Design and applications of Monte Carlo methods based on piecewise deterministic Markov processes

Grazzi, S.

DOI
10.4233/uuid:a555e23d-708d-4bd1-94ca-7a6906313ef2

Publication date
2023

Document Version
Final published version

Citation (APA)
Grazzi, S. (2023). *Design and applications of Monte Carlo methods based on piecewise deterministic Markov processes*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:a555e23d-708d-4bd1-94ca-7a6906313ef2

Design and applications of Monte Carlo methods based on piecewise deterministic Markov processes

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
Thursday, the 2nd of March 2023 at 3:00 p.m.
by

Sebastiano GRAZZI,
Master of Science in Quantitative Finance,
University of Bologna, Italy
born in Cattolica, Italy

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	*chairperson.*
Prof. dr. ir. F.H. van der Meulen|,	Department of Mathematics, Vrije Universiteit Amsterdam, *promotor.*
Dr. ir. G.N.J.C. Bierkens,	Delft Institute of Applied Mathematics (DIAM), Delft University of Technology, *co-promotor.*

Independent members:

Prof. dr. A.W. van der Vaart,	Delft Institute of Applied Mathematics (DIAM), Delft University of Technology.
Prof. dr. A. Beskos,	Department of Statistical Science, University College London.
Prof. dr. P. Fearnhead,	Department of Mathematics and Statistics, Lancaster University.
Prof. dr. M.R.H. Mandjes,	Korteweg-de Vries Institute for Mathematics, University of Amsterdam.
Prof. dr. ir. G.J. Jongbloed,	Delft Institute of Applied Mathematics (DIAM), Delft University of Technology, *reserve member.*

Other member:

Dr. M. Schauer,	Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, *external collaborator.*

Dedication

Although this thesis is, to a certain extent, just a collection of articles written during my Ph.D. studies at TU Delft, it holds immense personal value because it gathers the greatest efforts and the most important professional achievements of my life to date.

My research was primarily conducted individually, but the most inspiring and valuable moments that gave me the fuel and motivation to work persistently were those shared with others, such as my family, friends and colleagues. I dedicate this thesis to them.

First of all, I would like to thank the Dutch Research Council (NWO) and TU Delft for funding this Ph.D. I am very grateful to the whole group in Statistics and Applied Probability who created a positive and lively working environment. A special mention of appreciation goes to Geurt Jongbloed, whose daily energy and enthusiasm always lifted my spirits.

On a more personal level, I would like to thank my supervisors Frank van der Meulen and Joris Bierkens, who believed in me and guided me with exceptional expertise and professionalism. I extend my heartfelt gratitude to Moritz Schauer, a key member of my research team, who, without any obligation on his part, spent countless hours with me, filling the shoes of both a fantastic supervisor and a very dear friend: Moritz, your passion for research, your ingenious ways of looking at problems, your vision and your creativity inspired many of the ideas that shaped this thesis. Thank you for your invaluable contributions.

As life is not only made by work, these last 4 years have been made sweeter and more joyful by Matteo, Sid, Serena, Alessandro and Ardjen, my flatmates and close friends with whom I have shared almost all my time outside work. The bond we forged during this time is tight and unbreakable.

Another very important friend who made my weekends much more exciting is Dirk-Jan, with whom I share an unbounded passion for sailing: thank you DJ for letting me in your crew.

A big thank you also goes to my friends and colleagues Marc, Bart, Andrea, Lasse, Paul, Simone, Martina, Giorgio and Tomas.

Finally, but most importantly, this Ph.D. would not have been possible without Alèxia, my parents Giulia and Paolo and my siblings Isotta, Emilio, Niccolò and Federico, who gave me unconditional support at all times.

Thank you for making this journey so special!

Sebastiano, March 2023

***Cover page**: the figure in the cover is a stylised phase space plot of a sample realisation of the Zig-Zag sampler featuring boundary conditions as detailed in Chapter 1.5.1*

This page is intentionally left blank.

Contents

List of Symbols

$(X_i)_{i=1,2,\ldots,N}$	ordered set with elements $X_1, X_2, \ldots, X_N$
$\{X_i\}_{i=1,2,\ldots,N}$	unordered set with elements $X_1, X_2, \ldots, X_N$
i.i.d.	'independent and identically distributed'
$\xrightarrow{\text{d}}$	convergence in distribution
$\xrightarrow{\text{a.s.}}$	convergence almost surely
$\mu(f^\ell)$	$\int f^\ell \mathrm{d}\mu$, $f\colon \mathcal{X} \to \mathbb{R}$, $(\mathcal{X}, \mathcal{A}, \mu)$, $\ell \geq 1$.
σ_f^2	$\mu(f^2) - \mu(f)^2$
$\sim$	'with distribution'
$\approx$	'approximately equal to'
$\mathcal{B}(E)$	Borel σ-algebra on E
$\mathcal{M}(E)$	Borel measurable functions $f\colon E \to \mathbb{R}$
$C(E)$	$\{f \in \mathcal{M}(E)\colon f \text{ is continuous}\}$
$C_b(E)$	$\{f \in C(E)\colon f \text{ is bounded}\}$
$C_o(E)$	$\{f \in C(E)\colon f \text{ vanishes at infinity}\}$
$C^i(E)$	$\{f\colon E \to \mathbb{R}\colon\ f \text{ is } i\text{th times continuously differentiable}\}$
$\mathcal{L}(X)$	distribution (or law) of X
$\mathcal{O}(f(n))$	'order $f(n)$'
$\mathcal{N}_d(c, \Sigma)$	d-dimensional Normal distribution centered at $c \in \mathbb{R}^d$ and with covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$
$\mathcal{N}(c, \sigma^2)$	1-dimensional Normal distribution centered at $c \in \mathbb{R}$ and with variance $\sigma^2 \in \mathbb{R}^+$
$\text{Unif}(A)$	Uniform random variable on the set A
$\text{Exp}(\lambda)$	Exponential random variable with rate λ
$\text{IPP}(f)$	First event time of an inhomogeneous Poisson process with rate $f\colon \mathbb{R}^+ \to \mathbb{R}^+$ (equation (4.7), page 81)
$\delta_x(\cdot)$	Dirac measure at x.
$\mathbb{E}(\cdot)$, $\mathbb{V}(\cdot)$	Expectation and variance
$y = x[A : c]$	$y_i = c$ if $i \in A$, $y_i = x_i$ otherwise. Here $y, x \in \mathbb{R}^d$, $c \in \mathbb{R}$ and $A \subset \{1, 2, \ldots, d\}$.

Chapter 1

Introduction

We investigate some aspects of the design of piecewise deterministic Markov processes (PDMPs) for Monte Carlo simulation in Bayesian inference problems. The aim of this chapter is to present the theory which motivates the study of PDMPs and to highlight our contributions. In this chapter we show that PDMPs naturally arise as limits of discrete time Markov chains and share remarkable properties which are desirable for Monte Carlo simulations. The illustration in Section 1.5.1 summarises many of the innovations. The presentation of the background material is largely inspired by Roberts and Rosenthal (2004), Geyer (1998), Diaconis, Holmes, and Neal (2000) and Rosenthal (2003).

1.1 Motivation

For a given probabilty measure μ on a measurable space $\mathcal{X}$, consider the problem of computing expectations

$$\mu(f) := \int_{\mathcal{X}} f(x)\mu(\mathrm{d}x) \tag{1.1}$$

for functions $f\colon \mathcal{X} \mapsto \mathbb{R}$ for which $\mu(f) < \infty$. This problem is commonly encountered in *Bayesian inference* where X is a random variable that corresponds to the unknown parameter in a statistical model and takes values on a parameter space $\mathcal{X}$. The Bayesian paradigm assigns to X a *posterior* measure μ on $\mathcal{X}$. The measure μ depends on the data, through a log-likelihood $\ell(x)$ (with this notation, we omit the dependence of ℓ on the data) and a prior measure μ_0, which takes into account information on model parameter prior to the data, such as the set of values that X can take, sparsity and smoothness assumptions; the relation between μ, μ_0 and ℓ is obtained by applying the *Bayes formula* and here is assumed to be

$$\mu(\mathrm{d}x) = C \exp(\ell(x))\mu_0(\mathrm{d}x), \tag{1.2}$$

with $C = (\int_{\mathcal{X}} \exp(\ell(x))\mu_0(\mathrm{d}x))^{-1}$ being the constant of normalization which does not depend on x. The Bayesian inferential procedure requires point estimates which are of the form of equation (1.1), e.g. the posterior mean (taking $f(x) = x_i$ for $i = 1, 2, \dots$) and posterior probabilities of a given set $A \in \mathcal{B}(\mathcal{X})$ (with $f(x) = \mathbf{1}_A(x)$, $\mathbf{1}$ being the indicator function). It is often the case that $\mu(f)$ cannot be derived analytically, hence numerical techniques must be employed. This is a frequently encountered situation in Bayesian inference where, in most cases, the constant C on the right hand-side of equation (1.2) cannot be computed analytically. It becomes now apparent that Bayesian inference problems are often strongly related to the problem of numerical integration. Throughout the thesis, we often refer to μ as the *target measure*, which is a general probability measure from which we want to estimate the right hand-side in equation (1.1) and we often do not make any explicit connection with the posterior distribution in the Bayesian framework.

It is natural to think of applying ordinary numerical integration methods, for example approximating μ by evaluating μ with a quadrature rule on a finite partition with elements in $\mathcal{X}$ (see for example Thisted 1988, Chapter 5 for details). This approach has several caveats and most notably it suffers from the curse of dimensionality, i.e. if $\mathcal{X} \subset \mathbb{R}^d$, the number of discretization points needed for a given numerical precision grows exponentially with the dimensions d.

Popular alternatives fall within the name of *Monte Carlo methods*, where Monte Carlo refers to integration methods involving the simulation of random variables.

Remark 1.1.1. *There are other important methods used for estimating $\mu(f)$ which are not treated in this thesis. Most notably, methods based on* importance sampling *(see e.g. Robert and Casella 1999, Section 3),* Sequential Monte Carlo *methods (see e.g. Doucet, De Freitas, Gordon, et al. 2001) and* Quasi-Monte Carlo *methods (see e.g. Niederreiter 1992).*

1.2 Markov chain Monte Carlo

Monte Carlo methods are integration methods which are based on the simulation of a sequence of random variables $X_1, X_2, \dots$ jointly defined on a arbitrary probability space and each one taking values on $\mathcal{X}$. For a fixed $N > 0$, define sample averages as

$$\hat{\mu}_N(f) := \frac{1}{N} \sum_{i=1}^{N} f(X_i), \tag{1.3}$$

for functions $f \colon \mathcal{X} \to \mathbb{R}$. Monte Carlo methods are devised such that the law of large numbers and the central limit theorem holds as follow.

Definition 1.2.1. (Law of Large Numbers (LLN)) *For a fixed function $f \colon \mathcal{X} \to \mathbb{R}$ and a measure μ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, with $\mu(f) < \infty$, a sequence of random variables*

$X_1, X_2, \ldots$ satisfies the Law of Large Numbers (LLN) if

$$\hat{\mu}_N(f) \xrightarrow{a.s} \mu(f), \qquad \text{as } N \to \infty. \tag{1.4}$$

Definition 1.2.2. (Central Limit Theorem) *For a fixed function $f\colon \mathcal{X} \to \mathbb{R}$ and a measure μ such that $\sigma_f^2 = \mu(f^2) - \mu(f)^2 < \infty$, a sequence of random variables $X_1, X_2, \ldots$ satisfies the Central Limit Theorem (CLT) if*

$$\sqrt{N}(\hat{\mu}_N(f) - \mu(f)) \xrightarrow{d} \mathcal{N}(0, \sigma_f^2(\infty)) \qquad \text{as } N \to \infty, \tag{1.5}$$

for some $0 < \sigma_f^2(\infty) < \infty$.

If $(X_i)_{i=1,2,\ldots,N}$ can be simulated by means of a computer, then, by the LLN, we can use the sample average $\hat{\mu}_N(f)$ as an (asymptotically) unbiased estimator for $\mu(f)$ and we can use the CLT and choose N appropriately to control the statistical error between $\hat{\mu}_N(f)$ and $\mu(f)$. This is the core idea of Monte Carlo methods. We are now ready to give the first concrete example of a sequence of random variables satisfying CLT and LLN:

Example 1.2.3. (A sequence of i.i.d random variables) *Consider a sequence of independent and identically distributed (i.i.d) $\mathcal{X}$-valued random variables $X_1, X_2, \ldots$, each distributed according to μ. Provided that $\sigma_f^2 < \infty$, the sequence trivially satisfies both the LLN and CLT with $\sigma_f^2(\infty) = \sigma_f^2$.*

Example 1.2.3 is very attractive in principle but often not applicable as it is rarely possible to directly simulate a sequence of mutually independent random variables with a given distribution μ. Hence, we now weaken the conditions made in Example 1.2.3 by considering a homogeneous Markov Chain $(X_i)_{i=1,2,\ldots}$ (allowing now for a dependent sequence of random variables). A Markov chain is completely characterized by the distribution of its initial value X_1 and the transition kernel $\mathcal{Q}\colon \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$, where, for every $x \in \mathcal{X}$, $\mathcal{Q}(x, \cdot)$ is a probability measure on $\mathcal{X}$ with $\mathcal{Q}(x, A)$ being the probability to jump from x to the set $A \in \mathcal{B}(\mathcal{X})$ in one step. These two components also gives a simple recipe to simulate a trajectory: simulate the first random variable $x_1 \sim \mathcal{L}(X_1)$ and iteratively $x_{i+1} \sim \mathcal{Q}(x_i, \cdot)$. Define the n-step transition kernel $\mathcal{Q}^n(x, \cdot) = \mathbb{P}(X_n \in \cdot \mid X_0 = x)$ and the total variation distance between two measures μ and ν on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ as

$$\|\mu - \nu\|_{TV} := \max_{A \in \mathcal{B}(\mathcal{X})} |\mu(A) - \nu(A)|.$$

In the context of MCMC methods, it is customary to consider Markov chains which are *geometrically ergodic*.

Definition 1.2.4. (Geometrically ergodic Markov chains) *A Markov chain is gemoterically ergodic if*

$$\|\mathcal{Q}^n(x,\cdot)-\mu\|_{TV} \leq M(x)\rho^n, \quad \mu-a.s.\ x\in\mathcal{X} \tag{1.6}$$

for some $\rho\in[0,1)$ *and with* $\sup_x M(x)<\infty$.

Intuitively, this condition implies that the marginal distribution of the process convergences to the target distribution as $N\to\infty$, so a straightforward implication is that μ must be the unique stationary distribution of the process. The reason for such customary assumption becomes evident with the following proposition (Ibragimov and Linnik 1971, Theorem 18.5.3):

Proposition 1.2.5. (CLT for geometrically ergodic Markov chains) *A geometrically ergodic Markov chain satisfies the CLT (equation (1.5)) whenever* $\mu(|f|^{2+\delta})<\infty$, *for some* $\delta>0$.

The connection between geometrically ergodic Markov chains and the CLT is rather subtle and far from trivial, as the former result involves marginal measures of a Markov chain, while the latter is a result of the whole chain, see Geyer (1998, Chapter 4.1.) for a discussion and other similar results. An important research area in MCMC is to assess the quantitative convergence of the marginal distributions of Markov chains (see Roberts and Rosenthal 2004)). This analysis serves in practise as a guidance to choose the *burn-in time* which corresponds to the number of iterations needed for the Markov chain to reach its stationary distribution and is used in practise to exclude these first iterations when computing the estimator of equation (1.3).

It remains unclear how to devise a geometrically ergodic Markov chain. A necessary condition is that the transition kernel $\mathcal{Q}$ of the Markov chain is invariant to a target μ, that is

$$\int_{x\in\mathcal{X}} \mu(\mathrm{d}x)\mathcal{Q}(x,\mathrm{d}y)=\mu(\mathrm{d}y).$$

We now introduce a simple recipe to construct a chain which is invariant to a target μ which is based on the concept of *reversibility*:

Definition 1.2.6. (Reversible Markov chains) *A Markov chain is* μ*-reversible if*

$$\mu(\mathrm{d}x)\mathcal{Q}(x,\mathrm{d}y)=\mu(\mathrm{d}y)\mathcal{Q}(y,\mathrm{d}x) \tag{1.7}$$

Equation (1.7) is referred as the *detailed balance condition* and can be informally interpreted as follows: the probability for the process to be at A and move to B, for any two regions $A,B\in\mathcal{B}(\mathcal{X})$, is equal to the probability to be at B and move to A.

Proposition 1.2.7. *A* μ*-reversible Markov chain is* $\mu-$*invariant.*

This is straightforward as

$$\int_{x\in\mathcal{X}} \mu(\mathrm{d}x)Q(x,\mathrm{d}y) = \int_{x\in\mathcal{X}} \mu(\mathrm{d}y)Q(y,\mathrm{d}x) = \mu(\mathrm{d}y).$$

Hence, by devising Markov chains satisfying the detailed balance condition as in equation (1.7), we automatically know that the process preserves the measure μ. This is the starting point for many popular MCMC methods, most notably the celebrated (and well studied) Metropolis-Hasting algorithm (Metropolis et al. 1953, Hastings 1970) which turns out to be geometrically ergodic for targets μ with exponentially light tails (Jarner and Hansen 2000). Reversible Markov chains also offer a simplified theoretical analysis as $\mathcal{Q}$ in this case is a self-adjoint operator in the Hilbert space $L_0^2(\mu) = \{f\colon \mathcal{X} \to \mathbb{R} \mid \pi(f^2) < \infty \text{ and } \pi(f) = 0\}$ with inner product $(f,g) = \int fg\mathrm{d}\mu$ and with a real spectrum which can be used to estimate the asymptotic variance $\sigma^2(\infty)$ in the CLT and the convergence rates of geometrically ergodic chains (right hand-side of equation (1.6)). See Rosenthal (2003, Section 3) for details. The point here is that Monte Carlo methods based on reversible Markov chains are simple to devise and provide a simplified analysis, which explains why they became so popular. However, in the past two decades it has been noticed (for example in Diaconis, Holmes, and Neal (2000)) that the detailed balance condition introduces "diffusive behaviour" of the underline process. Here, a process with 'diffusive behaviour' refers to a process which locally resembles a *Random Walk* and it requires a number of iterations of $\mathcal{O}(N^2)$ in order to cross regions with distance of $\mathcal{O}(N)$ (see Diaconis, Holmes, and Neal 2000, Section 1 for details).

As this was considered sub-optimal, a new class of more sophisticated Markov chains referred as *non-reversible* (as opposed to the reversible) was proposed and analysed; a far from exhaustive list of references is Geyer and Mira (2000), Diaconis, Holmes, and Neal (2000), Chen and Hwang (2013), Bierkens and Roberts (2017), Andrieu and Livingstone (2019). This new class of Markov chains does not exhibit diffusive behaviour and satisfies the *skew-detailed balanced condition:*

Definition 1.2.8. (Skew-detailed balanced condition) *For an involution $\mathcal{S}: \mathcal{X} \to \mathcal{X}$ and a measure μ on $\mathcal{X}$, such that $\mathcal{S} \circ \mathcal{S} = I$ and $\mu(\mathcal{S}^{-1}(\mathrm{d}x)) = \mu(\mathrm{d}x)$, a Markov chain satisfies the skew-detailed balance condition relative to a measure μ, if*

$$\mathcal{Q}(x,\mathrm{d}y)\mu(\mathrm{d}x) = \mathcal{Q}(\mathcal{S}^{-1}(y),\mathcal{S}^{-1}(\mathrm{d}x))\mu(\mathrm{d}y)$$

Proposition 1.2.9. *A Markov chain satisfying the skew-detailed balance condition relative to μ is μ-invariant.*

The skew-detailed balance condition reduces to the detailed balance condition upon taking $\mathcal{S} = I$. However, the non-reversible Markov chains considered here are defined on an augmented space of position x and velocity v and the involution

considered in the literature is $\mathcal{S}(x, v) = \{(x, -v)\}$ which can be understood as a time reversal operator. With this perspective, the skew-detailed balance condition can be seen as a detailed balance condition with time reversal dynamics (for any two sets $A, B \in \mathcal{B}(\mathcal{X})$, the probability of the process to be in A and jump to B is equivalent to the probability to be in B and jump in A for the time-reverted process).

Next, we look at a toy model which captures the fundamental differences between reversible and non-reversible chains and anticipated the use of piecewise deterministic Markov processes for Monte Carlo sampling.

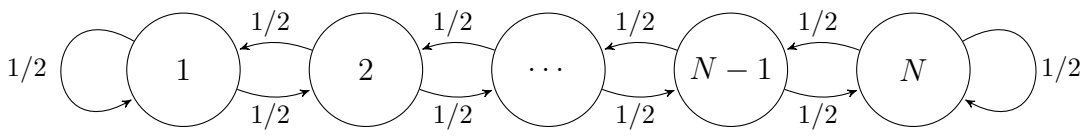
1.3 Random walks on a finite state space, a case study

In this section we recall a simple example which was initially analyzed in Diaconis, Holmes, and Neal (2000). This example is a pioneering work containing the fundamental idea which motivated the study and use of piecewise deterministic Markov processes for Monte Carlo methods, that is *lifting* the state space of the underlying Markov chain with a velocity component and breaking the *detailed balance condition*. When compared with their reversible counterpart, the lifted Markov chains are shown to

- converge faster to their invariant measure in terms of the total variation distance between the marginal measures of the chain at any iteration $n > 1$ and the target measure (Diaconis, Holmes, and Neal 2000);
- reduce the asymptotic variance $\sigma^2(\infty)$ of the CLT (Chen and Hwang 2013).

1.3.1 Random walks

Consider a *Random walk* $(X_i)_{i=1,2,\dots}$ on the space $\mathcal{X} := \{1, 2, \dots, N\}$ i.e. a Markov chain which starts at a given point $x_1 \in \mathcal{X}$ and with transition probabilities $\mathcal{Q}(x, x \pm 1) = \frac{1}{2}$ for all $x \in \mathcal{X} \setminus \{1, N\}$ and $\mathcal{Q}(1, 1) = \mathcal{Q}(1, 2) = \mathcal{Q}(N, N) = \mathcal{Q}(N, N-1) = \frac{1}{2}$ as illustrated with this graph:

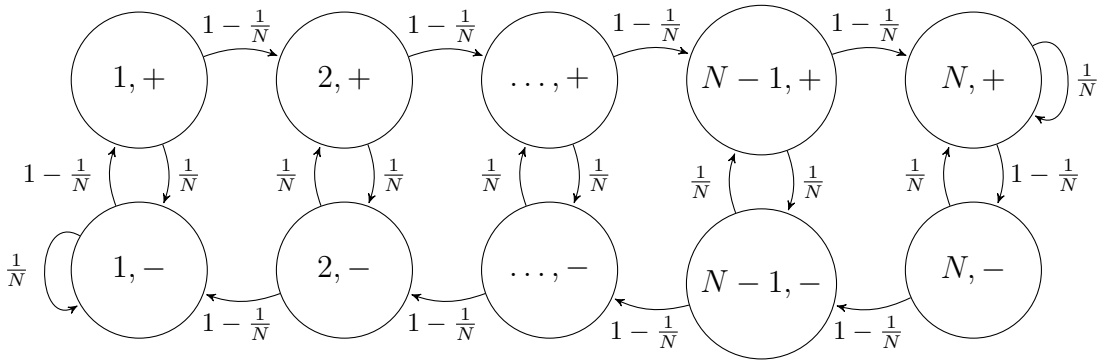


The Markov Chain satisfies the detailed balance condition relative to the measure $\mu = \text{Unif}(\mathcal{X})$ and it is geometrically ergodic (this can be easily checked as $\mathcal{X}$ is finite, see Roberts and Rosenthal 2004, Section 3.4). The symmetry in the detailed balance condition manifests itself on the trace of the chain with the process having

a tendency of backtracking and exploring the state space in a diffusive manner (see Figure 1.1, left panel). Ignoring the boundaries, by the central limit theorem, X_n is approximately a Gaussian random variable centered at x_1 and with variance equal to n. Hence $\mathbb{P}(|X_n - x_1| > c) = \mathcal{O}(\sqrt{n})$, for any $c > 0$. This heuristically implies that the process takes $\mathcal{O}(N^2)$ iterations to explore the full state space $\mathcal{X}$ which has size $|\mathcal{X}| = N$. Formal and quantitative results estimating convergence results of this chain may be found in Levin and Peres 2017, Example 12.3.1 and Example 12.11 for a Markov chain which is topologically equivalent to the one considered here.

1.3.2 Lifted random walks

Consider the Markov Chain taking values on a lifted state space $\mathcal{X} \times \{-, +\}$, where the first component is intended as the position and the second the velocity of the chain. The transition probabilities are shown by the following graph:



Essentially, at every iteration, the chain keeps moving in the same direction with high probability and switches direction with small probability or when hitting the boundary $\{(1, -), (N, +)\}$. One can check that the transition kernel $\mathcal{Q}$ satisfies the skew-detailed balance condition relative to the measure $\pi \otimes \rho$ with $\pi = \text{Unif}(\mathcal{X})$ and $\rho = \text{Unif}(\{-, +\})$, with involution $\mathcal{S}(x, \pm) = \mathcal{S}(x, \mp)$.

In contrast to the the Random Walk presented in Section 1.3.1, the lifted random walk does not satisfy the detailed balance condition and, most importantly, it is shown in Diaconis, Holmes, and Neal 2000 that it converges in total variation in $\mathcal{O}(N)$ iterations (as opposed to $\mathcal{O}(N^2)$ of its reversible counterpart, see Diaconis, Holmes, and Neal 2000, Theorem 1 for details). The faster exploration of the lifted random walk compared to the ordinary random walk is visible in Figure 1.1.

1.3.3 PDMPs as limits of lifted random walks

Consider the lifted random walk on the state space $\mathcal{X}^N \times \{+, -\}$ with

$$\mathcal{X}^N := \{\frac{1}{N}, \frac{2}{N}, \ldots, 1\},\ N > 1.$$

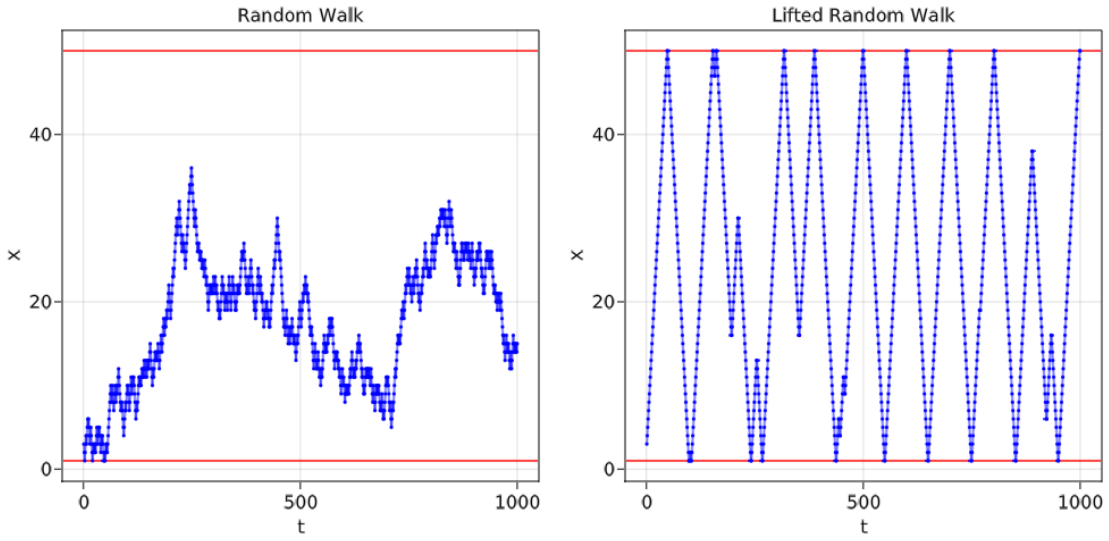


Figure 1.1: 1000 iterations of the random walk (left) and lifted random walk (right) on the space $\{1, 2, \ldots, 50\}$ and with initial position $x = 3$ ($x = 3$, $v = +$ respectively). Both chains are stationary to the measure $\text{Unif}(\mathcal{X})$.

The probability for this process to travel between two points $(x, y) \in \mathcal{X}^N$ with distance $c := |x - y|$ without changing its velocity component is $(1 - \frac{1}{N})^{cN}$. By taking the limit as N goes to infinity we have that

$$\lim_{N\to\infty} \left(1 - \frac{1}{N}\right)^{cN} = \exp(-c) = \mathbb{P}(Z > c) \text{ for } Z \sim \text{Exp}(1).$$

Heuristically, the limiting process for N going to infinity can be seen as a *continuous-time* Markov process with position and velocity components (X_t, V_t) which moves with piecewise constant velocity $V_t \in \{+1, -1\}$ in the space $[0, 1]$ and switches velocity sign at exponentially distributed times or when hitting the boundary $\{(0, -1), (1, +1)\}$. As such, the limit of the lifted random walk has piecewise-deterministic trajectory and a collection of random times which changes its velocity. Moreover, similarly as before, by the LLN and CLT, the estimator

$$\frac{1}{T}\int_0^T f(X_s)\mathrm{d}s$$

can be used to estimate expectations $\pi(f)$. Here T is the final clock of the process and $f \in \{g\colon \mathcal{X} \to \mathbb{R} \mid \pi(g) < \infty\}$. This is the core of PDMP samplers, which are continuous-time piecewise deterministic Markov processes on the augmented space of position and velocity, characterised by deterministic trajectories and a finite

collection of random times which act by changing the velocity component of process in order to target the correct distribution. A rigorous derivation of this heuristic limit is presented in Miclo and Monmarché (2013, Section 3). In a similar vein, Bierkens and Roberts (2017) derived more sophisticated PDMPs as limits of a more general class of 1 dimensional lifted Markov chains. For a general and detailed treatment of PDMPs, see Davis (1993).

The algorithms used to simulate PDMPs are fundamentally different from the algorithms used for the simulation of discrete-time Markov chains and in the literature (in particular in statistical mechanics, see Michel, Kapfer, and Krauth 2014) takes the name of *event-driven* algorithms, as they simulate and save only the state and the arrival time of the random events that modifies the deterministic dynamics (in this case only the exponentially distributed random times). The full trajectory can then be deterministically extrapolated from the saved events.

In the next section, we give a brief overview of the d-dimensional Zig-Zag sampler (Bierkens, Fearnhead, and Roberts 2019), a prominent and successful example of a PDMP sampler used for Monte Carlo integration. Although many results presented in this thesis are valid for general PDMP samplers, most of them are stated only for the Zig-Zag sampler and its extensions.

1.4 Standard d-dimensional Zig-Zag sampler

The standard d-dimensional Zig-Zag sampler is defined in the augmented space of position and velocity $\mathbb{R}^d \times \{-1, +1\}^d$ with elements denoted by $z = (x, v)$. The process dynamics are defined recursively and can be decomposed coordinate-wise. That is, for the process at time s and position $(X(s), V(s))$.

- Define the first random event time $\tau \geq s$ as the minimum of d random event times $\tau_1, \tau_2, \ldots, \tau_d$. The process moves with deterministic dynamics
$$(X(t), V(t)) = (X(s) + V(s)(t-s), V(s)) \quad \text{for } s \leq t < \tau.$$
- At time τ, the process changes its velocity by setting for $j = 1, 2, \ldots, d$,
$$V_j(\tau) = \begin{cases} -V_j(\tau-) & \text{if } j = i, \\ V_j(\tau-) & \text{otherwise,} \end{cases}$$
with $i = \arg\min_j(\tau_j)$.

It can be shown (see for example Bierkens, Fearnhead, and Roberts 2019) that if the inter-arrival times $(\tau_1 - s), \ldots, (\tau_d - s)$, are chosen to be the first event times of inhomogeneous Poisson processes with distribution
$$\mathbb{P}\left(\tau_i - s \geq t\right) = \exp\left(-\int_0^t \lambda(X(s) + V(s)z, V(s))\mathrm{d}z\right)$$

with

$$\lambda_i(x, v) = \max(0, v_i \partial_{x_i} \Psi(x)),$$

for a function $\Psi \in C^1(\mathbb{R}^d)$, then the process is invariant to the measure $\mu(\mathrm{d}x, \mathrm{d}v) = \pi(\mathrm{d}x) \otimes \rho(\mathrm{d}v)$, where $\rho = \mathrm{Unif}(\{-1, +1\}^d)$

$$\pi(\mathrm{d}x) = C \exp(-\Psi(x))\mathrm{d}x. \tag{1.8}$$

Under mild assumptions on Ψ, the Zig-Zag is geometrically ergodic (see Bierkens, Roberts, and Zitt 2019). Other theoretical results have been recently derived such as *i)* a spectral analysis of the generator of the process (Bierkens and Lunel 2022) which provides quantitative bounds on the convergence of the marginal measure of the process to the target; *ii)* a large deviation principle of the process characterizing the large deviations of the empirical measure of the process to its target (Bierkens, Nyquist, and Schlottke 2021); *iii)* diffusion limit results (Bierkens, Kamatani, and Roberts 2018) which show how the process scales (behaves) in high dimensions.

1.5 Contributions and outline of the thesis

1.5.1 Extensions of PDMPs for constrained spaces and discontinuous targets

PDMP samplers can be used for targeting a wide class of multi-dimensional measures. In this section, we give an example which informally illustrates the rich class of PDMPs considered and highlights some of the contributions of this thesis.

We run the Zig-Zag sampler featuring a rich behaviour given by random events of different nature. Figure 1.2 (cover of the thesis) displays the first two coordinates of a simulated trajectory. Below, we informally distinguish and comment on each of those random events and link them to their specific role in targeting an (artificially chosen) measure on $\mathbb{R}^d$, with piecewise-smooth density proportional to $\exp(\Psi(x))$ with

$$\Psi(x) = -x'\Gamma x + \sum_{i=1}^{d} \mathbf{1}_{(x_i > 1/2)} c + \sum_{i=1}^{\lfloor d/2 \rfloor} \log(x_{2i-1}) \tag{1.9}$$

relative to a reference measure

$$\prod_{i=1}^{d} \left(\mathbf{1}_{(x_i \in [0,1])} \mathrm{d}x_i + \delta_0(\mathrm{d}x_i - \frac{1}{4}) \right) \tag{1.10}$$

for a parameter $c > 0$ and a matrix $\Gamma = 1.3I + C0.5$, where each element $C_{i,j}$, $i, j = 1, 2, \ldots, d$ is 0 with probability 0.9 and an independent realization from $\mathcal{N}(0, 1)$ otherwise. For this simulation, we fixed d the dimensionality $d = 80$.

- *(Random events and repelling walls)* Analogously to the standard Zig-Zag samplers, the process changes direction at random times by switching every time the sign of only one velocity component as described in Section 1.4. This produces changes in direction and allows the process to target the smooth components of $\Psi(x)$ (the first and the third term of (1.9)).

  In this example, the position of the odd coordinates $\{(x_{i2-1}, v_{i2-1})\}_{i=1,2,\dots}$ can be arbitrarily close to 0, yet without ever touching 0 giving rise to repelling walls. This is because the density vanishes on those hyper-planes.

- *(Sticky floors)* All coordinates $\{(x_i, v_i)\}_{i=1,2,\dots}$, upon hitting $\frac{1}{4}$, "stick" in that point for an exponentially distributed time. This corresponds to momentarily setting the ith velocity component to 0 and allows the process to spend positive time in hyper-planes of the form

$$\bigotimes_{i=1}^{d} \{E_i \mid E_i = \{\frac{1}{4}\} \text{ or } E_i = \mathbb{R}\},$$

  (with some coefficients exactly equal to 1/4). Sticky floors allow the process to target mixtures of continuous and atomic components and, in this example, allow to change the reference measure from a d-dimensional Lebesgue measure to (1.10).

- *(Soft walls)* All coordinates $\{(x_i, v_i)\}_{i=1,2,\dots}$, upon hitting $\frac{1}{2}$ from below, switch their velocity with some probability. This allows the process to target densities which have discontinuities. In this example, the target density is discontinuous at $\frac{1}{2}$ in every component and the behaviour of the process at $\frac{1}{2}$ allows the process to target the second term in (1.9)).

- *(Hard walls)* The process switches always velocity at the boundaries $\{(x_{i2}, v_{i2}) = (0, -1)\}_{i=1,2,\dots}$ and $\{((x_i, v_i) = (1, +1)\}_{i=1,2,\dots}$. This allows the process to explore only the regions in $\mathbb{R}^d$ supported by the measure.

Each bullet point in this list will be formalized and described in details in the subsequent chapters.

This is a constructed example which is of interest for multiple reasons: *i)* an efficient and local implementation of the Zig-Zag sampler can be adopted which greatly profits of the local dependence structure of μ implied by the sparse form of Γ in (1.9) (see Chapter 2, Section 2.4 for more details); *ii)* the continuous and atomic components of the reference measure (equation (1.10)) makes the sampling problem not trivial. A mixture of atomic and continuous components arises naturally for example in Bayesian variable selection with spike-and-slab priors. By including sticky events, PDMPs can efficiently sample from such mixture measures, see Chapter 4

for more details; *iii)* As $\lim_{y \downarrow 0} \Psi(x[2i\colon y]) = \infty$, for $i = 1, 2, \ldots$, the gradient of the log-likelihood explodes thus complicating the application of gradient-based Markov chain Monte Carlo methods; *iv)* the discontinuities at $1/2$ and boundaries at 0 and 1 deteriorate the performance of ordinary MCMC (see Neal et al. 2011) and complicates the application of gradient-based methods, as the gradient is not defined at discontinuity. In Section 5 we give a simple framework to address piecewise smooth densities efficiently with PDMPs.

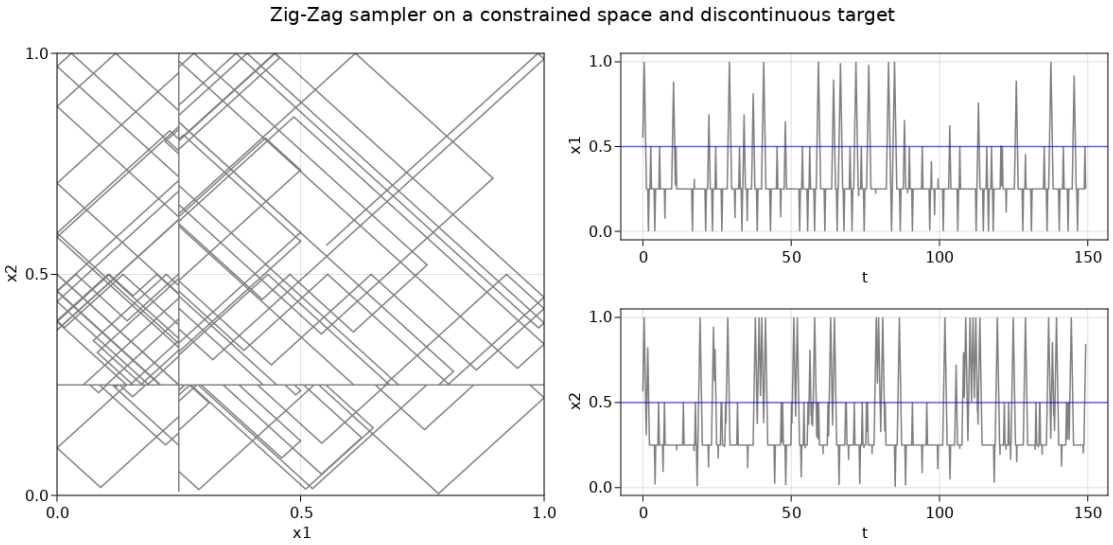


Figure 1.2: (x_1-x_2) phase space plot (left) and trace plots (right) of the first 2 coordinates of a Zig-Zag trajectory sampling a general density f supported in $[0,1]^d$ with discontinuity at $1/2$ (yellow line) in each coordinate with density vanishing at $x_2 = 0$. The reference measure has a Dirac mass at $1/4$ in each coordinate.

1.5.2 Overview of Chapter 2-3

In Chapter 2, we propose a method for sampling one-dimensional diffusion bridges were the diffusion is defined as a solution to the Itô stochastic differential equation

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \quad X_0 = x,\, X_T = y,\, t \in [0, T],$$

where $(W_t)_{t>0}$ is a standard one-dimensional Wiener process.

Similar to the Lévy-Ciesielski construction of a Brownian motion, we expand the diffusion path in a Faber-Schauder basis (see Figure 2.3 in Chapter 2). The coefficients within the basis are sampled using the Zig-Zag sampler, a particular PDMP sampler. A key innovation is the use of the *fully local* algorithm for the Zig-Zag sampler that allows to exploit the sparsity structure implied by the dependency

graph of the coefficients. Furthermore we use the *exact subsampling* technique to approximate the likelihood given by the Girsanov theorem which, in this setting, does not have a closed from. This differs from the standard use of the subsampling technique for PDMP samplers in regression problems for subsampling data points. We illustrate the performance of the proposed methods in a number of examples.

In Chapter 3, we introduce the Boomerang sampler as a novel class of PDMP samplers. The methodology begins by representing the target density as a density, e^{-U}, with respect to a prescribed (usually) Gaussian measure and constructs a continuous trajectory consisting of a piecewise elliptical path. The method moves from one elliptical orbit to another according to a rate function which can be written in terms of U. We demonstrate that the method is easy to implement and demonstrate empirically that it can outperform existing benchmark piecewise deterministic Markov processes such as the bouncy particle sampler and the Zig-Zag. We demonstrate theoretically and empirically that we can construct a control-variate subsampling boomerang sampler which is exact (i.e. target the correct distribution) and which possesses remarkable scaling properties in the large data limit.

The Boomerang sampler is particularly well suited to sample diffusion bridges with the methodology presented in Chapter 2. This is because the likelihood of a diffusion bridge is expressed as a density relative to a high dimensional Gaussian measure. As a key application, we illustrate a factorised version of the Boomerang sampler for the simulation of diffusion bridges and we highlight the advantages of this sampler compared to the standard Zig-Zag sampler, as used in Chapter 2.

1.5.3 Overview of Chapter 4

In chapter 4, we construct a new class of efficient Monte Carlo methods based on PDMPs suitable for inference in high dimensional sparse models, i.e. models for which there is prior knowledge that many coordinates are likely to be exactly 0. This is achieved with the fairly simple idea of endowing existing PDMP samplers with "sticky" coordinate axes, coordinate planes etc. Upon hitting those subspaces, an event is triggered during which the process *sticks* to the subspace, this way spending some time in a sub-model. This results in *non-reversible* jumps between different (sub-)models. While we show that PDMP samplers in general can be made sticky, we mainly focus on the Zig-Zag sampler. We show the method outperforms other existing methods by comparing scaling results of the algorithm and mixing times in relation with an established method for variable selection. The computational efficiency of our method (and implementation) is established through numerical experiments where both the sample size and the dimension of the parameter space are large.

1.5.4 Overview of Chapter 5

In Chapter 5 we include some results for the applications of PDMP samplers for discontinuous densities and constrained spaces (spaces with particular boundary conditions). We present the framework for using PDMPs in such setting and study two important applications:

- *(Spread of infectious diseases)* The model considered is known as SINR (susceptible-infected-notified-removed) and is used for modelling the spread of infectious disease in a population (see Jewell et al. 2009). The goal is to sample the posterior measure of infected times of a population of size N conditioned on the observation of the notification and removal times of population individuals up to a certain time horizon T. We combine the PDMP for piecewise smooth densities with the framework presented in Bierkens et al. (2023) for adding/removing efficiently in continuous time *occult infected individuals* (infected individuals which have not been detected up to the time T) by means of introducing sticky events which are events after which the process sticks to lower dimensional hyper-planes for some random time. In this case, the target density presents discontinuities relative to the order of the infection times, notification times and removal times of each individual and have a reference measure which is a mixture of Lebesgue and Dirac components.

- *(Hard-spheres with teleportation)* We consider a hard-sphere model in statistical mechanics, see Krauth (2006, Chapter 2) for an overview. We take N particles, each one taking values in $\mathbb{R}^d$. Denote the configuration of all particles by $x = \{x^{(i)} \in \mathbb{R}^d \colon 1 \leq i \leq N\}$ where we identify the ith particle as $x^{(i)} = x_{[(i-1)d+1,id]}$ and consider a given invariant measure $\mu^\star$ supported on $\mathbb{R}^{dN}$. We assume that each particle is a hard-sphere centered at $x^{(i)}$ with radius $r_i > 0$, $i = 1, 2, \ldots, N$ and consider the conditional invariant measure

$$\mu(\mathrm{d}x) \propto \mu^\star(\mathrm{d}x)\mathbf{1}_{x \in A}.$$

  with $A = \bigcap_{i=1}^{N} \bigcap_{\substack{j=1,2,\ldots,N, \\ j \neq i}} A_{i,j}$ and

$$A_{i,j} = \{x \in \mathbb{R}^{dN} \colon \|x^{(i)} - x^{(j)}\| \geq (r_i + r_j)\},$$

  that is, the measure $\mu^\star$ conditioned on the space where all hard-spheres do not overlap. The restriction for the process to be outside the region A creates boundaries which slow down the exploration of the state space. In order to enhance the exploration of the process, we modify the dynamics of PDMPs by introducing *teleportation* schemes allowing the process to make jumps in between boundaries of the space which are chosen conveniently in order to enhance the exploration of the state space. Teleportation schemes such as

the ones introduced here can be also used for applications where the target is supported on disconnected regions or distant regions which are difficult to reach with the standard PDMP dynamics which are continuous in space.

1.6 Publications and preprints

The results presented in Chapter 2 are joint work with Frank van der Meulen (TU Delft), Joris Bierkens (TU Delft) and Moritz Schauer (Chalmers University of Technology and University of Gothenburg) and are published as

> J. Bierkens, S. Grazzi, F. van der Meulen, and M. Schauer. "A piecewise deterministic Monte Carlo method for diffusion bridges". In: *Statistics and Computing* 31.3 (2021), pp. 1–21.

Chapter 3 is written in collaboration with Joris Bierkens, Gareth Roberts (University of Warwick) and Kengo Kamatani (Osaka University) and resulted in the publication

> J. Bierkens, S. Grazzi, K. Kamatani, and G. Roberts. "The Boomerang Sampler". In: *International conference on machine learning*. PMLR. 2020, pp. 908–918.

The material presented in Chapter 4 is joint work with Frank van der Meulen, Joris Bierkens and Moritz Schauer and it is published as

> J. Bierkens, S. Grazzi, F. v. d. Meulen, and M. Schauer. "Sticky PDMP samplers for sparse and local inference problems". In: *Statistics and Computing* 33.1 (2023), p. 8. To appear in: *Statistics and Computing*.

Chapter 5 derived in the paper

> J. Bierkens, S. Grazzi, M. Schauer, and G. Roberts. "Methods and applications of PDMP samplers with boundary conditions", *In preparation.*

The paper is currently in preparation.

This page is intentionally left blank.

Chapter 2

A PDMP sampler for diffusion bridges

2.1 Introduction

Diffusion processes are an important class of continuous time probability models which find applications in many fields such as finance, physics and engineering. They naturally arise by adding Gaussian random perturbations (white noise) to deterministic systems. We consider diffusions described by a one-dimensional stochastic differential equation of the form

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \mathrm{d}W_t, \quad X_0 = u, \tag{2.1}$$

where $(W_t)_{t\geq 0}$ is a driving scalar Wiener process defined in some probability space and b is the *drift* of the process. The solution of equation (2.1), assuming it exists, is an instance of one-dimensional time-homogeneous diffusion. We aim to sample X on $[0, T]$ conditional on $\{X_T = v\}$, also known as a *diffusion bridge*.

One driving motivation for studying this problem is estimation for discretely observed diffusions. Here, one assumes observations $\mathcal{D} = \{x_{t_1}, \ldots, x_{t_N}\}$ at observations times $t_1 < \ldots < t_N$ are given and interest lies in estimation of a parameter θ appearing in the drift b. It is well known that this problem can be viewed as a missing data problem as in Roberts and Stramer (2001), where one iteratively imputes the missing paths conditional on the parameter and the observations, and then the parameter conditional on the "full" continuous path. Due to the Markov property, the missing paths in between subsequent observations can be sampled independently and each of such segments constitutes a diffusion bridge. As this application requires sampling iteratively many diffusion bridges, it is crucial to have a fast algorithm for this step. We achieve this by adapting the Zig-Zag sampler for the simulation of diffusion bridges. The Zig-Zag sampler is an innovative non-reversible and rejection-free Markov process Monte Carlo algorithm which can exploit the structure present

in this high-dimensional sampling problem. It is based on simulating a piecewise deterministic Markov process (PDMP). To the best of our knowledge, this is the first application of PDMPs for diffusion bridge simulation. This method also illustrates the use of a local version of the Zig-Zag sampler in a genuinely high dimensional setting (arguably even an infinite dimensional setting).

The problem of diffusion bridge simulation has received considerable attention over the past two decades, see for example Bladt, Sørensen, et al. (2014), Beskos, Papaspiliopoulos, Roberts, et al. (2006), Meulen and Schauer (2017), Mider et al. (2019), Bierkens, Meulen, and Schauer (2020) and references therein. This far from exhaustive list of references includes methods that apply to a more general setting than considered here, such as multivariate diffusions, conditioning on partial observations and hypo-elliptic diffusions. Among the methods that can be applied, most of the methodologies available are of the acceptance-rejection type and scale poorly with respect to some parameters of the diffusion bridge. For example, if the proposed path is not informed by the target distribution, the probability of accepting the path depends strongly on the discrepancy between the proposed path and the target diffusion bridge measure and usually scales poorly as the time horizon of the diffusion bridge T grows. In contrast, gradient based techniques which compute informed proposals (e.g. Metropolis-adjusted Langevin algorithm), require the evaluation of the gradient of the target distribution, which, in this case, is a path integral that has to be generally computed numerically and its computational cost is of order T, leading to computational limitations. The present work aims to alleviate such restrictions through the use of a rejection-free method and an exact subsampling technique which reduces the cost of evaluating the gradient. On a more abstract level, our method can be viewed as targeting a probability distribution which is obtained by a push-forward of Wiener measure through a change of measure. It then becomes apparent that the studied problem of diffusion bridge simulation is a nicely formulated non-trivial example problem within this setting to study the potential of simulation based on PDMPs. Our results open new paths towards applications of the Zig-Zag for high dimensional problems.

2.1.1 Approach

In this section we present the main ideas used in this paper.

Brownian motion expanded in the Faber-Schauder basis

Our starting point is the Lévy-Ciesielski construction of Brownian Motion. Define $\bar{\phi}(t) = \sqrt{t}$, $\phi_{0,0}(t) = \sqrt{T}\left((t/T)\mathbf{1}_{[0,T/2]}(t) + (1-t/T)\mathbf{1}_{(1/2,1]}(t)\right)$ and set

$$\phi_{i,j}(t) = 2^{-i/2}\phi_{0,0}(2^i t - jT), \quad \text{for} \quad i = 0, 1, ..., \quad j = 0, 1, ...2^i - 1.$$

If $\bar{\xi}$ is standard normal and $\{\xi_{i,j}\}$ is a sequence of independent standard normal random variables (independent of $\bar{\xi}$), then

$$X^N(t) = \bar{\phi}(t)\bar{\xi} + \sum_{i=0}^{N}\sum_{j=0}^{2^i-1} \xi_{i,j}\phi_{i,j}(t) \tag{2.2}$$

converges almost surely on $[0,T]$ (uniformly in t) to a Brownian motion as $N \to \infty$ (see e.g. Section 1.2 of McKean 1969). The basis formed by $\bar{\phi}$ and $\{\phi_{i,j}\}$ is known as the Faber-Schauder basis (see Figure 2.1). The larger i, the smaller the support of $\phi_{i,j}$, reflecting that higher order coefficients represent the fine details of the process. A Brownian bridge starting in u and ending in v can be obtained by fixing $\bar{\xi} = v/\sqrt{T}$ and adding the function $\bar{\bar{\phi}}(t)u = (1 - t/T)u$ to (2.2). By sampling $\xi^N := (\xi_{0,0}, \xi_{1,0}, ..., \xi_{N,2^N-1})$ (which in this case are standard normal), approximate realisations of a Brownian bridge can be obtained.

Zig-Zag sampler for diffusion bridges

Let $\mathbb{Q}^u$ denote the Wiener measure on $C[0,T]$ with initial value $X_0 = u$ (cf. section 2.4 of Karatzas and Shreve 1991) and let $\mathbb{P}^u$ denote the law on $C[0,T]$ of the diffusion in (2.1). Under mild conditions on b, the two measures are absolutely continuous and their Radon-Nikodym derivative $\frac{\mathrm{d}\mathbb{P}^u}{\mathrm{d}\mathbb{Q}^u}$ is given by the Girsanov formula. Denote by $\mathbb{P}^{u,v_T}$ and $\mathbb{Q}^{u,v_T}$ the measures of the diffusion bridge and the Wiener bridge respectively, both starting at u and conditioned to hit a point v at time T. Applying the Bayes' law for conditional expectations (Klebaner 2005, Chapter 10) we obtain:

$$\frac{\mathrm{d}\mathbb{P}^{u,v_T}}{\mathrm{d}\mathbb{Q}^{u,v_T}}(X) = \frac{q(0,u,T,v)}{p(0,u,T,v)}\frac{\mathrm{d}\mathbb{P}^u}{\mathrm{d}\mathbb{Q}^u}(X), \tag{2.3}$$

where p and q are the transition densities of X under $\mathbb{P}, \mathbb{Q}$ respectively so that for $s < t$, $p(s,x,t,y)\mathrm{d}y = P(X_t \in \mathrm{d}y \mid X_s = x)$. As p is intractable, the Radon-Nikodym derivative for the diffusion bridge is only known up to proportionality constant. The main idea now consists of rewriting the Radon-Nikodym derivative in (2.3), evaluating it in X^N and running the Zig-Zag sampler for ξ^N targeting this density. Technicalities to actually get this to work are detailed in Section 2.3. A novelty is the introduction of a *local* version of the Zig-Zag sampler, analogously to the *local bouncy particle sampler* (Bouchard-Côté, Vollmer, and Doucet 2018). This allows for exploiting the sparsity in the dependence structure of the coefficients of the Faber-Schauder expansion efficiently, resulting in a reduction of the complexity of the algorithm. The methodology we propose is derived for one dimensional diffusion processes with unit diffusivity. However, diffusions with state-dependent diffusivity can be transformed to this setting using the Lamperti transform (an example is given in Subsection 2.5.3). In Subsection 2.6.1 we generalize the method to multivariate

diffusion processes with unit diffusivity, assuming the drift to be a conservative vector field.

2.1.2 Contributions of the paper

The Faber-Schauder basis offers a number of attractive properties:

(a) The coefficients of a diffusions have a structural conditional independence property (see Section 2.4 and Appendix A.1) which can be exploited in numerical algorithms to improve their efficiency.

(b) A diffusion bridge is obtained from the unconditioned process by simply fixing the coefficient $\bar{\xi}$.

(c) It will be shown (see for example Figure 2.8) that the non-linear component of the diffusion process is typically captured by coefficients ξ_{ij} in equation (2.2) for which i is small. This allows for a low dimensional representation of the process and yet a good approximation. Therefore, the approximation error caused by leaving out fine details is equally divided over $[0, T]$, contrary to approaches where a proxy for the diffusion bridge is simulated by Euler discretisation of an SDE governing its dynamics. In the latter case, the discretisation error accumulates over the interval on which the bridge is simulated.

(d) It is very convenient from a computational point of view as each function is piecewise linear with compact support.

We adopt the Zig-Zag sampler (Bierkens, Fearnhead, and Roberts 2019) which is a sampler based on the theory of piecewise deterministic Markov processes (see Fearnhead et al. 2018, Bouchard-Côté, Vollmer, and Doucet 2018, Andrieu and Livingstone 2019, Andrieu et al. 2018). The main reasons motivating this choice are:

(a) The partial derivatives of the log-likelihood of a diffusion bridge measure usually appear as a path integral that has to be computed numerically (introducing consequently computational burden derived by this step and its bias). The Zig-Zag sampler allows us to replace the gradient of the log-likelihood with an unbiased estimate of it without introducing bias in the target measure. This is done in Subsection 2.4.4 with the subsampling technique which was presented in Bierkens, Fearnhead, and Roberts (2019) for applications for which the evaluation of the log-likelihood is expensive due to the size of the dataset.

(b) In the same spirit as the local Bouncy Particle Sampler of Bouchard-Côté, Vollmer, and Doucet (2018) and Peters and With (2012), the *local* and the

fully local Zig-Zag sampler introduced in Section 2.4 reduces the complexity of the algorithm improving its efficiency with respect to the standard Zig-Zag Algorithm as the dimensionality of the target distribution increases (see Subsection 2.6.2). This opens the way to high dimensional applications of the Zig-Zag sampler when the dependency graph of the target distribution is not fully connected and when using subsampling. The factorization of the log-likelihood and the local method we proposed is reminiscent of other work such as e.g. Faulkner et al. (2018), Michel, Tan, and Deng (2019) and Monmarché et al. (2020).

(c) The method is a rejection-free sampler, differing from most of the methodologies available for simulating diffusion bridges.

(d) The Zig-Zag sampler is defined and implemented in continuous time, eliminating the choice of tuning parameters appearing for example in the proposal density of the Metropolis-Hastings algorithm. This advantage comes at the cost of a more complicated method which relies upon bounding from above rates which are model specific and often difficult to derive (see Section 2.5 for our specific applications).

(e) The process is non-reversible: as shown, for example, in Diaconis, Holmes, and Neal (2000), non-reversibility generally enhances the speed of convergence to the invariant measure and mixing properties of the sampler. For an advanced analysis on convergences results for this class of non-reversible processes, we refer to the articles Andrieu and Livingstone (2019) and Andrieu et al. (2018).

The local Zig-Zag sampler relies on the conditional independence structure of the coefficients only. This translates to other settings than diffusion bridge sampling, or other choices of basis functions. For this reason, Section 2.4 describes the algorithms of the sampler in their full generality, without referring to our particular application. A documented implementation of the algorithms used in this manuscript can be found in Schauer and Grazzi (2021).

2.1.3 Outline

In Section 2.2 we set some notation and recap the Zig-Zag sampler. In Section 2.3 we expand a diffusion process in the Faber-Schauder basis and prove the aforementioned conditional dependence. The simulation of the coefficients ξ^N presents some challenges as it is high dimensional and its density is expressed by an integral over the path. We give two variants of the Zig-Zag algorithm which enables sampling in a high dimensional setting. In particular, in Section 2.4 we present the local and fully local Zig-Zag algorithms which exploit a factorization of the joint density

(Appendix A.1) and a subsampling technique which, in this setting, is used to avoid the evaluation of the path integral appearing in the density (which otherwise would severely complicate the implementation of the sampler). In Section 2.5 we illustrate our methodology using a variety of examples, validate our approach and compare the Zig-Zag sampler with other benchmark MCMC algorithms. We conclude by sketching the extension of our method to multi-dimensional diffusion bridges, carrying out an informal scaling analysis and providing several remarks for future research (Section 2.6 and Section 2.7).

2.2 Preliminaries

Throughout, we denote by ∂_i the partial derivative with respect to the coefficient ξ_i, the positive part of a function f by $(f)^+$, the ith element and the Euclidean norm of a vector x respectively by $[x]_i$ and $\|x\|$. The cardinality of a countable set A is denoted by $|A|$.

2.2.1 Notation for the Faber-Schauder basis

To graphically illustrate the Faber-Schauder basis, a construction of a Brownian motion with the representation of the basis functions is given in Figure 2.1. The Faber-Schauder functions are piecewise linear with compact support. The length of the support and the height of the function is determined by the first index while the second index determines the location. All basis functions with first index i are referred to as *level* i basis functions. For convenience, we often swap between double and single indexing of Faber-Schauder functions. Denote the double indexing with (i, j) and the single indexing with n. We go from one to the other through the transformations

$$i = \lfloor \log_2(n) \rfloor, \qquad j = n - 2^i, \qquad n = 2^i + j;$$

where $\lfloor \cdot \rfloor$ denotes the floor function. The basis with truncation level N has $M := 2^{N+1} - 1$ coefficients. Let ξ^N denote the vector of coefficients up to level N, i.e.

$$\xi^N := (\xi_{0,0}, \xi_{1,0}, ..., \xi_{N,2^N-1}) \in \mathbb{R}^M \tag{2.4}$$

and let $X^{\xi^N} := X^N$ when we want to stress the dependencies of X^N on the coefficients ξ^N. Using double indexing, we denote by $S_{i,j} = \operatorname{supp} \phi_{i,j}$.

2.2.2 The Zig-Zag sampler

A *piecewise deterministic Markov process* (Davis 1993) is a continuous-time process with behaviour governed by random jumps at points in time, but deterministic evolution governed by an ordinary differential equation in between those times (yielding

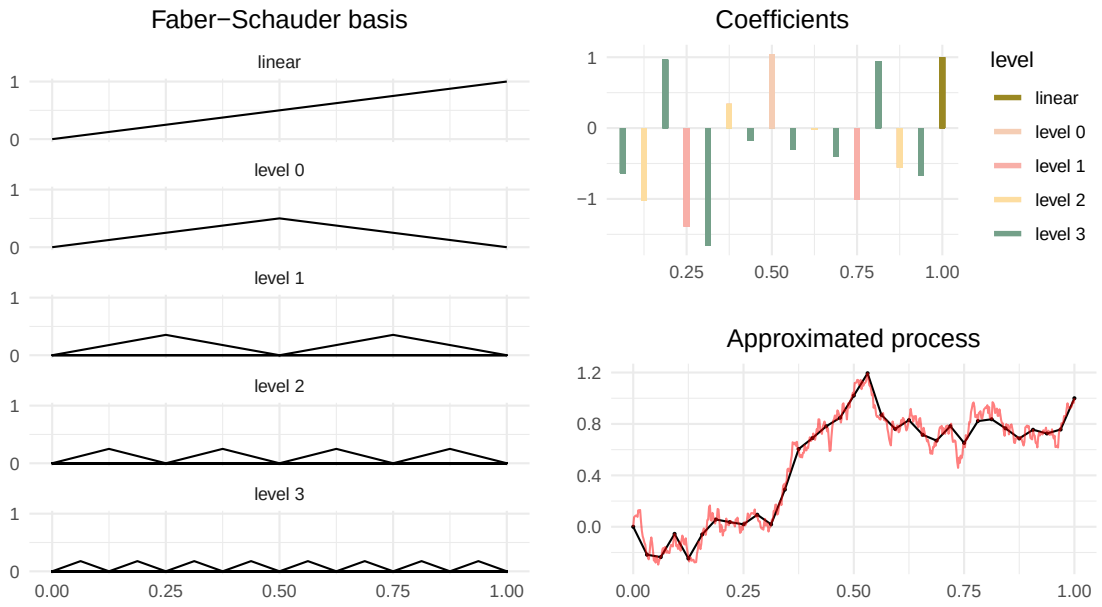


Figure 2.1: Lévy-Ciesielski construction of a Brownian motion on $(0, 1)$. On the left the Faber-Schauder basis functions up to level $N = 3$, on the top-right the values of the corresponding coefficients located at the peak of their relative FS basis function and on the bottom-right the resulting approximated Brownian path X^N (black line) compared with a finer approximation (red line). The truncated sum defines the process in $2^{N+1} + 1$ finite dyadic points (black dots) with linear interpolation in between points. A finer approximation corresponds to Brownian fill-in noise between any two neighboring dyadic points.

piecewise-continuous realizations). If the differential equation can be solved in closed form and the random event times can be sampled exactly, then the process can be simulated in continuous time without introducing any discretization error (up to floating number precision) making it attractive from a computational point of view.

By a careful choice of the event times and deterministic evolution, it is possible to create and simulate an ergodic and non-reversible process with a desired unique invariant distribution (Fearnhead et al. 2018). The Zig-Zag sampler (Bierkens, Fearnhead, and Roberts 2019) is a successful construction of such a processes. We now recap the intuition and the main steps behind the Zig-Zag sampler.

The *one-dimensional* Zig-Zag sampler is defined in the *augmented space* $(\xi, \theta) \in \mathbb{R} \times \{+1, -1\}$, where the first coordinate is viewed as the position of a moving particle and the second coordinate as its velocity. The dynamics of the process $t \mapsto (\xi(t), \theta(t))$ (not to be confused with the time indexing the diffusion process) are as follows: starting from $(\xi(0), \theta(0))$,

(a) its flow is deterministic and linear in its first component with direction $\theta(0)$ and constant in its second component until an event at time τ occurs. That is, $(\xi(t), \theta(t)) = (\xi(0) + t\theta(0), \theta(0))$, $0 \leq t \leq \tau$.

(b) At an event time τ, the process changes the sign of its velocity, i.e. $(\xi(\tau), \theta(\tau)) = (\xi(\tau-), -\theta(\tau-))$.

The event times are simulated from an inhomogeneous Poisson process with specified rate $\lambda\colon (\mathbb{R} \times \{1, -1\}) \to \mathbb{R}^+$ such that $P(\tau \in [t, t+\epsilon]) = \lambda(\xi(t), \theta(t))\epsilon + o(\epsilon)$, $\epsilon \downarrow 0$.

The *d-dimensional* Zig-Zag sampler is conceived as the combination of d one-dimensional Zig-Zag samplers with rates $\lambda_i(\xi, \theta)$, $i = 1, ..., d$, where the rates create a coupling of the independent coordinate processes. The following result provides a sufficient condition for the d-dimensional Zig-Zag sampler to have a particular d-dimensional target density π as invariant distribution. Assume that the target d-dimensional distribution has strictly positive density with respect to the Lebesgue measure i.e.

$$\pi(\mathrm{d}\xi) \propto \exp(-\psi(\xi))\mathrm{d}\xi, \qquad \xi \in \mathbb{R}^d.$$

Define the *flipping function* as $F_i(\theta) = (\theta_1, ..., -\theta_i, ..., \theta_d)$, for $\theta \in \{-1, +1\}^d$. For any $i = 1, ..., d$ and $(\xi, \theta) \in \mathbb{R}^d \times \{1, -1\}^d$, the Zig-Zag process with Poisson rates satisfying

$$\lambda_i(\xi, \theta) - \lambda_i(\xi, F_i(\theta)) = \theta_i \partial_i \psi(\xi), \tag{2.5}$$

has π as invariant density. Condition (2.5) is derived in the supplementary material of Bierkens, Fearnhead, and Roberts (2019). Condition (2.5) is equivalent to

$$\lambda_i(\xi, \theta) = (\theta_i \partial_i \psi(\xi))^+ + \gamma_i(\xi) \tag{2.6}$$

for some $\gamma_i(\xi) \geq 0$. Throughout, we set $\gamma_i(\xi) = 0$ because generally the algorithm is more efficient for lower Poisson event intensity (see for example Andrieu and Livingstone 2019, Subsection 5.4).

Assume the target density is $\pi(\xi) = c\widetilde{\pi}(\xi)$. The process targets the specific distribution function through the Poisson rate λ which is a function of the gradient of $\xi \mapsto \psi(\xi) = -\log(\widetilde{\pi}(\xi))$, so that any proportionality factor of the density disappears. Throughout we refer to the function ψ as the *energy function*. As opposed to standard Markov chain Monte Carlo methods, the process is not reversible and it is defined in continuous time.

Example 2.2.1. *Consider a d-dimensional Gaussian random variable with mean $\mu \in \mathbb{R}^d$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Then*

- $\pi(\xi) \propto \exp\left(-(\xi - \mu)'\Sigma^{-1}(\xi - \mu)/2\right)$,
- $\partial_k \psi(\xi) = [\Sigma^{-1}(\xi - \mu)]_k$,

- $\lambda_k(\xi, \theta) = (\theta_k[\Sigma^{-1}(\xi - \mu)]_k)^+$.

Notice that if Σ is diagonal, then $\lambda_k(\xi, \theta) = 0$ whenever the process is directed towards the mean so that no jump occurs in the kth component when one of the following conditions is satisfied: $(\theta_k = -1, \xi_k - \mu_k \geq 0)$ or $(\theta_k = 1, \xi_k - \mu_k \leq 0)$. In Figure 2.2 we simulate a realization of the Zig-Zag sampler targeting a univariate standard normal random distribution.

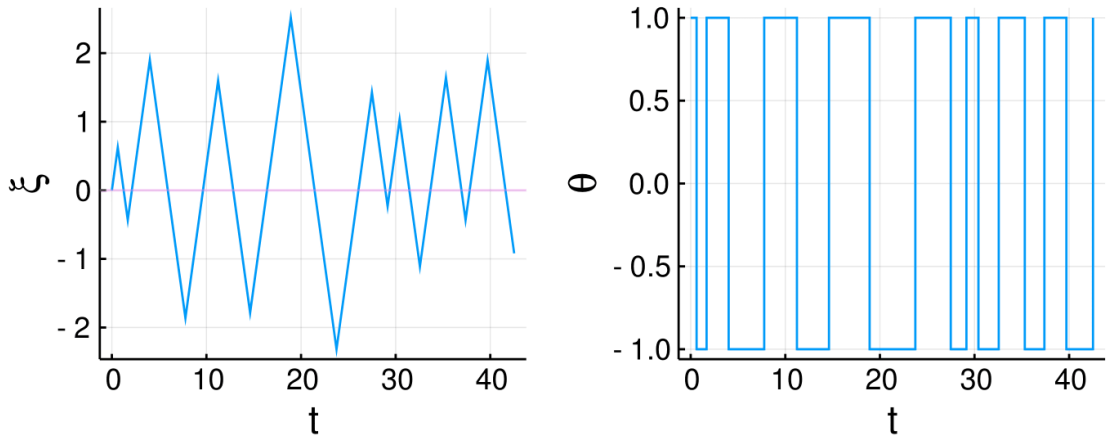


Figure 2.2: One dimensional Zig-Zag targeting a Gaussian random variable $\mathcal{N}(0, 1)$. Left: $t \mapsto \xi(t)$, right: $t \mapsto \theta(t)$.

Algorithm 1 shows the standard implementation of the Zig-Zag sampler. Given a fixed time $t \geq 0$ and a position $(\xi(t), \theta(t))$, the first event time τ^* after t is determined by taking the minimum of event times $\tau_1, \tau_2, \ldots, \tau_d$ simulated according to the Poisson rates $\lambda_i, i = 1, 2, ..., d$. At event time τ^*, the velocity vector becomes $\theta(\tau^*) = F_{i^*}(\theta(t))$, with $i^* = \arg\min(\tau_1, \ldots, \tau_d)$. The algorithm iterates this step moving forward each time until the next simulated event time exceeds the final clock τ_{final}.

Although we consider the velocities for each dimension of a d-dimensional Zig-Zag process to be either 1 or -1, these can be taken to be any non-zero values $(\theta_i, -\theta_i)$ for $i = 1, ..., d$. A finetuning of $\theta_1, ..., \theta_N$ can improve the performance of the sampler. Note that the only challenge in implementing Algorithm 1 lies on the simulation of the waiting times which correspond to the simulation of the first event time of d inhomogeneous Poisson processes (IPPs) with rates $\lambda_1, \lambda_2, ..., \lambda_d$ which are functions of the state space (ξ, θ) of the process. Since the flow of the process is linear and deterministic, the Poisson rates are known at each time and are equal to

$$\lambda_i(t; \xi, \theta) = \lambda_i(\xi + t\theta, \theta), \qquad i = 1, 2, ..., d.$$

To lighten the notation, we write $\lambda_i(t) := \lambda_i(t; \xi, \theta)$ when ξ, θ are fixed. Given an initial position ξ and velocity θ, the waiting times $\tau_1, ..., \tau_d$ are computed by finding the roots for x of the equations

$$\int_0^x \lambda_i(s)\mathrm{d}s + \log(u_i) = 0, \qquad i = 1, 2, ..., d, \tag{2.7}$$

where $(u_i)_{i=1,2,...,d}$ are independent realisations from the uniform distribution on $(0, 1)$. When it is not possible to find roots of equation (2.7) efficiently, for example in closed form, it suffices to find upper bounds for the rate functions for which this is possible; Subsection 2.4.4 treats this problem for our particular setting. The linear evolution of the process and the jumps of the velocities are always trivially computed and implemented.

Algorithm 1 returns a *skeleton* of values corresponding to the position of the process at the event times. From these values, it is straightforward to reconstruct the continuous path of the Zig-Zag sampler. Given a sample path of the Zig-Zag sampler from 0 to τ_{final}, we can obtain a sample from the target distribution in the following way:

- Denote by $\xi(\tau)$ the value of the vector ξ at the Zig-Zag clock $\tau < \tau_{\text{final}}$. Fixing a sample frequency $\Delta\tau$, we can produce a sample from the density π by taking the values of the random vector ξ at time $\tau_{\text{burn-in}} + \Delta\tau, \tau_{\text{burn-in}} + 2\Delta\tau, ..., \tau_{\text{final}}$ where $\tau_{\text{burn-in}}$ is the initial burn-in time taken to ensure that the process has reached its stationary regime. Throughout the paper, we create samples using this approach.

2.2.3 Zig-Zag sampler for Brownian bridges

The previous subsections contain all ingredients necessary to run the Zig-Zag sampler in a finite dimensional projection of the Brownian bridge measure $\mathbb{Q}^{0,v}$ on the interval $[0, T]$. We fix $\bar{\xi}$ to v and run the Zig-Zag sampler for ξ^N as defined in (2.4) targeting a multivariate normal distribution. Figure 2.3 shows 100 samples obtained from one sample run of the Zig-Zag sampler where the coefficients are mapped to samples paths using (2.2). The final clock of the Zig-Zag is set to $\tau_{\text{final}} = 500$ with initial burning $\tau_{\text{burn-in}} = 10$.

Both Brownian motion and the Brownian bridge are special in that all coefficients in the Faber-Schauder basis are independent. Of course, these processes can directly be simulated without need of a more advanced method like the Zig-Zag sampler. However, for a diffusion process with nonzero drift this property is lost. Nevertheless, we will see that when the process is expanded in the Faber-Schauder basis, many coefficients are still *conditionally* independent. This implies that the dependency graph of the joint density of the coefficients is sparse. We will show in Section 2.4

Algorithm 1 Standard d-dimensional Zig-Zag sampler (Bierkens, Fearnhead, and Roberts 2019)

procedure ZIGZAG($\tau_{\text{final}}, \xi, \theta$)
  Initialise $k = 1$, $t = 0$
  $\tau_j \sim \text{IPP}(\lambda_j(\cdot; \xi, \theta)), \quad j = 1, ..., d$ ▷ Draw from Inhomogeneous Poisson process (IPP)
  while $t \leq \tau_{\text{final}}$ **do**
    $\tau^*, i^* \leftarrow \text{findmin}(\tau_1, .., \tau_d)$
    Update: $\xi \leftarrow \xi + \theta(\tau^* - t)$
    Update: $\theta_{i^*} \leftarrow -\theta_{i^*}; \quad t \leftarrow \tau^*$
    Save $\xi^{(k)} \leftarrow \xi; \quad t^{(k)} \leftarrow t$
    for $j = 1, ..., d$ **do**
      $\tau_j \sim t + \text{IPP}(\lambda_j(\cdot; \xi, \theta))$
    end for
    $k \leftarrow k + 1$
  end while
  return Skeletons $(\xi^{(l)}, t^{(l)})_{l=1,...,k-1}$
end procedure

how this property can be exploited efficiently using the Zig-Zag sampler in its local version.

2.3 Faber-Schauder expansion of diffusion processes

We extend the results of Section 2.2 to one-dimensional diffusions governed by the SDE in (2.1). Although the density is defined in infinite dimensional space, in this section we justify both intuitively and formally that the diffusion can be approximated to arbitrary precision by considering a finite dimensional projection of it.

The intuition behind using the Faber-Schauder basis is that, under mild assumptions on the drift function b, any diffusion process behaves locally as a Brownian motion. Expanding the diffusion process with the Faber-Schauder functions, this notion translates to the existence of a level N such that the random coefficients at higher levels which are associated to the Faber-Schauder basis are approximately independent standard normal and independent from ξ^N under the measure $\mathbb{P}$.

Define the function $Z_t \colon \mathbb{R}^+ \times C[0,T] \to \mathbb{R}^+$ given by

$$Z_t(X) = \exp\left(\int_0^t b(X_s)\mathrm{d}X_s - \frac{1}{2}\int_0^t b^2(X_s)\mathrm{d}s\right) \tag{2.8}$$

where the first integral is understood in the Itô sense and $X \equiv (X_s,\, s \in [0,T])$.

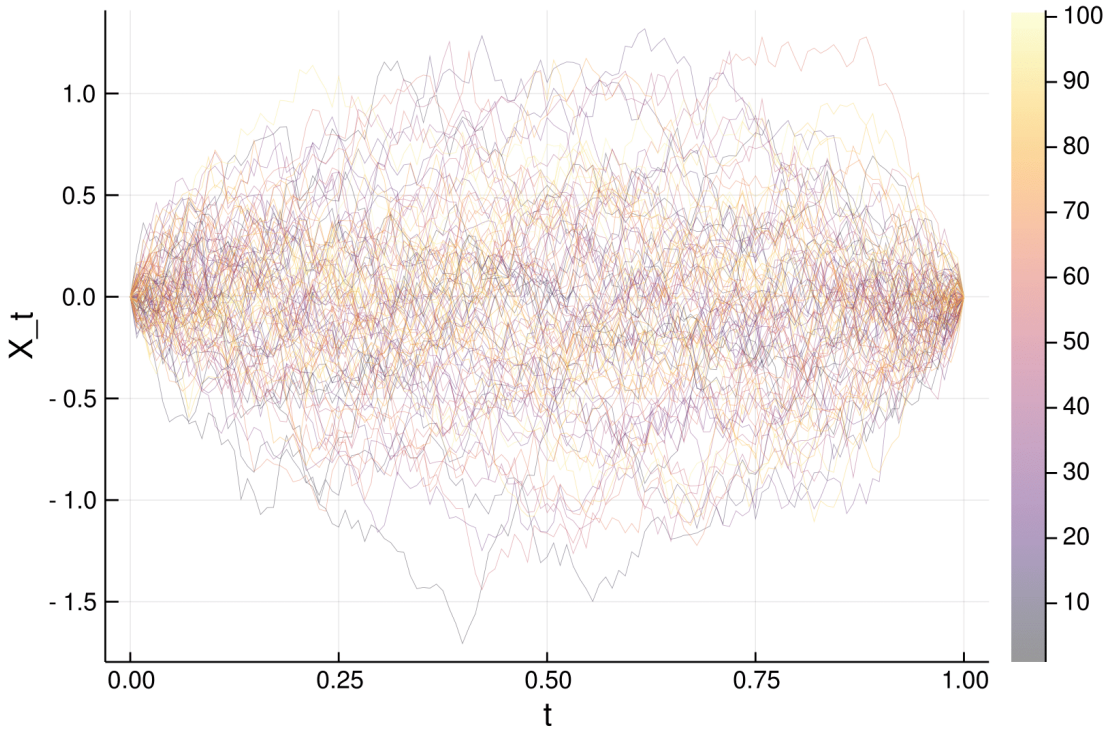


Figure 2.3: 100 samples from the Brownian bridge measure starting at 0 and hitting 0 at time 1 obtained by one run of the Zig-Zag sampler targeting the coefficients relative to the measure expanded with the Faber-Schauder basis. The resolution level is fixed to $N = 6$ and the Zig-Zag clock to $\tau_{\text{final}} = 500$ and initial burn in $\tau_{\text{burn-in}} = 10$.

Assumption 2.3.1. *Z_t is a $\mathbb{Q}$-martingale.*

For sufficient conditions for verifying that this assumption applies, we refer to Remark 2.3.6, Remark 2.3.9 and Liptser, Aries, and Shiryaev (2013), Chapter 6.

Theorem 2.3.2. *(Girsanov's theorem) If Assumption 2.3.1 is satisfied,*

$$\frac{\mathrm{d}\mathbb{P}^u}{\mathrm{d}\mathbb{Q}^u}(X) = Z_T(X). \tag{2.9}$$

Moreover, a weak solution of the stochastic differential equation exists which is unique in law.

Proof. This is a standard result in stochastic calculus (see Liptser, Aries, and Shiryaev 2013, Section 6). □

As we consider diffusions on $[0,T]$ with T fixed, we denote $Z(X) := Z_T(X)$. Due to the appearance of the stochastic Itô integral in $Z(X)$, we cannot substitute for X its truncated expansion in the Faber-Schauder basis. Clearly, whereas the approximation has finite quadratic variation, X has not. Assuming that b is differentiable and applying Itô's lemma to the function $B(x) = \int_0^x b(s)\mathrm{d}s$, the stochastic integral can be replaced and equation (2.8) is rewritten as

$$Z(X) = \exp\left(B(X_T) - B(X_0) - \frac{1}{2}\int_0^T \left(b^2(X_s) + b'(X_s)\right)\mathrm{d}s\right), \tag{2.10}$$

where b' is the derivative of b.

Definition 2.3.3. *Let X be a diffusion governed by (2.1). Let X^N be the process derived from X by setting to zero all coefficients of level exceeding N in its Faber-Schauder expansion (see equation (2.2)). Set*

$$Z^N(X) = \exp\left(B\left(X_T^N\right) - B\left(X_0^N\right) - \frac{1}{2}\int_0^T \left[b^2\left(X_s^N\right) + b'\left(X_s^N\right)\right]\mathrm{d}s\right).$$

We define the approximating measure $\mathbb{P}_N$ by the change of measure

$$\frac{\mathrm{d}\mathbb{P}_N^u}{\mathrm{d}\mathbb{Q}^u}(X) = \frac{Z^N(X)}{c_N}, \tag{2.11}$$

where $c_N = \mathbb{E}_{\mathbb{Q}}\left(Z^N(X)\right)$.

Note that the measure $\mathbb{P}_N^u$ associated to the approximated stochastic process is still on an infinite dimensional space and such that the joint measure of random coefficients ξ^N is different from the one under $\mathbb{Q}^u$ while the remaining coefficients stay independent standard normal and independent from ξ^N. This is equivalent to approximating the diffusion process at finite dyadic points with Brownian noise fill-in in between every two points. We now fix the final point v_T by setting $\bar{\xi} = v_T$. Define the *approximated stochastic bridge* with measure $\mathbb{P}_N^{u,v_T}$ in an analogous way of equation (2.11), so that

$$\frac{\mathrm{d}\mathbb{P}_N^{u,v_T}}{\mathrm{d}\mathbb{Q}^{u,v_T}}(X) = \frac{Z^N(X)}{c_N^{v_T}}. \tag{2.12}$$

where $c_N^{v_T} = \mathbb{E}_{\mathbb{Q}^{u,v_T}}\left(Z^N(X)\right)$. The following is the main assumption made.

Assumption 2.3.4. *The drift b is continuously differentiable and $b^2 + b'$ is bounded from below.*

Theorem 2.3.5. *If Assumptions 2.3.1 and 2.3.4 are satisfied, then $\mathbb{P}_N^{u,v_T}$ converges weakly to $\mathbb{P}^{u,v_T}$ as $N \to \infty$.*

Proof. In the following we lighten the notation by omitting the initial point u from the notation, which will be assumed fixed to $u = x_0$. We wish to show that $\mathbb{P}_N^{v_T}$ converges weakly to $\mathbb{P}^{v_T}$ as $N \to \infty$. This is equivalent to showing that $\int f \mathrm{d}\mathbb{P}_N^{v_T} \to \int f \mathrm{d}\mathbb{P}^{v_T}$ for all bounded and continuous functions f. Write $c_\infty^{v_T} = p(0, x_0, T, v_T)/q(0, x_0, T, v_T)$. By equation (2.3) and (2.9),

$$\mathbb{E}_{\mathbb{Q}^{v_T}} Z(X) = \mathbb{E}_{\mathbb{Q}^{v_T}} \frac{d\mathbb{P}^{x_0}}{d\mathbb{Q}^{x_0}} = c_\infty^{v_T} \mathbb{E}_{\mathbb{Q}^{v_T}} \left[\frac{d\mathbb{P}^{v_T}}{d\mathbb{Q}^{v_T}} \right] = c_\infty^{v_T}$$

and we have that

$$\begin{aligned}
&\left| \int f \mathrm{d}\mathbb{P}_N^{v_T} - \int f \mathrm{d}\mathbb{P}^{v_T} \right| \\
&= \left| \int f \left(\frac{Z^N}{c_N^{v_T}} - \frac{Z}{c_\infty^{v_T}} \right) \mathrm{d}\mathbb{Q}^{v_T} \right| \\
&\leq \|f\|_\infty \int \left| \frac{Z^N(X)}{c_N^{v_T}} - \frac{Z(X)}{c_\infty^{v_T}} \right| \mathrm{d}\mathbb{Q}^{v_T}(X) \\
&\leq \|f\|_\infty \left(\frac{1}{c_N^{v_T}} \int \left| Z^N(X) - Z(X) \right| \mathrm{d}\mathbb{Q}^{v_T}(X) + \int Z(X) \left| \frac{1}{c_N^{v_T}} - \frac{1}{c_\infty^{v_T}} \right| \mathrm{d}\mathbb{Q}^{v_T}(X) \right) \\
&\leq \|f\|_\infty \left(\frac{1}{c_N^{v_T}} \int \left| Z^N(X) - Z(X) \right| \mathrm{d}\mathbb{Q}^{v_T}(X) + \left| \frac{c_\infty^{v_T}}{c_N^{v_T}} - 1 \right| \right) \qquad (2.13)
\end{aligned}$$

where we used Assumption 2.3.1 for applying the change of measure between the conditional measures. Notice that $Z^N(X) = Z(X^N)$. The mapping $X \mapsto Z(X)$, as a function acting on $C(0,T)$ with uniform norm, is continuous, since B, b, and b' are continuous. Therefore, it follows from the Lévy-Ciesielski construction of Brownian motion (see Section 2.1.1) and the continuous mapping theorem that

$$Z^N(X) \to Z(X) \qquad \mathbb{Q}^{v_T} - a.s.$$

Now notice that, under conditional measures $\mathbb{Q}^{v_T}$ and $\mathbb{P}^{v_T}$, the term $B(X_T) - B(X_0)$ is fixed. By the assumptions on b and b', Z is a bounded function and by dominated convergence we get that

$$\lim_{N \to \infty} \mathbb{E}_{\mathbb{Q}}^{v_T} |Z^N(X) - Z(X)| = 0$$

giving convergence to zero of the first term in (2.13). This implies that also the constant $c_N := \mathbb{E}_{\mathbb{Q}}^{v_T} |Z^N(X)|$ converges to $\mathbb{E}_{\mathbb{Q}}^{v_T} |Z(X)| = c_\infty^{v_T}$ so that all the terms in (2.13) converge to 0. □

We now list some technical conditions for the process to satisfy Assumptions 2.3.1 and 2.3.4.

Remark 2.3.6. *If $|b(x)| \leq c(1+|x|)$, for some positive constant c, then Assumption 2.3.1 is satisfied.*

Proof. See Liptser, Aries, and Shiryaev (2013), Section 6, Example 3 (b). □

Remark 2.3.7. *If b is globally Lipschitz and continuously differentiable, then Assumptions 2.3.1 and 2.3.4 are satisfied.*

Proof. Assumption 2.3.4 is trivially satisfied. By Remark 2.3.6, also Assumption 2.3.1 is satisfied. □

In Subsection 2.5.3 we will present an example where the drift b is not globally Lipschitz, yet Assumption 2.3.4 is satisfied.

Assumption 2.3.8. *There exists a non-decreasing function $h : [0,\infty) \to [0,\infty)$ such that $B(x) \leq h(|x|)$ and*

$$\int_0^\infty \exp(h(x) - x^2/(2T))\, dx < \infty.$$

The above integrability condition is for example satisfied if $h(|x|) = c(1+|x|)$ for some $c > 0$.

Remark 2.3.9. *If Assumptions 2.3.4 and 2.3.8 hold, then Assumption 2.3.1 is satisfied.*

Proof. By Subsection 3.5 in Karatzas and Shreve (1991), (Z_t) is a local martingale. Say $b'(x) + b^2(x) \geq -2C$, where $C \geq 0$. Using the assumptions, we have

$$Z_t = \exp\left(B(X_t) - B(X_0) - \tfrac{1}{2}\int_0^t \{b'(X_s) + b^2(X_s)\}\, ds\right) \leq A\exp(Ct)\exp(h(|X_t|)),$$

with constant $A = \exp(-B(X_0))$. Then

$$\sup_{t\in[0,T]} Z_t \leq A \sup_{t\in[0,T]} \exp(Ct)\exp(h(|X_t|)) \leq A\exp(CT)\exp\left(h\left(\max_{t\in[0,T]}|X_t|\right)\right).$$

By Lemma 2.3.10, below

$$\mathbb{E}\sup_{t\in[0,T]} Z_t \leq A\exp(CT)\,\mathbb{E}\exp(h(\max_{t\in[0,T]}|X_t|)) < \infty.$$

Then for a sequence of stopping times (τ_k) diverging to infinity such that $(Z_t^{\tau_k})_{0\leq t\leq T}$ is a martingale for all k, we have

$$\mathbb{E}Z_0 = \mathbb{E}Z_0^{\tau_k} = \mathbb{E}Z_t^{\tau_k} \to \mathbb{E}Z_t$$

as $k \to \infty$ by dominated convergence. □

Lemma 2.3.10. *Suppose $h\colon [0,\infty) \to [0,\infty)$ is non-decreasing. Let $N_T = \max_{0\le t\le T} |X_t|$ where (X_t) is a Brownian motion. Then*

$$\mathbb{E} \exp h(N_T) \le 4 \int_0^\infty \frac{1}{\sqrt{2\pi T}} \exp(h(x) - x^2/(2T))\, dx.$$

Proof. The maximum $M_T = \max_{0\le t\le T} X_t$ of a Brownian motion is distributed as the absolute value of a Brownian motion and thus has density function $\frac{2}{\sqrt{2\pi T}} \exp(-x^2/(2T))$, see Karatzas and Shreve (1991), Subsection 2.8. We have $\mathbb{P}(N_T \ge y) \le 2\mathbb{P}(M_T \ge y)$ from which the result follows. □

Finally we mention that Theorem 2.3.5 can be generalized in the following way to diffusions without a fixed end point.

Proposition 2.3.11. *If Assumption 2.3.4 is satisfied and B is bounded, then $\mathbb{P}_N$ converges weakly to $\mathbb{P}$.*

The proof follows the same steps of the one of Theorem 2.3.5. In this case we need to pay attention on B, as for unconditioned process, the final point is not fixed. If B is bounded, then Assumption 2.3.8 is satisfied. By Remark 2.3.9 also Assumption 2.3.1 is satisfied so that we can apply Theorem 2.3.2 for the change of measure. Finally, by the assumptions on b and B, the function Z is bounded and by dominated convergence we get that

$$\lim_{N\to\infty} \mathbb{E}_{\mathbb{Q}} |Z^N(X) - Z(X)| = 0.$$

2.4 A local Zig-Zag algorithm with subsampling for high-dimensional structured target densities

In Subsection 2.4.4 we will show that the task of sampling diffusion bridges boils down to the task of sampling a high-dimensional vector $\xi^N \in \mathbb{R}^M$ under the measure $\mathbb{P}_N^{u,v_T}$. Define by P_{ξ^N} the distribution of the vector ξ^N. Under the target measure,

$$P_{\xi^N}(\mathrm{d}\xi^N) = \pi(\xi^N)\mathrm{d}\xi^N.$$

We take the density π to be the M-dimensional invariant density (target density) for the Zig-Zag sampler. An efficient implementation of piecewise deterministic Monte Carlo methods, including the local and fully local Zig-Zag sampler can be found in Schauer and Grazzi (2021).

2.4.1 Subsampling technique

In our setting, the integral appearing in the Girsanov formula (2.10) poses difficulties when finding the root of equation (2.7) and would require numerical evaluation of the integral, hence also introducing a bias. By adapting the subsampling technique presented in Bierkens, Fearnhead, and Roberts (2019) (Section 4) we avoid this problem altogether (see Subsection 2.4.4). In general this technique requires

(a) unbiased estimators for $\partial_i \psi$ i.e. random functions $\partial_i \widetilde{\psi}_i(\xi, U_i)$ such that

$$E_{U_i}[\partial_i \widetilde{\psi}_i(\xi, U_i)] = \partial_i \psi(\xi),$$

for all i and ξ. These random functions create new (random) Poisson rates given by

$$\widetilde{\lambda}_i(t; \xi, \theta; U_i) = (\theta_i \partial_i \widetilde{\psi}(\xi(t), U_i))^+, \qquad i = 1, 2, ..., d, \tag{2.14}$$

whose evaluation becomes feasible and computationally more efficient compared to the original Poisson rates given by equation (2.6).

(b) upper bounds $\bar{\lambda}_i : (\mathbb{R}^+ \times \mathbb{R}^d \times \{-1, +1\}^d) \to \mathbb{R}^+$ for all $i = 1, ..., d$ such that for any point (ξ, θ) and $t \geq 0$ we have

$$P\left(\widetilde{\lambda}_i(t; \xi, \theta; U_i) \leq \bar{\lambda}_i(t; \xi, \theta)\right) = 1. \tag{2.15}$$

As we show in Algorithm 2 and in Section 2.5, these upper bounds are used for finding the roots of equation (2.7).

Algorithm 2 gives the algorithm for the Zig-Zag sampler with subsampling. It can be proved (see Bierkens, Fearnhead, and Roberts 2019) that the Zig-Zag sampler with subsampling has the same invariant distribution as its original and therefore does not introduce any bias. Note that we slightly modified the algorithm from Bierkens, Fearnhead, and Roberts (2019) in order to reduce its complexity. In particular it is sufficient to draw new waiting times and to save the coordinates only when the *if* condition at the *subsampling step* of Algorithm 2 is true.

2.4.2 Local Zig-Zag sampler

Subsection 3.1 of Bouchard-Côté, Vollmer, and Doucet (2018) proposes a local algorithm for the *Bouncy Particle Sampler* which is a process belonging to the class of piecewise deterministic Markov processes. Similar ideas apply to our setting.

Assumption 2.4.1. *The Poisson rate λ_i for a d-dimensional target distribution is a function of the coordinates $N_i \subset \{1, \ldots, d\}$,*

$$\lambda_i(s; \xi, \theta) = \lambda_i(s; \xi_k, \theta_k : k \in N_i).$$

Algorithm 2 d-dimensional Zig-Zag sampler with subsampling

procedure ZIGZAG_WS($\tau_{\text{final}}, \xi, \theta$)
  Initialise $k = 1$, $t = 0$
  $\tau_j \sim \text{IPP}(\bar{\lambda}_j(\cdot; \xi, \theta))$, $j = 1, ..., d$
  while $t \leq \tau_{\text{final}}$ **do**
    $\tau^*, i^* \leftarrow \text{findmin}(\tau_1, ..., \tau_d)$
    $\xi^{old} \leftarrow \xi$
    Update: $\xi \leftarrow \xi + \theta(\tau^* - t)$
    Update: $\Delta t \leftarrow \tau^* - t; \quad t \leftarrow \tau^*$
    $U_{i^*} \sim \text{Law}(U_{i^*}), V \sim \text{Unif}(0, 1)$
    if $V \leq \widetilde{\lambda}_{i^*}(0, \xi, \theta, U_{i^*})/\bar{\lambda}_{i^*}(\Delta t; \xi^{old}, \theta)$ **then** ▷ Subsampling step
      Save $\xi^{(k)} \leftarrow \xi, \quad t^{(k)} \leftarrow t$
      $k \leftarrow k + 1$
      $\theta_{i^*} \leftarrow -\theta_{i^*}$
      for $j \in \{1, \ldots, d\} \setminus \{i^*\}$ **do**
        $\tau_j \sim t + \text{IPP}(\bar{\lambda}_j(\cdot; \xi, \theta))$
      end for
    else
      $\tau_{i^*} \sim t + \text{IPP}(\bar{\lambda}_{i^*}(\cdot; \xi, \theta))$
    end if
  end while
  return Skeletons $(\xi^{(l)}, t^{(l)})_{l=1,2,...,k-1}$
end procedure

Recall that by the definition of λ_i (see equation (2.6)), the ith partial derivative of the negative loglikelihood determines the sets N_i. Now let us suppose that the first event time τ is triggered by the coordinate i so that at event time, the velocity θ_i is flipped. For all λ_k which are not function of this coordinate ($k \notin N_i$), we have

$$\lambda_k^{old}(\tau + s) = \lambda_k^{new}(s),$$

which implies that the waiting times drawn before τ, are still valid after switching the velocity i. This allows us to rescale the previous waiting time and reduce the number of computations at each step. The sets $N_1, ..., N_d$ are connected to the factorisation of the target distribution and define its conditional dependence structure. Indeed, take a d-dimensional target distribution with the following decomposition

$$\pi(\xi) = \prod_{i=1}^{N} \pi_i(\xi^{(i)})$$

where $\xi^{(i)} := \{\xi_j : j \in \Gamma_i\}$ and $\Gamma_i \subset \{1, 2, ..., N\}$ defines a subset of indices. We have that

$$-\partial_k \log(\pi(\xi)) = -\sum_{i=1}^{N} \partial_k \log \pi_i(\xi^{(i)}), \quad k = 1, ..., d$$

where the ith term in the sum is equal to 0 if $k \notin \Gamma_i$. Since the Poisson rates (2.6) are defined through the partial derivatives, the factorisation defines the sets $N_1, ..., N_d$ of Assumption 2.4.1.

Algorithm 3 shows the implementation of the local sampler which exploits any conditional independence structure so that the complexity of the algorithm scales well with the number of dimensions.

The local Zig-Zag sampler simplifies to independent one-dimensional Zig-Zag processes if the coefficients are pairwise independent coefficients, as it was the case in the example of sampling a Brownian motion or Brownian bridge (see Subsection 2.2.3). On the other hand, it defaults to Algorithm 1 when the dependency graph is fully connected, that is if $N_i = \{1, \dots, d\}, \forall i$.

Algorithm 3 d-dimensional local Zig-Zag sampler

Input: The bounds $\bar{\lambda}_i$ depend only on ξ_k, θ_k, for $k \in N_i$
procedure ZIGZAG_LOCAL($\tau_{\text{final}}, \xi, \theta$)
  Initialise $k = 1$, $t = 0$
  $\tau_j \sim \text{IPP}(\lambda_j(\cdot; \xi, \theta))$, $j = 1, ..., d$
  while $t \leq \tau_{\text{final}}$ **do**
    $\tau^*, i^* \leftarrow \text{findmin}(\tau_1, ..., \tau_d)$
    Update: $\xi \leftarrow \xi + \theta(\tau^* - t)$
    Update: $\theta_{i^*} \leftarrow -\theta_{i^*}$; $t \leftarrow \tau^*$
    Save $\xi^{(k)} \leftarrow \xi$; $t^{(k)} \leftarrow t$
    $k \leftarrow k + 1$
    for j in N_{i^*} **do** ▷ Local step
      $\tau_j \sim t + \text{IPP}(\lambda_j(\cdot; \xi, \theta))$
    end for
  end while
  return Skeletons $(\xi^{(l)}, t^{(l)})_{l=1,...,k-1}$
end procedure

2.4.3 Fully local Zig-Zag sampler

Combining the subsampling technique and the local ZZ can lead to a further reduction of the complexity of the algorithm. Indeed the bounds for the Poisson rates

might induce sparsity as $\bar{\lambda}_i$ can be function of few coordinates (see for example Subsection 2.5.2). This means that, after flipping θ_i, $\bar{\lambda}_j^{old}(\tau+t) = \bar{\lambda}_j^{new}(t)$ for almost all $j \neq i$ making the if statement in the *local step* of Algorithm 3 almost always satisfied and improving the efficiency of the algorithm. This means that, after flipping θ_i, we have that $\bar{\lambda}_j^{old}(\tau+t) = \bar{\lambda}_j^{new}(t)$ for almost all $j \neq i$ or, in other words, the cardinality of the set N_i in the *local step* of Algorithm 3 is small. Furthermore, the evaluation of $\widetilde{\lambda}_i(t,\xi,\theta)$ and $\bar{\lambda}_i(t,\xi,\theta)$ for $i = 1, 2, ..., d$ does not necessarily require to access the location of all the coordinates ξ_j so that, by assigning an independent time for each coordinate and updating only the coordinates needed for the evaluation of $\widetilde{\lambda}_i$ and $\bar{\lambda}_i$, the algorithm can be made more efficient. This is shown in the fully local ZZ sampler (Algorithm 4) where $\bar{N}_i, \widetilde{N}_i(U_i)$ define respectively the subset and the random subset of the coordinates required for the evaluation of $\bar{\lambda}_i(\cdot;\xi,\theta)$ and $\widetilde{\lambda}_i(\cdot;\xi,\theta;U_i)$.

2.4.4 Sampling diffusion bridges

In order to employ the Zig-Zag sampler to simulate from the bridge measure we choose the truncation level N in equation (2.2). Then, under $\mathbb{P}_N^{u,v_T}$

$$\pi(\mathrm{d}\xi^N) \propto Z^N(X) \exp\left(\frac{-\|\xi^N\|^2}{2}\right) \mathrm{d}\xi^N.$$

This is a straightforward consequence of the change of measure in (2.12) and the Lévy-Ciesielski construction.

We need to make one further assumption:

Assumption 2.4.2. *The drift b of the diffusion process is twice differentiable.*

Assumption 2.4.2 is necessary in order to compute the ξ_k-partial derivative of the energy function, which becomes

$$\partial_k \psi(\xi^N) = \frac{1}{2} \int_{S_k} h_k(s;\xi^N)\mathrm{d}s + \xi_k, \tag{2.16}$$

where

$$h_k(s;\xi^N) = \phi_k(s)\left(2b(X_s^N)b'(X_s^N) + b''(X_s^N)\right).$$

As the index k in the Faber-Schauder basis function gets larger, both the magnitude of ϕ_k and the size of its support decrease so that typically $\int h_k(s;\xi^N)\mathrm{d}s$ gets smaller and $\partial_k\psi(\xi) \approx \xi_k$ which corresponds to the partial derivative of the energy function of a standardized Gaussian random variable with independent components. This justifies one more time the intuition that for high levels i, the random variables ξ_{ij}, $j = 1, ..., 2^i - 1$ are approximately normally distributed and almost independent from the other random coefficients.

In order to avoid the evaluation of the integral appearing in (2.16) and the difficulty of drawing a Poisson time from its corresponding rate (2.6), we employ the subsampling technique. Considering ξ^N nonrandom, we take as an unbiased estimator for $\partial_k \psi(\xi_N)$ the (random) function

$$\frac{1}{2}|S_k| h_k(U_k; \xi^N) + \xi_k, \tag{2.17}$$

where $U_k \sim \text{Unif}(S_k)$ and as the bounding intensity rate

$$\bar{\lambda}_k(t, \xi^N, \theta^N) = \frac{1}{2}|S_k||\theta_k|\bar{\Phi}_k f(\xi^N(t)) + (\theta_k \xi_k(t))^+, \quad \xi^N \in \mathbb{R}^M, \tag{2.18}$$

where $\bar{\Phi}_k = \max_s(\phi_k(s))$ and $f(\xi^N) \geq \left|2b(X_s^{\xi^N})b'(X_s^{\xi^N}) + b''(X_s^{\xi^N})\right|$, $\forall s \in [0,T]$, $\xi^N \in \mathbb{R}^M$. The subsampling technique avoids the numerical computation of the time integral (2.16), thus avoiding a numerical bias and reducing the computational effort from $\mathcal{O}(T)$ (for fixed discretization size) to $\mathcal{O}(1)$. The variance of this unbiased estimator can be reduced by averaging over multiple independent uniform draws or similar strategies (see for example Section 2.5.4), albeit at the cost of additional computations. In Section 2.5 we show specifically for each numerical experiment how we derived the Poisson upper bounds $\bar{\lambda}_i$.

The compact support of the Faber-Schauder functions induce a sparse dependency structure on the target measure π. Indeed, X_t only depends on those values of $\xi_{l,k}$ for which $t \in S_{l,k}$. See Figure 2.4 for an illustration. It is easy to see that this implies that $\frac{\partial \psi(\xi^N)}{\partial \xi_{(i,j)}}$ depends only on those $\xi_{(k,l)}$ for which the interior of $S_{i,j} \cap S_{k,l}$ is non-empty. In particular, define the relation $\xi_{i,j} \ll \xi_{k,l}$ to hold if $S_{k,l} \subset S_{i,j}$. If this happens, then we refer to $\xi_{i,j}$ as the *ancestor* of $\xi_{k,l}$ (and conversely $\xi_{k,l}$ as the *descendant*). Then the sets in Assumption 2.4.1 (using double indexing) can be chosen as $N_{i,j} = \{\xi_{h,d} \colon \xi_{h,d} \ll \xi_{i,j} \vee \xi_{h,d} \gg \xi_{i,j}\}$ with cardinality $|N_{i,j}| = 2^{N-i+1} + i - 1$, where N is the truncation level. Formally, $N_{i,j}$ are the neighborhoods of the interval graph induced by $((S_{i,j} \colon i \in \{1, 2, \ldots, N\},\ j \in \{0, 1, \ldots, 2^i - 1\}))$ with vertices $\{(i,j) \colon i \in \{1, 2, \ldots, N\},\ j \in \{0, 1, \ldots, 2^i - 1\}\}$, where there is an edge between (i,j) and (l,k) if the interior of $S_{i,j} \cap S_{k,l}$ is non-empty (see Figure A.1). The factorization of the partial derivatives leads to a specific dependency structure of the coefficients under the target diffusion bridge measure: the coefficient $\xi_{i,j}$ is conditionally independent of the coefficient $\xi_{k,l}$ if $S_{i,j} \cap S_{k,l} = \emptyset$ conditionally on the set of common ancestors $(\xi_{m,n} \colon \xi_{m,n} \ll \xi_{i,j} \wedge \xi_{m,n} \ll \xi_{k,l})$. This argument is made more formal by decomposing the likelihood function in Appendix A.1.

2.5 Numerical results

We show numerical results for three representative examples. In general, when applying our method, we start from a model (2.1), devise a representation of the

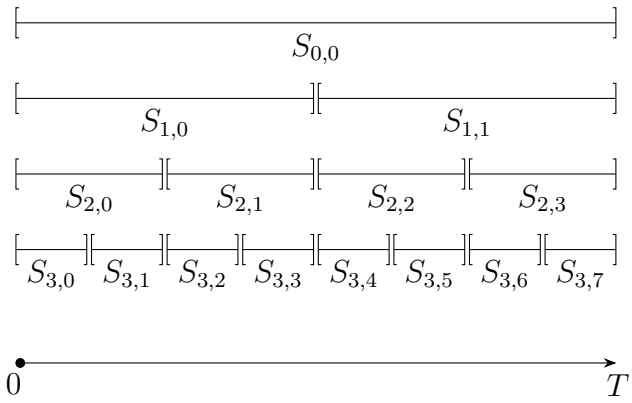


Figure 2.4: Support of the Faber-Schauder functions $(\phi_{i,j} : i \in \{0, 1, \dots, N\}, j = \{0, 1, \dots, 2^i - 1\}$ with $N = 3$. The coefficient $\xi_{i,j}$ is independent of the coefficient $\xi_{k,l}$ conditionally on the set of common ancestors $(\xi_{m,n} \colon S_{m,n} \cap S_{i,j} \neq \emptyset \wedge S_{m,n} \cap S_{k,l} \neq \emptyset)$ if $S_{i,j} \cap S_{k,l} = \emptyset$.

approximate diffusion bridge (2.12), that we sample using generic implementations of algorithms 1-4 from our package, which are easily adapted to the task of sampling the coefficients of the Faber-Schauder expansion. To this end, we provide the k-th partial derivative of the energy function (2.16) or an upper bound to the Poisson rate (2.18) as argument for the sampler, as well as the sets $N_{i,j}$ as given in Section 2.4.4. The reader is referred to the file `faberschauder.jl` in the public repository `https://github.com/SebaGraz/ZZDiffusionBridge/src` for the implementation of the expansion and for the generic implementation of the different variants of the Zig-Zag sampler to our package (Schauer and Grazzi 2021).

The first class of diffusion processes considered are diffusions with linear drift function (Subsection 2.5.1). This is a special case, where our method does not require the subsampling technique described in Subsection 2.4.1 and only Algorithm 3 has been employed. Notice that for this class, the transition kernel of the conditioned process is known. In Subsection 2.5.2, we apply our method for diffusions which substantially differ from Brownian motions, being highly non-linear and multimodal and therefore creating challenging bridge distributions for standard MCMC. Here we use the the fully local algorithm (Algorithm 4). In the specific example considered, the implementation of the Zig-Zag sampler is facilitated by the drift function and its derivatives being bounded and therefore a bounded Poisson rate for the subsampling technique is available. In view of this, we choose for the third numerical experiment a diffusion with unbounded drift (Subsection 2.5.3). For all the models, Assumptions 2.3.1, 2.3.4 and 2.4.2 are immediate to verify and Assumption 2.4.1 is satisfied. For each experiment, the burn-in $\tau_{\text{burn-in}}$ and final clock τ_{final} are manually tuned by inspecting the trace of ξ^N and ensuring that the process reached station-

arity before $\tau_{\text{burn-in}}$ and fully explore the state space before the final clock τ_{final}. The computations are performed with a conventional laptop with a 1.8GHz intel core i7-8550U processor and 8GB DDR4 RAM. We wrote the program in Julia 1.4.2 which allows profiling and optimizing the code for high performance. The program is publicly available on GitHub at `https://github.com/SebaGraz/ZZDiffusionBridge` where the reader can follow the documentation to reproduce the results.

2.5.1 Linear diffusions

A linear stochastic differential equation conditioned to hit a final point v_T has the form

$$\mathrm{d}X_t = (\alpha + \beta X_t)\mathrm{d}t + \mathrm{d}W_t, \qquad X_0 = u, X_T = v_T \tag{2.19}$$

for some $(\alpha, \beta) \in \mathbb{R}^2$. Assumptions 2.3.1, 2.3.4 and 2.4.2 can be easily verified. In this case the energy function of the target distribution is

$$\psi(\xi^N) = C_1 - \ln(Z^N(X)) + \frac{\|\xi^N\|^2}{2} = C_2 + \frac{1}{2}\int_0^T \left(\beta^2 \left(X_t^{\xi^N}\right)^2 + 2\alpha\beta X_t^{\xi^N}\right) \mathrm{d}t + \frac{\|\xi^N\|^2}{2},$$

for some constant C_1, C_2. Note that ψ is a quadratic function of ξ, which means that the target density is still Gaussian under $\mathbb{P}_N^{u,v_T}$. It follows that

$$\partial_{\xi_k}\psi(\xi^N) = \int_{t\in S_k} \phi_k(t)\left(\beta^2\left(\bar{\bar{\phi}}(t)u + \bar{\phi}(t)v_T/\sqrt{T} + \sum_{j\in N_k}\phi_j\xi_j\right) + \alpha\beta)\right)\mathrm{d}t + \xi_k.$$

Interchanging the integral and the sum, this becomes

$$\partial_{\xi_k}\psi(\xi^N) = \beta^2\left(\bar{\bar{\Phi}}_k u + \bar{\Phi}_k v_T/\sqrt{T} + \sum_{j\in N_k}\Phi_{jk}\xi_j\right) + \alpha\beta\Phi_k + \xi_k,$$

where $\Phi_k = \int \phi_k \mathrm{d}t$, $\Phi_{jk} = \int \phi_k\phi_j \mathrm{d}t$, $\bar{\Phi}_k = \int \bar{\phi}\phi_k \mathrm{d}t$ and $\bar{\bar{\Phi}} = \int \bar{\bar{\phi}}\phi_k \mathrm{d}t$. This is a linear function of ξ^N and, for each i, the event times with rates λ_i, see (2.6), can be directly simulated without upper bounds. Figure 2.5 shows samples from the resulting diffusion bridge measure with $\alpha = -5, \beta = -1$ obtained with this method running the Zig-Zag sampler for $\tau_{\text{final}} = 1000$, with a burn-in time of $\tau_{\text{burn-in}} = 10$. The closed form of the expansion of linear processes, or more generally, reciprocal linear processes, with the Faber-Schauder basis was also found and used in Meulen, Schauer, and Waaij (2018) for the problem of nonparametric drift estimation of diffusion processes. The results are validated by computing analytically the density of the random variable $X_{T/2}$ (which, for the linear case, is known in close form) and comparing this with its empirical density obtained from one sample of the Zig-Zag process (see Figure 2.7, left panel).

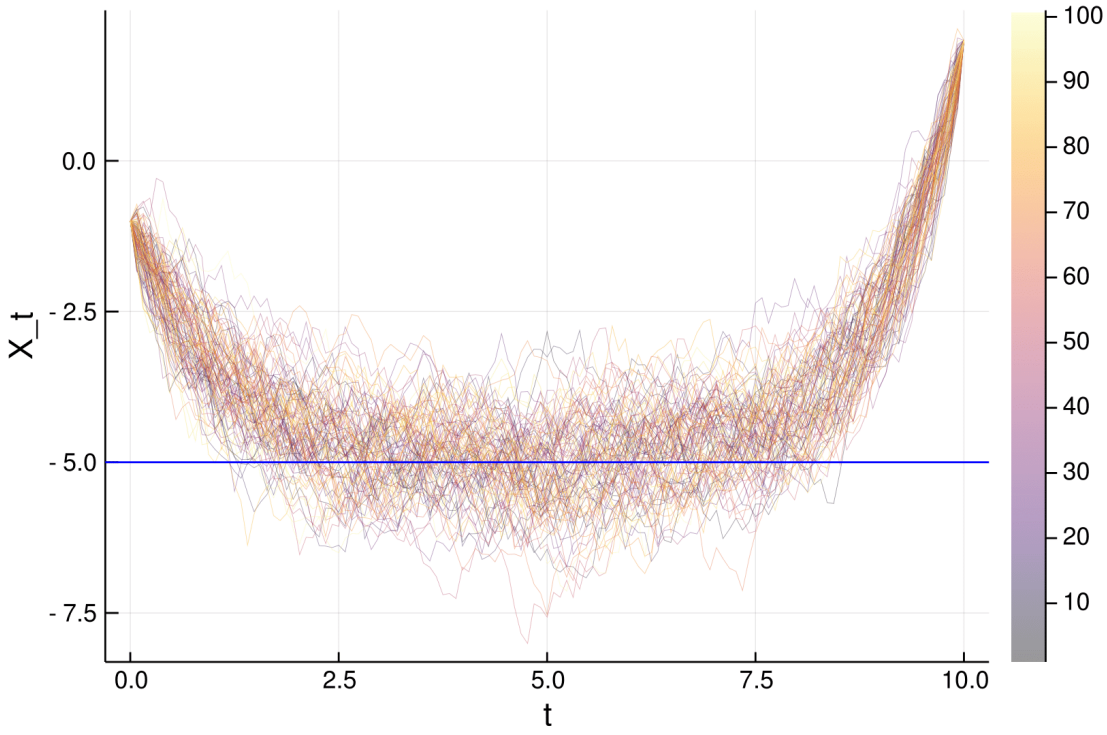


Figure 2.5: Simulation of the diffusion bridge measure (100 samples) given by equation (2.19) starting at -1.0 and conditioned to hit 2.0 at $T = 10$. $\alpha = -5.0, \beta = -1.0$ which is equivalent to a mean reverting process with mean reversion at $x = -5$ (straight line). The truncation level is $N = 6$, final clock $\tau_{\text{final}} = 1000$ and burn-in $\tau_{\text{burn-in}} = 10$.

2.5.2 Non-linear multi-modal diffusions

The stochastic differential equation considered here has the form

$$\mathrm{d}X_t = \alpha \sin(X_t)\mathrm{d}t + \mathrm{d}W_t, \qquad X_0 = u, X_T = v_T \tag{2.20}$$

for some $\alpha \geq 0$. When $\alpha = 0$ the process is a standard Brownian motion while for positive α, the process is attracted to its stable points $(2k-1)\pi$, $k \in \mathbb{N}$. Assumption 2.3.1, 2.3.4, 2.4.2 follow from drift, its primitive and its derivative being globally bounded. Fixing N, the energy function is given by

$$\psi(\xi^N) = \frac{\alpha}{2}\int_0^T \left(\alpha \sin^2(X_t^{\xi^N}) + \cos(X_t^{\xi^N})\right)\mathrm{d}t + \frac{\|\xi^N\|^2}{2}.$$

Using trigonometric identities, we obtain that

$$\partial_{\xi_k}\psi(\xi^N) = \frac{1}{2}\int_{S_k} \phi_k(t)\left(\alpha^2 \sin\left(2X_t^{\xi^N,k}\right) - \alpha \sin\left(X_t^{\xi^N,k}\right)\right)\mathrm{d}t + \xi_k$$

where $X_t^{\xi^N,k} := \bar{\bar{\phi}}(t)u + \bar{\phi}(t)v_T/\sqrt{T} + \sum_{j\in N_k}\phi_j(t)\xi_j$. To avoid the need to find the roots of equation (2.7) we apply the subsampling technique described in Subsection 2.4.1. Since the drift and its derivatives are bounded, we can easily find the following upper bound for (2.14):

$$\bar{\lambda}_k(t) = |\theta_k|a_1 + (\theta_k\xi_k(t))^+, \tag{2.21}$$

with $a_1 = \bar{\Phi}_k S_k(\alpha^2+\alpha)/2$, $\bar{\Phi}_k = \max(\phi_k)$ and $\xi_k(t) = \xi_k+\theta_k t$. In this case, the upper bound $\bar{\lambda}_i$ is a function only of the coefficient ξ_i. Figure 2.6 shows the results obtained with this method setting $\alpha = 0.7$. For this diffusion, the non-linearity and multiple modes make the mixing of the Zig-Zag sampler slower so we set $\tau_{\text{final}} = 10000$ and burn-in $\tau_{\text{burn-in}} = 10$.

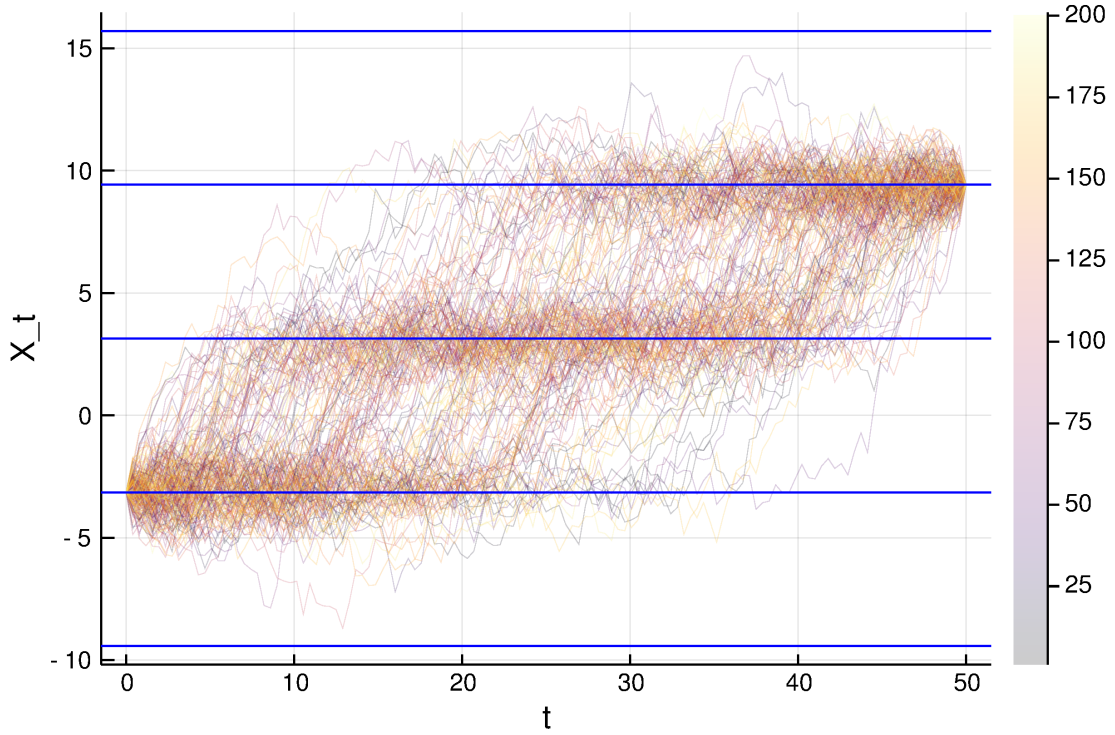


Figure 2.6: Simulation of the diffusion bridge measure (200 samples) given by equation (2.20) with $\alpha = 0.7$ starting at $-\pi$ at time 0 and hitting 3π at $T = 50$. Truncation level $N = 6$, final clock $\tau_{\text{final}} = 10000$ and burn-in 10. The straight horizontal lines are the attraction points of the process.

Analysing the goodness of the empirical diffusion bridge distribution obtained is a difficult task since the true conditional distribution is not known in a tractable form. We start by checking if some geometrical properties of the diffusion bridge

distributions are preserved in the simulations. For example, in Figure 2.6, it can be noticed that the diffusion is attracted to the stable points $\pm\pi, \pm 3\pi, ...$, and symmetric (geometrically speaking, after rotation) around the vertical axes $t = T/2$. We furthermore validate our method by simulating forward diffusion processes, using Euler discretization in a fine grid, and retaining only the paths which end in a ϵ-ball of a certain point at time T (ϵ-ball forward simulation). If the final point is such that the probability of ending in this ϵ-ball is high enough, we can create in this way a sample from the approximated bridge and compare it to the samples obtained from the Zig-Zag. The right panel of Figure 2.7 shows the joint empirical distribution with the two methods of the first quarter and third quarter random variables. Finally, Figure 2.8 illustrates that the marginal distribution of the coefficients in higher levels is approximately Gaussian and the non-linearity of the process is absorbed by the coefficients in low levels.

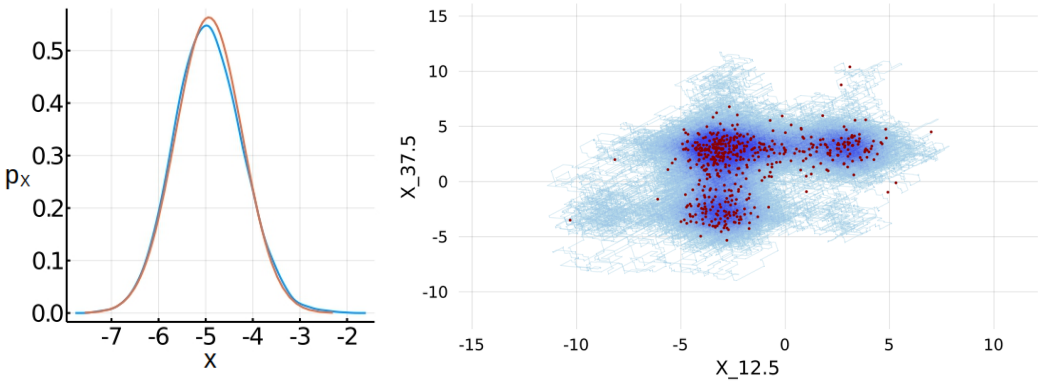


Figure 2.7: On the left panel: comparison between empirical distribution (blue line, computed with a kernel estimator) and the exact distribution (red line) of the midpoint random variable X_5 for the linear diffusion (equation 2.19) with $a = -5$ and $b = -1$. The empirical distribution has been extracted from the same experiment shown in Figure 2.5. On the right panel: comparison between the joint distribution of the variables $X_{T/4}$ and $X_{3T/4}$ of the process given in equation (2.20) starting at $-\pi$ and hitting π at $T = 50$. The scatter plot with red dots are obtained with ϵ-ball Euler simulation with $\epsilon = 0.1$ and discretization $\Delta t = 0.0005$ while the blue continuous path is the Zig-Zag path.

2.5.3 Diffusions with unbounded drift

Here we consider *stochastic exponential logistic models*. For this class, the process grows exponentially with rate r until it reaches its saturation point K. Its dynamics are perturbed by noise which grows as the population grows. The resulting stochastic

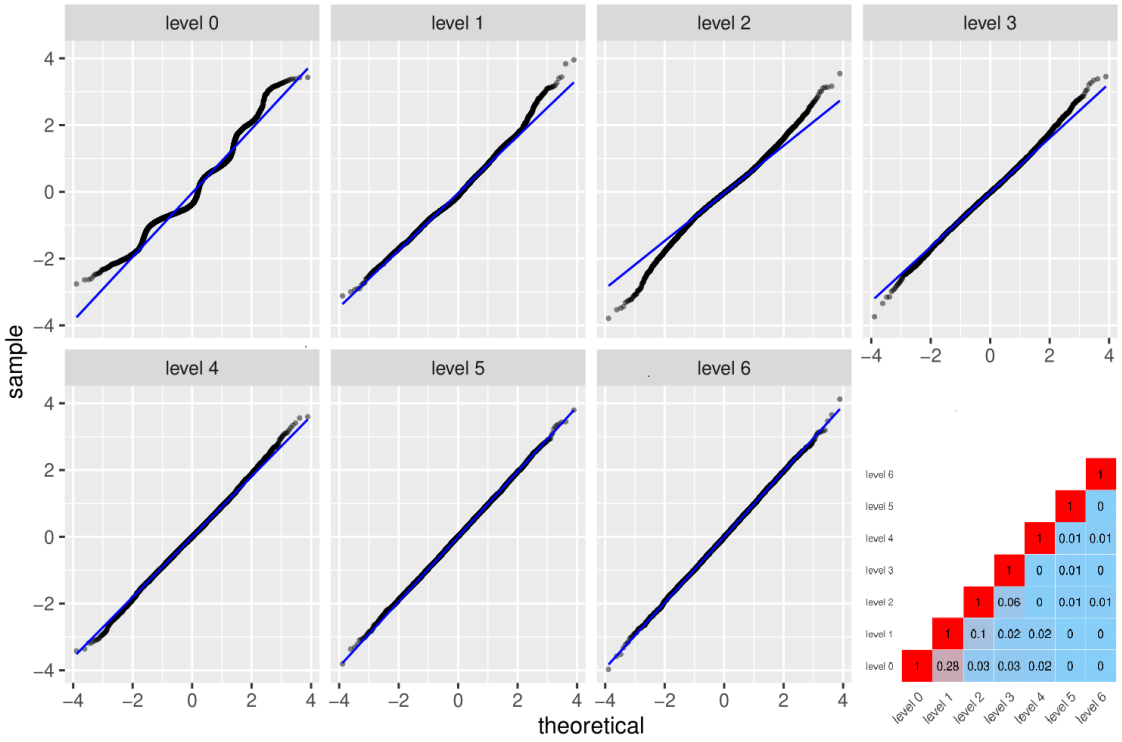


Figure 2.8: Q-Q (quantile-quantile) plot against standard normal distributions of the sample path of 7 coefficients respectively at level $0, 1, 2, 3, 4, 5, 6$ targeting the conditional bridge measure given by equation (2.20) with $\alpha = 0.7$ and initial point $u = 0$ and final point $v = 0$ at $T = 100$. On the bottom right panel, the heatmap of the absolute value of the sample correlation between the coefficients at different levels. The blue straight lines correspond to the marginal measures of the coefficients relatively to a Brownian bridge.

differential equation takes the form

$$\mathrm{d}Y_t = rY_t(1 - Y_t/K)\mathrm{d}t + \beta Y_t \mathrm{d}W_t, \qquad X_0 = u > 0, \quad X_T = v_T > 0. \tag{2.22}$$

We can transform the process in order to get a new process with unitary diffusivity $\sigma = 1$ (Lamperti transform with $X_t = -\log(Y_t)/\beta$). The transformed differential equation becomes

$$\mathrm{d}X_t = (c_1 + c_2 e^{-\beta X_t})\mathrm{d}t + \mathrm{d}W_t, \qquad X_0 = -\log(u)/\beta, \quad X_T = -\log(v)/\beta.$$

with $c_1 = \beta/2 - r/\beta$ and $c_2 = r/(\beta K)$. Note that the drift function b of the transformed process is not global Lipschitz continuous. Nevertheless Assumptions 2.3.4 and 2.4.2 are satisfied and by Remark 2.3.9, also Assumption 2.3.1 is verified. In

this case, the partial derivative of the energy function is given by

$$\partial_k \psi(\xi^N) = \frac{1}{2} \int_{S_k} \phi_k(s) \left(a_1 e^{-\beta X_s^{\xi^N}} - a_2 e^{-2\beta X_s^{\xi^N}} \right) \mathrm{d}s + \xi_k,$$

where $a_1 = 2r^2/(\beta K)$, $a_2 = a_1/K$. As before, it is not possible to simulate directly the first event time using the Poisson rates given by equation (2.6). The subsampling technique requires an upper bound for the unbiased estimator (2.14). Define the following quantities

$$b_k^{(1)} := \inf_{s \in S_k} \left\{ \bar{\bar{\phi}}(s) u_0 + \bar{\phi}(s) v_T / \sqrt{T} + \sum_{i \in N_k} \phi_i(s) \xi_i \right\}, \qquad b_k^{(2)} := \inf_{s \in S_k} \left\{ \sum_{i \in N_k} \phi_i(s) \theta_i \right\}.$$

For any $a, b, c \in \mathbb{R}$, $(a+b+c)^+ \leq (a)^+ + (b)^+ + (c)^+$ and hence a valid upper bound for the Poisson rate (2.14) is given by

$$\bar{\lambda}_k(t) = \lambda_k^{(1)}(t) + \lambda_k^{(2)}(t) + \lambda_k^{(3)}(t) \tag{2.23}$$

with

$$\lambda_k^{(1)}(t) = \max\left(0, \theta_k \xi_k(t)\right),$$
$$\lambda_k^{(2)}(t) = \max\left(0, \frac{1}{2} \theta_k \bar{\phi}_k S_k z_k^{(1)} e^{-\beta_k^\star t}\right),$$
$$\lambda_k^{(3)}(t) = \max\left(0, -\frac{1}{2} \theta_k \bar{\phi}_k S_k z_k^{(2)} e^{2\beta_k^\star t}\right)$$

and

$$z_k^{(1)} = a_1 \exp(-\beta b_k^{(1)}),\ z_k^{(2)} = z_k^{(1)} \exp(-\beta b_k^{(1)}),\ \beta_k^\star = -\beta b_k^{(2)},\ \bar{\phi}_i = \max_s \phi_i(s).$$

Using the *superposition theorem* (see for example Grimmett and Stirzaker 2001), we can simulate a waiting time with Poisson rate (2.23) by means of simulating three waiting times according to the Poisson rates $\lambda_k^{(1)}, \lambda_k^{(2)}, \lambda_k^{(3)}$ and then take the minimum of the three realizations. Since at any time $t > 0$, either $\lambda_k^{(2)}(t)$ or $\lambda_k^{(3)}(t)$ is 0, we just need to evaluate two waiting times. Figure 2.9 shows the results obtained with our method for this process. The final clock of the Zig-Zag sampler is set to $T^\star = 1000$ and initial burn-in time $\tau_{\text{burn-in}} = 10$.

2.5.4 Numerical comparisons

In this section we benchmark the fully local Zig-Zag sampler against the Metropolis-adjusted Langevin algorithm (MALA) (Roberts and Tweedie 1996), Hamiltonian Monte Carlo (HMC) (Duane et al. 1987) and another well known PDMP, the Bouncy

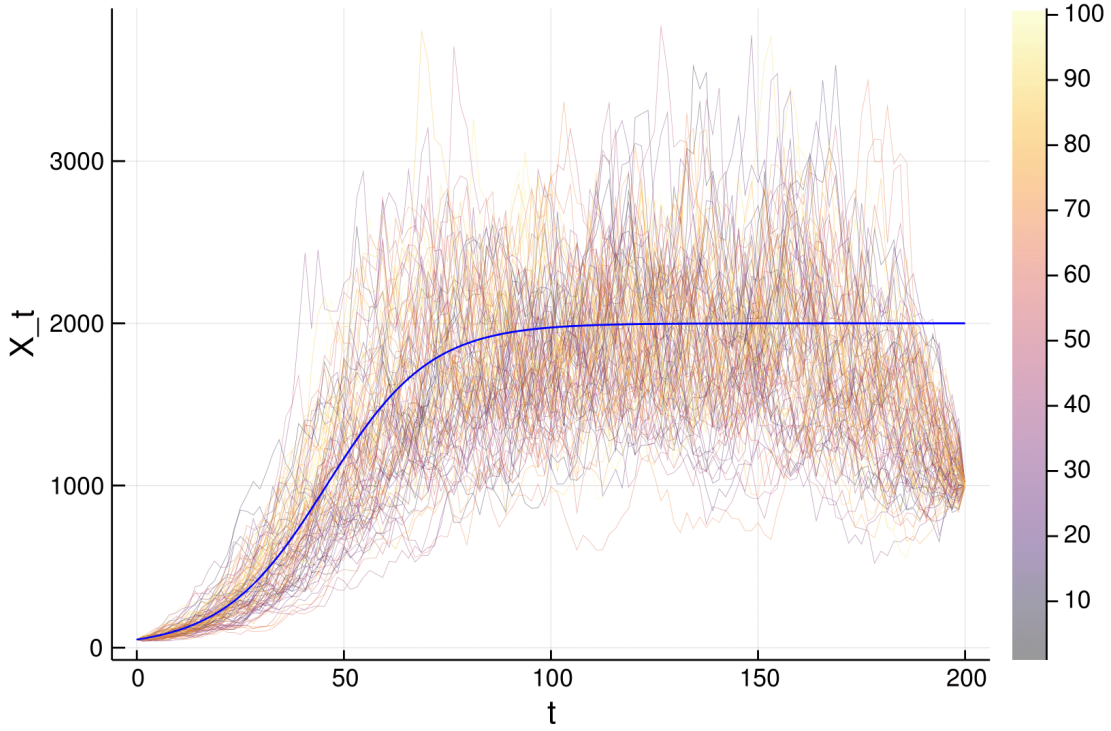


Figure 2.9: Simulation of the diffusion bridge measure (100 samples) given by the logistic growth model (equation (2.22)) with parameters $K = 2000, r = 0.08, \beta = 0.1$, starting at the value 50 and hitting 1000 at time 200. Truncation level $N = 6$, final clock $\tau_{\text{final}} = 1000$ and burn-in $\tau_{\text{burn-in}} = 10$. The blue smooth line is the solution of the deterministic logistic model without final condition.

particle sampler (Bouchard-Côté, Vollmer, and Doucet 2018). The Bouncy Particle sampler can use the exact subsampling technique in a very similar way as explained in Subsection 2.4.1. According to the scaling limit results obtained in Bierkens, Kamatani, and Roberts (2018), the Zig-Zag is more efficient compared to the Bouncy Particle sampler in a high dimensional setting when the conditional dependency graph corresponding to the target measure exhibits sparsity (which clearly is the case here). The MALA sampler is a well known discrete-time Markov chain Monte Carlo method which performs informed updates through the gradient of the target distribution. HMC is considered a state-of-the-art algorithm. In contrast to PDMPs, for HMC and MALA the gradient needs to be fully evaluated and no subsampling methods can be exploited. Thus, the integral in (2.16) needs to be computed numerically, introducing bias. Furthermore, contrary to PDMPs, the resulting Markov chain is reversible. We study the performance of the samplers for the stochastic differential equation (2.20) with $u, v = 0$ and the time horizon $T = 100$ and we let

α vary. As α increases, the target distribution on the coefficients presents higher peaks and valleys and is therefore a challenging distribution for general Markov chain Monte Carlo methods. We fix the refreshment rate of the Bouncy Particle sampler to 1 to avoid a degenerate behaviour and implement the MALA algorithm with adaptive step size over 250,000 iterations. We used the automatically tuned dynamic integration time HMC Algorithm (Betancourt 2018) with $3,000$ iterations and with diagonal mass matrix and integrator step size both adaptively tuned in a warm-up phase of $2,000$ iterations, with the latter adapted using a dual-averaging algorithm (Hoffman and Gelman 2014) with target acceptance statistic of 0.8. The algorithm is provided in the package `AdvancedHMC.jl` (Ge, Xu, and Ghahramani 2018). The integral appearing in the gradient of the energy function is computed for the MALA sampler and for the HMC sampler numerically with a simple Euler integration scheme over 2^{N+1} points, where N is the truncation level which is fixed to 6 for all the experiments. The final clock for the PDMPs is $T' = 25,000$. We also include the numerical results of two variants of the Zig-Zag sampler:

(ZZv1) where the partial derivative in (2.16) is estimated by averaging over multiple independent realizations of (2.17), with the number of realizations is proportional to the length of the range of the integral in (2.16);

(ZZv2) where the partial derivative in (2.16) is estimated by decomposing the range of the integral into N subintervals (with N proportional to the length of the range of the integral) and evaluating the integrand at a random point drawn inside each subinterval.

These variants of the Zig-Zag have been proposed after noticing that the coefficients at low levels are the ones deviating the most from normality and the partial derivative with respect to those coefficients have larger support. This suggests that refining the estimates of the partial derivative of the energy function only with respect to those coefficients can be beneficial and improve the performances of the PDMPs. Figure 2.10 shows the results obtained. The fully local Zig-Zag and its variants always outperform the Bouncy Particle sampler, the MALA and the HMC with respect to the statistics considered, namely the mean, median and minimum of the effective sample size computed for each coefficient of the Faber-Schauder expansion and the effective sample size of the coefficient $\xi_{0,0}$, which gives the middle point $X_{T/2}$ and, as shown in Figure 2.10, is one of the most difficult coefficients to sample.

2.6 Extensions

In this section we briefly sketch the extension of the approach presented in Section 2.3 to a class of multi-dimensional diffusion bridges. Then we study the scaling

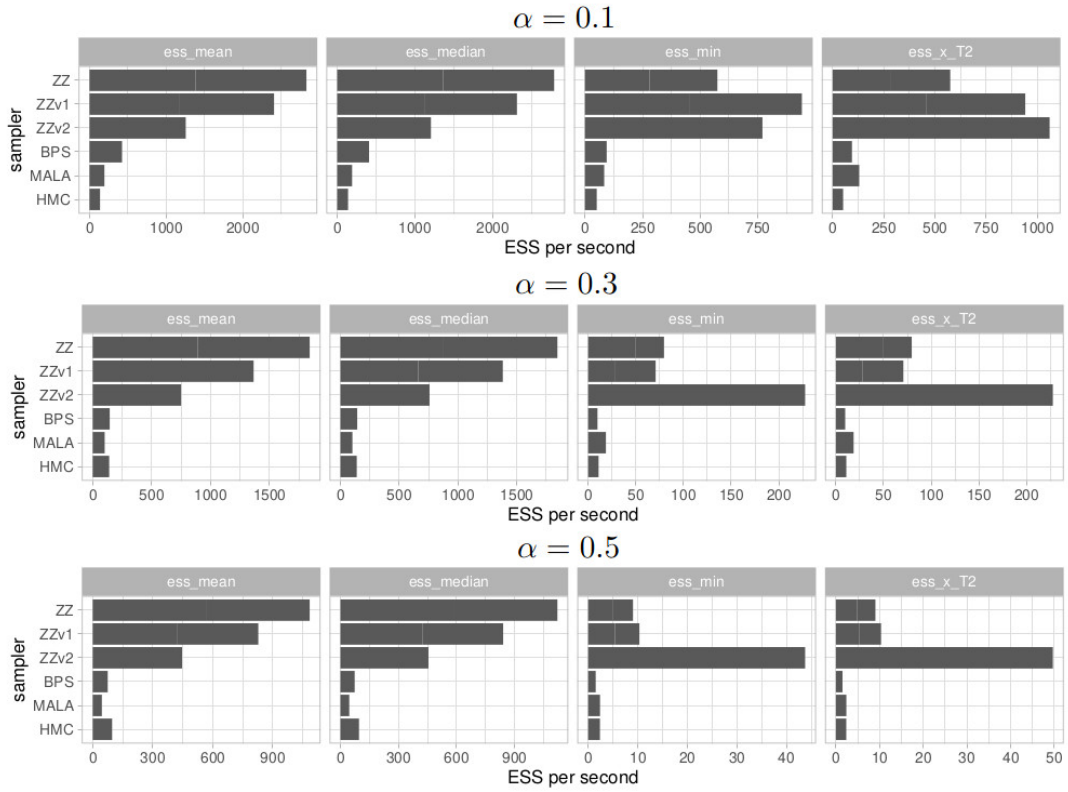


Figure 2.10: Performance comparison of the fully local Zig-ZaZ (ZZ), its variants (ZZv1 and ZZv2), the Bouncy Particle sampler with Subsampling (refreshment rate set to 1), MALA and HMC sampler. The performance measure considered here are respectively the effective sample size (ESS) of the middle point $X_{T/2}$, the median and the minimum of the ESS over the dimension of the coefficients of the expansion. The target diffusion bridge with drift $b(x) = \alpha \sin(x)$ with $u, v = 0$ and $T = 50$ and truncation level $N = 6$. The final clock for the PDMPs is set to $T' = 25000$, the number of iterations for the MALA is set to be 250000 with adaptive time step targeting the acceptance rate 0.6 (Roberts and Rosenthal 1998) and the number of iteration for the HMC is 3000 with the algorithm fine-tuned by the package `AdvancedHMC.jl`. All the quantities are normalized by the runtime of execution. The asymptotic variance estimate used for computing the ESS is obtained using batch means. Notice that, while the subsampling technique adopted for the piecewise deterministic Monte Carlo methods does not introduce bias on the target distribution, the numerical integration adopted for the MALA and HMC samplers introduces bias on the target distribution.

properties of the algorithm with respect to three quantities: the time horizon of the

diffusion bridge T, the truncation level N and the dimensionality of the diffusion bridge d.

2.6.1 Multivariate diffusion bridge

Consider a d-dimensional diffusion bridge given the stochastic differential equation

$$\mathrm{d}X_t = \nabla B(X_t)\mathrm{d}t + \mathrm{d}W_t, \quad X_0 = u, X_T = v_T, \quad u, v_T \in \mathbb{R}^d,$$

where $(W_t)_{t\geq 0}$ is a d-dimensional Wiener process and $\nabla B : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a conservative vector field, i.e. the gradient of some scalar-valued function B. Denote its law by $\mathbb{P}^{u,v_T}$. Similarly to equation (2.10), under mild assumption on $\nabla B(X_t)$, we can write the change of measure between $\mathbb{P}^{u,v_T}$ and the standard d dimensional Wiener bridge measure $\mathbb{Q}^{u,v_T}$ as

$$\frac{\mathrm{d}\mathbb{P}^{u,v_T}}{\mathrm{d}\mathbb{Q}^{u,v_T}}(X) = C \exp\left\{ B(X_T) - B(X_0) - \frac{1}{2}\int_0^T \|b(X_t)\|^2 + \Delta B(X_t)\,\mathrm{d}t \right\},$$

where $b = \nabla B$, ΔB is the Laplacian of B and C is a normalization constant which depends on u, v_T and T. It is straightforward to derive an equivalent approximated measure as done in equation (2.12) and prove Theorem 2.3.11 in the multi-dimensional setting. In this case the d dimensional diffusion bridge measure is approximated by the same truncated expansion of equation (2.2) with coefficients $\xi_{i,j}, i = 0, ..., N;\ j = 0, ..., 2^N$ which now are d dimensional random vectors. The total dimensionality of the target density for diffusion bridges becomes $d(2^{N+1} - 1)$. Similarly to the one dimensional case, Proposition A.1.1 holds (the proof follows in a similar fashion of the proof of Proposition A.1.1 and is omitted for brevity). The Poisson rates $\lambda^k_{i,j}$ (where, $k \in \{1, ..., d\}$ defines the coordinate of the d dimensional process) are functions of the sets $N^k_{i,j}$ which have maximum admissible size $|N^k_{i,j}| = d(2^{N-i+1} + i - 1) \leq d(2^{N+1} - 1)$ so that Assumption 2.4.1 holds.

2.6.2 Scaling for large T, N, d

The following scaling analysis serves as preliminary work for future explorations. The expected run time of the fully local Zig-Zag sampler (Algorithm 4) is intimately related with the number of Poisson event times for a fixed final clock τ_{final} and the conditional independence structure appearing in the target measure. The former is determined by the size of the Poisson bounding rates $\bar{\lambda}_1, ..., \bar{\lambda}_M$ while the latter is defined by the sets $N_1, ..., N_M$ and determine the complexity of the *local step* of Algorithm 3.

Remark 2.6.1. *For a fixed position and velocity, the Poisson bounding rates used in the Zig-Zag sampler with subsampling (Algorithm 2) for diffusion bridges are of*

the form $\bar{\lambda}_{i,j} = C_1 T^{3/2} 2^{-3i/2} + C_2$, $i = 0, 1, ..., N$; $j = 0, 1, ..., 2^i - 1$, *for some terms* C_1 *and* C_2 *which do not depend on* i *and* T.

Proof. For every $i = 0, 1, ..., N$; $j = 0, 1, ..., 2^i - 1$, the time horizon T and scaling index i enter in the bounding rates of (2.18) through the terms $S_{i,j}$ and $\bar{\phi}_{i,j}$. The first term is of $\mathcal{O}(T 2^{-i})$ and the second one is of $\mathcal{O}(\sqrt{T} 2^{-i/2})$. □

Proposition 2.6.1 helps understanding how the complexity of the algorithm scales as T grows and as the truncation level N grows. As T grows, the Poisson rates increase with order $T^{3/2}$ so that the total number of Poisson events for a fixed Zig-Zag clock increases with the same order.

Furthermore, as the truncation level N grows, the change of measure affects less and less the coefficients in high levels and the partial derivative of the energy function goes to zero with rate $2^{-3N/2}$) implying that the for large N, $\bar{\lambda}_{N,j} \approx C_2 = (\xi_{N,j}\theta_{N,j})^+$ (which is the Poisson rate for the Brownian bridge). As a consequence, the Poisson processes of the coefficients in high levels (i large) will be approximately independent with all the other coefficients and not function of the level i so that the complexity of Algorithm 4 scales approximately linearly with the number of mesh points. This is opposed to the standard Zig-Zag algorithm (Algorithm 1) which does not take advantage of the approximate independence of the coefficients in high levels so that the $2^{N+1} - 1$ waiting times have to be renovated at every reflection of each coefficient.

The scaling result under mesh refinement (when N grows) is unsatisfactory as the algorithm deteriorates when the resolution of the path increases. A partial solution can be obtained by letting the absolute value of the marginal velocities $|\theta_{N,j}|$ decrease as N increases. This would enhance the scaling property of the algorithm under mesh refinement at the cost of a slow mixing of high level components. An alternative solution is considered in Bierkens et al. (2020) where the authors enhance the scaling property of the algorithm by replacing the Zig-Zag algorithm with the *Factorised Boomerang sampler*. The Factorised Boomerang sampler differs form the Zig-Zag by having curved trajectories which are invariant to a prescribed Gaussian measure. This allows the process to sample from the Gaussian measure (Brownian bridge measure) at barely no cost. However, the main drawback of the factorized Boomerang sampler is the current limiting techniques for simulating Poisson times given the curved trajectories which lead to Poisson upper bounds which are not tight.

Finally, when the dimensionality of the diffusion bridge is $d \gg 1$, both the dimensionality of the target density of the Zig-Zag sampler and the sets $N_{i,j}^k$ for $i = 0, ..., N$; $j = 0, ..., 2^i - 1$; $k = 1, ..., d$ grow linearly with d so that, in general, we expect the computational time to grow with rate d^2. When the drift of the multidimensional bridge presents a sparse structure, i.e. not all coordinates of the differential equation interact directly with each other, as common in the high dimensional case arising from discretized stochastic partial differential equations (e.g.

Mider, Schauer, and Meulen 2020, Section 6), the size of those sets reduces considerably until the extreme case of d independent diffusion bridges where the sets $N_{i,j}^k$ are not anymore a function of d and clearly the complexity grows linearly with the dimensionality d.

2.7 Conclusions

In this paper we have introduced a new method for the simulation of diffusion bridges which substantially differs from existing methods by using the Zig-Zag sampler and the basis of representation adopted. We motivated both choices and presented the method and its implementation. The resulting simulated bridge measures are shown to be close to the real measures, even for low dimensional approximations and bridges which are highly non-linear. We took advantage of the subsampling technique and a local version of the Zig-Zag to sample high dimensional approximation to conditional measures of diffusions with intractable transition densities. The subsampling technique is a key property in favour of using piecewise deterministic Monte Carlo methods for diffusion bridges (and whenever the target measure is expressed as an integral that requires numerical evaluation). However, the main limitation found for these methods is that they rely on upper bounds of the Poisson rates which are model-specific. Upper bounds for PDMC are easily derived in situations where the log-likelihood has a bounded Hessian. In our setting this means that we wish for the function $b^2(x) - b'(x)$ to have bounded second derivative. In other cases, tailor made bounds need to be derived which can be substantially more complicated. Furthermore, the performance of these samplers can be affected if the upper bounds are too large.

In conclusion, this is the first time (to our knowledge) the Zig-Zag has been employed in a high dimensional practical setting. We claim that the promising results will open research toward applications of the Zig-Zag for high dimensional problems. We mention below some possible extensions of the methodology proposed which are left for future research:

(a) The hierarchical structure of the Faber-Schauder basis suggests that the Zig-Zag should explore the space at different velocities to achieve optimal performance. Unfortunately, it is not immediately clear how to tune the velocity vector;

(b) In Section 2.6 we anticipated the possibility to simulate multidimensional diffusion bridges. In order to generalize the results presented in this paper, we assumed the drift being a conservative vector field. In order to relax this limiting assumption, new convergence results have to be derived which deal explicitly with the stochastic integral appearing in equation (2.8).

(c) The driving motivation for proposing this methodology is to perform parameter estimation of a discretely observed diffusion model. For this purpose, the Zig-Zag sampler runs jointly on the augmented path space given by the coefficients ξ and the parameter space Θ.

Algorithm 4 Implementation of the d-dimensional fully local Zig-Zag sampler

Input: The bounds $\bar{\lambda}_i$ depend only on ξ_k, θ_k, for $k \in \bar{N}_i$ and the random Poisson rates $\widetilde{\lambda}_i$ (eq. (2.14)) depends only on U_i (the randomizing argument of $\widetilde{\partial}_i\psi$) and ξ_k, θ_k for $k \in \widetilde{N}_i(U_i)$

procedure ZIGZAG_FULLY_LOCAL($\tau_{\text{final}}, \xi, \theta$)

Initialise: $k = 1$, $t = \mathbf{0} \in \mathbb{R}^d$, $\tau^* = 0$

$\tau_j \sim \text{IPP}(\bar{\lambda}_j(\cdot; \xi, \theta))$, $j = 1, ..., d$

while $\max(t) \leq \tau_{\text{final}}$ **do**

$\tau_{i^*}^{old} \leftarrow \tau^*, \quad \xi_{i^*}^{old} \leftarrow \xi_{i^*}$

$\tau^*, i^* \leftarrow \text{findmin}(\tau_1, ..., \tau_d)$

$U_{i^*} \sim \text{Law}(U_{i^*})$

for j in $\bar{N}_{i^*} \cup \widetilde{N}_{i^*}(U_{i^*})$ **do**

Update: $\xi_j \leftarrow \xi_j + \theta_j(\tau^* - t_j)$

Update: $t_j \leftarrow \tau^*$

end for

$V \sim \text{Unif}(0, 1)$

if $V \leq \widetilde{\lambda}_{i^*}(0; \xi, \theta; U_{i^*}) / \bar{\lambda}_{i^*}(\tau^* - \tau_{i^*}^{old}; \xi^{old}, \theta)$ **then**

Update: $\theta_{i^*} \leftarrow -\theta_{i^*}$

Update: $k \leftarrow k + 1$

Save: $i^{(k)} \leftarrow i^*, s^{(k)} \leftarrow \tau^*, \xi^{(k)} \leftarrow \xi_{i^*}$

for n in $\left(\bigcup_{j \in \bar{N}_{i^*}} \bar{N}_j\right) \setminus \left(\bar{N}_{i^*} \cup \widetilde{N}_{i^*}(U_{i^*})\right)$ **do**

Update: $\xi_n \leftarrow \xi_n + \theta_n(\tau^* - t_n)$

Update: $t_n \leftarrow \tau^*$

end for

for j in $\bar{N}_{i^*} \setminus \{i^*\}$ **do**

$\tau_j \sim \tau^* + \text{IPP}(\bar{\lambda}_j(\cdot; \xi, \theta))$

$\tau_j^{old} \leftarrow \tau^*, \quad \xi_j^{old} \leftarrow \xi_j$

end for

end if

$\tau_{i^*} \sim \tau^* + \text{IPP}(\bar{\lambda}_{i^*}(\cdot; \xi, \theta))$

end while

return reflection tuples $((i^{(l)}, s^{(l)}, \xi^{(l)}))_{l=1,...,k}$

end procedure

Chapter 3

The Boomerang sampler

3.1 Introduction

Markov chain Monte Carlo remains the gold standard for asymptotically exact (ie bias-free) Bayesian inference for complex problems in Statistics and Machine Learning; see for example Brooks et al. 2011. Yet a major impediment to its routine implementation for large data sets is the need to evaluate the target density (and possibly other related functionals) at each algorithm iteration.

Partly motivated by this, in recent years there has been a surge in the development of innovative piecewise deterministic Monte Carlo methods (PDMC, most notably the Bouncy Particle Sampler (BPS) Bouchard-Côté, Vollmer, and Doucet 2018 and the Zig-Zag Sampler (ZZ) Bierkens, Fearnhead, and Roberts 2019), as a competitor for classical MCMC algorithms such as Metropolis-Hastings and Gibbs sampling. We refer to Fearnhead et al. 2018 for an accessible introduction to the PDMC setting. The primary benefits of these methods are the possibility of *exact subsampling* and *non-reversibility*. Exact subsampling refers to the possibility of using only a subset of the full data set (or even just a single observation) at each iteration of the algorithm, without introducing bias in the output of the algorithm Fearnhead et al. 2018. Non-reversibility is a property of MCMC algorithms related to a notion of direction of the algorithm, reducing the number of backtracking steps, thus reducing the diffusivity of the algorithm and reducing the asymptotic variance; as analyzed e.g. in Diaconis, Holmes, and Neal 2000; Andrieu and Livingstone 2019.

The current key proponents BPS and ZZ of the PDMC paradigm share the following description of their dynamics. The process moves continuously in time according to a constant velocity over random time intervals, which are separated by 'switching events'. These switching events occur at stochastic times at which the velocity, or a component of it, is either reflected, or randomly refreshed. The direction of a reflection, and the random time at which it occurs, is influenced by the target probability distribution.

In this paper we explore the effect of modifying the property of constant velocity. By doing so we introduce the Boomerang Sampler which has dynamics of the simple form $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}$, $\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -\boldsymbol{x}$. Similar ideas were introduced in Vanetti et al. 2017 and termed Hamiltonian-BPS, a method which can be seen as a special case of our approach. We generalise the Hamiltonian-BPS algorithm in three important ways.

1. We relax a condition which restricts the covariance function of the auxiliary velocity process to be isotropic. This generalisation is crucial to ensure good convergence properties of the algorithm.

2. Furthermore we extend the Boomerang Sampler to allow for exact subsampling (as introduced above), thus permitting its application efficiently for large data sets.

3. We also introduce a factorised extension of the sampler which has important computational advantages in the common situation where the statistical model exhibits suitable conditional dependence structure.

Our method also has echoes of the elliptical slice sampler Murray, Adams, and MacKay 2010 which has been a successful discrete-time MCMC method especially within machine learning applications. Both methods are strongly motivated by Hamiltonian dynamics although there are also major differences in the two approaches. Finally we mention some other PDMP methods with non-linear dynamics such as Randomized HMC Bou-Rabee and Sanz-Serna 2017; Deligiannidis, Paulin, and Doucet 2018, and others Markovic and Sepehri 2018; Terenin and Thorngren 2018.

We shall study the Boomerang Sampler and two subsampling alternatives theoretically by analysing the interaction of Bayesian posterior contraction, data size (n) and subsampling schemes in the regular (smooth density) case. We shall show that no matter the rate of posterior contraction, a suitably constructed subsampled Boomerang sampler achieves an $\mathcal{O}(n)$ advantage over non-subsampled algorithms.

At the same time, we show that for the (non-subsampled) Boomerang Sampler, the number of switching events, and thus the computational cost, can be reduced by factor $\mathcal{O}(1/d)$ (where d is the number of dimensions) relative to other piecewise deterministic methods, thanks to the deterministic Hamiltonian dynamics of the Boomerang Sampler.

We illustrate these analyses with empirical investigations in which we compare the properties of Boomerang samplers against other PDMC benchmarks demonstrating the superiority of subsampled Boomerang for sufficiently large data size for any fixed dimension in the setting of logistic regression. We shall also give an empirical study to compare the Boomerang Sampler with its competitors as dimension

increases. Finally, as a potentially very useful application we describe the simulation of diffusion bridges using the Factorised Boomerang Sampler, demonstrating substantial computational advantages compared to its natural alternatives.

Notation

For $\boldsymbol{a} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}$ a positive definite matrix in $\mathbb{R}^{d\times d}$ we write $\mathcal{N}(\boldsymbol{a}, \boldsymbol{\Sigma})$ for the Gaussian distribution in $\mathbb{R}^d$ with mean $\boldsymbol{a}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\langle \cdot, \cdot \rangle$ denote the Euclidean inner product in $\mathbb{R}^d$. We write $(a)_+ := \max(a, 0)$ for the positive part of $a \in \mathbb{R}$, and we write $\langle \cdot, \cdot \rangle_+ := (\langle \cdot, \cdot \rangle)_+$ for the positive part of the inner product.

3.2 The Boomerang Sampler

The Boomerang Sampler is a continuous time, piecewise deterministic Markov process (PDMP) with state space $S = \mathbb{R}^d \times \mathbb{R}^d$. The two copies of $\mathbb{R}^d$ will be referred to as the *position space* and the *velocity space*, respectively. Our primary interest is in sampling the position coordinate, for which the auxiliary velocity coordinate is a useful tool for us.

Let μ_0 denote a Gaussian measure on S specified by $\mu_0 = \mathcal{N}(\boldsymbol{x}_\star, \boldsymbol{\Sigma}) \otimes \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a positive definite matrix in $\mathbb{R}^{d\times d}$. Often we take $\boldsymbol{x}_\star = \boldsymbol{0}$ to shorten expressions, which can be done without loss of generality by a shift in the position coordinate. The measure μ_0 will be referred to as the *reference measure*. The Boomerang Sampler is designed in such a way that it has stationary probability distribution μ with density $\exp(-U(\boldsymbol{x}))$ relative to μ_0. Equivalently, it has density

$$\exp\left(-U(\boldsymbol{x}) - \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_\star)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{x}_\star) - \tfrac{1}{2}\boldsymbol{v}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{v}\right)$$

relative to the Lebesgue measure $\mathrm{d}\boldsymbol{x} \otimes \mathrm{d}\boldsymbol{v}$ on $\mathbb{R}^d \times \mathbb{R}^d$. We assume that this density has a finite integral. The marginal distribution of μ with respect to $\boldsymbol{x}$ is denoted by Π.

The Boomerang process moves along deterministic trajectories $(\boldsymbol{x}_t, \boldsymbol{v}_t) \in \mathbb{R}^d \times \mathbb{R}^d$ which change direction at random times. The deterministic trajectories satisfy the following simple ordinary differential equation:

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{v}_t, \quad \frac{\mathrm{d}\boldsymbol{v}_t}{\mathrm{d}t} = -(\boldsymbol{x}_t - \boldsymbol{x}_\star), \tag{3.1}$$

with explicit solution $\boldsymbol{x}_t = \boldsymbol{x}_\star + (\boldsymbol{x}_0 - \boldsymbol{x}_\star)\cos(t) + \boldsymbol{v}_0 \sin(t)$, $\boldsymbol{v}_t = -(\boldsymbol{x}_0 - \boldsymbol{x}_\star)\sin(t) + \boldsymbol{v}_0 \cos(t)$. Note that $(\boldsymbol{x}, \boldsymbol{v}) \mapsto \langle \boldsymbol{x} - \boldsymbol{x}_\star, \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{x}_\star)\rangle + \langle \boldsymbol{v}, \boldsymbol{Q}\boldsymbol{v}\rangle$ is invariant with respect to the flow of (3.1) for any symmetric matrix $\boldsymbol{Q}$. In particular the flow of (3.1) preserves the Gaussian measure μ_0 on S.

Given an initial position $(\boldsymbol{x}_0, \boldsymbol{v}_0) \in S$, the process moves according to the motion specified by (3.1), resulting in a trajectory $(\boldsymbol{x}_t, \boldsymbol{v}_t)_{t\geq 0}$, until the first event occurs. The distribution of the first reflection event time T is specified by

$$\mathbb{P}(T \geq t) = \exp\left(-\int_0^t \lambda(\boldsymbol{x}_s, \boldsymbol{v}_s)\,\mathrm{d}s\right),$$

where $\lambda : S \to [0, \infty)$ is the *event rate* and is specified as

$$\lambda(\boldsymbol{x}, \boldsymbol{v}) = \langle \boldsymbol{v}, \nabla U(\boldsymbol{x})\rangle_+. \tag{3.2}$$

For $\boldsymbol{x} \in \mathbb{R}^d$ we define the *contour reflection* $\boldsymbol{R}(\boldsymbol{x})$ to be the linear operator from $\mathbb{R}^d$ to $\mathbb{R}^d$ given, for $(\boldsymbol{x}, \boldsymbol{v}) \in S$, by

$$\boldsymbol{R}(\boldsymbol{x})\boldsymbol{v} = \boldsymbol{v} - \frac{2\langle \nabla U(\boldsymbol{x}), \boldsymbol{v}\rangle}{|\boldsymbol{\Sigma}^{1/2}\nabla U(\boldsymbol{x})|^2}\boldsymbol{\Sigma}\nabla U(\boldsymbol{x}). \tag{3.3}$$

Importantly the reflection satisfies

$$\langle \boldsymbol{R}(\boldsymbol{x})\boldsymbol{v}, \nabla U(\boldsymbol{x})\rangle = -\langle \boldsymbol{v}, \nabla U(\boldsymbol{x})\rangle \tag{3.4}$$

and

$$|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}(\boldsymbol{x})\boldsymbol{v}| = |\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}|, \tag{3.5}$$

which are key in establishing that the resulting Boomerang Sampler has the correct stationary distribution.

At the random time T at which a switch occurs, we put $\boldsymbol{v}_T := \boldsymbol{R}(\boldsymbol{x}_{T-})\boldsymbol{v}_{T-}$, where we use the notation $\boldsymbol{y}_{t-} := \lim_{s\uparrow t} \boldsymbol{y}_s$. The process then starts afresh according to the dynamics (3.1) from the new position $(\boldsymbol{x}_T, \boldsymbol{v}_T)$. Additionally, at random times generated by a homogeneous Poisson process with rate $\lambda_{\text{refr}} > 0$ the velocity is refreshed, i.e. at such a random time T we independently draw $\boldsymbol{v}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. This additional input of randomness guarantees that the Boomerang Sampler can visit the full state space and is therefore ergodic, as is the case for e.g. BPS Bouchard-Côté, Vollmer, and Doucet 2018.

In Section 1 of the Supplement we define the generator of the Boomerang Sampler, which can in particular be used to prove that μ is a stationary distribution for the Boomerang process, and which can be used in subsequent studies to understand its probabilistic properties.

Remark 3.2.1 (On the choice of the reference measure)**.** *In principle we can express* any *probability distribution* $\Pi(\mathrm{d}\boldsymbol{x}) \propto \exp(-E(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}$ *as a density relative to* μ_0 *by defining*

$$U(\boldsymbol{x}) = E(\boldsymbol{x}) - \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_\star)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{x}_\star). \tag{3.6}$$

As mentioned before we can take μ_0 to be identical to a Gaussian prior measure in the Bayesian setting. Alternatively, and this is an approach which we will adopt in this paper, we may choose μ_0 to be a Gaussian approximation of the measure Π which may be obtained at relatively small computational cost in a preconditioning step.

3.2.1 Factorised Boomerang Sampler

As a variation to the Boomerang Sampler introduced above we introduce the Factorised Boomerang Sampler (FBS), which is designed to perform well in situations where the conditional dependencies in the target distribution are sparse. For simplicity we restrict to the case with a diagonal reference covariance $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2)$.

The deterministic dynamics of the FBS are identical to those of the standard Boomerang Sampler, and given by (3.1). The difference is that every component of the velocity has its own switching intensity. This is fully analogous with the difference between BPS and ZZ, where the latter can be seen as a factorised Bouncy Particle Sampler. In the current case, this means that as switching intensity for the i-th component of the velocity we take

$$\lambda_i(\boldsymbol{x}, \boldsymbol{v}) = (v_i \partial_i U(\boldsymbol{x}))_+,$$

and once an event occurs, the velocity changes according to the operator $\boldsymbol{F}_i(\boldsymbol{v})$ given by

$$\boldsymbol{F}_i(\boldsymbol{v}) = \Big(v_1, \dots, v_{i-1}, -v_i, v_{i+1}, \dots, v_d\Big)^\top.$$

Also, the velocity of each component is refreshed according to $v_i \sim \mathcal{N}(0, \sigma_i^2)$ at rate $\lambda_{\mathrm{refr},i} > 0$.

Note that the computation of the reflections has a computational cost of $\mathcal{O}(1)$, compared to the reflections in (3.3) being at least of $\mathcal{O}(d)$, depending upon the sparsity of $\boldsymbol{\Sigma}$. The sparse conditional dependence structure implies that the individual switching intensities $\lambda_i(\boldsymbol{x})$ are in fact functions of a subset of the components of $\boldsymbol{x}$, contributing to a fast computation. This feature can be exploited by an efficient 'local' implementation of the FBS algorithm which reduces the number of Poisson times simulated by the algorithm (similar in spirit to the local Bouncy Particle Sampler Bouchard-Côté, Vollmer, and Doucet 2018 and the local Zig-Zag sampler in Bierkens et al. 2021). In Section 3.3.2 we will briefly comment on the dimensional scaling of FBS. As an illustration of a realistic use, FBS will be applied to the simulation of diffusion bridges in Section 3.4.2.

3.2.2 Subsampling with control variates

Let $E(\boldsymbol{x})$ be the energy function, i.e., negative log density of Π with respect to the Lebesgue measure. Consider the setting where $E(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} E^i(\boldsymbol{x})$, as is often the case in e.g. Bayesian statistics or computational physics. (Let us stress that n represents a quantity such as the number of interactions or the size of the data, and *not* the dimensionality of $\boldsymbol{x}$, which is instead denoted by d.) Using this structure, we introduce a subsampling method using the Gaussian reference measure as a tool for the efficient construction of the Monte Carlo method.

Relative to a Gaussian reference measure with covariance $\boldsymbol{\Sigma}$ centred at $\boldsymbol{x}_\star$, the negative log density is given by (3.6). Let us assume

$$\boldsymbol{\Sigma} = [\nabla^2 E(\boldsymbol{x}_\star)]^{-1} \tag{3.7}$$

for a reference point $\boldsymbol{x}_\star$. In words, the curvature of the reference measure will agree around $\boldsymbol{x}_\star$ with the curvature of the target distribution. We can think of $\boldsymbol{x}_\star$ as the mean or mode of an appropriate Gaussian approximation used for the Boomerang Sampler. Note however that we shall not require that $\nabla E(\boldsymbol{x}_\star) = \boldsymbol{0}$ for the sampler and its subsampling alternatives to work well, although some restrictions will be imposed in Section 3.3.1. In this setting it is possible to employ a subsampling method which is exact, in the sense that it targets the correct stationary distribution. This is an extension of methodology used for subsampling in other piecewise deterministic methods, see e.g. Fearnhead et al. 2018 for an overview.

Assume for notational convenience that $\boldsymbol{x}_\star = \boldsymbol{0}$. As an unbiased estimator for the log density gradient of U we could simply take

$$\widetilde{\nabla U(\boldsymbol{x})} = \nabla E^I(\boldsymbol{x}) - \nabla^2 E(\boldsymbol{0})\boldsymbol{x}, \tag{3.8}$$

where I is a random variable with uniform distribution over $\{1, \dots, n\}$. We shall see in Proposition 3.3.1 that this estimator will lead to weights which increase with n and therefore we shall consider a control variate alternative.

Therefore also consider the control variate gradient estimator $\widehat{\nabla U(\boldsymbol{x})} = \boldsymbol{G}^I(\boldsymbol{x})$, where, for $i = 1, \dots, n$,

$$\boldsymbol{G}^i(\boldsymbol{x}) = \nabla E^i(\boldsymbol{x}) - \nabla^2 E^i(\boldsymbol{0})\boldsymbol{x} - \nabla E^i(\boldsymbol{0}) + \nabla E(\boldsymbol{0}). \tag{3.9}$$

Taking the expectation with respect to I,

$$\begin{aligned}
&\mathbb{E}_I \widehat{\nabla U(\boldsymbol{x})} \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\{\nabla E^i(\boldsymbol{x}) - \nabla^2 E^i(\boldsymbol{0})\boldsymbol{x} - \nabla E^i(\boldsymbol{0}) + \nabla E(\boldsymbol{0})\right\} \\
&= \nabla E(\boldsymbol{x}) - \nabla^2 E(\boldsymbol{0})\boldsymbol{x} = \nabla U(\boldsymbol{x}),
\end{aligned}$$

so that $\widehat{\nabla U(\boldsymbol{x})}$ is indeed an unbiased estimator for $\nabla U(\boldsymbol{x})$. In Section 3.3 we shall show that $\widehat{\nabla U(\boldsymbol{x})}$ has significantly superior scaling properties for large n than $\widetilde{\nabla U(\boldsymbol{x})}$.

Remark 3.2.2. *In various situations we can find a reference point $\boldsymbol{x}_\star$ such that $\nabla E(\boldsymbol{x}_\star) = \mathbf{0}$, in which case the final term in (3.9) vanishes. We include the term here so that it can accommodate the general situation in which $\nabla E(\boldsymbol{x}_\star) \neq \mathbf{0}$.*

Upon reflection, conditional on the random draw I, we reflect according to

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

The Boomerang Sampler that switches at the random rate $\widehat{\lambda(\boldsymbol{x},\boldsymbol{v})} = \langle \boldsymbol{v}, \widehat{\nabla U(\boldsymbol{x})}\rangle_+$, and reflects according to $\boldsymbol{R}^I$ will preserve the desired target distribution in analogy to the argument found in Bierkens, Fearnhead, and Roberts 2019.

3.2.3 Simulation

The implementation of the Boomerang Sampler depends crucially on the ability to simulate from a nonhomogeneous Poisson process with a prescribed rate. In this section we will make a few general comments on how to achieve these tasks for the Boomerang Sampler and for the Subsampled Boomerang Sampler.

Suppose we wish to generate the first event according to a switching intensity $\lambda(\boldsymbol{x}_t, \boldsymbol{v}_t)$ where $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ satisfy (3.1). This is challenging because it is non-trivial to generate points according to time inhomogeneous Poisson process, but also the function $\lambda(\boldsymbol{x}_t, \boldsymbol{v}_t)$ may be expensive to evaluate. It is customary in simulation of PDMPs to employ the technique of *Poisson thinning* to generate an event according to a deterministic rate function $\overline{\lambda}(t) \geq 0$, referred to as *computational bound*, such that $\lambda(\boldsymbol{x}_t, \boldsymbol{v}_t) \leq \overline{\lambda}(t)$ for all $t \geq 0$. The function $\overline{\lambda}(t)$ should be suitable in the sense that we can explicitly simulate T according to the law

$$\mathbb{P}(T \geq t) = \exp\left(-\int_0^t \overline{\lambda}(s)\, ds\right).$$

After generating T from this distribution, we accept T as a true switching event with probability $\lambda(\boldsymbol{x}_T, \boldsymbol{v}_T)/\overline{\lambda}(T)$. As a consequence of this procedure, the first time T that gets accepted in this way is a Poisson event with associated intensity $\lambda(\boldsymbol{x}_t, \boldsymbol{v}_t)$.

In this paper we will only consider bounds $\overline{\lambda}(t)$ of the form $\overline{\lambda}(t; \boldsymbol{x}_0, \boldsymbol{v}_0) = a(\boldsymbol{x}_0, \boldsymbol{v}_0) + tb(\boldsymbol{x}_0, \boldsymbol{v}_0)$. We will call the bound *constant* if $b(\boldsymbol{x}, \boldsymbol{v}) = 0$ for all $(\boldsymbol{x}, \boldsymbol{v})$, and *affine* otherwise. As a simple example, consider the situation in which $|\nabla U(\boldsymbol{x})| \leq m$ for all $\boldsymbol{x}$. In this case we have

$$\lambda(\boldsymbol{x}, \boldsymbol{v}) = \langle \boldsymbol{v}, \nabla U(\boldsymbol{x})\rangle_+ \leq m|\boldsymbol{v}| \leq m\sqrt{|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2}.$$

Since the final expression is invariant under the dynamics (3.1), we find that

$$\lambda(\boldsymbol{x}_t, \boldsymbol{v}_t) \leq m\sqrt{|\boldsymbol{x}_0|^2 + |\boldsymbol{v}_0|^2}, \quad t \geq 0,$$

which gives us a simple constant bound.

In the case of subsampling the switching intensity $\widehat{\lambda(\boldsymbol{x}, \boldsymbol{v})}$ is random. Still, the bound $\overline{\lambda}(t; \boldsymbol{x}_0, \boldsymbol{v}_0)$ needs to be an upper bound for all random realizations of $\widehat{\lambda(\boldsymbol{x}, \boldsymbol{v})}$. In the case we use the unbiased gradient estimator $\widehat{\nabla U(\boldsymbol{x})} = \boldsymbol{G}^I$ of (3.9), we can bound e.g.

$$\widehat{\lambda(\boldsymbol{x}, \boldsymbol{v})} \leq \sup_{i,\boldsymbol{x}} |\boldsymbol{G}^i(\boldsymbol{x})||\boldsymbol{v}| \leq \sup_{i,\boldsymbol{x}} |\boldsymbol{G}^i(\boldsymbol{x})|\sqrt{|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2},$$

assuming all gradient estimators $\boldsymbol{G}^i$ are globally bounded. We will introduce different bounds in detail in Section 2 of the Supplement.

3.3 Scaling for large data sets and large dimension

3.3.1 Robustness to large n

In this section, we shall investigate the variability of the rates induced by the Boomerang Sampler and its subsampling options. The size of these rates is related to the size of the upper bounding rate Poisson process used to simulate them. Moreover, the rate of the upper bounding Poisson rate is proportional to the number of density evaluations, which in turn is a sensible surrogate for the computing cost of running the algorithm.

As in Section 3.2.2, we describe E as a sum of n constituent negative log-likelihood terms: $E(\boldsymbol{x}) = -\sum_{i=1}^n \ell_i(\boldsymbol{x})$. (In the notation above we are just setting $\ell_i(\boldsymbol{x}) = -nE_i(\boldsymbol{x})$.) Under suitable regularity conditions, the target probability measure Π satisfies posterior contraction around $\boldsymbol{x} = \boldsymbol{0}$ at the rate η, that is for all ϵ there exists $\delta > 0$ such that $\Pi(B_{n^{-\eta}\delta}(\boldsymbol{0})) > 1 - \epsilon$ where $B_r(\boldsymbol{0})$ denotes the ball of radius r centred at $\boldsymbol{0}$. As a result of this, we typically have velocities of order $n^{-\eta}$ ensuring that the dynamics in (3.1) circles the state space in $\mathcal{O}(1)$ time.

The various algorithms will have computational times roughly proportional to the number of likelihood evaluations, which in turn depends on the event rate (and its upper bound). Therefore we shall introduce explicitly the subsampling bounce rates corresponding to the use of the unbiased estimators in (3.8) and (3.9).

$$\widetilde{\lambda(\boldsymbol{x}, \boldsymbol{v})} = \langle \boldsymbol{v}, \widetilde{\nabla U(\boldsymbol{x})} \rangle_+ \; ; \qquad \widehat{\lambda(\boldsymbol{x}, \boldsymbol{v})} = \langle \boldsymbol{v}, \widehat{\nabla U(\boldsymbol{x})} \rangle_+ \; .$$

To simplify the arguments below, we also assume that ℓ_i has all its third derivatives uniformly bounded, implying that all third derivative terms of E are bounded by a

constant multiple of n. This allows us to write down the expansion

$$\begin{aligned}\nabla U(\boldsymbol{x}) &= \nabla E(\mathbf{0}) + \nabla^2 E(\mathbf{0})\boldsymbol{x} - \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \mathcal{O}(n|\boldsymbol{x}|^2) \\ &= \nabla E(\boldsymbol{x}) \\ &= \nabla E(\mathbf{0}) + \mathcal{O}(n|\boldsymbol{x}|^2)\ . \end{aligned} \tag{3.10}$$

Similarly we can write

$$\begin{aligned}\widehat{\nabla U(\boldsymbol{x})} &= n\nabla \ell_I(\boldsymbol{x}) - n\nabla^2 \ell_I(\mathbf{0})\boldsymbol{x} - n\ell_I(\mathbf{0}) \\ &\quad + \nabla E(\mathbf{0}) - \boldsymbol{\Sigma}^{-1}\boldsymbol{x} \\ &= \nabla E(\mathbf{0}) + \mathcal{O}(n|\boldsymbol{x}|^2)\ . \end{aligned} \tag{3.11}$$

using the same Taylor series expansion.

We can now use this estimate directly to obtain bounds on the event rates. We summarise this discussion in the following result.

Proposition 3.3.1. *Suppose that $\boldsymbol{x},\ \boldsymbol{v} \in B_{n^{-\eta}\delta}(\mathbf{0})$ for some δ, and under the assumptions described above, we have that*

$$\lambda(\boldsymbol{x}, \boldsymbol{v}) \leq \mathcal{O}\left(n^{-\eta}(|\nabla E(\mathbf{0})| + n^{1-2\eta})\right) \tag{3.12}$$

$$\widetilde{\lambda(\boldsymbol{x}, \boldsymbol{v})} \leq \mathcal{O}\left(|\nabla E(\mathbf{0})|\right) + \mathcal{O}(n)) \tag{3.13}$$

$$\widehat{\lambda(\boldsymbol{x}, \boldsymbol{v})} \leq \mathcal{O}\left(n^{-\eta}(|\nabla E(\mathbf{0})| + n^{1-2\eta})\right) \tag{3.14}$$

Thus the use of $\widehat{\nabla U(\boldsymbol{x})}$ does not result in an increased event rate (in order of magnitude). There is therefore an $\mathcal{O}(n)$ computational advantage obtained from using subsampling due to each target density valuation being $\mathcal{O}(n)$ quicker.

Proposition 3.3.1 shows that as long as the reference point $\boldsymbol{x}^*$ (chosen to be $\mathbf{0}$ here for convenience) is chosen to be sufficiently close to the mode so that $|\nabla E(\mathbf{0})|$ is at most $\mathcal{O}(n^{1-2\eta})$, then we have that

$$\lambda(\boldsymbol{x}, \boldsymbol{v}) = \widehat{\lambda(\boldsymbol{x}, \boldsymbol{v})} = \mathcal{O}\left(n^{1-3\eta}\right)\ .$$

Note that this rate can go to 0 when $\eta > 1/3$. In particular in the regular case where Bernstein von-Mises theorem holds, we have $\eta = 1/2$. In this case the rate of jumps for the Boomerang can recede to 0 at rate $n^{-1/2}$ so long as $|\nabla E(\mathbf{0})|$ is at most $\mathcal{O}(1)$).

3.3.2 Scaling with dimension

In this section, we will discuss how the Boomerang Sampler has an attractive scaling property for high dimension. This property is qualitatively similar to the preconditioned Crank-Nicolson algorithm Diaconis, Holmes, and Neal 2000; Beskos et al.

2008 and the elliptical slice sampler Murray, Adams, and MacKay 2010 which take advantage of the reference Gaussian distribution.

The dimensional complexity of BPS and ZZ was studied in Bierkens, Kamatani, and Roberts 2018; Deligiannidis, Paulin, and Doucet 2018; Andrieu et al. 2018. For the case of an isotropic target distribution, the rate of reflections per unit of time is constant for BPS and proportional to d for ZZ with unit speeds in all directions. On the other hand, the time until convergence is of order d for the BPS and 1 for ZZ. Therefore, the total number of reflections required for convergence of these two algorithms is of the same order which grows linearly with dimension.

For the Boomerang Sampler we consider the following setting. Consider reference measures $\mu_{0,d}(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{v}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_d) \otimes \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_d)$ for increasing dimension d, where for every $d = 1, 2, \ldots$, $\boldsymbol{\Sigma}_d$ is a d-dimensional positive definite matrix. Relative to these reference measures we consider a sequence of potential functions $U_d(\boldsymbol{x})$. Thus relative to Lebesgue measure our target distribution $\Pi_d(\mathrm{d}\boldsymbol{x})$ has density $\exp(-E_d(\boldsymbol{x}))$, where $E_d(\boldsymbol{x}) = U_d(\boldsymbol{x}) + \frac{1}{2}\langle \boldsymbol{x}, \boldsymbol{\Sigma}_d^{-1}\boldsymbol{x}\rangle$. Let $\mathbb{E}_d$ denote expectation with respect to $\Pi_d(\mathrm{d}\boldsymbol{x}) \otimes \mathcal{N}(0, \boldsymbol{\Sigma}_d)(\mathrm{d}\boldsymbol{v})$. We assume that the sequence (U_d) satisfies

$$\sup_{d=1,2,\ldots} \mathbb{E}_d[|\boldsymbol{\Sigma}_d^{1/2}\nabla U_d(\boldsymbol{x})|^2] < \infty, \tag{3.15}$$

The condition (3.15) arises naturally for instance in the context of Gaussian regression, spatial statistics, Bayesian inverse problems as well as the setting of the diffusion bridge simulation example described in detail in Section 3.4.2.

Furthermore we assume that the following form of the Poincaré inequality holds,

$$\mathbb{E}_d[f_d(\boldsymbol{x})^2] \leq \frac{1}{C^2}\mathbb{E}_d[|\boldsymbol{\Sigma}_d^{1/2}\nabla f_d(\boldsymbol{x})|^2] \tag{3.16}$$

with constant $C > 0$ independent of dimension, and where $f_d : \mathbb{R}^d \to \mathbb{R}$ is any mean zero differentiable function. A sufficient condition for (3.16) to hold is

$$C^2 I \preceq \boldsymbol{\Sigma}_d^{1/2}\nabla^2 E_d(\boldsymbol{x})\boldsymbol{\Sigma}_d^{1/2} = \boldsymbol{\Sigma}_d^{1/2}\nabla^2 U_d(\boldsymbol{x})\boldsymbol{\Sigma}_d^{1/2} + I$$

by the classical Brascamp-Lieb inequality Brascamp and Lieb 1976; Bakry, Gentil, and Ledoux 2014; note that it may also hold in the non-convex case, see e.g. Lorenzi and Bertoldi 2007, Section 8.6.

Under the stated assumptions we argue that (i) the expected number of reflections per unit time scales as $\mathcal{O}(1)$ with respect to dimension, and (ii) within a continuous time interval that scales as $\mathcal{O}(1)$, the Boomerang Sampler mixes well. Claims (i) and (ii) are provided with a heuristic motivation in the Supplement. A rigorous proof for this claim remains part of our future work.

In the ideal but non-sparse scenario, the computational cost of the event time calculation for the Boomerang Sampler is thus expected to be a factor d smaller

compared to BPS and ZZ assuming that the cost per event is the same for these algorithms. However, this comparison is unrealistic since in general we can not simulate reflections directly. In practice, we need to use the thinning method as discussed in Section 3.2.3. The thinning method introduces a significant amount of shadow events (which are rejected after inspection), and the true events usually represent a small portion relative to the number of shadow events. As a result there can be a high cost for calculating shadow events even when the number of true events is small.

For the FBS, the expected number of events per unit of time is $\sum_{i=1}^{d} \mathbb{E}_d[(v_i \partial_i U(\boldsymbol{x}))_+]$. Under the hypothesis above, this is of $\mathcal{O}(d^{1/2})$. Thus, the number of events is much bigger than that of the Boomerang. However, as in the case of ZZ, under a sparse model assumption, the cost of calculation per jump is of constant order whereas it is of the order of d for the Boomerang Sampler. Therefore, the Factorised Boomerang Sampler should outperform the Boomerang Sampler for this sparse setup.

3.4 Applications and experiments

3.4.1 Logistic regression

As a suitable test bed we consider the logistic regression inference problem. Given predictors $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)}$ in $\mathbb{R}^d$, and outcomes $z^{(1)}, \ldots, z^{(n)}$ in $\{0, 1\}$, we define the log likelihood function as

$$\ell(\boldsymbol{x}) = -\sum_{i=1}^{n} \left\{ \log(1 + e^{\boldsymbol{x}^\top \boldsymbol{y}^{(i)}}) - z^{(i)} \boldsymbol{x}^\top \boldsymbol{y}^{(i)} \right\}.$$

Furthermore we impose a Gaussian prior distribution over $\boldsymbol{x}$ which for simplicity we keep fixed to be a standard normal distribution throughout these experiments. As a result we arrive at the negative log target density

$$E(\boldsymbol{x}) = \sum_{i=1}^{n} \left\{ \log(1 + e^{\boldsymbol{x}^\top \boldsymbol{y}^{(i)}}) - z^{(i)} \boldsymbol{x}^\top \boldsymbol{y}^{(i)} \right\} + \tfrac{1}{2} \boldsymbol{x}^\top \boldsymbol{x}.$$

As a preprocessing step when applying the Boomerang Sampler, and all subsampled methods, we find the mode $\boldsymbol{x}_\star$ of the posterior distribution and define $\boldsymbol{\Sigma}$ by (3.7). We apply the Boomerang Sampler, with and without subsampling. These samplers are equipped with an affine computational bound and a constant computational bound respectively, both discussed in Section 2 of the Supplement (the affine bound is usually preferred over a constant bound, but a useful affine bound is not available in the subsampling case).

We compare the Boomerang to both BPS and ZZ with and without subsampling. In all subsampling applications we employ appropriate control variance techniques

to reduce the variability of the random switching intensities, as discussed in Section 3.2.2. Furthermore in the dimension dependent study we include the Metropolis adjusted Langevin algorithm (MALA) for comparison. Throughout these experiments we use Effective Sample Size (ESS) per second of CPU time as measure of the efficiency of the methods used. ESS is estimated using the Batch Means method, where we take a fixed number of 50 batches for all our estimates. ESS is averaged over the dimensions of the simulation and then divided by the runtime of the algorithm to obtain "average ESS per second" (other ESS summaries could also have been used). The time horizon is throughout fixed at $10,000$ (with 10,000 iterations for MALA). For ZZ and BPS the magnitude of the velocities is rescaled to be comparable on average with Boomerang, to avoid unbalanced runtimes of the different algorithms. In Figures 3.1 and 3.2 the boxplots are taken over 20 randomly generated experiments, where each experiment corresponds to a logistic regression problem with a random (standard normal) parameter, based on randomly generated data from the model.[1] The refreshment rates for BPS and the Boomerang Samplers are taken to be 0.1.

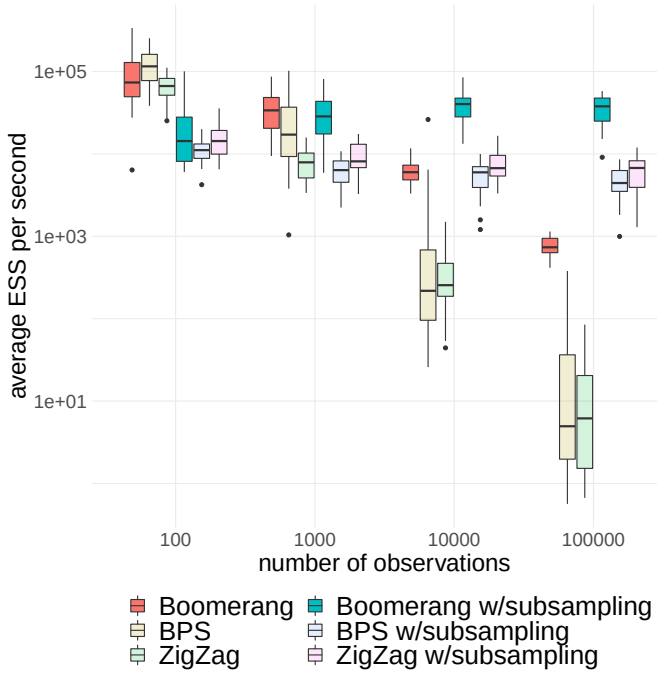


Figure 3.1: Scaling of Boomerang Sampler compared to other PDMC methods for the logistic regression problem of Section 3.4.1 as a function of the number of observations. Here $d = 2$.

[1]The code used to carry out all of the experiments of this paper may be found online at `https://github.com/jbierkens/ICML-boomerang`.

The Boomerang Sampler is seen to outperform the other algorithms, both in terms of scaling with dimension as with respect to an increase in the number of observations. For a fixed dimension, the subsampling algorithms will clearly outperform the non-subsampling algorithms as number of observations n grows. In particular, the ESS/sec stays fixed for the subsampled algorithms, and decreases as $\mathcal{O}(n)$ for the non-subsampled versions. In this case, we did not include the MALA algorithm since we observed its complexity strongly deteriorating as the number of observations increases. For a large number of observations ($n \geq 10,000$, $d = 2$) we see that the Boomerang Sampler (with and without subsampling) accepts almost none of the proposed switches. This means that effectively we are sampling from the Gaussian reference measure. This observed behaviour is in line with the scaling analysis in Section 3.3.1.

In the second experiment we let the dimension d grow for a fixed number of observations. The subsampling algorithms currently do not scale as well as the non-subsampled versions. For practical purposes we therefore only consider non-subsampled algorithms for the comparison with respect to dimensional dependence. For the dimensions $d \leq 32$ we tested the Boomerang outperforms MALA, but it seems empirically that MALA has a better scaling behaviour with dimension. Note that MALA needs careful tuning to exhibit this good scaling. We remark that the beneficial scaling properties of the underlying Boomerang Process as discussed in Section 3.3.2 may be adversely affected by suboptimal computational bounds. We are optimistic that the dimensional scaling of subsampled algorithms can be further improved by designing better computational bounds.

In all cases the necessary preprocessing steps can be done very quickly. In particular the plots are not affected by including (or excluding) the preprocessing time in the computation of ESS/sec.

3.4.2 Diffusion bridges

In Bierkens et al. 2021 the authors introduce a framework for the simulation of diffusion bridges (diffusion processes conditioned to hit a prescribed endpoint) taking strong advantage of the use of factorised piecewise deterministic samplers. This invites the use of the Factorised Boomerang Sampler (FBS). We consider time-homogeneous one-dimensional conditional diffusion processes (diffusion bridges) of the form

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \mathrm{d}W_t, \quad X_0 = u, \quad X_T = v$$

where W is a scalar Brownian motion and b satisfies some mild regularity conditions (see Bierkens et al. 2021 for details). This simulation problem is an essential building block in Bayesian analysis of non-linear diffusion models with low frequency observations Roberts and Stramer 2001.

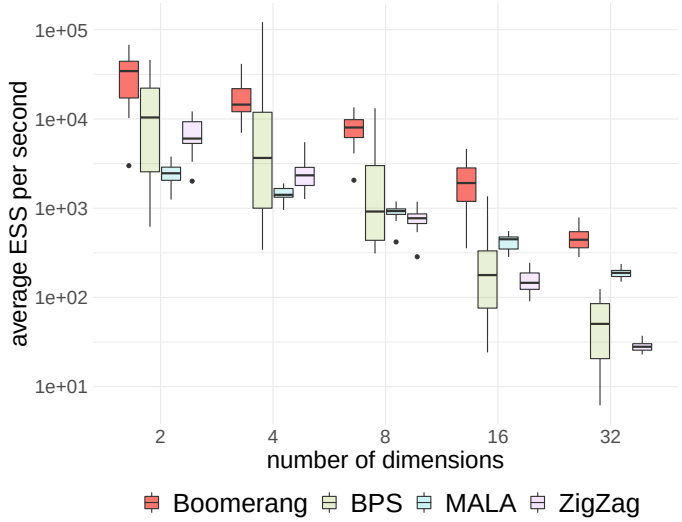


Figure 3.2: Scaling of Boomerang Sampler compared to other PDMC methods and MALA for the logistic regression problem of Section 3.4.1 as a function of the number of dimensions. Here the number of observations is $n = 1,000$.

We consider the approach of Bierkens et al. 2021 where the diffusion path on $[0, T]$ is expanded with a truncated Faber Schauder basis as

$$X_t^N = \bar{\bar{\phi}}(t)u + \bar{\phi}(t)v + \sum_{i=0}^{N}\sum_{j=0}^{2^i-1}\phi_{i,j}(t)x_{i,j}.$$

Here,

$$\begin{aligned}
\bar{\phi}(t) &= t/T, \qquad \bar{\bar{\phi}}(t) = 1 - t/T, \\
\phi_{0,0}(t) &= \sqrt{T}\big((t/T)\mathbf{1}_{[0,T/2]}(t) + (1 - t/T)\mathbf{1}_{(T/2,T]}(t)\big), \\
\phi_{i,j}(t) &= 2^{-i/2}\phi_{0,0}(2^i t - jT) \quad i \geq 0, \quad 0 \leq j \leq 2^i - 1,
\end{aligned}$$

are the Faber-Schauder functions and N is the truncation of the expansion. In Bierkens et al. 2021, ZZ is used to sample the corresponding coefficients $\boldsymbol{x} := (x_{0,0}, ..., x_{N,2^N-1})$ which have a density measure written with respect to a standard Gaussian reference measure (corresponding to a Brownian bridge measure in the path space, see Bierkens et al. 2021 for details). In particular we have that

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(\boldsymbol{x}, \boldsymbol{v}) \propto \exp\left\{-\frac{1}{2}\int_0^T \left(b^2(X_s^N) + b'(X_s^N)\right)\mathrm{d}s\right\} \tag{3.17}$$

where b' is the derivative of b and $\mu_0 = \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \otimes \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ with $\boldsymbol{I}$ the $2^{N+1}-1$ dimensional identity matrix. The measure given by (3.17) has a remarkable conditional

independence property (Proposition 2, Bierkens et al. 2021) and the coefficients $x_{i,j}$, for i large, responsible for the local behaviour of the process, are approximately independent standard Gaussian, reflecting the fact that, locally, the process behaves as a Brownian motion.

In Bierkens et al. 2021 the authors device a local implementation of ZZ which optimally exploits the sparse conditional independence structure of the target distribution, alleviating the computational costs in high dimensional setting (e.g. of a high truncation level N). Since the Girsanov density (3.17) is expressed relative to a standard normal distribution on the coefficients $\boldsymbol{x}$, the Factorised Boomerang Sampler is a natural candidate for a further reduction in computational cost, by reducing the required number of simulated events, in particular at the higher levels where the coefficients have approximately a Gaussian distribution. This will allow for a further increase of the truncation level N and/or faster computations at a fixed truncation levels.

We consider the the class of diffusion bridges with drift equal to

$$b(x) = \alpha \sin(x),\ \alpha \geq 0. \tag{3.18}$$

The higher α, the stronger is the attraction of the diffusion paths to the stable points $(2k-1)\pi, k \in \mathbb{N}$ while for $\alpha = 0$ the process reduces to a Brownian bridge with $\mu = \mu_0$. Equivalently to Bierkens et al. 2021, we use subsampling as the gradient of the log density in (3.17) involves a time integral that cannot be solved analytically in most of the cases. The unbiased estimator for $\partial_{x_{i,j}} U(\boldsymbol{x})$ is the integrand evaluated at a uniform random point multiplied by the range of the integral. The Poisson bounding rates used for the subsampling can be found in the Supplement, Section 5.

Figure 3.3 shows the resulting bridges for $\alpha = 1$, starting at $u = -\pi$ and hitting $v = 3\pi$ at final time $T = 50$ after running the FBS with clock $T^\star = 20000$, as simulated on a standard desktop computer. The refreshment rate relative to each coefficient $x_{i,j}$ is fixed to $\lambda_{\text{refr},i,j} = 0.01$ and the truncation of the expansion is $N = 6$.

In Figure 3.4, we compare the performances of the Boomerang Sampler and ZZ by computing the average number of reflections (y-axis on a log-scale) for the coefficients $x_{i,j}$ at each level (x-axis). The number of reflections is understood as a measure of complexity of the algorithm. We repeat the experiment for $\alpha = 0.5$ and $\alpha = 0$ (where $\mu = \mu_0$) and fix the truncation level to be $N = 10$ which corresponds to a $2047 + 2047$ dimensional space for $(\boldsymbol{x}, \boldsymbol{v})$. For a fair comparison we set the expected ℓ^1 norm of the velocities and the time horizon of the two samplers to be the same. In both cases, the average number of reflections converges to the average number of reflections under μ_0 (dashed lines) indicating that the coefficients at high levels are approximately standard normally distributed but while ZZ requires a fixed number of reflections for sampling from $\mu = \mu_0$, the Boomerang does not, allowing to high resolutions of the diffusion bridges at lower computational cost.

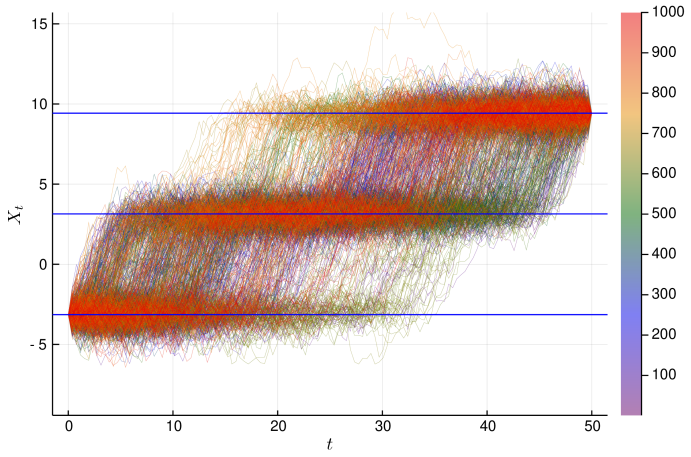


Figure 3.3: 1000 diffusion bridges with drift equal to (3.18) with $\alpha = 1$, $u = -\pi$, $v = 3\pi$, $T = 50, L = 6$ sampled with the FBS with time horizon $T^\star = 20,000$ and refreshment rates $\lambda_{\text{refr},i} = 0.01$ for all i. The straight horizontal lines are the attraction points.

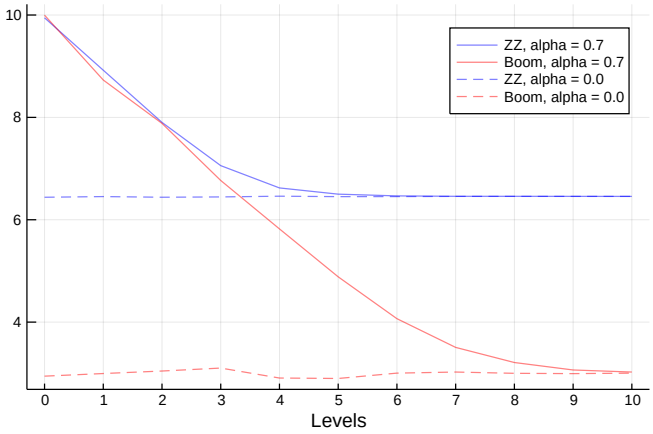


Figure 3.4: Average number of reflections (on a log-scale) for the coefficients $x_{i,j}$ at the level $i = 0, 1, .., 10$ for the diffusion bridge given by (3.18) with $\alpha = 0.5$ (solid lines) and $\alpha = 0.0$ (dashed lines) for the Zig-Zag Sampler (blue lines) and the Factorised Boomerang Sampler (red lines) with $T^\star = 2,000$ and Boomerang refreshment rates $\lambda_{\text{refr},i} = 0.01$ for all i.

3.4.3 Dependence upon reference measure

In a final experiment we investigate the dependence of the performance of the Boomerang Sampler upon the choice of reference measure. For this we consider a simple setting in which the target distribution is a standard normal distribution

in d dimensions. However, instead of using the standard normal distribution as reference measure, we perturb it in two ways: (i) we vary the component-wise variance σ^2 of the reference measure, and (ii) we vary the mean $\boldsymbol{x}_\star$ of the reference measure. Specifically, we choose a reference measure $\mathcal{N}(\boldsymbol{x}_\star, \boldsymbol{\Sigma}) \otimes \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, which we choose in case (i) to be $\boldsymbol{x}_\star = \boldsymbol{0}, \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, and in case (ii), $\boldsymbol{x}_\star = \alpha(1, \dots, 1)^\top$, $\boldsymbol{\Sigma} = \boldsymbol{I}$. As performance measure we use the ESS per second for the quantity $|\boldsymbol{x}|^2$. We use refreshment rate 0.1 for Boomerang, and we compare to the Bouncy Particle Sampler, with refreshment 1.0, with both samplers run over a time horizon of 10,000. In Figure 3.5 the results of this experiment are displayed for varying σ^2, and in Figure 3.6 the results are displayed for varying $\boldsymbol{x}_\star$. The box plots are taken over 20 experiments of the Boomerang Sampler, which are compared to a single run of the Bouncy Particle Sampler (dashed line).

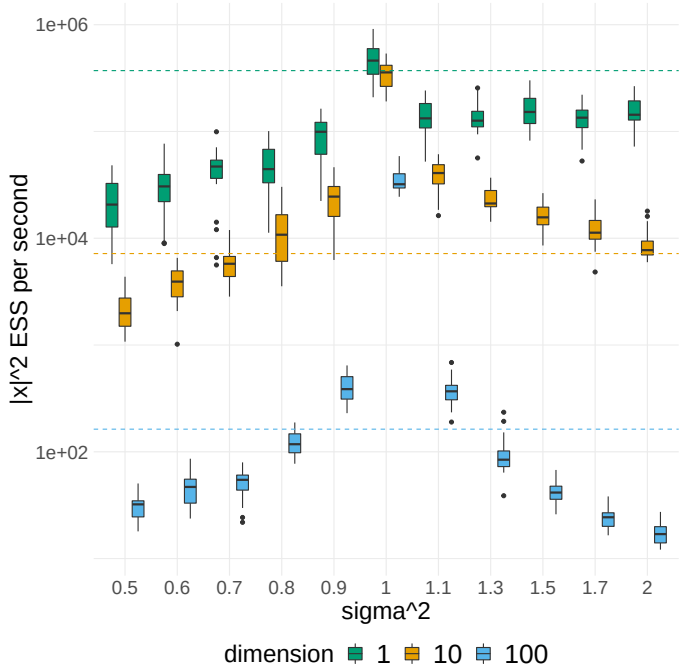


Figure 3.5: Effect of perturbing the variance of the reference measure. As reference measure we choose $\mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \otimes \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, where σ^2 is varied from 0.5 to 2.0.

In this setting, the Boomerang Sampler significantly outperforms the BPS, although the performance is seen to depend upon the choice of reference measure. Note however that the dependencies of $\boldsymbol{\Sigma}$ on σ^2 and of $\boldsymbol{x}_\star$ upon α scale as $\operatorname{trace} \boldsymbol{\Sigma} = \sigma^2 d$ and $\|\boldsymbol{x}_\star\| = \alpha d^{1/2}$ respectively, so that in high dimensional cases the sensitivity on $\boldsymbol{x}_\star$ and $\boldsymbol{\Sigma}$ may be more moderate than might appear from Figures 3.5 and 3.6.

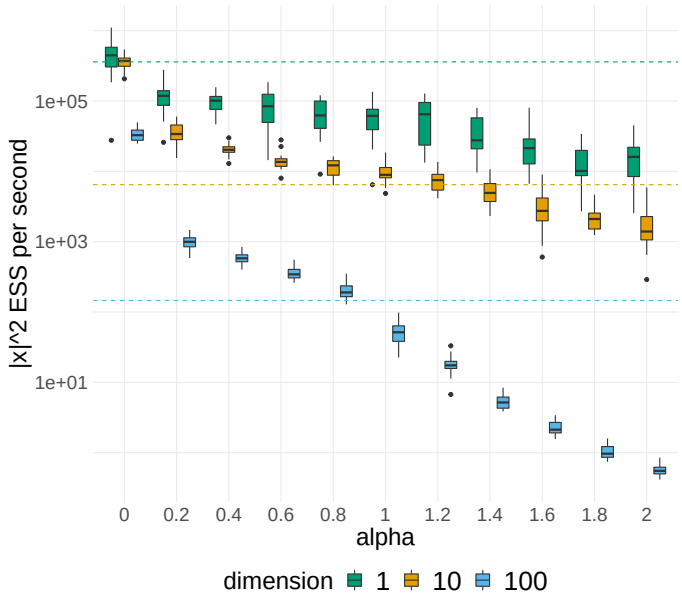


Figure 3.6: Effect of perturbing the mean of the reference measure. As reference measure we choose $\mathcal{N}(\alpha \mathbf{1}, \boldsymbol{I}) \otimes \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, where α is varied from 0.0 to 2.0.

3.5 Conclusion

We presented the Boomerang Sampler as a new and promising methodology, outperforming other piecewise deterministic methods in the large n, moderate d setting, as explained theoretically and by performing a suitable benchmark test. The theoretical properties of the underlying Boomerang Sampler in high dimension are very good. However currently a large computational bound and therefore a large number of rejected switches are hampering the efficiency. We gave a numerical comparison which demonstrates that Boomerang performs well against its natural competitors; however one should be cautious about drawing too many conclusions about the performance of the Boomerang without a more comprehensive simulation study. Further research is required to understand in more detail the dependence upon e.g. reference covariance $\boldsymbol{\Sigma}$, centering position $\boldsymbol{x}_\star$, refreshment rate, computational bounds and the choice of efficiency measure.

We furthermore introduced the Factorised Boomerang Sampler and illustrated its ability to tackle a challenging simulation problem using an underlying sparse structure.

An important direction for further research is the improvement of the computational bounds, in particular with the aim of having good scaling with dimension of subsampled algorithms. Related to this it is important to gain a better understand-

ing of the relative optimality of subsampled versus non-subsampled algorithms in the large n, large d case.

This page is intentionally left blank.

Chapter 4

Sticky PDMPs for variable selection

4.1 Introduction

4.1.1 Overview

Consider the problem of simulating from a measure μ on $\mathbb{R}^d$ that is a mixture of atomic and continuous components. A key application is Bayesian inference for sparse problems and variable selection under a spike-and-slab prior μ_0 of the form

$$\mu_0(\mathrm{d}x) = \prod_{i=1}^{d} \left(w_i \pi_i(x_i)\mathrm{d}x_i + (1-w_i)\delta_0(\mathrm{d}x_i)\right). \tag{4.1}$$

Here, $w_i \in [0,1]$, $\pi_1, \pi_2, \ldots, \pi_d$ are densities with respect to the Lebesgue measure referred to as *slabs* and δ_0 denotes the Dirac measure at zero. For sampling from μ, it is common to construct and simulate a Markov process with μ as invariant measure. Routinely used samplers such as the Hamiltonian Monte Carlo sampler (Duane et al. 1987) cannot be applied directly due to the degenerate nature of μ. We show that "ordinary" samplers based on piecewise deterministic Markov processes (PDMPs) can be adapted to sample from μ by introducing *stickiness*.

In piecewise deterministic Markov processes, the state space is augmented by adding to each coordinate x_i a velocity component v_i, doubling the dimension of the state space. They are characterized by piecewise deterministic dynamics between event times, where event times correspond to changes of velocities. PDMPs have received recent attention because they have good mixing properties (they are non-reversible and have 'momentum', see e.g. Andrieu and Livingstone 2019), they take gradient information into account and they are attractive in Bayesian inference scenarios with a large number of observations because they allow for subsampling of the observations without creating bias (Bierkens, Fearnhead, and Roberts 2019, Bierkens et al. 2020).

We introduce "sticking event times", which occur every time a coordinate of the process state hits 0. At such a time that particular component of the state freezes for an independent exponentially distributed time with a specifically chosen rate equal to $|v_i|\kappa_i$, for some $\kappa_i > 0$ which depends on μ. This corresponds to temporarily setting the marginal velocity to 0: the process "sticks to (or freezes at) 0" in that coordinate, while the other coordinates keep moving, as long as they are not stuck themselves. After the exponentially distributed time the coordinate moves again with its original velocity, see Figure 4.1 for an illustration of the sticky version of the Zig-Zag sampler (Bierkens, Fearnhead, and Roberts 2019). By this we mean that the dynamics of a ordinary PDMP are adjusted such that the process can spend a positive amount of time at the origin, at the coordinate axes and at the coordinate (hyper-)planes by sticking to 0 in each coordinate for a random time span whenever the process hits 0 in that particular coordinate. By restoring the original velocity of each coordinate after sticking at 0, we effectively generate *non-reversible jumps between states with different sets of non-zero coordinates*. In the Bayesian context this corresponds to having non-reversible jumps between models of varying dimensionality.

This allows us to construct a piecewise deterministic process that has a prespecified measure μ as invariant measure, which we assume to be of the form

$$\mu(\mathrm{d}x) = C_\mu \exp(-\Psi(x)) \prod_{i=1}^{d} \left(\mathrm{d}x_i + \frac{1}{\kappa_i} \delta_0(\mathrm{d}x_i) \right) \tag{4.2}$$

for some differentiable function Ψ, normalising constant $C_\mu > 0$ and positive parameters $\kappa_1, \kappa_2, \ldots, \kappa_d$. Here the Dirac masses are located at 0, but generalizations are straightforward. The resulting samplers and processes are referred to as *sticky samplers* and *sticky piecewise deterministic Markov processes* respectively. The proportionality constant C_μ is assumed to be unknown while $(\kappa_i)_{i=1,\ldots,d}$ are known. This is a natural assumption; suppose a statistical model with parameter x and log-likelihood $\ell(x)$ (notationally, we drop the dependence of ℓ on the data). Under the spike-and-slab prior defined in Equation (4.1), the posterior measure is of the form of Equation (4.2) with

$$\Psi(x) = C - \ell(x) - \sum_{i=1}^{d} \log(\pi_i(x_i)), \quad \kappa_i = \frac{w_i}{1 - w_i} \pi_i(0) \tag{4.3}$$

where C, independent of x, can be chosen freely for convenience. A popular choice for π_i is a Gaussian density centered at 0 with standard deviation σ_i. In this case, as $w/(1-w) \approx w$ for $w \approx 0$, κ_i depends linearly on w_i/σ_i in the sparse setting.

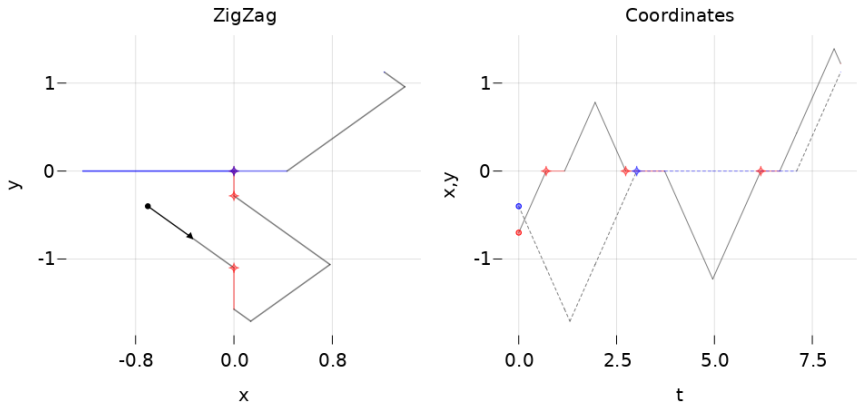


Figure 4.1: 2-dimensional Sticky Zig-Zag sampler with initial position $(-0.75, -0.4)$ and initial velocity $(+1, -1)$. On the left panel, a trajectory on the (x, y)-plane of the Sticky Zig-Zag sampler. The sticky event times relative to the x (respectively y) coordinate and the trajectories with the x (respectively y) stuck at 0 are marked with a blue (respectively red) cross and line. On the right panel, the trajectories of each coordinate against the time using the same (color-) scheme. The trajectory of y is dashed.

Relevant quantities useful for model selection, such as the posterior probability of a model excluding the first variable

$$\mu(\{0\} \times \mathbb{R}^{d-1}) = C_\mu \int \exp(-\Psi(x)) \frac{1}{\kappa_1} \delta_0(\mathrm{d}x_1) \prod_{i=2}^{d} \left(\mathrm{d}x_i + \frac{1}{\kappa_i} \delta_0(\mathrm{d}x_i) \right)$$

cannot be directly computed if C_μ is unknown. However, given a trajectory $(x(t))_{0 \le t \le T}$ of a PDMP with invariant measure μ, the quantity $\mu(\{0\} \times \mathbb{R}^{d-1})$ can be approximated by the ratio T_0/T where $T_0 = \mathrm{Leb}\{0 \le t \le T \colon x_1(t) = 0\}$. This simple, yet general idea requires the user only to specify $\{\kappa_i\}_{i=1}^d$ and Ψ as in Equation (4.2). Moreover, the posterior probability that a collection of variables are all jointly equal to zero can be estimated in a similar way by computing the fraction of time that all corresponding coordinates of the process are simultaneously zero and, more generally, expectations of functionals with respect to the posterior can be estimated from the simulated trajectory.

4.1.2 Related literature

The main purpose of this paper is to show how "ordinary" PDMPs can be adjusted to sample from the measure μ as defined in (4.2). The numerical examples illustrate its applicability in a wide range of applications. One specific application that has

received much attention in the statistical literature is variable selection using a spike-and-slab prior. For the linear model, early contributions include Mitchell and Beauchamp (1988) and George and McCulloch (1993). Some later contributions for hierarchical models derived from the linear model are Ishwaran and Rao (2005), Guan and Stephens (2011), Zanella and Roberts (2019) and Liang, Livingstone, and Griffin (2021). These works have in common that samples from the posterior are obtained from Gibbs sampling and can be implemented in practise only in specific cases (when the Bayes factors between (sub-)models can be explicitly computed). A general and common framework for MCMC methods for variable selection was introduced in Green (1995) and Green and Hastie (2009) and referred to as *reversible jump MCMC*.

Methods that scale better (compared to Gibbs sampling) with either the sample size or dimension of the parameter can be obtained in different ways. Firstly, rather than sampling from the posterior one can *approximate* the posterior within a specified class, for example using variational inference. As an example, Ray, Szabo, and Clara (2020) adopt this approach in a logistic regression problem with spike-and-slab prior. Secondly, one can try to obtain sparsity using a prior which is not of spike-and-slab type. For example, Griffin and Brown (2021) consider Gibbs sampling algorithms for the linear model with priors that are designed to promote sparseness, such as the Laplace or horseshoe prior (on the parameter vector). While such methods scale well with dimension of data and parameter, these target a different problem: the posterior is not of the form (4.2). That is, the posterior itself is not sparse (though derived point estimates may be sparse and the posterior itself may have good properties when viewed from a frequentist perspective). Moreover, part of the computational efficiency is related to the specific model considered (linear or logistic regression model) and, arguably, a generic gradient-based MCMC method would perform poorly on such measures since the gradient of the (log-)density near 0 in each coordinate explodes to account for the change of mass in the neighborhood of 0 induced by the continuous spike component of the prior.

A recent related work by Chevallier, Fearnhead, and Sutton (2020) addresses variable selection problems using PDMP samplers. The different approach taken in that paper is based on the framework of reversible jump (RJ) MCMC as proposed in Green (1995). A comparison between Chevallier, Fearnhead, and Sutton (2020) and our work may be found in Appendix C.3.

4.1.3 Contributions

- We show how to construct sticky PDMP samplers from ordinary PDMP samplers for sampling from the measure in Equation (4.2). This extension allows for informed exploration of sparse models and does not require any additional

tuning parameter. We rigorously characterise the stationary measure of the sticky Zig-Zag sampler.

- We analyse the computational efficiency of the sticky Zig-Zag sampler by studying its complexity and mixing time.

- We demonstrate the performance of the sticky Zig-Zag sampler on a variety of high dimensional statistical examples (e.g. the example in Section 4.4.2 has dimensionality 10^6).

The Julia package `ZigZagBoomerang.jl` (Schauer and Grazzi 2021) implements efficiently the sticky PDMP samplers from this article for general use.

4.1.4 Outline

Section 4.2 formally introduces sticky PDMP samplers and gives the main theoretical results for the sticky Zig-Zag sampler. In Section 4.2.4 we explain how the sticky Zig-Zag sampler may be applied to subsampled data, allowing the algorithm to access only a fraction of data at each iteration, hence reducing the computational cost from $\mathcal{O}(N)$ to $\mathcal{O}(1)$, where N is the sample size. In Section 4.3 we extend the Gibbs sampler for variable selection for target measures of the form of Equation (4.2). We analyse and compare the computational complexity and the mixing times of both the sticky Zig-Zag sampler and the Gibbs sampler. Section 4.4 presents five statistical examples with simulated data and analyses the outputs after applying the algorithms considered in this article. In Section 4.5 both limitations and promising research directions are discussed.

There are five appendices. The derivation of our theoretical results is given in Appendix C.1. Appendix C.2 extends some of the theoretical results for two other sticky samplers: the sticky version of the Bouncy particle sampler (Bouchard-Côté, Vollmer, and Doucet 2018) and the Boomerang sampler (Bierkens et al. 2020), the latter having Hamiltonian deterministic dynamics invariant to a prescribed Gaussian measure. Appendix C.3 contains a self-contained discussion with heuristic arguments and simulations which highlight the differences between the sticky PDMPs and the method of Chevallier, Fearnhead, and Sutton (2020). Appendix C.4 complements Section 4.3 with the details of the derivations of the main results and by presenting local implementations of the sticky Zig-Zag sampler that benefit of a sparse dependence structure between the coordinates of the target measure. Appendix C.5 contains some of the details of the numerical examples of Section 4.4.

4.1.5 Notation

The ith element of the vector $x \in \mathbb{R}^d$ is denoted by x_i. We denote

$$x_{-i} := (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) \in \mathbb{R}^{d-1}$$

. Write

$$(x[k\colon y])_i := \begin{cases} x_i & i \neq k, \\ y & i = k. \end{cases}$$

and $[x]_A := (x_i)_{i \in A} \in \mathbb{R}^{|A|}$ for a set of indices $A \subset \{1, 2, \ldots, d\}$ with cardinality $|A|$. We denote by $\sqcup$ the disjoint union between sets and the positive and negative part of a real-valued function f by $f^+ := \max(0, f)$ and $f^- := \max(0, -f)$ respectively so that $f = f^+ - f^-$. For a topological space E, let $\mathcal{B}(E)$ denote the Borel σ-algebra on E. Denote by $\mathcal{M}(E)$ the class of Borel measurable functions $f\colon E \to \mathbb{R}$ and let $C(E) = \{f \in \mathcal{M}(E)\colon f \text{ is continuous }\}$. For a measure $\mu(\mathrm{d}x, \mathrm{d}y)$ on a product space $\mathcal{X}, \mathcal{Y}$, we write the marginal measure on $\mathcal{X}$ by $\mu(\mathrm{d}x) = \int_{\mathcal{Y}} \mu(\mathrm{d}x, \mathrm{d}y)$.

4.2 Sticky PDMP samplers

In what follows, we formally describe the sticky PDMP samplers (Section 4.2.1) and give the main theoretical results obtained for the sticky Zig-Zag sampler (Section 4.2.3). Section 4.2.4 extends the sticky Zig-Zag sampler with subsampling methods.

4.2.1 Construction of sticky PDMP samplers

The state space of the the sticky PDMPs contains two copies of zero for each coordinate position. This construction allows a coordinate process arriving at zero from below (or above) to spend an exponentially distributed time at zero before jumping to the "other" zero and continuing the dynamics. Formally, let $\overline{\mathbb{R}}$ be the disjoint union $\overline{\mathbb{R}} = (-\infty, 0^-] \sqcup [0^+, \infty)$ with the natural topology[1] τ, where we use the notation 0^-, 0^+ to distinguish the zero element in $(-\infty, 0]$ from the zero element in $[0, \infty)$. The process has *càdlàg*[2] trajectories in the locally compact state space $E = \overline{\mathbb{R}}^d \times \mathcal{V}$, where $\mathcal{V} \subset \mathbb{R}^d$. Pairs of position and velocity will typically be denoted by $(x, v) \in \overline{\mathbb{R}}^d \times \mathcal{V}$. A trajectory reaching zero in a coordinate from below (with positive velocity) or from above (with negative velocity) spends time at the closed

[1] A function $f\colon \overline{\mathbb{R}} \to \mathbb{R}$ is continuous if both restrictions to $(\infty, 0^-]$ and $[0^+, \infty)$ are continuous. If $f(0^-) = f(0^+)$, we write $f(0)$.

[2] I.e., trajectories that are continuous from the right, with existing limits from the left.

end of the half open interval $(-\infty, 0^-]$ or $[0^+, \infty)$, respectively. For $i = 1, \ldots, d$ we define the associated 'frozen boundary' $\mathfrak{F}_i \subset E$ for the ith coordinate as

$$\mathfrak{F}_i := \{(x, v) \in E \colon x_i = 0^-,\ v_i > 0 \ \text{ or } \ x_i = 0^+,\ v_i < 0\}.$$

Thus the ith coordinate of the particle is sticking to zero (or frozen), if the state of the particle belongs to the ith frozen boundary $\mathfrak{F}_i$.

Sometimes, we abuse notation by writing $(x_i, v_i) \in \mathfrak{F}_i$ when $(x, v) \in \mathfrak{F}_i$ as the set $\mathfrak{F}_i$ has restrictions only on x_i, v_i. The closed endpoints of the half-open intervals are somewhat reminiscent of sticky boundaries in the sense of Liggett (2010, Example 5.59). Denote by $\alpha \equiv \alpha(x, v)$ the set of indices of active coordinates corresponding to state (x, v), defined by

$$\alpha(x, v) = \{i \in \{1, 2, \ldots, d\} \colon (x, v) \notin \mathfrak{F}_i\} \tag{4.4}$$

and its complement $\alpha^c = \{1, 2, \ldots, d\} \setminus \alpha$. Furthermore define a jump or *transfer mapping* $T_i \colon \mathfrak{F}_i \to E$ by

$$T_i(x, v) = \begin{cases} (x[i \colon 0^+], v) & \text{if } x_i = 0^-, v_i > 0, \\ (x[i \colon 0^-], v) & \text{if } x_i = 0^+, v_i < 0. \end{cases}$$

The sticky PDMPs on the space E are determined by their infinitesimal characteristics: their dynamics are determined by random state changes happening at random jump times of a time inhomogeneous Poisson process with intensity depending on the state of the process, and a deterministic flow governed by a differential equation in between. The state changes are characterised by a Markov kernel $\mathcal{Q} \colon E \times \mathcal{B}(E) \to [0, 1]$, at random times sampled with state dependent intensity $\lambda \colon E \to [0, \infty)$. The deterministic dynamics are determined coordinate-wise by the integral equation

$$(x_i(t), v_i(t)) = (x_i(s), v_i(s)) + \int_s^t \xi_i(x_i(r), v_i(r))\mathrm{d}r, \quad i = 1, 2, \ldots, d, \tag{4.5}$$

with ξ_i being state dependent with form

$$\xi_i(x, v) = \begin{cases} \bar{\xi}_i(x_i, v_i) & (x_i, v_i) \notin \mathfrak{F}_i \\ (0, 0) & (x_i, v_i) \in \mathfrak{F}_i, \end{cases} \tag{4.6}$$

for functions $\bar{\xi}_i \colon \overline{\mathbb{R}} \times \mathbb{R} \to \overline{\mathbb{R}} \times \mathbb{R}$ which depend on the specific PDMP chosen and corresponds to the coordinate-wise dynamics of the ordinary PDMP while the second case in Equation (4.6) captures the behaviour of the ith coordinate when it sticks at 0.

For PDMP samplers, we typically have $\bar{\xi}_i = \bar{\xi}_j$ for all $i, j \in 1, \ldots, d$ and we have different types of state changes given by Markov kernels $\mathcal{Q}_1$, $\mathcal{Q}_2$, ..., for example refreshments of the velocity, reflections of the velocity, unfreezing of a coordinate etc. If each transition is triggered by its individual independent Poisson clock with intensity $\lambda_1, \lambda_2, \ldots$, then $\lambda = \sum_i \lambda_i$, and $\mathcal{Q}$ itself can be written as the mixture

$$\mathcal{Q}((x,v),\cdot) = \sum_i \frac{\lambda_i((x,v))}{\lambda((x,v))} \mathcal{Q}_i((x,v),\cdot).$$

With that, the dynamics of the sticky PDMP sampler $t \mapsto (X(t), V(t))$ are as follows: starting from $(x, v) \in E$,

1. its flow in each coordinate is deterministic and continuous until an event happens. The deterministic dynamics are given by (4.5). Upon hitting $\mathfrak{F}_i$, the ith coordinate process freezes, captured by the state dependence of (4.6).

2. A frozen coordinate "unfreezes" or "thaws" at rate equal to $\kappa_i|v_i|$ by jumping according to the transfer mapping T_i to the location $(0^+, v_i)$ (or $(0^-, v_i)$) outside $\mathfrak{F}_i$ and continuing with the *same* velocity as before. That is, on hitting $\mathfrak{F}_i$, the ith coordinate process freezes for an independent exponentially distributed time with rate $\kappa_i|v_i|$. This constitutes a non-reversible move between models of different dimension. The corresponding transition $\mathcal{Q}_{i,\text{thaw}}$ is the Dirac measure at $\delta_{T_i(x,v)}$ and the intensity component $\lambda_{i,\text{thaw}}$ equals $\kappa_i|v_i|\mathbf{1}_{\mathfrak{F}_i}$.

3. An inhomogeneous Poisson process λ_{refl} with rate depending on Ψ triggers the reflection events. At a reflection event time, the process changes its velocities according to its reflection rule $\mathcal{Q}_{\text{refl}}$ in such a way that the process is invariant to the measure μ.

4. Refreshment events can be added, where, at exponentially distributed interarrival times, the velocity changes according to a refreshment rule leaving the measure μ invariant. Refreshments are sometimes necessary for the process to be ergodic.

The resulting stochastic process (X_t, V_t) is a sticky PDMP with dynamics $\mathcal{Q}$, λ, φ, initialised in $(X(\tau_0), V(\tau_0))$. Let $s \to \varphi(s, x, v)$ be the deterministic solution of (4.5) starting in (x, v). Set $\tau_0 = 0$ and the initial state $(X(\tau_0), V(\tau_0)) \in E$. A sample of a sticky PDMP is given by the recursive construction in Algorithm 5.

In what follows, we focus our attention on the Sticky Zig-Zag sampler and defer to Appendix C.2 the details of the Bouncy Particle sampler and the Boomerang samplers.

Algorithm 5 PDMP samplers: recursive construction

Given the current state $(X(\tau_k), V(\tau_k))$ at time τ_k

1. Sample independently Δ_k as the first event time of an inhomogeneous Poisson process. We denote $\Delta_k \sim \text{IPP}(s \to \lambda(\varphi(s, X(\tau_k), V(\tau_k)))$, for

$$\mathbb{P}\left(\Delta_k \geq t\right) = \exp\left(-\int_0^t \lambda(\varphi(s, X(\tau_k), V(\tau_k))\mathrm{d}s\right). \tag{4.7}$$

2. Let $\tau_{k+1} = \tau_k + \Delta_k$ and set for $t \in [\tau_k, \tau_{k+1})$

$$(X(t), V(t)) = \varphi(t - \tau_k, X(\tau_k), V(\tau_k)).$$

3. Let

$$(X(\tau_{k+1}), V(\tau_{k+1})) \sim \mathcal{Q}(\varphi(\Delta_k, X(\tau_k), V(\tau_k)), \cdot).$$

4.2.2 Sticky Zig-Zag sampler

A trajectory of the Sticky Zig-Zag sampler has piecewise constant velocity which is an element of the set $\mathcal{V} = \{v \colon |v_i| = a_i, \forall i \in \{1, 2, \ldots, d\}\}$ for a fixed vector a. For each index i, the deterministic dynamics of Equation (4.6) are determined by the function $\bar{\xi}_i(x_i, v_i) = (v_i, 0)$. The reflection rate λ_{refl} is *factorised* coordinate-wise and the reflection event for the ith coordinate is determined by the inhomogeneous rate

$$\lambda_{i,\text{refl}}(x, v) = \mathbb{1}_{i \in \alpha(x,v)}(v_i \partial_i \Psi(x))^+. \tag{4.8}$$

At reflection time of the ith coordinate, the transition kernel $\mathcal{Q}_{i,\text{refl}}$ acts deterministically by flipping the sign of the ith velocity component of the state: $(x_i, v_i) \to (x_i, -v_i)$. As shown in Bierkens, Roberts, and Zitt (2019), the Zig-Zag sampler does not require refreshment events in general to be ergodic.

4.2.3 Theoretical aspects of the Sticky Zig-Zag sampler

A theoretical analysis of the sticky Zig-Zag sampler is given in Appendix C.1.1. In this section we review key concepts and state the main results.

The stationary measure of a PDMP is studied by looking at the extended generator of the process which is an operator characterising the process in terms of local martingales - see Davis (1993, Section 14) for details. The extended generator is - as the name suggests - an extension of the infinitesimal generator of the process (defined for example in Liggett 2010, Theorem 3.16) in the sense that it acts on a larger class of functions than the infinitesimal generator and it coincides with the

infinitesimal generator when applied to functions in the domain of the infinitesimal generator.

A general representation of the extended generator of PDMPs is given in Davis (1993, Section 26), while the infinitesimal generator of the ordinary Zig-Zag sampler is given in the supplementary material of Bierkens, Fearnhead, and Roberts (2019). Here, we highlight the main results we have derived for the sticky Zig-Zag sampler.

Recall $t \to \varphi(t, x, v)$ denotes the deterministic solution of (4.5) starting in (x, v) and τ is the natural topology on E. Define the operator $\mathcal{A}$ with domain

$$\mathcal{D}(\mathcal{A}) = \{f \in \mathcal{M}(E) : t \mapsto f(\varphi(t, x, v)) \ \tau\text{-absolutely continuous } \forall (x, v) \text{ and } \forall i : \lim_{t \downarrow 0} f(x[i\colon 0^+ + t], \cdot) = f(x[i\colon 0^+], \cdot), \lim_{t \downarrow 0} f(x[i\colon 0^- - t], \cdot) = f(x[i\colon 0^-], \cdot)\}$$

by $\mathcal{A}f(x, v) = \sum_{i=1}^{d} \mathcal{A}_i f(x, v)$ with

$$\mathcal{A}_i f(x, v) = \begin{cases} a_i \kappa_i \left(f(T_i(x, v)) - f(x, v)\right) & (x, v) \in \mathfrak{F}_i, \\ v_i \partial_{x_i} f(x, v) + \lambda_i(x, v) \left(f(x, v[i : -v_i]) - f(x, v)\right) & \text{else.} \end{cases}$$

Proposition 4.2.1. *The extended generator of the d-dimensional Sticky Zig-Zag process is given by $\mathcal{A}$ with domain $\mathcal{D}(\mathcal{A})$.*

Proof. See Appendix C.1.4. □

Notice that, the operator $\mathcal{A}$ restricted on $D = \{f \in C_c^1(E), \mathcal{A}f \in C_b(E)\}$ coincides with the infinitesiaml generator of the ordinary Zig-Zag process restricted on D, see Proposition C.1.5, Appendix C.1.4 for details.

Theorem 4.2.2. *The d-dimensional Sticky Zig-Zag sampler is a Feller process and a strong Markov process in the topological space (E, τ) with stationary measure*

$$\mu(\mathrm{d}x, \mathrm{d}v) = \frac{1}{C} \sum_{u \in \mathcal{V}} \exp(-\Psi(x)) \prod_{i=1}^{d} \left(\mathrm{d}x_i + \frac{1}{\kappa_i} \left(\mathbb{1}_{v_i > 0} \, \delta_{0^-}(\mathrm{d}x_i) + \mathbb{1}_{v_i < 0} \, \delta_{0^+}(\mathrm{d}x_i) \right) \delta_u(\mathrm{d}v) \right), \tag{4.9}$$

for some normalization constant $C > 0$.

Proof. The construction of the process and the characterization of the extended generator and its domain of the d-dimensional Sticky Zig-Zag process can be found in Appendix C.1.1. We then prove that the process is Feller and strong Markov (Appendix C.1.2 and Appendix C.1.3). By Liggett (2010, Theorem 3.37), μ is a stationary measure if, for all $f \in D$, $\int \mathcal{L} f \mathrm{d}\mu = 0$. This last equality is derived in Appendix C.1.5. □

Theorem 4.2.3. *Suppose Ψ satisfies Assumption C.1.2. Then the sticky Zig-Zag process is ergodic and μ is its unique stationary measure.*

Proof. See Appendix C.1.6. □

The following remark establishes a formula for the recurrence time of the Sticky Zig-Zag to the null model, and may serve as guidance in design of the probabilistic model or the choice of the parameter κ_i, here assumed for simplicity to be all equal.

Remark 4.2.4. (Recurrence time of the Sticky Zig-Zag to zero) *The expected time to leave the position* $\mathbf{0} = (0, 0, \ldots, 0)$ *for a* d*-dimensional Sticky Zig-Zag with unit velocity components is* $\frac{1}{\kappa d}$ *(since each coordinate leaves 0 according to an exponential random variable with parameter* κ*). A simple argument given in Appendix C.1.7 shows that the expected time of the process to return to the null model is*

$$\frac{1 - \mu(\{\mathbf{0}\})}{d\kappa\mu(\{\mathbf{0}\})}. \tag{4.10}$$

4.2.4 Extension: sticky Zig-Zag sampler with subsampling method

Here we address the problem of sampling a d-dimensional target measure when the log-likelihood is a sum of N terms, when d and N are large. Consider for example a regression problem where both the number of covariates and the number of experimental units in the dataset are large. In this situation full evaluation of the log-likelihood and its gradient is prohibitive. However, PDMP samplers can still be used with the exact subsampling technique (e.g. Bierkens, Fearnhead, and Roberts 2019) as this allows for substituting the gradient of the log-likelihood (which is required for deriving the reflection times) by an estimate of it which is cheaper to evaluate, without introducing any bias on the output of the sampler.

The subsampling technique for Sticky Zig-Zag samplers requires to find an unbiased estimate of the gradient of Ψ in (4.2). To that end, assume the following decomposition:

$$\partial_{x_i}\Psi(x) = \left(\sum_{j=1}^{N_i} S(x, i, j)\right), \quad \forall x \in \overline{\mathbb{R}}^d,\ i = 1, 2, \ldots, d, \tag{4.11}$$

for some scalar valued function S. This assumption on Ψ is satisfied for example for the setting with a spike-and-slab prior and a likelihood that is a product of factors, such as for likelihoods of (conditionally) independent observations.

For fixed (x, v) and $x^* \in \mathbb{R}^d$, for each $i \in \alpha(x, v)$ the random variable

$$N_i \left(S(x, i, J) - S(x^*, i, J)\right) + \partial_{x_i}\Psi(x^*), \quad J \sim \text{Unif}(\{1, 2, \ldots, N_i\})$$

is an unbiased estimator for $\partial_{x_i}\Psi(x)$. Define the Poisson rates

$$\widetilde{\lambda}_{i,j}(x, v) = (v_i N_i(S(x, i, j) - S(x^*, i, j)) + v_i\partial_{x_i}\Psi(x^*))^+$$

and, for each $i \in \alpha$, define the bounding rate

$$\overline{\lambda}_i(t, x, v) \geq \widetilde{\lambda}_{i,j}(\varphi(t, x, v)), \quad t \geq 0, \forall j \in \{1, 2, \ldots, N_i\},$$

which is specified by the user and such that Poisson times with inhomogeneous rate $\tau \sim \text{IPP}(s \to \overline{\lambda}_i(s, x, v))$ can be simulated (see Appendix C.4.2 for details on the simulation of Poisson times).

The Sticky Zig-Zag with subsampling has the following dynamics:

- the deterministic dynamics and the sticky events are identical to the ones of the Sticky Zig-Zag sampler presented in Section 4.2.3;
- a *proposed* reflection time equals $\min_{i \in \alpha(x,v)} \tau_i$, with $\{\tau_i\}_{i \in \alpha(x,v)}$ being independent inhomogeneous Poisson times with rates $s \to \overline{\lambda}_i(s, x, v)$;
- at the proposed reflection time τ triggered by the ith Poisson clock, the process reflects its velocity according to the rule $(x, v) \to (x, v[i, -v_i])$ with probability $\widetilde{\lambda}_{i,J}(\varphi(\tau, x, v))/\overline{\lambda}_i(\tau, x, v)$ where $J \sim \text{Unif}(\{1, 2, \ldots, N_i\})$.

Proposition 4.2.5. *The Sticky Zig-Zag with subsampling has a unique stationary measure given by Equation* (4.9).

The proof of Proposition 4.2.5 follows with a similar argument made in the proof of Bierkens, Fearnhead, and Roberts (2019, Theorem 4.1). The number of computations required by the Sticky Zig-Zag with subsampling to compute the next event time with respect to the quantity N is $\mathcal{O}(1)$ (since $\partial_{x_i}\Psi(x^*)$ can be precomputed). This advantage comes at the cost of introducing 'shadow event times', which are event times where the velocity component does not reflect. In case the posterior density satisfies a Bernstein-von-Mises theorem, the advantage of using subsampling over the standard samplers has been empirically shown and informally argued for in Bierkens, Fearnhead, and Roberts (2019, Section 5) and Bierkens et al. (2020, Section 3) for large N and when choosing x^* to be the mode of the posterior density.

4.3 Performance comparisons for Gaussian models

In this section we discuss the performance of the Sticky Zig-Zag sampler in comparison with a Gibbs sampler. The sticky Zig-Zag sampler includes new coordinates randomly but uses gradient information to find which coordinates are zero. By comparing to a Gibbs sampler that just proposes models at random, we show that it is an efficient scheme of exploration. As the Gibbs sampler requires closed form expression of Bayes factors between different (sub-)models (Equation (4.12) below),

we consider Gaussian models. The comparison is motivated by considering two samplers that do not require model specific proposals or other tuning parameters. In specific cases such as the target models considered below, the Gibbs sampler could be improved by carefully choosing a problem-specific proposal kernel in between (sub-)models, see for example Zanella and Roberts (2019) and Liang, Livingstone, and Griffin (2021) – something we don't consider here.

The comparison is primarily in relation to the dimension d, average number of active particles and sample size N of the problem. It is well known that the performance of a Markov chain Monte Carlo method is given by both the computational cost of simulating the algorithm and the convergence properties of the underlying process. In Section 4.3.2 we consider both these aspects and compare the results obtained for the sticky Zig-Zag sampler with those relative to the Gibbs sampler. The results are summarised in Table 4.1 and Table 4.2. The technical details of this section are given in Appendix C.4.

4.3.1 Gibbs sampler

We can use a set of active indices α to define a model, as the corresponding set of non-zero values in $\mathbb{R}^d$:

$$\mathcal{M}_\alpha := \{x \in \mathbb{R}^d \colon x_i = 0, i \notin \alpha\} \quad \text{for } \alpha \subset \{1, 2, \dots, d\}.$$

For every set of indices $\alpha \subset \{1, 2, \dots, d\}$ and for every j, the Bayes factors relative to two neighbouring (sub-)models (those differing by only one coefficient) for a measure as in Equation (4.2) are given by

$$B_j(\alpha) = \frac{\mu(\mathcal{M}_{\alpha \cup \{j\}})}{\mu(\mathcal{M}_{\alpha \setminus \{j\}})} = \frac{\kappa_j \int_{\mathbb{R}^{|\alpha \cup \{j\}|}} \exp(-\Psi(y)) \mathrm{d}x_{\alpha \cup \{j\}}}{\int_{\mathbb{R}^{|\alpha \setminus \{j\}|}} \exp(-\Psi(z)) \mathrm{d}x_{\alpha \setminus \{j\}}}, \tag{4.12}$$

where $y = \{x \in \mathbb{R}^d \colon x_i = 0,\ i \notin (\alpha \cup \{j\})\}$, $z = \{x \in \mathbb{R}^d \colon x_i = 0,\ i \notin (\alpha \setminus \{j\})\}$. The Gibbs sampler starting in (x, α), with $x_i \neq 0$ only if $i \in \alpha$ for some set of indices $\alpha \subset \{1, 2, \dots, d\}$, iterates the following two steps:

1. Update α by choosing randomly $j \sim \mathrm{Unif}(\{1, 2, \dots, d\})$ and set $\alpha \leftarrow \alpha \cup \{j\}$ with probability p_j where p_j satisfies $p_j/(1 - p_j) = B_j(\alpha)$, otherwise set $\alpha \leftarrow \alpha \setminus \{j\}$.

2. Update the free coefficients x_α according to the marginal probability of x_α conditioned on $x_i = 0$ for all $i \in \alpha^c$.

In Appendix C.4.1, we give an analytical expressions for the right hand-side of Equation (4.12) and the conditional probability in step 2 when Ψ is a quadratic function of x. For logistic regression models, neither step 1 nor step 2 can be

directly derived and the Gibbs samplers makes use of a further auxiliary Pólya-Gamma random variable ω which has to be simulated at every iteration and makes the computations of step 1 and step 2 tractable, conditionally on ω (see Polson, Scott, and Windle 2013 for details).

4.3.2 Runtime analysis and mixing times

The ordinary Zig-Zag sampler can greatly profit in the case of models with a sparse conditional dependence structure between coordinates by employing local versions of the standard algorithm as presented in Bierkens et al. (2021). In Appendix C.4.2 we discuss how to simulate sticky PDMPs and derive similar local algorithms relative to the sticky Zig-Zag. Also the Gibbs sampler algorithm, as described in Section 4.3.1, benefits when the conditional dependence structure of the target is sparse. In Appendix C.4.3 we analyse the computational complexity of both algorithms. In the analysis, we drop the dependence on (x, v) and we assume that the size of $\alpha(t) := \{i \colon x_i(t) \neq 0\}$ fluctuates around a typical value p in stationarity. Thus p represents the number of non-zero components in a typical model, and can be much smaller than d in sparse models.

Table 4.1 summarises the results obtained of both algorithms in terms of the sample size N and p when the conditional dependence structure between the coordinates of the target is full and the sub-sampling method presented in Section 4.2.4 cannot be employed (left-column) and when there is sparse dependence structure and subsampling can be employed (right-column). Our findings are validated by numerical experiments in Section 4.4 (Figure 4.5, Figure 4.9).

Algorithm	Worst case	Best case
Sticky Zig-Zag	$p^2 N$	p
Gibbs sampler	$p(p^2 + N)$	$p(\sqrt{p} + N)$

Table 4.1: Computational scaling of the Sticky Zig-Zag algorithm and the Gibbs sampler for variable selection for p and sample size N. Worst case is when the target density does not present any conditional independence structure and the subsampling method for the Sticky Zig-Zag cannot be employed; best case when the target measure presents a relevant conditional independence structure and subsampling can be employed.

We now turn our focus on the mixing time of both the underlying processes. Given the different nature of dependencies of the two algorithms, a rigorous and theoretical comparison of their mixing times is difficult and outside the scope of this work. We therefore provide an heuristic argument for two specific scenarios where we let both algorithms be initialized at $x \sim \mathcal{N}_d(0, I) \in \mathbb{R}^d$, hence in the full model,

and assume that the target μ assigns most of its probability mass to the null model $\mathcal{M}_\emptyset$. Then we derive the expected hitting time to $\mathcal{M}_\emptyset$ for both processes. The two scenarios differ as in the former case the target μ is supported in every sub-model so that the process can reach the point $(0, 0, \ldots, 0)$ by visiting any sequence of sub-models while in the latter case the measure μ is supported in a single nested sequence of sub-models. Details of the two scenarios are given in Appendix C.4.4. Table 4.2 summarizes the scaling results (in terms of dimensions d) derived in the two cases considered.

Algorithm	μ supported on every model	μ supported on a nested sequence
Sticky Zig-Zag	$\log(d)$	d
Gibbs sampler	$d\log(d)$	d^2

q

Table 4.2: Scaling relative to the dimension d of the expected time (number of iteration for the Gibbs sampler) to travel from the full model (initialized as a standard Gaussian random variable) to the null model (which is the mode of the target). The results are for targets which are supported in every model and for targets supported on a single sequence of nested sub-models.

4.4 Examples

In this section we apply the Sticky Zig-Zag sampler and, when possible, compare its performance with the Gibbs sampler in five different problems of varying nature and difficulty:

4.4.1 *(Learning networks of stochastic differential equations)* A system of interacting agents where the dynamics of each agent are given by a stochastic differential equation. We aim to infer the interactions among agents. This is an example where the likelihood does not factorise and the number of parameters increases quadratically with the number of agents. We demonstrate the Sticky Zig-Zag sampler under a spike-and-slab prior on the parameters that govern the interaction and compare this with the Gibbs sampler.

4.4.2 *(Spatially structured sparsity)* An image denoising problem where the prior incorporates that a large part of the image is black (corresponding to sparsity), but also promotes positive correlation among neighbouring pixels. Specifically, this examples illustrates that the Sticky Zig-Zag sampler can be employed in high dimensional regimes (the showcase is in dimension one million) and for sparsity promoting priors other than factorised priors such as spike-and-slab priors.

4.4.3 *(Sampling from a bimodal target)* A multi-modal model based on Gaussian increments of an unknown parameter; this is a constructed example where the Gibbs sampler fails to mix while the Sticky Zig-Zag sampler mixes well.

4.4.4 *(Logistic regression)* The logistic regression model where both the number of covariates and the sample size are large, while assuming the coefficient vector to be sparse. This is an non-Gaussian optimal scenario where the Sticky Zig-Zag sampler can be employed with subsampling technique achieving $\mathcal{O}(1)$ scaling with respect to the sample size.

4.4.5 *(Estimating a sparse precision matrix)* The setting where N realisations of independent Gaussian vectors with precision matrix of the form XX' are observed. Sparsity is assumed on the off-diagonal elements of the lower-triangular matrix X. What makes this example particularly interesting is that the gradient of the log-likelihood explodes in some hyper-planes, complicating the application of gradient-based Markov chain Monte Carlo methods.

In all cases we simulate data from the model and assume the parameter to be sparse (i.e. most of its elements are assumed to be zero) and high dimensional. In case a spike-and-slab prior is used, the slabs are always chosen to be zero-mean Gaussian with (large) variance σ_0^2. The sample sizes, parameter dimensions and additional difficulties such as correlated parameters or non-linearities which are considered in this section illustrate the computational efficiency of our method (and implementation) in a wide range of settings. In all examples we used either the local or the fully local algorithm of the Sticky Zig-Zag as detailed in Appendix C.4.2 with velocities in the set $\mathcal{V} = \{-1, +1\}^d$. Comparisons with the Gibbs sampler are possible for Gaussian models and the logistic regression model. Our implementation of the Gibbs sampler is taking advantage of model sparsity. Because of its computational overhead, when such comparisons are included, the dimensionality of the problems considered has been reduced. The performance of the two algorithms is compared by running the two algorithms for approximately the same computing time. As performance measure we consider the squared error as a function of the computing time:

$$c \mapsto \mathcal{E}_{\mathrm{s}}(c) := \sum_{i=1}^{d} (p_i^{\mathrm{s}}(c) - \overline{p}_i)^2, \tag{4.13}$$

where c denotes computing time (we use c rather than t as the latter is used as time index for the Zig-Zag sampler). In the displayed expression, we first compute $\overline{p}_i$, which is an approximation to the posterior probability of the ith coordinate being nonzero. This quantity can either be obtained by running the Sticky Zig-Zag sampler or the Gibbs sampler (if applicable) for a very long time. As we show the Sticky Zig-Zag sampler to converge faster, especially in high dimensional problems, we use

this sampler in approximating this value. We stress that the same result could be obtained by running the Gibbs sampler for a very long time. More precisely, we compute for each coordinate of the Sticky Zig-Zag sampler the fraction of time it is nonzero. In $\mathcal{E}_{\mathrm{s}}(c)$, the value of $\overline{p}_i$ is compared to $p_i^{\mathrm{s}}(c)$ which is the fraction of time (or fraction of samples in case of the Gibbs sampler) where x_i is nonzero using computational budget c and sampler 's'. All the experiments were carried out with a conventional laptop with Intel core i5-10310 processor and 16GB DDR4 RAM. Pre-processing time and memory allocation of both algorithms are comparable.

4.4.1 Learning networks of stochastic differential equations.

In this example we consider a stochastic model for p autonomously moving agents ("boids") in the plane. The dynamics of the location of the ith agent is assumed to satisfy the stochastic differential equation

$$\mathrm{d}U_i(s) = -\lambda U_i(s)\mathrm{d}s + \sum_{j\neq i} x_{i,j}(U_j(s) - U_i(s))\mathrm{d}s + \sigma \mathrm{d}W_i(s), \qquad 1 \leq i \leq p \quad (4.14)$$

where, for each i, $(W_i(s))_{0\leq s\leq T}$ is an independent 2-dimensional Wiener process. We assume the trajectory of each agent is observed continuously over a fixed interval $[0, T]$. This implies $\sigma > 0$ can be considered known, as it can be recovered without error from the quadratic variation of the observed path. For simplicity we will also assume the mean-reversion parameter $\lambda > 0$ to be known. Let $x = \{x_{i,j} \colon i \neq j\} \in \mathbb{R}^{p^2-p}$ denote the unknown parameter. If $x_{i,j} > 0$, agent i has the tendency to follow agent j, on the other hand, if $x_{i,j} < 0$, agent i tends to avoid agent j. Hence, estimation of x aims at inferring which agent follows/avoids other agents. We will study this problem from a Bayesian point of view assuming sparsity of x, incorporated via the prior using a spike and slab prior. This problem has been studied previously in Bento, Ibrahimi, and Montanari (2010) using ℓ_1-regularised least squares estimation.

Motivation for studying this problem can be found in Reynolds (1987) and the presentation at JuliaCon 2020 by Jesse Bettencourt (2020). An animation of the trajectories of the agents in time can be found at Grazzi and Schauer (2021).

Suppose $U_i(s) = (U_{i,1}(s), U_{i,2}(s))$ and let

$$Y(s) = (U_{1,1}(s), \ldots, U_{p,1}(s), U_{1,2}(s), \ldots, U_{p,2}(s))$$

denote the vector obtained upon concatenation of all x-coordinates and y-coordinates of all agents. Then, it follows from Equation (4.14) that $\mathrm{d}Y(s) = C(x)Y(s)\mathrm{d}s + \sigma \mathrm{d}W(s)$, where $W(s)$ is a Wiener process in $\mathbb{R}^{2p}$. Here, $C(x) = \mathrm{diag}(A(x), A(x))$

where

$$A(x) = \begin{bmatrix} -\lambda - \overline{x}_1 & x_{1,2} & x_{1,3} & \dots \\ x_{2,1} & -\lambda - \overline{x}_2 & x_{2,3} & \\ x_{3,1} & & \ddots & \\ \vdots & & & \end{bmatrix}$$

with $\overline{x}_i = \sum_{j \neq i} x_{i,j}$. If $\mathbb{P}_x$ denotes the measure on path space of $Y_T := (Y(s),\, s \in [0,T])$ and $\mathbb{P}_0$ denotes the Wiener-measure on $\mathbb{R}^{2p}$, then it follows from Girsanov's theorem that

$$\ell(x) := \log \frac{\mathbb{P}_x}{\mathbb{P}_0}(Y_T) = \frac{1}{\sigma^2} \int_0^T (C(x)Y(s))' \mathrm{d}Y(s) - \frac{1}{2\sigma^2} \int_0^T \|C(x)Y(s)\|^2 \mathrm{d}s. \quad (4.15)$$

As we will numerically only be able to store the observed sample path on a fine grid, we approximate the integrals appearing in the log-likelihood $\ell(x)$ using a standard Riemann-sum approximation of Itô integrals (see e.g. Rogers and Williams 2000a, Ch. IV, sec. 47) and time integrals. We assume x to be sparse which is incorporated by choosing a spike-and-slab prior for x as in Equation (4.1). The posterior measure is of the form of (4.2) with κ and $\Psi(x)$ as in (4.3). As $x \mapsto \Psi(x)$ is quadratic, the reflection times of the Sticky Zig-Zag sampler can be computed in closed form.

Numerical experiments: In our numerical experiments we fix $p = 50$ (number of agents), $T = 200$ (length of time-interval), $\sigma = 0.1$ (noise-level) and $\lambda = 0.2$ (mean-reversion coefficient). We set the parameter x such that each agent has one agent that tends to follow and one agent that tends to avoid. Hence, for every i, we set $x_{i,j}$ to be zero for all $j \neq i$, except for 2 distinct indices $j_1, j_2 \sim \mathrm{Unif}(\{1,2,\dots,d\} \setminus i)$ with $x_{i,j_1} x_{i,j_2} < 0$. The parameter x is very sparse and it is highly nontrivial to recover its value. We then simulate Y_T using Euler forward discretization scheme, with step-size equal to 0.1 and initial configuration $Y(0) \sim \mathcal{N}_{2p}(0, I)$.

The prior weights $w_1 = w_2 = \cdots = w_d$ (w_i being the prior probability of the ith coordinate to be nonzero) are conveniently chosen to equal the proportion of non-zero elements in the true (data-generating) parameter vector x. The variance of each slab was taken to be $\sigma_0^2 = 50$. We ran the Sticky Zig-Zag sampler with final clock 500, where the algorithm was initialized in the full-model with no coordinate frozen at 0 at the posterior mean of the Gaussian density proportional to Ψ.

Figure 4.2 shows the discrepancy between the parameters used during simulation (ground truth) and the estimated posterior median. In this figure, from the (sticky) Zig-Zag trajectory of each element $x_{i,j}$ $(i \neq j)$ we collected their values at time $t_i = i0.1$ and subsequently computed the median of the those values. We conclude that all parameters which are strictly positive (coloured in pink) are recovered well. At the bottom of the figure (black points and crosses), 25 are incorrectly identified

as either being zero or negative. In this experiment, the Sticky Zig-Zag sampler outperforms the Gibbs sampler considerably.

In Figure 4.3 we compare the performance of the Sticky Zig-Zag sampler with the Gibbs sampler. Here, all the parameters (including initialisation) are as above, except now the number of agents is taken as $p = 20$. Both $c \mapsto \mathcal{E}_{\text{Zig-Zag}}(c)$ and $c \mapsto \mathcal{E}_{\text{Gibbs}}(c)$, with c denoting the computational budget, are computed for $c \in [0, 10]$. For this, the final clock of the Zig-Zag was set to 10^4 and the number of iterations for the Gibbs sampler was set to 1.2×10^4. For obtaining $\bar{p}_i$ the Sticky Zig-Zag sampler was run with final clock 5×10^4 (taking approximately 50 seconds computing time).

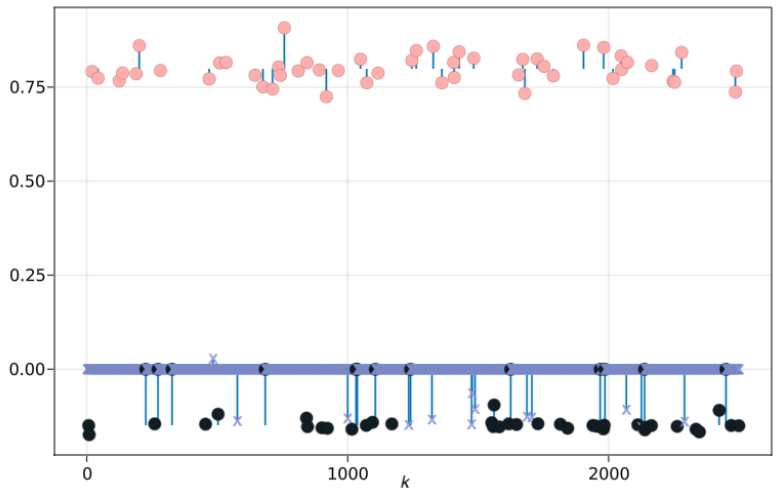


Figure 4.2: Posterior median estimate of x_k (where k can be identified with (i, j)) versus k computed using the Sticky Zig-Zag sampler. Thin vertical lines indicate distance to the truth. True zeros are plotted with the symbol $\times$, others are plotted as points. With $p = 50$ agents, the dimension of the problem is $d = 2450$

4.4.2 Spatially structured sparsity

We consider the problem of denoising a spatially correlated, sparse signal. The signal is assumed to be an $n \times n$-image. Denote the observed pixel value at location (i, j) by $Y_{i,j}$ and assume

$$Y_{i,j} = x_{i,j} + Z_{i,j}, \quad Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2), \qquad i, j \in \{1, \dots, n\}.$$

The "true signal" is given by $x = \{x_{i,j}\}_{i,j}$ and this is the parameter we aim to infer, while assuming σ^2 to be known. We view x as a vector in $\mathbb{R}^d$, with $d = n^2$ but use both linear indexing x_k and Cartesian indexing $x_{i,j}$ to refer to the component

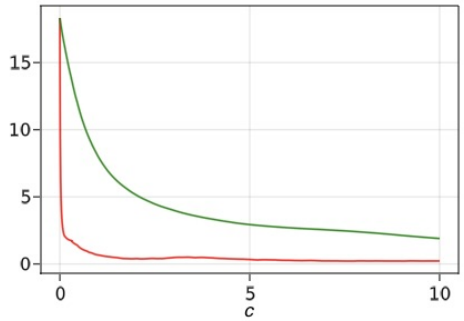


Figure 4.3: Squared error of the marginal inclusion probabilities (Equation 4.13) $c \to \mathcal{E}_{\text{zig-zag}}(c)$ (red) and $c \to \mathcal{E}_{\text{gibbs}}(c)$(green) where c represent the computing time in seconds. With $p = 20$ agents the dimension of the problem is $p(p-1)/2 = 380$.

at index $k = n(i-1)+j$. The log-likelihood of the parameter x is given by $\ell(x) = C + \sigma^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_{i,j} - Y_{i,j}|^2$, with C a constant not depending on x.

We consider the following prior measure

$$\mu_0(\mathrm{d}x) = \exp\left(-\frac{1}{2}x'\Gamma x\right)\prod_{i=1}^{d}\left(\mathrm{d}x_i + \frac{1}{\kappa}\delta_0(\mathrm{d}x_i)\right).$$

The Dirac masses in the prior encapsulate sparseness in the underlying signal and an appropriate choice of Γ can promote smoothness. Overall, the prior encourages *smoothness, sparsity and local clustering of zero entries and non-zero entries*. As a concrete example, consider $\Gamma = c_1\Lambda + c_2 I$ where Λ is the graph Laplacian of the pixel neighbourhood graph: the pixel indices i, j are identified with the vertices $V = \{(i,j)\colon\ (i,j) \in \{1,\dots,n\}^2\}$ of the $n \times n$ -lattice with edges $E = \{\{v, v'\} : (v, v') = ((i,j),(i',j')) \in V^2, |i-i'|+|j-j'| = 1\}$ (using the set notation for edges). Thus, edges connect a pixel to its vertical and horizontal neighbours. Then

$$\lambda_{v,v'} = \begin{cases} \text{degree}(v) & v = v' \\ -1 & \{v, v'\} \in E \\ 0 & \text{otherwise} \end{cases}$$

and $\Lambda = (\Lambda_{k,l})_{k,l\in\{1,\dots,n^2\}}$ with $\Lambda_{(i-1)n+j,(k-1)n+l} = \lambda_{(i,j),(k,l)}$, for $\quad i,j,k,l \in \{1,\dots,n\}$.

This is a prior which is applicable in similar situations as the fused Lasso in Tibshirani et al. (2005).

Numerical experiments: We assume that pixel (i,j) corresponds to a physical location of size $\Delta_1 \times \Delta_2$ centered at $u(i,j) = u_0 + (i\Delta_1, j\Delta_2) \in \mathbb{R}^2$. To numerically illustrate our approach, we use a heart shaped region given by $x_{i,j} = 5\max(1 -$

$h(u(i,j)),0)$ where $h\colon \mathbb{R}^2 \to [0,\infty)$ is defined by $h(u_1,u_2) = u_1^2 + \left(\frac{5u_2}{4} - \sqrt{|u_1|}\right)^2$, $u_0 = (-4.5,-4.1)$, $n = 10^3$ and $\Delta_1 = \Delta_2 = 9/n$. In the example, about 97% of the pixels of the truth are black. The dimension of the parameter equals 10^6. Figure 4.4, top-left, shows the observation Y with $\sigma^2 = 0.5$ and the ground truth.

As the ordinary Sticky Zig-Zag sampler would require storing and ordering 1 million elements in the priority queue we ran the Sticky Zig-Zag sampler with sparse implementation as detailed in Remark C.4.1. For this example, we have $\Psi(x) = \ell(x) + 0.5x'\Gamma x$. We took $c_1 = 2, c_2 = 0.1$ in the definition of Γ and chose the parameters $\kappa_1 = \kappa_2 = \cdots = \kappa_d = 0.15$ for the smoothing prior. The reflection times are computed by means of a thinning scheme, see Appendix C.5.2 for details. We set the final clock of the Sticky Zig-Zag sampler to 500. Results from running the sampler are summarized in Figure 4.4.

In Figure 4.5, the runtimes of the Sticky Zig-Zag sampler and Gibbs sampler are shown (in a log-log scale) for different values of n^2 (dimensionality of the problem), the final clock was fixed to $T = 500$ (10^3 iteration for the Gibbs sampler). All the other parameters are kept fixed as described above. The results agree well with the scaling results of Table 4.1, rightmost column.

In Figure 4.6 we show $t \to \mathcal{E}_{\text{Zig-Zag}}(t)$ and $t \to \mathcal{E}_{\text{Gibbs}}(t)$ for t ranging from 0 to 5, in case $n = 20$. Both samplers were initialized at the posterior mean of the Gaussian density proportional to Ψ (hence, in the full-model with no coordinates set to 0). In this experiment, the Sticky Zig-Zag sampler outperforms the Gibbs sampler considerably.

4.4.3 Sampling from a bimodal target

In this section we present an example where the Gibbs sampler has substantial difficulties exploring the state-space, while the Sticky Zig-Zag sampler performs well. Fix the functions $\mu^{(j)}\colon \mathbb{R} \to \mathbb{R}$ with $j = 1,2,3$ and grid points $t_0 < t_1 < \cdots < t_n$. Let $x_{i,j} = \mu^{(j)}(t_i)$. Rather than considering the statistical model with observations $Y_{i,j} \sim \mathcal{N}(x_{i,j},\sigma^2)$ (corresponding to a standard linear regression) we consider the model with observations $\bar{Y}_{i,j} \sim \mathcal{N}(x_{i,j} - x_{i-1,j},\sigma^2)$, corresponding to observing increments of each $\mu^{(j)}$ with error. The likelihood of $\{\bar{Y}_{1,j},\ldots,\bar{Y}_{n,j}\}_{j=1,2,3}$ is given by (we omit dependence on the observations in the notation)

$$\ell(x) = C - \frac{1}{2\sigma^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(x_{i,j} - x_{i-1,j} - \bar{Y}_{i,j})^2, \qquad x = (x_0,x_1,\ldots,x_n),$$

for some constant C which does not depend on x. As the likelihood is invariant under mapping each $x_{i,j}$ to $x_{i,j} + c_j$, it is clear that there is no unique maximum likelihood estimator. However, upon specification of a prior distribution, the posterior distribution is well defined. We impose a spike-and-slab prior on x (equation (4.1)), with

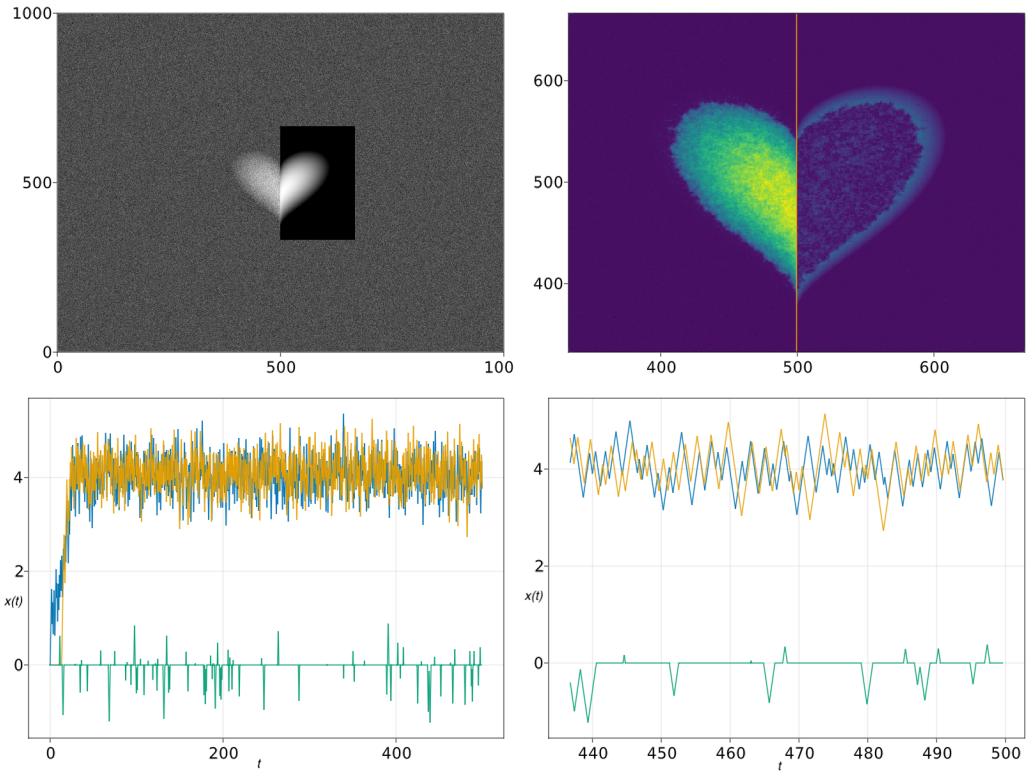


Figure 4.4: Top-left: observed $1\,000 \times 1\,000$ image of a heart corrupted with white noise, with part of the ground truth inset. Top-right, left half: posterior mean estimated from the trace of the Sticky Zig-Zag sampler (detail). Top-right, right half: mirror image showing the absolute error between the posterior mean and the ground truth in the same scale (color gradient between blue (0) and yellow (maximum error)). Bottom: trace plot of 3 coordinates; on the left the full trajectory is shown whereas on the right only the final 60 time units are displayed. The traces marked with blue and orange lines belong to neighbouring coordinates (highly correlated) from the center, the trace marked with green belongs to a coordinate outside the region of interest.

each slab centered at zero and with variance σ_0^2, thereby encouraging that flat parts of the curves $t \mapsto \mu^{(j)}(t)$ are located at the horizontal axis. Suppose the ground truth for $\mu := (\mu^{(1)}, \mu^{(2)}, \mu^{(3)})$ is given by

$$\mu^{(j)}(t_i) = c_j + \begin{cases} 0 & \text{if } j \neq 2 \\ 6t_i + 2\pi\cos(t_i) & \text{if } j = 2 \end{cases}, \quad t_i = 0, \frac{1}{n}2\pi, \ldots, \frac{n-1}{n}2\pi, 2\pi.$$

for values $c_j \in \mathbb{R}$, $j = 1, 2, 3$. It is clear from the description that with this choice of μ and prior, the blocks of coordinates $\{x_{i,1}\}_{i=0,1\ldots,n}$ and $\{x_{i,3}\}_{i=0,1\ldots,n}$ are likely to be

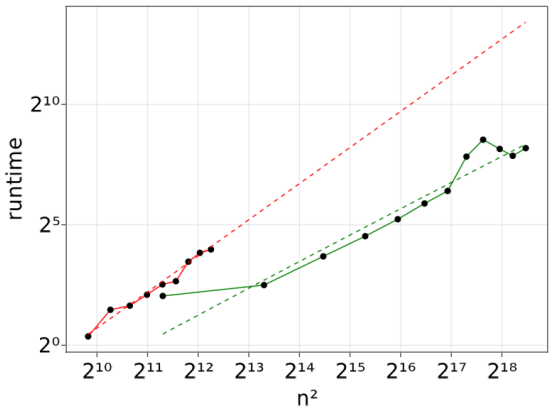


Figure 4.5: Runtime comparison of the Sticky Zig-Zag sampler (green) and the Gibbs sampler (red) for the example in Subsection 4.4.2. The horizontal axis displays the dimension of the problem, which is n^2. The vertical axis shows runtime in seconds. The runtime is evaluated at $n^2 = 50^2, 100^2, \ldots, 600^2$ for the sticky Zig-Zag sampler and at $n^2 = 40^2, 45^2, \ldots, 70^2$ for the Gibbs sampler. Both plots are on a log-log scale. The dashed curves shows the theoretical scaling (including a log-factor for the priority queue insertion): $x \mapsto c_1 x \log(x)$ (green) and $x \mapsto c_2 x^{3/2}$ (orange), with c_1 and c_2 chosen conveniently.

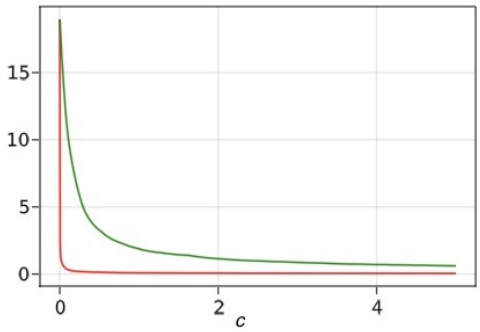


Figure 4.6: Squared error of the marginal inclusion probabilities (Equation 4.13) $c \to \mathcal{E}_{\text{zig-zag}}(c)$ (red) and $t \to \mathcal{E}_{\text{gibbs}}(c)$(green) where c represent the computational time in seconds; right-panel: zoom-in near 0. Here the dimension of the problem is $n^2 = 400$.

0 and are mutually independent from the block $\{x_{i,2}\}_{i=0,1\ldots,n}$ and therefore acting as 'background' noise under the posterior (similarly to the black pixel in Section 4.4.2). The block $\{x_{i,2}\}_{i=0,1\ldots,n}$ will present bi-modality under the posterior as $t \to \mu^{(2)}(t)$ has 2 flat parts and the corresponding coefficients of each flat part are individually encouraged to concentrate aposteriori at the horizontal axis. Moreover, traversing

from one mode to the other mode requires many parameters to be changed from zero to nonzero (and vice versa) simultaneously. The Gibbs sampler proposes a new (sub-)model uniformly at random, i.e., without using information from the data, and has difficulties to explore the posterior in this example. In contrast, this example shows that the Sticky Zig-Zag uses data information in exploring the models and tends to set those variables to zero which are small under the model currently explored (compare to the discussion in Section 4.3).

Numerical experiment: We set $n = 39$ and simulated data $\bar{Y}_{i,j}$ with $\sigma = 0$ (hence $\bar{Y}_{i,j} = \mu^{(j)}(t_i) - \mu^{(j)}(t_{i-1})$) for all $i = 1, 2, \ldots, n$ and $j = 1, 2, 3$, whereas in the likelihood we took $\sigma = 0.6$. The variance of the prior slabs are set to $\sigma_0^2 = 6.0$. The dimension of the problem is $3(n+1) = 120$. Figure 4.7 shows the results for the Sticky Zig-Zag sampler (with $T = 5 \times 10^5$) and the Gibbs sampler (with 7×10^4 iterations) which have been simulated for a similar computing time with burn-in of the first 0.3 segment of the trajectory for the Sticky Zig-Zag (first 0.3 fraction of iterations for the Gibbs sampler). Both algorithms were initialized at $x(0) \sim \mathcal{N}_{3(n+1)}(0, I)$.

4.4.4 Logistic regression

Suppose $\{0, 1\} \ni Y_i \mid x \sim \mathrm{Ber}(\psi(x^T a_i))$ with $\psi(u) = (1 + e^{-u})^{-1}$. $a_i \in \mathbb{R}^d$ denotes a vector of covariates and $x \in \mathbb{R}^d$ a parameter vector. Assume $Y_1, \ldots, Y_N$ are independent, conditionally on x. The log-likelihood is equal to

$$\ell(x) = \sum_{j=1}^{N} \left(\log\left(1 + e^{\langle a_j, x \rangle}\right) - y_j \langle a_j, x \rangle \right)$$

We assume a spike-and-slab prior of (4.1) with zeromean Gaussian slabs and (large) variance σ_0^2. Then the posterior can be written as in Equation (4.2), with Ψ and κ as in Equation (4.3).

Numerical experiments: We consider two categorical features with 30 levels each and 5 continuous features. For each observation, an independent random level of each discrete feature and a random value of the continuous features, $\mathcal{N}(0, 0.1^2)$ is drawn. Let the design matrix $A \in \mathbb{R}^{N \times d}$ be the matrix where the i-th row is the vector a_i. A includes the levels of the discrete features in dummy encoding and the interaction terms between them also in dummy encoding scaled by 0.3 (960 columns), and the continuous features in the final 5 columns. This implies that the dimension of the parameter equals $d = 965$. We then generate $N = 50d = 48250$ observations using as ground truth sparse coefficients obtained by setting $x_i = z_i \xi_i$ where $z_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(0.1)$ and $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 5^2)$, where $\{z_i\}$ and $\{\xi_i\}$ are independent.

We run the sticky ZigZag with subsampling and bounding rates derived in Appendix C.5.1. We chose $w_1 = w_2 = \cdots = w_d = 0.1$ and $\sigma_0^2 = 10^2$ and ran the

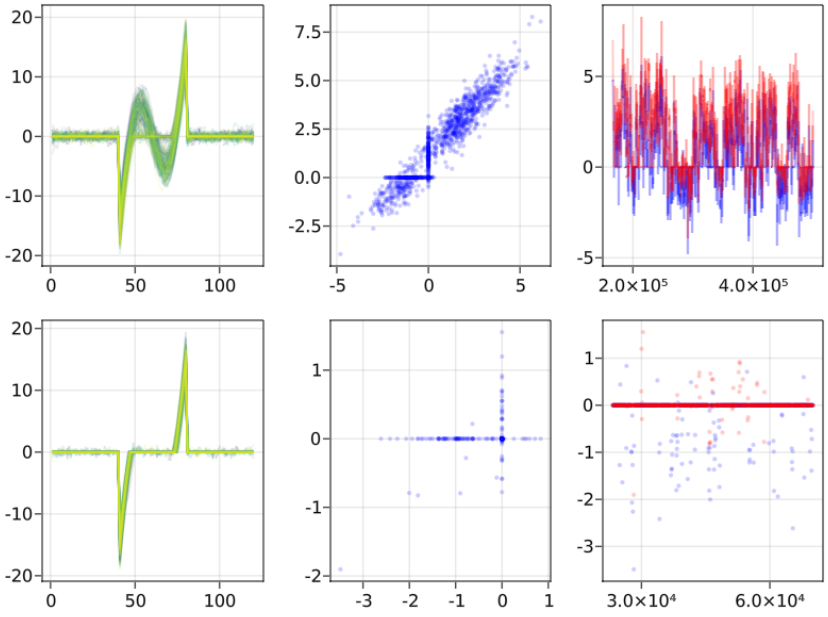


Figure 4.7: Visualization of the output of model with Gaussian increments. Top: Sticky Zig-Zag sampler. Bottom: Gibbs sampler. Left: trace of the vector $(\{x_{i,1}\}_{i=0,1,\ldots,n}, \{x_{i,2}\}_{i=0,1,\ldots,n}, \{x_{i,3}\}_{i=0,1,\ldots,n})$ (the color gradient from yellow to blue indicates the sampling time). Center: (x_{48}-x_{49}) phase portrait (for clarity, we sub-sampled points of the Zig-Zag trajectories at every 250 unit times and sub-sampled points of the Gibbs sampler at every 70 iterations). Right: traces of x_{48}(blue) and x_{49}(red). Here, the dimension of the problem is $3(n+1) = 120$. Here it is clear that, with the given computational budget, the Gibbs sampler fails to mix between different sub-models.

Sticky Zig-Zag sampler for 100 time-units. The implementation makes use of a sparse matrix representation of A, speeding up the computation of inner products $\langle a_j, x\rangle$. Figure 4.8 reveals that while perfect recovery is not obtained (as was to be expected), most nonzero/zero features *are* recovered correctly.

In a second numerical experiment we compare the computing time of the Sticky Zig-Zag sampler and Gibbs sampler (as proposed in Polson, Scott, and Windle 2013) as we vary the number of observations (N). In this case, we reduce the dimension of the parameter by restricting to 2 categorical variables, including their pairwise interactions, augmented by 3 "continuous" predictors (leading to the parameter vector $x \in \mathbb{R}^9$). For each sample size N we ran the Gibbs sampler for 1000 iterations and the Sticky Zig-Zag sampler for 1000 time units. Our interest here is not to compare the computing time of the samplers for a fixed value of N, but rather the scaling of each algorithm with N. Figure 4.9 shows that the computing time for the

Sticky Zig-Zag sampler is roughly constant when varying N. On the contrary, the computing time increases linearly with N for the Gibbs sampler. This is consistent with the theoretical scaling results presented in Table 4.1 (rightmost column). We remark that qualitatively similar results would be obtained if we would have fixed the number of iterations of the Gibbs sampler and endtime of the Zig-Zag sampler to different values.

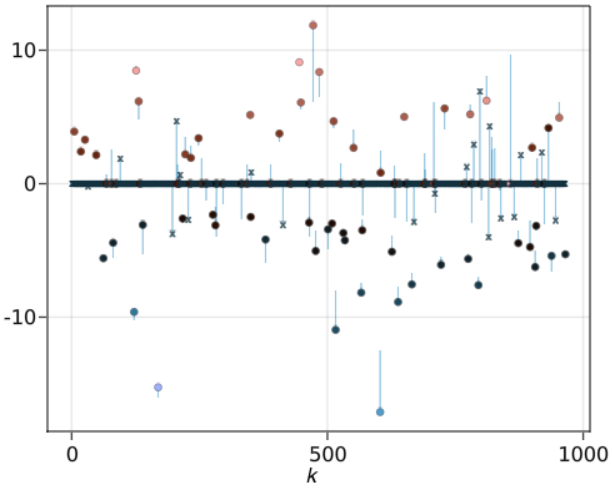


Figure 4.8: Results for the logistic regression coefficients derived with the Sticky Zig-Zag sampler with subsampling. Description as in caption of Figure 4.2. The dimension of this problem is $d = 965$.

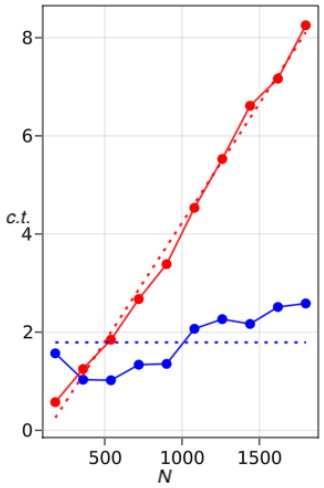


Figure 4.9: Logistic regression example: computing time in seconds versus number of observations. Solid red line: Gibbs samplers with 10^3 iterations. Solid blue line: Sticky Zig-Zag samplers with subsampling ran for 10^3 time units. The dashed lines correspond to the scaling results displayed in Table 4.1. Here, the dimension of the problem is fixed to $d = 9$.

4.4.5 Estimating a sparse precision matrix

Consider

$$Y_i \mid X \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p\left(0, (XX')^{-1}\right), \quad i = 1, 2, \ldots, N$$

for some unknown lower triangular sparse matrix $X \in \mathbb{R}^{p \times p}$. We aim to infer the lower-triangular elements of X which we concatenate to obtain the parameter vector

$x := \{X_{i,j} : 1 \leq j \leq i \leq p\} \in \mathbb{R}^{p(p+1)/2}$. This class of problems is important as the precision matrix XX' unveils the conditional independence structure of Y, see for example Shi, Ghosal, and Martin (2021), and reference therein, for details.

We impose a prior measure on x of the product form $\mu_0(\mathrm{d}x) = \bigotimes_{i=1}^{p} \bigotimes_{j=1}^{i} \mu_{i,j}(\mathrm{d}x_{i,j})$ where

$$\mu_{i,j}(\mathrm{d}x_{i,j}) = \begin{cases} \pi_{i,j}(x_{i,j})\mathbf{1}_{(x_{i,j}>0)}\mathrm{d}x_{i,j} & i = j, \\ w\pi_{i,j}(x_{i,j})\mathrm{d}x_{i,j} + (1-w)\delta_0(\mathrm{d}x_{i,j}) & i \neq j, \end{cases}$$

and $\pi_{i,j}$ is the univariate Gaussian density with mean $c_{i,j} \in \mathbb{R}$ and variance $\sigma_0^2 > 0$.

This prior induces sparsity on the lower-triangular off-diagonal elements of X while preserving strict positive definiteness of XX' (as the elements on the diagonal are restricted to be positive).

The posterior in this example is of the form

$$\mu(\mathrm{d}x) \propto \exp(-\Psi(x)) \Big(\bigotimes_{i=1}^{p} \bigotimes_{j=1}^{i-1} (\mathrm{d}x_{i,j} + \frac{1}{\kappa_{i,j}} \delta_0(\mathrm{d}x_{i,j})) \Big) \bigotimes_{k=1}^{p} \mathrm{d}x_{k,k}$$

with

$$\Psi(x) = \frac{1}{2} \sum_{i=1}^{N} Y_i' XX' Y_i - N \sum_{i=1}^{p} \log(x_{i,i}) + \sum_{i=1}^{p} \sum_{j=1}^{i-1} \frac{(x_{i,j} - c_{i,j})^2}{2\sigma_0^2} + \sum_{i=1}^{p} \frac{(x_{i,i} - c_{i,i})^2}{2\sigma_0^2}$$

and $\kappa_{i,j} = \pi_{i,j}(0)w/(1-w)$. In particular, the posterior density is not of the form as given in Equation (4.2), as the diagonal elements cannot be zero and have a marginal density relative to the Lebesgue measure, while the off-diagonal elements are marginally mixtures of a Dirac and a continuous component. Notice that, for any $i = 1, 2, \ldots, p$, as $x_{i,i} \downarrow 0$, $\exp(-\Psi(x))$ vanishes and $\nabla\Psi(x) \to \infty$. This makes the sampling problem challenging for gradient-based algorithms.

Numerical experiments: We apply the Sticky Zig-Zag sampler where the reflection times are computed by using a thinning and superposition scheme for inhomogeneous Poisson processes, see Appendix C.5.3 for the details.

We simulate realisations $y_1, \ldots, y_N$ with precision matrix XX' a tri-diagonal matrix with diagonal $(0.5, 1, 1, \ldots, 1, 1, 0.5) \in \mathbb{R}^p$ and off-diagonal $(-0.3, -0.3, \ldots, -0.3) \in \mathbb{R}^{p-1}$. In the prior we chose $\sigma_0^2 = 10$ and $c_{i,j} = \mathbf{1}_{(i=j)}$ and for $1 \leq j \leq i \leq p$ and $w = 0.2$.

We fixed $N = 10^3$ and $p = 200$ and ran the Sticky Zig-Zag sampler for 600 time-units. We initialized the algorithm at $x(0) \sim \mathcal{N}_{p(p+1)/2}(0, I)$ and set a burn-in of 10 unit-time. The left panel of Figure 4.10 shows the error between XX' (the ground truth) and $\overline{X}\,\overline{X}'$ where $\overline{X}$ is posterior mean of the lower triangular matrix estimated with the sampler. The error is concentrated on the non-zero elements of the matrix while the zero elements are estimated with essentially no error. The right panel

of Figure 4.10 shows the trajectories of two representative non-zero elements of X. The traces show qualitatively that the process converges quickly to its stationary measure. In this case, comparisons with the Gibbs sampler are not possible as there is no closed form expression for the Bayes factors of Equation (4.12).

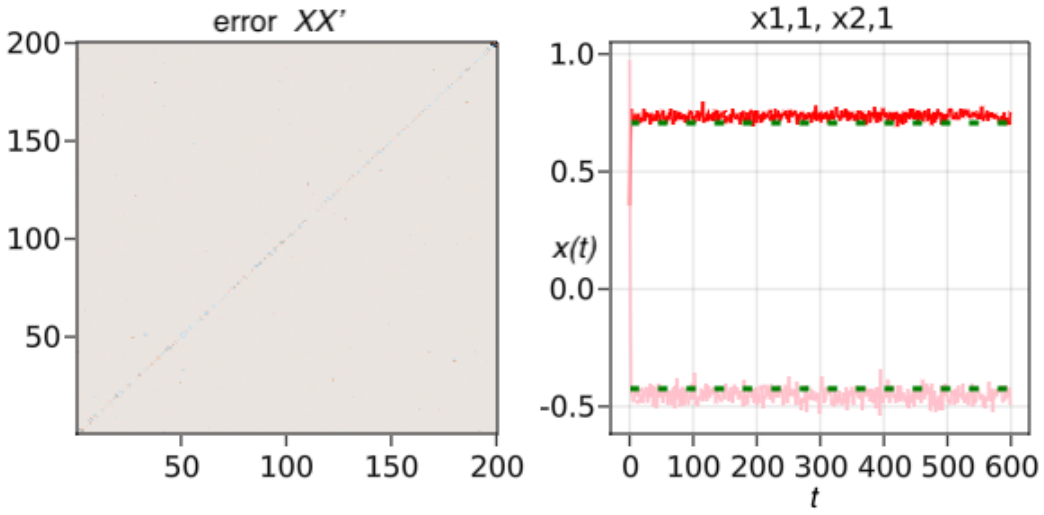


Figure 4.10: Left: error between the true precision matrix and the precision matrix obtained with the estimated posterior mean of the lower-triangular matrix (colour gradient between white (no error) and black (maximum error)). Right: traces of two non-zero coefficients ($x_{1,1}$ in red and $x_{2,1}$ in pink) of the lower triangular matrix. Dashed green lines are the ground truth. Here, the dimension of each vector Y_i is $p = 200$ and the dimension of the problem is $p(p+1)/2 = 20\,100$.

4.5 Discussion

The sticky Zig-Zag sampler inherits some limitations from the ordinary Zig-Zag sampler:

Firstly, if it is not possible to simulate the reflection times according to the Poisson rates in Equation (4.8), the user needs to find and specify upper bounds of the Poisson rates from which it is possible to simulate the first event time (see Appendix C.4.2 for details). This procedure is referred to as *thinning* and remains the main challenge when simulating the Zig-Zag sampler. Furthermore, the efficiency of the algorithm deteriorates if the upper bounds are not tight.

Secondly, the Sticky Zig-Zag sampler, due to its continuous dynamics, can experience difficulty traversing regions of low density, in particular it will have difficulty reaching 0 in a coordinate if that requires passing through such a region.

Finally, the process can set to 0 (and not 0) only one coordinate at a time, hence failing to be ergodic for measures not supported on neighbouring sub-models. For example, consider the space $\mathbb{R}^2$ and assumes that the process can visit either the origin $(0, 0)$ or the full space $\mathbb{R}^2$ but not the coordinate axes $\{0\} \times \mathbb{R} \cup \mathbb{R} \times \{0\}$. Then the process started in $\mathbb{R}^2$ hits the origin with probability 0, hence failing to explore the subspace $(0, 0)$.

In what follows, we outline promising research directions deferred to future work.

4.5.1 Sticky Hamiltonian Monte Carlo

The ordinary Hamiltonian Monte Carlo (HMC) process as presented by Neal et al. (2011) can be seen as a piecewise deterministic Markov processes with deterministic dynamics equal to

$$\dot{x} = v, \qquad \dot{v} = -\nabla\Psi(x) \tag{4.16}$$

where $\nabla\Psi$ is the gradient of the negated log-density relative to the Lebesgue measure. At random exponential times with constant rate, the velocity component is refreshed as $v \sim \mathcal{N}(0, I)$ (similarly to the refreshment events in the bouncy particle sampler). By applying the same principles outlined in Section 4.2, such process can be made sticky with Equation (4.2) as its stationary measure.

Unfortunately, in most cases, the dynamics in (4.16) cannot be integrated analytically so that a sophisticated numerical integrator is usually employed and a Metropolis-Hasting steps compensates for the bias of the numerical integrator (see Neal et al. 2011 for details). These two last steps makes the process effectively a discrete-time process and its generalization with sticky dynamics is not anymore trivial.

4.5.2 Extensions

The setting considered in this work does not incorporate some relevant classes of measures:

- Posteriors given by prior measures which freely choose prior weights for each (sub-)model. This limitation is mainly imputed to the parameter $\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_d)$ which here does not depend on the location component x of the state space. While the theoretical framework built can be easily adapted for letting κ depend on x, it is currently unclear to us the exact relationship between κ and the posterior measure in this more general setting.

- Measures which are not supported on neighbouring sub-models are also not covered here. To solve this problem, different dynamics for the process should be developed which allow the process to jump in space and set multiple coordinates to 0 (and not 0) at a time.

Chapter 5

Methods and applications of PDMP samplers with boundary conditions

5.1 Overview

5.1.1 Introduction

Markov Chain Monte Carlo (MCMC) methods are central tools used to derive asymptotically exact posterior measures in Bayesian inference and consist of simulating Markov chains which converge to the desired target measure. The performance of the method is determined by the convergence property of the chain and the computational cost of simulating it. Recent attention has been drawn by Monte Carlo methods based on continuous-time piecewise deterministic Markov processes (PDMP samplers), see Vanetti et al. (2017) for an overview. PDMP samplers are non-reversible Markov processes endowed with *momentum* and characterized by deterministic dynamics and a collection of random events. This results in a class of processes which have good mixing (fast convergence to the target measure, see for example Diaconis, Holmes, and Neal 2000), have lower asymptotic variance (see for example Chen and Hwang 2013), and can be simulated exactly in continuous time (up to floating point precision). Another attractive feature of PDMP samplers is that they allow to substitute the gradient of the target log-density with an unbiased estimate of it without introducing bias. This technique is refereed as *exact subsampling* and leads to efficient simulations in difficult scenarios e.g. it has been exploited in regression problems with large sample size (Bierkens, Fearnhead, and Roberts 2019, Bierkens et al. 2020, Bierkens et al. 2023) or for high dimensional problems with intractable densities (Bierkens et al. 2021).

5.1.2 Contribution

In this work

- We give a simple sufficient condition at the boundary in terms of skew-detailed balance for PDMP samplers to target efficiently measures with densities which are discontinuous over the boundary and measures supported on a constrained space. The framework considered allows the process to speed-up and to jump in space upon crossing a boundary. We demonstrate that every coordinate can be endowed by sticky points for targeting measures which are also mixtures of continuous and atomic components.

- We apply PDMP samplers with boundary conditions in two examples: for sampling the latent space of infected times with unknown infected population size in the SIR model with notifications and for sampling the invariant measure in hard-sphere models.

5.1.3 Related literature

Our approach generalizes Bierkens et al. (2018) which defines the Bouncy Particle Sampler for target densities on restricted domains. The use of speed-up functions was initially considered in Vasdekis and Roberts (2021) for sampling efficiently heavy tailed distributions with PDMP samplers. We extend speed-up functions by considering piecewise discontinuous functions. We combine our framework with the sticky events presented in Bierkens et al. 2023 for targeting measures which are also mixture of continuous and atomic component. The framework considered allows to define a boundary which acts as teleportation portals, allowing the process to jump in space along the given boundary. The teleportation portals defined in this work are reminiscent to the approach presented in Moriarty, Vogrinc, and Zocca (2020) for sampling measures in a non-convex and disconnected space with a Metropolis-Hastings algorithm. Another approach based on Hamiltonian Monte Carlo (HMC) sampler for sampling discontinuous densities is given by Nishimura, Dunson, and Lu (2020). In contrast with HMC, the PDMPs presented here can be simulated in continuous time, without relying on discretization methods.

A relevant related work is given by Chevallier et al. (2021). In this work, PDMPs for piecewise-smooth densities are presented together with a detailed and rigorous theory for PDMPs with boundary conditions. Our work and this paper complement each other as our focus is on applications of these methods.

5.2 PDMP samplers with boundaries

In this section, we present an overview of PDMP samplers with boundary conditions, we review the known sufficient conditions which allow the process to target the smooth components of the density and establish general conditions at the boundary based on *skew-detailed balance* to target the discontinuity of the density at the boundary and to define teleportation portals.

We consider the problem of sampling from a measure μ on $\mathbb{R}^d = \sqcup_{i\in K}\Omega_i$ where $\Omega_i, i \in K$ are disjoint subsets of $\mathbb{R}^d$ indexed by a countable set K and μ is of the form

$$\mu(\mathrm{d}x) = C\exp(-\Psi(x))\mathrm{d}x \tag{5.1}$$

for some constant of proportionality C and a function $\Psi(x)$ which is assumed to be differentiable on the interior of each Ω_i, $i \in K$. For some region $A \subset \mathbb{R}^d$, $\Psi(x)$ can be infinity for $x \in A$, corresponding to a region which is not supported by μ.

We assume that each boundary $\partial\Omega_i, i \in K$ is a $(d-1)$-dimensional piecewise-smooth manifold. For every point $x \in \partial\Omega_j$, $j \in K$, we denote $\Psi(x,j) = \lim_{t\downarrow 0}\Psi(\gamma_j(t))$ where $t \to \gamma_j(t)$ is any curve in Ω_j with $\gamma_j(0) = x$ and assume that these limits exist and do not depend on the curve γ_j. We denote the corresponding outward normal vector by $n(x,j)$. With this setting, for $x \in \{\partial\Omega_j \cap \partial\Omega_i,\, i,j \in K\}$ we have that $n(x,i) = -n(x,j)$ and if Ψ is discontinuous at that point, then $\Psi(x,i) \neq \Psi(x,j)$.

Next, we build PDMP samplers which can target μ.

5.2.1 Building blocks of PDMPs with boundaries

Denote the space of possible velocities of the process by $\mathcal{V} \subset \mathbb{R}^d$ and, for all $i \in K$, the augmented spaces $E^\circ = (\bigsqcup_{i\in K}\Omega_i \setminus \partial\Omega_i) \times \mathcal{V}$ and $\partial E = (\bigcup_{i\in K}\partial\Omega_i) \times \mathcal{V}$, with elements given by the tuple $z = (x,v)$. PDMPs takes values in the state space

$$E = \left(E^\circ \sqcup \partial E^-\right)$$

with

$$\partial E^- = \{(x,v) \in \partial E\colon \langle v, n(x)\rangle < 0\} \tag{5.2}$$

and boundary

$$\partial E^+ = \{(x,v) \in \partial E\colon \langle v, n(x)\rangle > 0\}. \tag{5.3}$$

Note that ∂E^+ is the set of position and velocity for which the process reaches the boundary, while the elements of ∂E^- are those for which the process leaves the boundary.

PDMPs are characterized by a finite collection of random events and deterministic dynamics in between those events as follows:

- Let $\phi : E \times \mathbb{R} \to E$ be the deterministic flow of the process with differential form
$$\frac{\mathrm{d}\phi(z_0, t)}{\mathrm{d}t} = (vs(\phi_x(z_0, t)), 0), \quad \phi(z_0, 0) = z_0 \tag{5.4}$$
with ϕ_x being the position component of ϕ and for a speed-up function s.

  The dynamics in equation (5.4) generate straight lines in the position coordinate with speed proportional to the function $s\colon \mathbb{R}^d \to \mathbb{R}^+$. Generalization to dynamics other than straight lines are possible, for example one might consider PDMPs with Hamiltonian dynamics invariant to Gaussian measures (Bierkens et al. 2020).

- A collection of random event times $\tau_1, \tau_2, \dots$ determines the times where the process changes velocity component. These are computed recursively and are determined by a rate function $\lambda\colon E \to \mathbb{R}^+$. The first event time τ_1 of the process starting at $z_0 \in E$ coincides with the first event time of an inhomogeneous Poisson process with rate $t \to \lambda(\phi(z_0, t))$ and therefore satisfies
$$\mathbb{P}(\tau_1 > t) = \exp\left(-\int_0^t \lambda(\phi(z_0, s))\mathrm{d}s\right). \tag{5.5}$$
  Hereafter, we write $\tau_1 \sim \mathrm{IPP}(t \to \lambda(\phi(z_0, t)))$.

- Two kernels $\mathcal{Q}_{\partial E^+}, \mathcal{Q}_E$ determine the behaviour of the process respectively when approaching the boundary and at random event times. The two kernels are combined in a Markov kernel $\mathcal{Q}\colon E \sqcup \partial E^+ \times \mathcal{B}(E) \to [0, 1]$ defined as
$$\mathcal{Q}(z, \cdot) = \begin{cases} \mathcal{Q}_E(z, \cdot) & z \in E \\ \mathcal{Q}_{\partial E^+}(z, \cdot) & z \in \partial E^+. \end{cases}$$

Similar to Vasdekis and Roberts (2021), the speed-up function $s(x)$ is allowed to generate exploding dynamics, that is dynamics for which the process escape to infinity in finite time. This will not be problematic, as long as we make sure that a random event switches velocity before the process escapes to infinity. This condition is reflected by the following assumption:

Assumption 5.2.1. *(speed growth condition, Vasdekis and Roberts 2021, Assumption 3.1)* $\lim_{\|x\|\to\infty} \exp(-\Psi(x))s(x) = 0$.

We assume a speed-up function $s(x)$ of the form
$$s(x) = s_{\mathrm{c}}(x)s_{\mathrm{j}}(x)$$
where $s_{\mathrm{c}}(x)$ is continusly differentiable while $s_{\mathrm{j}}(x)$ is piecewise constant with jumps at the boundaries $\partial\Omega$.

A necessary condition for PDMPs to target the measure $\mu \otimes \rho$, where ρ is the marginal invariant measure of the velocity component while μ is the target measure we are interested to sample from, is that, for functions in

$$\mathcal{A} = \{f \in C_c(E); \quad t \to f(\phi(z,t)) \text{ is absolutely continuous } \forall z \in E; \\ f(z) = \int_{\partial E^-} f(z')\mathcal{Q}_{\partial E^+}(z, \mathrm{d}z'), \quad \forall z \in \partial E^+\} \tag{5.6}$$

the following equality holds

$$\int_E \mathcal{L}f \, \mathrm{d}(\mu \otimes \rho) = 0 \tag{5.7}$$

where $\mathcal{L}$ is the extended generator of the process. For PDMPs, $\mathcal{L}$ is known and its expression is given in Appendix D.1. For a more detailed derivation of the invariant measure of PDMP samplers with boundary conditions, see Chevallier et al. (2021).

Next we impose sufficient conditions for (5.7) to hold. In particular we will see that conditions on different components of PDMPs serve to target different components of the measure in equation (5.1). In particular, in Section 5.2.2, we present standard conditions on $(\lambda, s_c, \mathcal{Q}_{E^\circ})$ imposed for ordinary PDMPs (see for example Vasdekis and Roberts 2021) which allow the process to target the differentiable part of the density, while in Section 5.2.3, new conditions on $(s_j, \mathcal{Q}_{\partial E^+})$ are imposed which allow to define teleportation portals and to target discontinuous densities.

Remark 5.2.1. (Extensions with sticky components) *The methodology can be extended for targeting a measure on $\mathbb{R}^d$ which has a piecewise-smooth density relative to a reference measures which is a mixture of Dirac and Lebesgue components of the form*

$$\prod_{i=1}^{d} \left(\mathrm{d}x_i + \frac{1}{\kappa_i} \delta_{c_i}(\mathrm{d}x_i) \right)$$

for some elements $c_i \in \mathbb{R}, \kappa_i > 0$, $i = 1, 2, \ldots, d$. This is achieved by combining the behaviour at discontinuity with sticky events *which are triggered when each coordinate x_i hits c_i and during which the coordinate x_i sticks at c_i for an exponential time with rate κ_i. After this time, the dynamics of that coordinate are restored. The sticky components are introduced and presented in detail Bierkens et al. (2023).*

5.2.2 Review of sufficient conditions on the interior

For our PDMP samplers the kernel $\mathcal{Q}_E$ acts by only changing the velocity component, while leaving the position unchanged. Hence, with abuse of notation, we let $\mathcal{Q}_E \colon E \times \mathcal{B}(\mathcal{V}) \to [0,1]$ be a kernel acting only on the velocity component. Throughout, we distinguish between two different classes of events: *reflections* and

refreshments, which are defined by rates and kernels $(\lambda_{\mathrm{b}}, \mathcal{Q}_{E,\mathrm{b}})$ and $(\lambda_{\mathrm{r}}, \mathcal{Q}_{E,\mathrm{r}})$, respectively. The former ensures that the process targets the right measure while the latter ensures ergodicity of the process. Then, for $z \in E$, $\lambda(z) = \lambda_{\mathrm{r}}(z) + \lambda_{\mathrm{b}}(z)$ and

$$\mathcal{Q}_E(z, \cdot) = \frac{\lambda_{\mathrm{r}}(z)}{\lambda_{\mathrm{r}}(z) + \lambda_{\mathrm{b}}(z)} \mathcal{Q}_{E,\mathrm{r}}(z, \cdot) + \frac{\lambda_{\mathrm{b}}(z)}{\lambda_{\mathrm{r}}(z) + \lambda_{\mathrm{b}}(z)} \mathcal{Q}_{E,\mathrm{b}}(z, \cdot).$$

Next, we makes assumptions on both the refreshment and the reflection events. Recall that the desired target measure of the process takes the form $C \exp(-\Psi(x))\mathrm{d}x\rho(\mathrm{d}v)$, for some constant of normalization C.

Assumption 5.2.2. *(Conditions on refreshments) Let $\lambda_{\mathrm{r}}(x, v) \geq 0$ be a positive function which does not depend on its second argument. Furthermore, let $\mathcal{Q}_{\mathrm{E,r}}$ be invariant to ρ, i.e.*

$$\int_{v \in \mathcal{V}} \rho(\mathrm{d}v) \mathcal{Q}_{E,\mathrm{r}}((x, v), \mathrm{d}v') = \rho(\mathrm{d}v'), \quad \forall x \in \{y \in \mathbb{R}^d \colon (y, v) \in E\}.$$

It is customary for PDMP samplers to set $\mathcal{Q}_{E,\mathrm{r}}((x, v), \cdot) = \rho(\cdot)$ and a fixed rate $\lambda_{\mathrm{r}}(x, v) = c \geq 0$.

Assumption 5.2.3. *(Conditions on reflections) For all $(x, v) \in E$ and for a PDMP with continuous speed-up function $s_c(x)$, let $\lambda_{\mathrm{b}} \colon E \to \mathbb{R}^+$ satisfy*

$$\lambda_{\mathrm{b}}(x, v) - \lambda_{\mathrm{b}}(x, -v) = \langle v, A(x) \rangle$$

with

$$A(x) = s_c(x) \nabla \Psi(x) - \nabla s_c(x).$$

Let $\mathcal{Q}_{E,\mathrm{b}}$ satisfy

$$\int_{v \in \mathcal{V}} \rho(\mathrm{d}v) \lambda(x, v) \mathcal{Q}_{E,\mathrm{b}}((x, v), \mathrm{d}v') = \lambda(x, -v') \rho(\mathrm{d}v'), \quad \forall x \in \{y \in \mathbb{R}^d \colon (y, v) \in E\}. \tag{5.8}$$

The next proposition shows that if we make the assumptions above and compute the left-hand side of equation (5.7), we are left with an integral over ∂E.

Proposition 5.2.2. *Consider a PDMP sampler satisfying Assumption 5.2.2-5.2.3. Then*

$$\int_E \mathcal{L} f \mathrm{d}(\mu \otimes \rho) = \int_{(x,v) \in \partial E} f(x, v) \langle n(x), v \rangle s_j(x) \mu(\mathrm{d}x) \rho(\mathrm{d}v), \quad f \in \mathcal{A}, \tag{5.9}$$

where $\mathcal{L}$ is the extended generator of the process and $\mathcal{A}$ is given in (5.6).

Proof. See Appendix D.1.1. □

Next, we will show that the right hand-side of equation (5.9) can be made equal to 0 by detailing the behaviour of PDMPs at the boundary ∂E^+.

5.2.3 Sufficient conditions at the boundary

Here, we state a fairly general condition for the kernel $\mathcal{Q}_{\partial E^+}$ which guarantees that the process is invariant at the boundary to the target density given in equation (5.1).

Recall that a map $\mathcal{S}\colon \mathcal{X} \to \mathcal{X}$ is an involution if $\mathcal{S} \circ \mathcal{S} = I$, where I stands for the identity map. We now give an important definition, followed by the main assumption made at the boundary:

Definition 5.2.4. (Skew detailed balance condition) *For an involution $\mathcal{S} : \mathcal{X} \to \mathcal{X}$, a kernel $\mathcal{Q}\colon (\mathcal{X}, \mathcal{B}(\mathcal{X})) \to [0,1]$ satisfies the skew detailed balance condition relative to a measure μ on $\mathcal{X}$, if*

$$\mathcal{Q}(z, \mathrm{d}z')\mu(\mathrm{d}z) = \mathcal{Q}(\mathcal{S}^{-1}(z'), \mathcal{S}^{-1}(\mathrm{d}z))\mu(\mathcal{S}^{-1}(\mathrm{d}z')).$$

Assumption 5.2.5. *Let $\mathcal{Q}_{\partial E^+} : \partial E^+ \times \mathcal{B}(\partial E^-) \to [0,1]$ satisfy the skew detailed balance condition relative to the signed measure $\nu(\mathrm{d}x, \mathrm{d}v) = \langle n(x), v\rangle s_j(x)\mu(\mathrm{d}x)\rho(\mathrm{d}v)$ defined on ∂E with involution $\mathcal{S}(x,v) = (x,-v)$.*

Notice that, contrary to the ordinary application of skew detailed balance, ν is a signed measure and $\nu(\mathrm{d}z) = -\nu(\mathcal{S}^{-1}(\mathrm{d}z))$.

Proposition 5.2.3. *Consider a PDMP sampler satisfying Assumption 5.2.2-5.2.3-5.2.5. Then $\mu \otimes \rho$ is a stationary measure of the process.*

Proof. If $\mu \otimes \rho$ is a stationary measure of the PDMP, then we must have that $\int \mathcal{L}f \mathrm{d}(\mu \otimes \rho) = 0$, $f \in \mathcal{A}$. By Proposition 5.2.2 and the boundary condition in (5.6), we have that

$$\begin{aligned}
\int_E \mathcal{L}f \mathrm{d}(\mu \otimes \rho) &= \int_{z \in \partial E^+} \int_{p \in \partial E^-} \mathcal{Q}_{\partial E}(z, \mathrm{d}p) f(p) \nu(\mathrm{d}z) \\
&\quad + \int_{\partial E^-} f(z)\nu(\mathrm{d}z) \\
&= \int_{p \in \partial E-} f(p) \int_{z \in \partial E^+} \mathcal{Q}_{\partial E^+}(z, \mathrm{d}p)\nu(\mathrm{d}z) \qquad (5.10) \\
&\quad + \int_{\partial E^-} f(z)\nu(\mathrm{d}z) \\
&= \int_{\partial E^-} f(p)\nu(\mathcal{S}(\mathrm{d}p)) + \int_{\partial E^-} f(z)\nu(\mathrm{d}z) = 0 \qquad (5.11)
\end{aligned}$$

where in (5.10) we applied Fubini for interchanging integrals and in (5.11) we used Assumption 5.2.5. □

For simplicity, in this article, we focus on specific transition kernels at the boundary which satisfy Assumption 5.2.5 and will be used for the two main applications in Section 5.3. The transition kernels considered take the form

$$\begin{aligned}\mathcal{Q}_{\partial E^+}((x,v),\mathrm{d}(y,w)) = \alpha(x,y)\mathcal{T}(x,\mathrm{d}y)\mathcal{R}_1((x,v),y,\mathrm{d}w) \\ + \left(1-\int \alpha(x,y)\mathcal{T}(x,\mathrm{d}y)\right)\delta_x(\mathrm{d}y)\mathcal{R}_2((x,v),\mathrm{d}w).\end{aligned} \tag{5.12}$$

In this setting, the process, upon hitting a boundary $(x,v)\in\partial E^+$, jumps to (y,w) with $y\sim\mathcal{T}(x,\cdot)$ and $w\sim\mathcal{R}_1((x,v),y,\cdot)$ with probability $\alpha(x,y)$ and reflects at the boundary otherwise, by setting a new velocity according to $q\sim\mathcal{R}_2((x,v),\cdot)$, see Algorithm 6 for its implementation. We now specify in details each term in equation (5.12).

The kernel $\mathcal{T}\colon \partial\Omega\times\mathcal{B}(\partial\Omega)\to[0,1]$ acts on the boundary of the position component of the process and such that for every $x\in\partial\Omega$, $\mathcal{T}(x,\dot{})$ is absolutely continuous with respect to $\mathcal{T}(y,\cdot)$, almost surely for every $y\sim\mathcal{T}(x,\cdot)$. $\alpha(x,y)=\min(1,R(x,y))$ and $R(\cdot,\cdot)$ is the Radon-Nikodym derivative on the product space $(\partial\Omega\times\partial\Omega,\mathcal{B}(\partial\Omega\times\partial\Omega))$ defined as

$$R(x,y)=\frac{\nu(\mathrm{d}y)\mathcal{T}(y,\mathrm{d}x)}{\nu(\mathrm{d}x)\mathcal{T}(x,\mathrm{d}y)} \tag{5.13}$$

where $\nu(\mathrm{d}x)=\exp(-\Psi(x))s_j(x)\mathrm{d}x$. Finally, for every $(x,v)\in\partial E^+$ and $y\in\mathcal{T}(x,\cdot)$, we define two kernels acting on the velocity components: $\mathcal{R}_1\colon\{(x,v)\in\partial E^+\}\times\{y\in\partial\Omega\}\times\mathcal{B}(\partial\mathcal{V}^-(y))\to[0,1]$ and $\mathcal{R}_2\colon\{(x,v)\in\partial E^+\}\times\mathcal{B}(\partial\mathcal{V}^-(x))\to[0,1]$, with $\partial\mathcal{V}^-(x)=\{v\in\mathcal{V}\colon(x,v)\in\partial E^-\}$, which satisfy

$$\langle n(x),v\rangle=-\langle n(y),W\rangle,\qquad \text{a.s. for } W\sim\mathcal{R}_1((x,v),y,\cdot)$$

and

$$\langle n(x),v\rangle=-\langle n(x),W\rangle,\qquad \text{a.s. for } W\sim\mathcal{R}_2((x,v),\cdot).$$

Remark 5.2.4. *The jump component of the speed-up function s_j can be tuned in order to reduce the probability to reflect at the boundary (see equation (5.13)) and in some cases can be chosen to completely off-set that probability so that the process always crosses the boundary. In this case, there is no need to specify a reflection rule of the velocity at the boundary. This is key whenever there is no good choice to reflect the velocity at the boundary rather than flipping completely the velocity vector, hence generating undesirable* back-tracking *effects of the underlying Markov process.*

Next, we give two concrete examples of PDMP samplers with boundaries that will be used in the applications of Section 5.3.

Algorithm 6 Behaviour of PDMPs at the Boundary

For $(x, v) \in \partial E^+$:

- Propose a point $y \in \partial\Omega$ as $y \sim \mathcal{T}(x, \cdot)$.
- Simulate $u \sim \text{Unif}([0, 1])$.
- If $\alpha(x, y) > u$, set the new state to $(x', v') = (y, w)$ with $w \sim \mathcal{R}_1((x, v), y, \cdot)$.
- Otherwise set the state $(x', v') = (x, q)$ with $q \sim \mathcal{R}_2((x, v), \cdot)$.
- return the new state (x', v').

5.2.4 Example: d-dimensional Zig-Zag sampler for discontinuous densities

The d-dimensional Zig-Zag sampler is a PDMP sampler defined in the augmented space of position and velocity $\mathbb{R}^d \times \{-1, +1\}^d$. For the process in the state $z \in E$, the first random reflection time is given by $\tau = \min(\tau_1, \tau_2, \ldots, \tau_d)$ where for $i = 1, 2, \ldots, d$, $\tau_i \sim \text{IPP}(t \to \lambda_{i,\text{b}}(\phi(z, t)))$ and

$$\lambda_{i,\text{b}}(x, v) = \max(v_i(s_c(x)\partial_{x_i}\Psi(x) - \partial_{x_i}s_c(x)), 0). \tag{5.14}$$

At random time τ, the velocity component changes as $v \to v[k; -v_k]$ where $k = \arg\min(\tau_1, \tau_2, \ldots, \tau_d)$. For more details on the standard Zig-Zag sampler see Bierkens et al. (2018) and Vasdekis and Roberts (2021).

Here we extend the process for target densities which are discontinuous at the boundary as given by equation (5.1). The invariant measure of the process is then $\mu(\mathrm{d}x) \times \text{Unif}(\{-1, +1\}^d)$. We assume that for every $(x, i) \in \partial\Omega_i, i \in K$ there is a unique $j \in K$ such that the point $(x, j) \in \partial\Omega_j$. This assumption is similar to Chevallier et al. (2021, Assumption 1, (iii)).

We link those points by the function $\kappa\colon \partial\Omega \to \partial\Omega$ with $\kappa(x, i) = (x, j)$. We set $\mathcal{T}((x, i), \cdot) = \delta_{\kappa(x,i)}(\cdot)$. As in this case $n(x, i) = -n(\kappa(x, i))$ we set

$$\mathcal{R}_1((x, i, v), \kappa(x, i), \mathrm{d}w) = \delta_v(\mathrm{d}w)$$

and $\mathcal{R}_2((x, i, v), \cdot) = \delta_{v'}(\cdot)$ with

$$v'_\ell = \begin{cases} -v_\ell & n_\ell(x) \neq 0, \\ v_\ell & n_\ell(x) = 0. \end{cases}$$

The process, upon hitting a boundary $(x, i) \in \partial\Omega$, crosses the boundary with probability $\alpha((x, i), \kappa(x, i))$ without changing the velocity component and reflects otherwise by switching the sign of only the components which are not orthogonal to $n(x, i)$.

5.2.5 Example: d-dimensional Bouncy Particle Sampler with teleportation on constrained spaces

The d-dimensional Bouncy Particle Sampler (BPS) is defined in the space of position and velocity $\mathbb{R}^d \times \mathbb{R}^d$. The velocity component has a marginal invariant measure equal to $\rho(\cdot) = \mathcal{N}_d(0, I)$. For an initial state $z \in E$, the first random reflection time is distributed as $\tau \sim \mathrm{IPP}(t \to \lambda_{\mathrm{b}}(\phi(z,t)))$ with

$$\lambda_{\mathrm{b}}(x,v) = \max(\langle v, \nabla\Psi(x)\rangle s_c(x) - \langle v, \nabla s_c(x)\rangle, 0).$$

At reflection time, the process changes velocity according to the kernel $\mathcal{Q}_{E,\mathrm{b}}((x,v),\cdot) = \delta_{R_\Psi(x,v)}$ for

$$R_\Psi(x,v) = v - 2\frac{\langle v, \Psi(x)\rangle}{\|\Psi(x)\|}\Psi(x).$$

At random exponential times with rate $\lambda_{\mathrm{r}}(x,v) = c > 0$, the process refreshes its velocity by drawing a new velocity $v' \sim \rho$. See Bouchard-Côté, Vollmer, and Doucet (2018) for an overview of the standard BPS. In this example, we extend the BPS for targets of the form

$$\mu(\mathrm{d}x) = \exp(-\Psi(x))\mathbf{1}_A \mathrm{d}x \tag{5.15}$$

for a set A with a $(d-1)$ piecewise-smooth boundary ∂A and a function $\Psi \in \mathcal{C}^1(\mathbb{R}^d)$. This corresponds to a smooth target density on a constrained space given by the set A.

While the standard BPS for constrained spaces as presented in Bierkens et al. (2018) would reflect the velocity every time the process reaches the boundary ∂A, in our setting, the kernel $\mathcal{T}\colon \partial\Omega \times \mathcal{B}(\partial\Omega) \to [0,1]$, for some $\partial\Omega \supseteq \partial A$ allows the process to jump in space when hitting ∂A, effectively creating teleportation portals. To that end, we set $\mathcal{R}_2(z,\cdot) = \delta_{R_n(z)}(\cdot)$ with

$$R_n(x,v) = v - 2\frac{\langle v, n(x)\rangle}{\|n(x)\|}n(x) \tag{5.16}$$

while $\mathcal{R}_1((x,v),y,\cdot) = \delta_{v'}(\cdot)$ with $v'(x,y,v) = U(n(x), -n(y))v$. Here, for any $v \in \mathbb{R}^d$, $U\colon \mathbb{R}^d \times \mathbb{R}^d \to SO(d)$ is a measurable function taking values in the rotation group $SO(d) := \{U \in \mathbb{R}^{d\times d} : U'U = 1, \det(U) = 1\}$, such that $|U(q,w)v| = |v|$ and $\langle v, q\rangle = \langle U(q,w)v, w\rangle$, for every $q, w \in \mathbb{R}^d$.

This framework allows the process to make jumps in $\partial\Omega$ and to visit disconnected regions or distant regions which are difficult to reach with continuous paths. The problem of sampling from a conditional measure as in equation (5.15) arises for example in the simulation of extreme events. In this case, standard Markov Chain Monte Carlo methods can fail to explore the full measure because of the inability of the chain to traverse subsets of measure (close to) 0.

5.3 Applications

In this section we motivate our work by applying the PDMPs with boundaries as described in Section 5.2 for sampling the latent space of infection times in the SIR model with notifications and for sampling the invariant measure of hard-sphere problems.

5.3.1 SIR model with notifications

The model presented here is inspired by the setting established in Jewell et al. (2009) for modelling the spread of infectious disease in a population. Here we combine the PDMP for piecewise smooth densities with the framework presented in Bierkens et al. (2023) for adding/removing efficiently in continuous time occult infected individuals (infected individuals which have not been notified up to the observation time) by means of introducing sticky events which are events after which the process sticks to lower dimensional hyper-planes for some random time.

Consider an infection process $\{Y(t) \in \{S, I, N, R\}^d \colon 0 < t < T\}$ on a population of size d. Each coordinate $Y_i(t)$ takes values

$$y_i(t) = \begin{cases} S & \text{if } i \text{ is } \textit{susceptible} \text{ at time } t, \\ I & \text{if } i \text{ is } \textit{infected} \text{ at time } t, \\ N & \text{if } i \text{ is } \textit{notified} \text{ at time } t, \\ R & \text{if } i \text{ is } \textit{removed} \text{ at time } t. \end{cases}$$

Each coordinate-process $(Y_i(t))_{t>0}$ is allowed to change state in the following direction: $S \to I \to N \to R$.

For every individuals on $i = 1, 2, \dots, d$, define

$$\tau_i = \inf\{t > 0 \colon y_i(t) = I\}, \quad \tau_i^\star = \inf\{t > 0 \colon y_i(t) = N\}, \quad \tau_i^\circ = \inf\{t > 0 \colon y_i(t) = R\}$$

respectively for the first infection time, notification time, removing time of individual i with the convention that $\tau_i^\star(\tau_i^\circ) = \infty$ if $\tau_i^\star(\tau_i^\circ) \geq T$. We observe $\tau^\star$ (the notification times) and τ° (the removing times) and we are interested on recovering the minimum between the infection times and the observation time T: $x := (x_i = \tau_i \wedge T \colon i = 1, 2, \dots, d)$. We this convention, $x_i = T$ when an individual i has not been infected before observation time T, hence is susceptible at observation time.

For every $i = 1, 2, \dots, d$, individual i changes its state from S to I according to an inhomogeneous Poisson process with rate $t \to \beta_i(y(t))$ for a function $\beta_i \colon \{S, I, R, N\}^d \to \mathbb{R}^+$ usually referred as the *infectious pressure* on i. As $\{y(t) \colon 0 \leq t \leq T\}$ can be recovered by knowing $(x, \tau^\star, \tau^\circ)$, we write $\beta_i(x) := \beta_i(y(x_i))$ (with this notation, we omit the dependence of $\beta_i(x)$ on $(\tau^\star, \tau^\circ)$ which are known and

fixed throughout). For each pair (i, j) with $i \neq j$, define the *infection rate*

$$\beta_{i,j}(x) = \begin{cases} C_{i,j}, & x_i < x_j \leq \tau_i^\star \\ \gamma C_{i,j} & \tau_i^\star < x_j \leq \tau_i^\circ \\ 0 & \text{otherwise,} \end{cases}$$

where $\gamma \in (0, 1)$ is the factor of reduction of the infectivity after notification time and $C_{i,j} := d(i, j)\vartheta_i\xi_j$ is a measure of infectivity of i towards j. Here $d\colon \{1, 2, \ldots, d\}^2 \to \mathbb{R}^+$ is an inverse distance metric between two individuals and $\vartheta_i, \xi_i > 0$ are seen respectively as the infectivity and susceptibility baselines of individual i (see Jewell et al. 2009 for more details). The infectious pressure on individual i is given by

$$\beta_j(x) = \sum_{i \neq j} \beta_{i,j}(x).$$

Denote the set of individuals which have been notified before time T by $\mathcal{N}_{T^-} := \{i\colon \tau_i^\star < T\}$ and by $\mathcal{N}_{T^-}^c := \{1, 2, \ldots, d\} \setminus \mathcal{N}_T$ its complementary. We assume that the delay between infection and notification $(\tau_i^\star - \tau_i)$ of individual i is a random variable with density f and distribution F.

The population infection time $x = (x_1, x_2, \ldots, x_d)$ is then distributed according to

$$\mu(\mathrm{d}x) \propto \bigotimes_{i=1}^{d} \rho_i(x)\mu_i(\mathrm{d}x_i). \tag{5.17}$$

where $\rho_i(x)\mu_i(\mathrm{d}x_i)$ can be heuristically interpreted as the distribution of the ith infection time. For $i \in \mathcal{N}_{T^-}^c$,

$$\begin{aligned} \rho_i(x) &= (1 - F(T - x_i))\beta_i(x) \exp\left(-B_i(x)\right), \\ \mu_i(\mathrm{d}x_i) &= \mathbf{1}_{(0 \leq x_i \leq T)}\mathrm{d}x_i + \kappa_i(x)\delta_T(\mathrm{d}x_i) \end{aligned} \tag{5.18}$$

with $\kappa_i(x) = \frac{1}{\beta_i(x)}$ and $B_i(x) = \int_0^T \beta_i(x[i; s])\mathrm{d}s$. See Appendix D.2.1 for the details of the derivation of the measure above. Here the point mass at $x_i = T$ absorbs the event that individual i has not been infected before time T while the remaining part of the density represents the event for the individual i to be infected but not notified before time T (in such case the infected individual i is often referred as *occult*). For $i \in \mathcal{N}_{T^-}$,

$$\begin{aligned} \rho_i(x) &= \beta_i(x) \exp\left(-B_i(x)\right) f(\tau_i^\star - x_i), \\ \mu_i(\mathrm{d}x_i) &= \mathbf{1}_{(0 \leq x_i \leq \tau_i^\star)}\mathrm{d}x_i, \end{aligned} \tag{5.19}$$

see Appendix D.2.1 for details.

We apply the Zig-Zag sampler for discontinuous densities as presented in Section 5.2.4, with no speed-up function ($s(x) = 1$) and with sticky events (Remark 5.2.1) to target the measure μ. Below, we summarize the behaviour of the process:

- The process reflects velocity randomly in space according to the gradient of the continuous component of the density of μ. The reader may find in Appendix D.2.2 the explicit computations of the reflection times.

- For every $i = 1, 2, \ldots, d$, the process hits the boundary when the coordinate x_i hits a element of the vectors $x_{-i}, \tau^\star, \tau^\circ$. This is because $x_i \to \beta_i(x)$ is discontinuous on those points. For the boundary corresponding to two coordinates of x colliding, the process bounces off the discontinuity with some probability by changing the sign of the velocity of both coordinates, while if the discontinuity corresponds to a coordinate of x colliding with a notification/removal time, the process bounces off just by changing the sign of the velocity of that coordinate. Upon hitting a given boundary, the process traverses the discontinuity without changing its velocity with some probability. In particular, each coordinate-process (x_i, v_i) for $i = 1, 2, \ldots, d$ never crosses the boundary $(x_i, v_i) = (\tau_i^\star, +1)$ and $(x_i, v_i) = (0, -1)$.

- Each particle (x_i, v_i) with $i \in \mathcal{N}_T^c$ sticks at T for an exponential time with rate equal to $\beta_i(x)/2$ (the factor $\frac{1}{2}$ originates since the Dirac measure is located at the boundary of the interval), upon hitting T. The sticking time of the particle x_i corresponds to the individual indexed by i being susceptible and not infected.

Numerical experiment

We fix $d = 50$ and set $\xi_1, \xi_2, \ldots, \xi_d, \vartheta_1, \ldots, \vartheta_d$ to be the realization of i.i.d. random variables distributed according to Unif$([0, 1])0.9+0.7$ and set $d(i, j) = 0.4(\mathbf{1}_{(|i-j|\leq 5)})$. We assume that the $\tau_i^\star - \tau_i \sim \text{Exp}(0.3)$.

To generate a synthetic dataset, we simulated forward the model up to time $T = 5.0$, setting $\tau_{25} = 0$, see Figure 5.1 for visualizing the dynamics of each individual. Before time T, 47 individuals have been infected, 28 individual have been notified, 18 individuals have been removed.

We fix $x_{25} = \tau_{25} = 0$ and simulate the $(d-1)$-sticky Zig-Zag sampler with final clock $T^\star = 500$. Figure 5.2 shows the marginal posterior densities and the final segment of the Zig-Zag trajectory relative to coordinates 11, 13, 38, 49. Those individuals have a different status at time T: susceptible, occult (infected but not notified), notified and removed.

5.3.2 Hard-sphere models

As an application for teleportation which is relevant in statistical mechanics, we consider the sampling problem in hard-sphere models. This class of models motivated the first Markov chain Monte Carlo method in the pioneering work of Metropolis

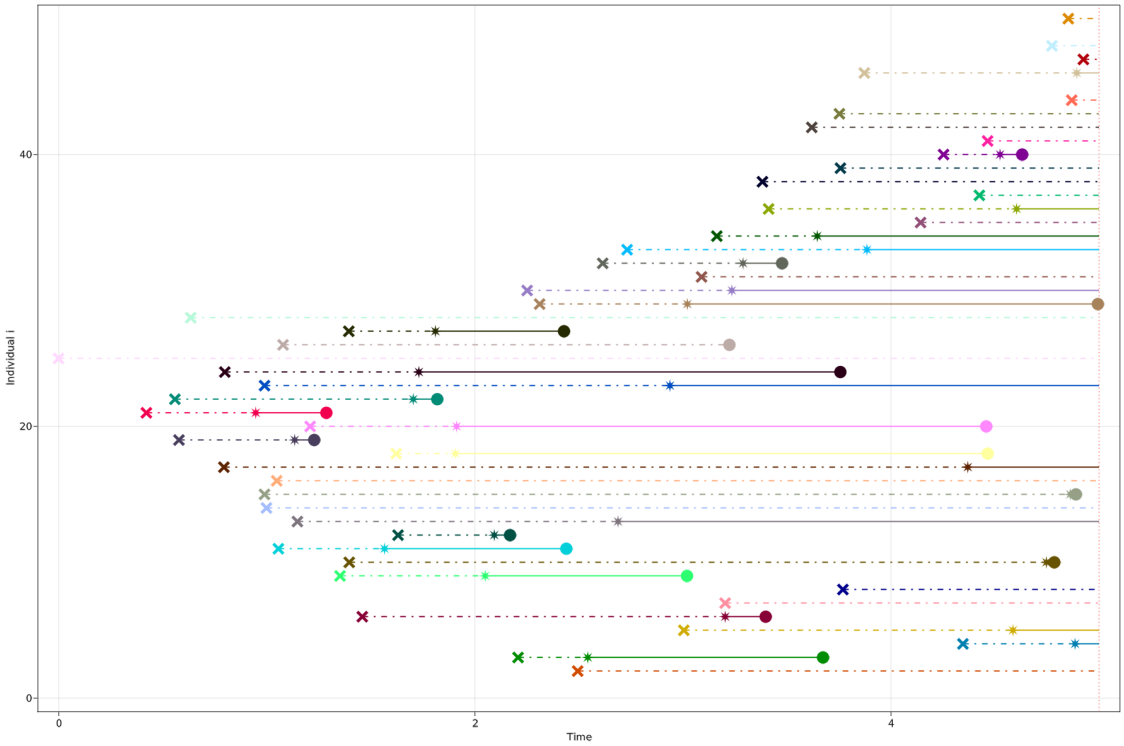


Figure 5.1: Simulated data from the SINR model of the population y-axis (conditional on the infection time of individual 25 to be 0). The symbols $\times, \star, \circ$ indicates respectively the unobserved infection times, the observed notification times and the observed removal times for each individual. The dotted lines indicate the unobserved time between infection and notification times.

et al. (1953). More recently, PDMP samplers have been employed for this class of problems for example in Michel, Tan, and Deng (2019) and Monemvassitis, Guillin, and Michel (2022). For an overview of hard-sphere models, see Krauth (2006, Chapter 2). For a survey of MCMC methods used for sampling from hard-sphere models, see Faulkner and Livingstone (2022). For related models, see Møller, Huber, and Wolpert (2010).

We consider N particles, each one taking values in $\mathbb{R}^d$. Denote the configuration of all particles by $x = \{x^{(i)} \in \mathbb{R}^d \colon 1 \leq i \leq N\}$ where we identify the location of the ith particles by $x^{(i)} = x_{[(i-1)d+1,id]}$. Consider a measure $\mu^\star(\mathrm{d}x) = \exp(-\Psi(x))\mathrm{d}x$, where $\Psi(x) = \sum_{i=1}^N \Psi_0(x^{(i)})$, for a smooth function Ψ_0 supported on $\mathbb{R}^d$. We assume that each particle $i = 1, 2, \ldots, d$ is a hard-sphere centered in $x^{(i)}$ and with radius $r_i > 0$ and consider the conditional invariant measure

$$\mu(\mathrm{d}x) \propto \mu^\star(\mathrm{d}x)\mathbf{1}_{x \in A}$$

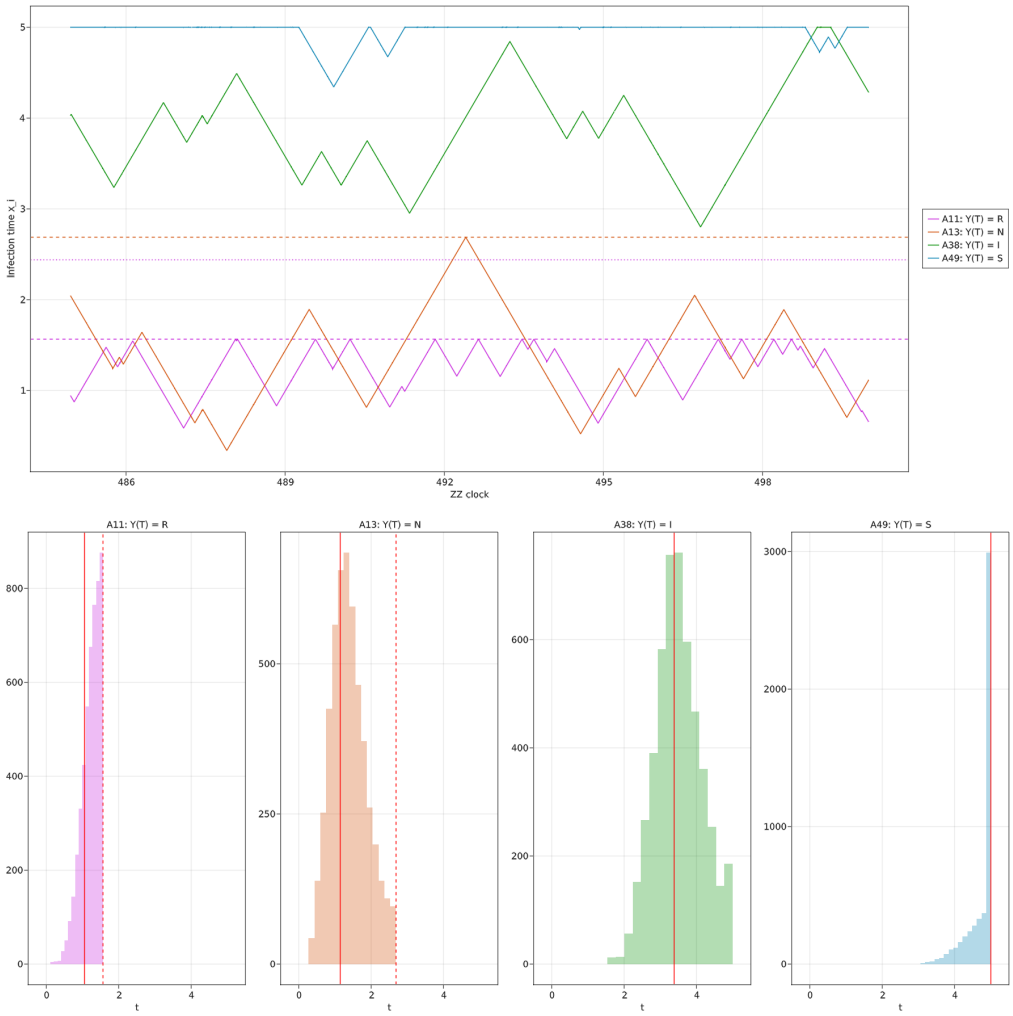


Figure 5.2: Top panel: Final 25 time units of the coordinate-process relative to the infection times of individuals 11 (pink), 13 (orange), 38 (green), 49 (blue) which, at observation time T had a different status (susceptible, infected but not notified, notified and removed). Dashed lines corresponds to the notification times, dotted lines corresponds to the removing time. Bottom panel: marginal densities of the infection times of those individuals estimated with the full trajectory of the Zig-Zag.

with $A = \bigcap_{i=1}^{N} \bigcap_{\substack{j=1, \\ j \neq i}}^{N} A_{i,j}$ and

$$A_{i,j} = \{x \in \mathbb{R}^{dN} \colon \|x^{(i)} - x^{(j)}\| \geq (r_i + r_j)\},$$

that is the measure $\mu^\star$ conditioned on the space where all the hard-spheres do not overlap. The restriction for the process to in A creates boundaries which slow down the exploration of the state space for standard PDMP samplers as they must reflect

the velocity vector when hitting the set A^c, thus never crossing the boundary, see Bierkens et al. (2018) for a detailed description of standard PDMPs on restricted domains.

We apply the (Nd)-dimensional Bouncy Particle Sampler described in Section 5.2.5 with no speed-up ($s(x) = 1$). To that end, we define the boundary of the process

$$\partial E^+ = \{(x, v) \in \partial\Omega \times \mathbb{R}^{dN} \colon \langle n(x), v\rangle > 0\}$$

where $\partial\Omega = \bigcup_{i=1}^N \bigcup_{\substack{j=1 \\ j\neq i}}^N \partial\Omega_{i,j}$ and

$$\partial\Omega_{i,j} = \{x \in \mathbb{R}^{dN} \colon |x^{(i)} - x^{(j)}| = (r_i + r_j)\}.$$

We define the kernel $\mathcal{T}(x, \cdot) = \delta_{\kappa(x)}(\cdot)$ for a function $\kappa \colon \partial\Omega \to \partial\Omega$. For $x \in \partial\Omega_{i,j}$, we set

$$[\kappa(x)]^{(\ell)} = \begin{cases} x^{(j)} + \frac{x^{(i)} - x^{(j)}}{r_i + r_j}(r_i - r_j) & \text{if } \ell = i \\ x^{(i)} + \frac{x^{(j)} - x^{(i)}}{r_i + r_j}(r_j - r_i) & \text{if } \ell = j \\ x^{(\ell)} & \text{otherwise} \end{cases}$$

which attempt to swap the location of the balls i and j by moving the smaller hard-sphere more than the larger hard-sphere, while preserving the location of the extremities of the two hard-spheres. The teleportation is successful with non-zero probability only if $\kappa(x) \in A$, that is, if after teleportation, no hard-spheres overlaps, see Figure 5.3 for an illustration. Other possible choices of teleportation portals are possible, see Appendix D.3 for a discussion.

The Bouncy Particle Sampler with teleportation at the boundary behaves as follows. For any point $(x, v) \in \partial E^+$ and for $\ell = 1, 2, \ldots, N$:

- propose to teleport to $y = \kappa(x)$;
- if $\kappa(x) \in A$, then with probability $\alpha(x, \kappa(x))$, set the new state equal to $(\kappa(x), w)$ with $w = R_n(\kappa(x), -v)$;
- otherwise reflect the velocity at the boundary and set the new state equal to (x, w) with $w = R_n(x, v)$,

where R_n is defined in equation (5.16).

Numerical experiment

We fix $N = 6$ and $d = 2$. We let $r_i \sim 2.0 + 1.5\text{Unif}([0, 1])$. We set

$$\Psi_0(x) = \frac{1}{4}\|x\|^2, \quad x \in \mathbb{R}^d$$

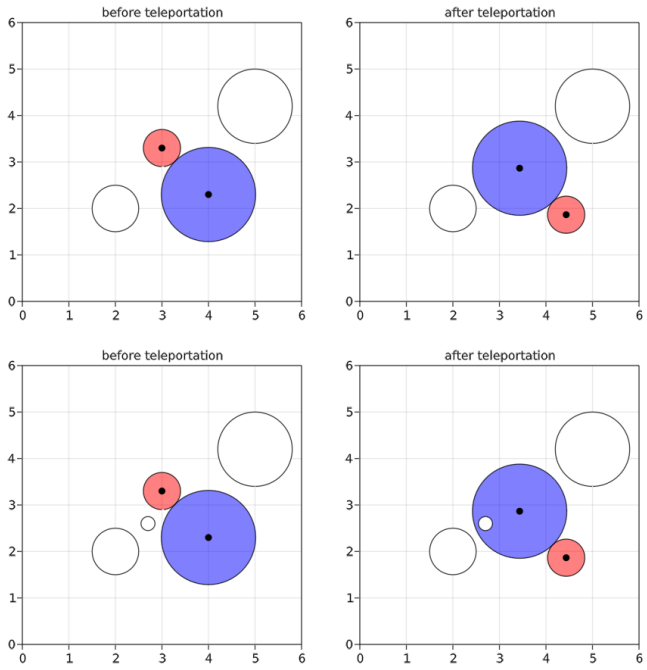


Figure 5.3: Illustration of the teleportation portals. Left panels: two configurations $x\colon (x, v) \in \partial E^+$. Right panels: $y = \tau(x)$. Top panels: the proposed state after teleportation is valid, i.e. $\tau(x) \in A$. Bottom panels: the proposed state after teleportation is not valid i.e. $\tau(x) \in A^c$.

and compare the performance of the standard BPS on constrained space as presented in Bierkens et al. (2018) and the BPS with teleportation as presented in Section 5.2.5. We initialize both the samplers in a valid configuration, see Figure 5.4. Both samplers have refreshment rate $\lambda_{\mathrm{r},E}(z) = 0.01$.

Figure 5.5 compares the trace of the functional $\langle x_i, x_j \rangle$ where $i, j \in \{1, 2, \ldots, N\}$ are the indices of the hard-spheres with largest radius: $r_i = \max(r), r_j = \max(r_{-i})$ obtained when running the two samplers with final clock equal to 2,000. The animations showing the evolution of the hard-spheres according to the dynamics of the standard BPS and BPS with teleportation may be found at `https://github.com/SebaGraz/hard-sphere-model`.

5.4 Discussion

The material presented in this chapter offers some open problems.

Firstly, the theoretical framework presented in Section 5.2 introduces a piecewise constant speed-up function s_j which, as discussed in Remark 5.2.4, can be tuned

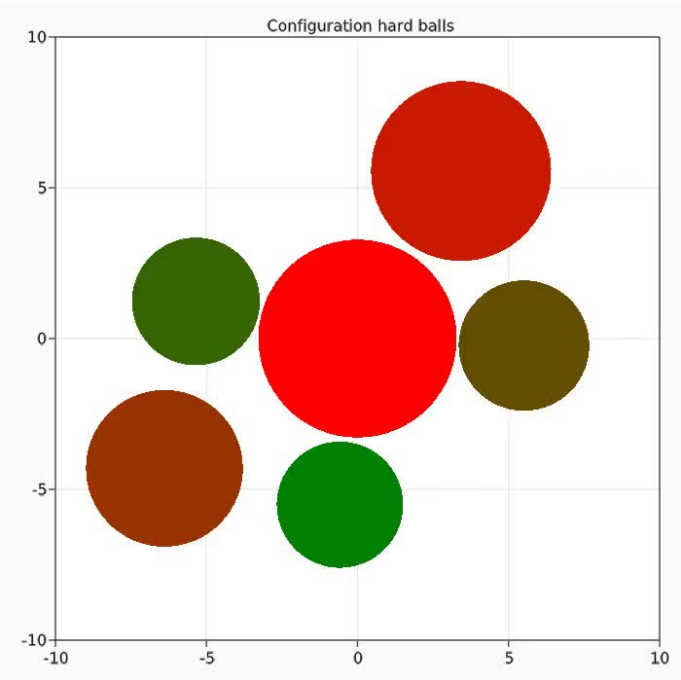


Figure 5.4: Initial configuration $x(0)$ for the standard BPS and BPS with teleportation. Here $x(0) \in A$.

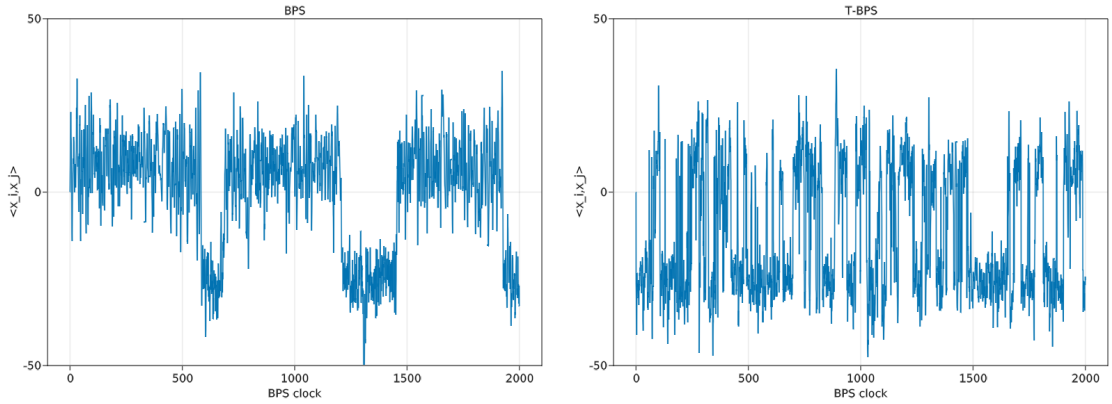


Figure 5.5: trace $t \to \langle x_i(t), x_j(t) \rangle$ where t is the clock of the Bouncy Particle Sampler (BPS) and i, j are the indices of the two hard-sphere with largest radius. Left: BPS without teleportation. Right: BPS with teleportation.

to increase the probability to cross the boundary (provided that Assumption 5.2.1 is satisfied). In Section 5.3 we set $s_j(x) = 1$ and we do investigate the benefits of tuning the speed-up function. It was proven in Vasdekis and Roberts (2021) that

the speed-up function is beneficial for heavy tailed distribution and we expect it can improve the performance of the sampler for multi-modal densities.

Secondly, in Section 5.3.1, we use PDMPs for sampling the latent space of infection times in the SINR model. In contrast with the method proposed for example in Jewell et al. 2009, this framework allows to set the state of each individual in $\mathcal{N}_T^c$ from susceptible to infected (and vice-versa) continuously in time, without an acceptance-rejection step and furthermore without any tailored proposal kernel. Mixing times of the Zig-Zag sampler and scaling of the algorithm's complexity used for this application are not analyzed and numerical comparisons with other existing MCMC methods are not performed.

Finally, in Appendix D.3, we discuss some possible choices of teleportation portals for hard-sphere models. The list of teleportation portals considered is far from exhaustive and a rigorous study on the choice of teleportation portals for hard-sphere models is not present.

This page is intentionally left blank.

Appendix A

Supplement of Chapter 2

A.1 Factorization of the diffusion bridge measure

Here we derive rigorously the conditional independence structure of the coefficients which arise from the compact support of the Faber-Schauder functions as shown in Figure 2.4. Recall that the relation $\xi_{i,j} \ll \xi_{k,l}$ holds if $S_{k,l} \subset S_{i,j}$ and in that case we refer to $\xi_{i,j}$ as the *ancestor* of $\xi_{k,l}$ (and conversely $\xi_{k,l}$ as the *descendant*). Notice that each coefficient is both descendant and ancestor of itself.

Proposition A.1.1. *(Conditional independence structure) Denote the set of common ancestors of $\xi_{i,j}$ and $\xi_{k,l}$ by $A_{(i,j;k,l)} := \{\xi_{h,d} \colon \xi_{h,d} \ll \xi_{k,l} \wedge \xi_{h,d} \ll \xi_{i,j}\}$. Under $\mathbb{P}_N^{v_T}$, $\xi_{i,j}$ is conditionally independent from $\xi_{k,l}$, given the set $A_{(i,j;k,l)}$, whenever the interior of the supports of their basis function are disjoint that is neither $\xi_{i,j} \ll \xi_{k,l}$ nor $\xi_{k,l} \ll \xi_{i,j}$ is satisfied.*

Proof. For $i = 1, ..., N$; $j = 1, ..., 2^i - 1$, define the vectors of ancestors and descendants of $\xi_{i,j}$ as $\xi^{(i,j)} := \{\xi_{m,n} \colon \xi_{m,n} \ll \xi_{i,j} \vee \xi_{m,n} \gg \xi_{i,j}\}$. Assume, without loss of generality, that $i \leq k$ and consider two coefficients $\xi_{i,j}, \xi_{k,l}$. We factorize $Z^N(X)$ by partitioning the integration interval $[0, T]$ in a sequence of sub-intervals $S_{k,0}, S_{k,1}, ..., S_{k,2^k-1}$ so that

$$Z^N(X) = \prod_{p=1}^{2^k-1} f_{k,p}(\xi^{(k,p)}). \tag{A.1}$$

Here

$$f_{k,p}(\xi^{(k,p)}) = \exp\left(B(X^N_{\max S_{k,p}}) - B(X^N_{\min S_{k,p}}) - \frac{1}{2}\int_{S_{k,p}} b^2\left(X_s^{N;k,p}\right) + b'\left(X_s^{N;k,p}\right) \mathrm{d}s \right).$$

with

$$X_s^{N;k,p} = \bar{\bar{\phi}}(s)u + \bar{\phi}(s)v_T/\sqrt{T} + \sum_{(i,j)\colon \xi_{i,j} \ll \xi_{k,p}} \phi_{i,j}(s)\xi_{i,j}$$

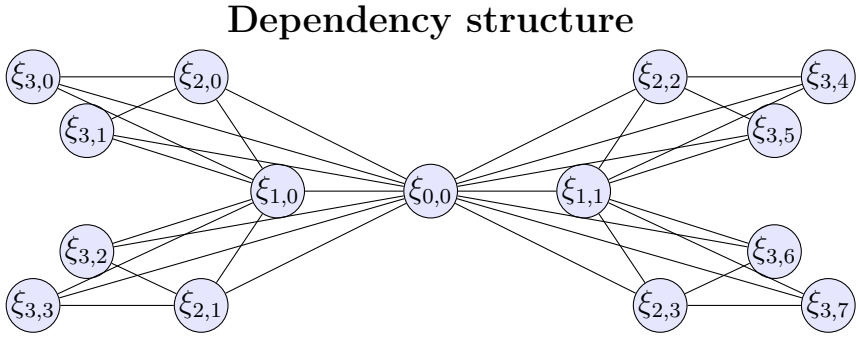


Figure A.1: Graphical representation of the dependency structure of the random vector of the coefficients under $\mathbb{P}_N^{u,v_T}$. $\xi_{i,j} \perp\!\!\!\perp \xi_{k,l}$ conditionally on the vertices which have a direct hedge to both $\xi_{i,j}$ and $\xi_{k,l}$ if $\xi_{i,j}$ does not have a direct edge to $\xi_{k,l}$. The dependency graph is a *chordal graph*.

and we used that $X_s^N = X_s^{N;i,j}$ when $s \in S_{i,j}$, $X_T^N = \bar{\phi}(T)v_T/\sqrt{T}$ and $X_0^N = \bar{\bar{\phi}}(0)u$. Now just notice that, under this factorization, the only factor which is a function of $\xi_{k,l}$ is $f_{k,l}(\xi^{(k,l)})$. Here, if $\xi_{i,j} \not\ll \xi_{k,l}$ then $\xi^{(k,l)}$ does not contain $\xi_{i,j}$. Conversely, the factors containing $\xi_{i,j}$ are those $f_{k,p}(\xi^{(k,p)})$ such that $\xi_{i,j} \ll \xi_{k,p}$ with $p = 0, 1, ..., 2^k - 1$. If $\xi_{i,j} \not\ll \xi_{k,l}$, none of the vectors $\xi^{(k,p)}$ contains $\xi_{k,l}$. Since, under the measure $\mathbb{Q}^{u,v_T}$, the random variables in the vector ξ^N are pairwise independent, the factorization on $Z^N(X)$ defines the dependency structure of the vector ξ^N under $\mathbb{P}_N^{v_T}$ so that $\xi_{i,j}$ and $\xi_{k,l}$ are independent conditionally on their common coefficients given by the set $A_{(i,j;k,l)}$. □

More intuitively, the factorization of $Z(X)$ gives rise to the dependency graph displayed in Figure A.1 which shows that the coefficients in high *levels* (i large) are coupled with just few other coefficients and conditionally independent from all the remaining. The conditional independence of the coefficients implies that the partial derivatives of the energy function (and consequently the Poisson rates given by equation (2.6)) are functions of only few coefficients in the sense of Assumption 2.4.1. In particular the sets in Assumption 2.4.1 (using double indexing) can be chosen as $N_{i,j} = \{\xi_{h,d} \colon \xi_{h,d} \ll \xi_{i,j} \vee \xi_{h,d} \gg \xi_{i,j}\}$ with size $|N_{i,j}| = 2^{N-i+1} + i - 1$, where N is the truncation level.

Appendix B

Supplement of Chapter 3

B.1 Generator and stationary distribution

B.1.1 Boomerang Sampler

For simplicity take $\boldsymbol{x}_\star = \mathbf{0}$. The generator of the Boomerang Sampler is defined by

$$\begin{aligned}\mathcal{L}\psi(\boldsymbol{x},\boldsymbol{v}) = &\langle \boldsymbol{v}, \nabla_{\boldsymbol{x}}\psi(\boldsymbol{x},\boldsymbol{v})\rangle - \langle \boldsymbol{x}, \nabla_{\boldsymbol{v}}\psi(\boldsymbol{x},\boldsymbol{v})\rangle \\ &+ \lambda(\boldsymbol{x},\boldsymbol{v})\left(\psi(\boldsymbol{x},\boldsymbol{R}(\boldsymbol{x})\boldsymbol{v}) - \psi(\boldsymbol{x},\boldsymbol{v})\right) \\ &+ \lambda_{\text{refr}}\left(\int_{\mathbb{R}^d}\psi(\boldsymbol{x},\boldsymbol{w})\phi(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} - \psi(\boldsymbol{x},\boldsymbol{v})\right),\end{aligned}$$

for any compactly supported differentiable function ψ on S, where ϕ is the probability density function of $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$.

Taking $\lambda(\boldsymbol{x},\boldsymbol{v})$ and $\boldsymbol{R}(\boldsymbol{x})$ as in Eqs. (2) and (3) of the paper respectively, we will now verify that $\int_S \mathcal{L}\psi\, d\mu = 0$ for all such functions ψ, and for μ being the measure on S with density $\exp(-U(\boldsymbol{x}))$ relative to μ_0. This then establishes that the Boomerang Sampler has stationary distribution μ. A complete proof also requires verification that the compactly supported, differentiable functions form a core for the generator, which is beyond the scope of this paper. For a discussion of this topic for archetypal PDMPs see Holderrieth 2019.

First we consider the terms involving the partial derivatives of ψ. By partial integration, we find

$$\begin{aligned}&\int_S \langle \boldsymbol{v}, \nabla_{\boldsymbol{x}}\psi(\boldsymbol{x},\boldsymbol{v})\rangle - \langle \boldsymbol{x}, \nabla_{\boldsymbol{v}}\psi(\boldsymbol{x},\boldsymbol{v})\rangle\, \mu(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{v}) \\ &= \int_S \psi(\boldsymbol{x},\boldsymbol{v})\langle \boldsymbol{v}, \nabla U(\boldsymbol{x})\rangle\, \mu(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{v})\end{aligned}$$

Next we inspect the term representing the switches occurring at rate $\lambda(\boldsymbol{x},\boldsymbol{v})$. By Eq. (5) of the paper, the coordinate transform $\boldsymbol{w} = \boldsymbol{R}(\boldsymbol{x})\boldsymbol{v}$ (for fixed $\boldsymbol{x}$) leaves the

measure $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ over the velocity component invariant. Using this observation, we find that

$$\begin{aligned}
&\int_S \lambda(\boldsymbol{x}, \boldsymbol{v})(\psi(\boldsymbol{x}, \boldsymbol{R}(\boldsymbol{x})\boldsymbol{v}) - \psi(\boldsymbol{x}, \boldsymbol{v}))\, \mu(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{v}) \\
&= \int_S \lambda(\boldsymbol{x}, \boldsymbol{R}(\boldsymbol{x})\boldsymbol{w})\psi(\boldsymbol{x}, \boldsymbol{w})\, \mu(d\boldsymbol{x}, d\boldsymbol{w}) \\
&\quad - \int_S \lambda(\boldsymbol{x}, \boldsymbol{v})\psi(\boldsymbol{x}, \boldsymbol{v})\, \mu(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{v}) \\
&= \int_S [\lambda(\boldsymbol{x}, \boldsymbol{R}(\boldsymbol{x})\boldsymbol{v}) - \lambda(\boldsymbol{x}, \boldsymbol{v})]\psi(\boldsymbol{x}, \boldsymbol{v})\, \mu(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{v}).
\end{aligned}$$

Using Eq. (2) and (4) of the paper, and the identity $(-a)_+ - (a)_+ = -a$, it follows that this expression is equal to

$$\begin{aligned}
&\int_S [\langle \boldsymbol{R}(\boldsymbol{x})\boldsymbol{v}, \nabla U(\boldsymbol{x})\rangle_+ - \langle \boldsymbol{v}, \nabla U(\boldsymbol{x})\rangle_+]\, \psi(\boldsymbol{x}, \boldsymbol{v})\, \mu(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{v}) \\
&= -\int_S \langle \boldsymbol{v}, \nabla U(\boldsymbol{x})\rangle \psi(\boldsymbol{x}, \boldsymbol{v})\, \mu(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{v}).
\end{aligned}$$

Finally by changing the order of integration, it can be shown that

$$\int_S \lambda_{\mathrm{refr}} \left(\int_{\mathbb{R}^d} \psi(\boldsymbol{x}, \boldsymbol{v})\phi(\boldsymbol{v})\, \mathrm{d}\boldsymbol{v} - \psi(\boldsymbol{x}, \boldsymbol{v}) \right) \mu_0(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{v}) = 0.$$

Adding all terms yields that $\int_S \mathcal{L}\psi \, \mathrm{d}\mu = 0$.

B.1.2 Factorised Boomerang Sampler

The Factorised Boomerang Sampler has generator

$$\begin{aligned}
\mathcal{L}\psi(\boldsymbol{x}, \boldsymbol{v}) = &\langle \boldsymbol{v}, \nabla_{\boldsymbol{x}} \psi(\boldsymbol{x}, \boldsymbol{v})\rangle - \langle \boldsymbol{x}, \nabla_{\boldsymbol{v}} \psi(\boldsymbol{x}, \boldsymbol{v})\rangle \\
&+ \sum_{i=1}^{d} \lambda_i(\boldsymbol{x}, \boldsymbol{v})(\psi(\boldsymbol{x}, \boldsymbol{F}_i(\boldsymbol{v})) - \psi(\boldsymbol{x}, \boldsymbol{v})) \\
&+ \lambda_{\mathrm{refr}} \left(\int \psi(\boldsymbol{x}, \boldsymbol{w})\phi(\boldsymbol{w})\, d\boldsymbol{w} - \psi(\boldsymbol{x}, \boldsymbol{v}) \right).
\end{aligned}$$

Verifying stationarity of μ is done analogously to the case of the non-factorised Boomerang Sampler, but now has to be carried out componentwise.

B.2 Computational bounds

Suppose $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ satisfies the Hamiltonian dynamics ODE of Eq. (1) in the paper, starting from $(\boldsymbol{x}_0, \boldsymbol{v}_0)$ in $\mathbb{R}^d \times \mathbb{R}^d$. Throughout we assume $U : \mathbb{R}^d \to \mathbb{R}$ is a twice continuously differentiable function with Hessian matrix $\nabla^2 U$. Furthermore we assume without loss of generality that $\boldsymbol{x}_\star = \mathbf{0}$. First we consider bounds for switching intensities of the form $\lambda(\boldsymbol{x}, \boldsymbol{v}) = \langle \boldsymbol{v}, \nabla U(\boldsymbol{x}) \rangle_+$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ we use $\|\boldsymbol{A}\|$ to denote the matrix norm induced by the Euclidean metric.

Lemma B.2.1 (Constant bound)**.** *Suppose there exists a constant $M > 0$ such that for all $\boldsymbol{x} \in \mathbb{R}^d$ we have the global bound*

$$\|\nabla^2 U(\boldsymbol{x})\| \leq M.$$

Define $m := |\nabla U(\mathbf{0})|$. Then for all $t \geq 0$,

$$\lambda(\boldsymbol{x}_t, \boldsymbol{v}_t) \leq \frac{M}{2}(|\boldsymbol{x}_0|^2 + |\boldsymbol{v}_0|^2) + m\sqrt{|\boldsymbol{x}_0|^2 + |\boldsymbol{v}_0|^2}. \tag{B.1}$$

Proof. We have the following estimate on the switching intensity.

$$\begin{aligned}\lambda(\boldsymbol{x}, \boldsymbol{v}) &= \langle \boldsymbol{v}, \nabla U(\boldsymbol{x}) \rangle_+ \\ &\leq \langle \boldsymbol{v}, \nabla U(\mathbf{0}) \rangle_+ + \int_0^1 |\langle \boldsymbol{v}, \nabla^2 U(\boldsymbol{x}s)\boldsymbol{x} \rangle| \, \mathrm{d}s.\end{aligned}$$

We may bound the inner product in the integrand as follows.

$$\begin{aligned}|\langle \boldsymbol{v}, \nabla^2 U(\boldsymbol{y})\boldsymbol{x} \rangle| &\leq \|\nabla^2 U(\boldsymbol{x})\| \; |\boldsymbol{v}| \; |\boldsymbol{x}| \\ &\leq M \left(\frac{|\boldsymbol{v}|^2 + |\boldsymbol{x}|^2}{2} \right)\end{aligned}$$

by the Cauchy–Schwarz inequality. Also

$$|\langle \boldsymbol{v}, \nabla U(\mathbf{0}) \rangle| \leq m|\boldsymbol{v}| \leq m\sqrt{|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2}.$$

Combining these estimates and the fact that $|\boldsymbol{x}_t|^2 + |\boldsymbol{v}_t|^2$ is invariant under the dynamics of Eq. (1) in the paper yields the stated result. □

Lemma B.2.2 (Affine bound)**.** *Suppose $\|\nabla^2 U(\boldsymbol{x})\| \leq M$ for all $\boldsymbol{x} \in \mathbb{R}^d$, and let $m = |\nabla U(\mathbf{0})|$. Then for a solution $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ to Eq. (1) of the paper with $\lambda(\boldsymbol{x}, \boldsymbol{v}) = \langle \boldsymbol{v}, \nabla U(\boldsymbol{x}) \rangle_+$, we have for all $t \geq 0$*

$$\lambda(\boldsymbol{x}_t, \boldsymbol{v}_t) \leq \left(a(\boldsymbol{x}_0, \boldsymbol{v}_0) + t b(\boldsymbol{x}_0, \boldsymbol{v}_0) \right)_+,$$

where

$$\begin{aligned}a(\boldsymbol{x}, \boldsymbol{v}) &= \langle \boldsymbol{v}, \nabla U(\boldsymbol{x}) \rangle_+, \quad \textit{and} \\ b(\boldsymbol{x}, \boldsymbol{v}) &= M\left(|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2\right) + m\sqrt{|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2}.\end{aligned}$$

Proof. By the Hamiltonian dynamics,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\langle \boldsymbol{v}_t, \nabla U(\boldsymbol{x}_t)\rangle \\
&= -\langle \boldsymbol{x}_t, \nabla U(\boldsymbol{x}_t)\rangle + \langle \boldsymbol{v}_t, \nabla^2 U(\boldsymbol{x}_t)\boldsymbol{v}_t\rangle \\
&= -\langle \boldsymbol{x}_t, \nabla U(\boldsymbol{0})\rangle - \int_0^1 \langle \boldsymbol{x}_t, \nabla^2 U(s\boldsymbol{x}_t)\boldsymbol{x}_t\rangle \,\mathrm{d}s \\
&\qquad + \langle \boldsymbol{v}_t, \nabla^2 U(\boldsymbol{x}_t)\boldsymbol{v}_t\rangle \\
&\leq |\boldsymbol{x}_t||\nabla U(\boldsymbol{0})| + M\left(|\boldsymbol{x}_t|^2 + |\boldsymbol{v}_t|^2\right).
\end{aligned}$$

Using that $|\boldsymbol{x}_t|^2 + |\boldsymbol{v}_t|^2$ is invariant under the dynamics yields the stated result. □

Lemma B.2.3. *Suppose $|\nabla U(\boldsymbol{y})| \leq C$ for all $\boldsymbol{y} \in \mathbb{R}^d$. Then, for all trajectories $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ satisfying Eq. (1) of the paper we have*

$$\lambda(\boldsymbol{x}_t, \boldsymbol{v}_t) \leq C\sqrt{|\boldsymbol{x}_0|^2 + |\boldsymbol{v}_0|^2}.$$

Proof. We have

$$\lambda(\boldsymbol{x}, \boldsymbol{v}) \leq C|\boldsymbol{v}| \leq C\sqrt{|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2},$$

and the latter expression is constant along trajectories. □

Analogously we have the following useful bound for the Factorized Boomerang Sampler.

Lemma B.2.4. *Suppose $U : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Suppose there exist constants $c_1, \ldots, c_d$ such that, for all $\boldsymbol{y} \in \mathbb{R}^d$ and $i = 1, \ldots, d$, we have*

$$|\partial_i U(\boldsymbol{x})| \leq c_i \quad \textit{for all } \boldsymbol{x},\ i.$$

Then

$$\lambda_i(\boldsymbol{x}_t, \boldsymbol{v}_t) \leq c_i\sqrt{|x_0^i|^2 + |v_0^i|^2}.$$

Lemma B.2.5. *Suppose for all i we have that*

$$\sqrt{\sum_j \partial_i\partial_j U(\boldsymbol{x})^2} \leq M_i,$$

and

$$|\partial_i U(\boldsymbol{0})| \leq m_i.$$

Then

$$\lambda_i(\boldsymbol{x}_t, \boldsymbol{v}_t) \leq (a_i(\boldsymbol{x}_0, \boldsymbol{v}_0) + b_i(\boldsymbol{x}_0, \boldsymbol{v}_0)t)^+$$

where

$$a_i(\boldsymbol{x}, \boldsymbol{v}) = (v^i \partial_i U(\boldsymbol{x}))^+$$

$$b_i(\boldsymbol{x}, \boldsymbol{v}) = \sqrt{(x^i)^2 + (v^i)^2} \left(m_i + M_i \sqrt{|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2} \right).$$

Proof. We compute

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} v_t^i \partial_i U(\boldsymbol{x}_t) \\
&= -x_t^i \partial_i U(\boldsymbol{x}_t) + v_t^i \sum_{j=1}^d \partial_i \partial_j U(\boldsymbol{x}_t) v_t^j \\
&= -x_t^i \partial_i U(\boldsymbol{0}) - \int_0^1 x_t^i \sum_{j=1}^d \partial_i \partial_j U(s\boldsymbol{x}_t) x_t^j \mathrm{d}s \\
&\quad + v_t^i \sum_{j=1}^d \partial_i \partial_j U(\boldsymbol{x}_t) v_t^j \\
&\leq \sqrt{(x_t^i)^2 + (v_t^i)^2} |\partial_i U(\boldsymbol{0})| + M_i |x_t^i| |\boldsymbol{x}_t| + M_i |v_t^i| |\boldsymbol{v}_t| \\
&\leq \sqrt{(x_t^i)^2 + (v_t^i)^2} |\partial_i U(\boldsymbol{0})| \\
&\quad + M_i / 2 \left(\alpha (|x_t^i|^2 + |v_t^i|^2) + (1/\alpha)(|\boldsymbol{x}_t|^2 + |\boldsymbol{v}_t|^2) \right).
\end{aligned}$$

Optimising over α, and using that $|x_t^i|^2 + |v_t^i|^2$ is constant along Factorised Boomerang Trajectories, yields the stated result. □

B.2.1 Computational bounds for subsampling

In the case of subsampling we use the unbiased estimator of Eq. (9) of the paper.

Lemma B.2.6. *Suppose that for some positive definite matrix $\boldsymbol{Q}$ we have that, for all i, and $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^d$,*

$$\nabla^2 E^i(\boldsymbol{y}_1) - \nabla^2 E^i(\boldsymbol{y}_2) \preceq \boldsymbol{Q}, \tag{B.2}$$

where $A \preceq B$ means $B - A$ is positive semidefinite. Suppose $\widehat{\nabla U(\boldsymbol{x})}$ is given by Eq. (9) of the paper, and $\nabla E(\boldsymbol{0}) = \boldsymbol{0}$. Along a trajectory $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ satisfying the Hamiltonian dynamics of Eq. (1) of the paper, we have, for all $t \geq 0$, that

$$\langle \boldsymbol{v}_t, \widehat{\nabla U(\boldsymbol{x}_t)} \rangle \leq \tfrac{1}{2} (|\boldsymbol{Q}^{1/2} \boldsymbol{x}_0|^2 + |\boldsymbol{Q}^{1/2} \boldsymbol{v}_0|^2), \quad a.s.$$

where the almost sure statement is with respect to all random (subsampling) realisations of the switching intensity.

Remark B.2.1. *Lemma (B.2.6) is easily extended to the case in which* $\nabla E(\mathbf{0}) \neq \mathbf{0}$. *In this case we have*

$$\langle \boldsymbol{v}_t, \widehat{U(\boldsymbol{x}_t)} \rangle \leq \tfrac{1}{2}(|\boldsymbol{Q}^{1/2}\boldsymbol{x}_0|^2 + |\boldsymbol{Q}^{1/2}\boldsymbol{v}_0|^2) + (|\boldsymbol{v}_0|^2 + |\boldsymbol{x}_0|^2)^{1/2}|\nabla E(\mathbf{0})|, \quad a.s.$$

Remark B.2.2. *In practice one may wish to take* $\boldsymbol{Q}$ *to be a diagonal matrix, which reduces the computation of the computational bound to a* $\mathcal{O}(d)$ *computation instead of* $\mathcal{O}(d^2)$. *For example one could take* $\boldsymbol{Q} = c\boldsymbol{I}$ *for a suitable constant* $c > 0$ *such that (B.2) is satisfied.*

Remark B.2.3 (Affine bound for subsampling is strictly worse)**.** *When we try to obtain an affine bound, of the form*

$$\widehat{\lambda(\boldsymbol{x}_t, \boldsymbol{v}_t)} \leq a(\boldsymbol{x}_0, \boldsymbol{v}_0) + b(\boldsymbol{x}_0, \boldsymbol{v}_0),$$

then it seems we cannot avoid an expression for a *of the form of the bound in Lemma B.2.6. As a consequence, the affine bound is strictly worse than the constant bound.*

Proof (of Lemma B.2.6). Suppose we have $I = i$ for the random index I in Eq. (9) of the paper. We compute

$$\begin{aligned}
&\langle \boldsymbol{v}_t, \widehat{\nabla U(\boldsymbol{x}_t)} \rangle \\
&= \langle \boldsymbol{v}_t, \nabla E^i(\boldsymbol{x}_t) - \nabla^2 E^i(\mathbf{0})\boldsymbol{x}_t - \nabla E^i(\mathbf{0}) \rangle \\
&= \langle \boldsymbol{v}_t, \int_0^1 \nabla^2 E^i(s\boldsymbol{x}_t)\boldsymbol{x}_t \, \mathrm{d}s - \nabla^2 E^i(\mathbf{0})\boldsymbol{x}_t \rangle.
\end{aligned}$$

Then we may continue the above computation to find, using Lemma B.2.7 below, that

$$\begin{aligned}
\langle \boldsymbol{v}_t, \widehat{\nabla U(\boldsymbol{x}_t)} \rangle &= \int_0^1 \langle \boldsymbol{v}_t, [\nabla^2 E^i(s\boldsymbol{x}_t) - \nabla^2 E^i(\mathbf{0})]\boldsymbol{x}_t \rangle \, \mathrm{d}s \\
&\leq \int_0^1 |\boldsymbol{Q}^{1/2}\boldsymbol{v}_t| \, |\boldsymbol{Q}^{1/2}\boldsymbol{x}_t| \, \mathrm{d}s \\
&\leq \tfrac{1}{2}(|\boldsymbol{Q}^{1/2}\boldsymbol{v}_t|^2 + |\boldsymbol{Q}^{1/2}\boldsymbol{x}_t|^2).
\end{aligned}$$

Since $\frac{1}{2}(|\boldsymbol{Q}^{1/2}\boldsymbol{v}_t|^2 + |\boldsymbol{Q}^{1/2}\boldsymbol{x}_t|^2)$ is invariant under the dynamics, the stated conclusion follows. □

Lemma B.2.7. *Suppose* $\boldsymbol{M}, \boldsymbol{P} \in \mathbb{R}^{d\times d}$ *are symmetric matrices with* $\boldsymbol{P}$ *positive definite and such that* $-\boldsymbol{P} \preceq \boldsymbol{M} \preceq \boldsymbol{P}$. *Then* $\langle \boldsymbol{M}\boldsymbol{y}, \boldsymbol{z} \rangle \leq |\boldsymbol{P}^{1/2}\boldsymbol{y}| \, |\boldsymbol{P}^{1/2}\boldsymbol{z}|$ *for all* $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^{d\times d}$.

Proof. Taking $\boldsymbol{y} = \boldsymbol{P}^{-1/2}\boldsymbol{x}$, we find

$$|\langle \boldsymbol{P}^{-1/2}\boldsymbol{M}\boldsymbol{P}^{-1/2}\boldsymbol{x}, \boldsymbol{x}\rangle| = |\langle \boldsymbol{M}\boldsymbol{y}, \boldsymbol{y}\rangle| \leq \langle \boldsymbol{P}\boldsymbol{y}, \boldsymbol{y}\rangle = |\boldsymbol{x}|^2,$$

which establishes that $\|\boldsymbol{P}^{-1/2}\boldsymbol{M}\boldsymbol{P}^{-1/2}\| \leq 1$. Using this observation we arrive at

$$\langle \boldsymbol{M}\boldsymbol{y}, \boldsymbol{z}\rangle \leq \underbrace{\|\boldsymbol{P}^{-1/2}\boldsymbol{M}\boldsymbol{P}^{-1/2}\|}_{\leq 1} |\boldsymbol{P}^{1/2}\boldsymbol{y}|\,|\boldsymbol{P}^{1/2}\boldsymbol{z}|.$$

□

B.3 Scaling with dimension

In Section 3.2 of the paper, we discuss the scaling of the Boomerang Sampler with dimension. The argument in that section is self contained, but relies on the observation that the change of $E_d(\boldsymbol{x}_t)$ over a time interval of order 1 is at least of order $d^{1/2}$. Here we motivate this observation.

In the following arguments, we assume stationarity of the process for simplicity. Let U_d, $\boldsymbol{\Sigma}_d$, E_d, Π_d, $\mathbb{E}_d$ be as described in Section 3.2 of the manuscript. For simplicity and without loss of generality we assume that $E_d(\boldsymbol{x})$ is normalised as $\mathbb{E}_d[E_d(\boldsymbol{x})] = 0$. Furthermore, for simplicity we assume that $\mathbb{E}_d[\boldsymbol{x}] = \boldsymbol{0}$ although this condition can be relaxed.

As discussed we suppose that the sequence (U_d) satisfies

$$\sup_{d\in\mathbb{N}} \mathbb{E}_d[|\boldsymbol{\Sigma}_d^{1/2}\nabla U_d(\boldsymbol{x})|^2] \leq \kappa \tag{B.3}$$

for some $\kappa > 0$. Furthermore, we assume that the following form of the Poincaré inequality is satisfied for $\Pi_d(\mathrm{d}\boldsymbol{x}) \propto \exp(-E_d(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$:

$$C\ \mathbb{E}_d\left[f_d(\boldsymbol{x})^2\right]^{1/2} \leq \mathbb{E}_d\left[|\boldsymbol{\Sigma}_d^{1/2}\nabla f_d(\boldsymbol{x})|^2\right]^{1/2} \tag{B.4}$$

for some constant $C > 0$ not depending on d, and any differentiable function $f_d : \mathbb{R}^d \to \mathbb{R}$ with mean 0 and finite variance.

By (B.3) the expected number of reflections per unit time $\mathbb{E}_d[\langle \boldsymbol{v}, \nabla U_d(\boldsymbol{x})\rangle_+]$ is bounded with respect to dimension. However the process mixes well in a single time unit under suitable regularity conditions as we will discuss now.

By applying (B.4) to $f_d(\boldsymbol{x}) = (\boldsymbol{\Sigma}_d^{-1/2}\boldsymbol{x})_i$, where $\boldsymbol{v}_i$ denotes the i-th coordinate of $\boldsymbol{v}$, we have $C^2\mathbb{E}_d[|\boldsymbol{\Sigma}_d^{-1/2}\boldsymbol{x}|^2] \leq \mathbb{E}_d[\mathrm{trace}(\boldsymbol{\Sigma}_d^{-1/2}\boldsymbol{\Sigma}_d\boldsymbol{\Sigma}_d^{-1/2})] = d$, using the stated assumption $\mathbb{E}_d[\boldsymbol{x}] = \boldsymbol{0}$.

Also by (B.4) and by Minkowski's inequality,

$$\begin{aligned}\mathbb{E}_d[E_d(\boldsymbol{x})^2]^{1/2} &\le C^{-1}\mathbb{E}_d[|\boldsymbol{\Sigma}_d^{1/2}\nabla E_d(\boldsymbol{x})|^2]^{1/2} \\ &= C^{-1}\mathbb{E}_d[|\boldsymbol{\Sigma}_d^{1/2}\nabla U_d(\boldsymbol{x}) + \boldsymbol{\Sigma}_d^{-1/2}\boldsymbol{x}|^2]^{1/2} \\ &= C^{-1}(\kappa^{1/2} + C^{-1}d^{1/2}) = \mathcal{O}(d^{1/2}).\end{aligned}$$

If $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ satisfies the ODE Eq. (1) of the paper, the unit time difference $E_d(\boldsymbol{x}_t) - E_d(\boldsymbol{x}_0)$ is

$$\int_0^t \langle \nabla E_d(\boldsymbol{x}_s), \boldsymbol{v}_s\rangle \mathrm{d}s \approx \int_0^t \langle \boldsymbol{\Sigma}_d^{-1}\boldsymbol{x}_s, \boldsymbol{v}_s\rangle \mathrm{d}s.$$

Here, the difference between the left- and the right-hand sides is $\int_0^t \langle \boldsymbol{\Sigma}_d^{1/2}\nabla U(\boldsymbol{x}_s), \boldsymbol{\Sigma}_d^{-1/2}\boldsymbol{v}_s\rangle \mathrm{d}s$ which is of order $d^{1/2}$ under the assumption of stationarity by (B.3) and the Cauchy-Schwarz inequality, using that $\mathbb{E}_d[|\boldsymbol{\Sigma}_d^{-1/2}\boldsymbol{v}_s|^2] = d$. The right-hand may be simplified to

$$\begin{aligned}&\int_0^t \langle \boldsymbol{\Sigma}_d^{-1}(\boldsymbol{x}_0 \cos s + \boldsymbol{v}_0 \sin s), -\boldsymbol{x}_0 \sin s + \boldsymbol{v}_0 \cos s\rangle \mathrm{d}s \\ &= A_0 \int_0^t 2 \sin s \ \cos s \ \mathrm{d}s + B_0 \int_0^t (\cos^2 s - \sin^2 s) \mathrm{d}s \\ &= A_0(1 - \cos 2t)/2 + B_0(\sin 2t)/2\end{aligned}$$

where $A_0 = (\langle \boldsymbol{v}_0, \boldsymbol{\Sigma}_d^{-1}\boldsymbol{v}_0\rangle - \langle \boldsymbol{x}_0, \boldsymbol{\Sigma}_d^{-1}\boldsymbol{x}_0\rangle)/2$ and $B_0 = \langle \boldsymbol{x}_0, \boldsymbol{\Sigma}_d^{-1}\boldsymbol{v}_0\rangle$. Then A_0 and B_0 are uncorrelated since $\boldsymbol{\Sigma}_d^{-1/2}\boldsymbol{v}_0$ follows the standard normal distribution. Also, $\mathbb{E}_d[A_0^2] \ge \mathrm{Var}(A_0) \ge \mathrm{Var}(\langle \boldsymbol{v}_0, \boldsymbol{\Sigma}_d^{-1}\boldsymbol{v}_0\rangle) = 2d$. Therefore,

$$\begin{aligned}\mathbb{E}_d[|E_d(\boldsymbol{x}_t) - E_d(\boldsymbol{x}_0)|^2] &\gtrsim \mathbb{E}_d[A_0^2]\left(\frac{1 - \cos 2t}{2}\right)^2 \\ &\ge 2d\left(\frac{1 - \cos 2t}{2}\right)^2.\end{aligned}$$

Thus the change of $E_d(\boldsymbol{x}_t)$ over a term interval of $\mathcal{O}(1)$ is of order $d^{1/2}$ whereas $E_d(\boldsymbol{x}_t)$ itself has the same order. These informal arguments suggest that dynamics of the Boomerang sampler in a finite time interval sufficiently changes the log density even in high dimension. However, further study should be made in this direction.

B.4 Logistic regression

We assume a prior distribution $\pi_0(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ on $\mathbb{R}^d$. Given predictors $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)}$ in $\mathbb{R}^d$, and outcomes $z^{(1)}, \ldots, z^{(n)}$ in $\{0, 1\}$, we obtain the negative

log posterior distribution as

$$E(\boldsymbol{x}) = \sum_{i=1}^{n} \left\{ \log(1 + e^{\boldsymbol{x}^\top \boldsymbol{y}^{(i)}}) - z^{(i)} \boldsymbol{x}^\top \boldsymbol{y}^{(i)} \right\} + |\boldsymbol{x}|^2 / 2\sigma^2.$$

We then have

$$\nabla E(\boldsymbol{x}) = \boldsymbol{x}/\sigma^2 + \sum_{i=1}^{n} \boldsymbol{y}^{(i)} \left[\frac{e^{\boldsymbol{x}^\top \boldsymbol{y}^{(i)}}}{1 + e^{\boldsymbol{x}^\top \boldsymbol{y}^{(i)}}} - z^{(i)} \right],$$
$$\nabla^2 E(\boldsymbol{x}) = \boldsymbol{I}/\sigma^2 + \sum_{i=1}^{n} \frac{\boldsymbol{y}^{(i)} (\boldsymbol{y}^{(i)})^\top e^{\boldsymbol{x}^\top \boldsymbol{y}^{(i)}}}{\left(1 + e^{\boldsymbol{x}^\top \boldsymbol{y}^{(i)}}\right)^2}.$$

In the experiments in this paper we take a flat prior, i.e. $\sigma^2 = \infty$.

Let

$$\boldsymbol{x}_\star = \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} E(\boldsymbol{x}).$$

We take $\boldsymbol{\Sigma}^{-1} = \nabla^2 E(\boldsymbol{x}_\star)$. We have $U(\boldsymbol{x}) = E(\boldsymbol{x}) - (\boldsymbol{x} - \boldsymbol{x}_\star)^\top \nabla^2 E(\boldsymbol{x}_\star)(\boldsymbol{x} - \boldsymbol{x}_\star)/2$, which is a difference of two positive definite matrices. Using the general inequality $a \mapsto |a|/(1+a)^2 \leq 1/4$, we find

$$-\tfrac{1}{4} \sum_{i=1}^{n} \boldsymbol{y}^{(i)} (\boldsymbol{y}^{(i)})^\top \preceq \nabla^2 U(\boldsymbol{x}) \preceq \tfrac{1}{4} \sum_{i=1}^{n} \boldsymbol{y}^{(i)} (\boldsymbol{y}^{(i)})^\top.$$

We then simply have

$$\|\nabla^2 U(\boldsymbol{y})\| \leq M := \tfrac{1}{4} \| \sum_{i=1}^{n} \boldsymbol{y}^{(i)} (\boldsymbol{y}^{(i)})^\top \|.$$

These observations may be applied in conjunction with the lemmas of Section 2 in this supplement to obtain useful constant and affine computational bounds for the switching intensities.

B.5 Diffusion bridge simulation

We consider diffusion bridges of the form

$$\mathrm{d}X_t = \alpha \sin(X_t)\mathrm{d}t + \mathrm{d}W_t, \quad X_0 = u, X_T = v, t \in [0, T] \tag{B.5}$$

where W is a scalar Brownian motion and $\alpha \geq 0$. The diffusion path is expanded with a truncated Faber-Schauder basis such that

$$X_t^N = \bar{\bar{\phi}}(t)u + \bar{\phi}(t)v + \sum_{i=0}^{N} \sum_{j=0}^{2^i - 1} \phi_{i,j}(t) x_{i,j},$$

where N is the truncation of the expansion and

$$\begin{aligned}\bar{\phi}(t) &= t/T, \qquad \bar{\bar{\phi}}(t) = 1 - t/T,\\ \phi_{0,0}(t) &= \sqrt{T}\big((t/T)\mathbf{1}_{[0,T/2]}(t) + (1 - t/T)\mathbf{1}_{(T/2,T]}(t)\big),\\ \phi_{i,j}(t) &= 2^{-i/2}\phi_{0,0}(2^i t - jT) \quad i \geq 0, \quad 0 \leq j \leq 2^i - 1,\end{aligned}$$

are the Faber-Schauder functions. As shown in Bierkens et al. 2021, the measure of the coefficients corresponding to (B.5) is derived from the Girsanov formula and given by

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(\boldsymbol{x}, \boldsymbol{v}) \propto \exp\left\{\frac{-\alpha}{2}\int_0^T \left(\alpha \sin^2(X_s^N) + \cos(X_s^N)\right) \mathrm{d}s\right\}$$

where $\mu_0 = \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \otimes \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ with $\boldsymbol{I}$ the $2^{N+1} - 1$ dimensional identity matrix. By standard trigonometric identities we have that

$$\partial_{x_{i,j}} U(\boldsymbol{x}) = \frac{\alpha}{2}\int_{S_{i,j}} \phi_{i,j}(t)\left(\alpha \sin\left(2X_t^N\right) - \sin\left(X_t^N\right)\right) \mathrm{d}t$$

where $S_{i,j}$ is the support of the basis function $\phi_{i,j}$. Similarly to Bierkens et al. 2021, for each i, j, we use subsampling and consider the unbiased estimator for $\partial_{x_{i,j}} U(\boldsymbol{x})$ given by

$$\widehat{\partial_{x_{i,j}} U(\boldsymbol{x})} = S_{i,j}\phi_{i,j}(\tau_{i,j})\left(\alpha^2 \sin\left(2X_{\tau_{i,j}}^N\right) - \alpha \sin\left(X_{\tau_{i,j}}^N\right)\right)$$

where $\tau_{i,j}$ is a uniform random variable on $S_{i,j}$. This gives Poisson rates $\widehat{\lambda_{i,j}(\boldsymbol{x}, \boldsymbol{v})} = \langle \boldsymbol{v}, \widehat{\partial_{x_{i,j}} U(\boldsymbol{x})}\rangle_+$. In this case, for all i, j, $|\widehat{\partial_{x_{i,j}} U(\boldsymbol{x})}|$ is globally bounded, say by $m_{i,j}$. We use the constant Poisson bounding rates given, in similar spirit as in Section 2.3 of the paper, by

$$\overline{\lambda}_{i,j}(\boldsymbol{x}_t, \boldsymbol{v}_t) = m_{i,j}\sqrt{|x_0^{i,j}|^2 + |v_0^{i,j}|^2},$$

where we used that $t \to |x_t^{i,j}|^2 + |v_t^{i,j}|^2$ is constant under the Factorised Boomerang trajectories. Similarly to Bierkens et al. 2021, the FBS gains computational efficiency by a local implementation which exploits the fact that each $\overline{\lambda}_{i,j}(\boldsymbol{x}, \boldsymbol{v})$ is a function of just the coefficient $x_{i,j}$ (see Bierkens et al. 2021, Algorithm 3, for an algorithmic description of the local implementation of a factorised PDMP).

Appendix C

Supplement of Chapter 4

C.1 Details of the Sticky Zig-Zag sampler

C.1.1 Construction

In this section we discuss how the Sticky Zig-Zag can be constructed as a *standard* PDMP in the sense of Davis (1993). The construction is a bit tedious, but the underlying idea is simple: the Sticky Zig-Zag process has the dynamics of a ordinary Zig-Zag process until it reaches a freezing boundary $\mathfrak{F}_i = \{(x, v) \in E \colon x_i = 0^-,\, v_i > 0 \text{ or } x_i = 0^+,\, v_i < 0\}$ of $E = \overline{\mathbb{R}}^d \times \mathcal{V}$, with $\overline{\mathbb{R}} = (-\infty, 0^-] \sqcup [0^+, \infty)$ which has two copies of 0. Then it immediately changes dynamics and evolves as a lower dimensional ordinary Zig-Zag process *on the boundary*, at least until an unfreezing event happens or upon reaching yet another freezing boundary in the domain of the restricted process.

Davis' construction allows a standard PDMP to make instantaneous jumps at boundaries of open sets, but puts restrictions on further behaviour at that boundary. We circumvent these restrictions by first splitting up the space $\mathbb{R}^d \times \mathcal{V}$ into disconnected components in a way somewhat different than the construction of E as presented in Section 4.2. Only at a later stage we recover the definition of E.

Define the set

$$K = \{\to\circ, \circ\to, \leftarrow\circ, \circ\leftarrow, \overset{\leftarrow}{\circ}, \overset{\rightarrow}{\circ}\}$$

and

$$|K| = \{\circ, \leftarrow\circ\to, \to\circ\leftarrow\}$$

(note that $|K|$ does not denote the cardinality of the set K). Define the functions $k \colon \mathbb{R} \times \mathbb{R} \to K$ and $|k| \colon \mathbb{R} \times \mathbb{R} \to |K|$ by

(x,v)	$k(x,v)$	at (x,v) the process is...	$\lvert k\rvert(x,v)$
$x>0, v>0$	$\circ\rightarrow$	...moving away from 0 with positive velocity	$\leftarrow\circ\rightarrow$
$x<0, v<0$	$\leftarrow\circ$	...moving away from 0 with negative velocity	$\leftarrow\circ\rightarrow$
$x>0, v<0$	$\circ\leftarrow$	...moving toward 0 with negative velocity	$\rightarrow\circ\leftarrow$
$x<0, v>0$	$\rightarrow\circ$	...moving toward 0 with positive velocity	$\rightarrow\circ\leftarrow$
$x=0, v>0$	$\overset{\rightarrow}{\circ}$	...at 0 with positive velocity	$\circ$
$x=0, v<0$	$\overset{\leftarrow}{\circ}$	...at 0 with negative velocity	$\circ$

If $(x,v)\in\mathbb{R}^d\times\mathcal{V}$, then extend $k\colon \overline{\mathbb{R}}^d\times\mathcal{V}\to K^d$ and $|k|\colon \overline{\mathbb{R}}^d\times\mathcal{V}\to |K|^d$ by applying the map k and $|k|$ coordinatewise.

For each $\ell\in K^d$ define

$$\widetilde{E}_\ell^\circ = \{(\ell,x,v)\colon k(x,v)=\ell\}$$

Note that for $\ell\neq\ell'$ the sets $\widetilde{E}_\ell^\circ$ and $\widetilde{E}_{\ell'}^\circ$ are disjoint. The set $\widetilde{E}_\ell^\circ$ is open under the metric introduced in Davis (1993), p.58, which sets the distance between two points (ℓ,x,v) and (ℓ',x',v') to 1 if $\ell\neq\ell'$. We denote the induced topology on $\widetilde{E}$ by $\widetilde{\tau}$. $\widetilde{E}_\ell^\circ$ is a subset of $\mathbb{R}^{2d}$ of dimension $d_\ell=\sum_{i=1}^d \mathbb{1}_{|\ell_i|\neq\circ}$, since the velocities are constant in E_ℓ° and the position of the components i where $\ell_i=\circ$ are constant as well in $\widetilde{E}_\ell^\circ$ ($\widetilde{E}_\ell^\circ$ is isomorphic to an open subset of $\mathbb{R}^{d_\ell}$).

The sets which contain a singleton, i.e. $|\widetilde{E}_\ell^\circ|=1$, are those sets $\widetilde{E}_\ell^\circ$ such that $|\ell_i(x,v)|=\circ$ for all $i=1,2,\dots,d$ and are open as they contain one isolated point, but will have to be treated a bit differently. Then $\widetilde{E}^\circ=\bigcup_{\ell\in K^d}\widetilde{E}_\ell^\circ$ is the tagged space of open subsets of $\mathbb{R}^{d_\ell}$ used in Davis (1993, Section 24).

$\widetilde{E}^\circ$ separates the space into isolated components of varying dimension. In each component, the Sticky Zig-Zag process behaves differently and essentially as a lower dimensional Zig-Zag process.

Let $\partial\widetilde{E}_\ell^\circ$ denote the boundary of $\widetilde{E}_\ell^\circ$ in the embedding space $\mathbb{R}^{d_\ell}$ (where the velocity components are constant in $\widetilde{E}_\ell^\circ$), with elements written (ℓ,x,v). Some points in $\partial\widetilde{E}_\ell^\circ$ will also belong to the state space $\widetilde{E}$ of the Sticky Zig-Zag process, but only the entrance-non-exit boundary points:

$$\widetilde{E}=\bigcup_\ell \widetilde{E}_\ell,\quad \widetilde{E}_\ell=\widetilde{E}_\ell^\circ\cup\{(\ell,x,v)\in\partial\widetilde{E}_\ell^\circ\colon x_i=0\Rightarrow|\ell_i|\neq\rightarrow\circ\leftarrow \text{ for all } i\}.$$

(This corresponds to the definition of the state space in Davis 1993, Section 24, only that we use knowledge of the flow.)

The remaining part of the boundary is

$$\Gamma=\bigcup_\ell\Gamma_\ell\subset\bigcup_\ell\partial\widetilde{E}_\ell^\circ,\quad \Gamma_\ell=\{(\ell,x,v)\in\partial\widetilde{E}_\ell^\circ,\exists i\colon x_i=0,|\ell_i|=\rightarrow\circ\leftarrow\},$$

with $\widetilde{E}\cap\Gamma = \varnothing$ so that Γ is not part of the state space $\widetilde{E}$. Any trajectory approaching Γ, jumps back into $\widetilde{E}$ just before hitting Γ. If $\widetilde{E}_\ell^\circ$ is a singleton ($|\widetilde{E}_\ell^\circ| = 1$), then $\Gamma_\ell = \emptyset$ and $\widetilde{E}_\ell = \widetilde{E}_\ell^\circ$ (atoms).

Lemma C.1.1. *A bijection $\iota\colon \widetilde{E} \to E$ is given by*

$$\iota((\ell, \widetilde{x}, v)) = (x, v)$$

where

$$x_i = \begin{cases} 0^+\,(0^-) & \ell_i = \overleftarrow{\circ}\ (\ell_i = \vec{\circ}) \\ 0^+\,(0^-) & \ell_i = \circ\!\rightarrow\ (\ell_i = \leftarrow\!\circ),\ \widetilde{x}_i = 0 \\ \widetilde{x}_i & \textit{otherwise.} \end{cases}$$

Proof. Recall that $\alpha(x,v) := \{i \in \{1,2,\dots,d\}\colon (x,v) \notin \mathfrak{F}_i\}$ and α^c denotes its complement. First of all, notice that $\iota(\widetilde{E}) \subset E$. Now let $(x,v) \in E$ be given. We construct $e \in \widetilde{E}$ such that $(x,v) = \iota(e)$. If there is at least one $x_j = 0^\pm$ with $j \notin \alpha(x,v)$, then take $e = (\ell, \widetilde{x}, v) \in \widetilde{E} \setminus \widetilde{E}^\circ$ as follows (entrance-non-exit boundary): for $i \in \alpha^C$ we have $|\ell_i| = \circ$, $\widetilde{x}_i = 0$, while for all $i \in \alpha$ with $x_i = 0^\pm$, we have $|\ell_i| = \leftarrow\!\circ\!\rightarrow$, $\widetilde{x}_i = 0$. Then $\iota(e) = (x,v)$. Otherwise, $e = (k(\widetilde{x}, v), \widetilde{x}, v)) \in \widetilde{E}^\circ$ (interior of an open set) and $\iota(e) = (x,v)$ where $\widetilde{x}_i = 0$ for all $i \in \alpha(x,v)$ and $\widetilde{x}_i = x_i$ otherwise.

□

Having constructed the state space, we proceed with the process dynamics. Firstly, the deterministic flow (locally Lipschitz for every $\ell \in K$) is determined by the functions $\widetilde{\phi}_\ell\colon [0,\infty) \times \widetilde{E}_\ell^\circ \to \widetilde{E}_\ell^\circ$ which for the sticky ZigZag process are given by

$$\widetilde{\phi}(t, \ell, x, v) = (\ell, x', v), \quad \forall (\ell, x, v) \in E,$$

with $x_i + v_i t(\mathbb{1}_{|\ell_i| \neq \circ}), i = 1, 2, \dots, d$ and determines the vector fields

$$\mathfrak{X}_\ell \widetilde{f}(\ell, x, v) = \sum_{i=1}^{d} \mathbb{1}_{|\ell_i| \neq \circ} v_i \partial_{x_i} f(\ell, x, v), \quad f \in C^1(\widetilde{E}).$$

Sometimes we write $\widetilde{\phi}_k(t, x, v) = \widetilde{\phi}(t, k, x, v)$ for convenience. Next, further state changes of the process are instantaneous, deterministic jumps from the boundary Γ into $\widetilde{E}$

$$\mathcal{Q}^{\mathrm{f}}(((\ell, x, v), \cdot)) = \delta_{(k(x,v),x,v)}, \quad (\ell, x, v) \in \Gamma$$

and random jumps at random times corresponding to unfreezing events

$$\mathcal{Q}^{\mathrm{s}}((\ell, x, v), \cdot) = \frac{\sum_i \lambda_i^{\mathrm{s}}(\ell, x, v) \delta_{(\ell[i\colon \ell_i'], x, v)}}{\sum_i \lambda_i^{\mathrm{s}}(i, x, v)}$$

with $\ell_i' = \circ\!\rightarrow$ if $\ell_i = \vec{\circ}$ and $\ell_i' = \leftarrow\!\circ$ if $\ell_i = \overleftarrow{\circ}$, and random reflections

$$\mathcal{Q}^{\mathrm{r}}((\ell, x, v), \cdot) = \frac{\sum_i \lambda_i^{\mathrm{r}}(\ell, x, v)\delta_x \delta_{v[i:\,-v_i]}\delta_\ell}{\sum_i \lambda_i^{\mathrm{r}}(\ell, x, v)}$$

with

$$\lambda_i^{\mathrm{s}}(\ell, x, v) = \mathbb{1}_{|\ell_i|=\circ}\kappa_i$$

and

$$\lambda_i^{\mathrm{r}}(\ell, x, v) = \mathbb{1}_{\ell_i \neq \circ}\left((v_i \partial_i \Psi(x))^+ + \lambda_{0,i}(x)\right), \quad i = 1, 2, \ldots, d.$$

Then $\lambda\colon \widetilde{E} \to \mathbb{R}^+$

$$\lambda(\ell, x, v) = \sum_{i=1}^{d} \lambda_i^{\mathrm{r}}(\ell, x, v) + \lambda_i^{\mathrm{s}}(i, x, v)$$

and a Markov kernel $\mathcal{Q}\colon (\widetilde{E} \cup \Gamma, \mathcal{B}(\widetilde{E} \cup \Gamma)) \to [0, 1]$ by

$$\mathcal{Q}((\ell, x, v), .) = \begin{cases} \frac{\sum_i \lambda_i^{\mathrm{r}}(\ell,x,v)}{\lambda(\ell,x,v)}\mathcal{Q}^{\mathrm{r}}((\ell, x, v), .) + \frac{\sum_i \lambda_i^{\mathrm{s}}(\ell,x,v)}{\lambda(\ell,x,v)}\mathcal{Q}^{\mathrm{s}}((\ell, x, v), .) & (\ell, x, v) \in \widetilde{E}, \\ \mathcal{Q}^{\mathrm{f}}((\ell, x, v), .) & (\ell, x, v) \in \Gamma. \end{cases}$$

Proposition C.1.1. *$\mathfrak{X}, \lambda, \mathcal{Q}$ satisfy the* standard conditions *given in Davis (1993, Section 24.8), namely*

- *For each $\ell \in K$, $\mathfrak{X}_\ell$ is a locally Lipschitz continuous vector field and determines the deterministic flow $\widetilde{\phi}_\ell : \widetilde{E}_\ell \to \widetilde{E}_\ell$ of the PDMP.*
- *$\lambda\colon \widetilde{E} \to \mathbb{R}^+$ is measurable and such that $t \to \lambda(\widetilde{\phi}_\ell(t, x, v))$ is integrable on $[0, \epsilon(\ell, x, v))$, for some $\epsilon > 0$, for each ℓ, x, v.*
- *$\mathcal{Q}$ is measurable and such that $\mathcal{Q}((\ell, x, v), \{(\ell, x, v)\}) = 0$*
- *The expected number of events up to time t, starting at (ℓ, x, v) is finite for each $t > 0, \forall(\ell, x, v) \in \widetilde{E}$*

To see the latter, remember that for any initial point $(\ell, x, v) \in \widetilde{E}$, the deterministic flow (without any random event) hits Γ at most d times before reaching the singleton $(0, 0, \ldots, 0)$ and being constant there.

C.1.2 Strong Markov property

Proposition C.1.2. *(Part of Theorem 4.2.2) Let $(\widetilde{Z}_t)$ be a Zig-Zag process on $\widetilde{E}$ with characteristics $\mathfrak{X}, \lambda, \mathcal{Q}$. Then $Z_t = \iota(\widetilde{Z}_t)$ is a strong Markov process.*

Proof. By Davis (1993), Theorem 26.14, the domain of the extended generator of the process $(\widetilde{Z}_t)$ with characteristics $\mathfrak{X}, \lambda, \mathcal{Q}$ is

$$\mathcal{D}(\widetilde{\mathcal{A}}) = \{f \in \mathcal{M}(\widetilde{E}); t \to f(\widetilde{\phi}_\ell(t,x,v)) \ \widetilde{\tau}\text{-absolutely continuous } \forall (\ell,x,v) \in \widetilde{E}, \\ t = [0, t_\Gamma(\ell,x,v)); \ f(\ell,x,v) = f(\kappa(x,v),x,v), \quad (\ell,x,v) \in \Gamma\},$$

with

$$t_\Gamma(\ell,x,v) = \inf\{0 \le t \colon \widetilde{\phi}_\ell(t,x,v) \in \widetilde{\Gamma}\}$$

and

$$\widetilde{\mathcal{A}} f(\ell,x,v) = \mathfrak{X}_\ell f(\ell,x,v) + \lambda(\ell,x,v) \int_{\widetilde{E}} (f(\ell',x',v') - f(\ell,x,v)) Q(\ell,x,v,\mathrm{d}(\ell,x,v)).$$

The strong Markov property of $(\widetilde{Z}_t)$ follows by Davis (1993), Theorem 25.5. Denote by $(\widetilde{P}_t)_{t\ge 0}$ the Markov transition semigroup of $(\widetilde{Z}_t)$ and let $(P_t)_{t\ge 0}$ be a family of probability kernels on E and such that for any bounded measurable function $f \colon E \to \mathbb{R}$ and any $t \ge 0$,

$$\widetilde{P}_t(f \circ \iota) = (P_t f) \circ \iota.$$

Then $(P_t)_{t\ge 0}$ is the Markov transition semigroup of the process $Z_t = (\iota(\widetilde{Z}_t))$. By Rogers and Williams (2000b), Lemma 14.1, and since any stopping time for the filtration of $(\widetilde{Z}_t)$ is a stopping time for the filtration of (Z_t), Z_t is a *strong* Markov process.

□

C.1.3 Feller property

Given an initial point $\ell, x, v \in \widetilde{E}$, let

$$t_{\Gamma_1}(\ell,x,v) = \inf\{0 \le t \colon \widetilde{\phi}_\ell(t,x,v) \in \widetilde{\Gamma}\}$$

and define the *extended deterministic flow* $\widetilde{\varphi} \colon \widetilde{E} \to \widetilde{E}$ by setting $\varphi(0,\ell,x,v) = (\ell,x,v)$ and recursively by

$$\widetilde{\varphi}(t,\ell,x,v) = \begin{cases} \widetilde{\varphi}_\ell(t,x,v) & t < t_{\Gamma_1}, \\ \widetilde{\varphi}(t - t_{\Gamma_1}, k(x',v'), x', v') & t \ge t_{\Gamma_1} \end{cases}$$

with $(\ell',x',v') = \lim_{t\to t_{\Gamma_1}} \widetilde{\varphi}_\ell(t,x,v) \in \Gamma$.

Observe that $t \to \iota(\widetilde{\varphi}(t,\ell,x,v))$ is continuous on (E,τ). Define also

$$\Lambda(t,\ell,x,v) = \int_0^t \lambda(\widetilde{\varphi}(s,\ell,x,v))\mathrm{d}s.$$

Notice that, while $(\ell, x, v) \to \lambda(\ell, x, v)$ has discontinuities at the boundaries Γ, $(\ell, x, v) \to \Lambda(\ell, x, v)$ is continuous. Denote by T_1 the first random event (so excluding the deterministic jumps). Then for functions $f \in B(\widetilde{E})$ and $\psi \in B(\mathbb{R}^+ \times \widetilde{E})$, set $z(t) = (\ell(t), x(t), v(t))$ and define

$$\widetilde{G}\psi(t, \ell, x, v) = E[f(z(t))\mathbb{1}_{t<T_1} + \psi(t - T_1, z(t))\mathbb{1}_{t\geq T_1}].$$

We have that

$$\widetilde{G}\psi(t, \ell, x, v) = f(\widetilde{\varphi}(t, \ell, x, v)) \times \mathcal{T} \tag{C.1}$$

with

$$\mathcal{T} = \sum_i \int_0^t \mathbf{1}_{t\in[t_i^\Gamma, t_{i+1}^\Gamma)} \int_{x',v'} \psi(t-s, \ell, x, v)\mathcal{Q}((\ell, \mathrm{d}x', \mathrm{d}v'), \widetilde{\varphi}(s, \ell, x, v))$$
$$\lambda(\widetilde{\varphi}(s, \ell, x, v))e^{-\Lambda(s,\ell,x,v)}\mathrm{d}s.$$

The Feller property holds if, for each fixed t and for $f \in C_b(E)$, we have that $(x, v) \to P_t f(x, v)$ is continuous (and bounded follows easily). This is what we are going to prove below, by making a detour in the space $\widetilde{E}$, using the bijection ι and adapting some results found in Davis (1993, Section 27), for the process $\widetilde{Z}_t$.

Theorem C.1.3. *(Part of Theorem 4.2.2) Z_t is a Feller process.*

Proof. Take $f \in C_b(\widetilde{E})$ such that $f \circ \iota \in C_b(E)$. Call those functions on $\widetilde{E}$ τ-continuous. We want to show that $\widetilde{P}$ preserves τ-continuity. Notice that τ-continuous functions on $\widetilde{E}$ are such that

$$\lim_{t\to t_\Gamma} f(\widetilde{\varphi}(t, \ell, x, v)) = f(\widetilde{\varphi}(t_\Gamma, \ell, x, v))), \quad (\ell, x, v) \in \widetilde{E}.$$

For τ-continuous functions f and for a fixed t, the first term on the right hand side of (C.1) $(\ell, x, v) \to f(\widetilde{\varphi}(t, \ell, x, v))$ is clearly continuous. Also the second term is continuous since is of the form of an integral of a piecewise continuous function. Therefore, for any $t \geq 0$, $\psi(t, \cdot) \in B(\widetilde{E})$ and τ-continuous function f, we have that $(\ell, x, v) \to \widetilde{G}\psi(t, \ell, x, v)$ is continuous. Clearly, the (similar) operator

$$\widetilde{G}_n\psi_\ell(t, x, v) = E_x[f(\widetilde{\varphi}_\ell(t, x, v))\mathbb{1}_{t<T_n} + \psi(t - T_n, \widetilde{\varphi}_\ell(t, x, v))\mathbb{1}_{t\geq T_n}],$$

with T_n denoting the nth random time, is continuous as well for any fixed n, t, $\psi(t, \cdot) \in B(\widetilde{E})$ and τ-continuous function f. By applying Lemma 27.3 in Davis (1993) we have that for any $\psi(t, \cdot) \in B(\widetilde{E})$

$$|\widetilde{G}_n\psi_\ell(t, x, v) - \widetilde{P}_t f(x, v)| \leq 2\max(\|\psi\|\|f\|)P(t \geq T_n).$$

Finally, if λ is bounded, then we can bound $P(t \geq T_n)$ by something which does not depend on (ℓ, x, v) and goes to 0 as $n \to \infty$ so that $\widetilde{G}_n\psi \to \widetilde{P}_t f$ uniformly on $\ell, x, v \in \widetilde{E}$ under the supremum norm. This shows that, for any t, $\widetilde{P}_t$ (and therefore P_t) preserves τ-continuity. □

Remark C.1.4. *The proof of the Feller and Markov property follow similarly for the Bouncy Particle and the Boomerang sampler.*

C.1.4 The extended generator of Z_t

Let $f \in \mathcal{D}(\mathcal{A})$ if $\widetilde{f} \in \mathcal{D}(\widetilde{\mathcal{A}})$ and $f \circ \iota = \widetilde{f}$. Then $f \in \mathcal{D}(\mathcal{A})$ are τ-absolutely continuous functions along full deterministic trajectories on E:

$$\begin{aligned}\mathcal{D}(\mathcal{A}) = \{f \in \mathcal{M}(E); t \to f(\varphi(t,x,v)) \ \tau\text{-absolutely continuous } \forall (x,v);\\ \lim_{t\to 0} f(x[i\colon 0^+ + t], v) = f(x[i\colon 0^+], v);\\ \lim_{t\to 0} f(x[i\colon 0^- - t], v) = f(x[i\colon 0^-], v)\}.\end{aligned}$$

For those functions $f \in \mathcal{D}(\mathcal{A})$ with $f \circ \iota = \widetilde{f}$ we have that

$$\widetilde{\mathcal{A}}\widetilde{f}(\ell, \widetilde{x}, v) = \mathcal{A}f(x,v) = \sum_{i=1}^{N} \mathcal{A}_i f(x,v)$$

with

$$\mathcal{A}_i f(x,v) = \begin{cases} \kappa_i(f(T_i(x,v)) - f(x,v)) & (x,v) \in \mathfrak{F}_i, \\ v_i \partial_{x_i} f(x,v) + \lambda_i(x,v)(f(x, v[i\colon -v_i]) - f(x,v)), & \text{otherwise,} \end{cases}$$

and

$$\lambda_i(x,v) = (v_i \partial_i \Psi(x))^+ + \lambda_{0,i}(x), \quad i = 1, 2, \ldots, d,$$

for positive functions $\lambda_{0,i}$.

Denote the space of compactly supported functions on E which are continuously differentiable in their first argument by $C_c^1(E)$. Define $C_b(E) = \{f \in C(E)\colon f \text{ is bounded}\}$ and $D = \{f \in C_c^1(E), \mathcal{A}f \in C_b(E)\}$. The following proposition shows that the operator $\mathcal{A}$ restricted to D coincides with the infinitesimal generator of the ordinary Zig-Zag process restricted to D.

Proposition C.1.5. *We have*

$$\begin{aligned}D = \{f \in C_c^1(E)\colon v_i\kappa_i\left(f(T_i(x,v)) - f(x,v)\right)\\ = v_i\partial_i f(x,v) + \lambda_i(x,v)(f(x, v[i\colon -v_i])) - f(x,v)), (x,v) \in \mathfrak{F}_i \text{ for all } i = 1, \ldots, d\}.\end{aligned}$$

For $f \in D$, $\mathcal{A}f = \mathcal{L}f$, where $\mathcal{L}f = \sum_{i=1}^d \mathcal{L}_i f$ with

$$\mathcal{L}_i f(x,v) = v_i \partial_{x_i} f(x,v) + \lambda_i(x,v)\left(f(x, v[i\colon -v_i]) - f(x,v)\right).$$

Proposition C.1.6. *(Proposition 4.2.1) The extended generator of the process $(Z(t))$ is given by $\mathcal{A}$ with domain $\mathcal{D}(\mathcal{A})$.*

Proof. This is to verify that if $f \in \mathcal{D}(\widetilde{\mathcal{A}})$ and $\widetilde{\mathcal{A}}$ solve the martingale problem, i.e are such that

$$f(\ell(t), x(t), v(t)) - f(\ell, x, v) + \int_0^t \mathcal{A}f(\ell(s), x(s), v(s)\mathrm{d}s, \quad \forall (\ell, x, v) \in \widetilde{E}$$

is a local martingale (Davis 1993, Section 24) on $\widetilde{E}$, then $f \circ \iota \colon f \in \mathcal{D}(\widetilde{\mathcal{A}})$ and $\mathcal{A}$ solve the martingale problem on E (for any local martingale Z_t on $\widetilde{E}$, $\iota(Z_t)$ is a local martingale on E). □

By the Feller property, the extended generator is an extension of the generator defined as

$$\mathcal{L}f(x, v) := \lim_{t \downarrow 0} \frac{\mathbb{E}[f(X_t, V_t) \mid X_0 = x, V_0 = v] - f(x, v)}{t}$$

for a sufficient regular class of functions f for which this limit exists uniformly in x (see Liggett 2010, Section 3, for more details). Then, $D = \{f \in \mathcal{D}(\mathcal{A}) \colon f \in C_b^1, \mathcal{A}f \in C_b(E)\}$ is a core for $\mathcal{A}$ (as in Liggett 2010, Definition 3.31). Let $\mathcal{L}$ be the restriction of $\mathcal{A}$ on D. By Liggett (2010, Theorem 3.37), μ is a stationary measure if, for all $f \in D$:

$$\int \mathcal{L}f \mathrm{d}\mu = 0.$$

C.1.5 Remaining part of the proof

Invariant measure of the Sticky Zig-Zag process: We check here that the sticky d-dimensional Zig-Zag process as presented in Section 4.2.3 taking values in E with discrete velocities in $\mathcal{V} = \{v \colon |v_i| = a_i, \forall i \in \{1, 2, \ldots, d\}\}$ and with extended generator $\mathcal{A}$ is such that

$$\int \mathcal{L}f(x, v)\mu(\mathrm{d}x, \mathrm{d}v) = 0$$

for all $f \in D = \{f \in C_c^1(E), \mathcal{A}f \in C_b(E)\}$. Here, $\mathcal{L}$ is the extended generator $\mathcal{A}$ restricted to D (See Poposition (C.1.5)). For any-1 $f \in D$, define $\lambda_i^+ := \lambda_i(x, v[i \colon, a_i])$, $\lambda_i^- := \lambda_i(x, v[i \colon, -a_i])$, $f_i^+ := f(x, v[i \colon a_i])$, $f_i^- := f(x, v[i \colon -a_i])$, $f_i^+(y) := f(x[i \colon y], v[i \colon a_i])$, $f_i^-(y) := f(x[i \colon y], v[i \colon -a_i])$, . Also write the measure $\rho(\mathrm{d}x_i, v_i) := \mathrm{d}x_i + \frac{1}{\kappa}\left(\mathbb{1}_{v_i<0}\delta_0^+(\mathrm{d}x_i) + \mathbb{1}_{v_i>0}\delta_0^-(\mathrm{d}x_i)\right)$. We see that

$$
\begin{aligned}
\int \mathcal{L}_i f \mathrm{d}\mu = & \\
& \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} \left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-} \right) \left(a_i \partial_{x_i} f_i^+ + \lambda_i^+ (f_i^- - f_i^+) \right) \exp(-\Psi(x)) \mathrm{d}x_i \prod_{j\neq i} \rho(\mathrm{d}x_j, v_j) \right) \\
& + \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} \left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-} \right) \left(-a_i \partial_{x_i} f_i^- + \lambda_i^- (f_i^+ - f_i^-) \right) \exp(-\Psi(x)) \mathrm{d}x_i \prod_{j\neq i} \rho(\mathrm{d}x_j, v_j) \right) \\
& + \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} a_i \left(f_i^+(0^+) - f_i^+(0^-) \right) \exp(-\Psi(x[i\colon 0])) \prod_{j\neq i} \rho(\mathrm{d}x_j, v_j) \right) \\
& + \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} -a_i \left(f_i^-(0^-) - f_i^-(0^+) \right) \exp(-\Psi(x[i\colon 0])) \prod_{j\neq i} \rho(\mathrm{d}x_j, v_j) \right).
\end{aligned}
$$

By integration by parts we have that $\left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-} \right) \left(\partial_{x_i} f(x,v) \exp(-\Psi(x)) \right) \mathrm{d}x_i$ is equal to

$$
\begin{aligned}
& \left(f(x[i\colon 0^-], v) - f(x[i\colon 0^+], v) \right) \exp(-\Psi(x[i\colon 0])) \\
& \qquad + \left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-} \right) \left(\partial_i \Psi(x) f(x,v) \exp(-\Psi(x)) \right) \mathrm{d}x_i
\end{aligned}
$$

so that $\int \mathcal{L}_i f \mathrm{d}\mu$ is equal to

$$
\begin{aligned}
& \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} \left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-} \right) \left(a_i \partial_{x_i} \Psi(x) + \lambda_i^+ - \lambda_i^- \right) f_i^- \exp(-\Psi(x)) \mathrm{d}x_i \prod_{j\neq i} \rho(\mathrm{d}x_j, v_j) \right) \\
& + \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} \left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-} \right) \left(-a_i \partial_{x_i} \Psi(x) + \lambda_i^- - \lambda_i^+ \right) f_i^+ \exp(-\Psi(x)) \mathrm{d}x_i \prod_{j\neq i} \rho(\mathrm{d}x_j, v_j) \right) \\
& + \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} a_i \left(f_i^+(0^+) - f_i^+(0^-) \right) \exp(-\Psi(x[i\colon 0])) \prod_{j\neq i} \rho(\mathrm{d}x_j, v_j) \right) \\
& + \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} -a_i \left(f_i^-(0^-) - f_i^-(0^+) \right) \exp(-\Psi(x[i\colon 0])) \prod_{j\neq i} \rho(\mathrm{d}x_j, v_j) \right) \\
& + \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} a_i \left(f_i^+(0^-) - f_i^+(0^+) \right) \exp(-\Psi(x[i\colon 0])) \prod_{j\neq i} \rho(\mathrm{d}x_j, v_j) \right) \\
& + \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} -a_i \left(f_i^-(0^+) - f_i^-(0^-) \right) \exp(-\Psi(x[i\colon 0])) \prod_{j\neq i} \rho(\mathrm{d}x_j, v_j) \right) = 0,
\end{aligned}
$$

where we used that $-v_i\partial_i\Psi(x) + \lambda_i(x,v) - \lambda_i(x, F_i(v)) = 0$, $\forall (x,v) \in E$.

C.1.6 Ergodicity of the sticky Zig-Zag process

In this section, we prove that the sticky Zig-Zag is ergodic. As the argument partially relies on the ergodicity results of the ordinary Zig-Zag sampler (Bierkens, Roberts, and Zitt 2019), we start by making similar assumptions on Ψ as appearing in that paper.

Assumption C.1.2. (Assumptions of Bierkens, Roberts, and Zitt 2019, Theorem 1) *Let Ψ satisfy the following conditions:*

- $\Psi \in \mathcal{C}^3(\mathbb{R}^d)$,
- Ψ *has a non degenerate local-minimum,*
- *For some constants $c > d$, $c' \in \mathbb{R}$, $\Psi(x) > c\ln(|x|) - c'$, for all $x \in \mathbb{R}^d$.*

For every set $\alpha \subset \{1, 2, \dots, d\}$, we define the sub-space $\mathcal{M}_\alpha = \{x \in \mathbb{R}^d \colon x_i = 0,\, i \notin \alpha\}$ and define the $|\alpha|$-dimensional ordinary Zig-Zag process $(Z_t^{(\alpha)})_{t\geq 0}$, with $|\alpha| \leq d$, on the sub-space $\mathcal{M}_\alpha \times \{-1,+1\}^\alpha$ and with reflection rates $\lambda_i(x,v) = \max(0, v_i\partial_i\Psi(x))$, $x \in \mathcal{M}_\alpha$, $i \in \alpha$.

Proposition C.1.7. *Suppose Ψ satisfies Assumption C.1.2. Then for every set $\alpha \subset \{1, 2, \dots, d\}$, $(Z_t^{(\alpha)})_{t\geq 0}$ is ergodic with unique invariant measure with density $\exp(-\Psi(x))|_{\mathcal{M}_\alpha}$ relative to Leb$(\mathcal{M}_\alpha)(\mathrm{d}x) \otimes$ Uniform$(\{-1,+1\}^\alpha)(\mathrm{d}v)$. Furthermore, some skeleton chain of each process is irreducible.*

Proof. If Assumption C.1.2 holds on $\mathbb{R}^d$, then it holds on any the sub-space $\mathcal{M}_\alpha$, $\alpha \subset \{1, 2, \dots, d\}$, for functions $x \mapsto \Psi(x)$, $x \in \mathcal{M}_\alpha$. Proposition C.1.7 follows from the ergodic theorem of ordinary Zig-Zag processes (Bierkens, Roberts, and Zitt 2019, Theorem 1 and Theorem 5). □

Next, we show that, for any initial position $(x,v) \in E$, the sticky Zig-Zag process is Harris recurrent to the set where all coordinates are stuck at 0. Denote the measure $\overline{\delta}_0(\mathrm{d}x, \mathrm{d}v) = \bigotimes_{i=1}^d (\delta_{0^+,-1}(\mathrm{d}x_i, \mathrm{d}v_i) + \delta_{0^-,+1}(\mathrm{d}x_i, \mathrm{d}v_i))$, the set $\mathfrak{S} = \cap_{i=1}^d \mathfrak{F}_i$ and the first hitting time $\tau_A = \inf\{t > 0 \colon Z_t \in A\}$, where $Z_t = (X_t, V_t)$ is the sticky Zig-Zag process.

Proposition C.1.8. (Harris recurrence) *Suppose Ψ satisfies Assumption C.1.2. Then for any initial state $Z_0 = z_0 \in E$, we have that $\mathbb{P}(\tau_S < \infty) = 1$, for any $S \subset \mathfrak{S}$.*

Proof. Let $x_0 \in \mathcal{M}_\alpha$ for an arbitrary $\alpha \subset \{1, 2, \dots, d\}$. Denote the random time of the first stuck coordinate x_i, $i \in \alpha^c$ leaving zero by $T_1 \sim \text{Exp}(\sum_{j\in\alpha^c} \kappa_j) > 0$. Denote the random time of the first 'free' coordinate x_i, $i \in \alpha$ hitting zero by T_2.

Notice that T_1 is independent of the trajectory on the subspace $\mathcal{M}_\alpha$. and the sticky Zig-Zag process behaves as an ordinary $|\alpha|$-dimensional Zig-Zag process in the subspace $\mathcal{M}_\alpha$ for time $t \in [0, \min(T_2, T_1)]$. By Proposition C.1.7, T_2 is finite and $\mathbb{P}(T_2 < T_1) > 0$. By using the Markov structure of the process and iterating the same argument for a sequence of sub-models $\mathcal{M}_{\alpha_2}$, $\mathcal{M}_{\alpha_3}$, $\dots$, $\mathcal{M}_{\alpha_{|\alpha|-1}}$, with $|\alpha_j|+1 = |\alpha_{j+1}|$, we conclude that $P(\tau_{\mathfrak{S}} < \infty) = 1$.

Now, consider a subset $S \subset \mathfrak{S}$ and a random element from S. Without loss of generality, we may assume this element to be $s_0 = ((0^-, \dots, 0^-), (+1, \dots, +1))$. Next, we show that $\mathbb{P}(\tau_{\{s_0\}} < \infty) = 1$.

Let $\tau_{\mathfrak{S}}$ be the hitting time to the set $\mathfrak{S}$ of the sticky Zig-Zag $Z(t)_{t>0}$. Denote by $\beta := \{i\colon Z_i(\tau_{\mathfrak{S}}) \neq [s_0]_i\} \subset \{1, 2, \dots, d\}$ the set of indices for which the coordinate $Z_i(\tau_{\mathfrak{S}})$ is stuck on the other copy of zero. At time $Z(\tau_{\mathfrak{S}})$ the process will stay in the null model for a time $\Delta T \sim \text{Exp}(\sum_{j=1}^d \kappa_j)$. At time $T + \Delta T$ a coordinate $i \in \beta$ is released with positive probability $\kappa_i / \sum_j \kappa_j$. Conditional on ΔT and on the event that the coordinate i is released at time $T + \Delta T$, the sticky Zig-Zag behaves as a 1 dimensional ordinary Zig-Zag sampler until time $\tau_{\mathfrak{S}} + \Delta T + \min(\Delta T_1, \Delta T_2)$, where, similarly as before, $\Delta T_1 \sim \text{Exp}(\sum_{j\neq i} \kappa_j)$ (and it is independent from the trajectory of the free coordinate) and ΔT_2 is the hitting time to 0 of the coordinate process $Z_i(\tau_{\mathfrak{S}} + \Delta T + t)_{t>0}$. By Proposition C.1.7, ΔT_2 is finite and $\mathbb{P}(\Delta T_2 < \Delta T_1) > 0$. By using the Markovian structure of the process and iterating this argument for all $i \in \beta$ we conclude that $\mathbb{P}(\tau_{\{s_0\}} < \infty) = 1$, hence $\mathbb{P}(\tau_S < \infty) = 1$. □

By Meyn and Tweedie (1993, Theorem 6.1), the sticky Zig-Zag sampler is ergodic if it is Harris recurrent with invariant probability μ and if some skeleton of the chain is irreducible. For the latter condition, notice that once the process reaches the null model, it will stay there for a time $\Delta T \sim \text{Exp}(\sum_{j=1}^d \kappa_j)$ and $\mathbb{P}(\Delta T > \Delta) > 0$ for any $\Delta > 0$. Hence, any skeleton $Z^{(\Delta)} = (Z(0)), Z(\Delta), Z(2\Delta), \dots)$ (with $\Delta > 0$) is irreducible relative to the measure $\overline{\delta}_0$.

C.1.7 Recurrence time of the sticky Zig-Zag to 0

The recurrent time to the point $\mathbf{0} = (0, 0, \dots, 0)$ is derived with a simple heuristic argument. We assume the sticky Zig-Zag to have unit velocity components and to be ergodic with stationary measure μ. Clearly, the expected time to leave $\mathbf{0}$ is $(\kappa d)^{-1}$ since each coordinate leaves 0 according to an independent exponential random variable with parameter κ. Denote by τ_0 the recurrent time to 0, i.e. the random time spent outside $\mathbf{0}$ before returning to $\mathbf{0}$. By ergodicity, the expectation

of τ_0 must satisfy the following equation

$$\frac{(\kappa d)^{-1}}{\mu(\{\mathbf{0}\})} = \frac{\mathbb{E}[\tau_0]}{1-\mu(\{\mathbf{0}\})}.$$

C.2 Other sticky PDMP samplers

Here we extend the results presented in Section 4.2.3 for two other Sticky PDMP samplers: the sticky version of the Bouncy particle sampler (Bouchard-Côté, Vollmer, and Doucet 2018) and the Boomerang sampler (Bierkens et al. 2020), the latter having Hamiltonian deterministic dynamics invariant to a prescribed Gaussian measure. To visually assess the difference in sample paths, we show in Figure C.1 a typical realization of the Sticky Zig-Zag sampler, Sticky Bouncy particle sampler and Sticky Boomerang sampler.

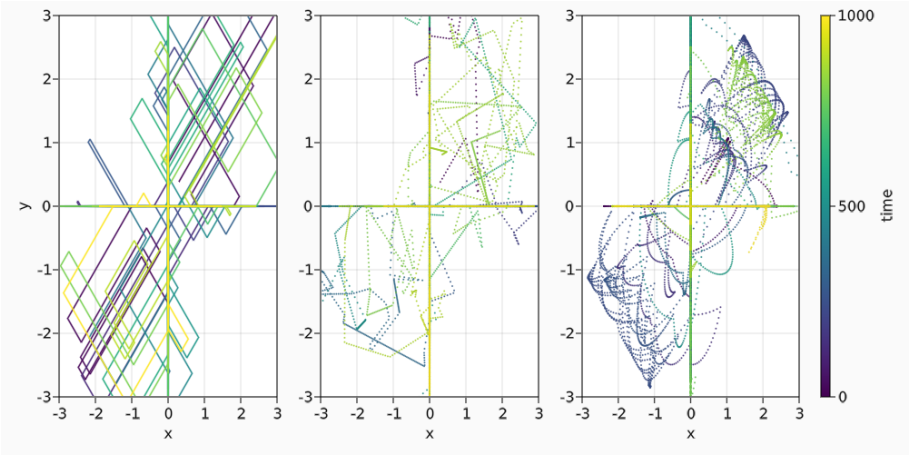


Figure C.1: (x-y) phase portraits, of 3 different sticky PDMP samplers targeting the measure of Equation (4.2) with $\exp(-\Psi)$ being a mixture of two bivariate Gaussian densities centred respectively in the first and the third quadrant of the x-y axes. Left: Sticky Zig-Zag sampler. Middle: sticky Bouncy Particle sampler with refreshment rate equal to 0.1. Right: sticky Boomerang sampler with refreshment rate equal to 0.1. For all the samplers, $\kappa_1 = \kappa_2 = 0.1$ and the final clock was set to $T = 10^3$. As the sticky Bouncy Particle sampler and the Boomerang sampler don't have constant speed, we marked their continuous trajectories in the phase plots with dots. The distance of dots indicates the speed of traversal.

C.2.1 Sticky Bouncy Particle sampler

The inner product and the norm operator in the subspace determined by A is denoted by $\langle x, v\rangle_A := \sum_{i\in A} x_i v_i$ and $\|x\|_A := \sum_{i\in A} x_i^2$ with the convention that $\langle\cdot,\cdot\rangle_{\{1,2,\ldots,d\}} = \langle\cdot,\cdot\rangle$ and $\|\cdot\|_{\{1,2,\ldots,d\}} = \|\cdot\|$. The deterministic dynamics of the sticky Bouncy Particle process are identical to that of the Sticky Zig-Zag process, having piecewise constant velocity. For each $i \in \{1, 2, \ldots, d\}$, when the process hits a state $(x, v) \in \mathfrak{F}_i$, the ith coordinate (x_i, v_i) sticks for an exponentially distributed time with rate equal to

$\kappa_i|v_i|$ while the other coordinates continue their flow until a reflection or refreshment event happens. A reflection occurs with an inhomogeneous rate equal to

$$\lambda(x,v) = \max(0, \langle v, \nabla\Psi(x)\rangle_\alpha),$$

where α is as defined in Equation (4.4). At reflection time the process jumps with a contour reflection of the active velocities with respect to $\nabla\Psi$:

$$(R_\Psi(x,v)v)_i = \begin{cases} v_i & i \notin \alpha(x,v) \\ v_i - 2\frac{\langle\nabla\Psi(x),v\rangle_\alpha}{\|\nabla\Psi(x))\|_\alpha^2}\partial_i\Psi(x) & \text{else.} \end{cases}$$

Similarly to the ordinary Bouncy Particle sampler, the sticky Bouncy Particle sampler refreshes its velocity component at exponentially distributed times with homogeneous rate equal to λ_{ref}. This is necessary for avoiding pathological behaviour of the process (see Bouchard-Côté, Vollmer, and Doucet 2018). At refreshment times, each coordinate renews its velocity component independently according to the following refreshment rule

$$v_i' \sim \begin{cases} Z_i & (x,v) \notin \mathfrak{F}_i, \\ \text{sign}(v_i)|Z_i| & (x,v) \in \mathfrak{F}_i, \end{cases} \tag{C.2}$$

where $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, independently of all random quantities. The refreshment rule coincides with the refreshment rule given in the ordinary Bouncy Particle sampler algorithm Bouchard-Côté, Vollmer, and Doucet (2018) for the coordinates whose index is in the set α. For the components which are stuck at 0, the refreshment rule renews the velocity without changing its sign. This prevents the possibility for the ith stuck component to jump out the set $\mathfrak{F}_i$ (changing its label from frozen to active at refreshment time).

The extended generator of the sticky Bouncy Particle sampler is given by

$$\mathcal{A}f(x,v) = \sum_{i=1}^d \mathcal{G}_i f(x,v) + \lambda(x,v)(f(x, R_\Psi(x,v)v) - f(x,v)) \\ + \lambda_{\text{ref}} \int (f(x,w) - f(x,v))\, \varrho_{x,v}(w)\mathrm{d}w$$

and

$$\mathcal{G}_i f(x,v) = \begin{cases} |v_i|\kappa_i \left(f(T_i(x,v)) - f(x,v)\right) & (x,v) \in \mathfrak{F}_i \\ v_i \partial_{x_i} f(x,v) & \text{else,} \end{cases}$$

where

$$\varrho_{x,v}(w) = \rho(w_{\alpha(x,v)}) \prod_{i \in \alpha(x,v)^c} 2\rho(w_i)\mathbb{1}_{v_i w_i > 0},$$

for sufficient regular functions $f\colon E \to \mathbb{R}$ in the extended domain of the generator. Here, $\rho(y)$ is the standard normal density function evaluated at y.

Proposition C.2.1. *The d-dimensional sticky Bouncy Particle sampler is invariant to the measure*

$$\mu(\mathrm{d}x, \mathrm{d}v) = \frac{1}{C}\rho(v)\mathrm{d}v \exp(-\Psi(x)) \prod_{i=1}^{d} \left(\mathrm{d}x_i + \frac{1}{\kappa_i} \left(\mathbb{1}_{v_i>0}\delta_{0^-}(\mathrm{d}x_i) + \mathbb{1}_{v_i<0}\delta_{0^+}(\mathrm{d}x_i) \right) \right) \tag{C.3}$$

for some normalization constant C.

Proof. The transition kernel $R_\Psi(x)$ satisfies the following properties:

$$\langle \nabla\Psi(x), R_\Psi(x,v)v \rangle_\alpha = -\langle \nabla\Psi(x), v \rangle_\alpha$$

and

$$\|R_\Psi(x,v)v\|^2 = \|v\|^2_{\alpha^c} + \|R_\Psi(x,v)v\|^2_\alpha = \|v\|^2_{\alpha^c} + \|v\|^2_\alpha = \|v\|^2$$

so, $\rho(R^A_\Psi(x)v) = \rho(v)$ ($\rho(x)$ here denotes the standard Gaussian density evaluated at x). Furthermore λ satisfies

$$-\langle v, \nabla\Psi(x) \rangle_\alpha + \lambda(x,v) - \lambda(x, R_\Psi(x,v)v) = 0, \quad \forall (x,v) \in E. \tag{C.4}$$

Let us check that the process satisfies $\int \mathcal{L}f(x,v)\mu(\mathrm{d}x, \mathrm{d}v) = 0$, for all $f \in D = \{f \in C^1_c(E), \mathcal{A}f \in C_b(E)\}$ where $\mathcal{L}$ is the extended generator $\mathcal{A}$ restricted to D.

First let us fix some notation: denote $f_i(y) = f(x[i\colon y], v)$, $Rf(x,v) = f(x, R_\Psi(x,v)v)$ and $R\lambda(x,v) = \lambda(x, R_\Psi(x,v)v)$. Also write $\delta_0(\mathrm{d}x_i, v_i) := \mathbb{1}_{v_i<0}\delta_{0^+}(\mathrm{d}x_i) + \mathbb{1}_{v_i>0}\delta_{0^-}(\mathrm{d}x_i)$ and $\Delta_i f(x,v) := f(x[i\colon 0^+], v) - f(x[i\colon 0^-], v))$. We

have this preliminary result:

$$\begin{aligned}
\int \sum_{i=1}^{d} \mathcal{G}_i f \mathrm{d}\mu &= \frac{1}{C} \sum_i \int \left(\mathcal{G}_i f \exp(-\Psi(x))(\mathrm{d}x_i + \frac{1}{\kappa_i}\delta_0(\mathrm{d}x_i)) \right) \\
&\qquad \prod_{j \neq i} \left(\mathrm{d}x_j + \frac{1}{\kappa_j}\delta_0(\mathrm{d}x_j, v_j) \right) \rho(v)\mathrm{d}v \\
&= \frac{1}{C} \sum_i \int (v_i \partial_{x_i} f \exp(-\Psi(x))\mathrm{d}x_i + v_i \Delta_i f \exp(-\Psi(x))\delta_0(\mathrm{d}x_i)) \\
&\qquad \prod_{j \neq i} \left(\mathrm{d}x_j + \frac{1}{\kappa_j}\delta_0(\mathrm{d}x_j, v_j) \right) \rho(v)\mathrm{d}v \qquad \text{(C.5)} \\
&= \frac{1}{C} \sum_i \int (v_i \partial_{x_i} \Psi(x) f(x,v) \exp(-\Psi(x))\mathrm{d}x_i) \prod_{j \neq i} \left(\mathrm{d}x_j + \frac{1}{\kappa_j}\delta_0(\mathrm{d}x_j, v_j) \right) \rho(v)\mathrm{d}v \qquad \text{(C.6)} \\
&= \frac{1}{C} \sum_{A \subset \{1,\dots,d\}} \left(\sum_{i \in A} \left(\int v_i \partial_{x_i} \Psi(x) f(x,v) \exp(-\Psi(x))\mathrm{d}x_A \right) \prod_{j \in A^c,} \frac{1}{\kappa_j}\delta_0(\mathrm{d}x_j, v_j) \right) \qquad \text{(C.7)} \\
&= \frac{1}{C} \sum_{A \subset \{1\dots,d\}} \int \langle v, \nabla \Psi(x[A^c\colon 0]) \rangle_A f(x[A^c\colon 0], v) \exp(-\Psi(x[A^c\colon 0]))\mathrm{d}x_A \prod_{j \in A^c} \frac{1}{\kappa_j} \rho(v)\mathrm{d}v
\end{aligned}$$

Here from (C.5) to (C.6) we used integration by parts in the two half planes $(\infty, 0^+]$ and $[0^-, -\infty)$. For the equivalence of (C.6) to (C.7) note that placing $|A|$ balls in d numbered boxes and marking one of them (say the ball in box i) is equivalent to placing a marked ball in box i and distributing the remaining unmarked balls over the remaining boxes. Also notice that

$$\begin{aligned}
\int \lambda_{\mathrm{ref}} \int (f(x,w) - f(x,v)) \varrho(w)\mathrm{d}w\mathrm{d}\mu = \\
\frac{1}{C} \sum_{A \subset \{1,2,\dots,d\}} \lambda_{ref} \int (f(x,w) - f(x,v)) \exp(-\Psi(x))\mathrm{d}x_A \\
\times \prod_{i \in A^c} \frac{1}{\kappa_i} \delta_{0^-}(\mathrm{d}x_i) \mathbb{1}_{v_i > 0} \mathbb{1}_{w_i > 0} 2^{|A^c|} \rho(v)\rho(w)\mathrm{d}v\mathrm{d}w \\
+ \frac{1}{C} \sum_{A \subset \{1,2,\dots,d\}} \lambda_{\mathrm{ref}} \int (f(x,w) - f(x,v)) \exp(-\Psi(x))\mathrm{d}x_A \\
\times \prod_{i \in A^c} \frac{1}{\kappa_i} \delta_{0^+}(\mathrm{d}x_i) \mathbb{1}_{v_i < 0} \mathbb{1}_{w_i < 0} 2^{|A^c|} \rho(v)\rho(w)\mathrm{d}v\mathrm{d}w,
\end{aligned}$$

which is equal to 0 by symmetry between v and w. Then

$$\begin{aligned}
\int \mathcal{L}f \mathrm{d}\mu = \frac{1}{C} \sum_{A\subset\{1\dots,d\}} \int \langle v, \nabla\Psi(x[A^c\colon 0])\rangle_A \exp(-\Psi(x[A^c\colon 0])) f(x[A^c\colon 0], v)\mathrm{d}x_A \\
\prod_{j\in A^c} \frac{1}{\kappa_j}\rho(v)\mathrm{d}v + \int (\lambda(x, R_\Psi(x,v)) - \lambda(x,v)) f(x,v)\mu(\mathrm{d}x, \mathrm{d}v) \\
= \frac{1}{C} \sum_{A\subset\{1\dots,d\}} \int \langle v, \nabla\Psi(x[A^c : 0])\rangle_A \exp(-\Psi(x[A^c\colon 0])) \qquad \text{(C.8)} \\
f(x[A^c\colon 0], v)\mathrm{d}x_A \prod_{j\in A^c} \frac{1}{\kappa_j}\rho(v)\mathrm{d}v \\
+ \frac{1}{C} \sum_{A\subset\{1,\dots,d\}} \int \left(\lambda(x[A^c\colon 0], R_\Psi v) - \lambda(x[A^c\colon 0], v)\right) f(x[A^c\colon 0], v) \qquad \text{(C.9)} \\
\exp(-\Psi(x[A^c\colon 0]))\mathrm{d}x_A \prod_{j\in A^c} \frac{1}{\kappa_j}\rho(v)\mathrm{d}v \\
= 0,
\end{aligned}$$

where in Equation (C.8)-(C.9) we used a change of variable $v' = R_\Psi(x,v)v$ and property (C.4). □

Remark C.2.2. *In more generality, the transition kernel at refreshment times can be chosen as follows: with two refreshment transition densities q^A and q^F such that $q^A(w_A \mid v_A)\rho(v_A)$ and $q^F(w_F \mid v_F)\rho(v_F)$ for each $A \sqcup F = \{1,\dots,d\}$ are symmetric densities in w, v, the refreshment kernel*

$$\varrho_{x,v}(dy, dw) = q^A(w_{\alpha(x,v)} \mid w_{\alpha(x,v)}) q^F(w_{\alpha^c(x,v)} \mid w_{\alpha^c(c,v)}) \delta_{\mathcal{F}(x,v,w)}(\mathrm{d}y)\mathrm{d}w$$

where

$$(\mathcal{F}(x,v,w))_i = \begin{cases} 0^- & \text{if } x_i = 0^+, v_i < 0, w_i > 0, \\ 0^+ & \text{if } x_i = 0^-, v_i > 0, w_i < 0, \\ x_i & \text{else} \end{cases}$$

leaves the target measure μ invariant.

The transition kernels given in Remark C.2.2 satisfy the Equation $\lambda_{\mathrm{ref}} \int f(x,w) - x(x,v)\varrho_{x,w}\mathrm{d}w\mathrm{d}\mu = 0$ and therefore, by similar computations as in the proof of Proposition C.2.1, leave μ invariant. For example, the preconditioned Crank-Nicolson scheme Cotter et al. (2013) falls withing this setting.

C.2.2 Sticky Boomerang sampler

The sticky Boomerang sampler has Hamiltonian dynamics prescribed by the vector field $\bar{\xi}_i(x_i, v_i) = (v_i, -x_i)$ with close-form solution

$$(x_i(t), v_i(t)) = (\cos(t)x_i(0) + \sin(t)v_i(0), -x_i(0)\sin(t) + \cos(t)v_i(0)), \tag{C.10}$$

and is invariant to a prescribed Gaussian measure centered in 0. Define $U(x)$ such that

$$U(x) = \Psi(x) - \frac{1}{2}x'\Sigma^{-1}x$$

for a positive semi-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$. Consider for example the application in Bayesian inference with spike-and-slab prior (Equation (4.1)) where $\{\pi_i\}_{i=1}^d$ are centered Gaussian densities with variance σ_i^2. Then a natural choice is $\Sigma = \text{Diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$.

Similarly to the sticky Bouncy Particle sampler, the process reflects its velocity at an inhomogeneous rate given by

$$\lambda(x, v) = \langle v, \nabla U(x)\rangle_\alpha^+$$

with reflection specified by the transition kernel

$$(R_U(x, v)v)_i = \begin{cases} v_i & i \notin \alpha \\ v_i - 2\frac{\langle \nabla U(x), v\rangle_\alpha}{\|\nabla\Sigma^{1/2}U(x)),\|_\alpha^2}\langle \Sigma_{[i,:]}, \nabla U(x)\rangle_\alpha & \text{else} \end{cases}$$

and refreshes the velocity at exponentially distributed times with rate equal to λ_{ref} according to the rule given in Equation (C.2).

Proposition C.2.3. *The d-dimensional sticky Boomerang sampler is invariant to the measure in Equation* (C.3).

Proof. The extended generator of the sticky d-dimensional Boomerang process is given by

$$\mathcal{A}f(x, v) = \sum_{i=1}^d \mathcal{G}_i f(x, v) + \lambda(x, v)(f(x, R_U(x, v)v) - f(x, v)) + \lambda_{\text{ref}} \int (f(x, w) - f(x, v))\, \varrho_{x,v}(w)\mathrm{d}w$$

and

$$\mathcal{G}_i f(x, v) = \begin{cases} |v_i|\kappa_i \left(f(T_i(x, v)) - f(x, v)\right) & (x, v) \in \mathfrak{F}_i \\ v_i\partial_{x_i} f(x, v) + x_i\partial_{v_i} f(x, v) & \text{else,} \end{cases}$$

where

$$\varrho_{x,v}(w) = \rho(w_{\alpha(x,v)}) \prod_{i \in \alpha(x,v)^c} 2\rho(w_i)\mathbb{1}_{v_i w_i > 0},$$

$\rho(y)$ being the standard normal density function evaluated at y and for sufficient regular functions $f\colon E \to \mathbb{R}$ in the extended domain of the generator. Then, define $D = \{f \in C_c^1(E), \mathcal{A}f \in C_b(E)\}$ and $\mathcal{L}$ as the extended generator $\mathcal{A}$ restricted to D. The component of the extended generator $(x,v) \to \partial_{x_i} f(x,v) + x_i \partial_{v_i} f(x,v)$ produces Hamiltonian dynamics (see Equation (C.10)) preserving any Gaussian measure centered on 0. Notice that the $R_U(x)$ satisfies

$$\langle \nabla U(x), R_U(x)v \rangle_{\alpha(x,v)} = -\langle \nabla U(x), v \rangle_{\alpha(x,v)}$$

and that

$$\|\Sigma^{-1/2} R_U(x) v\| = \|\Sigma^{-1/2} v\|.$$

Then one can check that $\int \mathcal{L}f(x,v)\mu(\mathrm{d}x, \mathrm{d}v) = 0$ by carrying out similar computations as in the proof of Proposition C.2.1. □

A variant of the sticky Boomerang sampler is the sticky factorised Boomerang sampler (being the sticky version of the factorised Boomerang sampler introduced in Bierkens et al. 2020). Here the process has the same dynamics, refreshment rule and sticky events of the sticky Boomerang process but has a different reflection rate and reflection rule. Similarly to the Sticky Zig-Zag process, the first reflection time of the sticky factorised Boomerang sampler is given by the minimum of $|\alpha(x,v)|$ Poisson times $\{\tau_j\colon j \in \alpha(x,v)\}$ with $\tau_j \sim \mathrm{IPP}(t \to \lambda_j(\varphi(t,x,v))$ and $\lambda_j(x,v) = (\partial_{x_j} U(x) v_j)^+$. Likewise the Sticky Zig-Zag process, at the reflection time the process reflects its velocity by changing the sign of the ith component $v \to v[i\colon -v_i]$ where $i = \arg\min\{\tau_j\colon j \in \alpha(x,v)\}$. As shown in Bierkens et al. (2020) the factorised Boomerang sampler can outperform the Boomerang sampler when $\partial_{x_i} U$ is function of few coordinates.

C.3 Comparison between reversible jump PDMPs and sticky PDMPs

In this appendix, we discuss the differences between the sticky PDMPs and RJ (Reversible Jumps) PDMPs presented in Chevallier, Fearnhead, and Sutton (2020) which, similarly to us, addresses variable selection problems using PDMP samplers.

The approach taken in Chevallier, Fearnhead, and Sutton (2020) is based on the framework of reversible jump (RJ) MCMC as proposed in Green (1995) and its derivation is therefore substantially different from our approach. Nonetheless, the samplers have certain similarities. The dynamics of both the RJ PDMPs in Chevallier, Fearnhead, and Sutton (2020) and the sticky PDMPs proposed in this paper allow each coordinate to stick at 0 for an exponential time. The rate of the exponential time of the sticky PDMPs depends only on the velocity component of each coordinate, while the rate of RJ PDMPs can depend on the current state of the process. The latter is slightly more general as it allows to choose freely a prior weight on the Dirac measure for each possible model (while our approach allows to choose freely a prior weight on the Dirac measure of each possible coordinate). An important difference between the two methods is the behaviour of the process after the particle sticks at 0: the velocity of the coordinate of the sticky PDMPs is restored to its previous value while for RJ PDMPs, a new velocity is drawn independently to the previous one. The former action introduces non-reversible jumps between models while the latter reversible jumps and a random walk behaviour when jumping between models. This simple, yet substantial, difference leads to two different limiting behaviour of the two processes when the number of Dirac measures increases. The limiting behaviour of both processes is unvelied below in Appendix C.3.2 through numerical simulations: while the Sticky Zig-Zag converges to ordinary Zig-Zag, the RJ Zig-Zag asymptotically exhibits diffusive behaviour.

For RJ PDMPs, the random walk behaviour is mitigated by introducing a tuning parameter p which allows each coordinate to stick at 0 only a fraction of times when hitting 0 (and compensating for this by down-scaling the rate of the exponential waiting time when the coordinate sticks). The parameter p is tuned to be equal to 0.6 based on empirical criteria. In Appendix C.3.1 we investigate the possibility to introduce the tuning parameter p in the Sticky Zig-Zag sampler and, based on a heuristic argument and a simulation study, we concluded that it is not beneficial for us.

C.3.1 Heuristics for the choice of p

Here we investigate the possibility of introducing the parameter p to the Sticky Zig-Zag sampler. This parameter was originally introduced in Chevallier, Fearnhead, and Sutton (2020). Based on the heuristic argument and the simulation study given

below, we conclude that the introduction of p does not improve the performance of the Sticky Zig-Zag sampler.

The parameter p defines the probability for a coordinate to stick at 0 when it hits 0. By introducing this parameter, the times of the particles stuck at 0 has to be rescaled by a factor of p in order target the right measure.

Consider a trajectory $\{z_t \colon 0 < t < T\}$ of the one dimensional ordinary Zig-Zag sampler (without stickiness) targeting a given measure. In this case, one could create a trajectory of the Sticky Zig-Zag process retrospectively just by adding constant segments equal to 0, every time the process hits 0 with random length equal to XY, with $X \sim \text{Ber}(p)$ and $Y \sim \text{Exp}(\kappa/p)$, X independent from Y. Then, if the trajectory z_t hits 0 N-times, the total occupation time of the sticky process in 0 is Gamma-distributed with shape parameters $\frac{N}{p}$ and inverse scale parameter $p\kappa$ (in variable selection, this would correspond to the posterior probability of the sub-model without the coefficient). While the mean of this random variable is constant for every p, its variance is $\frac{N}{\kappa p}$ and is minimized when $p = 1$.

Based on the aforementioned heuristics, it appears not useful to introduce the parameter p for the Sticky Zig-Zag. This claim is supported by simulations presented in Figure C.2, where we vary p from 0.1 (top) to 1.0 (bottom) for a 20 dimensional Gaussian density with pairwise correlation equal to 0.99 and relative to the measure

$$\prod_{i=1}^{d} \big(\mathrm{d}x_i + c \sum_{j \in \mathbb{N}} \delta_{j*0.01}(\mathrm{d}x_i)\big), \tag{C.11}$$

with $c = 1.0$. In Figure C.2, left panels, the traces are more erratic when p is small and the process traverses the space in less time when p is large (notice the different ranges of the vertical axis). In Figure C.2, right panels, the phase portrait of the first two coordinates is shown. By visual inspection it is possible to notice that the phase portrait fails to be symmetric on the axis $x_1 = -x_2$ for p small while it succeeds for $p = 1$ (notice again the different ranges of the axes), hence suggesting that Zig-Zag sampler has a better mixing for $p = 1$.

C.3.2 Limiting behaviour

Here we show the different limiting behaviour between the RJ-PDMP samplers and the sticky PDMP samplers as the number of Dirac measures increases.

The limiting behaviour of the two samplers significantly differ because after every time a coordinate sticks at a point mass, the sticky PDMP sampler preserves the velocity component while RJ PDMP sampler has to refresh a new independent velocity. We illustrate the limiting behaviour of the two samplers through simulations where we let the Sticky Zig-Zag and the RJ-Zig-Zag sampler (with $p = 0.6$) target a 20-dimensional measure with a Gaussian density with pairwise correlation

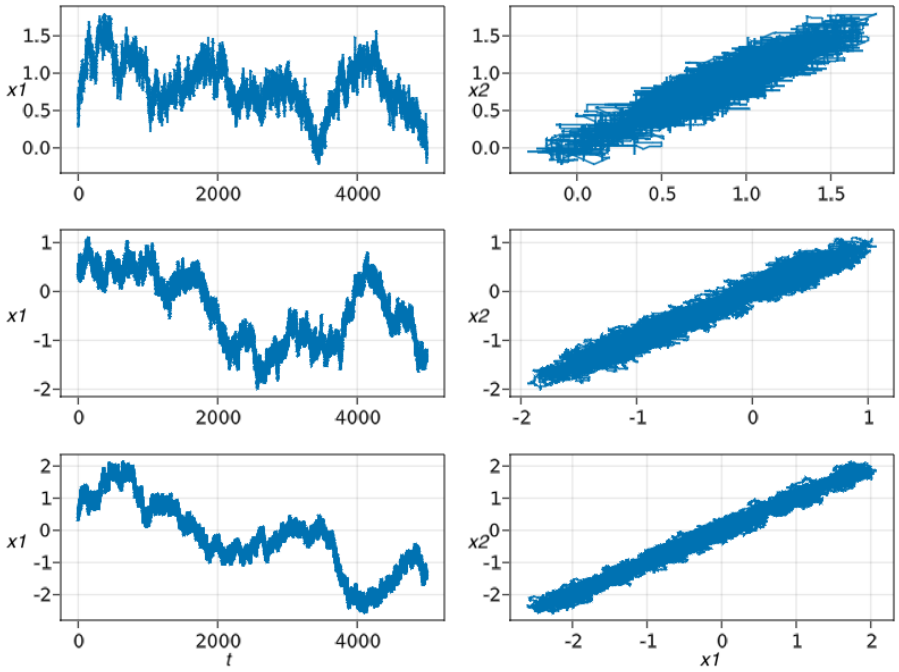


Figure C.2: x_1 trace plots (left) and x_1-x_2 phase portraits (right) of the Sticky Zig-Zag samplers with final clock $T = 50^3$ with p equal to 0.1 (top), 0.5 (center), 1.0 (bottom). The target measure has a Gaussian density with pairwise correlation equal to 0.99 relative to the reference measure of Equation (C.11). By comparing the symmetry of the empirical measures along the diagonal and the range of the coordinates, one can conclude that the algorithm performs best for $p = 1$.

equal to 0 (Figure C.3) and 0.99 (Figure C.4) relative to the reference measure of Equation (C.11) with $c = 10$. While the Sticky Zig-Zag sampler resemble an ordinary Zig-Zag sampler, the RJ-PDMP sampler has a limiting diffusive behaviour and appears to explore the space less efficiently than the sticky PDMP sampler (see the range of the axes and the symmetries of the measure around the axis $x_2 = -x_1$).

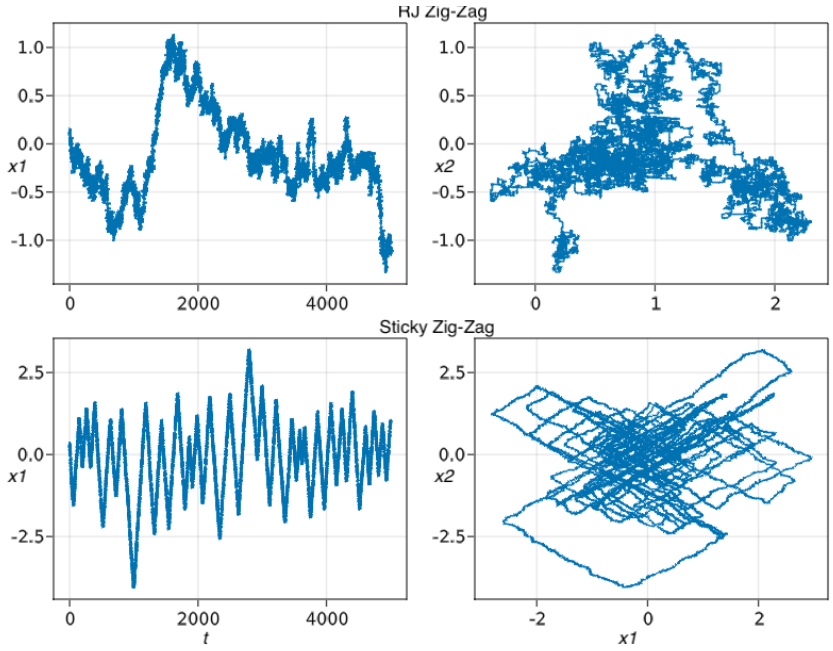


Figure C.3: Comparison between RJ Zig-Zag samplers (first row) and Sticky Zig-Zag samplers (second row) targeting a 20 dimensional measure with Gaussian density with pairwise correlation equal to 0.0 and relative to the reference measure in Equation (C.11). Column 1: trace plot of the first coordinate. Column 2: trace plot of the second column. In all cases $T = 10^4$. By looking at the range of each coordinate, it is clear that the Sticky Zig-Zag mixes faster than its reversible counterpart.

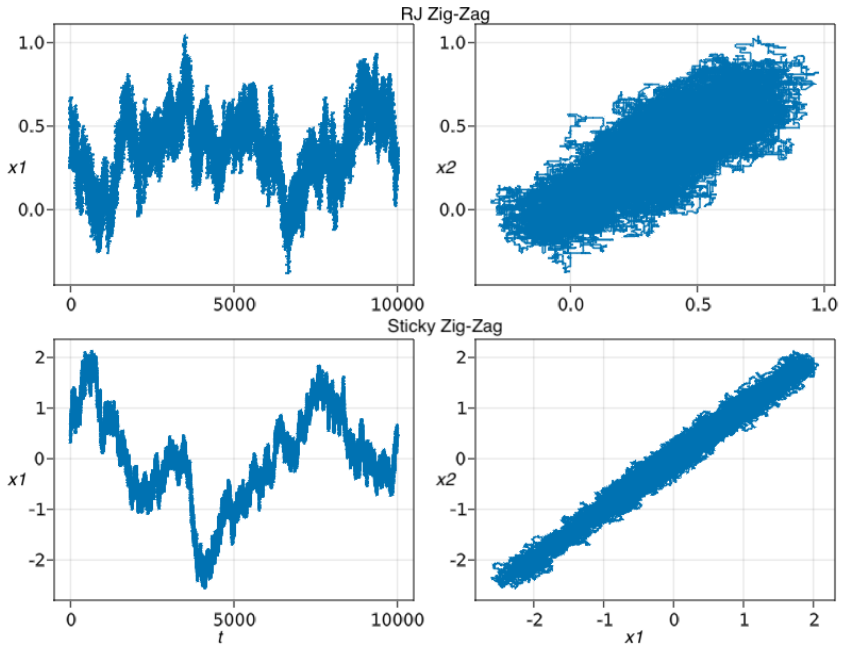


Figure C.4: Same description as in Figure C.3, except now for a Gaussian measure with pairwise correlation equal to 0.99. By looking for example at the symmetry along the axis $x_2 = -x_1$ and the ranges of the coordinates, it is clear that the Sticky Zig-Zag outperforms the RJ Zig-Zag.

C.4 Details of Section 4.3

C.4.1 Bayes factors for Gaussian models

Let $(X, Y) \sim N(\mu, \Gamma^{-1})$, written in block form as

$$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \Gamma_x & \Gamma_{xy} \\ \Gamma'_{xy} & \Gamma_y \end{bmatrix}.$$

Denote the density of (X, Y) evaluated at (x, y) by $\phi([x, y];\, \mu,\, \Gamma^{-1})$. Let

$$X \mid (Y = y) \sim \mathcal{N}(\mu_{x|y},\, \Gamma_x^{-1}) \tag{C.12}$$

be the marginal density of X given $Y = y$, where $\mu_{x|y} = \mu_x - \Gamma_x^{-1}\Gamma_{xy}(y - \mu_y)$. Assume Γ_x to be positive definite and let the marginal density of Y be

$$\int \phi([x, y];\, \mu,\, \Gamma^{-1})\mathrm{d}x = (2\pi)^{\frac{d_x - d}{2}} |\Gamma|^{\frac{1}{2}} |\Gamma_x|^{-\frac{1}{2}} \exp\left(\frac{1}{2}\mu'_{x|y}\Gamma_x\mu_{x|y} - \frac{1}{2}[-\mu_x, y - \mu_y]'\Gamma[-\mu_x, y - \mu_y]\right) \tag{C.13}$$

where d_x is the size of X.

We are now ready to compute the corresponding Bayes factors of two neighbouring (sub-)models as in Equation (4.12) when Ψ is a quadratic function. For every set of indices $\alpha \subset \{1, 2, \ldots, d\}$ and for every j, the Bayes factors relative to two neighbouring (sub-)models (those differing by only one coefficient) for a measure as in Equation (4.2) are given by

$$B_j(\alpha) = \frac{\mu(\mathcal{M}_{\alpha\cup\{j\}})}{\mu(\mathcal{M}_{\alpha\setminus\{j\}})} = \frac{\kappa_i \int_{\mathbb{R}^{|\alpha\cup\{j\}|}} \exp(-\Psi(y))\mathrm{d}x_{\alpha\cup\{j\}}}{\int_{\mathbb{R}^{|\alpha\setminus\{j\}|}} \exp(-\Psi(z))\mathrm{d}x_{\alpha\setminus\{j\}}}, \tag{C.14}$$

where $y = \{x \in \mathbb{R}^d \colon x_i = 0,\, i \notin (\alpha \cup \{j\})\}$, $z = \{x \in \mathbb{R}^d \colon x_i = 0,\, i \notin (\alpha \setminus \{j\})\}$. Since Ψ is quadratic, we can write $\exp(-\Psi(x)) = C\phi(x;\, \mu,\, \Gamma^{-1})$ for some parameters C, μ, Γ. By using both Equation C.12 and Equation C.13 we have that the right hand side of Equation (C.14) is equal to

$$\kappa_i \sqrt{\frac{2\pi|\Gamma_{x_1}|}{|\Gamma_{x_2}|}} \exp\left(\frac{1}{2}(\mu'_{x_1|y_1=\mathbf{0}}\Gamma_{x_1}\mu_{x_1|y_1=\mathbf{0}} - \mu'_{x_2|y_2=\mathbf{0}}\Gamma_{x_2}\mu_{x_2|y_2=\mathbf{0}})\right)$$

where $x_1 = x_{\alpha_{-j}\cup\{j\}}$, $x_2 = x_{\alpha_{-j}\setminus\{j\}}$, $y_1 = x_{\alpha^c_{-j}\setminus\{j\}}$, $y_2 = x_{\alpha^c_{-j}\cup\{j\}}$. Furthermore, by Equation C.12, the random variable at step 2 of the Gibbs sampler presented in Section 4.3.1 can be simulated as $X_\alpha|(X_{\alpha^c} = \mathbf{0}) \sim \mathcal{N}(\mu_{x_\alpha|x_{\alpha^c}=\mathbf{0}}, \Gamma_{x_\alpha})$.

C.4.2 Simulating sticky PDMPs and sticky Zig-Zag samplers

Sticky samplers can be implemented recursively by modifying appropriately the ordinary PDMP samplers so to include sticky events as introduced in Section 4.2. We discuss how to integrate local implementations of the algorithms to increase the sampler's performance in case of a sparse dependence structure in the target measure and in case of local upper bounding rates.

Although PDMPs have continuous trajectories, the algorithm computes and saves only a finite collection of points (which we refer to as the skeleton of the continuous trajectory) corresponding to the positions, velocities and times where the deterministic dynamics of the process change. In between those points, the continuous trajectory can be deterministically interpolated.

In case the ith partial derivative of the negative score function is a sum of N_i terms, which is the case for example in regression problems, subsampling techniques can be employed as described in Section 4.2.4.

Computing Poisson times for PDMPs

As PDMPs move deterministically (and with simple dynamics) in between event times, the main computational challenge consists of simulating those times. Given an initial position (x, v), the distribution of the time until the next event is specified in (4.7). A sample from this distribution can be found by solving for τ' in the equation

$$\int_0^{\tau'} \lambda(\varphi(s, x, v))\mathrm{d}s = t, \quad t = \mathrm{Exp}(1). \tag{C.15}$$

We then write that $\tau' \sim \mathrm{IPP}(\lambda(\varphi(\cdot, x, v))$. When it is not possible to find the root of Equation (C.15) in closed form, it suffices to find upper bounds $\overline{\lambda}$ for the rate functions which satisfies, for any $(x, v) \in E$ and for some $\Delta = \Delta(x, v) > 0$

$$\overline{\lambda}(t, x, v) \geq \lambda(\varphi(t, x, v)), \quad \Delta \geq t \geq 0, \tag{C.16}$$

for which this is possible and use the thinning scheme: Let $\tau' \sim \mathrm{IPP}(\bar{\lambda}(\cdot, x, v))$; if $\tau' > \Delta$ then the proposed time is rejected and a new time has to be drawn as $\tau' \sim \mathrm{IPP}(\bar{\lambda}(\cdot, \phi(\Delta, x, v)))$. We *accept* the proposed time with probability $\lambda(\phi(\tau', x, v))/\bar{\lambda}(\tau', x, v)$. This scheme is referred as *adaptive thinning* in Bouchard-Côté, Vollmer, and Doucet (2018). More sophisticated and potentially efficient thinning schemes have been proposed, see Sutton and Fearnhead (2021). The simulation of unfreezing times is easier: once the i-th component hits zero then it sticks at zero for a time that is exponentially distributed with parameter $\kappa_i|v_i|$.

For the ordinary d-dimensional Zig-Zag and the factorised Boomerang sampler (these samplers are called *factorised PDMPs* in Bierkens et al. 2020), the reflection time is factorised as the minimum of d independent clocks $\tau_1, \tau_2, \ldots, \tau_d$ where $\tau_i \sim$

IPP($\lambda_i(\varphi(\cdot, x, v))$ for $i = 1, 2, \dots, d$. The first reflection time of the d-dimensional sticky factorised samplers is obtained instead by finding the minimum of $|\alpha| < d$ independent clocks with the same rates λ_i of the ordinary factorised sampler, but only for the active coordinates $i \in \alpha(x, v)$.

If $\partial_{x_i}\Psi$ (an estimate of $\partial_{x_i}\Psi$ when using subsampling) or the upper bound $\overline{\lambda}$ depends on fewer coordinates, then the evaluation of each reflection time is cheaper. The fully local implementation presented in Bierkens et al. (2021) exploits these two features once in proposing the reflection time and once for deciding whether to accept. Below, we discuss in more details the algorithm of Sticky Zig-Zag sampler with local upper bounds and with subsampling.

Local implementation:

Assume that the sets $\overline{A}_i$ and $\overline{\lambda}_i$ are such that

$$\overline{\lambda}_i(t, x, v) = f_i(t, x_{\overline{A}_i}), \quad \forall x, \text{ for } i = 1, 2, \dots, d$$

for some $f_i \colon \mathbb{R}^+ \times \mathbb{R}^{|\overline{A}_i|} \to \mathbb{R}^+$ with $\overline{A}_i \subset \{1, 2, \dots, d\}$. Given an initial position (x, v) and random times $\tau_j \sim \text{IPP}(t \to \overline{\lambda}_j(t, x, v))$, for $i \in \alpha$, denote by $i = \arg\min_{j \in \alpha(x,v)} \tau_j$ and $\tau = \min_{j \in \alpha(x,v)} \tau_j$ the first proposed reflection time. According to the thinning procedure for Poisson processes, the process flips the ith coordinate with probability $\lambda_i(\varphi(\tau, x, v))/\overline{\lambda}_i(\tau, x, v)$. If the process flips the ith velocity, then the Poisson rates $\{\overline{\lambda}_j \colon j \in \alpha, \overline{A}_j \not\ni i\}$ continue to be valid upper bounds so that the corresponding reflection times do not need to be renewed (see Bierkens et al. 2021, Section 4, for implementation details).

In general, when the ith particle freezes at 0 or was stuck at 0 and gets released, the reflection times $\{\tau_j \colon i \in \overline{A}_j\}$ have to be renewed. However this is not always the case, as there are applications, such as the one in Section 4.4.4, for which the upper bounding rates $\{\overline{\lambda}_i\}_{i=1}^d$ continue to be valid upper bounds when one or more particles hit 0 and therefore the waiting times computed before the particles hit 0 are still valid.

Fully local implementation:

Consider now the decomposition of $\partial_{x_i}\Psi$, $i = 1, 2, \dots, d$ given in Equation (4.11) and such that

$$S(x, i, j) = f_{i,j}(x_{\widetilde{A}_{i,j}}), \quad \forall x, \text{ for } (i, j) \in \{1, 2, \dots, d\} \times \{1, 2, \dots, N_i\}$$

for some $f_{i,j} \colon \mathbb{R}^{|\widetilde{A}_{i,j}|} \to \mathbb{R}$ with $\widetilde{A}_{i,j} \subset \{1, 2, \dots, d\}$.

The fully local implementation of the Sticky Zig-Zag with subsampling profits from local upper bounds and local gradient estimators by assigning an independent

time for each coordinate, thus evolving the flow of only the coordinates which are required at each step and by stacking $\{\tau_j \wedge \tau_j^\star : j \in \alpha\}$, with τ_j being a proposed reflection time and $\tau_j^\star$ the hitting time to 0, and the unfreezing times $\{\tau_j^\circ : j \in \alpha^c\}$ in an ordered queue. For a documented implementation, see Schauer and Grazzi (2021).

Given an initial point (x, v) and if $i = \arg\min(\tau_j : j \in \alpha(x, v))$ is the coordinate of the first proposed reflection time $\tau = \min(\tau_j : j \in \alpha(x, v))$, the sampler reflects the velocity of the ith coordinate with probability $\widetilde{\lambda}_{i,J}(x_{\widetilde{A}_i}(\tau), v)/\overline{\lambda}(\tau, x, v)$ with $J \sim \text{Unif}(\{1, 2, \ldots, N_i\})$. Hence, it is only required to update the position of the coordinates with index in $\widetilde{A}_{i,J} \setminus \alpha^c(x, v)$. Then,

- if the ith velocity flips, then the algorithm needs to update only the waiting times $\{\tau_j : j \in \alpha, \overline{A}_j \ni i\}$ (as described in Appendix C.4.2) and, to this end, needs to update the position of the coordinates with index $\{k \in \overline{A}_j \setminus \alpha^c(x, v) : i \in \overline{A}_j\}$;
- in the other case, when the ith velocity does not change (shadow event), only τ_i has to be renewed so that only the particles in $\overline{A}_i$ have to be updated.

Remark C.4.1. *(Sparse implementation.) When the dimensionality d is large, inserting each waiting time in a ordered queue and initializing the state space can be computationally expensive. If for example the product $k_i|v_i|$ is equal for all i, an alternative efficient and sparse implementation is possible. Here we simulate the sticky time for each frozen coordinate by means of simulating the overall sticky time from the exponential distribution with rate $\sum_{i \in \alpha^c} \kappa_i |v_i|$ (which has to be renewed every time a new particle sticks at 0) and selecting the particle to unfreeze uniformly from the set α^c. A further improvement can be obtained by representing x as a sparse vector and saving only the location of the active particles $\{x_i : i \in \alpha\}$.*

C.4.3 Runtimes of the algorithms

We will now compute typical runtimes for the Gaussian model, assuming a decomposition

$$\Psi(x) = (x - \mu)'\Gamma(x - \mu) = \sum_{i=1}^{N} (x - \mu_i)'\Gamma_i(x - \mu_i) + c,$$

so that N captures the dependence on the number of observations in a Bayesian setting.

Sticky Zig-Zag sampler:

The computational cost of simulating PDMP samplers is intimately related with the number of random times generated. This, in turn, depends on the intensity of the

rate λ of the underlying Poisson process. For any initial position and velocity (x, v), the total rate of the Sticky Zig-Zag sampler is equal to

$$\lambda(x, v) = \sum_{i \in \alpha} \lambda_i(x, v) + \sum_{i \in \alpha^c} |v_i| \kappa_i \tag{C.17}$$

where, as before, $\alpha = \{i \colon x_i \neq 0\}$. In the following analysis, we drop the dependence on (x, v) and we assume that the size of $\alpha(t) := \{i \colon x_i(t) \neq 0\}$ fluctuates around a typical value p in stationarity. Thus p represents the number of non-zero components in a typical model, and can be much smaller than d in sparse models.

We consider the sticky Zig-Zag with local implementation as in Remark C.4.1 where we assume $\kappa := \kappa_1 = \kappa_2 = \cdots = \kappa_d$. We ignore logarithmic factors, e.g., for priority queue insertion. In the analysis below we distinguish between the computational costs of reflection events and unfreezing events.

The number of reflection and unfreezing events per unit time interval are respectively $\mathcal{O}(p)$ and $\mathcal{O}((d-p)\kappa)$ per unit time; see Equation (C.17). Once either a reflection or unfreezing event happens, we have to recompute between $\mathcal{O}(1)$ and $\mathcal{O}(p)$ new reflection event times (depending on the elements of $\overline{A}_i \cap \alpha$; see Appendix C.4.2). Finally, each newly computed reflection event time for the particles $i \in \alpha$ requires a computation ranging from $\mathcal{O}(1)$ to $\mathcal{O}(N)$. The complexity $\mathcal{O}(1)$ can be achieved using the subsampling technique (Section 4.2.4) in ideal scenarios (Bierkens, Fearnhead, and Roberts 2019). Table 4.1 in Section 4.3 summarizes the overall scaling complexity of the Sticky Zig-Zag algorithm for the quantities p and N.

Gibbs sampler:

At each iteration, the Gibbs sampler algorithm requires the evaluation of the Bayes factors which involves the inversion of a square matrix of dimension $p \times p$. This can be efficiently obtained with a Cholesky decomposition of a sub-matrix of Γ. This is a computation of $\mathcal{O}(p^3)$ when Γ is full; a lower order is possible when Γ is sparse. For example, in the example in Section 4.4.2, the complexity of this operation is $\mathcal{O}(p^{3/2})$. This is followed by computing sufficient statistics in step 2 of Section 4.3.1 which involves the inversion of a triangular matrix which is $\mathcal{O}(|\alpha^2|)$ ($\mathcal{O}(1)$ if the Cholesky factor is sparse) in addition to an operation of order pN (for example in linear or logistic regression). It is important to notice that if Γ is sparse, its Cholesky factors might not be. Our finding are summarized in Table 4.1 in Section 4.3 and validated by the numerical experiments of Section 4.4 (Figure 4.5, Figure 4.9).

C.4.4 Mixing

Next to the complexity per iteration, we should also understand the time the underlying process needs to explore the state space and to reach its stationary measure.

Given the different nature of dependencies of the two algorithms, a rigorous and theoretical comparison of their mixing times is difficult. We therefore provide a heuristic argument for two specific scenarios.

Let both algorithms be initialized at $x \sim \mathcal{N}_d(0, I)$ with all non-zero coordinates ($\alpha^c = \emptyset$) and assume that the target μ assigns most of its probability mass to the null model $\mathcal{M}_\emptyset$. Consider the following scenarios:

- *A measure supported in every model* and such that for any two models $\mathcal{M}_{\alpha_i}$ and $\mathcal{M}_{\alpha_j}$ with $\alpha_i \neq \alpha_j$, we have $\mu(\mathcal{M}_{\alpha_i}) > \mu(\mathcal{M}_{\alpha_j})$ if $|\alpha_i| < |\alpha_j|$. The Sticky Zig-Zag will be directed to the null model, each coordinate with speed 1, so that the first visit of the null set happens with an expected time $\mathcal{O}(\max_i(|x_i|))$ which is of $\mathcal{O}(\log d)$ if x is standard Gaussian. On the other hand, the Gibbs sampler, at every iteration, randomly picks a coordinate and, if this is a non-zero coordinate, succeeds to set that coordinate to zero. Denote by τ_α the (random) number of iterations needed for the algorithm to set any non-zero coordinate to zero, when exploring a model $\mathcal{M}_\alpha$. Then $\mathbb{E}(\tau_\alpha) = d/|\alpha|$ which ranges from 1 (when $\mathcal{M}_\alpha$ is the full model) to d (for any sub-model with only one non-zero coordinate). Consider any sequence $\mathcal{M}_{\alpha_1}, \mathcal{M}_{\alpha_2}, \ldots, \mathcal{M}_{\alpha_{d-1}}$ of models with $|\alpha_j| + 1 = |\alpha_{j+1}|$ (decreasing size) and with $\mathcal{M}_{\alpha_1}$ begin the full model. By adding the expected number of iterations at each of those model, we conclude that the process started at x in the full model, is expected to reach the null model in $\sum_{i=1}^{d} d/i$ iterations which is of $\mathcal{O}(d \log(d))$.

- *A measure supported on a single nested sequence of sub-models,* up to the full model: i.e. for a model $\mathcal{M}_{\alpha_j}$, with $\mu(\mathcal{M}_{\alpha_j}) \neq 0$ there is only one sub-model $\mathcal{M}_{\alpha_i} \subset \mathcal{M}_{\alpha_j}$ with $|\alpha_i| + 1 = |\alpha_j|$ and the smaller model again has more probability mass $\mu(\mathcal{M}_{\alpha_i}) > \mu(\mathcal{M}_{\alpha_j})$. By a similar argument as above, the first expected visit time of the null model is of $\mathcal{O}(\sum_{i=1}^{d} |x_i|) = \mathcal{O}(d)$ for the Sticky Zig-Zag, while for the Gibbs sampler the expected number of steps is d^2.

Table 4.2 in Section 4.3 summarizes the scaling results derived in the two cases considered above.

C.5 Details of Section 4.4

C.5.1 Logistic regression

Similar computations for the bounds of the Poisson rates of the Zig-Zag sampler applied to logistic regressions can be found in the supplementary material of Bierkens, Fearnhead, and Roberts (2019). Given a posterior density of the form of Equation (4.2) with

$$\Psi(x) = \sum_{j=1}^{N} \left(\log\left(1 + e^{\langle A_{[j,:]}, x\rangle}\right) - y_j \langle A_{[j,:]}, x\rangle\right) + \frac{1}{2\sigma^2}\|x\|^2$$

we use the Sticky Zig-Zag subsampler presented in Section 4.2.4. To that end, define $U(x) = \Psi(x) - \frac{1}{2\sigma^2}\|x\|^2$. We decompose the partial derivatives of U as follow:

$$\partial_{x_i} U(x) = \sum_{j \in \Gamma_i} S(x, i, j)$$

with sets $\Gamma_i = \{j \in \{1, 2, \ldots, N\} \colon A_{j,i} \neq 0\}$ and

$$S(x, i, j) = \left(\frac{A_{[j,i]} e^{\langle A_{[j,:]}, x\rangle}}{1 + e^{\langle A_{[j,:]}, x\rangle}} - y_j A_{[j,i]}\right).$$

Then, for all $i = 1, 2, \ldots, p$ and any $x' \in \mathbb{R}^p$, if $\mathrm{J} \sim \mathrm{Unif}(\Gamma_k)$, the estimator $[|\Gamma_i|(S(x, i, J) - S(x', i, J)] + \partial_{x_i} U(x^*) + \sigma^{-2} x_i$ is unbiased for $\partial_{x_i}\Psi(x)$. Notice that the partial derivative of $S(x, k, j)$ is bounded:

$$\partial_{x_i}(S(x, k, j)) = \frac{A_{[j,k]} A_{[j,i]} e^{\langle A_{[j,:]}, x\rangle}}{\left(1 + e^{\langle A_{[j,:]}, x\rangle}\right)^2} \leq \frac{1}{4} A_{[j,k]} A_{[j,i]},$$

which means that for $i = 1, 2, \ldots, d$

$$|S(x, i, j) - S(x', i, j)| \leq C_i \|x - x'\|_p, \quad p \geq 1,\ j \in \Gamma_i,\ x, x' \in \mathbb{R}^d,$$

with

$$C_k = \frac{1}{4} \max_{j=1,..,N} |A_{[j,k]}|\, \|A_{j,:}\|_2.$$

Then given an initial position $(x, v) \in E$, tuning parameter x' and for any $t \geq 0$, write $(x(t), v(t)) = \varphi(t, x, v)$ with $i \in \alpha(x, v)$:

$$\begin{aligned}
\widetilde{\lambda}_i(x(t), v(t)) &= \left(v_i \left(\partial_{x_i} U(x') + \sigma^{-2} x_i(t) + |\Gamma_i|(S(x(t), i, j) - S(x', i, j))\right)\right)^+ \\
&\leq (v_i(\partial_{x_i} U(x') + \sigma^{-2}(x_i + v_i t)))^+ \\
&\qquad + |v_i||\Gamma_i| \left(|S(x(t), i, j) - S(x, i, j)| + |S(x, i, j) - S(x', i, j)|\right) \\
&\leq (v_i(\partial_{x_i} U(x') + \sigma^{-2}(x_i + v_i t))^+ + |v_i||\Gamma_i| C_i \left(t\|v\|_p + \|x - x'\|_p\right).
\end{aligned}$$

Thus we set

$$\lambda_i(t, x, v) = v_i(a_i(x, v) + b_i(x, v)t)$$

where $a_i(x, v) = (v_i(\partial_i U(x') + \sigma^{-2}x_i))^+ + C_i|\Gamma_i||v_i|\|x - x'\|_p$ and $b_i(x, v) = |v_i|C_i|\Gamma_i|\|v\|_p + v_i^2\sigma^{-2}$. We choose x' to be the posterior mode of $\exp(-\Psi)$, which in this case is unique and easily found with the Newton's method since the function $\exp(-\Psi)$ is convex. Given an initial position (x, v), suppose the particle $j \neq i$ gets frozen at time $\tau \geq 0$. Then for $t \geq \tau$ we have that $\| \int_0^t v(t)\mathrm{d}t\|_p = \tau\|v\|_p + (t - \tau)\|v'\|_p \leq t\|v\|_p$, with $v' = v[j\colon 0]$. This implies that the Poisson times drawn before the jth coordinate gets stuck are still valid upper bounds after time τ. The same argument follows easily for $n \geq 1$ coordinates getting stuck at 0.

C.5.2 Spatially structured sparsity

For this application, we use the thinning scheme as presented in Appendix C.4.2. The bounding rates are of the form

$$\bar{\lambda}_i(t, x(t_0), v(t_0)) = (c + v_i(t_0)\partial_{x_i}\Psi(x(t_0))^+ \tag{C.18}$$

for $t \in [0, \Delta]$ with $\Delta = 1/c$. To see this, define the Lipschitz growth bound $L_{x,v,\Delta}$ as

$$P(\sup_{0<t<\Delta} \frac{1}{t}|V_i(t)\partial_{x_i}\Psi(X(t))| \leq L_{x,\Delta} \mid X(0) = 0, V(0) = v) = 1, \quad i = 1, 2, \ldots, d,$$

which gives an explicit expression for c in Equation (C.18) as

$$c - L_\Delta\Delta = 0 \Rightarrow \Delta = 1/c,$$

such that the inequality (C.16) holds. With $L_\Delta = \sup_x L_{x,v,\Delta}$, in this application we have that

$$L_\Delta = \sup_{v,t} |\partial_t\partial_{x_i}\Psi(x + tv)| = c_2 + 8c_1 + 1/\sigma^2$$

with c_1, c_2 defined in Section 4.4.2. With this given choice, in the simulations of Section 4.4.2, the ratio between the accepted reflection times and the proposed reflection times was 0.357. Here we used the local implementation of the Sticky Zig-Zag given by Appendix C.4.2 (with sets $\overline{A_i} = i$ for all i) in conjunction with the sparse algorithm as in Remark C.4.1.

C.5.3 Sparse precision matrix

By write $\Psi(x)\bigotimes_{i=1}^p \bigotimes_{j=1}^i (\mathrm{d}x_{i,j} + \frac{1}{\kappa}\delta_0(\mathrm{d}x_{i,j})\mathbf{1}_{(i\neq j)})$ and we have that

$$\partial_{x_{i,j}}\Psi(x) = (YY')_{(i,:)}X_{(:,j)} + \gamma_{i,j}(x_{i,j} - c_{i,j}) - \mathbf{1}_{(i=j)}\left(\frac{N}{x_{i,j}}\right). \tag{C.19}$$

Note that, for any initial position and velocity (x, v), the reflection times of the Sticky Zig-Zag with rates $\lambda_{i,j}(\phi(t, x, v)) = (v_i \partial_{x_{i,j}} \Psi(x + vt))^+$ can be computed exactly for the off-diagonal elements and via a thinning scheme for the diagonal elements where

$$\lambda_{i,i}(\phi(t, x, v) \leq \overline{\lambda}_{i,i}(t, x, v) + \overline{\overline{\lambda}}_{i,i}(t, x, v), \quad t > 0, \forall i.$$

Here $\overline{\lambda}_{i,i}(t, x, v) = (v_{i,i}(YY'_{i,:}(X_{:,i} + vt) + \gamma_{i,i}(x_{i,i} + vt - c_{i,i})))^+$ and $\overline{\overline{\lambda}}_{i,i}(t, x, v) = \left(-v_{i,i}\frac{N}{x_{i,i}+v_{i,i}t}\right)$ and a Poisson time form the bounding rate is simulated as $\min(\tau_1, \tau_2)$ where $\tau_1 \sim \text{IPP}(s \to \overline{\lambda}_{i,i}(s, x, v))$ and $\tau_2 \sim \text{IPP}(s \to \overline{\overline{\lambda}}_{i,i}(s, x, v))$.

This page is intentionally left blank.

Appendix D

Supplement of Chapter 5

D.1 Extended generator of PDMPs with boundaries

The extended generator for PDMPs with boundaries is given in Davis (1993, Section 26) and, for the PDMP samplers considered in Section 5.2, takes the form

$$\begin{aligned}\mathcal{L}f(x,v) = \langle v, \nabla f(x)\rangle s(x) + \lambda_{\mathrm{b}}(x,v)\int \mathcal{Q}_{E,\mathrm{b}}((x,v),\mathrm{d}z)(f(z)-f(x,v)) \\ +\lambda_{\mathrm{r}}(x,v)\int \mathcal{Q}_{E,\mathrm{r}}((x,v),\mathrm{d}z)(f(z)-f(x,v)).\end{aligned}$$

Here we let $\mathcal{L}$ act on functions f in the set

$$\begin{aligned}\mathcal{A} = \{f \in C_c(E); \quad & t \to f(\phi(z,t)) \text{ is absolutely continuous } \forall z \in E; \\ & f(z) = \int_{\partial E^-} f(z')Q_{\partial E^+}(z,\mathrm{d}z'), \quad \forall z \in \partial E^+\}.\end{aligned}$$

The set $\mathcal{A}$ is contained in the domain of the extended generator $\mathcal{D}(\mathcal{L})$ given in Davis (1993, Section 26).

D.1.1 Proof Proposition 5.2.2

By applying the divergence theorem, we have that, for $f \in \mathcal{A}$,

$$\int_E \mathcal{L}f \mathrm{d}(\mu \otimes \rho) = \int_E \left(\langle v, s_c(x)\nabla\Psi(x) - \nabla s_c(x)\rangle - \lambda_{\mathrm{b}}(x,v)\right) f(x,v)\rho(\mathrm{d}v)\mu(\mathrm{d}x) \tag{D.1}$$

$$+ \int_E \int_E \lambda_{\mathrm{b}}(z)\mathcal{Q}_{E,\mathrm{b}}(z,\mathrm{d}z')f(z')(\rho\otimes\mu)(\mathrm{d}z) \tag{D.2}$$

$$+ \int_E \lambda_{\mathrm{r}}(z) \int_E \mathcal{Q}_{E,\mathrm{r}}(z,\mathrm{d}z')(f(z') - f(z))(\rho\otimes\mu)(\mathrm{d}z) \tag{D.3}$$

$$+ \int_{\partial E} \langle v, n(x)\rangle f(x,v)s(x)\rho(\mathrm{d}v)\mu(\mathrm{d}x). \tag{D.4}$$

Then, by Assumption 5.2.3, the right hand-side of (D.1) cancel with the term in (D.2). Furthermore, by Assumption 5.2.2, the refreshment kernel is invariant to ρ so the term in (D.3) is also equal to 0. We are left with the term in (D.4), proving Proposition 5.2.2.

D.2 SIR with notifications

D.2.1 Derivation of the measure in Section 5.3.1

We now derive the terms $\rho_i(x)\mu_i(\mathrm{d}x_i)$ (equation (5.18) and equation (5.19)), which can be heuristically interpreted as the distribution of the infection times X_i relative to those individuals $i \in \{1, 2, \ldots, d\}$ which have not been notified up to time T and those which have been notified before time T. To ease the notation, we drop the index i and consider random times $X \in \mathbb{R}^+$, $\tau^\star \in \mathbb{R}^+$, where $\tau^\star = x + \sigma$, with $\sigma \geq 0$ a random variable independent of x. Suppose X has density $g(t) = \beta(t)\exp(-B(t))$, where $B(t) = \int_0^t \beta(s)\,ds$, and σ has cdf F and pdf f, both with support on $[0, \infty)$.

We wish to determine, for $t \geq 0$, $T \geq 0$,

$$\mathbb{P}(X \wedge T \leq t \mid \tau^\star \geq T).$$

Clearly if $t \geq T$ then this conditional probability is equal to one. It remains to compute, for $t < T$

$$\mathbb{P}(X \leq t \mid \tau^\star \geq T).$$

We compute

$$\begin{aligned}
\mathbb{P}(X \leq t \mid \tau^\star \geq T) &= \frac{\mathbb{P}(X < t, \tau^\star \geq T)}{\mathbb{P}(\tau^\star \geq T)} \\
&= \frac{\mathbb{P}(X < t, X + \sigma \geq T)}{\mathbb{P}(X + \sigma \geq T)} \\
&= \frac{\int_0^t g(r)\mathbb{P}(\sigma > T - r)\, dr}{\int_0^T g(r)\mathbb{P}(\sigma > T - r)\, dr + \int_T^\infty g(r)\, dr} \\
&= \frac{\int_0^t g(r)(1 - F(T - r))\, dr}{\int_0^T g(r)(1 - F(T - r))\, dr + \int_T^\infty g(r)\, dr} \\
&= \frac{\int_0^t \beta(r)\exp(-B(r))(1 - F(T - r))\, dr}{\int_0^T \beta(r)\exp(-B(r))(1 - F(T - r))\, dr + \exp(-B(T))}.
\end{aligned}$$

We see that the random variable $x \wedge T$, conditional on $\tau^\star \geq T$, has a Lebesgue density

$$h(t) = \frac{\beta(t)\exp(-B(t))(1 - F(T - t))}{\int_0^T \beta(r)\exp(-B(r))(1 - F(T - r))\, dr + \exp(-B(T))}$$

on $[0, T]$, and an atomic component of mass

$$1 - \int_0^T h(t)\, dt = \frac{\exp(-B(T))}{\int_0^T \beta(r)\exp(-B(r))(1 - F(T - r))\, dr + \exp(-B(T))}$$

at T. This measure is equal to the measure in equation (5.18).

The measure in equation (5.19) of the infection times for notified individuals is easy to derive as

$$P(X < t \,|\, \tau^\star = c) = \begin{cases} 0 & t < 0, \\ C\int_0^t g(s)f(c - s)\mathrm{d}s & 0 \leq t < c, \\ 1 & c \leq t, \end{cases}$$

for some constant C, giving the desired result.

D.2.2 Computing reflections times

The target measure in equation 5.17 can be rewritten as

$$\mu(\mathrm{d}x) \propto L(x)\mu_0(\mathrm{d}x)$$

where

$$L(x) \propto \exp\left(-\sum_{i=1}^{N} B_i(x)\right)\left(\prod_{i\in\mathcal{N}_T} f(\tau_i^\star - x_i)\beta_i(x)\right)\left(\prod_{i\in\mathcal{N}_T^c}(1-F(T-x_i))\beta_i(x)\right)$$

and

$$\mu_0(\mathrm{d}x) = \left(\prod_{i\in\mathcal{N}_T}\left(\mathbf{1}_{(0\le x_i\le T)}\mathrm{d}x_i + \kappa_i\delta_T(\mathrm{d}x_i)\right)\right)\left(\prod_{i\in\mathcal{N}_T^c}\mathbf{1}_{(0\le x_i\le\tau_i^\star)}\mathrm{d}x_i\right).$$

In the experiments in Section 5.3.1 we assumed that f and F are respectively the density and the distribution of an exponential r.v. with parameter β. Then $\partial_{x_i}\log(f(\tau_i^\star - x_i)) = \partial_{x_i}\log(1-F(T-x_i)) = \beta$.

The first event time of the ith clock of the Zig-Zag process is a Poisson clock with rate $\beta_{i,\mathrm{b}} = (v_i\partial_{x_i}(-\log L(x)))^+$ where

$$\begin{aligned}\partial_{x_k}(-\log L(x)) &= \sum_j \partial_{\tau_k} B_j(x) - \sum_{i\in\mathcal{N}_T}\partial_{x_k}\log f(\tau_i^\star - x_i) - \sum_{i\in\mathcal{N}_T^c}\partial_{x_k}\log(1-F(T-x_i))\\ &= \sum_j \partial_{x_k} B_j(x) - \beta\end{aligned}$$

where

$$B_j(x) = \sum_{i\neq j}\int_0^T \beta_{i,j}(x[j;t])\mathrm{d}t = \sum_{i\neq j} C_{i,j}((\tau_i^\star\wedge x_j - x_i\wedge x_j) + \gamma(\tau_i^\circ\wedge x_j - \tau_i^\star\wedge x_j).$$

Hence the partial derivative of the negative log-likleihood is

$$\partial_{x_k}(-\log L(x)) = \sum_i \partial_{x_k} B_i(x) - \beta = \sum_{i\neq k}\partial_{x_k}\left(B_{k,i}(x) + B_{i,k}(x)\right) - \beta = \sum_{i\neq k}\mathcal{G}_{i,k}(x) - \beta$$

with

$$\mathcal{G}_{i,k}(x) = \begin{cases} -C_{k,i} & \tau_k < x_i,\\ C_{i,k} & x_i < x_k < \tau_i^\star,\\ \gamma C_{i,k} & \tau_i^\star < x_k < \tau_i^\circ,\\ 0 & \tau_i^\circ < x_k.\end{cases}$$

Conditioned on the process not hitting a boundary (a discontinuity), the rates $t\to\lambda_{i,\mathrm{b}}(\phi(t,z))$, $i=1,2,\ldots,d$ (equation 5.14) of the Zig-Zag sampler are constant. Hence, if the process is at $z\in E$, we can efficiently simulate the first reflection time simply as $\tau = \min(\tau_1,\tau_2,\ldots,\tau_d)$, where $\tau_i\sim\mathrm{IPP}(t\to\lambda_{i,\mathrm{b}}(\phi(z,t))$, for $i=1,2,\ldots,d$.

D.3 Teleportation rules for hard-sphere models

Recall that we aim to sample from the measure

$$\mu(\mathrm{d}x) \propto \exp(-\sum_{i=1}^{N} \Psi_0(x^{(i)}))\mathbf{1}_A(x)\mathrm{d}x, \quad \Psi \in \mathcal{C}^1(\mathbb{R}^d)$$

with $A = \bigcap_{i=1}^{N} \bigcap_{\substack{j=1,2,\dots,N,\\ j\neq i}} A_{i,j}$ and

$$A_{i,j} = \{x \in \mathbb{R}^{dN} \colon \|x^{(i)} - x^{(j)}\| \geq (r_i + r_j)\}.$$

We define the boundary for the process as

$$\partial E^+ = \{(x, v) \in \partial\Omega \times \mathbb{R}^{dN} \colon \langle n(x), v\rangle > 0\}$$

where $\partial\Omega = \bigcup_{i=1}^{N} \bigcup_{\substack{j=1\\ j\neq i}}^{N} \partial\Omega_{i,j}$ and

$$\partial\Omega_{i,j} = \{x \in \mathbb{R}^{dN} \colon |x^{(i)} - x^{(j)}| = (r_i + r_j)\}.$$

We consider only deterministic teleportation rules of the form $\mathcal{T}(x, \cdot) = \delta_{\kappa(x)}(\cdot)$ for $(x, v) \in \partial E^+$ and informally discuss 3 different choices for κ, each one being efficient in different scenarios. Here for efficient teleportation rules we mean those that accept the new proposed points with high probability. To that end, we want the teleportation rule

- to swap the centers of two colliding hard-spheres and, after that, to move their center as little as possible to have high acceptance probability,
- to move the total volume of hard-spheres as little as possible in order to reduce the probability that some hard-spheres overlap after teleportation.

In the following, we let $(x, v) \in \partial E^+$ with $|x^{(i)} - x^{(j)}| = r_i + r_j$ for some i, j.

D.3.1 Swap the centers

Set, for $\ell = 1, 2, \dots, N$

$$[\kappa(x)]^{(\ell)} = \begin{cases} x^{(j)} & \text{if } \ell = i \\ x^{(i)} & \text{if } \ell = j \\ x^{(\ell)} & \text{otherwise} \end{cases}$$

This teleportation swaps $x^{(i)}$ with $x^{(j)}$ (see Figure D.1, top panels) and has the advantage that if $\kappa(x) \in A$, then the new point is accepted with probability 1 as $\alpha(x, \kappa(x)) = 1$.

When the size of two hard-spheres substantially differs, as it is the case when $r_i \approx 0$ and r_j is large, the large hard-sphere has to be moved as much as the small hard-sphere so that the total volume of hard-spheres which is moved when applying κ is large. In the scenario of many hard-spheres surrounding these two hard-spheres, this teleportation rule leads to invalid proposed configurations with $\kappa(y) \notin A$ with high probability.

We describe another teleportation rule which is effective when $r_i \approx 0$ and r_j is large.

D.3.2 Move only the smaller hard-sphere

When $r_i \ll r_j$, we propose to move only the smaller hard-sphere by setting, for $\ell = 1, 2, \dots, N$

$$[\kappa(x)]^{(\ell)} = \begin{cases} 2x^{(j)} - x^{(i)} & \text{if } \ell = i \\ x^{(\ell)} & \text{otherwise} \end{cases}$$

see Figure D.1, middle panels for an illustration. The aim here is to not move the larger hard-sphere and therefore maximize the probability that $\kappa(x) \in A$ when the two hard-spheres are surrounded by other hard-spheres. This comes at the cost of having possibly low probability to accept the new points.

Finally we introduce a third teleportation rule, which is the one used in Section 5.3.2 and can bee seen as a compromise between the teleportation rules described in Section D.3.1 and Section D.3.2.

D.3.3 Move the smaller hard-sphere more than the larger one

For $\ell = 1, 2, \dots, N$, we set

$$[\kappa(x)]^{(\ell)} = \begin{cases} x^{(j)} + \frac{x^{(i)} - x^{(j)}}{r_i + r_j}(r_i - r_j) & \text{if } \ell = i \\ x^{(i)} + \frac{x^{(j)} - x^{(i)}}{r_i + r_j}(r_j - r_i) & \text{if } \ell = j \\ x^{(\ell)} & \text{otherwise,} \end{cases}$$

for $\alpha \in [0, 1]$. This teleportation rule moves the smaller hard-sphere more and the larger hard-sphere in order to increase the probability that $\kappa(x) \in A$ while trying to move as little as possible the center of the hard-spheres, see Figure D.1, bottom panel. Notice that, if $r_i = r_j$, this teleportation coincides with the teleportation in Section D.3.1.

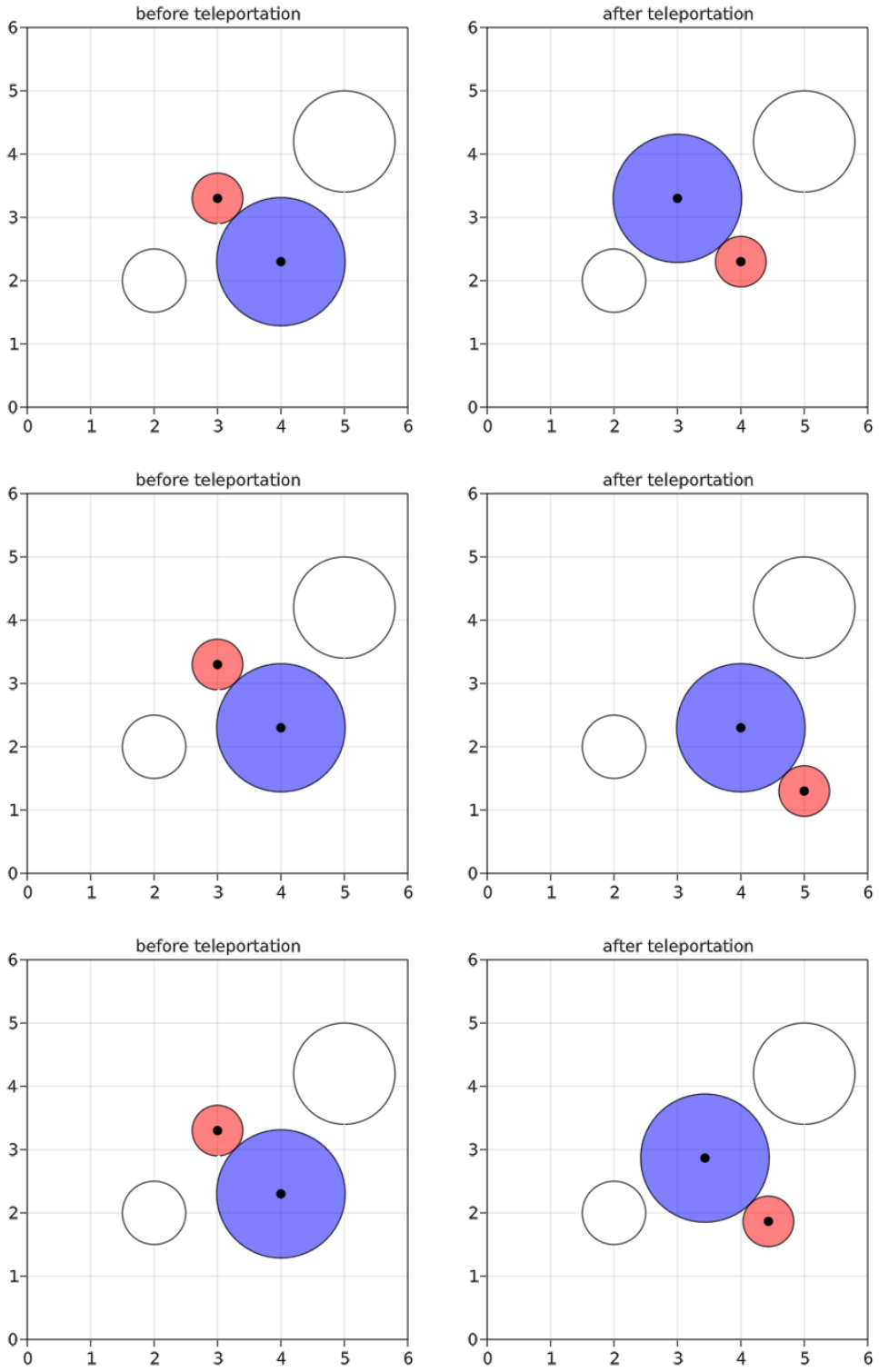


Figure D.1: Illustration of 3 different teleportation rules (from Top to Bottom panel). Left panels: configuration before teleportation; right panels: configuration after teleportation. Top panels: teleportation by swapping the centers of 2 hard-spheres. Middle panels: teleportation by moving only the smaller hard-sphere. Bottom-panel: teleportation which moves more the smaller hard-sphere than the bigger one.

This page is intentionally left blank.

Bibliography

[1] C. Andrieu, A. Durmus, N. Nüsken, and J. Roussel. *Hypocoercivity of Piecewise Deterministic Markov Process-Monte Carlo*. 2018. arXiv: 1808.08592.

[2] C. Andrieu and S. Livingstone. "Peskun-Tierney ordering for Markov chain and process Monte Carlo: beyond the reversible scenario". In: (2019). arXiv: 1906.06197.

[3] D. Bakry, I. Gentil, and M. Ledoux. "Analysis and Geometry of Markov Diffusion Operators". In: *Grundlehren der mathematischen Wissenschaften* (2014). ISSN: 2196-9701. DOI: 10.1007/978-3-319-00227-9. URL: http://dx.doi.org/10.1007/978-3-319-00227-9.

[4] J. Bento, M. Ibrahimi, and A. Montanari. "Learning Networks of Stochastic Differential Equations". In: (2010). arXiv: 1011.0415.

[5] A. Beskos, O. Papaspiliopoulos, G. O. Roberts, et al. "Retrospective exact simulation of diffusion sample paths with applications". In: *Bernoulli* 12.6 (2006), pp. 1077–1098.

[6] A. Beskos, G. O. Roberts, A. Stuart, and J. Voss. "MCMC methods for diffusion bridges". In: *Stoch. Dyn.* 8.3 (2008), pp. 319–350. URL: https://doi.org/10.1142/S0219493708002378.

[7] M. Betancourt. *A Conceptual Introduction to Hamiltonian Monte Carlo*. 2018. arXiv: 1701.02434 [stat.ME].

[8] J. Bierkens, P. Fearnhead, and G. Roberts. "The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data". In: *Ann. Statist.* 47.3 (June 2019), pp. 1288–1320.

[9] J. Bierkens, S. Grazzi, K. Kamatani, and G. Roberts. "The Boomerang Sampler". In: *International conference on machine learning*. PMLR. 2020, pp. 908–918.

[10] J. Bierkens, S. Grazzi, F. van der Meulen, and M. Schauer. "A piecewise deterministic Monte Carlo method for diffusion bridges". In: *Statistics and Computing* 31.3 (2021), pp. 1–21.

[11] J. Bierkens, S. Grazzi, F. v. d. Meulen, and M. Schauer. "Sticky PDMP samplers for sparse and local inference problems". In: *Statistics and Computing* 33.1 (2023), p. 8.

[12] J. Bierkens, S. Grazzi, M. Schauer, and G. Roberts. "Methods and applications of PDMP samplers with boundary conditions".

[13] J. Bierkens, K. Kamatani, and G. O. Roberts. *High-dimensional scaling limits of piecewise deterministic sampling algorithms*. 2018. arXiv: 1807.11358.

[14] J. Bierkens and S. M. V. Lunel. "Spectral analysis of the zigzag process". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 58. 2. Institut Henri Poincaré. 2022, pp. 827–860.

[15] J. Bierkens, F. van der Meulen, and M. Schauer. "Simulation of elliptic and hypo-elliptic conditional diffusions". In: *Advances in Applied Probability* 52.1 (2020), pp. 173–212. DOI: 10.1017/apr.2019.54.

[16] J. Bierkens, P. Nyquist, and M. C. Schlottke. "Large deviations for the empirical measure of the zig-zag process". In: *The Annals of Applied Probability* 31.6 (2021), pp. 2811–2843.

[17] J. Bierkens and G. Roberts. "A piecewise deterministic scaling limit of lifted Metropolis–Hastings in the Curie–Weiss model". In: *The Annals of Applied Probability* 27.2 (2017), pp. 846–882.

[18] J. Bierkens, G. O. Roberts, and P.-A. Zitt. "Ergodicity of the zigzag process". In: *The Annals of Applied Probability* 29.4 (2019), pp. 2266–2301.

[19] J. Bierkens et al. "Piecewise deterministic Markov processes for scalable Monte Carlo on restricted domains". In: *Statistics & Probability Letters* 136 (2018), 148–154. ISSN: 0167-7152. DOI: 10.1016/j.spl.2018.02.021. URL: http://dx.doi.org/10.1016/j.spl.2018.02.021.

[20] M. Bladt, M. Sørensen, et al. "Simple simulation of diffusion bridges with application to likelihood inference for diffusions". In: *Bernoulli* 20.2 (2014), pp. 645–675.

[21] N. Bou-Rabee and J. M. Sanz-Serna. "Randomized Hamiltonian Monte Carlo". In: *Ann. Appl. Probab.* 27.4 (2017), pp. 2159–2194. DOI: 10.1214/16-AAP1255. URL: https://doi.org/10.1214/16-AAP1255.

[22] A. Bouchard-Côté, S. J. Vollmer, and A. Doucet. "The Bouncy Particle Sampler: A Non-Reversible Rejection-Free Markov Chain Monte Carlo Method". In: *Journal of the American Statistical Association* 113.522 (2018), pp. 855–867.

[23] H. J. Brascamp and E. H. Lieb. "On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation". In: *J. Functional Analysis* 22.4 (1976), pp. 366–389. DOI: 10.1016/0022-1236(76)90004-5. URL: https://doi.org/10.1016/0022-1236(76)90004-5.

[24] S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, eds. *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2011, pp. xxvi+592. ISBN: 978-1-4200-7941-8. DOI: 10.1201/b10905. URL: https://doi.org/10.1201/b10905.

[25] T.-L. Chen and C.-R. Hwang. "Accelerating reversible Markov chains". In: *Statistics & Probability Letters* 83.9 (2013), pp. 1956–1962. ISSN: 0167-7152. DOI: https://doi.org/10.1016/j.spl.2013.05.002. URL: https://www.sciencedirect.com/science/article/pii/S0167715213001533.

[26] A. Chevallier, P. Fearnhead, and M. Sutton. *Reversible Jump PDMP Samplers for Variable Selection*. 2020. arXiv: 2010.11771.

[27] A. Chevallier, S. Power, A. Q. Wang, and P. Fearnhead. "PDMP Monte Carlo methods for piecewise-smooth densities". In: *arXiv preprint arXiv:2111.05859* (2021).

[28] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. "MCMC methods for functions: modifying old algorithms to make them faster". In: *Statistical Science* (2013), pp. 424–446.

[29] M. H. A. Davis. *Markov models and optimization*. Vol. 49. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1993.

[30] G. Deligiannidis, D. Paulin, and A. Doucet. "Randomized Hamiltonian Monte Carlo as Scaling Limit of the Bouncy Particle Sampler and Dimension-Free Convergence Rates". In: *preprint arXiv:1808.04299* (2018). arXiv: 1808.04299. URL: http://arxiv.org/abs/1808.04299.

[31] P. Diaconis, S. Holmes, and R. M. Neal. "Analysis of a nonreversible Markov chain sampler". In: *Annals of Applied Probability* (2000), pp. 726–752.

[32] A. Doucet, N. De Freitas, N. J. Gordon, et al. *Sequential Monte Carlo methods in practice*. Vol. 1. 2. Springer, 2001.

[33] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. "Hybrid Monte Carlo". In: *Physics letters B* 195.2 (1987), pp. 216–222.

[34] M. F. Faulkner and S. Livingstone. "Sampling algorithms in statistical physics: a guide for statistics and machine learning". In: *arXiv preprint arXiv:2208.04751* (2022).

[35] M. F. Faulkner, L. Qin, A. C. Maggs, and W. Krauth. "All-atom computations with irreversible Markov chains". In: *The Journal of Chemical Physics* 149.6 (Aug. 2018), p. 064113. ISSN: 1089-7690. DOI: 10.1063/1.5036638. URL: http://dx.doi.org/10.1063/1.5036638.

[36] P. Fearnhead, J. Bierkens, M. Pollock, and G. O. Roberts. "Piecewise Deterministic Markov Processes for Continuous-Time Monte Carlo". In: *Statist. Sci.* 33.3 (Aug. 2018), pp. 386–412. DOI: 10.1214/18-STS648. URL: https://doi.org/10.1214/18-STS648.

[37] H. Ge, K. Xu, and Z. Ghahramani. "Turing: a language for flexible probabilistic inference". In: *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*. 2018, pp. 1682–1690. URL: http://proceedings.mlr.press/v84/ge18b.html.

[38] E. I. George and R. E. McCulloch. "Variable selection via Gibbs sampling". In: *Journal of the American Statistical Association* 88.423 (1993), pp. 881–889.

[39] C. J. Geyer. "Markov chain Monte Carlo lecture notes". In: *Course notes, Spring Quarter* 80 (1998).

[40] C. J. Geyer and A. Mira. "On non-reversible Markov chains". In: *Monte Carlo methods (Toronto, ON, 1998)* 26 (2000), pp. 95–110.

[41] S. Grazzi and M. Schauer. *Boid animation*. https://youtu.be/O1VoURPwVLI. Youtube, 2021. URL: https://youtu.be/O1VoURPwVLI.

[42] P. J. Green. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* (1995), p. 16.

[43] P. J. Green and D. I. Hastie. "Reversible jump MCMC". In: *Genetics* 155.3 (2009), pp. 1391–1403.

[44] J. E. Griffin and P. J. Brown. "Bayesian global-local shrinkage methods for regularisation in the high dimension linear model". In: *Chemometrics and Intelligent Laboratory Systems* (2021), p. 104255.

[45] G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford University Press, 2001.

[46] Y. Guan and M. Stephens. "Bayesian variable selection regression for genome-wide association studies and other large-scale problems". In: *The Annals of Applied Statistics* 5.3 (2011), pp. 1780–1815.

[47] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: (1970).

[48] M. D. Hoffman and A. Gelman. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1593–1623.

[49] P. Holderrieth. "Cores for Piecewise-Deterministic Markov Processes used in Markov Chain Monte Carlo". In: (2019). arXiv: `1910.11429`. URL: `http://arxiv.org/abs/1910.11429`.

[50] I. A. Ibragimov and Y. V. Linnik. *Independent and stationary sequences of random variables*. With a supplementary chapter by I. A. Ibragimov and V. V. Petrov, Translation from the Russian edited by J. F. C. Kingman. Wolters-Noordhoff Publishing, Groningen, 1971, p. 443.

[51] H. Ishwaran and J. S. Rao. "Spike and slab variable selection: Frequentist and Bayesian strategies". In: *The Annals of Statistics* 33.2 (2005), 730–773. ISSN: 0090-5364.

[52] S. F. Jarner and E. Hansen. "Geometric ergodicity of Metropolis algorithms". In: *Stochastic processes and their applications* 85.2 (2000), pp. 341–361.

[53] C. P. Jewell, T. Kypraios, P. Neal, and G. O. Roberts. "Bayesian analysis for emerging infectious diseases". In: *Bayesian analysis* 4.3 (2009), pp. 465–496.

[54] JuliaCon 2020 by Jesse Bettencourt. *"JuliaCon 2020 | Boids: Dancing with Friends and Enemies"*. 2020. URL: `https://www.youtube.com/watch?v=8gS6wejsGsY`.

[55] I Karatzas and S. E. Shreve. "Brownian motion and stochastic calculus". In: *Graduate texts in Mathematics* 113 (1991).

[56] F. C. Klebaner. *Introduction to stochastic calculus with applications*. World Scientific Publishing Company, 2005.

[57] W. Krauth. *Statistical mechanics: algorithms and computations*. Vol. 13. OUP Oxford, 2006.

[58] D. A. Levin and Y. Peres. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc., 2017.

[59] X. Liang, S. Livingstone, and J. Griffin. "Adaptive random neighbourhood informed Markov chain Monte Carlo for high-dimensional Bayesian variable Selection". In: *arXiv preprint arXiv:2110.11747* (2021).

[60] T. M. Liggett. *Continuous time Markov processes*. Vol. 113. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2010.

[61] R. Liptser, B. Aries, and A. Shiryaev. *Statistics of Random Processes: I. General Theory*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2013. ISBN: 9783662130438.

[62] L. Lorenzi and M. Bertoldi. *Analytical methods for Markov semigroups*. Vol. 283. Pure and Applied Mathematics (Boca Raton). Chapman & Hall/CRC, Boca Raton, FL, 2007, pp. xxxii+526.

[63] J. Markovic and A. Sepehri. "Bouncy Hybrid Sampler as a Unifying Device". In: *preprint arXiv:1802.04366* (2018). URL: http://arxiv.org/abs/1802.04366.

[64] H. P. McKean. *Stochastic integrals*. Vol. 353. American Mathematical Soc., 1969.

[65] N. Metropolis et al. "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[66] F. van der Meulen and M. Schauer. "Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals". In: *Electron. J. Statist.* 11.1 (2017), pp. 2358–2396. DOI: 10.1214/17-EJS1290. URL: https://doi.org/10.1214/17-EJS1290.

[67] F. van der Meulen, M. Schauer, and J. van Waaij. "Adaptive nonparametric drift estimation for diffusion processes using Faber–Schauder expansions". In: *Statistical Inference for Stochastic Processes* 21.3 (2018), pp. 603–628.

[68] S. P. Meyn and R. L. Tweedie. "Stability of Markovian processes II: Continuous-time processes and sampled chains". In: *Advances in Applied Probability* 25.3 (1993), pp. 487–517.

[69] M. Michel, S. C. Kapfer, and W. Krauth. "Generalized event-chain Monte Carlo: Constructing rejection-free global-balance algorithms from infinitesimal steps". In: *The Journal of chemical physics* 140.5 (2014), p. 054116.

[70] M. Michel, X. Tan, and Y. Deng. "Clock Monte Carlo methods". In: *Physical Review E* 99.1 (Jan. 2019). ISSN: 2470-0053. DOI: 10.1103/physreve.99.010105. URL: http://dx.doi.org/10.1103/PhysRevE.99.010105.

[71] L. Miclo and P. Monmarché. "Étude spectrale minutieuse de processus moins indécis que les autres". In: *Séminaire de probabilités xlv*. Springer, 2013, pp. 459–481.

[72] M. Mider, M. Schauer, and F. van der Meulen. *Continuous-discrete smoothing of diffusions*. 2020. arXiv: 1712.03807.

[73] M. Mider et al. *Simulating bridges using confluent diffusions*. 2019. arXiv: 1903.10184.

[74] T. J. Mitchell and J. J. Beauchamp. "Bayesian variable selection in linear regression". In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1023–1032.

[75] J. Møller, M. L. Huber, and R. L. Wolpert. "Perfect simulation and moment properties for the Matérn type III process". In: *Stochastic Processes and their Applications* 120.11 (2010), pp. 2142–2158.

[76] A. Monemvassitis, A. Guillin, and M. Michel. "PDMP characterisation of event-chain Monte Carlo algorithms for particle systems". In: *arXiv preprint arXiv:2208.11070* (2022).

[77] 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". In: *The Journal of Chemical Physics* 153.2 (July 2020), p. 024101. ISSN: 1089-7690. DOI: 10.1063/5.0005060. URL: http://dx.doi.org/10.1063/5.0005060.

[78] J. Moriarty, J. Vogrinc, and A. Zocca. "A Metropolis-class sampler for targets with non-convex support". In: (2020).

[79] I. Murray, R. Adams, and D. MacKay. "Elliptical slice sampling". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Y. W. Teh and M. Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, 2010, pp. 541–548. URL: http://proceedings.mlr.press/v9/murray10a.html.

[80] R. M. Neal et al. "MCMC using Hamiltonian dynamics". In: *Handbook of markov chain monte carlo* 2.11 (2011), p. 2.

[81] H. Niederreiter. *Random number generation and quasi-Monte Carlo methods*. SIAM, 1992.

[82] A. Nishimura, D. B. Dunson, and J. Lu. "Discontinuous Hamiltonian Monte Carlo for discrete parameters and discontinuous likelihoods". In: *Biometrika* 107.2 (2020), 365–380. ISSN: 1464-3510. DOI: 10.1093/biomet/asz083. URL: http://dx.doi.org/10.1093/biomet/asz083.

[83] E. A. J. F. Peters and G. de With. "Rejection-free Monte Carlo sampling for general potentials". In: *Physical Review E* 85.2 (Feb. 2012). ISSN: 1550-2376. DOI: 10.1103/physreve.85.026703. URL: http://dx.doi.org/10.1103/PhysRevE.85.026703.

[84] N. G. Polson, J. G. Scott, and J. Windle. "Bayesian inference for logistic models using Pólya–Gamma latent variables". In: *Journal of the American statistical Association* 108.504 (2013), pp. 1339–1349.

[85] K. Ray, B. Szabo, and G. Clara. "Spike and slab variational Bayes for high dimensional logistic regression". In: (2020). arXiv: 2010.11665.

[86] C. W. Reynolds. "Flocks, Herds and Schools: A Distributed Behavioral Model". In: *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*. Association for Computing Machinery, 1987, 25–34.

[87] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Vol. 2. Springer, 1999.

[88] G. O. Roberts and J. S. Rosenthal. "General state space Markov chains and MCMC algorithms". In: *Probability surveys* 1 (2004), pp. 20–71.

[89] G. O. Roberts and J. S. Rosenthal. "Optimal scaling of discrete approximations to Langevin diffusions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1 (1998), pp. 255–268.

[90] G. O. Roberts and O. Stramer. "On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm". In: *Biometrika* 88.3 (2001), pp. 603–621.

[91] G. O. Roberts and R. L. Tweedie. "Exponential Convergence of Langevin Distributions and Their Discrete Approximations". In: *Bernoulli* 2.4 (1996), pp. 341–363. ISSN: 13507265. URL: `http://www.jstor.org/stable/3318418`.

[92] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*. Vol. 2. Cambridge university press, 2000.

[93] L. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales: Volume 1, Foundations*. Cambridge Mathematical Library. Cambridge University Press, 2000.

[94] J. S. Rosenthal. "Asymptotic variance and convergence rates of nearly-periodic Markov chain Monte Carlo algorithms". In: *Journal of the American Statistical Association* 98.461 (2003), pp. 169–177.

[95] M. Schauer and S. Grazzi. *mschauer/ZigZagBoomerang.jl: v0.6.0*. Version v0.6.0. Mar. 2021. URL: `https://doi.org/10.5281/zenodo.4601534`.

[96] W. Shi, S. Ghosal, and R. Martin. "Bayesian estimation of sparse precision matrices in the presence of Gaussian measurement error". In: *Electronic Journal of Statistics* 15.2 (2021), pp. 4545–4579.

[97] M. Sutton and P. Fearnhead. "Concave-Convex PDMP-based sampling". In: *arXiv preprint arXiv:2112.12897* (2021).

[98] A. Terenin and D. Thorngren. "A Piecewise Deterministic Markov Process via (r, θ) swaps in hyperspherical coordinates". In: *preprint arXiv:1807.00420* (2018). arXiv: `1807.00420`. URL: `http://arxiv.org/abs/1807.00420`.

[99] R. Thisted. *Elements of Statistical Computing: NUMERICAL COMPUTATION*. Elements of Statistical Computing v. 1. Taylor & Francis, 1988. ISBN: 9780412013713. URL: https://books.google.nl/books?id=b9uoIswZUcUC.

[100] R. Tibshirani et al. "Sparsity and smoothness via the fused lasso". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1 (2005), pp. 91–108.

[101] P. Vanetti, A. Bouchard-Côté, G. Deligiannidis, and A. Doucet. "Piecewise-Deterministic Markov Chain Monte Carlo". In: (2017). arXiv: 1707.05296 [stat.ME].

[102] G. Vasdekis and G. O. Roberts. "Speed Up Zig-Zag". In: *arXiv preprint arXiv:2103.16620* (2021).

[103] G. Zanella and G. Roberts. "Scalable importance tempering and Bayesian variable selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.3 (2019), pp. 489–517.

This page is intentionally left blank.

Summary (Dutch)

Markov Chain Monte Carlo methoden zijn de meest gebruikte algoritmen voor exacte Bayesiaanse inferentie. Deze methoden bestaan uit het simuleren van een Markov-keten die convergeert naar een gewenste Bayesiaanse a-posteriori kansmaat. Het gesimuleerde traject van deze Markov-keten kan vervolgens ook gebruikt worden om verwachtingen van functies ten opzichte van die gewenste maat te benaderen. Wij beschouwen Monte Carlo methoden gebaseerd op Piecewise deterministic Markov processes (PDMP samplers). PDMP samplers zijn continue-tijd stochastische processen die per constructie niet reversibel zijn. Deze niet-reversibiliteitseigenschap kan de prestaties van simulatiemethoden verbeteren, zowel wat betreft convergentie naar stationariteit als in termen van asymptotische variantie. In Hoofdstuk 1 introduceren wij PDMPs. Hoofdstuk 2 gaat over de simulatie van een-dimensionale diffusiebruggen. De voorgestelde methodologie berust op het uitbreiden van de ruimte van de diffusiebruggen met een geschikte afgeknotte basis en het toepassen van de Zig-Zag sampler op de hoog-dimensionale coëfficiëntenruimte. In Hoofdstuk 3 introduceeen we de Boomerang-sampler, een nieuwe PDMP-sampler die beter presteert dan bestaande PDMP-samplers voor stationaire verdelingen uitgedrukt in termen van hoog-dimensionale Gaussische kansmaten. De Boomerang sampler volgt elliptische deterministische trajecten wat voordelig is voor verdelingen met een Gaussische referentiemaat. Een belangrijke toepassing is de simulatie van diffusiebruggen met de in Hoofdstuk 2 geïntroduceerde methode, aangezien de niet-genormaliseerde dichtheid een hoog-dimensionale Gaussische referentiemaat heeft. In Hoofdstuk 4 construeren we een nieuwe klasse efficiënte Monte Carlo methoden op basis van PDMP's die geschikt zijn voor inferentie van hoog-dimensionale verdelingen met zowel continue als atomaire componenten. Dit wordt bereikt met het vrij eenvoudige idee om bestaande PDMP-samplers te voorzien van "plakkende" coördinaat-assen, coördinaatvlakken. Bij het raken van die deelruimten treedt een gebeurtenis op waarbij het proces in die deelruimte vast blijft, waardoor het proces enige tijd in een deelmodel doorbrengt. Tenslotte presenteert Hoofdstuk 5 enkele resultaten over de toepassing van PDMP-samplers met randvoorwaarden. De belangrijkste motiverende toepassingen zijn gebaseerd op het SIR-model in de epidemiologie, dat wordt gebruikt om de verspreiding van ziekten te beschrijven, en op modellen van harde bollen die van belang zijn in de statistische mechanica.

This page is intentionally left blank.

Summary (English)

Markov Chain Monte Carlo methods are the most popular algorithms used for exact Bayesian inference problems. These methods consist of simulating a Markov chain which converges to a desired Bayesian posterior measure and use the simulated trajectory to approximate expectations of functionals relative to that measure. We consider Monte Carlo methods based on Piecewise deterministic Markov processes (PDMP samplers). PDMP samplers are continuous-time processes that are non-reversible by construction. Non-reversibility may improve the performance of sampling methods, both in terms of convergence to stationarity and asymptotic variance. In Chapter 1 we give a concise presentation which motivates and introduces PDMPs. Chapter 2 is about the simulation of one-dimensional diffusion bridges. The methodology proposed relies on expanding the space of diffusion bridges with a suitable truncated basis and applying the Zig-Zag sampler on the high-dimensional coefficient space. In Chapter 3 we introduce the Boomerang sampler as a new PDMP sampler which outperforms existing PDMP samplers for target measures expressed in terms of high dimensional Gaussian measures. The Boomerang sampler has elliptical deterministic dynamics which preserves Gaussian measures at barely no cost. A key application is the simulation of diffusion bridges with the method introduced in Chapter 2 as the unnormalised density is relative to a high dimensional Gaussian measure. In chapter 4, we construct a new class of efficient Monte Carlo methods based on PDMPs suitable for inference in high dimensional mixtures of continuous and atomic components. This is achieved with the fairly simple idea of endowing existing PDMP samplers with "sticky" coordinate axes and coordinate hyper-planes. Upon hitting those subspaces, an event is triggered during which the process *sticks* to the subspace, this way spending some time in a sub-model. Finally, Chapter 5 presents some results on the application of PDMP samplers with boundary conditions. The key motivating applications are based on the SIR model in epidemiology used for describing the spread of diseases and hard-spheres models which are of interest in statistical mechanics.

This page is intentionally left blank.

Acknowledgement

The work presented in this thesis is part the research programme *Bayesian inference for high dimensional processes* with project number 613.009.034c, which is financed by the Dutch Research Council (NWO) under the *Stochastics – Theoretical and Applied Research* (STAR) grant and by the Delft Institute of Applied Mathematics (DIAM), TU Delft.

This page is intentionally left blank.

Curriculum vitæ

Sebastiano Grazzi was born on the 18th of May 1994 in Cattolica, Italy. He obtained a music degree in Oboe at *Conservatorio G. Rossini* (2012) and a high school diploma in classical studies at *Liceo Classico T. Mamiani* (2013). In 2016 he obtained a Bachelor's degree in Statistical Sciences at the *University of Bologna* and in 2018 he obtained a double Master's degree in mathematical finance at the *University of Bologna* and at *Ludwig Maximilian University of Munich* with final thesis titled: *"Rough Models for Realized Volatility"* under the supervision of Francesca Biagini and Umberto Cherubini. Between 2012 and 2016 he worked in the summer seasons as a lifeguard in Pesaro, as a sailing instructor in Ascona, Switzerland and as a crew member on a sailing boat in Portofino. During his highier education, he joined (with scholarship) the Erasmus Exchange programs at *Plymouth University* (2015-2016), at *Ludwig Maximilian University of Munich* (2018) and the Oversease Exchange program at the *University of Technology, Sydney* (2017). He also obtained a Post-Graduate Diploma in Mathematical Finance at the Department of Mathematics of the *University of Bologna* (2017-2018). From 2018 to 2022, he worked as a PhD candidate at Delft Institute of Applied Mathematics (DIAM), *Delft University of Technology* with project *"Bayesian inference for high dimensional diffusion processes"* under the supervision of Frank van der Meulen, Joris Bierkens and Moritz Schauer. From 2019 to 2022, he was also instructor of the course *Introduction to Statistics* of the Bachelor's program in mathematics. In September 2022, He was hired within the *CoSInES* grant (Computational Statistical Inference for Engineering and Security) at the *University of Warwick* as a Research Associate under the supervision of Gareth Roberts.

This page is intentionally left blank.

List of publications

Published articles:

Bayesian inference for SDE models: a case study for an excitable stochastic-dynamical model, with F. van der Meulen, M. Mider, M. Schauer. 2020. Nextjournal, 2020 [`https://nextjournal.com/Lobatto/FitzHugh-Nagumo`]

J. Bierkens, S. Grazzi, K. Kamatani, and G. Roberts. "The Boomerang Sampler". In: *International conference on machine learning.* PMLR. 2020, pp. 908–918

J. Bierkens, S. Grazzi, F. van der Meulen, and M. Schauer. "A piecewise deterministic Monte Carlo method for diffusion bridges". In: *Statistics and Computing* 31.3 (2021), pp. 1–21.

J. Bierkens, S. Grazzi, F. v. d. Meulen, and M. Schauer. "Sticky PDMP samplers for sparse and local inference problems". In: *Statistics and Computing* 33.1 (2023), p. 8. To appear in: *Statistics and Computing.*

In Preparation:

J. Bierkens, S. Grazzi, M. Schauer, and G. Roberts. "Methods and applications of PDMP samplers with boundary conditions", *In preparation.*