where*

$$\mathcal{M} = \{M \in \mathbb{R}^{d\times d} : M_{ii} \in [V^i_{\min}, V^i_{\max}],\ M_{jk} = 0 \text{ for all } j \neq k\}, \tag{5.13}$$

with $V_{\max} \geq V^i_{\max} \geq V^i_{\min} \geq V_{\min} > 0$ for each $i = 1, \dots, d$, and where Λ is a set of excess switching rates $\gamma : E \to \mathbb{R}^d_+$ such that for all $\gamma \in \Lambda$ it holds that

$$\gamma_i(x, \theta) \leq \gamma_{\max} \qquad \text{for all } (x, \theta) \in E,\ i = 1, \dots, d. \tag{5.14}$$

Let $\delta > 0$ and $\alpha > 0$ be such that $0 < (\delta\gamma_{\max})/V_{\min} < \alpha < 1$ and define $\phi(s) = \frac{1}{2}\text{sign}(s)\ln(1 + \delta|s|)$. Then the function

$$V(x, \theta) = \exp\left(\alpha\psi(x) + \sum_{i=1}^{d} \phi(\theta_i \partial_i \psi(x))\right) \tag{5.15}$$

is a simultaneous Lyapunov function for the family of ZZ processes, that is there exist $A_1 > 0$, $A_2 > 0$, a compact set $C \subset E$ such that

$$\mathcal{L}_{M,\gamma} V(x, \theta) \leq -A_1 V(x, \theta) + A_2 \mathbb{1}_C(x, \theta) \qquad \text{for all } (x, \theta) \in E,\ M \in \mathcal{M},\ \gamma \in \Lambda,$$

where A_1, A_2, C do not depend neither on M nor on γ (but depend on $\mathcal{M}$ and Λ).

Proof. The proof can be found in Appendix 5.B.4.2. □

If we wish to consider a more general class of positive-definite matrices, then we have to settle for the following, weaker result.

Lemma 5.22. *Let Assumption 5.13 hold. Consider a family of linearly transformed Zig-Zag processes with generators $\{\mathcal{L}_{M,\gamma} : M \in \mathcal{M}, \gamma \in \Lambda\}$, where $\mathcal{M} \subset \mathbb{R}^{d\times d}$ is a compact space of positive definite matrices and Λ is a space of excess switching rates $\gamma : E \to \mathbb{R}_+$ such that (5.14) is satisfied for some $\gamma_{\max}$. Let $\delta > 0$ and $\alpha > 0$ be such that $0 < \delta\gamma_{\max} < \alpha < 1$. Define for each $M \in \mathcal{M}$ the function*

$$V_M(x, \theta) = \exp\left(\alpha\psi(x) + \sum_{i=1}^{d} \phi(\theta_i \langle M_i, \nabla\psi(x)\rangle)\right), \tag{5.16}$$

where M_i denotes the i-th column of M and $\phi : \mathbb{R} \to \mathbb{R}$ was defined in Lemma 5.21. Then there are $A_1 > 0$, $A_2 > 0$, and a compact set $C \subset E$ such that for all $M \in \mathcal{M}$ the following simultaneous drift condition holds:

$$\mathcal{L}_{M,\gamma} V_M(x,\theta) \leq -A_1 V_M(x,\theta) + A_2 \mathbb{1}_C(x,\theta) \qquad \text{for all } (x,\theta) \in E.$$

In particular A_1, A_2, C do not depend neither on M nor on γ (but depend on $\mathcal{M}$ and Λ).

Proof. The proof can be found in Appendix 5.B.4.2. □

5.4.1.3 Finalising the proof of Theorem 5.14

Let us first consider the case of diagonal preconditioners. Let $\Delta t > 0$ be the discretisation step. Then by Lemma 5.20 for any set of the form $C = D \times V$, with D compact and $V \subseteq \Theta$, there exist $\delta > 0$, $n_0 := \inf\{n \in \mathbb{N} : n \geq \frac{t_0}{\Delta t}\}$, $\nu_M(\cdot)$ such that C is a uniform (ν_M, n_0, δ)-small set for the family of discretised processes. Observe that no conditions on Δt are required. Moreover, a simultaneous drift condition holds by Lemma 5.21 combined with Lemma 5.33 for any Δt. The condition $\mu(V) < \infty$ is satisfied by definition of the Lyapunov function V, and in fact it also holds that $\sup_{(x,\theta)\in C} V(x,\theta) < \infty$ by continuity of V in x. Therefore all the conditions of Assumption 5.6 are verified, which means the family is simultaneously geometrically ergodic and by Theorem 3 in [8] the containment condition is satisfied.

In the case of a non-diagonal transformation matrix, parts (a)-(c) of condition (1) in Theorem 5.11 are verified by Lemmas 5.20 and 5.22. It also holds that

$$\sup_{\{(x,\theta)\in C, M\in\mathcal{M}\}} V_M(x,\theta) < \infty$$

for (small) sets $C = D \times V$, with D compact and $V \subseteq \Theta$, because of continuity of each V_M in x and M, together with the fact that D and $\mathcal{M}$ are compact spaces (see Lemma 5.22 for the definition of $\{V_M\}_{M\in\mathcal{M}}$). Moreover $\mu(V_M) = \tilde{\mu}_M(\tilde{V}_M) < \infty$, where $\tilde{\mu}_M = \tilde{\pi}_M \otimes \mathrm{Unif}(\Theta)$ and $\tilde{V}_M$ is a Lyapunov function for a standard ZZ process with invariant measure $\tilde{\mu}_M$. Theorem 5.11 ensures that the containment condition holds true with no restriction on Δt.

Proposition 5.25 implies that, under the assumption that $p_n \to 0$ as $n \to \infty$, the adaptive strategy satisfies diminishing adaptation. Therefore ergodicity follows from diminishing adaptation and containment.

5.4.2 Proof of Theorem 5.18

In the next two sections we show respectively that condition (2) in Theorem 5.11 is verified for the family of preconditioned BPS and/or of BPS with refreshment rates in a compact set. In Section 5.4.2.3 we use these auxiliary results to show the theorem.

5.4.2.1 Simultaneous coupling inequality

The next lemma states that a simultaneous coupling inequality is satisfied for the BPS with adaptive preconditioner and/or adaptive refreshment rate. The proof is based on the proof of Lemma 12 in [64], which shows a coupling inequality result for the standard BPS.

Lemma 5.23. *Let condition (a) in Assumption 5.17 hold for the energy function ψ. Consider the family of BP processes $\{P^t_{M,\lambda_r} : M \in \mathcal{M}, \lambda_r \in \Lambda_r\}$, where $\mathcal{M}$ is a compact space of non-singular preconditioning matrices and $\Lambda_r = [\lambda_r^{\min}, \lambda_r^{\max}]$ is the set of refreshment rates, for some $0 < \lambda_r^{\min} \leq \lambda_r^{\max} < \infty$. Then for any compact set $K \subset \{(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^d : \|x\| + \|\theta\| \leq R\}$, with $R \geq 0$, there exists $\alpha > 0$ such that for all (x,θ), $(\tilde{x},\tilde{\theta}) \in K$, for all $t > 0$, and for all $M \in \mathcal{M}$ and $\lambda_r \in \Lambda_r$*

$$\|P^t_{M,\lambda_r}((x,\theta),\cdot) - P^t_{M,\lambda_r}((\tilde{x},\tilde{\theta}),\cdot)\|_{\mathrm{TV}} \leq 2\,(1-\alpha)\,.$$

In particular α is independent of M and λ_r.

Proof. The proof can be found in Appendix 5.B.5.1. □

5.4.2.2 Drift condition for the BPS

The second condition we need is uniformity of the constants in the drift condition for the family of preconditioned BPS and/or for the family of BPS with different refreshment rate. To this end, we go through the proof of Lemma 7 from [64] to show that this is indeed the case.

Lemma 5.24. *Consider a family of BP processes with generators $\{\mathcal{L}_{M,\lambda_r} : M \in \mathcal{M}, \lambda_r \in \Lambda_r\}$, where $\mathcal{M}$ is a compact space of non-singular matrices that act as preconditioners, λ_r is the refreshment rate and $\Lambda_r = [\lambda_{min}, \lambda_{max}]$ for some $0 < \lambda_{min} \leq \lambda_{max} < \infty$. Let Assumption 5.17 hold and let $\nu = \mathcal{N}(0, \mathbb{1}_d)$. Then there are $A_1, A_2 > 0$ and a class of functions $\{V_{M,\lambda_r} : M \in \mathcal{M}, \lambda_r \in \Lambda_r\}$ such that for each $M \in \mathcal{M}$ and $\lambda_r \in \Lambda_r$ it holds that*

$$\mathcal{L}_{M,\lambda_r} V_{M,\lambda_r}(x,\theta) \leq -A_1 V_{M,\lambda_r}(x,\theta) + A_2 \qquad \text{for all } (x,\theta) \in \mathbb{R}^d \times \mathbb{R}^d,$$

where in particular A_1, A_2 do not depend on M.

Proof. The proof can be found in Appendix 5.B.5.2. □

5.4.2.3 Finalising the proof of Theorem 5.18

Let $\Delta t > 0$ be a discretisation step. Lemma 5.24 gives the drift condition, that is condition (b) in Theorem 5.11. Then Lemma 5.23 implies that a coupling inequality holds for any compact set. Sets of the form $V_{M,\lambda_r}(x,\theta) + V_{M,\lambda_r}(x,\theta) \leq C_2$ are compact by definition of the class of Lyapunov functions $\{V_{M,\lambda_r} : M \in \mathcal{M}, \lambda_r \in \Lambda_r\}$. We are in particular free to choose the constant C_2 as large as we wish. Note also that the coupling inequality holds for all $t > 0$, hence there are no constraints on the choice

of Δt. Moreover $\mu(V_{M,\lambda_r}) = \tilde{\mu}_M(\tilde{V}_{M,\lambda_R}) < \infty$ for all $M \in \mathcal{M}$ and $\lambda_r \in \Lambda_r$. Here $\tilde{V}_{M,\lambda_R}$ is the Lyapunov function of a standard BPS with refreshment rate λ_r and target $\tilde{\mu}_M = \tilde{\pi}_M \times \nu$. The containment condition is thus verified as all conditions in part (2) of Theorem 5.11 hold. Proposition 5.25 implies the diminishing adaptation condition, and thus ergodicity follows.

5.4.3 Proving the diminishing adaptation condition

A key part of Theorem 5.5 is condition (b), i.e. the diminishing adaptation condition. For the adaptive scheme described in Section 5.2.4 the condition can be easily shown to be true as the adaptation happens with diminishing probability.

Proposition 5.25. *Consider the adaptive schemes in Section 5.2. In particular, assume that $\{p_n\}_{n\geq 0}$, i.e. the sequence of probabilities of updating the adaptation parameters, is such that $p_n \to 0$ as $n \to \infty$. Then the diminishing adaptation holds for any $t \geq 0$.*

Proof. Consider for example the adaptive BPS. Observe that $M_{n+1} = M_n$ and $\lambda_r^{n+1} = \lambda_r^n$ with probability $1 - p_{n+1}$ and thus

$$\|P^{\Delta t}_{M_{n+1},\,\lambda_r^{n+1}}((x,\theta),\cdot) - P^{\Delta t}_{M_n,\,\lambda_r^n}((x,\theta),\cdot)\|_{\mathrm{TV}} \leq 2p_{n+1} \to 0 \qquad \text{as } n \to \infty.$$

Thus the diminishing adaption holds. The same reasoning works for the adaptive ZZS. □

5.5 Numerical experiments

In this section we test the empirical performance of the adaptive schemes we defined in Section 5.2.4. All experiments are implemented in Julia and the corresponding codes can be found at `https://github.com/andreabertazzi/Adaptive_PDMC_samplers`. Let us state some settings that hold for all experiments below. The time horizon is set to $T = 10^5$ for all processes. This is large enough for the adaptive PDMC samplers to learn and take advantage of the covariance structure. When considered fixed, the refreshment rate of the BPS is taken to be $\lambda_r = 1$. The excess switching rate for ZZS is set to 0 in all experiments. The discretisation step is chosen to be $\Delta t = 0.5$, which in our experiments turns out to be a good choice for a wide range of targets. Moreover, we set adaptation times to be every $t_{\mathrm{adap}} = 2000$ continuous time units. The probability of adapting decays as $\mathcal{O}(\log\log n)$. Finally, no normalisation in the sense discussed at the end of Section 5.2.3 is employed. The performance measures we consider are the *effective sample size per second* (ESS/sec) for the mean and for the radius statistic $t(x) = \sum_{i=1}^d x_i^2$. In Sections 5.5.1 and 5.5.2 these are computed in continuous time as discussed in [23] by estimating the asymptotic variance with the batch means method, and the variance of the Monte Carlo estimate on the continuous time trajectories of the processes. On the other hand, in Section 5.5.3 we compute the

mean squared error (MSE) in discrete time for the sample mean and radius statistic and take advantage of the fact that in the large time horizon regime the MSE is approximately given by the asymptotic variance of the observable divided by the number of generated samples. This alternative way to compute the ESS avoids poor convergence of the batch means method in the multimodal case. Finally, in all settings we repeat the same task 20 times and report all the results in box-plots.

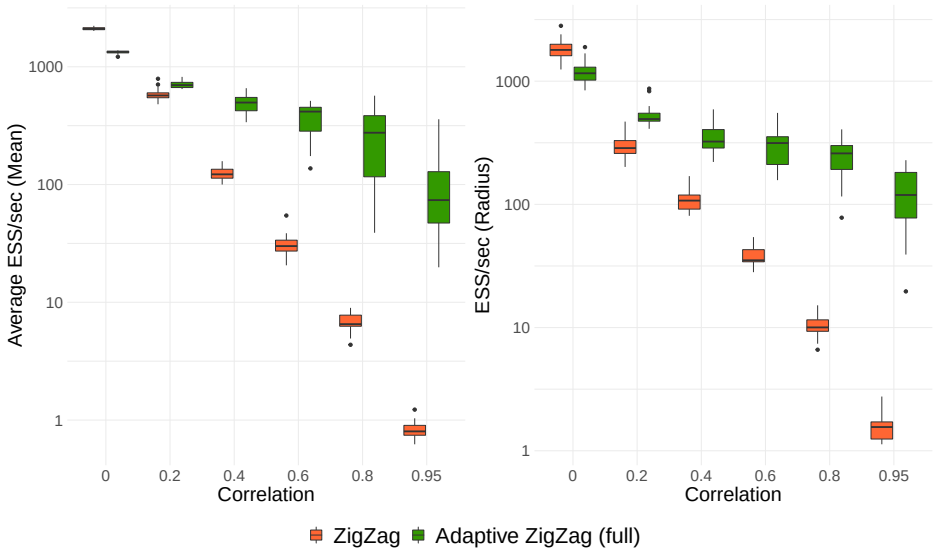
5.5.1 Multidimensional Gaussian target

In this section we focus on two different kinds of multivariate Gaussian target distributions. The first one, denoted by **MG1**, has unitary variances and correlation ρ between all components. Denoting the covariance matrix by Σ, this means that $\Sigma_{ii} = 1$ for each $i = 1, \dots, d$ and $\Sigma_{ij} = \rho$ for all $i \neq j$. We study how the adaptive PDMC algorithms compare to their non-adaptive counterparts for different values of ρ and different dimensionalities. In this setting we focus on adaptive algorithms that estimate the full covariance matrix. The second Gaussian target we consider has variances $0.5, 1, 5, 10, 15$ repeated depending on the dimension, together with a milder correlation between components. This setting is denoted as **MG2** and is useful to compare all kinds of adaptive algorithms we introduced.

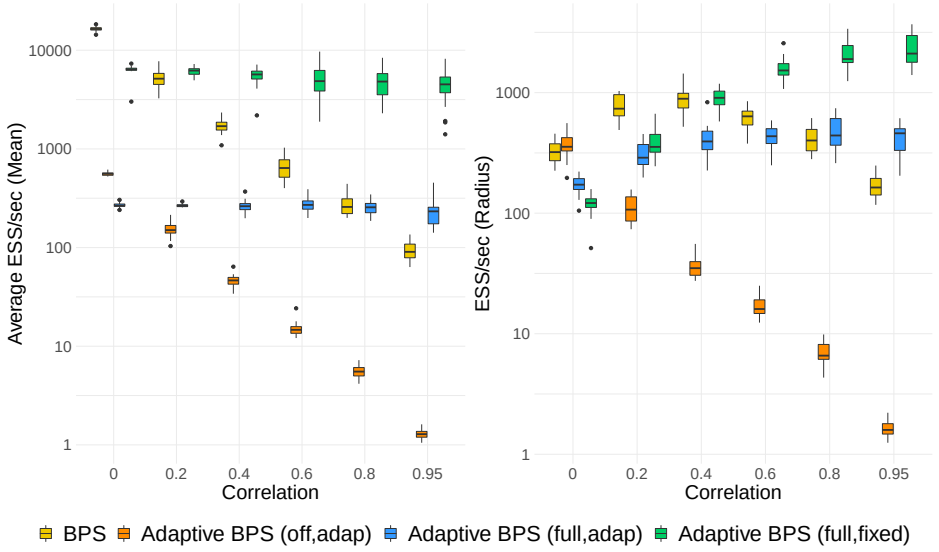
5.5.1.1 MG1 target

In the first experiment we consider a 50-dimensional **MG1** target for different values of ρ. In Figure 5.3 the average ESS/sec and the ESS/sec for the radius statistic are shown. As expected, the performance of the ZZS is degrading as the correlation increases. This behaviour is for the most part caused by the very large number of events that have to be simulated for very narrow targets. The adaptive ZZS successfully improves over this inconvenience and is stable with respect to the increasing correlation. The standard BPS with fixed refreshment rate shows a decaying average ESS/sec, while the ESS/sec for the radius statistic appears to increase as ρ grows up until $\rho = 0.4$ and then becomes smaller. This behaviour is likely due to the fact that the choice $\lambda_r = 1$ is more suited for the estimation of the radius in case of a more concentrated target rather than for a standard Gaussian. A similar behaviour is shown by the adaptive BPS's. Overall we notice a marked improvement for the BPS with adaptive preconditioner and fixed λ_r. Choosing to adapt only the refreshment turns out to be a detrimental decision when the target is correlated. Indeed the optimality criterion derived [27] assumes a standard Gaussian target.

In Figure 5.4 we study how the adaptive schemes compare to the standard ones for an **MG1** target with correlation $\rho = 0.8$ and increasing dimensionalities of target. The plots show that when the target is strongly correlated the effect of the adaptation shows no sign of diminishing. It also seems clear that the sampler of choice in this case should be the adaptive BPS with fixed, rather than adaptive, refreshment rate. This could be due to the fact that, when the refreshment rate is updated adaptively and the target is anisotropic, a too large λ_r is chosen at first due to the high number of reflections, thus slowing down the estimation of the covariance matrix. As suggested



(a) Comparison between the ZZS and the ZZS with adaptive preconditioner learning the **full** covariance matrix.



(b) Comparison between the BPS and various alternative adaptive BPS's. **Adaptive BPS (full,adap)** denotes the BPS that learns the entire covariance matrix (**full**) and with adaptive refreshment rate (**adap**). Adaptation of the preconditioner can be turned **off**, and similarly the refreshment rate can be **fixed**.

Figure 5.3: Results as a function of the correlation for **MG1** targets.

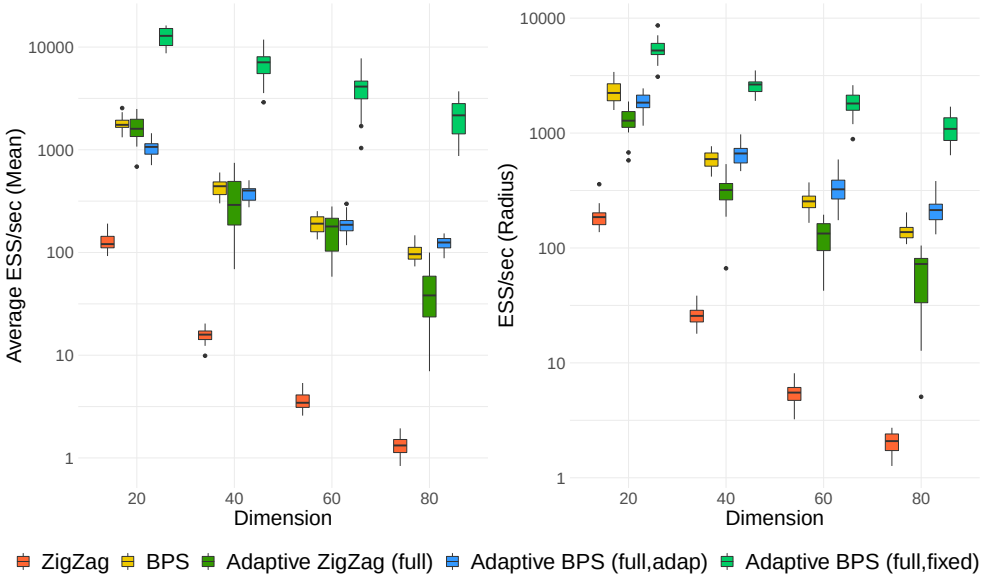
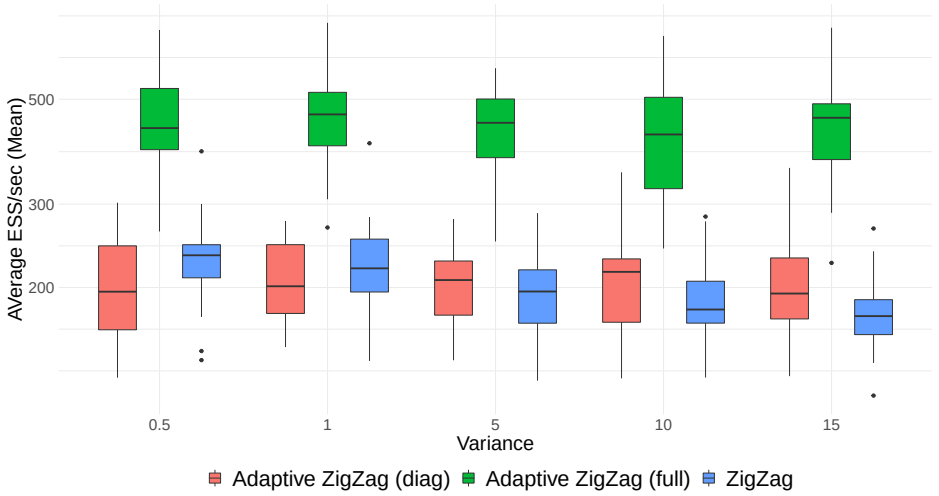


Figure 5.4: **MG1** target with $\rho = 0.8$ and different dimensionalities.

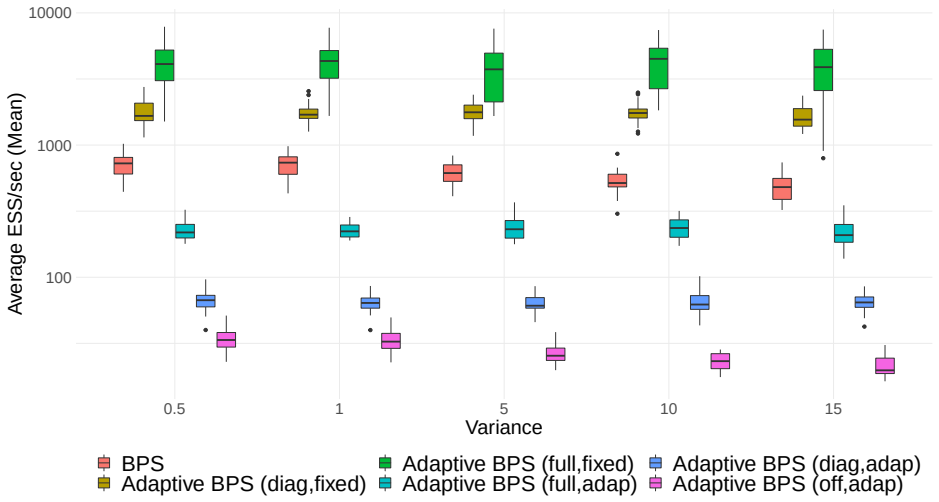
by a reviewer, one could avoid this issue by keeping the refreshment rate fixed until the estimate of the covariance stabilises, and only then starting to learn the optimal λ_r. It is worth pointing out that the performance of the BPS with adaptive preconditioner and refreshment improves compared to the BPS as the dimension increases. This is according to the theoretical results in [27], which are indeed obtained in the high dimensional limit. Therefore we expect that for a large d it is reasonable to apply both the transformation scheme and the tuning of λ_r.

5.5.1.2 MG2 target

Let us now consider a 50-dimensional **MG2** target with a mild correlation set to $\rho = 0.3$. Figure 5.5 shows the results for several adaptive PDMC samplers. The adaptive algorithms that learn the entire covariance matrix show the largest gain in terms of ESS/sec. For the BPS the choice of learning only the variance of each component of the target seems interesting, also in view of larger dimensions. As in the previous section, we observe that the adaptation of the refreshment rate seems to have a bad effect for anisotropic targets.



(a) Results of the **MG2** experiment for the ZZS. The option **diag** refers to the adaptive algorithm that learns only the diagonal of the covariance matrix.



(b) Results of the **MG2** experiment for the BPS.

Figure 5.5: Comparison of several samplers in the context of Section 5.5.1.2.

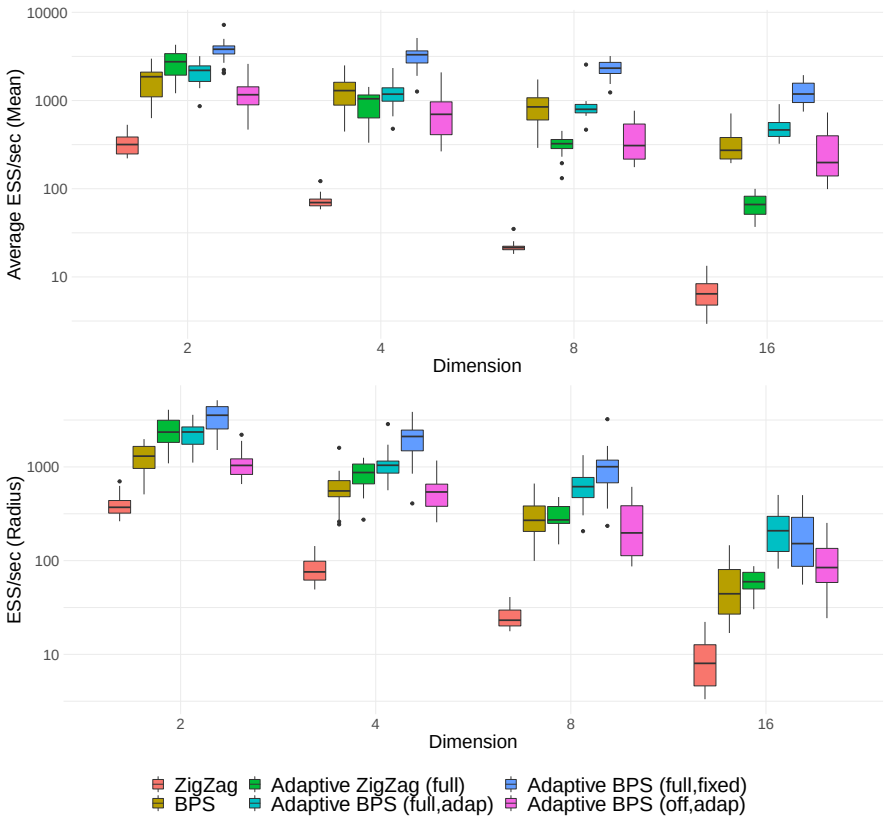


Figure 5.6: Logistic regression task of Section 5.5.2.

5.5.2 Logistic regression with correlated data

The next numerical experiment we consider is a Bayesian logistic regression task. In this setting for $j = 1, \dots, n_{\text{obs}}$ a binary output value $y_j \in \{0, 1\}$ has distribution

$$\mathbb{P}(Y_j = 1|\beta) = \frac{1}{1 + \exp\left(-\beta^T x_j\right)},$$

where $\{x_j\}_{j=1}^{n_{\text{obs}}}$ are known covariates, and $\beta \in \mathbb{R}^d$ is an unknown parameter. We take a flat prior and thus obtain the posterior

$$\pi(\beta|\{y_j\}_{j=1}^{n_{\text{obs}}}) \propto \prod_{j=1}^{n_{\text{obs}}} \frac{\exp\left(-y_j \beta^T x_j\right)}{1 + \exp\left(-\beta^T x_j\right)}.$$

We force correlation between some components of the parameter by taking, for $j = 1, \dots, n_{\text{obs}}$ and $i = 1, \dots, d$, $(x_j)_i = 1 + \varepsilon N_{ji}$, where $N_{ji} \sim \mathcal{N}(0, 1)$ and $\varepsilon = 0.1$. The

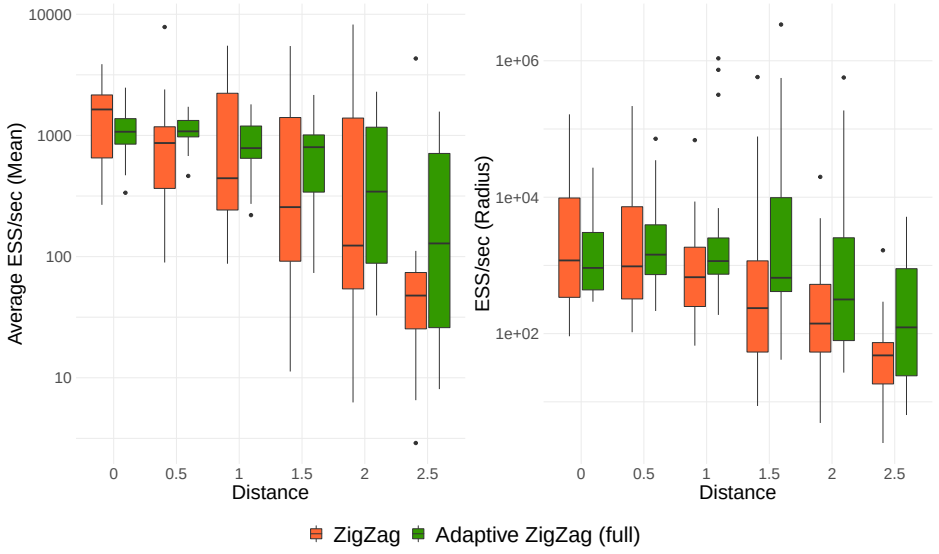
results of the experiment are reported in Figure 5.6, in which the samplers are tested with targets as above with $d = 2, 4, 8, 16$ and $n_{\text{obs}} = 1000$. The adaptation of the refreshment rate follows the alternative scheme discussed in Appendix 5.A.3. This scheme seems more stable as the refreshment rate is update gradually and cannot jump immediately to very large or small values. Although the dimensionality is small and the correlation is limited to a subset of the coordinates, we observe that the adaptive schemes outperform their standard counterparts.

5.5.3 Mixture of Gaussian distributions

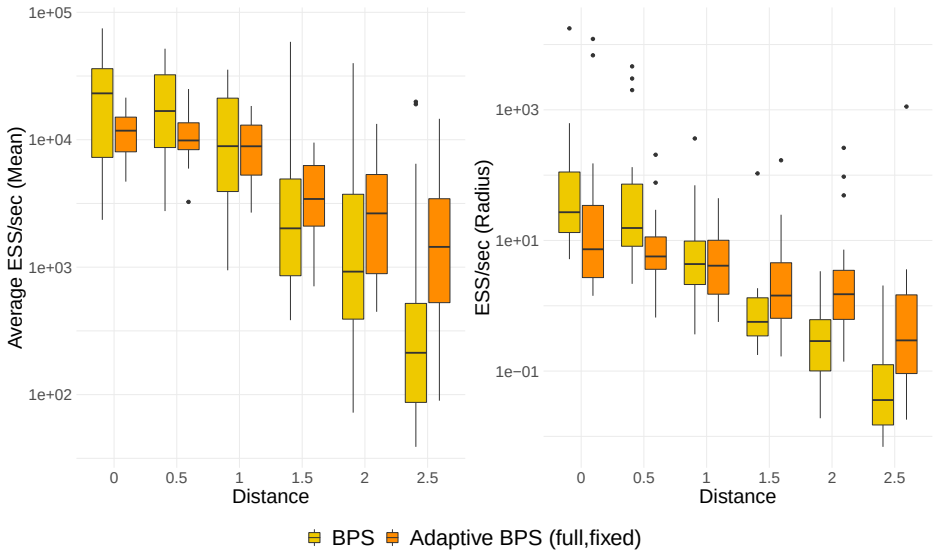
Consider a mixture of two 30-dimensional Gaussian distributions $\mathcal{N}(0_d, \Sigma)$ and $\mathcal{N}(\mu, \Sigma)$, both with weight $\frac{1}{2}$. Here we take Σ with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.25$ for $j \neq i$. Moreover we take $\mu = a \times (1, \dots, 1)$, where a is a parameter that determines the distance between the two means. We investigate the performance of the adaptive schemes compared to standard ones as a function of the parameter a. The details on the implementation of this experiment can be found in Appendix 5.A.4. In this experiment we test the robustness of the algorithm in a case in which the target distribution is cigar shaped, but multimodal. The time horizon for ZZS is 10^5, while for BPS it is 4×10^5. Figure 5.7 shows the results for the adaptive samplers in this setting. We observe that as the distance between the means increases the performance of the adaptive samplers improves over the standard ones. We remark that for small values of the distance a one cannot expect improvements of the adaptive algorithms as the correlation is small and thus the target is not very anisotropic.

5.6 Discussion

In this paper we proposed adaptive schemes to overcome two of the current issues with the BPS and ZZS. We have shown that the refreshment rate and the excess switching rate can be tuned on the fly as long as the updates or the probabilities of updating decrease to 0. With this approach the user does not have to worry about tuning the refreshment rate. A current limitation is that more theory or experiments are needed to determine a criterion that works well with anisotropic targets. In addition to this, we have proposed a way to make the PDMC samplers learn and take advantage of the covariance structure of the target. The theoretical results stated in Theorems 5.14 and 5.18 ensure that the adaptive samplers are ergodic and thus converge to the correct measure π. It is challenging to prove theoretical statements regarding performance improvements of the adaptive schemes over the standard PDMC samplers. However, the numerical experiments we conducted suggest that our adaptive algorithms can lead to a significant performance improvement when there are strong anisotropies. An alternative approach could be to use an optimisation algorithm to obtain an estimate of the covariance matrix by computing the Hessian of the target at its point of maximum. However, this optimisation step can be expensive and moreover for realistic problems it is not a given that this estimator is a good approx-



(a) Results for the ZZS with time horizon $T = 10^5$.



(b) Results for the BPS with time horizon $T = 4 \times 10^5$.

Figure 5.7: Numerical results for a mixture of two Gaussian distributions as described in Section 5.5.3.

imation of the posterior covariance. In particular in our theory we do not assume log convexity of the target, and also we do not assume that we are in the large sample regime where a Bernstein-von Mises theorem holds. In addition, the adaptive schemes discussed in this paper can be applied to the Boomerang sampler [25]. This would result in elliptical dynamics that are adapted to resemble the (unimodal) target at hand. Naturally, the adaptive algorithms should be run with a time horizon that is large enough to benefit from the adaptation. Two other important settings are the discretisation step and the time between two adaptations. Based on our experience with the experiments, we suggest $\Delta t = \mathcal{O}(10^{-1})$ and $t_{\text{adaps}} = \mathcal{O}(10^3)$. For concentrated targets, as for instance posteriors when there is a very large number of data points, it is suggested to choose both values small. We remark that in very high dimensional settings it may be unfeasible to let the samplers learn the full covariance matrix, as the computation of M entails calculating the square root of the empirical covariance matrix. In such cases we suggest either learning only the diagonal elements or blocks of the covariance. We remark that the adaptive PDMC algorithms with subsampling are applicable in the setting of tall data, that is when data-set is made of a large number of observations, but with a moderate dimensionality. In such settings subsampling can be shown empirically to result in an improved efficiency (in terms of ESS per second); a result which is backed by a heuristic argument, based on posterior contraction, i.e., the Bernstein-Von Mises theorem; see [23] for details. More research is necessary to understand in which situations subsampling can lead to improved efficiency, and in particular if improved efficiency is possible in cases for which the Bernstein-von Mises theorem does not apply; see also [9, 86] for a critical discussion of subsampling methods.

A question that one could naturally ask is how applicable this transformation scheme is in case of a multimodal target. The answer depends on the specific target at hand, but one can design a target as a mixture of Gaussian distributions for which applying the transformation scheme would not speed up the convergence of the sampler. However, when the target is multimodal it is possible for instance to use the adaptive PDMC samplers together with the framework proposed in [126]. In this framework, the adaptive PDMC samplers would be beneficial since the regions around each mode would be explored more efficiently by taking advantage of the covariance structure of the specific mode.

Finally, we remark that the idea of learning the covariance structure of the target on the fly could be applied to obtain adaptive versions of the Hamiltonian Monte Carlo (HMC) algorithm [118] and of the Metropolis Adjusted Langevin Algorithm (MALA) [139]. In particular both the HMC algorithm and the MALA are sensitive to correlation in the target and can thus benefit from a suitable preconditioner, as argued respectively in Section 4.1 of [118] and in [138]. Moreover, the preconditioner could be chosen to take advantage of the geometry of the target, as proposed in [75] for HMC and MALA. The preconditioner could be estimated adaptively with an appropriate adaptation strategy, together with similar ideas presented in this manuscript.

5.A Implementation of adaptive PDMC algorithms

The main issue in the exact simulation of PDMPs lies in the simulation of the switching times. It is generally proposed to use Poisson thinning to overcome this difficulty. The idea is to find upper bounds for the switching rates that are more tractable, then simulate a Poisson process with said rate and finally adjust with an acceptance-rejection step. Applying a preconditioning matrix to PDMPs results in a modification of the switching rates and thus the bound proposed in [23, 32] do not directly apply to our proposed algorithms. In the following two sections we give two important examples to motivate that it is possible to find bounds with similar ideas to those used for standard PDMC algorithms. In Appendix 5.A.3 we discuss a different adaptation strategy for the refreshment rate.

5.A.1 Dominated Hessian of the negative log-likelihood

Let us consider the case in which the Hessian of the negative log-likelihood is dominated by a positive definite matrix. Denoting the Hessian by $H_\psi(x) = (\partial_i \partial_j \psi(x))_{i,j=1}^d$, this means that for any $x \in \mathbb{R}^d$ it holds that $\langle H_\psi(x)u, u\rangle \leq \langle Qu, u\rangle$ for all $u \in \mathbb{R}^d$. Then we say that $-Q \preceq H_\psi(x) \preceq Q$. Assuming Q is symmetric, we have for all $u, v \in \mathbb{R}^d$ that $\langle u, H_\psi(x)v\rangle \leq \|u\|_2 \|Qv\|_2$.

Let us consider the switching rates of a ZZ process with preconditioner M, i.e. $\lambda_i^M(x, \theta) = (\theta_i \langle M_i, \nabla\psi(x)\rangle)_+$, where M_i is the i-th column of M. Remember that the deterministic trajectory of this process is $x(t) = x + M\theta t$, where (x, θ) is the initial condition. We have

$$\begin{aligned}\partial_i \psi(x(t)) &= \partial_i \psi(x) + \int_0^t \sum_{j=1}^d \partial_j \partial_i \psi(x(s))(M\theta)_j ds \\ &= \partial_i \psi(x) + \int_0^t \langle H_\psi(x(s))e_i, M\theta\rangle ds,\end{aligned}$$

where e_i, is the i-th vector of the canonical basis, i.e. with zeros in all components except for a 1 in the i-th component. Taking $a_i = \theta_i \langle M_i, \nabla\psi(x)\rangle$ and $b_i = \sqrt{d}\|M\|_2 \|QM_i\|_2$ it follows that

$$\begin{aligned}\lambda_i^M(x(t), \theta) &= \left(\theta_i \langle M_i, \nabla\psi(x)\rangle + \theta_i \sum_{j=1}^d M_{ji} \int_0^t \langle H_\psi(x(s))e_j, M\theta\rangle ds\right)_+ \\ &\leq \left(\theta_i \langle M_i, \nabla\psi(x)\rangle + \int_0^t \langle H_\psi(x(s))M_i, M\theta\rangle ds\right)_+ \\ &\leq (a_i + t\|M\theta\|_2 \|QM_i\|_2)_+ \leq (a_i + b_i t)_+.\end{aligned}$$

It is possible to have a tighter bound taking $b_i = \|QM_i\|_2 \|M\theta\|_2$, but this entails having to update b_i after every switching time, which would make the overall simulation more expensive. Note that in particular for a diagonal M we have that $\|M\theta\|_2 =$

$(\sum_{i=1}^d M_{ii}^2)^{1/2}$ and we can take $a_i = \theta_i M_{ii} \partial_i \psi(x)$ and $b_i = M_{ii}\|Q_i\|_2 (\sum_{i=1}^d M_{ii}^2)^{1/2}$. It is worth observing that $\|Q\|_2$ can be computed once before running the algorithm and then stored. Terms $\|M\|_2$ and $\|M_i\|_2$ can in principle be estimated by taking a maximum over the class of matrices $\mathcal{M}$. However, having loose computational bounds leads to a lower percentage of accepted proposed events and thus to an increased computational burden.

For BPS with preconditioner M we showed that $\lambda_M(x,\theta) = (\langle M\theta, \nabla\psi(x)\rangle)_+$. With computations similar to the ones above we obtain

$$\begin{aligned}\lambda_M(x(t),\theta) &= \left(\langle M\theta, \nabla\psi(x)\rangle + \int_0^t \langle M\theta, H_\psi(x(s))M\theta\rangle ds\right)_+ \\ &\leq \left(\langle M\theta, \nabla\psi(x)\rangle + t\langle M\theta, QM\theta\rangle\right)_+ \leq (a+bt)_+\end{aligned}$$

with $a = \langle M\theta, \nabla\psi(x)\rangle$ and $b = \langle M\theta, QM\theta\rangle$. Note that $M\theta$ is the velocity of the process and needs to be computed in any case to determine the trajectories.

5.A.2 Subsampling techniques

PDMC algorithms are particularly interesting because they allow for exact subsampling. This means that it is possible to modify PDMC samplers such that the correct invariant measure is maintained but without going through the entire data-set at each iteration. This was illustrated in [23, 32] for ZZS and BPS. Subsampling techniques are particularly helpful when the number of data points n is much larger than the dimensionality d of the posterior density function. This is an interesting scenario for the adaptive algorithms that are here introduced, as for a small d the additional computations necessary to learn and take the square root of (part of) the covariance matrix are not overpowering. In this section we explain how adaptive PDMC with subsampling can be implemented.

Suppose the partial derivatives of the posterior distribution can be written in the form

$$\partial_i \psi(x) = \frac{1}{n}\sum_{j=1}^n E_i^j(x), \tag{5.17}$$

where $E_i^j : \mathbb{R}^d \to \mathbb{R}^d$ are continuous mappings. This holds for example when the target satisfies

$$\psi(x) = \frac{1}{n}\sum_{j=1}^d \psi^j(x).$$

This is the case for instance when the target is a posterior density of a model with iid observations. In such cases one can choose $\psi^j(x) = -\log(\pi_0(x)) - n\log\left(l(y^j|x)\right)$, where π_0 is a prior of the parameters and l is the likelihood associated to observation y^j. Then the idea is to define a collection of switching rates along a trajectory.

$$m_i^j(t) = (\theta_i \langle M_i, E^j(x + M\theta t)\rangle)_+,$$

where $E^j(x) = (E_1^j(x), \dots, E_d^j(x))^T$. If we can find uniform bounds $H_i(t)$ such that $m_i^j(t) \leq H_i(t)$ for all j, then the subsampling procedure is successful. One can simulate IPPs with rates $H_i(t)$ and then accept the proposed switch of component i_0 with probability $m_{i_0}^J(t)/H_{i_0}(t)$ where $J \sim \text{Unif}\{1, \dots, n\}$. The next proposition shows that stationarity of π is preserved for the ZZS with subsampling also introducing a preconditioner.

Proposition 5.26. *The Zig-Zag sampler with subsampling and preconditioner M has invariant distribution $\mu = \pi \times Unif(\Theta)$. Moreover it coincides with a Zig-Zag process with switching rates*

$$\lambda_i^M(x, \theta) = \frac{1}{n} \sum_{j=1}^{n} (\theta_i \langle M_i, E^j(x) \rangle)_+ \quad \textit{for } (x, \theta) \in E,\ i = 1, \dots, d.$$

Proof. The statements can be derived following the proof of Theorem 4.1 in [23] and is thus omitted. □

An interesting situation is that of a negative log-likelihood with Lipschitz partial derivatives, i.e. $|\partial_i \psi(x) - \partial_i \psi(y)| \leq C_i \|x - y\|_p$. In this case we can take a reference point x^* and take

$$E^j(x) = \nabla \psi(x^*) + \nabla \psi^j(x) - \nabla \psi^j(x^*).$$

This allows us to obtain the following uniform bound on the switching rates

$$\begin{aligned} m_i^j(t) &= (\theta_i \langle M_i, \nabla \psi(x^*) + \nabla \psi^j(x + M\theta t) - \nabla \psi^j(x^*) \rangle)_+ \\ &\leq \left(\theta_i \langle M_i, \nabla \psi(x^*) \rangle + |\langle M_i, \nabla \psi^j(x + M\theta t) - \nabla \psi^j(x^*) \rangle|\right)_+ \\ &\leq \left(\theta_i \langle M_i, \nabla \psi(x^*) \rangle + \sum_{l=1}^{d} C_l |M_{li}| \|x + M\theta t - x^*\|_p\right)_+ \\ &\leq (\theta_i \langle M_i, \nabla \psi(x^*) \rangle + \langle |M_i|, C \rangle \|x - x^*\|_p + t \langle |M_i|, C \rangle \|M\theta\|_p)_+ . \end{aligned}$$

Therefore we can simulate d IPPs with rates $H_i(t) = (a_i + b_i t)_+$, where $a_i = \theta_i \langle M_i, \nabla \psi(x^*) \rangle + \langle |M_i|, C \rangle \|x - x^*\|_p$ and $b_i = d^{1/p} \langle |M_i|, C \rangle \|M\|_p$.

The same estimator for the gradient of the negative log-density can be used for a preconditioned BPS. With computations analogous to above we find

$$m_i^j(t) \leq (\langle M\theta, \nabla \psi(x^*) \rangle + \langle |M\theta|, C \rangle \|x - x^*\|_p + t \langle |M\theta|, C \rangle \|M\theta\|_p)_+ \leq H_i(t).$$

Although establishing ergodicity of these adaptive algorithms would be an interesting research question, we leave it for future research due to the lack of theoretical results for their standard counterparts.

5.A.3 A different adaptation strategy for the refreshment rate

The adaptive scheme for the refreshment rate described in Section 5.2.4 is a natural implementation of the results in [27]. However, it can be unstable when combined with an adaptive preconditioner. Here we define an alternative adaptive strategy and we prove that the diminishing adaptation condition from Theorem 5.5 is satisfied.

Assume that the refreshment rate of the BPS is constant in the position and velocity spaces. We update it iteratively as follows:

$$\lambda^r_{n+1} = \begin{cases} \lambda^r_n + q_{n+1} & \text{if } \frac{n_{\text{refresh}}(n+1)}{n_{\text{events}}(n+1)} < \lambda^*, \\ \lambda^r_n - q_{n+1} & \text{if } \frac{n_{\text{refresh}}(n+1)}{n_{\text{events}}(n+1)} > \lambda^*, \end{cases} \tag{5.18}$$

in which $n_{\text{refresh}}(n)$ and $n_{\text{events}}(n)$ are respectively the number of refreshments and the number of events up to time n, q_n is a positive, decreasing sequence such that $q_n \to 0$, and $\lambda^* = 0.7812$.

For the strategy in Section 5.2.4 we used that at time n there is a probability of adapting equal to p_n, such that $p_n \to 0$. In the remainder of this section we show that for the adaptive scheme described in (5.18) the diminishing adaptation condition holds even if the refreshment rate is updated with probability 1.

Lemma 5.27. *Let π be a probability distribution. Denote as P^t_γ the semigroup of a ZZ process with excess switching rate $\gamma : E \to \mathbb{R}^d_+$. Let γ_1, γ_2 be two bounded excess switching rates. Then for any $t \geq 0$*

$$\sup_{(x,\theta)\in E} \|\delta_{(x,\theta)} P^t_{\gamma_1} - \delta_{(x,\theta)} P^t_{\gamma_2}\|_{\text{TV}} \to 0 \qquad \text{as } \|\gamma_1 - \gamma_2\|_\infty \to 0. \tag{5.19}$$

Similarly, denote now by P^t_λ the semigroup of a BPS with refreshment rate $\lambda : E \to \mathbb{R}_+$. Let λ_1, λ_2 be two bounded refreshment rates. Then for any $t \geq 0$

$$\sup_{(x,\theta)\in E} \|\delta_{(x,\theta)} P^t_{\lambda_1} - \delta_{(x,\theta)} P^t_{\lambda_2}\|_{\text{TV}} \to 0 \qquad \text{as } \|\lambda_1 - \lambda_2\|_\infty \to 0. \tag{5.20}$$

Proof. Consider first the case of the BPS. Consider the generators of two BPS's with refreshment rates λ_1 and λ_2 and denote them respectively by $\mathcal{L}_{\lambda_1}$ and $\mathcal{L}_{\lambda_2}$. Now observe that $\mathcal{L}_{\lambda_2} = \mathcal{L}_{\lambda_1} + B$, where B is a bounded operator such that

$$Bf(x,\theta) = (\lambda_2(x,\theta) - \lambda_1(x,\theta)) \int_{\mathbb{R}^d} (f(x,\theta') - f(x,\theta))\, \nu(\mathrm{d}\theta').$$

In other words, $\mathcal{L}_{\lambda_2}$ is a perturbed version of $\mathcal{L}_{\lambda_1}$, with bounded perturbation B. Since $\mathcal{L}_{\lambda_1}$ and $\mathcal{L}_{\lambda_2}$ generate strongly continuous semigroups $(P^t_{\lambda_1})_{t\geq 0}, (P^t_{\lambda_2})_{t\geq 0}$, we can apply the reasoning in the proof of Corollary 1.11 in Chapter 3 of [67] to obtain that, for any $t > 0$, $\|P^t_{\lambda_1} - P^t_{\lambda_2}\| \to 0$ as $B \to 0$, which is equivalent to

$$\sup_{f\in\mathcal{C}_0(E), |f|\leq 1} \sup_{(x,\theta)\in E} |P^t_{\lambda_1} f(x,\theta) - P^t_{\lambda_2} f(x,\theta)| \to 0 \qquad \text{as } \|\lambda_1 - \lambda_2\|_\infty \to 0,$$

where $\mathcal{C}_0(E)$ is the space of continuous functions on E that vanish at infinity. Inverting the order of the two suprema and applying the Riesz-Markov representation theorem (see e.g. [7]) we obtain the result in (5.20).

Now consider the case of two ZZS with two different bounded excess switching rates $\gamma_1 : E \to \mathbb{R}_+^d$ and $\gamma_2 : E \to \mathbb{R}_+^d$. The perturbation is now given by

$$\tilde{B}f(x,\theta) = \sum_{i=1}^{d}((\gamma_2(x,\theta))_i - (\gamma_1(x,\theta))_i)(f(x,R_i\theta) - f(x,\theta)).$$

Then the result (5.19) follows by the same arguments as above. □

Proposition 5.28. *Consider the adaptive PDMC algorithms defined in Section 5.2.4, with the following modification. Let the adaptation of the refreshment rate be performed with probability* 1 *at adaptation times, but in such a way that*

$$\sup_{(x,\theta)\in E} |\lambda_r^{n+1}(x,\theta) - \lambda_r^n(x,\theta)| \leq \delta_{n+1},$$

where $(\delta_n)_{n\geq 1}$ *is such that* $\delta_n \to 0$ *as* $n \to \infty$*, and similarly for the refreshment rate of the ZZS. Then for any* $\Delta t > 0$

$$\lim_{n\to\infty}\left(\sup_{(x,\theta)\in E}\|P^{\Delta t}_{M_{n+1},\,\lambda_r^{n+1}}((x,\theta),\cdot) - P^{\Delta t}_{M_n,\,\lambda_r^n}((x,\theta),\cdot)\|_{\mathrm{TV}}\right) = 0 \quad \textit{in probability,}$$

and similarly for the ZZS.

Proof. By the triangle inequality for the total variation distance we obtain

$$\begin{aligned}
&\|\delta_{(x,\theta)}P^{\Delta t}_{M_{n+1},\,\lambda_r^{n+1}} - \delta_{(x,\theta)}P^{\Delta t}_{M_n,\,\lambda_r^n}\|_{\mathrm{TV}} \leq \\
&\leq \|P^{\Delta t}_{M_{n+1},\,\lambda_r^{n+1}}((x,\theta),\cdot) - P^{\Delta t}_{M_{n+1},\,\lambda_r^n}((x,\theta),\cdot)\|_{\mathrm{TV}} \\
&\quad + \|P^{\Delta t}_{M_{n+1},\,\lambda_r^n}((x,\theta),\cdot) - P^{\Delta t}_{M_n,\,\lambda_r^n}((x,\theta),\cdot)\|_{\mathrm{TV}}.
\end{aligned} \tag{5.21}$$

The inequality above allows us to deal with the two adaptations separately. The second term in the right hand side of (5.21) is handled by Proposition 5.25.

Now focus on the first term in the right hand side of (5.21), for which we want to apply Lemma 5.27. It is sufficient to observe that supremum norm of the perturbation goes to zero in probability when $n \to \infty$. Indeed the sequence of refreshment rates is Cauchy in probability, as for all n we have $\|\lambda_r^{n+1} - \lambda_r^n\|_\infty \leq \delta_{n+1}$. Convergence in probability follows from the fact that $(\delta_n)_{n\geq 0}$ is a convergent sequence and therefore the diminishing adaptation condition holds.

□

5.A.4 Details on the implementation of a Gaussian mixture target

Consider the target of Section 5.5.3, that is

$$\pi(x) \propto \lambda \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) + (1-\lambda)\exp\left(-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right).$$

Here we wish to find upper and lower bounds to the Hessian of the negative log-density, which can be used to simulate the adaptive ZZS and BPS following the approach of Appendix 5.A.1. Start by writing $w = \frac{1}{2}(\mu_2+\mu_1)$ and $v = \frac{1}{2}(\mu_2-\mu_1)$. Then we have

$$\begin{aligned}\pi(x) &\propto \exp\left(-\frac{1}{2}(x-w)^T\Sigma^{-1}(x-w)\right)\left[\alpha\exp\left(\mu_1^T\Sigma^{-1}x - w^T\Sigma^{-1}x\right)\right.\\ &\quad\left.+\,\beta\exp\left(\mu_2^T\Sigma^{-1}x - w^T\Sigma^{-1}x\right)\right]\\ &= \exp\left(-\psi_1(x)-\psi_2(x)\right)\end{aligned}$$

where

$$\alpha = \lambda\exp\left(\frac{1}{2}v^T\Sigma^{-1}(w+\mu_1)\right),$$
$$\beta = (1-\lambda)\exp\left(-\frac{1}{2}v^T\Sigma^{-1}(w+\mu_2)\right),$$

while

$$\psi_1(x) = \frac{1}{2}(x-w)^T\Sigma^{-1}(x-w)$$

and

$$\psi_2(x) = -\log\left[\alpha\exp(-(\Sigma^{-1}v)^Tx) + \beta\exp((\Sigma^{-1}v)^Tx)\right].$$

Clearly $\nabla^2\psi_1(x) = \Sigma^{-1}$. Write $m(x) = (\Sigma^{-1}v)^Tx$. Then $\nabla_x m(x) = \Sigma^{-1}v$. We have

$$\nabla\psi_2(x) = (\Sigma^{-1}v)\frac{\alpha\exp(-m(x)) - \beta\exp(m(x))}{\alpha\exp(-m(x)) + \beta\exp(m(x))},$$

and

$$\begin{aligned}\nabla^2\psi_2(x) &= (\Sigma^{-1}v)(\Sigma^{-1}v)^T\left[-1 + \frac{(\alpha\exp(-m(x)) - \beta\exp(m(x)))^2}{(\alpha\exp(-m(x)) + \beta\exp(m(x)))^2}\right]\\ &= (\Sigma^{-1}v)(\Sigma^{-1}v)^T\left[\frac{-4\alpha\beta}{(\alpha\exp(-m(x)) + \beta\exp(m(x)))^2}\right].\end{aligned}$$

It remains to find the minimum of

$$m \mapsto \frac{-4\alpha\beta}{(\alpha\exp(-m) + \beta\exp(m))^2}.$$

This is achieved at $m = \frac{1}{2}\ln\frac{\alpha}{\beta}$, yielding that

$$-\frac{1}{4}\left(\Sigma^{-1}(\mu_2-\mu_1)\right)\left(\Sigma^{-1}(\mu_2-\mu_1)\right)^T \preceq \nabla^2\psi_2(x) \preceq 0.$$

We conclude that

$$\Sigma^{-1} - \frac{1}{4}\left(\Sigma^{-1}(\mu_2-\mu_1)\right)\left(\Sigma^{-1}(\mu_2-\mu_1)\right)^T \preceq \nabla^2\psi(x) \preceq \Sigma^{-1}. \tag{5.22}$$

Remark 5.29 (Some special situations). Consider the case in which $v = \frac{1}{2}(\mu_2-\mu_1)$ is an eigenvector of Σ^{-1} with eigenvalue γ. Let Q denote the positive definite matrix with eigenvalues identical to those of Σ^{-1} along the directions orthogonal to v, and with eigenvalue $\max(\gamma, \gamma^2\|v\|^2 - \gamma)$ along v. Then $-Q \preceq \nabla^2\psi(x) \preceq Q$. In particular

$$\|\nabla^2\psi(x)\|_2 \le \max(\kappa, \gamma, \gamma^2\|v\|^2 - \gamma),$$

where κ denotes the maximal eigenvalue of Σ^{-1} restricted to the orthogonal complement of v.

In the special case for which $\Sigma = I_d$, we find for the lower bound

$$\nabla^2\psi \succeq I_d - \frac{1}{4}(\mu_2-\mu_1)(\mu_2-\mu_1)^T,$$

with eigenvalues 1 and $1 - \frac{1}{4}\|\mu_2-\mu_1\|_2^2$. So if $\|\mu_2-\mu_1\|_2^2 \le 2$, then $-I_d \preceq \nabla^2\psi \preceq I_d$. In general

$$\|\nabla^2\psi(x)\|_2 \le \max(1, 1/4\|\mu_2-\mu_1\|_2^2 - 1) \quad \text{for all } x \in \mathbb{R}^d.$$

5.B Proofs

5.B.1 Proofs of Section 5.2

Proof of Proposition 5.1. The extended generator $(\tilde{\mathcal{L}}_M, \mathcal{D}(\tilde{\mathcal{L}}_M))$ of the standard ZZ process is by [48, Definition 14.15] such that for all $\tilde{f} \in \mathcal{D}(\tilde{\mathcal{L}}_M)$

$$M_t^{\tilde{f}} = \tilde{f}(\Xi_t, \Theta_t) - \tilde{f}(\xi, \theta) - \int_0^t \tilde{\mathcal{L}}_M \tilde{f}(\Xi(s), \Theta(s))\mathrm{d}s \tag{5.23}$$

is a local martingale. For any $\tilde{f} \in \mathcal{D}(\tilde{\mathcal{L}}_M)$ we define a function $h(x,\theta) = \tilde{f}(\xi,\theta)$ with $x = M\xi$, for any $M \in \mathcal{M}$. Observe that $h(X_t, \Theta_t) = \tilde{f}(\Xi_t, \Theta_t)$ since $X_t = M\,\Xi_t$. The generator of the transformed process then satisfies Condition (5.23) provided that

$$\mathcal{L}_M h(x,\theta) = \tilde{\mathcal{L}}_M \tilde{f}(\xi,\theta) \qquad \text{for all } (x,\theta) \in E.$$

Therefore

$$\begin{aligned}\tilde{\mathcal{L}}_M \tilde{f}(\xi,\theta) &= \langle \theta, \nabla_\xi h(M\xi,\theta)\rangle + \sum_{i=1}^{d} \tilde{\lambda}_{M,i}(\xi,\theta)(h(M\xi,R_i\theta) - h(M\xi,\theta)) \\ &= \langle M\theta, \nabla_x h(x,\theta)\rangle + \sum_{i=1}^{d} \tilde{\lambda}_{M,i}(M^{-1}x,\theta)(h(x,R_i\theta) - h(x,\theta)) \\ &= \mathcal{L}_M h(x,\theta).\end{aligned}$$

□

Proof of Proposition 5.2. Following the approach in the proof of Theorem 2.2 in [23], we want to check that for any $M \in \mathcal{M}$, and any $h \in \mathcal{D}(\mathcal{L}_M)$, it holds that $\int_E \mathcal{L}_M h(x,v)\mathrm{d}\mu = 0$. By a change of variable and using the same reasoning as in the proof of Proposition 5.1, for any

$$\begin{aligned}\int \mathcal{L}_M h(x,\theta)\mathrm{d}\mu(x,\theta) &= \frac{1}{Z}\sum_{\theta\in\{-1,+1\}^d}\int \mathcal{L}_M h(x,\theta)\exp(-\psi(x))\mathrm{d}x \\ &= \frac{1}{\tilde{Z}}\sum_{\theta\in\{-1,+1\}^d}\int \tilde{\mathcal{L}}_M \tilde{f}(\xi,\theta)\exp(-\tilde{\psi}_M(\xi))\mathrm{d}\xi \\ &= \int \tilde{\mathcal{L}}_M \tilde{f}(\xi,\theta) d\tilde{\mu}_M(\xi,\theta) \;=\; 0,\end{aligned}$$

where $\tilde{\mu}_M = \tilde{\pi}_M \otimes \mathrm{Unif}(\{-1,+1\}^d)$ the last equality was obtained by the invariance of $\tilde{\mu}_M$ for the standard Zig-Zag process. □

Proof of Proposition 5.3. Following the same approach as in the proof of Proposition 5.1, for any $\tilde{f} \in \mathcal{D}(\tilde{\mathcal{L}}_M)$ we define a function $h(x,\theta) = \tilde{f}(\xi,\theta)$ with $x = M\xi$. The generator of the transformed process then satisfies Condition (5.23) provided that

$$\mathcal{L}_M h(x,\theta) = \tilde{\mathcal{L}}_M \tilde{f}(\xi,\theta) \qquad \text{for all } (x,\theta) \in E.$$

In this case

$$\begin{aligned}\tilde{\mathcal{L}}_M \tilde{f}(\xi,\theta) &= \langle \theta, \nabla_\xi h(M\xi,\theta)\rangle + \tilde{\lambda}_M(\xi,\theta)(h(M\xi,\tilde{R}(\xi)\theta) - h(M\xi,\theta)) \\ &\quad + \lambda_r(\xi,\theta)\int (h(M\xi,\theta') - h(M\xi,\theta))\nu(\mathrm{d}\theta') \\ &= \langle M\theta, \nabla_x h(x,\theta)\rangle + \tilde{\lambda}_M(M^{-1}x,\theta)(h(x,\tilde{R}(M^{-1}x)\theta) - h(x,\theta)) \\ &\quad + \lambda_r(M^{-1}x,\theta)\int (h(x,\theta') - h(x,\theta))\nu(\mathrm{d}\theta') \\ &= \mathcal{L}_M h(x,\theta).\end{aligned}$$

□

Proof of Proposition 5.4. As for Proposition 5.2, it suffices to show that for all $M \in \mathcal{M}$ $\int_E \mathcal{L}_M h(x,v)\mathrm{d}\mu = 0$. By a change of variable we obtain

$$\begin{aligned}\int \mathcal{L}_M h(x,\theta)\mathrm{d}\mu(x,\theta) &= \frac{1}{Z}\int_{\mathsf{X}}\int_{\mathcal{V}} \mathcal{L}_M h(x,\theta)\exp(-\psi(x))\nu(\theta)\mathrm{d}\theta\mathrm{d}x \\ &= \frac{1}{\tilde{Z}}\int_{\mathsf{X}}\int_{\mathcal{V}} \tilde{\mathcal{L}}_M \tilde{f}(\xi,\theta)\exp(-\tilde{\psi}_M(\xi))\nu(\theta)\mathrm{d}\theta\mathrm{d}\xi \\ &= \int \tilde{\mathcal{L}}_M \tilde{f}(\xi,\theta)d\tilde{\mu}(\xi,\theta) \;=\; 0,\end{aligned}$$

where $\tilde{\mu}_M = \tilde{\pi}_M \otimes \nu$ and the last equality follows by the invariance of $\tilde{\mu}_M$ for the BPS as shown in [32]. □

5.B.2 Proof of Theorem 5.9

Let us initially separate the proof under Assumption 5.7 and next under Assumption 5.8. First we show that in both cases the containment condition holds if the sequence $\{V_{\Gamma_n}(Z_n)\}_{n\geq 0}$ is bounded in probability.

5.B.2.1 The case of Assumption 5.7

Let $\gamma \in \mathcal{Y}$ and let C be the uniform $(\nu_\gamma, \delta, n_0)$-small set as defined in Assumption 5.7. As described in [134, 141] we define the following coupling between two chains with kernel P_γ. The two chains evolve independently each with kernel P_γ when they are outside of C. When both chains are inside C, then with probability δ we let the two chains couple after n_0 steps by drawing $X_{n+n_0} = Y_{n+n_0}$ from ν_γ, or with probability $(1-\delta)$ we independently draw $X_{n+n_0} \sim (P_\gamma^{n_0}(X_n,\cdot) - \delta\nu_\gamma(\cdot))/(1-\delta)$ and $Y_{n+n_0} \sim (P_\gamma^{n_0}(Y_n,\cdot) - \delta\nu_\gamma(\cdot))/(1-\delta)$. Then by taking $h_\gamma(x,y) = (V_\gamma(x) + V_\gamma(y))/2$ as suggested in the proof of Theorem 3 in [8] we have

$$\mathbb{E}_\gamma\left(h_\gamma(X_1,Y_1)\middle|X_0 = x, Y_0 = y\right) \leq \lambda h_\gamma(x,y) \qquad \text{for all } (x,y) \notin C \times C.$$

Define now $A(\gamma) \coloneqq \sup_{(x,y)\in C\times C}\mathbb{E}_\gamma\left(h_\gamma(X_{n_0},Y_{n_0})\middle|X_0 = x, Y_0 = y\right)$. Then by [141, Theorem 5] we have for each $\gamma \in \mathcal{Y}$ the following bound on the total variation distance

$$||P_\gamma^n(x,\cdot) - \mu(\cdot)||_{TV} \leq (1-\delta)^{\lfloor \frac{\sqrt{n}}{n_0} \rfloor} + \lambda^{n-\sqrt{n}n_0+1}(A(\gamma))^{\sqrt{n}-1}(V_\gamma(z) + \pi(V_\gamma))/2.$$

The dependence on γ in $A(\gamma)$ can be eliminated by noting that

$$A(\gamma) \leq \sup_{\gamma\in\mathcal{Y}}(A(\gamma)) \leq \lambda^{n_0}\sup_{\{\gamma\in\mathcal{Y},\, x\in C\}} V_\gamma(x) + bn_0 =: B,$$

which is independent of γ. In particular by our assumptions we have that $B < \infty$ and $\pi(V_\gamma) < \infty$ for each γ.

5.B.2.2 The case of Assumption 5.8

Define for each V_γ and any measurable function $\varphi : E \to \mathbb{R}$ and any $\beta \geq 0$ the norm

$$\|\varphi\|_{\beta,V_\gamma} = \sup_{z\in E} \frac{|\varphi(z)|}{1+\beta V_\gamma(z)}. \tag{5.24}$$

For measures μ_1, μ_2 on E such that, for each $\gamma \in \mathcal{Y}$, $\mu_1(V_\gamma) < \infty$, $\mu_2(V_\gamma) < \infty$, consider the V_γ weighted norm

$$\rho_\beta(\mu_1,\mu_2) = \sup_{\{\varphi:\|\varphi\|_{\beta,V_\gamma}\leq 1\}} (\mu_1(\varphi) - \mu_2(\varphi)). \tag{5.25}$$

Note that ρ_β depends on γ. Now for each γ we can apply [64, Theorem 24] and therefore there exist $\beta^* > 0$ and $\kappa \in (0,1)$ such that

$$\rho_{\beta^*}(\mu_1 P_\gamma, \mu_2 P_\gamma) \leq \kappa \rho_{\beta^*}(\mu_1,\mu_2),$$

where in particular κ and β^* depend only on α, λ, C_1 as defined in Assumption 5.8. Therefore κ and β^* do not depend on γ by uniformity of said constants. The norm (5.24) is such that if $\|\varphi\|_{0,V_\gamma} \leq 1$, then $\|\varphi\|_{\beta,V_\gamma} \leq 1$ for any $\beta > 0$. Choosing $\mu_1 = \delta_z$, $\mu_2 = \pi$, for any $z \in E$ we obtain

$$\begin{aligned}
\|P_\gamma^n(z,\cdot) - \pi(\cdot)\|_{\mathrm{TV}} &\leq \rho_{\beta^*}(\delta_z P_\gamma^n, \pi) \\
&\leq \kappa^n \rho_{\beta^*}(\delta_z,\pi) \\
&= \kappa^n \sup_{\{\varphi:\|\varphi\|_{\beta^*,V_\gamma}\leq 1\}} (\varphi(z) - \pi(\varphi)) \\
&\leq \kappa^n \left(2 + \beta^* V_\gamma(z) + \beta^* \pi(V_\gamma)\right)
\end{aligned}$$

Here we used (5.25) in the third equality, and twice (5.24) in the last inequality. Moreover by assumption $\pi(V_\gamma) < +\infty$ for all $\gamma \in \mathcal{Y}$.

5.B.2.3 Boundedness in probability of the sequence of Lyapunov functions

In both cases the containment condition is thus satisfied if the process $\{V_{\Gamma_n}(Z_n)\}_{n\geq 0}$ is bounded in probability. This is indeed implied by our assumption that Γ_n is updated only if $Z_n \in B$, where B is a compact set, as suggested in [44, 38, 126]. For the sake of completeness we report the main steps.

By [135, Lemma 3] it is enough to show that $\sup_{n\in\mathbb{N}} \mathbb{E}(V_{\Gamma_n}(Z_n)) < \infty$. By our assumptions and letting $D := \sup_{\{z\in B,\, \gamma\in\mathcal{Y}\}} V_\gamma(z) < \infty$

$$\begin{aligned}
&\mathbb{E}(V_{\Gamma_{n+1}}(Z_{n+1})|Z_n = z, \Gamma_n = \gamma) = \\
&\quad = \mathbb{E}(V_{\Gamma_{n+1}} Z_{n+1} \left(\mathbb{1}(Z_{n+1} \notin B) + \mathbb{1}(Z_{n+1} \in B)\right) | Z_n = z, \Gamma_n = \gamma) \\
&\quad \leq \mathbb{E}(V_{\Gamma_n}(Z_{n+1}) \mathbb{1}(Z_{n+1} \notin B) | Z_n = z, \Gamma_n = \gamma) + D \\
&\quad \leq \lambda V_\gamma(z) + b + D
\end{aligned}$$

Taking expectations on both sides one obtains

$$\mathbb{E}(V_{\Gamma_{n+1}}(Z_{n+1})) \leq \lambda \mathbb{E}(V_{\Gamma_n}(Z_n)) + b + D.$$

This shows that the sequence is contracting since $\lambda \in (0,1)$ and this is enough by [135, Lemma 2] to show our boundedness in probability of the sequence.

5.B.3 Proof of Theorem 5.11

To prove Theorem 5.11 it is sufficient to verify that the discretised process with time step Δt satisfies the assumptions of Theorem 5.9.

Let us first consider the assumptions in alternative (1) of Theorem 5.11. The uniform small set condition of the discretised process (Assumption 5.7(a)) follows by choosing $n_0 = \frac{t_0}{\Delta t}$. The geometric drift condition (Assumption 5.7(b)) is a consequence of Lemma 5.33.

Now consider the second set of assumptions, that is alternative (2) in the theorem. In this case we wish to show that Assumption 5.8 holds for the discretised process. Part (a) of Assumption 5.8 is immediately verified by the fact that $\Delta t = t_0$. For part (b), first observe that from Lemma 5.33 by our assumptions we have

$$P_\gamma^{\Delta t} V_\gamma(z) \leq e^{-A_1 \Delta t} V_\gamma(z) + \frac{A_2}{A_1}(1 - e^{-A_1 \Delta t}).$$

In the notation of Assumption 5.8(b) we have $\lambda = e^{-A_1 \Delta t}$ and $C_1 = \frac{A_2}{A_1}$. In particular we have $2C_1 < C_2$ since we assumed that $C_2 > 2A_2/A_1$.

The thesis now follows in both cases by Theorem 5.9.

5.B.4 Proof of Theorem 5.14

5.B.4.1 Minorisation condition for the ZZS

In this section we show that a simultaneous small set condition holds for the family of ZZ processes as stated in Lemma 5.20. The strategy of the proof is to reduce the d-dimensional minorisation condition to 1-dimensional conditions for every component of the process. In Lemma 5.32 we prove a simultaneous minorisation condition for the ZZ process in the one-dimensional case.

Assumption 5.30 (Assumption 3 in [21]). *Let $\psi \in \mathcal{C}^2(\mathbb{R})$. Define the switching rates of a 1-d ZZ process as $\lambda(x,\theta) = (\theta\psi'(x))_+$, where $x \in \mathbb{R}$ and $\theta \in \{-1,+1\}$. There exists $x_0 > 0$ such that*

$$\inf_{x \geq x_0} \lambda(x,+1) > \sup_{x \geq x_0} \lambda(x,-1),$$
$$\inf_{x \geq -x_0} \lambda(x,-1) > \sup_{x \leq -x_0} \lambda(x,+1).$$

Remark 5.31. Assumption 5.30 requires that there exists a point $x_0 \in \mathbb{R}$ such that the process has a strictly positive probability of changing direction if it is outside of and moving outwards of the set $[-x_0, +x_0]$. The lemma below states that a small set condition follows from this assumption.

Lemma 5.32. *Consider the family of 1-dimensional Zig-Zag processes with generators* $\{\mathcal{L}_m : m \in \mathcal{M}\}$ *where* $\mathcal{M} = [V_{\min}, V_{\max}]$ *for some* $0 < V_{\min} \leq V_{\max} < \infty$. *Thus for* $f \in \mathcal{D}(\mathcal{L}_m)$ *and* $m \in [V_{\min}, V_{\max}]$

$$\mathcal{L}_m f(x,\theta) = m\theta f'(x,\theta) + (m(\theta\psi'(x))_+ + \gamma(x))(f(x,-\theta) - f(x,\theta)).$$

Assume either of the two conditions:

- $\gamma(x) = 0$ *for all* $x \in \mathbb{R}$, *but Assumption 5.30 holds;*
- *there exists* $\gamma_{\min} > 0$ *such that* $\gamma(x) \geq \gamma_{\min}$ *for any* $x \in \mathbb{R}$, *and* $\gamma(\cdot)$ *is bounded on compact sets.*

Then for any set of the form $C = D \times V$, *where* $D \subset \mathbb{R}$ *is a compact set and* $V \subseteq \{-1, +1\}$, *there exists* $t_0 > 0$ *such that for any* $t \geq t_0$ *there are* $\delta > 0$, *and a probability measure* ν *on* E *such that*

$$P_m^t((x,\theta),\cdot) \geq \delta\nu(\cdot) \qquad \text{for all } (x,\theta) \in C \text{ and all } m \in \mathcal{M}.$$

Moreover, consider $\{\mathcal{L}_{m,\gamma} : m \in \mathcal{M}, \gamma \in \Lambda\}$, *where* Λ *is a family of switching rates* $\gamma : \mathbb{R} \to \mathbb{R}_+$ *such that for all* $\gamma \in \Lambda$ *it holds that* $0 < \gamma_{\min} \leq \gamma(x) \leq \gamma_{\max} < \infty$ *for all* $x \in \mathbb{R}$. *Then for any compact set* C *as above, there exists* $t_0 > 0$ *such that for any* $t \geq t_0$ *there are* $\delta > 0$, *and a probability measure* ν *on* E *that satisfy*

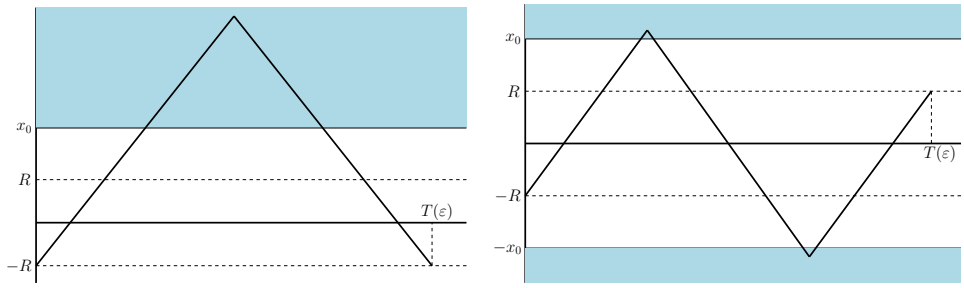
$$P_{m,\gamma}^t((x,\theta),\cdot) \geq \delta\nu(\cdot) \qquad \text{for all } (x,\theta) \in C,\ \gamma \in \Lambda,\ m \in \mathcal{M}.$$

In particular t_0 *depends on* $V_{\min}$, *and* δ *depends on* $V_{\min}, V_{\max}, \gamma_{\min}, \gamma_{\max}$, *but neither depends on* m *or* γ.

Proof. Let $R > 0$ and $C = [-R, R] \times V$, where $V \subseteq \{-1, +1\}$. Consider the family $\{\mathcal{L}_m : m \in \mathcal{M}\}$. For $m \in [V_{\min}, V_{\max}]$, the process $\mathcal{L}_m$ moves with velocity $m\theta$, so either $+m > 0$ or $-m < 0$. Consider first the case in which Assumption 5.30 is satisfied for some $x_0 > 0$. Note that as a consequence it is satisfied for any $\mathcal{L}_m$. We consider the case $x_0 > R$ because it is the most general. Indeed one is always free to take $\tilde{x}_0 > R > x_0$ and apply the reasoning below. Alternatively the case $x_0 \leq R$ follows by the same method of proof.

For $B = A \times V \subset E$ let $\nu(B) = (\mathrm{Leb}(A \cap [-R, R])/\mathrm{Leb}([-R, R])) \times (\delta_1(\theta) + \delta_{-1}(\theta))$ be a probability measure on E. Define $T = T(\varepsilon) := (4x_0 + 2R)/V_{\min} + \varepsilon$ where $\varepsilon > 0$. This choice of the time horizon allows the process to reach the region with strictly positive probability of a velocity flip for any initial position (x, θ) with $x \in [-R, R]$ and $\theta \in \{-1, +1\}$. The addition of $\varepsilon > 0$ is necessary to introduce a margin where the flip can take place. See Figure 5.8 for a visual aid.

We show that C is a uniform (ν, δ, T)-small set. Since the proof applies for all $\varepsilon > 0$, although resulting indifferent δ's, one is free to choose $t_0 = T(\varepsilon)$ freely. Let us consider the two cases below.



(a) First case. (b) Second case.

Figure 5.8: Illustration of the proof of Lemma 5.32.

First case: let the initial condition be $(x, +1)$ with $x \in [-R, R]$ and we want to be at time T in $B = A \times \{-1\}$ with $A \in \mathcal{B}([-R, +R])$ any Borel set. We can then use the following inequality

$$\mathbb{P}_{(x,+1)}((X_T, \Theta_T) \in B) \geq \mathbb{P}_{(x,+1)}(X_T \in A, E_1), \tag{5.26}$$

where E_1 is the event that exactly one velocity switch takes place. We are then in the case of Figure 5.8(a). Observe that by the choice of T the process has enough time to travel the longest path, i.e. from $(-R, +1)$ to $(-R, -1)$ with the smallest allowed velocity $V_{\min}$. In order to compute the r.h.s. in (5.26) one can compute $\mathbb{P}_{(x,+1)}(X_T \leq y, E_1)$ and then differentiate with respect to y. To do this we assume only one velocity switch and impose that $X_T \leq y$, resulting in the condition

$$X_T = x + mt - m(T - t) \leq y,$$

where t is the time at which the velocity switch takes place. By rearranging we obtain the condition $t \leq \frac{y-x}{2m} + \frac{T}{2} =: \bar{t}(y)$. Observe that for any $m \in [V_{\min}, V_{\max}]$ the process has enough time to reach the region where it is assumed there is a strictly positive probability of a velocity flip. Indeed using that $y \in [-R, R]$ we find

$$\bar{t}(y) \geq \frac{x_0 - x}{m} + \frac{y + x}{2m} + \frac{2x_0 + 2R}{2V_{\min}} + \frac{\varepsilon}{2} \geq \frac{x_0 - x}{m} + \frac{x_0}{V_{\min}} + \frac{\varepsilon}{2} > \frac{x_0 - x}{m}.$$

Therefore

$$\mathbb{P}_{(x,+1)}(X_T \leq y, E_1) = \int_0^{\bar{t}(y)} \lambda(x + ms, 1) \exp\left(-\int_0^s \lambda(x + mu, 1)\mathrm{d}u\right) \exp\left(-\int_0^{T-s} \lambda(x + ms - mu, -1)\mathrm{d}u\right) \mathrm{d}s.$$

After differentiation one obtains

$$\begin{aligned}
&\mathbb{P}_{(x,+1)}(X_T \in A, E_1) = \\
&\quad = \int_A \frac{1}{2m}\lambda(x + m\bar{t}(y), 1) \exp\left(-\int_0^{\bar{t}(y)} \lambda(x + mu, 1)\mathrm{d}u\right) \\
&\qquad \exp\left(-\int_0^{T-\bar{t}(y)} \lambda(x + m\bar{t}(y) - mu, -1)\mathrm{d}u\right)\mathrm{d}y \\
&\quad \geq \frac{1}{2V_{\min}} \int_A \lambda(x + m\bar{t}(y), 1) \exp(-\bar{t}(y)\lambda_{\max} - (T - \bar{t}(y))\lambda_{\max})\mathrm{d}y \\
&\quad = \frac{\exp(-\lambda_{\max}T)}{2V_{\min}} \int_A \lambda(x + m\bar{t}(y), 1)\mathrm{d}y \\
&\quad \geq 2\,\frac{\lambda_{\min}R\exp(-\lambda_{\max}T)}{V_{\min}} \left(\frac{1}{2} \cdot \frac{1}{2R}\int_A \mathrm{d}y\right) \\
&\quad = 2\,\frac{\lambda_{\min}R\exp(-\lambda_{\max}T)}{V_{\min}}\,\nu(B).
\end{aligned}$$

where

$$\begin{aligned}
\lambda_{\max} &\coloneqq \max_{\{x\in\tilde{C}_T, \theta\in\{-1,+1\}\}} (V_{\max}(\theta\psi'(x))_+ + \gamma(x)) < \infty, \\
\lambda_{\min} &\coloneqq \min\left\{\inf_{x\geq x_0}(V_{\min}\psi'(x))_+\,,\ \inf_{x\leq -x_0}(V_{\min}\psi'(x))_+\right\} > 0.
\end{aligned} \tag{5.27}$$

where $\tilde{C}_T \coloneqq [-R - V_{\max}T, R + V_{\max}T]$ is the set of points that can be reached in time T by a process starting in $x \in [-R, R]$ with velocity $V_{\max}$ (and thus for any value $m \in [V_{\min}, V_{\max}]$). Thus $\lambda_{\max}$ is the maximum switching rate achieved within time T.

Second case: again we consider an initial condition with positive velocity, but now also a positive velocity at time T, i.e. we consider a set $B = A \times \{+1\}$. Taking advantage of the same idea as before we use the bound

$$\mathbb{P}_{(x,+1)}((X_T, \Theta_T) \in B) \geq \mathbb{P}_{(x,+1)}(X_T \in A, E_2),$$

where E_2 is the event that exactly two switches take place. This case corresponds to Figure 5.8(b). Let t_1 and t_2 be the times of first and second switch respectively. Then $X_T \leq y$ when

$$X_T = x + mt_1 - m(t_2 - t_1) + m(T - t_2) \leq y,$$

which can be rearranged as $t_2 \geq t_1 + \frac{x-y}{2m} + \frac{T}{2} =: \overline{t_2}(y)$. We can obtain a bound for t_1 by imposing that $\overline{t_2}(y) < T$. The resulting condition is $t_1 < T/2 - (x-y)/(2m) =: \bar{t}_1(y)$. We observe that by definition of T it follows that for any $x, y \in [-R, R]$

$$\bar{t}_1(y) - \frac{x_0 - x}{m} \geq \frac{2x_0}{V_{\min}} - \frac{x_0 - x}{m} + \frac{\varepsilon}{2} > 0,$$

which means that there is enough time for the process to reach cyan region in Figure 5.8(b) and have the first velocity flip. The same holds true for the second velocity flip. Now compute the distribution function as

$$\mathbb{P}_{(x,+1)}(X_T \le y, E_2) = \int_{s=0}^{\bar{t}_1(y)} \int_{u=\tilde{t}_2(y)}^{T-s} \lambda(x+ms,1) \exp\left(-\int_0^s \lambda(x+ml,1)\mathrm{d}l\right) \lambda(x+ms-mu,-1) \exp\left(-\int_0^u \lambda(x+ms-ml,-1)\mathrm{d}l\right) \exp\left(-\int_0^{T-u-s} \lambda(x+ms-mu+ml,1)\mathrm{d}l\right)\mathrm{d}s\,\mathrm{d}u,$$

where $\tilde{t}_2(y) := \overline{t_2}(y) - t_1 = \frac{x-y}{2m} + \frac{T}{2}$. Then by differentiating we obtain

$$\begin{aligned}
&\mathbb{P}_{(x,+1)}(X_T \in A, E_2) = \\
&\quad = \int_A \int_0^{\bar{t}_1(y)} \frac{1}{2m} \lambda(x+ms,1) \exp\left(-\int_{l=0}^s \lambda(x+ml,1)\mathrm{d}l\right) \lambda(x+ws-w\tilde{t}_2(y),-1) \exp\left(-\int_{l=0}^{\tilde{t}_2(y)} \lambda(x+ms-ml,-1)\mathrm{d}l\right) \exp\left(-\int_{l=0}^{T-\tilde{t}_2(y)-s} \lambda(x+ms-m\tilde{t}_2(y)+ml,1)\mathrm{d}l\right)\mathrm{d}s\,\mathrm{d}y \\
&\quad \ge \frac{\exp(-\lambda_{\max}T)}{2V_{\min}} \int_A \int_0^{\bar{t}_1(y)} \lambda(x+ws,1)\lambda(x+ws-w\tilde{t}_2(y),-1)\,\mathrm{d}s\,\mathrm{d}y.
\end{aligned}$$

Since the integrand is non-negative we can lower bound this quantity by restricting the domain of integration corresponding to the s variable. A sensible choice is $(x_0 - x)/m \le s \le -(x_0+x)/m + \tilde{t}_2(y)$, as this would imply that both switches take place in the cyan region in Figure 5.8(b). Indeed for $s \in [(x_0-x)/m, -(x_0+x)/m + \tilde{t}_2(y)]$ we have that $x + ms \ge x_0$ and $x + ms - m\tilde{t}_2(y) \le -x_0$. Observe that

$$-\frac{x_0+x}{m} + \tilde{t}_2(y) = \frac{T}{2} + \frac{x-y}{2m} - \frac{x+x_0}{m} \le \frac{T}{2} + \frac{x-y}{2m} - \frac{x-y}{m} \le \bar{t}_1(y),$$

where we used that $-x_0 \le y$ since $y \in [-R, R]$. Combining this with the fact that $\bar{t}_1(y) > \frac{x_0-x}{m}$, we have shown that $[(x_0-x)/m, -(x_0+x)/m + \tilde{t}_2(y)] \subset [0, \bar{t}_1(y)]$. Therefore on this interval we can bound below the switching rates as follows

$$\begin{aligned}
&\mathbb{P}_{(x,+1)}(X_T \in A, E_2) \ge \\
&\quad \ge \frac{\exp(-\lambda_{\max}T)}{2V_{\min}} \int_A \int_{\frac{x_0-x}{m}}^{-\frac{x_0+x}{m}+\tilde{t}_2(y)} \lambda(x+ms,1)\lambda(x+ms-w\tilde{t}_2(y),-1)\mathrm{d}s\mathrm{d}y \\
&\quad \ge \frac{\exp(-\lambda_{\max}T)}{2V_{\min}} \lambda_{\min}^2 \int_A \int_{\frac{x_0-x}{m}}^{-\frac{x_0+x}{m}+\tilde{t}_2(y)} \mathrm{d}s\mathrm{d}y
\end{aligned}$$

$$\geq \frac{\exp(-\lambda_{\max}T)}{2V_{\min}}\lambda_{\min}^2 \int_A \frac{\varepsilon}{2}\mathrm{d}y$$
$$= \frac{R\exp(-\lambda_{\max}T)}{V_{\min}}\lambda_{\min}^2\ \varepsilon\,\nu(B).$$

By symmetry the same bounds hold also when the process has initial velocity is -1. Therefore for any $\varepsilon > 0$ we proved that $C = [-R, +R] \times \{-1, +1\}$ is a uniform (ν, δ, T)-small set with

$$\delta = \min\left\{2\,\frac{\lambda_{\min}R\exp(-\lambda_{\max}T)}{V_{\min}},\ \frac{\varepsilon\,\lambda_{\min}^2 R\exp(-\lambda_{\max}T)}{V_{\min}}\right\}. \tag{5.28}$$

Notice that there is an ε dependence also in $\lambda_{\min}$ and $\lambda_{\max}$, but there is no dependence on the specific value of m. To conclude the argument we observe that it is always possible to incorporate any compact set in a set of the same form as C and therefore all compact sets are uniformly small for the family of Zig-Zag processes.

When the switching rate is strictly positive but Assumption 5.30 does not hold, the same reasoning can be applied by taking $x_0 = 0$, $\lambda_{\min} = \gamma_{\min} > 0$, and $T = \frac{2R}{V_{\min}} + \varepsilon$ for some $\varepsilon > 0$. The switching rates can be upper-bounded because γ is bounded on compact sets. Thus the only difference is that in this case the process can switch at any time and does not need to escape a compact set to do so. One can again obtain the uniform small set condition with δ defined as in (5.28), and ν as above.

Finally, consider the family $\{\mathcal{L}_{m,\gamma} : m \in \mathcal{M}, \gamma \in \Lambda\}$. Choosing

$$\lambda_{\max} = \max_{\{x\in\tilde{C}_T, \theta\in\{-1,+1\}\}} (V_{\max}(\theta\psi'(x))_+) + \gamma_{\max}$$

and applying the same reasoning as in the case of a strictly positive refreshment shows the statement. □

Proof of Lemma 5.20. We first show that all compact sets are uniformly small for the family of standard Zig-Zag processes with targets in $\{\tilde{\pi}_M(\xi) = \pi(M\xi) : M \in \mathcal{M}\}$. Let C be a d-dimensional rectangle of the form $[-R, +R]^d \times V$ for some $R > 0$ and $V \subseteq \Theta$.

Let $M \in \mathcal{M}$ and $\gamma \in \Lambda$ and denote as $(\Xi_t^{M,\gamma}, \Theta_t^{M,\gamma})_{t\geq 0}$ the ZZ process that targets $\tilde{\pi}_M(\xi)$ and has switching rate $\tilde{\gamma}_M(\xi, \theta) = \gamma(M\xi, \theta)$. Let $\tilde{P}_{M,\gamma}$ be the corresponding semigroup. Then observe that for any rectangle $B = B_1 \times \cdots \times B_d$, with $B_i = [R_{i,1}, R_{i,2}]$ for some $R_{i,2} > R_{i,1}$ the following holds

$$\begin{aligned}\tilde{P}^t_{M,\gamma}((\xi,\theta), B) &= \mathbb{P}_{(\xi,\theta)}((\Xi_t^{M,\gamma}, \Theta_t^{M,\gamma}) \in B)\\ &= \prod_{i=1}^{d} \mathbb{P}_{(\xi,\theta)}\big((\Xi_t^{M,\gamma}, \Theta_t^{M,\gamma})_i \in B_i \,|\, (\Xi_t^{M,\gamma}, \Theta_t^{M,\gamma})_j \in B_j \text{ for } j > i\big).\end{aligned} \tag{5.29}$$

Then observe that, independently of the values at any time $s < t$ of the other components of the process, the i-th switching rate $\tilde{\lambda}_{M,i}(\xi,\theta) = (\theta_i \partial_i \psi(M\xi))_+ + \gamma_i(M\xi,\theta)$ satisfies the bounds

$$0 < \gamma_{\min} \leq \tilde{\lambda}_{M,i}(\xi,\theta) \leq \lambda_{\max} < \infty, \tag{5.30}$$

where we have defined

$$\lambda_{\max} \coloneqq \max_{i=1,\dots,d} \max_{M\in\mathcal{M}} \max_{\{\xi\in\tilde{C}_t,\theta\in\Theta\}} \left\{(\theta_i \partial_i \psi(M\xi))_+\right\} + \gamma_{\max}$$

with $\tilde{C}_t = [-R-t, R+t]^d$ is the set of reachable points in time t. Here $\lambda_{\max}$ is well defined because $\psi \in \mathcal{C}^1$. In particular neither of the bounds in (5.30) depend on the specific $\tilde{\pi}_M$ and thus hold for any $M \in \mathcal{M}$. Therefore, we can apply Lemma 5.32 with $t_0 = 2R + \varepsilon$ with $\varepsilon > 0$ to each component of the product (5.29). It then follows that for any $t \geq t_0$ there exists $\delta_1 > 0$ that satisfies

$$\tilde{P}^t_{M,\gamma}((\xi,\theta),B) \geq \prod_{i=1}^{d} \delta_1 \nu_1(B_i) = \delta_1^d\, \nu_d(B) = \delta_d\, \nu_d(B) \qquad \text{for all } (\xi,\theta) \in C, \tag{5.31}$$

where $\nu_d = \mathrm{Leb}_d(C) \times \mathrm{Unif}(\Theta)$ is the d-dimensional equivalent of ν_1 that was defined in Lemma 5.32. Most importantly, δ depends only on the bounds on the switching rates, which are uniform for all $\gamma \in \Lambda$. Therefore Equation (5.31) holds for the same δ and ν_d for all $M \in \mathcal{M}$, $\gamma \in \Lambda$, and any rectangle $B \subset C$. Since the set of rectangles is a π-system and generates the Borel σ-algebra, by the monotone class theorem (Theorem 6.2 in [85]) it follows that the lower bound of Equation (5.31) holds for any Borel set B. Since any compact set can be included in a large enough hypercube, for all sets $C = D \times V$, with D compact and $V \subseteq \Theta$, there are ν_d and $t_0 > 0$ such that for all $t \geq t_0$ there exists $\delta_d > 0$ for which C is uniformly (t, δ_d, ν_d)-small for the family of standard ZZ processes defined above.

Now we wish to translate this result to the family of linearly transformed Zig-Zag processes with excess switching rates in Λ and target π. Denote as $(X_t^{M,\gamma}, \Theta_t^{M,\gamma})_{t\geq 0}$ the process with generator $\mathcal{L}_{M,\gamma}$ and observe that for any $t > 0$ and any set $A = A_x \times V$, for a Borel set A_x, the event $\{(X_{M,\gamma}(t), \Theta_t^{M,\gamma}) \in A\}$ is equivalent to the event $\{(\Xi_t^{M,\gamma}, \Theta_t^{M,\gamma}) \in \tilde{A}_M\}$ for $\tilde{A}_M = \{(\xi,\theta) \in E : M\xi \in A_x,\, \theta \in V\}$. Assume $C \in E$ is a compact set and let $(x,\theta) \in C$. Then define $C_{\mathcal{M}} = \{(\xi,\theta) = (M^{-1}x,\theta) : (x,\theta) \in C,\, M \in \mathcal{M}\}$ which is itself a compact set and depends on $\mathcal{M}$, but not on the specific $M \in \mathcal{M}$. Then, for all $t \geq t_0$, with t_0 large enough, it holds that

$$\begin{aligned}
P^t_{M,\gamma}((x,\theta),A) &= \mathbb{P}_{(x,\theta)}((X_t^{M,\gamma}, \Theta_t^{M,\gamma}) \in A) \\
&= \mathbb{P}_{(\xi,\theta)}((\Xi_t^{M,\gamma}, \Theta_t^{M,\gamma}) \in \tilde{A}_M) \\
&\geq \delta \nu_d(\tilde{A}_M) && \text{for all } (\xi,\theta) \in C_{\mathcal{M}} \\
&= \delta \nu_M(A) && \text{for all } (x,\theta) \in C
\end{aligned}$$

with $\nu_M(A) = \nu_d(\tilde{A}_M)$ and δ is chosen such that $C_{\mathcal{M}}$ is a uniform (t,δ,ν)-small set for the family of standard ZZ processes with targets $\tilde{\pi}_M$ and switching rates $\tilde{\gamma}_M$.

Therefore there is no dependence neither on M nor on γ in t_0 and δ and the proof is concluded. □

5.B.4.2 Drift conditions for the Zig-Zag process

Proof of Lemma 5.21. We follow the proof of Lemma 11 in [24]. Let $M \in \mathcal{M}$ and $\gamma \in \Lambda$. Applying the generator to V, which was defined in Equation (5.15), we obtain

$$\begin{aligned}\left(\frac{\mathcal{L}_{M,\gamma}V}{V}\right)(x,\theta) = \sum_{i=1}^{d} M_{ii}\theta_i \left(\alpha\partial_i\psi(x) + \sum_{j=1}^{d}\theta_j\partial_{ij}\psi(x)\phi'(\theta_j\partial_j\psi(x))\right) \\ + \sum_{j=1}^{d}\left(M_{ii}\left(\theta_i\partial_i\psi(x)\right)_+ + \gamma_i(x,\theta)\right)\cdot \\ \cdot\left(\exp\left(\phi(-\theta_i\partial_i\psi(x)) - \phi(\theta_i\partial_i\psi(x))\right) - 1\right).\end{aligned}$$

Now consider $s = \theta_i\partial_i\psi(x) \geq 0$, then

$$\begin{aligned}M_{ii}\alpha s + (M_{ii}s + \gamma_i)(\exp\left(\phi(-s) - \phi(s)\right) - 1) = \\ = M_{ii}\left(\alpha s + \left(s + \frac{\gamma_i}{M_{ii}}\right)\left(\frac{1}{1+\delta s} - 1\right)\right) \\ = M_{ii}\left((\alpha-1)s + \frac{s - \gamma_i/M_{ii}}{1+\delta s}\right) \\ \leq -M_{ii}((1-\alpha)|s| + 1/\delta).\end{aligned}$$

In case $s = \theta_i\partial_i\psi(x) < 0$ we obtain

$$\begin{aligned}M_{ii}\alpha s + (M_{ii}s + \gamma_i)(\exp\left(\phi(-s) - \phi(s)\right) - 1) = M_{ii}\left(\alpha s + \frac{\gamma_i}{M_{ii}}(1 + \delta|s| - 1)\right) \\ \leq -M_{ii}\left(\alpha - \frac{\gamma_{\max}}{M_{ii}}\delta\right)|s|.\end{aligned}$$

Note that by assumption $(\alpha - \gamma_{\max}\delta/M_{ii}) \geq (\alpha - \gamma_{\max}\delta/V_{\min}) > 0$. For the remaining term the best we can do is to derive the following bound

$$\begin{aligned}\sum_{i,j} M_{ii}\theta_i\theta_j\partial_{ij}\psi(x)\phi'(\theta_j\partial_j\psi(x)) \leq \sum_{i,j} M_{ii}\phi'(\theta_j\partial_j\psi(x))|\partial_{ij}\psi(x)| \\ \leq V_{\max}\frac{\delta}{2}\sum_{i,j}|\partial_{ij}\psi(x)|,\end{aligned}$$

where we have used that $0 \leq \phi'(s) \leq \delta/2$. Finally we obtain for any $M \in \mathcal{M}$

$$\left(\frac{\mathcal{L}_{M,\gamma}V}{V}\right)(x,\theta) \leq -\min\left(1-\alpha, \alpha - \frac{\gamma_{\max}\delta}{V_{\min}}\right)V_{\min}\sum_{i=1}^{d}|\partial_i\psi(x)| \tag{5.32}$$

$$+ \frac{V_{\max} d}{\delta} + \frac{\delta}{2} V_{\max} \sum_{i,j} |\partial_{ij} \psi(x)|,$$

which is independent of the specific M and γ and can be made arbitrarily small outside of a sufficiently large compact set C by our assumptions on ψ. □

Proof of Lemma 5.22. Consider the change of variables $\xi = M^{-1}x$. Denote as $\tilde{\mathcal{L}}_{M,\gamma}$ the generator of a ZZ process with transformed stationary measure $\tilde{\pi}_M$ and transformed excess switching rate $\tilde{\gamma}_M(\xi, \theta) = \gamma(M\xi, \theta)$. Then transforming the function in (5.16) we obtain a Lyapunov function for the standard ZZ process with transformed target:

$$\tilde{V}_M(\xi, \theta) = \exp\left(\alpha \tilde{\psi}_M(\xi) + \sum_{i=1}^{d} \phi(\theta_i \partial_i \tilde{\psi}_M(\xi)) \right),$$

where $\tilde{\psi}_M(\xi) = \psi(M\xi)$. Since $\mathcal{M}$ is a compact space of positive definite linear transformations, Assumption 5.13 is satisfied by each $\tilde{\psi}_M(\cdot)$. Then, by the proof of Lemma 11 in [24], for any constant $A_1 > 0$ there exists a large enough ball $\tilde{B}_M \coloneqq \{(\xi, \theta) \in E : \theta \in \Theta, \xi \in B(0, \tilde{R}_M)\}$ such that

$$\tilde{\mathcal{L}}_{M,\gamma} \tilde{V}_M(\xi, \theta) \leq -A_1 \tilde{V}_M(\xi, \theta) \qquad \text{for all } (\xi, \theta) \notin \tilde{B}_M.$$

In particular $\tilde{B}_M$ does not depend on γ, but only on $\gamma_{\max}$. Observe that by the proof of Lemma 11 in [24] it follows that $\tilde{R}_M$ depends continuously on M. Indeed one can show that $(\tilde{\mathcal{L}}_{M,\gamma} \tilde{V}_M(\xi, \theta))/\tilde{V}_M(\xi, \theta)$ is smaller or equal than a sum of terms which depend continuously on the components of $\nabla \tilde{\psi}_M$ and $\nabla^2 \tilde{\psi}_M$. Continuity follows from the assumption that $\psi \in \mathcal{C}^2$ and because M does not appear in other ways. Proposition 5.1 then implies that $\mathcal{L}_{M,\gamma} V_M(x, \theta) = \tilde{\mathcal{L}}_{M,\gamma} \tilde{V}_M(\xi, \theta)$. Thus for each $M \in \mathcal{M}$

$$\mathcal{L}_{M,\gamma} V_M(x, \theta) \leq -A_1 V_M(x, \theta) \qquad \text{for all } (x, \theta) \notin B_M,$$

where $B_M \coloneqq \{(x, \theta) \in E : (M^{-1}x, \theta) \in \tilde{B}_M\}$ and A_1 does not depend on M. Finally, take C to be any ball that contains all sets B_M. Then C is bounded by continuity of $\tilde{R}_M$ in M and compactness of $\mathcal{M}$. By continuity of $\mathcal{L}_{M,\gamma} V_M(x, \theta)$ in x, θ, M it follows that

$$\mathcal{L}_{M,\gamma} V_M(x, \theta) \leq -A_1 V_M(x, \theta) + A_2 \mathbb{1}_C(x, \theta),$$

in which $A_2 = \max_{\{M \in \mathcal{M},\, \gamma \in \Lambda,\, (x,\theta) \in C\}} (\mathcal{L}_{M,\gamma} V_M(x, \theta) + A_1 V_M(x, \theta))$. In particular the maximum over $\gamma \in \Lambda$ does not cause problem as the γ's are uniformly bounded. Hence A_2 is independent of the specific M and γ. □

5.B.5 Proof of Theorem 5.18

5.B.5.1 Simultaneous coupling inequality

Proof of Lemma 5.23. Let $M \in \mathcal{M}$ and $\lambda_r \in \Lambda_r$. First note that for $(\xi, \theta) = (M^{-1}x, \theta)$ and $(\tilde{\xi}, \tilde{\theta}) = (M^{-1}\tilde{x}, \tilde{\theta})$ we have the equality

$$\|P^t_{M,\lambda_r}((x, \theta), \cdot) - P^t_{M,\lambda_r}((\tilde{x}, \tilde{\theta}), \cdot)\|_{\mathrm{TV}} = \|\tilde{P}^t_{M,\lambda_r}((\xi, \theta), \cdot) - \tilde{P}^t_{M,\lambda_r}((\tilde{\xi}, \tilde{\theta}), \cdot)\|_{\mathrm{TV}}, \quad (5.33)$$

where $\tilde{P}^t_{M,\lambda_r}$ is the transition kernel of a standard BPS with energy function $\tilde{\psi}_M(\xi) = \psi(M\xi)$ as defined in Figure 2, and refreshment rate λ_r.

Our strategy is thus to apply Lemma 12 in [64] to each semigroup $\tilde{P}^t_{M,\lambda_r}$ and then take advantage of Equation (5.33). Observe that it is possible to apply the lemma because part (a) of Assumption 3.9 implies that for all $M \in \mathcal{M}$

$$\int \|\nabla \tilde{\psi}_M(\xi)\| \tilde{\pi}_M(\mathrm{d}\xi) = \int \|\nabla \psi(x)\| \pi(\mathrm{d}x) < \infty,$$

for $\tilde{\pi}_M(\xi) = \exp(-\tilde{\psi}_M(\xi))/\tilde{Z}_M$, with $\tilde{\psi}_M(\xi) = \psi(M\xi)$ and $\tilde{Z}_M = Z/|\det(M)|$. Hence the integrability condition holds for each $\tilde{P}^t_{M,\lambda_r}$ with respect to the corresponding target $\tilde{\pi}_M$. Applying the lemma and Equation (5.33) it follows that for any compact set $K \subset \{(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^d : \|x\| + \|\theta\| \leq R\}$, with $R \geq 0$, for each $M \in \mathcal{M}$ and $\lambda_r \in \Lambda_r$ there exists $\alpha = \alpha(\lambda_r, M, K) > 0$ such that for all (x,θ), $(\tilde{x},\tilde{\theta}) \in K$, for all $t > 0$

$$\|P^t_{M,\lambda_r}((x,\theta),\cdot) - P^t_{M,\lambda_r}((\tilde{x},\tilde{\theta}),\cdot)\|_{\mathrm{TV}} \leq 2\,(1 - \alpha(\lambda_r, M, K))\,. \tag{5.34}$$

Observe that there is a dependence on M due to the fact that the set of initial conditions K is transformed to $K_M \coloneqq \{(\xi,\theta) \in \mathbb{R}^d \times \mathbb{R}^d : \|M^{-1}\xi\| + \|\theta\| \leq R\}$ in Equation (5.33). This translates to a dependence on M in the coefficient α, as a consequence of the dependence of α on set K_M. This inconvenience can be avoided if we choose α such that the coupling inequality is satisfied for all initial conditions $(\xi,\theta) \in K_{\mathcal{M}}$, where $R_{\mathcal{M}}$ is such that $K_{\mathcal{M}} \coloneqq \{(\xi,\theta) \in \mathbb{R}^d \times \mathbb{R}^d : \|\xi\| + \|\theta\| \leq R_{\mathcal{M}}\}$ satisfies $K_M \subset K_{\mathcal{M}}$ for all $M \in \mathcal{M}$. Thus for all $M \in \mathcal{M}$ it holds that if $(x,\theta) \in K$, then $(\xi,\theta) = (M^{-1}x,\theta) \in K_{\mathcal{M}}$.

The last thing to do is showing that there exists a constant $\alpha^* > 0$ independent of M and λ_r such that $\alpha(\lambda_r, M, K) \geq \alpha^*$ for all $M \in \mathcal{M}$ and $\lambda_r \in \Lambda_r$. This is indeed the case for the following reasons:

- The dependence on M appears then in the following term, which in the statement of Lemma 12 in [64] appears as a factor in $\alpha(\lambda_r, M, K)$:

$$g(r) = \mathbb{P}\left(E_3 \leq r\tilde{N} \sup_{\{\xi : \|\xi\| \leq (1+E_1/\lambda_r)R_{\mathcal{M}} + (r/\lambda_r)\tilde{N}\}} \|\nabla \tilde{\psi}_M(\xi)\|\right).$$

  Here $\tilde{N} = N + (1 + E_1/\lambda_r)R_{\mathcal{M}}$ for some $N > 0$, and E_1, E_3 are independent exponential random variables with parameter 1. Because $g(r)$ is a factor in $\alpha(\lambda_r, M, K)$, we wish to bound it from below, and hence we should bound the supremum of $\|\nabla \tilde{\psi}_M(\xi)\|$ from below. Denoting $\zeta(\lambda_r) = (1 + E_1/\lambda_r)R_{\mathcal{M}} + (r/\lambda_r)\tilde{N}$, this can be done as follows:

$$\begin{aligned}\sup_{\{\xi : \|\xi\| \leq \zeta(\lambda_r)\}} \|\nabla \tilde{\psi}_M(\xi)\| &= \sup_{\{x : \|M^{-1}x\| \leq \zeta(\lambda_r)\}} \|M^T \nabla \psi(x)\| \\ &\geq \left(\min_{M \in \mathcal{M}} \frac{1}{\|M^{-T}\|}\right) \sup_{\{x : \|x\| \leq \min_{M \in \mathcal{M}}(\|M\|)\zeta(\lambda_r)\}} \|\nabla \psi(x)\|,\end{aligned}$$

where we used that $\tilde{\psi}_M(\xi) = M^T \nabla\psi(x)$, that $\|M^{-1}x\| \geq \|x\|/\|M\|$, and the compactness of $\mathcal{M}$. This is sufficient to eliminate any dependence on M in α.

- Depending on the specific factors in the statement of the lemma, the switching rate λ_r can be conveniently bounded either above by $\lambda_r^{\max}$ or below by $\lambda_r^{\min}$ in order to bound $\alpha(\lambda_r, M, K)$ from below. This can be done in every term in which λ_r appears and it follows that the dependence on it can be easily eliminated.

We have thus shown that we can choose $\alpha^* > 0$ such that in the same setting of (5.34), and for all $M \in \mathcal{M}$ and all $\lambda_r \in \Lambda_r$

$$\|P^t_{M,\lambda_r}((x,\theta),\cdot) - P^t_{M,\lambda_r}((\tilde{x},\tilde{\theta}),\cdot)\|_{\mathrm{TV}} \leq 2\,(1-\alpha^*)\,.$$

□

5.B.5.2 Drift condition for the BPS

Proof of Lemma 5.24. We follow the same underlying idea that was used in the proof of Lemma 5.22. It is in fact sufficient that there exist $A_1, A_2 > 0$, both independent of M and λ_r, such that for all M and λ_r there exists a function $\tilde{V}_{M,\lambda_r}(\xi,\theta)$ such that

$$\tilde{\mathcal{L}}_{M,\lambda_r}\tilde{V}_{M,\lambda_r}(\xi,\theta) \leq -A_1\tilde{V}_{M,\lambda_r}(\xi,\theta) + A_2 \qquad \text{for all } (\xi,\theta) \in \mathbb{R}^d \times \mathbb{R}^d, \quad (5.35)$$

where $\tilde{\mathcal{L}}_{M,\lambda_r}$ is the generator of a standard BPS with target $\tilde{\psi}_M(\xi) = \psi(M\xi)$ and refreshment rate λ_r. Indeed, we can then define $V_{M,\lambda_r}(x,\theta) = \tilde{V}_{M,\lambda_r}(M^{-1}x,\theta)$ to obtain by Proposition 2.3

$$\mathcal{L}_{M,\lambda_r}V_{M,\lambda_r}(x,\theta) = \tilde{\mathcal{L}}_{M,\lambda_r}\tilde{V}_{M,\lambda_r}(\xi,\theta) \leq -A_1 V_{M,\lambda_r}(x,\theta) + A_2 \quad (5.36)$$

for all $(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^d$. In order to show the inequality in (5.35) we rely on [64, Lemma 7]. In particular we begin by showing that the constants $c_1, c_2, c_3, c_4 > 0$, $R > 0$ defined in [64, Assumption A8] can be chosen independently of the specific M for the class of standard BP samplers with targets $\tilde{\pi}_M$. This implies that A_1, A_2 can then be chosen to be the same for every $M \in \mathcal{M}$. The dependence on λ_r is dealt with as a second step. For ease of notation from now on we denote the corresponding energy function as $\psi_M(\xi)$. Let us restrict our attention to the case $\nu = \mathcal{N}(0, \mathbb{1}_d)$ and thus, using the notation in [64], we take $\ell(\xi) = 1$, $H(\|\theta\|) = \eta\|\theta\|^2$ for some $\eta \in (0,1)$ small enough such that $\int_{\mathbb{R}^d} \exp(\eta\|\theta\|^2)\nu(\mathrm{d}\theta) < \infty$, and $\overline{\psi}_M(\xi) = \psi_M^{\zeta}(\xi)$ for $\zeta \in (0,1)$ chosen as in part (c) of Assumption 3.9. Observe that η, ζ are uniform in M.

We start by considering c_1, which must be such that $\|\nabla_\xi\overline{\psi}_M(\xi)\| \geq c_1$ for all points outside of a ball of radius R_1. In particular we wish to show that there exist c_1, R_1 that satisfy the property above for all $M \in \mathcal{M}$. Observing that for $x = M\xi$ it holds that $\nabla_\xi\overline{\psi}_M(\xi) = \zeta\psi_M^{\zeta-1}(\xi)\nabla_\xi\psi_M(\xi)$ and $\nabla_\xi\psi_M(\xi) = M^T\nabla_x\psi(x)$, hence we have that for any $M \in \mathcal{M}$

$$\|\nabla_\xi\overline{\psi}_M(\xi)\| = \zeta\frac{\|\nabla_\xi\psi_M(\xi)\|}{\psi_M^{1-\zeta}(\xi)} \geq \frac{\zeta}{\|M^{-T}\|}\frac{\|\nabla_x\psi(x)\|}{\psi^{1-\zeta}(x)} \geq C\frac{\|\nabla_x\psi(x)\|}{\psi^{1-\zeta}(x)}, \quad (5.37)$$

where $C = \zeta \min_{M\in\mathcal{M}}(1/\|M^{-T}\|) > 0$. Assumption 5.17(c) implies that there exist $\tilde{c}_1, \tilde{R}_1 > 0$ such that $\|\nabla_x\psi(x)\|/\psi^{1-\zeta}(x) \geq \tilde{c}_1$ for any x such that $\|x\| \geq \tilde{R}_1$, and because $\mathcal{M}$ is a compact space, we can choose $c_1 = C\,\tilde{c}_1$ and $R_1 = \tilde{R}_1$, which are then independent of M. The constant c_2 must be such that $\ell(\xi) \leq c_2$, and because for a Gaussian ν, we may take as written above $\ell(\xi) = 1$. Therefore $c_2 = 1$ for each $M \in \mathcal{M}$.

Then, c_3 must be such that $(\|\nabla_\xi\psi_M(\xi)\|/\|\nabla_\xi\overline{\psi}_M(\xi)\|) \geq c_3$ for all ξ such that $\xi > R_3$ for some $R_3 > 0$. Indeed for $x = M\xi$ we have

$$\frac{\|\nabla_\xi\psi_M(\xi)\|}{\|\nabla_\xi\overline{\psi}_M(\xi)\|} = \frac{\|\nabla_\xi\psi_M(\xi)\|}{\zeta\psi_M^{\zeta-1}(\xi)\|\nabla_\xi\psi_M(\xi)\|} = \frac{1}{\zeta}\psi_M^{1-\zeta}(\xi) = \frac{1}{\zeta}\psi^{1-\zeta}(x).$$

By part (b) of Assumption 5.17 we have that $\lim_{\|x\|\to\infty}\psi(x) = +\infty$ and since $\zeta \in (0,1)$ there exist c_3 and R_3 large enough such that outside of a ball of radius R_3 we have $(\|\nabla_\xi\psi_M(\xi)\|/\|\nabla_\xi\overline{\psi}_M(\xi)\|) \geq c_3$ for any $M \in \mathcal{M}$.

It is left to show that c_4 can be chosen uniform. For $A_{x,M} \coloneqq \{\theta \in \mathbb{R}^d : \eta\|\theta\|^2 \leq 3\overline{\psi}_M(M^{-1}x)\}$, c_4 must be such that

$$\|\nabla^2\overline{\psi}_M(\xi)\|\left(\sup_{\theta\in A_{x,M}}\|\theta\|^2\right) \leq c_4 \qquad \text{for any } \xi \text{ such that } \|\xi\| > R_4 \tag{5.38}$$

for some $R_4 > 0$. Observing that for $\theta \in A_{x,M}$ it holds that $\sup_{\theta\in A_{x,M}}\|\theta\|^2 \leq \frac{3}{\eta}\psi_M^\zeta(\xi)$, we obtain

$$\begin{aligned}
&\|\nabla^2\overline{\psi}_M(\xi)\|\left(\sup_{\theta\in A_{x,M}}\|\theta\|^2\right) \leq \\
&\leq \Big(\zeta(1-\zeta)\psi_M^{\zeta-2}(\xi)\|\nabla_\xi\psi_M(\xi)(\nabla_\xi\psi_M(\xi))^T\| + \zeta\psi_M^{\zeta-1}(\xi)\|\nabla_\xi^2\psi_M(\xi)\|\Big)\frac{3}{\eta}\psi_M^\zeta(\xi) \\
&\leq 3\frac{\zeta(1-\zeta)}{\eta}\left(\frac{\|\nabla_\xi\psi_M(\xi)\|}{\psi_M^{1-\zeta}(\xi)}\right)^2 + 3\frac{\zeta}{\eta}\frac{\|\nabla_\xi^2\psi_M(\xi)\|}{\psi_M^{1-2\zeta}(\xi)} \\
&\leq 3\frac{\zeta(1-\zeta)}{\eta}\|M^T\|^2\left(\frac{\|\nabla_x\psi(x)\|}{\psi^{1-\zeta}(x)}\right)^2 + 3\|M\|\|M^T\|\frac{\zeta}{\eta}\frac{\|\nabla_x^2\psi(x)\|}{\psi^{1-2\zeta}(x)} \\
&\leq 3\frac{\zeta(1-\zeta)}{\eta}D^2\left(\frac{\|\nabla_x\psi(x)\|}{\psi^{1-\zeta}(x)}\right)^2 + 3D^2\frac{\zeta}{\eta}\frac{\|\nabla_x^2\psi(x)\|}{\psi^{1-2\zeta}(x)}.
\end{aligned} \tag{5.39}$$

In the second to last inequality we used that $\nabla_\xi^2\psi_M(\xi) = M^T\nabla_x^2\psi(x)M$ with $x = M\xi$, and in the last inequality we defined $D = \max_{M\in\mathcal{M}}\{\|M\| \vee \|M^T\|\}$. Part (c) of Assumption 5.17 implies that there exist $\tilde{c}_4, \tilde{R}_4$ such that

$$\frac{\|\nabla_x\psi(x)\|}{\psi^{1-\zeta}(x)} \leq \tilde{c}_4, \quad \frac{\|\nabla_x^2\psi(x)\|}{\psi^{1-2\zeta}(x)} \leq \tilde{c}_4 \qquad \text{for all } x \text{ such that } \|x\| \geq \tilde{R}_4.$$

As a consequence, because $\mathcal{M}$ is a compact space and by (5.39), there exist c_4, R_4 independent of M such that (5.38) holds for all $M \in \mathcal{M}$. It is now sufficient to take $R = \max\{R_1, R_3, R_4\}$, and we have shown that c_1, c_2, c_3, c_4, R can be picked uniformly for all standard BP samplers with energy function ψ_M. In particular there is no dependence on λ_r in all these constants.

It is important to observe that part (c) of Assumption 5.17 is satisfied by all BP samplers with targets in $\{\tilde{\pi}_M\}_{M\in\mathcal{M}}$ because we have bounds on $\|M\|$ and $\|M^{-1}\|$. Moreover, parts (a) and (b) of Assumption 5.17 are trivially verified because of the transformation scheme in Figure 5.2 that defines $\tilde{\psi}_M$. It is then possible to apply [64, Lemma 7] to each process to obtain the drift condition (5.35). Therefore for each $M \in \mathcal{M}$ and $\lambda_r \in \Lambda_r$ we have the Lyapunov function

$$\tilde{V}_{M,\lambda_r}(\xi,\theta) = \exp\Big(\kappa_{\lambda_r}\psi_M^{\zeta}(\xi)\Big)\varphi_{\lambda_r}\left(\frac{2}{rc_1}\langle\theta, \nabla_\xi\overline{\psi}_M(\xi)\rangle\right) + \exp\big(\eta\|\theta\|^2\big), \tag{5.40}$$

where $\eta \in (0,1)$ and r depend only the distribution at refreshments ν, $\kappa_{\lambda_r} \in (0,1)$ depends on the c_i's, and φ_{λ_r} is a positive, non decreasing, continuously differentiable function as defined in [64]. Notice that, as a consequence of the calculations above, κ_{λ_r} and φ_{λ_r} do not depend on the preconditioner. Applying our usual transformation scheme we obtain the functions

$$V_{M,\lambda_r}(x,\theta) = \exp\big(\kappa_{\lambda_r}\psi^{\zeta}(x)\big)\varphi_{\lambda_r}\left(\frac{2}{rc_1}\langle\theta, M^T\nabla_x\overline{\psi}(x)\rangle\right) + \exp\big(\eta\|\theta\|^2\big). \tag{5.41}$$

This means that for each $\tilde{V}_M$ as defined in (5.40), the same proof of [64, Lemma 7] can be followed and thus $\tilde{V}_M$ satisfies

$$\tilde{\mathcal{L}}_{M,\lambda_r}\tilde{V}_{M,\lambda_r}(\xi,\theta) \le -A_1(\lambda_r)\tilde{V}_{M,\lambda_r}(\xi,\theta) + A_2(\lambda_r) \qquad \text{for all } (\xi,\theta) \in \mathbb{R}^d \times \mathbb{R}^d. \tag{5.42}$$

In particular, A_1, A_2 both depend on λ_r, on c_1, c_2, c_3, c_4, R, and on other constants that depend only on ν and thus are independent of M. Therefore A_1, A_2 can be chosen uniformly in $M \in \mathcal{M}$.

The last step is now to eliminate the dependence on λ_r in A_1, A_2. Observe that $A_1(\lambda_r)$ is defined in the proof of [64, Lemma 7] as a minimum of several constants. It is enough for our needs to notice that the constants have a continuous dependence in λ_r and that we are considering $\lambda_r \in [\lambda_{\min}, \lambda_{\max}]$. On the other hand, $A_2(\lambda_r)$ can be chosen independently of λ_r by continuity of V_{M,λ_r} in λ_r. We can therefore choose A_1, A_2 independently both of M and λ_r, thus the wanted simultaneous drift condition follows by (5.36). □

5.B.6 Other technical results

The lemma below is needed to go from simultaneous, continuous time drift conditions to simultaneous, discrete time ones.

Lemma 5.33. *Consider a family of Markov processes with state space E and generators $\{\mathcal{L}_\gamma : \gamma \in \mathcal{Y}\}$. Assume the following conditions are verified:*

a) there exist $c_1 > 0$, $c_2 > 0$, a set $K \subset E$ all independent of γ, and a class of functions $\{V_\gamma : E \to [1, +\infty) : \gamma \in \mathcal{Y}\}$, such that for each $\gamma \in \mathcal{Y}$ the following drift condition holds

$$\mathcal{L}_\gamma V_\gamma(z) \leq -c_1 V_\gamma(z) + c_2 \mathbb{1}_K(z) \qquad \text{for all } z \in E. \tag{5.43}$$

b) the family of processes is such that for $K \subset E$ as in the previous point, a constant $t > 0$, and all $z \in E$, there exists a set $\tilde{K}(t) \subset E$ which depends on t such that on the event $z \notin \tilde{K}(t)$ it holds that $Z(t) \notin K$ almost surely for all $\gamma \in \mathcal{Y}$.

Then for any $t > 0$ there are $\lambda \in (0, 1)$ and $\kappa > 0$, both independent of γ, such that for each $\gamma \in \mathcal{Y}$

$$P_\gamma^t V_\gamma(z) \leq \lambda V_\gamma(z) + \kappa \mathbb{1}_{\tilde{K}}(z) \qquad \text{for all } z \in E.$$

In particular, if conditions (5.43) hold with $V \equiv V_\gamma$ for each γ, then

$$P_\gamma^t V(z) \leq \lambda V(z) + \kappa \mathbb{1}_{\tilde{K}}(z) \qquad \text{for all } z \in E.$$

Proof. Let $t > 0$ and $\gamma \in \mathcal{Y}$. As a first step, we apply Dynkin's formula to $f(z, t) = \exp(c_1 t) V_\gamma(z)$ as in the proof of Theorem 6.1 in [105] to obtain

$$e^{c_1 t} P_\gamma^t V_\gamma(z) = V_\gamma(z) + \int_0^t P_\gamma^s \left(\frac{\partial}{\partial s} + \mathcal{L}_\gamma \right) (e^{c_1 s} V_\gamma(z)) \, \mathrm{d}s.$$

By the product rule and the drift condition we can write the integrand on the right hand side as

$$\begin{aligned} P_\gamma^s \left(\frac{\partial}{\partial s} + \mathcal{L}_\gamma \right) (e^{c_1 s} V_\gamma(z)) &= e^{c_1 s} P_\gamma^s \mathcal{L}_\gamma V_\gamma(z) + c_1 e^{c_1 s} P_\gamma^s V_\gamma(z) \\ &\leq e^{c_1 s} P_\gamma^s \left(-c_1 V_\gamma(z) + c_2 \mathbb{1}_K(z) \right) + c_1 e^{c_1 s} P_\gamma^s V_\gamma(z) \\ &= c_2 e^{c_1 s} P_\gamma^s \mathbb{1}_K(z). \end{aligned}$$

Therefore by linearity of the integral

$$\begin{aligned} P_\gamma^t V_\gamma(z) &\leq e^{-c_1 t} V_\gamma(z) + c_2 \int_0^t e^{c_1 (s-t)} P_\gamma^s \mathbb{1}_K(z) \mathrm{d}s \\ &\leq e^{-c_1 t} V_\gamma(z) + c_2 \mathbb{1}_{\tilde{K}}(z) \int_0^t e^{c_1 (s-t)} \mathrm{d}s \\ &= e^{-c_1 t} V_\gamma(z) + \frac{c_2}{c_1} (1 - e^{-c_1 t}) \mathbb{1}_{\tilde{K}}(z). \end{aligned}$$

In the second inequality we took advantage of condition (b) to conclude that $P_\gamma^s \mathbb{1}_K(z) = \mathbb{P}_z(Z(s) \in K) \leq \mathbb{1}_{\tilde{K}}(z)$. It is then sufficient to take $\lambda = e^{-c_1 t}$, $\kappa = \frac{c_2}{c_1}(1 - e^{-c_1 t})$ to conclude the proof. Observe that λ and κ do not depend on γ (but depend on t). □

Remark 5.34. Condition (b) in Lemma 5.33 is for instance satisfied by a family of preconditioned Zig-Zag processes, if the preconditioner is taken from a compact set of positive definite matrices. Indeed this claim follows from the fact that all processes travel with almost surely bounded velocity. Therefore, it is possible to choose $\tilde{K}(t)$ by adding a suitable buffer zone around the set K, such that the set K is not reachable in time t starting outside $\tilde{K}(t)$.

Remark 5.35. Instead of drift conditions of the form (5.43), consider the following case

$$\mathcal{L}_\gamma V_\gamma(z) \leq -c_1 V_\gamma(z) + c_2 \qquad \text{for all } z \in E.$$

Without condition (b) in Lemma 5.33 we can still conclude by the same line of reasoning that there are $\lambda \in (0,1)$, $\kappa > 0$, both independent of γ, such that for each $\gamma \in \mathcal{Y}$

$$P_\gamma^t V_\gamma(z) \leq \lambda V_\gamma(z) + \kappa \qquad \text{for all } z \in E.$$

Chapter 6

Time transformations of PDMPs

6.1 Introduction

Consider a Markov process $(X_t)_{t\geq 0}$ on a state space E with stationary distribution with density $\mu(x) = \exp(-\psi(x))/Z$. The law of large numbers gives that *ergodic averages* converge to expectations from the stationary distribution: almost surely

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T f(X_t)\mathrm{d}t = \int f(x)\mu(\mathrm{d}x) =: \mu(f),$$

for any f that is μ-integrable. Taking $f = \mathbb{1}_A$ for a measurable set A, we find that the fraction of time that the process spends in A converges to $\mu(A) := \int \mathbb{1}_A(x)\mu(\mathrm{d}x)$. As a consequence, the process X_t can only stay for very little time in regions to which μ assigns low probability. When μ is a multimodal probability distribution this means that it will take a long time, and a lot of computational effort in the simulation, before the process leaves the neighbourhood of the mode it is currently exploring, as this means going through (and thus spending time) in a region of low density. For this reason, Markov chain Monte Carlo (MCMC) algorithms need a large computational budget to obtain a representative sample from multimodal target distributions. In this chapter we study the following idea: *increasing the speed of time* for the process X_t when it is in low density regions, we obtain that visits to such regions can become more frequent, as the process spends less time would there. In the multimodal setting, this gives that hops between modes can happen more often, or similarly in the case of heavy tailed targets the tails are explored more regularly. In mathematical terms, the concept of increasing or decreasing the speed of time depending on the state of the process corresponds to applying a *random time transformation* to another Markov

process. In this chapter we shall show that a Markov process X_t with stationary distribution μ corresponds to applying a time change to a suitable process Y_t with stationary distribution $\mu_s(y) \propto s(y)\mu(y)$. Here s plays the role of the *speed function* for the time variable and defines the random time transformation $r(t) = \int_0^t s(X_u)\mathrm{d}u$ which gives $X_t = Y_{r(t)}$. This connection gives that the paths of the two processes coincide, while the state of each process at a given time differs through the time change. The intuition is that Y_t is a well studied process, such as a Langevin diffusion or a PDMP from the MCMC literature, while X_t is a new process which differs from Y_t because of a time change. Importantly, the convergence properties of Y_t can be improved by choosing suitable speed functions, e.g. the rate of convergence of X_t to μ can be geometric even if Y_t is only (polynomially) ergodic. This motivates the use of time transformations in the context of MCMC algorithms based on continuous time Markov processes, as for instance PDMPs. In the case of PDMPs, [153] introduced a modification of the ZZS with non-constant speed. However, the authors interpreted the speed as norm of the velocity in space rather than speed of time. In this chapter we establish that the speed up Zig-Zag (SUZZ) of [153] falls into our framework and is in fact a time transformation of a standard ZZS with stationary distribution μ_s. This finding gives clarity on the choice of the speed function in SUZZ, for which until now there was no clear guideline. Notably, we shall argue that the SUZZ is uniformly ergodic for suitable speed functions taking advantage of existing results for the standard ZZS [24].

Since our framework is more general and includes any Markov process, in this chapter we introduce novel PDMPs which are obtained as time transformations of known PDMPs such as BPS [32], the Boomerang sampler [25], and the randomised Hamiltonian Monte Carlo process [30]. More generally, we study the idea of time changes of Markov processes extensively with the goal of translating properties of Y_t to X_t.

Organisation of the chapter

In Section 6.2 we give some intuition on the key ideas of the paper considering the ZZS and the SUZZ. In Section 6.3 we study time transformations of Markov processes in full generality, proving various theoretical results on the key properties that are needed for MCMC purposes. In Section 6.4 we study time transformations of PDMPs, introducing several new processes and giving a sufficient condition for uniform ergodicity. We conclude the chapter with a discussion.

6.2 Reasoning on the Zig-Zag process

6.2.1 The standard Zig-Zag process for multimodal and heavy tailed distributions

Let us start with a practical, one dimensional example, which was studied for the ZZS in [110]. Consider a double well potential ψ with two local minima at x_0 and

x_2 and a local maximum at x_1, where $x_0 < x_1 < x_2$. We shall refer to this target as multimodal target (MT). We wish to compute the probability that starting at x_0 with velocity $v_0 = +1$ the process successfully overcomes the potential barrier at x_1 in one go. Denoting the first event time of the process as $\tau := \inf\{t > 0 : V_t = -1\}$, this means that $X_\tau \geq x_1$. Let us compute the probability of this event assuming the excess switching rate is zero, i.e. $\gamma = 0$. Since the process moves with speed 1, this event takes place if $\tau > x_1 - x_0$. In particular $\lambda(x + t, 1) = \psi'(x + t)$ for $t \leq \tau$ as the potential is increasing. Therefore

$$\begin{aligned}\mathbb{P}_{(x_0,+1)}(X_\tau \geq x_1) &= \mathbb{P}_{(x_0,+1)}(\tau \geq x_1 - x_0) \\ &= \exp\left(-\int_0^{x_1-x_0} \psi'(x+t)\mathrm{d}t\right) \\ &= \exp(-(\psi(x_1) - \psi(x_0))).\end{aligned} \tag{6.1}$$

Thus this probability is exponentially decreasing in the difference of potential between x_1 and x_0, so the process will cross the low density region only sporadically. Intuitively, we can address this problem allowing the process to increase its speed when it is in the low density region. In this case, the process would spend less time in such region and therefore the crossing from one well to the other could happen more frequently.

Another case we shall consider is that of a d-dimensional unimodal target (UT). In this case we assume ψ is a differentiable function with global minimum at x_0, and moreover that the switching rates are strictly positive along dynamics $x_0 + vt$ for all $v \in \{\pm 1\}$ and $t > 0$, as is typically the case for isotropic targets. What we wish to study is the likelihood of exploring the tails of such target. This is of particular interest in the heavy tailed case, which was the main motivation for the SUZZ sampler [153]. Let us start the process is started at x_0 with velocity $v \in \{\pm 1\}$. Similarly to above, we now compute the probability that the ZZS reaches a certain distance to the centre x_0 before the first event, again denote by τ, takes place. For any $c > 0$, the process is at distance greater than c with probability

$$\mathbb{P}_{(x_0,v)}(|X_\tau - x_0| \geq c) = \mathbb{P}_{(x_0,v)}(\tau \geq c/\sqrt{d}) = \exp\left(-(\psi(x_0 + vc/\sqrt{d}) - \psi(x_0))\right). \tag{6.2}$$

Similarly to the MT case, we have a probability that is exponentially decreasing in the difference of potential. However, we would like the process to explore the tails as rapidly as possible instead of being stuck in the centre of the distribution for a large time. In particular, bad exploration of the tails can imply that the process is not geometrically ergodic, as is the case for ZZS with heavy tailed target [153].

6.2.2 The speed-up ZZS

The speed-up ZZS of [153] gives a modification of the ZZS which incorporates a speed function $s : \mathbb{R}^d \to \mathbb{R}_+$. Thus, the deterministic dynamics are given by the system of

ODEs

$$\begin{aligned} \mathrm{d}X_t &= s(X_t)V_t\mathrm{d}t, \\ \mathrm{d}V_t &= 0, \end{aligned} \tag{6.3}$$

for some initial conditions $(x, v) \in \mathbb{R}^d \times \{-1, +1\}^d$. Naturally the choice $s(x) = 1$ recovers the standard ZZS, where $x_t = x + vt$. The ODE (6.3) has solutions of the form $x_t = x + A(t)v_0$ for $A(t) = \int_0^t s(X_u)du$, as indeed the function s acts indeed as *speed* of the process and does not modify the direction. In the case of SUZZ, the switching rates that leave μ invariant are of the form

$$\begin{aligned} \lambda_i(x, v) &= (v_i(s(x)\partial_i\psi(x) - \partial_i s(x)))_+ \\ &= s(x)\left(v_i\left(\partial_i\psi(x) - \frac{\partial_i s(x)}{s(x)}\right)\right)_+, \end{aligned} \tag{6.4}$$

for $i = 1, \ldots, d$. In order to ensure that the process is non-explosive the speed function should satisfy the following condition, which excludes choices as $s(x) = \exp(a\psi(x))$ for $a \geq 1$.

Assumption 6.1. *It holds that* $\lim_{\|x\|\to\infty} \exp(-\psi(x))s(x) = 0$.

For a given speed function s, we now want to compare the SUZZ to ZZS in the MT and UT settings. Starting with the MT, we now compute the probability of overcoming the low density region that separates the two modes. We shall assume for the moment that s is such that the rate (6.4) is strictly positive in points (x, v) for which the switching rate of ZZS is strictly positive, that is

$$(v(s(x)\psi'(x) - s'(x)))_+ > 0 \iff (v\psi'(x))_+ > 0. \tag{6.5}$$

Denote as $t(x_1)$ the time that it takes the process to get from x_0 to x_1. We find

$$\begin{aligned} \mathbb{P}_{(x_0,+1)}(X_\tau \geq x_1) &= \exp\left(-\int_0^{t(x_1)} \lambda(x_u, +1)\mathrm{d}u\right) \\ &= \exp\left(-\int_0^{t(x_1)} ((s(x_u)\psi'(x_u) - s'(x_u)))_+\mathrm{d}u\right). \end{aligned}$$

Using the change of variables $\mathrm{d}x_t = s(x_t)\mathrm{d}t$, which comes from (6.3), it follows

$$\begin{aligned} \mathbb{P}_{(x_0,+1)}(X_\tau \geq x_1) &= \exp\left(-\int_{x_0}^{x_1} \psi'(x_s)\mathrm{d}x_u + \int_{x_0}^{x_1} \frac{s'(x_u)}{s(x_u)}\mathrm{d}x_u\right) \\ &= \frac{s(x_1)}{s(x_0)}\exp(-(\psi(x_1) - \psi(x_0))). \end{aligned} \tag{6.6}$$

Therefore we have obtained an expression that is the same as (6.1) apart from the factor $s(x_1)/s(x_0)$. In particular it is clear that this factor increases the probability of

crossing the low density region if $s(x_1) > s(x_0)$, while it decreases it if $s(x_1) < s(x_0)$. This respects the intuition that was given above: if the process speeds up in the low density region, then it spends less time there and thus it can visits such region more often without violating the stationarity with respect to π.

Let us now consider UT. With similar computations and by the gradient theorem we find

$$\mathbb{P}_{x_0,v}(|X_\tau - x_0| \geq c) = \frac{s(x_0 + vc/\sqrt{d})}{s(x_0)} \exp(-(\psi(x_0 + vc/\sqrt{d}) - \psi(x_0))). \tag{6.7}$$

Thus the process visits the tails if we choose a speed function that increases as the process leaves the origin.

A fundamental observation is that the probabilities (6.6) and (6.7) are identical to those of a ZZP with target distribution with potential $V(\cdot) = \psi(\cdot) - \ln s(\cdot)$. This suggests a deeper relation between the standard ZZS and the SUZZS, and one of the main results of this chapter is the characterisations of the connection between these two processes. In the next proposition, which is a corollary of the results that we shall obtain in the next sections, we prove that the SUZZS with invariant distribution μ is a time transformation of a standard ZZS with invariant distribution with density

$$\mu_s(x, v) = \frac{1}{Z_s} \exp(-V(x)) \frac{1}{2^d},$$

where $Z_s = \int \exp(-\psi(x)) s(x) dx$ is the normalisation constant.

Proposition 6.2. *Let $s : \mathbb{R}^d \to [\underline{s}, \infty)$ for some $\underline{s} > 0$. Let ψ and s satisfy Assumption 6.1. Let $(\Xi_t, \Theta_t)_{t \geq 0}$ be a standard Zig-Zag process with invariant measure μ_s, which is assumed to satisfy $Z_s < \infty$. Suppose the ZZP is non-explosive. Consider the SUZZ with speed s and invariant distribution μ and denote it as $(X_t, V_t)_{t \geq 0}$. Define the random functions*

$$r(t) = \int_0^t s(X_u) \mathrm{d}u, \qquad r^{-1}(t) = \int_0^t \frac{1}{s(Y_u)} \mathrm{d}u.$$

Then for all $t \geq 0$ it holds that

$$(X_t, V_t) = (\Xi_{r(t)}, \Theta_{r(t)}), \quad (\Xi_t, \Theta_t) = (X_{r^{-1}(t)}, V_{r^{-1}(t)}).$$

Remark 6.3. This result is different than what shown in [153], where the authors find a *space* transformation that connects the SUZZ with a particular Zig-Zag process that lives on a modified domain. In particular in [153] the relation holds only in the one dimensional case, while Proposition 6.2 is true in any dimension.

Remark 6.4. Observe that the skeleton chains of the SUZZ and of the corresponding ZZS coincide apart from the time of the random events, which are of course subject to the time change. As a consequence, one is free to simulate the event times with the simplest rate between those of ZZS and SUZZ.

The statement of the proposition provides a clear picture on how the choice of speed function affects the trajectories of the SUZZ. Indeed, the SUZZ follows the path of a standard ZZS with target μ_s and therefore Proposition 6.2 clarifies the effect of the choice of speed functions on the process. In particular, the intuitive choices we discussed previously, which assign larger speed in low density regions, correspond to a SUZZ which follows paths of a ZZS with heavier tails, or in general smaller differences in potential between minima and maxima in the multimodal setting. On the other hand, choices of s which go in the opposite direction discourage the process from entering low density regions, thus deteriorating the mixing of the process. Proposition 6.2 also confirms the intuition on Assumption 6.1 as discussed in [153]. If the assumption is not satisfied, the corresponding standard ZZS targets a potential that is null, or worse decreasing in the tails, and thus the Zig-Zag process will not switch once it reaches the tails.

6.2.2.1 Choices of speed function

Let us consider three examples of speed functions, which we shall compare also in the low temperature regime, that is when the potential is $\psi_\varepsilon = \psi/\varepsilon$ and ε is small. This is the interesting scenario in which the target becomes increasingly concentrated around its modes and traveling between two modes becomes thus more difficult.

Example 6.5. *Let $s(x) = \exp(a\psi(x))$ for some $a \in (0,1)$. Note that the choice $a = 1$ is not allowed by Assumption 6.1, while $a = 0$ would give the standard ZZS. Observe that this choice satisfies the assumption given in* (6.5)*, as the switching rates of SUZZ are of the form*

$$\lambda_i(x,v) = (1-a)\, s(x)\, (v_i \partial_i \psi(x))_+. \tag{6.8}$$

This choice of speed function corresponds to the approach of tempering*: the SUZZ is a time change of a ZZS with tempered target $\mu_s(x,v) \propto \exp(-(1-a)\psi(x))$. We also observe that rates of the corresponding standard ZZS given by Proposition 6.2 are $\lambda_i(x,v) = (1-a)(v_i\partial_i\psi(x))_+$, that is are weighted by $(1-a)$. This approach is particularly interesting in the multimodal setting, and indeed in the MT case temperature parameter ε we find*

$$\mathbb{P}_{(x_0,+1)}(X^\varepsilon_\tau \geq x_1) = \exp\left((1-a)\,\frac{\psi(x_1)-\psi(x_0)}{\varepsilon}\right).$$

Therefore as $\varepsilon \to 0$ the probability of travelling to the mode in x_2 is still exponentially decreasing in the potential difference, but the order of the exponential is now $(1-a)$ instead of 1. A similar result is obtained in the UT setting.

Example 6.6. *Another choice of interest is $s(x) = 1 + \psi(x)^p$ for some $p \geq 1$, assuming wlog $\psi(x) \geq 0$. When $p = 1$ the switching rates of SUZZ are $\lambda_i(x,v) = \psi(x)(v_i\partial_i\psi(x))_+$. In the MT setting with temperature ε we find*

$$\mathbb{P}_{(x_0,+1)}(X_\tau \geq x_1) = \frac{\varepsilon + \psi(x_1)}{\varepsilon + \psi(x_0)} \exp(-(\psi(x_1)-\psi(x_0))/\varepsilon)$$

For fixed ε this speed function improves the chances of reaching the other mode as $\psi(x_1) > \psi(x_0)$, while as $\varepsilon \to 0$ the effect becomes less relevant and does not improve the order of convergence to 0 of the probability. In this sense the choice of Example (6.5) seems preferable.

Example 6.7. *Let us consider a speed function that does not depend on the potential: $s(x) = 1 + |x - y_0|^2$ for some $y_0 \in \mathbb{R}^d$. For simplicity assume that the potentials of MT and UT satisfy (6.5). In the MT setting we find*

$$\mathbb{P}_{(x_0,+1)}(X_\tau \geq x_1) = \frac{1 + |x_1 - y_0|^2}{1 + |x_0 - y_0|^2} \exp(-(\psi(x_1) - \psi(x_0))).$$

It is clear that for most choices of y_0 this speed function decreases the chance of reaching the mode at x_2 starting from x_0. If $y_0 = x_0$, then such an event has higher probability than for ZZS, but then the opposite direction, i.e. x_2 to x_0, becomes less likely. Thus this choice of speed function will lead to poor exploration of the state space. In the d-dimensional UT case we find

$$\mathbb{P}_{(x_0,v)}(|X_\tau - x_0| \geq c) = \frac{1 + |x_0 - y_0 + vc/\sqrt{d}|^2}{1 + |x_0 - y_0|^2} \exp(-(\psi(x_0 + vc/\sqrt{d}) - \psi(x_0))),$$

which means the process explores the tails more often when $y_0 \approx x_0$. However, bad choices of y_0 can be harmful and give higher mixing time.

6.2.2.2 Eyring-Kramers formula for SUZZ

An informative result in the MT setting is the Eyring-Kramers (EK) formula, which gives the average time the process takes to overcome the energy barrier separating the two modes. The EK formula was derived for the one-dimensional ZZP in [110, Theorem 1.1] and here we extend this result to the SUZZ taking advantage of Proposition 6.2. In this context, one considers a tempered target $\psi_\varepsilon = \psi/\varepsilon$ and the interest is in the behaviour of the process as $\varepsilon \to 0$, which corresponds to the probability density being increasingly concentrated around the modes. We consider speed functions s_ε which satisfy the condition (6.5), as this ensures that the target of the standard ZZS connected to the SUZZ by Proposition 6.2 has again a double well potential.

Similarly to [110, Theorem 1.1], we initialise the SUZZ at $(x_0, -1)$ and we want to compute the expected value of

$$\tau := \inf\{t > 0 : X_t = x_1\}.$$

Proposition 6.8. *Consider a smooth one-dimensional double well potential $\psi > 0$ with $\psi''(x_0) > 0$ and let $s : \mathbb{R} \to [1, \infty)$ be a smooth function. Suppose Assumption 6.1 and condition (6.5) hold. Define $\psi_s(x) = \psi(x) - \ln s(x)$. Then the SUZZ with target $\mu_\varepsilon \propto \exp(-\psi/\varepsilon)$ and speed function $s_\varepsilon(x) = s(x)^{1/\varepsilon}$ satisfies*

$$\mathbb{E}_{(x_0,-1)}[\,\tau\,] \leq \sqrt{\frac{8\pi\varepsilon}{\psi_s''(x_0)}} \exp\left(\frac{\psi_s(x_1) - \psi_s(x_0)}{\varepsilon}\right)(1 + o_{\varepsilon\to 0}(1)).$$

Remark 6.9. A choice of speed function that is naturally of the form $s_\varepsilon(x) = s(x)^{1/\varepsilon}$ for the target ψ/ε is that of Example 6.5, that is $s(x) = \exp(a\psi(x))$ for $a \in (0,1)$. In this case, the statement of Theorem 6.8 gives

$$\mathbb{E}[\tau] = \sqrt{\frac{8\pi\varepsilon}{(1-a)\psi''(x_0)}} \exp\left((1-a)\frac{\psi(x_1)-\psi(x_0)}{\varepsilon}\right)(1+o_{\varepsilon\to 0}(1)).$$

This should be compared to the standard ZZS as considered in [110, Theorem 1.1], which corresponds to the choice $a = 0$. As ε goes to 0, we see that the hitting time of x_1 of SUZZ increases with a lower rate as that of ZZS. This gives further motivation to the speed function $s(x) = \exp(a\psi(x))$.

Proof of Theorem 6.8. By Proposition 6.2 we find that

$$\mathbb{E}_{(x_0,-1)}[\,\tau\,] = \mathbb{E}_{(x_0,-1)}[\,r^{-1}(\tau_{ZZ})\,],$$

where $\tau_{ZZ} = r(\tau)$ is the time at which the standard ZZS with potential $(\psi - \ln s)/\varepsilon$ reaches x_1. Because $s \geq 1$, we find by Proposition 6.2 that

$$\mathbb{E}_{(x_0,-1)}[\,\tau\,] \leq \mathbb{E}_{(x_0,-1)}[\,\tau_{ZZ}\,].$$

Under our assumptions we have that $\psi - \ln s$ is again a double well potential. Hence we can apply [110, Theorem 1.1] to the standard ZZS with potential $\psi - \ln s$ to obtain the statement of the theorem. □

6.3 Time transformations of Markov processes

In this section we develop a theory of time changed Markov processes, with emphasis on the properties that are relevant for MCMC algorithms. We lay the foundations of this framework taking advantage of the description of Section 1 of Chapter 6 in [68], then in Sections 6.3.1, 6.3.2, and 6.3.3 we give new results connecting properties of the original and time changed processes.

Let $(Y_t)_{t\geq 0}$ be a time homogeneous Markov process with state space E, which is the subset of a finite dimensional vector space. In this section, Y_t should be thought of as a well studied process, such as the overdamped Langevin diffusion or the Zig-Zag process. For $\underline{s} > 0$ introduce a continuously differentiable function $s : E \to [\underline{s}, \infty)$ which we assume is such that $s(Y_t)$ is almost surely bounded on bounded time intervals. This holds for instance if Y_t is non-explosive in the sense of [107] (see Section 6.3.2.1 for a definition). We introduce the stochastic process that satisfies the relation

$$X_t = Y_{r(t)}, \tag{6.9}$$

where

$$r(t) := \int_0^t s(X_u)\mathrm{d}u = \int_0^t s(Y_{r(u)})\mathrm{d}u. \tag{6.10}$$

Because s is lower bounded by a constant, with a change of variables we find

$$t = \int_0^t \frac{s(Y_u)}{s(Y_u)}\mathrm{d}u = \int_0^{r(t)} \frac{1}{s(Y_u)}\mathrm{d}u,$$

which gives the inverse time transformation

$$Y_t = X_{r^{-1}(t)} \quad \text{for} \quad r^{-1}(t) := \int_0^t \frac{1}{s(Y_u)}\mathrm{d}u. \tag{6.11}$$

Observe that by [68, Chapter 6, Theorem 1.1] the process X_t is well defined as

$$\inf\left\{t > 0 : \int_0^t \frac{1}{s(Y_u)}\mathrm{d}u = \infty\right\} = \infty,$$

which is a consequence of the assumption that s is lower bounded.

The last result from [68] that we need is a connection between the generators of X_t and Y_t. Recall from [68] that the extended generator of Y_t is the operator $\mathcal{L}$ that makes $f(Y_t) - \int_0^t \mathcal{L}f(Y_u)du$ a martingale with respect to the filtration generated by $(Y_t)_{t\geq 0}$. The set of functions for which this holds forms the domain of the generator, denoted as $\mathcal{D}(\mathcal{L})$. As a consequence, this is called martingale problem for $(\mathcal{L}, \mathcal{D}(\mathcal{L}))$. We shall assume throughout that the martingale problem for Y_t is well posed, that is it admits a unique solution. The following theorem connects the generators of $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$, which for the rest of the section denote the two Markov processes connected by (6.9).

Theorem 6.10 (Theorem 1.3, Chapter 6 in [68])**.** *Suppose $(Y_t)_{t\geq 0}$ is a Markov process with generator $(\mathcal{L}, \mathcal{D}(\mathcal{L}))$. Then, under the assumptions of this section, the process $(X_t)_{t\geq 0}$ satisfying* (6.9) *solves the martingale problem for $(s\mathcal{L}, \mathcal{D}(\mathcal{L}))$.*

This result is extremely useful to connect properties of the two processes, which are often obtained through conditions on the generator.

6.3.1 Stationarity, irreducibility, petite sets

We can easily relate the stationary measures of X_t and Y_t.

Proposition 6.11. *Let μ be a probability distribution on E. Assume $Z_s := \int_E s(y)\mu(\mathrm{d}y) < \infty$ and let μ_s be a probability distribution such that $\mu_s(A) = \int_A s(y)\mu(\mathrm{d}y)/Z_s$ for any measurable set A. Then μ_s is stationary for Y_t if and only if μ is stationary for X_t.*

Proof. Assume μ_s is stationary for Y_t. Then for all $f \in \mathcal{D}(\mathcal{L})$ it holds that $\int \mathcal{L}f\mathrm{d}\mu_s = 0$ (see e.g. [96, Theorem 3.37]). This implies that for all $f \in \mathcal{D}(\mathcal{L})$ we have $\int s\mathcal{L}f d\mu = 0$, which by Theorem 6.10 shows that μ is stationary for X_t. The reverse statement is identical. □

Let us consider the concept of *irreducibility*, which is important to obtain convergence of the law of the process to its stationary distribution. The process Y_t is irreducible if, for some measure φ, for any measurable set B such that $\varphi(B) > 0$ it holds $\mathbb{E}_y[\int_0^\infty \mathbb{1}_{Y_t \in B}\,\mathrm{d}t] > 0$, that is the process spends positive time in such sets. In the next proposition we show that irreducibility of X_t follows from irreducibility of Y_t.

Proposition 6.12. *Suppose $s(x) \geq \underline{s} > 0$ for any x and consider two processes Y_t and X_t related by (6.9). Then, X_t is irreducible if and only if Y_t is irreducible.*

Proof. Assume Y_t is ψ-irreducible and recall the relation (6.11). Then for any B such that $\psi(B) > 0$ we have by a change of variables

$$\mathbb{E}_y\left[\int_0^\infty \mathbb{1}_{Y_t \in B}\,\mathrm{d}t\right] = \mathbb{E}_y\left[\int_0^\infty \mathbb{1}_{X_{r^{-1}(t)} \in B}\mathrm{d}t\right] = \mathbb{E}_y\left[\int_0^\infty \mathbb{1}_{X_t \in B}\, s(X_t)\mathrm{d}t\right] > 0.$$

Because $s \geq \underline{s} > 0$ we have that the last inequality holds iff $\mathbb{E}_y\left[\int_0^\infty \mathbb{1}_{X_t \in B}\mathrm{d}t\right] > 0$, which shows X_t is ψ-irreducible. The reverse statement can be obtained observing that X_t is irreducible iff $\mathbb{E}_y\left[\int_0^\infty \mathbb{1}_{X_t \in B}\, s(X_t)\mathrm{d}t\right] > 0$ and then a change of variables gives irreducibility of Y_t. □

Let us now focus on the notion of *small set*, which is useful to establish ergodicity of a Markov process. A set C is (t_0, b, ν)-small for a process Y_t if there exist a constant $b > 0$ and a non-trivial probability measure ν on E such that for all $y \in C$ it holds that

$$P_{t_0}(y, \cdot) = \mathbb{P}_y(Y_{t_0} \in \cdot\,) \geq b\,\nu(\cdot).$$

A related, weaker condition is that of *petite set*: there exists a probability measure a on $(0, \infty)$ such that for all $y \in C$

$$\int a(\mathrm{d}t) P_t(y, \cdot) \geq b\,\nu(\cdot).$$

We are interested in proving that under suitable conditions a small or petite set for Y_t can be petite for the process X_t defined in (6.9). For this to hold, we require the following condition.

Assumption 6.13. *Consider sets C and $D \supset C$. There exist $t_0 > 0$, $b > 0$, $\varepsilon > 0$, a non-trivial measure ν such that for all $y \in C$ it holds*

$$\int_{t_0}^{t_0+\varepsilon} \mathbb{P}_y(Y_t \in A, Y_u \in D \text{ for all } u \in [0,t])\mathrm{d}t \geq b\nu(A) \qquad \text{for any } A. \tag{6.12}$$

In particular, ν and b are independent of t.

Equation (6.12) is satisfied e.g. when a small set condition holds uniformly in $t \in [t_0, t_0 + \varepsilon]$ combined with the event that the process does not leave D. The following result shows that Assumption 6.13 is sufficient to conclude that C is petite for the time transformed process X_t. The essential idea is that on the event in which Y_t does not leave the set D it is possible to bound $s(Y_t)$ and hence also the time transformation $r(t)$.

Lemma 6.14. *Suppose the process Y_t satisfies Assumption 6.13, and that s is lower bounded by $\underline{s} > 0$ and upper bounded on D by $\overline{s}_D < \infty$. Then the set C as in Assumption 6.13 is petite for the process X_t.*

Proof. Let a be the uniform distribution on the interval $[t_0/\overline{s}_D, (t_0+\varepsilon)/\underline{s}]$ and let $Z_a = (t_0+\varepsilon)/\underline{s} - t_0/\overline{s}_D$ denote its normalisation constant. For any $x \in C$ and for any measurable set A it holds that

$$\begin{aligned}\int a(t)P_t(x,A)\mathrm{d}t &\geq \frac{1}{Z_a}\int_{t_0/\overline{s}_D}^{(t_0+\varepsilon)/\underline{s}} \mathbb{P}_x(X_t \in A, X_u \in D \text{ for all } u \in [0,t])\mathrm{d}t \\ &= \frac{1}{Z_a}\mathbb{E}_x\left[\int_{t_0/\overline{s}_D}^{(t_0+\varepsilon)/\underline{s}} \mathbb{1}_{Y_{r(t)}\in A, Y_u \in D \text{ for all } u\in[0,r(t)]}\mathrm{d}t\right].\end{aligned}$$

In the last equation we applied Fubini's theorem. Now applying the time transformation $t' = r(t)$ we find

$$\begin{aligned}\int a(t)P_t(x,A)\mathrm{d}t &\geq \frac{1}{Z_a}\mathbb{E}_x\left[\int_{r(t_0/\overline{s}_D)}^{r((t_0+\varepsilon)/\underline{s})} \mathbb{1}_{Y_t\in A, Y_u\in D \text{ for all } u\in[0,t]}\frac{1}{s(Y_t)}\mathrm{d}t\right] \\ &\geq \frac{1}{Z_a\overline{s}_D}\mathbb{E}_x\left[\int_{r(t_0/\overline{s}_D)}^{r((t_0+\varepsilon)/\underline{s})} \mathbb{1}_{Y_t\in A, Y_u\in D \text{ for all } u\in[0,t]}\mathrm{d}t\right] \\ &\geq \frac{1}{Z_a\overline{s}_D}\mathbb{E}_x\left[\int_{t_0}^{t_0+\varepsilon} \mathbb{1}_{Y_t\in A, Y_u\in D \text{ for all } u\in[0,t]}\mathrm{d}t\right],\end{aligned}$$

where the last inequality is a consequence of the fact that on the event that Y_t stays in D it holds that $r(t_0/\overline{s}_D) \leq t_0$ and $r((t_0+\varepsilon)/\underline{s}) \geq (t_0+\varepsilon)$ almost surely, and hence we are effectively restricting the domain of integration. Finally, applying Assumption 6.13 we find

$$\begin{aligned}\int a(t)P_t(x,A)\mathrm{d}t &\geq \frac{1}{Z_a\overline{s}_D}\int_{t_0}^{t_0+\varepsilon} \mathbb{P}_x(Y_t \in A, Y_u \in D \text{ for all } u \in [0,t])\mathrm{d}t \\ &\geq \frac{b}{Z_a\overline{s}_D}\nu(A),\end{aligned}$$

which concludes the proof. □

Importantly, Assumption 6.13 is verified for the Zig-Zag process with strictly positive refreshment rate, for which we showed that compact sets satisfy a quantitative small set condition in Lemma 5.20.

Corollary 6.15. *Let (Y_t, W_t) be a ZZS with switching rates $\lambda_i(y,w) = (w_i\partial_i\psi(y))_+ + \gamma_i(x)$ for $i = 1,\dots,d$, where $0 < \underline{\gamma} \le \gamma_i(x) \le \overline{\gamma} < \infty$. Suppose ψ is continuously differentiable. Let (X_t, V_t) be the SUZZ defined by (6.9) with speed function s which is continuously differentiable and lower bounded by $\underline{s} > 0$. Then all sets of the form $C \times V$ for a compact set C and $V \subset \{\pm 1\}^d$ are petite for (X_t, V_t).*

Proof. By Lemma 5.20 any set of the form $C \times V$ for a compact set C and $V \subset \{\pm 1\}^d$ is $(t, b(t), \nu)$-small for the ZZS for any $t \ge t_0$, where t_0 is a large enough time and $b(t)$ is continuous in t. Now let $\varepsilon > 0$. Notice that the ZZS moves with constant velocity and therefore by time $t_0 + \varepsilon$ for any initial condition $(x,v) \in C$ the process cannot exit the set $D := \{(y,w) : \min_{x\in C}|x-y| \le \sqrt{d}(t_0+\varepsilon), w \in \{\pm 1\}\}$ by time $t_0 + \varepsilon$. It follows that Assumption 6.13 holds. Hence any $C \times V$ is petite for (X_t, V_t) by Lemma 6.14. □

6.3.2 Properties based on drift conditions

In this section we focus on properties that follow from drift conditions, such as non-explosivity and (polynomial, geometric, or uniform) ergodicity. The general drift condition we shall rely on is the following.

Assumption 6.16. *Consider the process Y_t with generator $(\mathcal{L}, \mathcal{D}(\mathcal{L}))$. There exists a function $V : \mathbb{R}^d \to [1,\infty]$ which is in the domain of the generator and satisfies the inequality*

$$\mathcal{L}V(z) \le -W(z)V(z) + m\mathbb{1}_C(z) \tag{6.13}$$

for some function $W > 0$, a constant $m \ge 0$, and a petite set C.

By Theorem 6.10, in this case the time transformed process has generator $\mathcal{L}_s = s\mathcal{L}$ and satisfies a drift condition

$$\mathcal{L}_s V(z) \le -s(z)W(z)V(z) + m\overline{s}_C\mathbb{1}_C(z), \tag{6.14}$$

where $\overline{s}_C := \max_{z\in C} s(z)$ and we are assuming s is bounded on C, which holds in our framework for instance if C is a compact set. As we shall discuss, the different properties we are after correspond to a specific form of $W(z)$, as proved in the paper by Meyn and Tweedie [107].

6.3.2.1 Non-explosivity, ergodicity

The first properties we focus on are non-explosivity and ergodicity. Let $(D_n)_{n\ge 0}$ be a family of open precompact sets such that $D_n \uparrow E$. Denote as τ_n the exit time of the process Y_t from the set D_n. Then Y_t is said to be *non-explosive* if $\lim_{n\to\infty}\tau_n = \infty$ almost surely. On the other hand, the process Y_t is said to be *ergodic* if $\lim_{t\to\infty}\|P_t(x,\cdot) - \mu(\cdot)\|_{TV} = 0$ for any initial condition $x \in E$. A drift condition that implies both non-explosivity and ergodicity of Y_t is given in the next assumption.

Assumption 6.17. *Y_t is irreducible, has stationary probability distribution $\mu_s \propto s\mu$, and has a petite set C, where C is compact. Moreover, it holds that*

$$\mathcal{L}V(y) \leq -cf(y) + m\mathbb{1}_C(y),$$

where V should be bounded on C and norm-like*, i.e. $V(y) \to \infty$ as $\|y\| \to \infty$, $f \geq 1$, and $c, m > 0$.*

The drift condition above corresponds to choosing $W(y) = cf(y)/V(y)$ in (6.13). Under Assumption 6.17, [107, Theorem 3.1] gives that Y_t is non-explosive, while to obtain ergodicity it is sufficient to apply [107, Theorem 4.2] and [106, Theorem 6.1]. Then we have the following result relating these properties to the time transformed process X_t.

Proposition 6.18. *Suppose Y_t satisfies Assumption 6.17. Suppose the set C is petite for the process X_t obtained by* (6.9)*. Assume moreover s is $\mathcal{C}^1$, lower bounded by a positive constant, and upper bounded on C. Then X_t is non-explosive and ergodic with respect to the measure μ.*

Proof. By Propositions 6.11 and 6.12 X_t is μ-stationary and irreducible. Then the results follow from the observing that by Theorem 6.10 X_t satisfies the drift condition in Assumption 6.17. Indeed X_t satisfies the drift condition

$$\mathcal{L}_s V(x) \leq -cs(x)f(x) + m\left(\max_{z\in C} s(z)\right)\mathbb{1}_C(x),$$

which taking advantage of $s \geq \underline{s}$ gives

$$\mathcal{L}_s V(x) \leq -c\underline{s}\tilde{f}(x) + m\left(\max_{z\in C} s(z)\right)\mathbb{1}_C(x),$$

for $\tilde{f}(x) = f(x)s(x)/\underline{s} \geq 1$. □

Remark 6.19. The assumption that V is norm-like is used to conclude that X_t is non-explosive, but it is not necessary to obtain ergodicity. In fact, this assumption can be stronger than needed, as non-explosivity is ensured by the condition $\mathcal{L}V \leq cV$ for some positive constant c and norm-like V. Therefore, with the assumption that X_t is non-explosive we obtain ergodicity of such process when Assumption 6.17 holds without requiring that V is norm-like.

6.3.2.2 Polynomial, geometric, and uniform ergodicity

We are now interested in the rate of convergence of the law of the process to its stationary distribution. Drift conditions as that in Assumption 6.16 are a fundamental technique to prove this type of result, see for instance [107, 57, 105]. In particular, the specific form of W in (6.13) determines the type of convergence. If $W(x) \geq m > 0$,

[57] shows that the convergence is exponential, that is there exist $\rho \in (0,1)$ and $B < \infty$ such that

$$\|P_t(x,\cdot) - \mu(\cdot)\|_V \leq BV(x)\rho^t,$$

where $\|\mu\|_V := \sup_{g \leq V} |\mu(g)|$. In particular, if V is upper bounded by a constant $\overline{V} < \infty$ the convergence is *uniform* in the initial condition, as the rhs is bounded by $B\overline{V}\rho^t$, which is independent of x. This is of course a desirable property as it signifies that the initial condition of the process does not affect the speed of convergence to the stationary distribution. Typically, Lyapunov functions are not bounded, as a bounded V implies that for any initial condition x the expected return time of the process to the small set C is bounded by a constant which does not depend on x (see [57, Theorem 6.1]). Therefore, this is hardly ever satisfied for typical Markov processes.

Alternatively, if (6.13) holds for $W = V^{-\alpha}$ for some $\alpha \in (0,1)$, then the process is polynomially ergodic in total variation distance (see [78]). This is particularly of interest in the case of heavy tailed stationary distributions, where common processes fail to have exponential convergence to their target.

In the next theorem we relate these types of convergence between two processes which are connected by (6.9). As we shall show, choosing a suitable speed function s can make the convergence geometric for X_t even if Y_t is only polynomially ergodic, while uniform ergodicity of X_t can be achieved if (6.13) holds for bounded V for a generic W.

Theorem 6.20. *Consider two irreducible, aperiodic Markov processes X_t and Y_t connected by* (6.9) *via a $\mathcal{C}^1$ speed function s. Suppose Y_t satisfies Assumption 6.16 and that set C is petite also for X_t. Assume s is lower bounded by a positive constant $\underline{s}$ and is upper bounded on C. Suppose also that $\mu_s \propto s\mu$ is a stationary probability distribution for Y_t, or equivalently μ is stationary for X_t. Then the following statements hold:*

1. *If $s(z) \geq l/W(z)$ for all $z \notin C$ and some constant $l > 0$, then X_t is geometrically ergodic. Moreover, X_t is uniformly ergodic if V is bounded.*
2. *If $s(z) \geq lV^{-\alpha}/W(z)$ for all $z \notin C$ and some constants $\alpha \in (0,1)$ and $l > 0$, then X_t is polynomially ergodic.*

Proof. The proof is a straightforward consequence of manipulations of the drift condition (6.13), which we already observed gives the drift condition (6.14). □

Remark 6.21. Let us make some remarks on the theorem. First, it needs to be stressed that, since X_t has a fixed stationary distribution μ, changing the speed s implies a change in the stationary distribution of Y_t and possibly in its Lyapunov function. Then, the statements give the necessary conditions on the speed function to obtain any type of convergence. Let us consider some cases:

- If $W(z)V(z) \geq 1$, then Y_t is ergodic, but we cannot conclude on its rate of convergence. Applying the theorem, we obtain polynomial or geometric convergence of X_t choosing respectively $s(z) \geq lV^{-\alpha}/W(z)$ and $s(z) \geq l/W(z)$.
- If $W(z)V(z) \geq V^{1-\alpha}(z)$, then Y_t is polynomially ergodic. By our theorem, we obtain geometric ergodicity of X_t choosing $s(z) \geq l/W(z)$.
- Uniform ergodicity of X_t follows as soon as V is bounded and we choose a speed function satisfying $s(z) \geq l/W(z)$.

Recall that exponential ergodicity of the ZZS was proved in [24] under Assumption 5.13. Clearly, exponential ergodicity of SUZZ follows when $\psi - \ln s$ satisfies Assumption 5.13 and also when s satisfies the conditions of Corollary 6.15.

6.3.3 Law of large numbers

One of the main interests in ergodic Markov processes is the estimation of expectations of observables wrt the stationary distribution, as indeed as T goes to infinity

$$\frac{1}{T}\int_0^T f(X_t)\mathrm{d}t \to \mu(f) \quad a.s.$$

Here it is convenient to denote $\mu(x) = \exp(-\psi(x))/Z$, where $\exp(-\psi(x))$ is the unnormalised distribution. Similarly, we denote the stationary distribution of the process Y_t, related to X_t by related by (6.11), as $\mu_s(y) = s(y)\exp(-\psi(y))/Z_s$. It might happen that the process X_t is difficult to simulate exactly, while we are able to simulate Y_t. In this scenario, a change of variables gives

$$\frac{1}{T}\int_0^T f(X_t)\mathrm{d}t = \frac{1}{T}\int_0^{r(T)} \frac{f(Y_t)}{s(Y_t)}\mathrm{d}t$$

which gives an estimator that only relies on the simulation of Y_t as this can also be rephrased as

$$\frac{1}{r^{-1}(T)}\int_0^T \frac{f(Y_t)}{s(Y_t)}\mathrm{d}t. \tag{6.15}$$

Let us verify this estimator converges to $\mu(f)$. Assuming Y_t satisfies a law of large numbers we find that

$$\frac{1}{T}\int_0^T \frac{f(Y_t)}{s(Y_t)}\mathrm{d}t \to \mu(f)\frac{Z}{Z_s}$$

and also

$$\frac{1}{T}r^{-1}(T) = \frac{1}{T}\int_0^T \frac{1}{s(Y_u)}\mathrm{d}u \to \mu_s(1/s) = \frac{Z}{Z_s}, \tag{6.16}$$

thus we find that (6.15) converges to $\mu(f)$ as required. When Z_s is known, (6.16) can be used to estimate Z, the normalising constant of μ.

For the settings in which the estimator in (6.15) is hard to compute, we can resort to a discretisation of the path by taking a step size δ and considering the Markov chain $\tilde{Y}_n = Y_{n\delta}$. Then (6.15) suggests estimating $\mu(f)$ with

$$\sum_{n=0}^{N} \frac{f(\tilde{Y}_n)}{s(\tilde{Y}_n)} \Bigg/ \sum_{n=0}^{N} \frac{1}{s(\tilde{Y}_n)}. \tag{6.17}$$

Naturally, the variance of this estimator is influenced by the choice of s.

The estimators (6.15) and (6.17) essentially coincide with the importance sampling approach we described in Section 2.1.3. Indeed, it is clear that (6.17) corresponds to drawing from a Markov chain which converges to μ_s and adjusting the weights of the samples suitably. The denominator in (6.17) then plays the same function of the denominator in the self-normalised importance sampling estimator (2.5).

6.4 Time transformations of piecewise deterministic Markov processes

In this section we study time transformations of PDMPs. This allows us to introduce novel processes based on existing PDMPs, modified to include a speed function.

Consider a PDMP $(Z_t)_{t\geq 0}$ taking values on a state space E, which is a subset of a finite dimensional vector space. Examples are $E = \mathbb{R}^d \times \mathbb{R}^d$ or $E = \mathbb{R}^d \times \{-1, +1\}^d$. The PDMP Z_t has deterministic dynamics described by the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi_t(z) = \Phi(\varphi_t(z)), \quad \varphi_0(z) = z, \qquad \text{for all } t \geq 0, z \in E, \tag{6.18}$$

where φ_t denote the integral curve of Φ, a smooth and globally Lipschitz vector field. We assume that φ_t leaves E invariant. The random event times are defined by a switching rate $\lambda : E \to [0, \infty)$, that is a continuous function, while the jump mechanism is described by the probability kernel Q. Therefore Z_t is a PDMP defined by the triple (Φ, λ, Q)[1]. This process should be intended as a well studied PDMP, such as the Zig-Zag process.

Let us now consider time transformations of the PDMP Z_t as given by (6.9). For a given speed function s, we shall denote the time transformed process as H_t. The next theorem, which is a corollary of Theorem 6.10, establishes the dynamics of H_t.

Theorem 6.22. *Let $s : E \to [\underline{s}, \infty)$ and let $(Z_t)_{t\geq 0}$ be a non-explosive PDMP (Φ, λ, Q) with stationary probability measure $\mu_s(z) \propto s(z)\exp(-\psi(z))$. Then the process $(H_t)_{t\geq 0}$ which satisfies $H_t = Z_{r(t)}$ for all $t \geq 0$, where $r(t) = \int_0^t s(H_u)\mathrm{d}u$, is*

[1]Note that, contrarily to previous chapters, here we identify the ODE with Φ instead of its integral.

a PDMP with characteristics $(s\Phi, s\lambda, Q)$ *and invariant measure* $\mu(z) \propto \exp(-\psi(z))$. *Moreover, the generator of* H_t *is given by*

$$\mathcal{L}_s f(z) = s(z)\langle \Phi(z), \nabla f(z)\rangle + s(z)\lambda(z)(Qf(z) - f(z)),$$

and its domain coincides with $\mathcal{D}(\mathcal{L})$.

Remark 6.23. The framework introduced by [48] only considers non-explosive deterministic dynamics, while in this case s typically gives an ODE finite explosion time. We can however apply Proposition 6.18 to determine make sure the time transformed process is non-explosive.

Although the theorem is an immediate corollary of Theorem 6.10, it is illustrative to derive the deterministic dynamics and the jumps of H_t, which follow by simple computations. First, we obtain the ODE governing the deterministic motion of H_t. Denote the flow map of H_t with initial condition z by $\varphi_t^H(z)$, which satisfies $\varphi_t^H(z) = \varphi_{r(t)}(z)$. Thus by the chain rule $\varphi_t^H(z)$ satisfies the ODE

$$\frac{d\varphi_t^H(z)}{dt} = s(\varphi_t^H(z))\Phi(\varphi_t^H(z)).$$

This shows that the deterministic motion of H_t is identified by $s\Phi$. Let us now consider the jump part of H_t. Denoting by λ_H the event rate of H_t, by τ_H the time of the first event for H_t, and noting that $r(t)$ is deterministic before an event has happened, we find by a change of variables

$$\mathbb{P}_z(\tau_H > t) = \exp\left(-\int_0^t \lambda_H(\varphi_u^H(z))\mathrm{d}u\right) = \exp\left(-\int_0^{r(t)} \frac{\lambda_H(\varphi_u(z))}{s(\varphi_u(z))}\mathrm{d}u\right).$$

Because H_t is a time transformation of Z_t, the processes follow the same paths and have random events when the same state is reached. Therefore, it must be that for any $t \geq 0$ it holds $\mathbb{P}_z(\tau_Z > r(t)) = \mathbb{P}_z(\tau_H > t)$. This holds iff $\lambda_H(z) = \lambda(z)s(z)$. Finally, note that the jump kernel is clearly unchanged by the time transformation. Therefore, we have obtained that H_t is a PDMP with characteristics $(s\Phi, s\lambda_Z, Q)$.

6.4.1 Examples

The SUZZ we encountered in Section 6.2.2 is a first example of time transformation of a known PDMP. Here, we define time transformations of several other PDMPs from the MCMC literature.

Example 6.24 (Bouncy Particle Sampler [32])**.** *Let us then define the BPS with speed* s, *denoted by* $(X_t, V_t)_{t\geq 0}$, *with a given invariant measure with density* $\mu(x, v) = \pi(x)\nu(v)$. *Following the recipe given by Theorem 6.22, we should time transform a standard BPS with target* $\mu_s(y, w) \propto s(y)e^{-\psi(y)}\nu(w)$. *Here we let* s *depend only on the position part as this seems the only scenario of interest, though the extension is*

straightforward. Therefore we obtain that the BPS with speed s is a PDMP with ODE (6.3)*, event rate*

$$\lambda_s(x,v) = \max(0, \langle v, s(x)\nabla\psi(x) - \nabla s(x)\rangle) + \lambda_r s(x), \tag{6.19}$$

and jump mechanism

$$\begin{aligned} Q_s((x,v),(dy,dw)) = &\frac{\max(0, \langle v, s(x)\nabla\psi(x) - \nabla s(x)\rangle)}{\lambda_s(x,v)}\, \delta_{(x,R_s(x)v)}(dy,dw) \\ &+ \frac{s(x)\lambda_r}{\lambda_s(x,v)}\, \delta_x(dy)\nu(dw), \end{aligned} \tag{6.20}$$

where the bounce mechanism corresponds to reflections off the potential $\psi_s(x) = \psi(x) - \ln s(x)$, i.e.

$$\begin{aligned} R_s(x)v &= v - 2\, \frac{\langle v, \nabla\psi(x) - \frac{\nabla s(x)}{s(x)}\rangle}{\left|\nabla\psi(x) - \frac{\nabla s(x)}{s(x)}\right|^2} \left(\nabla\psi(x) - \frac{\nabla s(x)}{s(x)}\right) \\ &= v - 2\, \frac{\langle v, s(x)\nabla\psi(x) - \nabla s(x)\rangle}{|s(x)\nabla\psi(x) - \nabla s(x)|^2} \left(s(x)\nabla\psi(x) - \nabla s(x)\right). \end{aligned} \tag{6.21}$$

Therefore, the reflection mechanism is different from that of the standard BPS with target μ. This is due to the fact that the reflection operator depends on the invariant distribution for the standard BPS, while this was not the case for the ZZS. A natural modification of the event rate is the choice $\lambda_r(x) = c/s(x)$, which makes the rate constant for the sped-up process.

Remark 6.25. The BPS is typically designed to have as stationary distributions for the velocity component either the uniform distribution on the unit sphere or the standard Gaussian distribution. Let us give an informal connection between the two alternatives through the lens of time transformations. Applying a time transformation with speed function $s(x,v) = 1/|v|$ to the BPS with Gaussian velocity gives a process with generator $\mathcal{L}_s f(x,v) = 1/|v|\, \mathcal{L}f(x,v)$. Such process evolves in the x-component exactly as a BPS with the uniform distribution on the unit sphere as stationary distribution for the velocity component. Indeed, the velocity is refreshed from the standard Gaussian distribution and is then normalised by its norm in the deterministic part (that is $\langle v/|v|, \nabla_x f(x,v)\rangle$ i.e. the process moves with velocity $v/|v|$) and in the part corresponding to reflections (that is $\langle v/|v|, \nabla\psi(x)\rangle_+(f(x,R(x)v) - f(x,v))$, which gives that reflections take place at the same point as for the original process).

Example 6.26 (Boomerang Sampler [25])**.** *We now wish to define the Boomerang sampler (X_t, V_t) with speed s and stationary distribution $\mu(x,v) \propto \exp(-\psi(x,v))$, where $\psi(x,v) = \psi(x) + (1/2)x^T\Sigma^{-1}x + (1/2)v^T\Sigma^{-1}v$. We find that (X_t, V_t) is the PDMP with characteristics $(s\Phi, \lambda_s, Q_s)$, where $\Phi(x,v) = (v, -(x - x_*))^T$ is the vector field defining the ODE of the standard Boomerang process, while λ_s and Q_s are*

respectively given by (6.19) *and* (6.20)*. Hence the deterministic motion is governed by*

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = s(X_t)V_t, \quad \frac{\mathrm{d}V_t}{\mathrm{d}t} = -s(X_t)(X_t - x_*).$$

Let us conclude the example with a brief discussion on a particular choice of speed function. Suppose we wish to sample from $\pi(x) \propto \exp(-\psi(x) - (1/2)x^T\Sigma^{-1}x)$*. Then a possible choice of speed function is* $s(x,v) = s(x) = \exp(\psi(x))$*, which satisfies the assumption* $\int s(x)\exp(-\psi(x,v))\mathrm{d}x\mathrm{d}v < \infty$*. In this case the corresponding Boomerang process with speed* s*, denoted* (X_t, V_t)*, is a time transformation of the standard Boomerang process with Gaussian invariant measure*

$$\mu_s(x,v) \propto \exp(-(1/2)x^T\Sigma^{-1}x - (1/2)v^T\Sigma^{-1}v).$$

For this process the rate of reflections is 0 *and therefore the motion is fully characterised by deterministic elliptical trajectories and velocity refreshments. As a consequence, the simulation of the paths of* (X_t, V_t) *does not require computing the gradient of* ψ*. Naturally the additional difficulty is that then the time transform* $r(t) = \int_0^t \exp(\psi(X_u))\mathrm{d}u$ *needs to be approximated if the correct state at time* t *of the Boomerang process with speed* s *needs to be known.*

Example 6.27 (Randomised Hamiltonian Monte Carlo [30])**.** *In the case of RHMC of [30] the jump mechanism is trivial and the target plays a role only in the deterministic dynamics. The dynamics of the standard RHMC with target* $\mu_s(y,w) \propto s(y)\exp(-\psi(y) - (1/2)|w|^2)$ *are described by the ODE*

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = W_t, \quad \frac{\mathrm{d}W_t}{\mathrm{d}t} = -\nabla\psi(Y_t) + \frac{\nabla s(Y_t)}{s(Y_t)}.$$

From Theorem 6.22 we conclude that the RHMC process with speed s *and invariant distribution with density* $\mu(x,v) \propto \exp(-\psi(x) - (1/2)|v|^2)$ *is a PDMP* (X_t, V_t) *with deterministic dynamics*

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = s(X_t)V_t, \quad \frac{\mathrm{d}V_t}{\mathrm{d}t} = -s(X_t)\nabla\psi(X_t) + \nabla s(X_t).$$

Such process moves on the level curves of the Hamiltonian $H(x,v) = -\psi(x) + \ln s(x) - (1/2)|v|^2$ *with speed given by* $s(x)$*. Naturally, the rate of refreshment should also be multiplied by the speed* s*. With analogous ideas it is possible to define a time transformed version of RHMC samplers with different choices kinetic energy, as e.g. in [120].*

6.4.2 Estimating expectations with the skeleton chain

In Section 6.3.3 we described how to estimate expectations from μ exploiting the relation between the two processes X_t and Y_t connected by a time transformation. Now, we consider a different approach which is specialised for PDMPs relying on

[42, Theorem 1]. Essentially, [42, Theorem 1] connects the invariant distribution of the PDMP Z_t to that of its skeleton chain, i.e. the discrete chain $(\tilde{Z}_n)$ given by the states of the PDMP Z_t right after each random event. In the context of two PDMPs connected by a time transformation, it is clear that their skeleton chains coincide: given the sequence of event times $\{\tau_k\}_{k\geq 0}$ for Z_t, it holds that $(Z_{\tau_k})_{k\geq 0} = (H_{r(\tau_k)})_{k\geq 0}$. An application of [42, Theorem 1] gives that the skeleton chain has invariant distribution $\tilde{\mu}(\cdot) \propto \int_E \lambda(z)Q(z,\cdot)\mu_s(\mathrm{d}z)$. From this property asymptotically exact estimators of observables can be computed using only the skeleton chain.

6.4.3 Conditions for uniform ergodicity

Theorem 6.20 gives general conditions on the speed function to obtain a time transformed process with wanted rate of convergence. In the case of PDMPs, the next theorem shows that uniform ergodicity follows also from an inequality of the type (6.13) where V is not bounded.

Theorem 6.28. *Consider a PDMP Z_t with characteristics (Φ, λ, Q) and invariant probability distribution $\mu_s \propto s\mu$. Let H_t be the corresponding time transformed PDMP $(s\Phi, s\lambda, Q)$. Assume there exists $V : \mathbb{R}^d \to [1,\infty)$ for which (6.13) holds for Z_t for some function W and a set C which is petite for both PDMPs. Suppose $s(z) \geq bV(z)/W(z)$ outside of C for some constant $b > 0$. Then H_t is uniformly ergodic: for some $\rho \in (0,1)$, $c > 0$*

$$\|P^t(z,\cdot) - \mu(\cdot)\|_V \leq c\rho^t \frac{1+2V(z)}{1+V(z)} \leq 2c\rho^t. \tag{6.22}$$

This theorem can be readily applied to design uniformly ergodic variants of the ZZS and BPS taking advantage of the results in [24, 51, 64].

Proof. Let $\tilde{V}(z) = V(z)/(1+V(z))$, where V is a Lyapunov function satisfying (6.13) for the PDMP (Φ, λ, Q), which has generator denoted by $\mathcal{L}$. Most importantly $\tilde{V}$ is bounded since $V > 0$. Applying the generator to $\tilde{V}$ we find

$$\mathcal{L}\tilde{V}(z) = \frac{1}{(1+V(z))^2}\langle \Phi(z), \nabla V(z)\rangle + \lambda(z)\int (\tilde{V}(y) - \tilde{V}(z))Q(z,\mathrm{d}y).$$

Now for the jump part we obtain

$$\lambda(z)\int (\tilde{V}(y) - \tilde{V}(z))Q(z,dy) = \lambda(z)\int \frac{V(y)-V(z)}{(1+V(y))(1+V(z))}Q(z,\mathrm{d}y)$$

Considering both the case $V(y) \geq V(z)$ and $V(z) \geq V(y)$ we obtain the inequality

$$\frac{V(y)-V(z)}{(1+V(y))(1+V(z))} \leq \frac{V(y)-V(z)}{(1+V(z))^2}.$$

Therefore applying our assumptions we find

$$\begin{aligned}\mathcal{L}\tilde{V}(z) &\leq \frac{1}{(1+V(z))^2}\Big(\langle\Phi(z),\nabla V(z)\rangle + \lambda(z)\int (V(y)-V(z))Q(z,\mathrm{d}y)\Big)\\ &= \frac{1}{(1+V(z))^2}\ \mathcal{L}V(z)\\ &\leq \frac{1}{(1+V(z))^2}\Big(-W(z)V(z) + m\mathbb{1}_C(z)\Big)\\ &\leq -\frac{W(z)}{1+V(z)}\,\tilde{V}(z) + m\mathbb{1}_C(z).\end{aligned}$$

Uniform ergodicity of the PDMP $(s\Phi, s\lambda, Q)$ then follows by Theorem 6.20(b) since outside of C we have $s(z) \geq bV(z)/W(z)$ for $b > 0$, which implies that $sW/(1+V) = bV/(1+V) \geq b/2$. □

Now we apply Theorem 6.28 to the ZZS, showing that for a speed function of the type $s(x) = \exp(\beta\psi(x))$ for $\beta \in (0,1)$ we obtain that SUZZ is uniformly ergodic.

Corollary 6.29. *Suppose $\psi > 0$ satisfies Assumption 5.13 and consider the speed function $s(x) = \exp(\beta\psi(x))$ for $\beta \in (0,1)$ and a refreshment rate γ such that $0 < \underline{\gamma} \leq \gamma(x) \leq \overline{\gamma} < \infty$. Then the SUZZ obtained by (6.9) with refreshment rate $s\gamma$ is uniformly ergodic.*

Proof. For $s(x) = \exp(\beta\psi(x))$ with $\beta \in (0,1)$ the SUZZ with target $\pi(x) \propto \exp(-\psi(x))$ corresponds to a time transformation of the standard ZZS with tempered target $\pi_s(y) \propto \exp(-(1-\beta)\psi(y))$. Clearly, potentials $(1-\beta)\psi$ for $\beta \in (0,1)$ satisfy Assumption 5.13 since we assume that ψ does. Therefore, under Assumption 5.13 we have by [24] that the standard ZZS with target π_s satisfies the drift condition (6.13) with $W(z) = l$ for some $l > 0$ and Lyapunov function

$$V(x,v) = \exp\left(\alpha(1-\beta)\psi(x) + \sum_{i=1}^{d}\phi(v_i(1-\beta)\partial_i\psi(x))\right),$$

where $\delta, \alpha > 0$ are such that $0 < \delta\overline{\gamma} < \alpha < 1$ in which $\overline{\gamma}$ is the maximum refreshment rate, and finally $\phi(s) = \frac{1}{2}\mathrm{sign}(s)\ln(1+\delta|s|)$. V is a Lyapunov function for arbitrarily small values of the constant α, as δ can be taken arbitrarily small. Therefore, it is now sufficient to choose $\alpha \in (0,1)$ such that outside of C

$$s(x) = \exp(\beta\psi(x)) \geq b\,V(x,v),$$

which is possible since by Assumption 5.13 we have ψ is the leading term in the exponent of V and the set C can accordingly be chosen large enough.

Moreover, any compact set is petite by Corollary 6.15 under our assumption and therefore by Theorem 6.28 the SUZZ is uniformly ergodic for our choice of speed function. □

6.5 Discussion

In this Chapter, we have introduced time transformations as an approach to improve convergence of Markov processes with challenging stationary distributions. According to our theory, substantial speed ups can be obtained with suitable choices of the speed function, s. A choice which appears particularly apt is $s(x) = \exp(a\psi(x))$ for $a \in (0, 1)$, which corresponds to targeting a tempered version of the target μ. Naturally, the parameter a determines the degree to which the target is "flattened" and is of fundamental importance to obtain a process which has the right balance between exploration of the state space and exploitation once a new mode is found. Investigations on the optimal value of a are left for future research.

An important theme which we did not treat in detail is that of the simulation of sped up PDMPs. A possibility is to simulate the standard PDMP Z_t and then compute the (now deterministic) function

$$r^{-1}(T) = \int_0^T \frac{1}{s(Z_t)} \mathrm{d}t.$$

This quantity will typically need to be estimated numerically, though we stress that this is a one dimensional integral which can be decomposed as

$$r^{-1}(T) = \sum_{\kappa=0}^{N_T-1} \int_0^{\tau_{\kappa+1}-\tau_\kappa} \frac{1}{s(\varphi_t(Z_{\tau_\kappa}))} \mathrm{d}t,$$

where τ_κ are the event times of the process, $0 = \tau_0 < \tau_1, < \cdots < \tau_{N_T}$, and N_T is the number of events of the process by time T. An alternative approach consists in approximating either Z_t or H_t with the techniques of Chapters 3 and 4. We leave a detailed analysis of this option as future research.

6.A Time transformations of Langevin diffusions

In this appendix, we apply the framework of Section 6.3 to the overdamped and underdamped Langevin diffusions, which are the basis for several MCMC algorithms.

Example 6.30 (Overdamped Langevin SDE)**.** *Consider the overdamped Langevin SDE*

$$\mathrm{d}Y_t = \left(-\nabla\psi(Y_t) + \frac{\nabla s(Y_t)}{s(Y_t)}\right)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$$

which is ergodic with respect to $\mu_s(y) \propto s(y)\exp(-\psi(y))$ under mild conditions. Then by Theorem 6.10 we find the the solution to

$$\mathrm{d}X_t = (-s(X_t)\nabla\psi(X_t) + \nabla s(X_t))\,\mathrm{d}t + \sqrt{2s(X_t)}\,\mathrm{d}B_t,$$

has invariant measure $\mu(x) \propto \exp(-\psi(x))$. *Moreover it holds that* $Y_t = X_{r(t)}$, *that is the paths of* X_t *and* Y_t *are identical apart from the time transformation. For* $s(x) = \exp(a\psi(x))$ *with* $a \in (0,1)$, *the process* X_t *was considered in [138].*

Example 6.31 (Underdamped Langevin SDE)**.** *Another well known and used process is the underdamped Langevin SDE, which in its simplest form is given by*

$$\begin{aligned} \mathrm{d}Y_t &= W_t \mathrm{d}t \\ \mathrm{d}W_t &= -\nabla\psi(Y_t)\mathrm{d}t - W_t\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \end{aligned}$$

This process has $\mu(y,w) \propto \exp(-\psi(y) - (1/2)|w|^2)$ *as stationary distribution under mild assumptions. Once again we wish to define a the underdamped Langevin SDE including a speed function* s. *Following Theorem 6.10 we find that this SDE is given by*

$$\begin{aligned} \mathrm{d}X_t &= s(X_t)V_t\mathrm{d}t \\ \mathrm{d}V_t &= -(s(X_t)\nabla\psi(X_t) - \nabla s(X_t))\mathrm{d}t - s(X_t)V_t\mathrm{d}t + \sqrt{2s(X_t)}\,\mathrm{d}B_t. \end{aligned}$$

Assuming this SDE is well posed, we find that the solution has μ *as stationary distribution.*

Bibliography

[1] A. B. Abdessalem, R. Azaïs, M. Touzet-Cortina, A. Gégout-Petit, and M. Puiggali. Stochastic modelling and prediction of fatigue crack propagation using piecewise-deterministic Markov processes. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 230(4):405–416, 2016.

[2] C. Andrieu and S. Livingstone. Peskun–tierney ordering for markovian monte carlo: beyond the reversible scenario. *Annals of Statistics*, 49(4):1958–1981, 2021.

[3] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18, 12 2008.

[4] C. Andrieu, P. Dobson, and A. Q. Wang. Subgeometric hypocoercivity for piecewise-deterministic Markov process Monte Carlo methods. *Electronic Journal of Probability*, 26:1 – 26, 2021.

[5] C. Andrieu, A. Durmus, N. Nüsken, and J. Roussel. Hypocoercivity of piecewise deterministic Markov process-Monte Carlo. *The Annals of Applied Probability*, 31(5):2478 – 2517, 2021.

[6] S. Apers, A. Sarlette, and F. Ticozzi. Characterizing limits and opportunities in speeding up markov chain mixing. *Stochastic Processes and their Applications*, 136:145–191, 2021.

[7] W. Arveson. Notes on measure and integration in locally compact spaces, 1996.

[8] Y. Bai, G. O. Roberts, and J. S. Rosenthal. On the containment condition for adaptive Markov chain Monte Carlo algorithms. *Advances and Applications in Statistics*, 21(1):1–54, 2011.

[9] R. Bardenet, A. Doucet, and C. C. Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47), 2017.

[10] A. Barker. Monte carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18(2):119–134, 1965.

[11] M. Benaïm, S. Le Borgne, F. Malrieu, and P.-A. Zitt. Qualitative properties of certain piecewise deterministic markov processes. *Annales de l'I.H.P. Probabilités et statistiques*, 51(3):1040–1075, 2015.

[12] H. C. Berg and D. A. Brown. Chemotaxis in escherichia coli analysed by three-dimensional tracking. *Nature*, 239(5374):500–504, 1972.

[13] É. Bernard, W. Krauth, and D. B. Wilson. Event-chain Monte Carlo algorithms for hard-sphere systems. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 80 5 Pt 2:056704, 2009.

[14] A. Bertazzi. Time transformations of piecewise-deterministic Markov processes. *In preparation*, 2023.

[15] A. Bertazzi and J. Bierkens. Adaptive schemes for piecewise deterministic Monte Carlo algorithms. *Bernoulli*, 28(4):2404 – 2430, 2022.

[16] A. Bertazzi, J. Bierkens, and P. Dobson. Approximations of Piecewise Deterministic Markov Processes and their convergence properties. *Stochastic Processes and their Applications*, 154:91–153, 2022.

[17] A. Bertazzi, P. Dobson, and P. Monmarché. Splitting schemes for second order approximations of piecewise-deterministic Markov processes. *arXiv.2301.02537*, 2023.

[18] J. Bierkens. Non-reversible Metropolis-Hastings. *Statistics and Computing*, 26 (6):1213–1228, 2016.

[19] J. Bierkens and A. Duncan. Limit theorems for the zig-zag process. *Advances in Applied Probability*, 49(3):791–825, 2017.

[20] J. Bierkens and S. M. V. Lunel. Spectral analysis of the zigzag process. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 58(2):827 – 860, 2022.

[21] J. Bierkens and G. Roberts. A piecewise deterministic scaling limit of lifted metropolis–hastings in the curie–weiss model. *The Annals of Applied Probability*, 27(2):846–882, 2017.

[22] J. Bierkens, A. Bouchard-Côté, A. Doucet, A. B. Duncan, P. Fearnhead, T. Lienart, G. Roberts, and S. J. 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.

[23] J. Bierkens, P. Fearnhead, and G. Roberts. The zig-zag process and super-efficient sampling for bayesian analysis of big data. *Annals of Statistics*, 47, 2019.

[24] J. Bierkens, G. O. Roberts, and P.-A. Zitt. Ergodicity of the zigzag process. *The Annals of Applied Probability*, 29(4):2266–2301, 2019.

[25] J. Bierkens, S. Grazzi, K. Kamatani, and G. Roberts. The boomerang sampler. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 908–918. PMLR, 2020.

[26] J. Bierkens, S. Grazzi, F. v. d. Meulen, and M. Schauer. Sticky PDMP samplers for sparse and local inference problems. *Statistics and Computing*, 33(1):8, 2022.

[27] J. Bierkens, K. Kamatani, and G. O. Roberts. High-dimensional scaling limits of piecewise deterministic sampling algorithms. *The Annals of Applied Probability*, 32(5):3361 – 3407, 2022.

[28] A. Bonfiglioli and R. Fulci. *Topics in noncommutative Algebra: The Theorem of Campbell, Baker, Hausdorff and Dynkin*, volume 2034. Springer, 01 2012. ISBN 978-3-642-22596-3.

[29] N. Bou-Rabee and A. Eberle. Markov chain Monte Carlo methods. *Lecture notes*, 2023.

[30] N. Bou-Rabee and J. M. Sanz-Serna. Randomized hamiltonian monte carlo. *The Annals of Applied Probability*, 27(4):2159–2194, 2017.

[31] N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of Applied Probability*, 30(3), Jun 2020.

[32] 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.

[33] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

[34] 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. ISBN 1581130678.

[35] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference on learning theory*, pages 300–323. PMLR, 2018.

[36] A. Chevallier, S. Power, A. Q. Wang, and P. Fearnhead. PDMP Monte Carlo methods for piecewise-smooth densities. *arXiv.2111.05859*, 2021.

[37] A. Chevallier, P. Fearnhead, and M. Sutton. Reversible Jump PDMP Samplers for Variable Selection. *Journal of the American Statistical Association*, 0(0): 1–13, 2022.

[38] C. Chimisov, K. Latuszynski, and G. Roberts. Adapting The Gibbs Sampler. *arXiv.1801.09299*, 2018.

[39] Cloez, Bertrand, Dessalles, Renaud, Genadot, Alexandre, Malrieu, Florent, Marguet, Aline, and Yvinec, Romain. Probabilistic and Piecewise Deterministic models in Biology. *ESAIM: Procs*, 60:225–245, 2017.

[40] C. Cocozza-Thivent. *Markov Renewal and Piecewise Deterministic Processes*. Springer, 2021.

[41] A. Corbella, S. E. F. Spencer, and G. O. Roberts. Automatic zig-zag sampling in practice. *arXiv:2206.11410*, 2022.

[42] O. L. V. Costa. Stationary Distributions for Piecewise-Deterministic Markov Processes. *Journal of Applied Probability*, 27(1):60–73, 1990.

[43] O. L. V. Costa and F. Dufour. Stability and Ergodicity of Piecewise Deterministic Markov Processes. *SIAM Journal on Control and Optimization*, 47(2): 1053–1077, 2008.

[44] R. V. Craiu, L. Gray, K. Łatuszyński, N. Madras, G. O. Roberts, and J. S. Rosenthal. Stability of adversarial markov chains, with an application to adaptive mcmc algorithms. *Ann. Appl. Probab.*, 25(6):3592–3623, 12 2015.

[45] D. Crisan and M. Ottobre. Pointwise gradient bounds for degenerate semigroups (of UFG type). In *Proc. R. Soc. A*, volume 472, page 20160442. The Royal Society, 2016.

[46] D. Crisan, P. Dobson, and M. Ottobre. Uniform in time estimates for the weak error of the euler method for sdes and a pathwise approach to derivative estimates for diffusion semigroups. *Transactions of the American Mathematical Society*, 374(5):3289–3330, 2021.

[47] A. Dassios and P. Embrechts. Martingales and insurance risk. *Communications in Statistics. Stochastic Models*, 5(2):181–217, 1989.

[48] M. Davis. *Markov Models & Optimization*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1993. ISBN 9780412314100.

[49] M. H. A. Davis. Piecewise-Deterministic Markov Processes: A General Class of Non-Diffusion Stochastic Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):353–388, 1984.

[50] V. De Bortoli, A. Durmus, M. Pereyra, and A. F. Vidal. Maximum likelihood estimation of regularization parameters in high-dimensional inverse problems: an empirical Bayesian approach. Part II: Theoretical analysis. *SIAM Journal on Imaging Sciences*, 13(4):1990–2028, 2020.

[51] G. Deligiannidis, A. Bouchard-Côté, and A. Doucet. Exponential ergodicity of the bouncy particle sampler. *The Annals of Statistics*, 47(3):1268–1287, 06 2019.

[52] 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. *The Annals of Applied Probability*, 31(6):2612 – 2662, 2021.

[53] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

[54] P. Diaconis and L. Miclo. On the spectral analysis of second-order Markov chains. *Annales de la Faculté des Sciences de Toulouse. Série VI. Mathématiques*, 3, 01 2013.

[55] P. Diaconis, S. Holmes, and R. M. Neal. Analysis of a nonreversible Markov chain sampler. *The Annals of Applied Probability*, 10(3):726 – 752, 2000.

[56] P. Dobson and J. Bierkens. Infinite Dimensional Piecewise Deterministic Markov Processes. *arXiv:2205.11452*, 2022.

[57] 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.

[58] S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

[59] A. B. Duncan, T. Lelièvre, and G. A. Pavliotis. Variance Reduction Using Nonreversible Langevin Samplers. *Journal of Statistical Physics*, 163(3):457–491, 2016.

[60] A. Durmus and E. Moulines. Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm. *arXiv:1605.01559*, 2016.

[61] A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551 – 1587, 2017.

[62] A. Durmus and É. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854 – 2882, 2019.

[63] A. Durmus, E. Moulines, and M. Pereyra. Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

[64] A. Durmus, A. Guillin, and P. Monmarché. Geometric ergodicity of the Bouncy Particle Sampler. *The Annals of Applied Probability*, 30(5):2069–2098, 10 2020.

[65] A. Durmus, A. Guillin, and P. Monmarché. Piecewise deterministic Markov processes and their invariant measures. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(3):1442 – 1475, 2021.

[66] P. Embrechts and H. Schmidli. Ruin estimation for a general insurance risk model. *Advances in Applied Probability*, 26(2):404–422, 1994.

[67] K.-J. Engel and R. Nagel. One-parameter semigroups for linear evolution equations. *Semigroup Forum*, 63:278–280, 06 2001.

[68] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[69] P. Fearnhead, J. Bierkens, M. Pollock, and G. O. Roberts. Piecewise deterministic markov processes for continuous-time monte carlo. *Statistical Science*, 33 (3):386–412, 08 2018.

[70] G. Fort and E. Moulines. Polynomial ergodicity of Markov transition kernels. *Stochastic Processes and their Applications*, 103(1):57–99, 2003.

[71] G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 39(6):3262–3289, 12 2011.

[72] G. Fort, E. Moulines, P. Priouret, and P. Vandekerkhove. A central limit theorem for adaptive and interacting Markov chains. *Bernoulli*, 20, 07 2011.

[73] A. E. Gelfand and A. F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410): 398–409, 1990.

[74] 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.

[75] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[76] 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. *hal-02436593*, 2020. working paper or preprint.

[77] P. Gustafson. A guided walk Metropolis algorithm. *Statistics and Computing*, 8(4):357–364, 1998.

[78] M. Hairer. Convergence of Markov Processes. *Lecture notes*, 2010.

[79] M. Hairer and J. C. Mattingly. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, volume 63 of *Progr. Probab.*, pages 109–117. Birkhäuser/Springer Basel AG, Basel, 2011.

[80] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

[81] M. Hildebrand. Rates of convergence of the Diaconis-Holmes-Neal Markov chain sampler. *Markov Processes and Related Fields*, 10(4):687–704, 2004.

[82] A. M. Horowitz. A generalized guided monte carlo algorithm. *Physics Letters B*, 268(2):247–252, 1991.

[83] E. Horton, A. E. Kyprianou, and D. Villemonais. Stochastic methods for the neutron transport equation I: Linear semigroup asymptotics. *The Annals of Applied Probability*, 30(6):2573 – 2612, 2020.

[84] J. Huggins and J. Zou. Quantifying the accuracy of approximate diffusions and Markov chains. In *Artificial Intelligence and Statistics*, pages 382–391. PMLR, 2017.

[85] J. Jacod and P. Protter. *Probability Essentials*. Springer Berlin Heidelberg, 2nd edition, 2004. ISBN 9783540438717.

[86] J. E. Johndrow, N. S. Pillai, and A. Smith. No Free Lunch for Approximate MCMC. *arXiv preprint arXiv:2010.12514*, 2020.

[87] P. E. Kloeden and E. Platen. Stochastic differential equations. In *Numerical Solution of Stochastic Differential Equations*, pages 103–160. Springer, 1992.

[88] W. Krauth. Event-Chain Monte Carlo: Foundations, Applications, and Prospects. *Frontiers in Physics*, 9, 2021.

[89] B. Leimkuhler and C. Matthews. Rational Construction of Stochastic Numerical Methods for Molecular Sampling. *Applied Mathematics Research eXpress*, 2013 (1):34–56, 06 2012.

[90] B. Leimkuhler, D. T. Margul, and M. E. Tuckerman. Stochastic, resonance-free multiple time-step algorithm for molecular dynamics with very large time steps. *Molecular Physics*, 111(22-23):3579–3594, 2013.

[91] B. Leimkuhler, C. Matthews, and G. Stoltz. The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. *IMA J. Numer. Anal.*, 36(1):13–79, 2016.

[92] 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.

[93] V. Lemaire, M. Thieullen, and N. Thomas. Exact Simulation of the Jump Times of a Class of Piecewise Deterministic Markov Processes. *Journal of Scientific Computing*, 2017.

[94] V. Lemaire, M. Thieullen, and N. Thomas. Thinning and multilevel Monte Carlo methods for piecewise deterministic (Markov) processes with an application to a stochastic Morris–Lecar model. *Advances in Applied Probability*, 52(1):138–172, 2020.

[95] P. Lewis and G. Shedler. Simulation of Nonhomogeneous Poisson Processes by Thinning, 1978.

[96] T. M. Liggett. *Continuous Time Markov Processes: An Introduction*. American Mathematical Society, 2010.

[97] J. S. Liu. Peskun's theorem and a modified discrete-state Gibbs sampler. *Biometrika*, 83(3):681–682, 09 1996.

[98] E. Löcherbach and P. Monmarché. Metastability for systems of interacting neurons. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 58 (1):343 – 378, 2022.

[99] J. Lu and L. Wang. On explicit L^2-convergence rate estimate for piecewise deterministic Markov processes. *arXiv:2007.14927*, jul 2020.

[100] E. Marinari and G. Parisi. Simulated Tempering: A New Monte Carlo Scheme. *Europhysics Letters*, 19(6):451, jul 1992.

[101] G. M. Martin, D. T. Frazier, and C. P. Robert. Computing Bayes: From Then 'Til Now', 2022.

[102] N. Metropolis. The beginning of the Monte Carlo method. *Los Alamos Science*, 15 (Special Issue, Stanisław Ulam 1909–1984):125–130, 1987.

[103] N. Metropolis and S. Ulam. The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[104] 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.

[105] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, 2 edition, 2009.

[106] 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.

[107] 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.

[108] M. Michel, S. C. Kapfer, and W. Krauth. Generalized event-chain Monte Carlo: Constructing rejection-free global-balance algorithms from infinitesimal steps. *The Journal of Chemical Physics*, 140(5), 2014.

[109] M. Michel, A. Durmus, and S. Sénécal. Forward Event-Chain Monte Carlo: Fast Sampling by Randomness Control in Irreversible Markov Chains. *Journal of Computational and Graphical Statistics*, 29(4):689–702, 2020.

[110] P. Monmarché. Piecewise deterministic simulated annealing. *ALEA*, 13(1): 357–398, 2016.

[111] P. Monmarché. Kinetic walks for sampling. *ALEA Lat. Am. J. Probab. Math. Stat.*, 17:491–530, 2020.

[112] P. Monmarché. High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion. *Electronic Journal of Statistics*, 15(2): 4117 – 4166, 2021.

[113] P. Monmarché, J. Weisman, L. Lagardère, and J.-P. Piquemal. Velocity jump processes: An alternative to multi-timestep methods for faster and accurate molecular dynamics simulations. *The Journal of Chemical Physics*, 153(2): 024101, 2020.

[114] C. Morris and H. Lecar. Voltage oscillations in the barnacle giant muscle fiber. *Biophysical Journal*, 35(1):193–213, 1981.

[115] I. Murray, R. P. Adams, and D. J. C. Mackay. Elliptical slice sampling. In *International Conference on Artificial Intelligence and Statistics*, 2009.

[116] R. Neal. Improving Asymptotic Variance of MCMC Estimators: Non-reversible Chains are Better. *arXiv:math/0407281*, 08 2004.

[117] R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705 – 767, 2003.

[118] R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.

[119] K. Neklyudov, M. Welling, E. Egorov, and D. Vetrov. Involutive MCMC: A Unifying Framework. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[120] A. Nishimura, Z. Zhang, and M. A. Suchard. Hamiltonian zigzag sampler got more momentum than its Markovian counterpart: Equivalence of two zigzags under a momentum refreshment limit. *arXiv:2104.07694*, 2021.

[121] F. Pagani, A. Chevallier, S. Power, T. House, and S. Cotter. NuZZ: numerical Zig-Zag sampling for general models. *arXiv:2003.03636*, 2020.

[122] K. Pakdaman, M. Thieullen, and G. Wainrib. Fluid limit theorems for stochastic hybrid systems with application to neuron models. *Advances in Applied Probability*, 42, 01 2010.

[123] G. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Springer, 2014.

[124] P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 12 1973.

[125] E. A. J. F. Peters and G. De With. Rejection-free Monte Carlo sampling for general potentials. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 85(2):1–5, 2012.

[126] E. Pompe, C. Holmes, and K. Łatuszyński. A framework for adaptive MCMC targeting multimodal distributions. *Ann. Statist.*, 48(5):2930–2952, 10 2020.

[127] R. Poncet. Generalized and hybrid Metropolis-Hastings overdamped Langevin algorithms. *arXiv.1701.05833*, 2017.

[128] K. Ramanan and A. Smith. Bounds on Lifting Continuous-State Markov Chains to Speed Up Mixing. *Journal of Theoretical Probability*, 31(3):1647–1678, 2018.

[129] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2004.

[130] C. Robert and G. Casella. A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science*, 26(1):102 – 115, 2011.

[131] 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.

[132] G. O. Roberts and J. S. Rosenthal. Convergence of Slice Sampler Markov Chains. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):643–660, 1999.

[133] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351 – 367, 2001.

[134] G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probab. Surveys*, 1:20–71, 2004.

[135] G. O. Roberts and J. S. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458–475, 2007.

[136] G. O. Roberts and J. S. Rosenthal. Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[137] G. O. Roberts and J. S. Rosenthal. Polynomial Convergence Rates of Piecewise Deterministic Markov Processes. *Methodology and Computing in Applied Probability*, 25(1):6, 2023.

[138] G. O. Roberts and O. Stramer. Langevin Diffusions and Metropolis-Hastings Algorithms. *Methodology and Computing in Applied Probability*, 2002.

[139] G. O. Roberts and R. L. Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 2(4):341–363, 1996.

[140] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[141] J. S. Rosenthal. Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *Journal of the American Statistical Association*, 90(430): 558–566, 1995.

[142] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[143] R. Rudnicki and M. Tyran-Kamińska. *Piecewise deterministic processes in biological models*, volume 1. Springer, 2017.

[144] J. M. Sanz-Serna and K. C. Zygalakis. Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations. *J. Mach. Learn. Res.*, 22:242:1–242:37, 2021.

[145] C. Sherlock and A. H. Thiery. A discrete bouncy particle sampler. *Biometrika*, 109(2):335–349, 02 2021.

[146] A. Singh and J. P. Hespanha. Stochastic hybrid systems for studying biochemical processes. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1930):4995–5011, 2010.

[147] D. W. Stroock. Some stochastic processes which arise from a model of the motion of a bacterium. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 28(4):305–315, 1974.

[148] M. Sutton, R. Salomone, A. Chevallier, and P. Fearnhead. Continuously-Tempered PDMP Samplers. *arXiv.2205.09559*, 2022.

[149] A. Thin, N. Kotelevskii, C. Andrieu, A. Durmus, E. Moulines, and M. Panov. Nonreversible MCMC from conditional invertible transforms: a complete recipe with convergence guarantees. *arXiv:2012.15550*, 2020.

[150] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8(1):1 – 9, 1998.

[151] K. S. Turitsyn, M. Chertkov, and M. Vucelja. Irreversible Monte Carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240(4):410–414, 2011.

[152] P. Vanetti, A. Bouchard-Côté, G. Deligiannidis, and A. Doucet. Piecewise-Deterministic Markov Chain Monte Carlo. *arXiv:1707.05296*, 2017.

[153] G. Vasdekis and G. O. Roberts. Speed Up Zig-Zag. *arXiv.2103.16620*, 2021.

[154] G. Vasdekis and G. O. Roberts. A note on the polynomial ergodicity of the one-dimensional Zig-Zag process. *Journal of Applied Probability*, 59(3):895–903, 2022.

[155] M. Vialaret and F. Maire. On the convergence time of some non-reversible Markov chain Monte Carlo methods. *Methodology and Computing in Applied Probability*, 2020.

[156] A. F. Vidal, V. De Bortoli, M. Pereyra, and A. Durmus. Maximum likelihood estimation of regularization parameters in high-dimensional inverse problems: An empirical Bayesian approach part i: Methodology and experiments. *SIAM Journal on Imaging Sciences*, 13(4):1945–1989, 2020.

[157] M. Vucelja. Lifting – A Nonreversible Markov Chain Monte Carlo Algorithm. *American Journal of Physics*, 84, 12 2014.

[158] J. Wallin and D. Bolin. Efficient Adaptive MCMC Through Precision Estimation. *Journal of Computational and Graphical Statistics*, 27(4):887–897, 2018.

[159] C. Wu and C. P. Robert. Coordinate sampler: a non-reversible Gibbs-like MCMC sampler. *Statistics and Computing*, 30(3):721–730, 2020.

[160] K. Xu, H. Ge, W. Tebbutt, M. Tarek, M. Trapp, and Z. Ghahramani. AdvancedHMC.jl: A robust, modular and efficient implementation of advanced HMC algorithms. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–10. PMLR, 2020.

Summary

Markov chain Monte Carlo (MCMC) algorithms are a widespread tool to approximate expectations of functions of interest with respect to a target probability distribution π using Markovian processes. The task of estimating such expectations arises naturally in various areas, ranging from Bayesian statistics and machine learning to statistical physics and the applied sciences. This thesis studies a novel class of MCMC methods called *piecewise deterministic Monte Carlo* (PDMC) algorithms, which are based on a family of stochastic processes called *piecewise deterministic Markov processes* (PDMPs) and introduce important methodological novelties compared to classical methods based on diffusions. A PDMP evolves according to prescribed deterministic dynamics specified by an ordinary differential equation (ODE) until a random clock rings, at which point the PDMP jumps to a new state obtained by drawing from a given Markov kernel. The substantial interest that PDMPs have received by the MCMC community is justified by two properties they possess: their natural *non-reversibility* and the possibility for *exact subsampling*. Non-reversibility is a property that results in faster MCMC algorithms, departing from the inefficient random walk-like exploration of reversible processes obtained with the classical Metropolis Hastings (MH) algorithm. Essentially, non-reversible Markov processes explore the state space in a systematic fashion thanks to the introduction of an additional momentum variable, which guides them towards areas where π is large. PDMPs are of particular interest in this sense because they are the first class of processes which give a general framework to design non-reversible processes, where plenty of freedom is left in the choice of the ODE and of the jump dynamics. On the other hand, exact subsampling refers to the fact that, when e.g. π is a *posterior distribution* obtained by a Bayesian statistics model, the law of the random times can be computed while accessing only a subset of the data-set. This property makes PDMPs innovative, as for previous algorithms subsampling could only be achieved at the cost of a biased estimate of the expectation of interest. Thus PDMPs can lead to faster, more efficient MCMC algorithms when dealing with data sets with a large number of observations, a situation that is ever more common in the applications.

This thesis studies methods to improve the applicability and the performance of MCMC algorithms based on PDMPs. First, we discuss the key ideas that lay the

foundations of the field of MCMC, spanning from the Metropolis-Hastings algorithm to PDMC methods, emphasising a common structure underlying most non-reversible MCMC algorithms studied in the literature. The rest of the thesis is divided in two parts, respectively treating *approximations* and *transformations* of PDMC algorithms.

In the first part we introduce several discretisation schemes that approximate a given PDMP and study the properties of the proposed algorithms in detail. This area is of fundamental importance to make PDMPs widely applicable, as indeed the PDMPs considered in the MCMC literature typically cannot be simulated exactly because of either complicated deterministic dynamics or because the random event times are distributed according to an exponential distribution with *non-homogeneous rate*. In the latter case, existing approaches to simulate the random event times are applicable exclusively when the rate is of simple form, a requirement that covers only toy models from the MCMC literature. In this thesis we introduce and study a wide variety of time discretisations of PDMPs of any order of accuracy, which can now be used as a basis for MCMC algorithms. We study two types of discretisations: the first kind is obtained generalising the principle behind classical *Euler schemes*, while the second is based on *splitting schemes*. In both settings, we establish the dependence of the error on the step size of the discretisation. For suitable Euler schemes we prove *uniform in time estimates* on the weak error, a particularly challenging result which gives that the error is fully controlled by the step size and does not depend on the time horizon. Moreover, for approximations of PDMPs obtained with Euler-based schemes we obtain error bounds in Wasserstein and total variation distance using the coupling approach. For our approximations based on splitting schemes we mainly focus on the Zig-Zag sampler (ZZS) and Bouncy Particle Sampler (BPS) and study the best splitting scheme in terms of bias in the invariant measure. For both samplers we obtain conditions ensuring existence and uniqueness of a stationary distribution for the approximation process, as well as exponential convergence to such a distribution. Importantly, we show that symmetric splitting schemes are of second order, although they only require one computation of the gradient of the negative log-likelihood per iteration. Another important novelty we introduce is the possibility to correct the introduced bias via a skew-reversible Metropolis-Hastings acceptance-rejection step. This allows us to design the first unbiased, PDMP-based MCMC algorithms that can be applied effortlessly to sample from any target probability distribution. Our numerical experiments show that the remarkable properties of PDMPs give their approximations excellent convergence properties improving over benchmark methods such as Hamiltonian Monte Carlo and the unadjusted Langevin algorithm.

The second part of the thesis concerns transformations of PDMPs. First, we discuss *space transformations* of PDMPs, in which case the main goal is to improve the performance of PDMC algorithms when the target distribution π is anisotropic. Our proposal is to design PDMC algorithms that learn *adaptively* the covariance structure of π and use this information to tune the velocity of the underlying PDMP, i.e. the directions that the PDMP is more likely to explore. Finding a good set of directions requires knowledge of the target π, and hence information on previous

positions of the process needs to be used. In a similar fashion, we introduce adaptive PDMC algorithms which automatically tune the refreshment rate of the process, i.e. the frequency at which the current velocity vector is replaced with an independent draw from a suitable distribution. For these algorithms we carefully study the convergence to the target distribution by establishing ergodicity, which is challenging for such non-homogeneous Markov processes. Moreover, we test our algorithms on some benchmark examples, on which we observe relevant improvements over the standard, non-adaptive samplers. In the last chapter of the thesis we consider *time transformations* of (piecewise deterministic) Markov processes, with an emphasis on improving the convergence of MCMC algorithms. In particular, we study the effect on the properties of a Markov process of a change of the speed of time, where importantly changes in speed depend on the state of the process. This notion can prove helpful in the context of multimodal target distributions, in which case we argue that communication between different modes can be improved by increasing the speed of time when the process is located in low density regions. We connect various properties of a process to those of a related time-changed process, such as a connection between the stationary distributions, the generators, non-explosivity, ergodicity and rate of convergence to the limiting distribution. For PDMPs we show that suitable time transformations can make a geometrically ergodic Markov process *uniformly ergodic*, a remarkable property which means that the initialisation of the process does not affect the speed of convergence. We apply our theorem to time transformations of the Zig-Zag process, demonstrating the applicability of our conditions. By applying this framework to PDMPs we define several novel processes which have dynamics depending on a user-chosen, interpretable speed function.

Samenvatting

Markovketen Monte Carlo (MCMC) algoritmes worden veel gebruikt om verwachtingswaarden van relevante functies met een gewenste kansverdeling π te benaderen met behulp van Markoviaanse processen. Het schatten van dergelijke verwachtingswaarden komt op natuurlijke wijze voor in verschillende vakgebieden, van Bayesiaanse statistiek en machine learning tot statistische fysica en toegepaste wetenschappen. Deze scriptie onderzoekt een nieuwe klasse van MCMC-methoden genaamd *stuksgewijs deterministische Monte Carlo* (PDMC) algoritmes, die gebaseerd zijn op een familie van stochastische processen genaamd *stuksgewijs deterministische markovprocessen* (PDMP's) en belangrijke methodologische vernieuwingen introduceren vergeleken met klassieke methoden op basis van diffusie. Een PDMP ontwikkelt zich volgens voorgeschreven deterministische dynamica beschreven door een gewone differentiaalvergelijking (ODE) totdat er een willekeurige wekker afgaat, waarop de PDMP naar een nieuwe toestand springt, die verkregen wordt door te trekken uit een gegeven markovkernel. De aanzienlijke interesse voor PDMP's vanuit de MCMC-gemeenschap is gerechtvaardigd door twee eigenschappen die zij bezitten: natuurlijke *irreversibiliteit* en de mogelijkheid tot *exacte subsampling*. Irreversibiliteit is een eigenschap die zorgt voor snellere MCMC-algoritmes, anders dan de inefficiënte random walk-achtige verkenning van reversibele processen verkregen met het klassieke Metropolis Hastings (MH) algoritme. In essentie verkennen irreversibele markovprocessen de toestandsruimte op een systematische manier dankzij de invoering van een extra impulswaarde, die hen naar gebieden stuurt waar π groot is. PDMP's zijn in dit opzicht van bijzonder belang omdat ze de eerste klasse van processen vormen die een algemeen kader bieden om irreversibele processen te ontwerpen, waarbij er veel vrijheid is in de keuze van de ODE en van de sprongdynamica. Aan de andere kant verwijst exacte subsampling naar het feit dat, wanneer π bijvoorbeeld een *posterior-verdeling* is, verkregen door middel van een Bayesiaans statistisch model, de wet van de willekeurige tijden kan worden berekend terwijl er slechts toegang is tot een deel van de dataset. Deze eigenschap maakt PDMP's innovatief, omdat bij eerdere algoritmen subsampling slechts bereikt kon worden ten koste van een vertekende schatting van de betreffende verwachtingswaarde. Daarom kunnen PDMP's leiden tot snellere, ef-

Translated to Dutch by Merel de Leeuw den Bouter.

ficiëntere MCMC-algoritmes voor datasets bestaande uit een groot aantal observaties, een situatie die steeds vaker voorkomt in de praktijk.

Deze thesis onderzoekt methoden om de toepasbaarheid en de prestaties van MCMC-algoritmes gebaseerd op PDMP's te verbeteren. Eerst bespreken we de belangrijkste ideeën die ten grondslag liggen aan het veld van MCMC, variërend van het Metropolis-Hastings-algoritme tot PDMC-methoden, waarbij we de gemeenschappelijke structuur benadrukken die ten grondslag ligt aan de meeste irreversibele MCMC-algoritmen die in de literatuur voorkomen. De rest van de thesis is verdeeld in twee delen, waarbij respectievelijk *benaderingen* en *transformaties* van PDMC-algoritmen worden behandeld.

In het eerste deel introduceren we verschillende discretisatieschema's die een gegeven PDMP benaderen en bestuderen we de eigenschappen van de voorgestelde algoritmen in detail. Dit onderzoeksgebied is van fundamenteel belang om PDMP's breed toepasbaar te maken, omdat de PDMP's die in de MCMC-literatuur worden beschouwd meestal niet exact kunnen worden gesimuleerd vanwege ofwel gecompliceerde deterministische dynamica of omdat de willekeurige gebeurtenistijden verdeeld zijn volgens een exponentiële verdeling met *niet-homogene snelheid*. In het laatste geval zijn bestaande benaderingen om de willekeurige gebeurtenistijden te simuleren uitsluitend toepasbaar wanneer de snelheid van een eenvoudige vorm is, een vereiste die alleen van toepassing is op modelvoorbeelden uit de MCMC-literatuur. In deze thesis introduceren en bestuderen we een grote verscheidenheid aan tijdsdiscretisaties van PDMP's van elke orde van nauwkeurigheid, die nu als basis kunnen dienen voor MCMC-algoritmen. We bestuderen twee soorten discretisaties: de eerste soort wordt verkregen door het principe achter de klassieke *Eulermethodes* te generaliseren, terwijl de tweede gebaseerd is op *splitsingsmethodes*. In beide gevallen stellen we de afhankelijkheid van de fout van de stapgrootte van de discretisatie vast. Voor geschikte Eulermethodes bewijzen we *uniform-in-tijd schattingen* van de zwakke fout, een bijzonder uitdagend resultaat dat aangeeft dat de fout volledig wordt bepaald door de stapgrootte en niet afhangt van de tijdsduur. Bovendien leiden we voor benaderingen van PDMP's verkregen met op Euler gebaseerde methodes foutgrenzen af in de Wasserstein- en totale variatie-afstand met behulp van de koppelingsaanpak. Voor onze benaderingen op basis van splitsingsmethodes richten we ons voornamelijk op de Zig-Zag sampler (ZZS) en de Bouncy Particle Sampler (BPS) en bestuderen we de beste splitsingmethode wat betreft bias in de stationaire verdeling. Voor beide samplers leiden we voorwaarden af die de aanwezigheid en uniciteit van een stationaire verdeling voor het benaderingsproces garanderen, evenals exponentiële convergentie naar deze verdeling. Belangrijk is dat we aantonen dat symmetrische splitsingsmethoden van de tweede orde zijn, hoewel ze slechts één berekening van de gradiënt van de negatieve logaritmische aannemelijkheidsfunctie per iteratie vereisen. Een andere belangrijke noviteit die we introduceren is de mogelijkheid om de geïntroduceerde bias te corrigeren via een scheef-reversibele Metropolis-Hastings acceptatie-weigeringsstap. Hierdoor kunnen we de eerste unbiased, op PDMP's gebaseerde MCMC-algoritmen ontwerpen die moeiteloos kunnen worden toegepast om te sampelen uit elke gewenste

kansverdeling. Onze numerieke experimenten tonen aan dat de opmerkelijke eigenschappen van PDMP's ervoor zorgen dat hun benaderingen uitstekende convergentie-eigenschappen hebben die beter zijn dan benchmarkmethoden zoals Hamiltonian Monte Carlo en het onaangepaste Langevin-algoritme.

Het tweede deel van de scriptie gaat over transformaties van PDMP's. Allereerst bespreken we *ruimtetransformaties* van PDMP's, waarbij het belangrijkste doel is om de prestaties van PDMC-algoritmen te verbeteren wanneer de gewenste verdeling π anisotroop is. Ons voorstel is om adaptieve PDMC-algoritmen te ontwerpen die de covariantiestructuur van π *adaptief* leren en deze informatie gebruiken om de snelheid van de onderliggende PDMP, d.w.z. de richtingen die de PDMP waarschijnlijk zal verkennen, af te stemmen. Het vinden van een goede set aan richtingen vereist kennis van π, en daarom moet informatie over eerdere posities van het proces worden gebruikt. Op een vergelijkbare manier introduceren we adaptieve PDMC-algoritmen die automatisch de verversingssnelheid van het proces, d.w.z. de frequentie waarmee de huidige snelheidsvector wordt vervangen door een onafhankelijke trekking uit een geschikte verdeling, afstemmen. Voor deze algoritmen bestuderen we zorgvuldig de convergentie naar de gewenste verdeling door ergodiciteit vast te stellen, wat uitdagend is voor dergelijke niet-homogene markovprocessen. Bovendien testen we onze algoritmen op enkele benchmarkvoorbeelden, waarbij we relevante verbeteringen ten opzichte van de standaard, niet-adaptieve samplers zien. In het laatste hoofdstuk van de scriptie bekijken we *tijdstransformaties* van (stuksgewijs deterministische) markovprocessen, met de nadruk op het verbeteren van de convergentie van MCMC-algoritmen. In het bijzonder bestuderen we het effect van een verandering in de tijdschaal op de eigenschappen van een markovproces, waarbij veranderingen in de tijdschaal afhankelijk zijn van de toestand van het proces. Dit kan nuttig zijn bij multimodale kansverdelingen, waarbij we betogen dat de communicatie tussen verschillende modi kan worden verbeterd door de snelheid van de tijd te verhogen wanneer het proces zich in gebieden met lage dichtheid bevindt. We leggen verbanden tussen verschillende eigenschappen van een proces en die van een gerelateerd tijdsgetransformeerd proces, zoals een verband tussen de stationaire verdelingen, de generatoren, niet-explosiviteit, ergodiciteit en snelheid van convergentie naar de limietverdeling. Voor PDMP's laten we zien dat geschikte tijdstransformaties van een geometrisch ergodisch markovproces een *uniform ergodisch* proces kunnen maken, een opmerkelijke eigenschap die inhoudt dat de initialisatie van het proces geen invloed heeft op de snelheid van convergentie. We passen onze stelling toe op tijdstransformaties van het Zig-Zag proces en tonen daarmee de toepasbaarheid van onze voorwaarden aan. Door dit kader toe te passen op PDMP's definiëren we verschillende nieuwe processen wier dynamica afhankelijk zijn van een door de gebruiker gekozen, interpreteerbare snelheidsfunctie.

Acknowledgements

I would like to express my gratitude to Joris Bierkens and Geurt Jongbloed for giving me the opportunity to pursue a PhD and for the guidance throughout this process in which I learnt a lot and grew as a person.

I would also like to thank the external members of the committee, Alain Durmus, Jeff Rosenthal, Gareth Roberts, Frank van der Meulen, Frank Redig, and Aad van der Vaart for taking the time to read this thesis and for participating to the ceremony. I am honoured to have you in my committee.

My sincere thanks go to my collaborators, Paul Dobson and Pierre Monmarché. Paul, it was a real pleasure to work together in front of a whiteboard, on Zoom, in Delft, in Edinburgh, in Austria, talking about couplings, weak errors, imaging experiments... I really hope many more (mathematical) adventures are yet to come.

The past four years have been unforgettable, thanks to the wonderful people I have had the pleasure of meeting and spending time with. I will not forget the coffee breaks and beers, the jokes, chats and rants about work, the great times at the office when at Van Mourik Broekmanweg 6, the epic foosball and beach volleyball games, all the hours spent brainstorming and organising events for our beloved PhD Forum, the awesome trips to conferences with skiing, swimming in the sea, hiking, dinners together, etc. Thank you all for the amazing memories!

My thanks also go to Marco Loog for introducing me to the world of research but also for the occasional confrontational questions about personal life decisions.

Michelle (and Annabelle), thank you for the essential help in the design of the cover.

I am thankful to the friends with whom I could forget about work and have a great time, Dario, Edo, Fra, Jack, Simo, and Taiyo.

To my family, thank you for your support.

Merel, it is hard to measure how much time you spent listening to my complaining before and during the writing of this thesis. Thank you for being there for me all the time and for taking good care of me. I am so glad I peeked at your answers in that infamous questionnaire, it was the best decision I have ever made.

Curriculum Vitae

Andrea Bertazzi was born in Milano, Italy, on January 31st 1995. He obtained his diploma of Liceo Scientifico at Istituto Gonzaga Milano in 2013. His career in mathematics started at Politecnico di Milano, where he received a Bachelor of Science in Mathematical Engineering in July 2016.

Afterwards, Andrea moved to the Netherlands to pursue a Master of Science in Applied Mathematics at TU Delft. He received his diploma in 2018 with a thesis on semi-supervised learning under the supervision of prof. Marco Loog. This great experience motivated him to prolong his stay in Delft and pursue a PhD in computational statistics supervised by dr. Joris Bierkens. Andrea's PhD, which started in 2019 and ended in 2023, was funded by the NWO Vidi grant "Zigzagging through computational barriers" with project number 016.Vidi.189.043.

From September 2023, Andrea will join École Polytechnique in Paris as a postdoctoral researcher under the supervision of prof. Éric Moulines, funded by an ERC grant.

Publications

Published

A. Bertazzi and J. Bierkens. Adaptive schemes for piecewise deterministic Monte Carlo algorithms. *Bernoulli*, 28(4):2404 – 2430, 2022

A. Bertazzi, J. Bierkens, and P. Dobson. Approximations of Piecewise Deterministic Markov Processes and their convergence properties. *Stochastic Processes and their Applications*, 154:91–153, 2022

Submitted

A. Bertazzi, P. Dobson, and P. Monmarché. Splitting schemes for second order approximations of piecewise-deterministic Markov processes. *arXiv.2301.02537*, 2023

In preparation

A. Bertazzi. Time transformations of piecewise-deterministic Markov processes. *In preparation*, 2023