Statistics and Computing (2023) 33:8
https://doi.org/10.1007/s11222-022-10180-5

ORIGINAL PAPER

Sticky PDMP samplers for sparse and local inference problems

Joris Bierkens[1] · Sebastiano Grazzi[2] · Frank van der Meulen[3] · Moritz Schauer[4,5]

Received: 28 June 2022 / Accepted: 10 November 2022

Abstract
We construct a new class of efficient Monte Carlo methods based on continuous-time piecewise deterministic Markov processes (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. Compared to the Gibbs sampler for variable selection, we heuristically derive favourable dependence of the Sticky Zig-Zag sampler on dimension and data size. The computational efficiency of the Sticky Zig-Zag sampler is further established through numerical experiments where both the sample size and the dimension of the parameter space are large.

Keywords Bayesian variable selection · Piecewise deterministic Markov process · Monte Carlo · Spike-and-slab · Big-data · High-dimensional problems · Non-reversible jump

1 Introduction

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

$$\mu_0(\mathrm{d}x) = \prod_{i=1}^{d} (w_i \pi_i(x_i)\mathrm{d}x_i + (1 - w_i)\delta_0(\mathrm{d}x_i)) \,. \tag{1.1}$$

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 et al. 2019a, 2020).

We introduce "sticking event times", which occur every time a coordinate of the process state hits 0. At such a time

✉ Sebastiano Grazzi
sebastiano.grazzi@warwick.ac.uk

1 Delft Institute of Applied Mathematics (DIAM), Delft University of Technology, Delft, The Netherlands

2 Department of Statistics, University of Warwick, Coventry, United Kingdom

3 Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

4 Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden

5 University of Gothenburg, Gothenburg, Sweden

Published online: 28 November 2022

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 Fig. 1 for an illustration of the sticky version of the Zig-Zag sampler (Bierkens et al. 2019a). 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 pre-specified measure μ as invariant measure, which we assume to be of the form

$$\mu(\mathrm{d}x) = C_\mu \exp(-\Psi(x)) \prod_{i=1}^{d} \left(\mathrm{d}x_i + \frac{1}{\kappa_i} \delta_0(\mathrm{d}x_i) \right) \tag{1.2}$$

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 Eq. (1.1), the posterior measure is of the form of Eq. (1.2) with

$$\Psi(x) = C - \ell(x) - \sum_{i=1}^{d} \log(\pi_i(x_i)),$$
$$\kappa_i = \frac{w_i}{1 - w_i} \pi_i(0) \tag{1.3}$$

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.

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 $T_0 = \mathrm{Leb}\{0 \le t \le T : x_1(t) = 0\}$. This simple, yet general idea requires the user only to specify $\{\kappa_i\}_{i=1}^d$ and Ψ as in Eq. (1.2). 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.

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 (1.2). 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 et al. (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 et al. (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 (1.2). 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

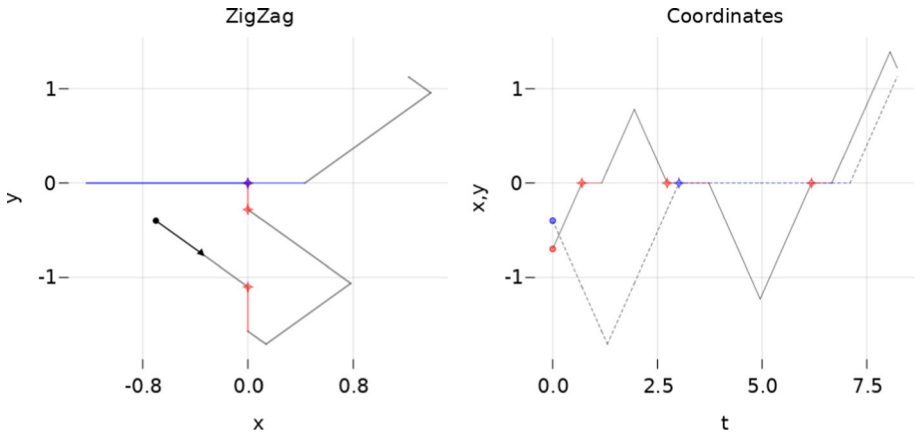

Fig. 1 Two-dimensional Sticky Zig-Zag sampler with initial position $(-0.75, -0.4)$ and initial velocity $(+1, -1)$. On the left panel, a trajectory on the (x, y)-plane of the Sticky Zig-Zag sampler. The sticky event times relative to the x (respectively y) coordinate and the trajectories with the x (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

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 et al. (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 et al. (2020) and our work may be found in Appendix C.

1.3 Contributions

- We show how to construct sticky PDMP samplers from ordinary PDMP samplers for sampling from the measure in Eq. (1.2). 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 Sect. 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.

1.4 Outline

Section 2 formally introduces sticky PDMP samplers and gives the main theoretical results for the sticky Zig-Zag sampler. In Sect. 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 Sect. 3 we extend the Gibbs sampler for variable selection for target measures of the form of Eq. (1.2). We analyse and compare the computational complexity and the mixing times of both the sticky Zig-Zag sampler and the Gibbs sampler. Section 4 presents four statistical examples with simulated data and analyses the outputs after applying the algorithms considered in this article. In Sect. 5 both limitations and promising research directions are discussed.

There are five appendices. The derivation of our theoretical results is given in Appendix A. Appendix B extends some of the theoretical results for two other sticky samplers: the sticky version of the Bouncy particle sampler (Bouchard-Côté et al. 2018) and the Boomerang sampler (Bierkens et al. 2020), the latter having Hamiltonian deterministic dynamics invariant to a prescribed Gaussian measure. Appendix C contains a self-contained discussion with heuristic arguments and simulations which highlight the differences between the sticky PDMPs and the method of Chevallier et al. (2020). Appendix D complements Sect. 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 E contains some of the details of the numerical examples of Sect. 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

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

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 $f: E \to \mathbb{R}$ and let $C(E) = \{f \in \mathcal{M}(E): f \text{ is continuous }\}$. 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 $\mu(\mathrm{d}x) = \int_{\mathcal{Y}} \mu(\mathrm{d}x, \mathrm{d}y)$.

2 Sticky PDMP samplers

In what follows, we formally describe the sticky PDMP samplers (Sect. 2.1) and give the main theoretical results obtained for the sticky Zig-Zag sampler (Sect. 2.3). Section 2.4 extends the sticky Zig-Zag sampler with subsampling methods.

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 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

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

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

$$\alpha(x, v) = \{i \in \{1, 2, \ldots, d\}: (x, v) \notin \mathfrak{F}_i\} \tag{2.1}$$

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

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

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 $\mathcal{Q}: E \times \mathcal{B}(E) \to [0, 1]$, at random times sampled with state dependent intensity $\lambda: E \to [0, \infty)$. The deterministic dynamics are determined coordinate-wise by the integral equation

$$(x_i(t), v_i(t)) = (x_i(s), v_i(s)) + \int_s^t \xi_i(x_i(r), v_i(r))\mathrm{d}r, \quad i = 1, 2, \ldots, d, \tag{2.2}$$

with ξ_i being state dependent with form

$$\xi_i(x, v) = \begin{cases} \bar{\xi}_i(x_i, v_i) & (x_i, v_i) \notin \mathfrak{F}_i \\ (0, 0) & (x_i, v_i) \in \mathfrak{F}_i, \end{cases} \tag{2.3}$$

for functions $\bar{\xi}_i: \overline{\mathbb{R}} \times \mathbb{R} \to \overline{\mathbb{R}} \times \mathbb{R}$ which depend on the specific PDMP chosen and corresponds to the coordinate-wise dynamics of the ordinary PDMP while the second case in Eq. (2.3) 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 $\lambda_1, \lambda_2, \ldots,$

[1] A function $f: \overline{\mathbb{R}} \to \mathbb{R}$ 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.

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 (2.2). Upon hitting $\mathfrak{F}_i$, the ith coordinate process freezes, captured by the state dependence of (2.3).
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 inter-arrival 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 (2.2) 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 1.

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

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 $\mathcal{V} = \{v : |v_i| = a_i, \forall i \in \{1, 2, \ldots, d\}\}$ for a fixed vector a. For each index i, the deterministic dynamics of Eq. (2.3) 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

Algorithm 1 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 $\Delta_k \sim \text{Poiss}(s \to \lambda(\varphi(s, X(\tau_k), V(\tau_k))))$, with
$$\mathbb{P}(\Delta_k \geq t) = \exp\left(-\int_0^t \lambda(\varphi(s, X(\tau_k), V(\tau_k)) \mathrm{d}s\right). \tag{2.4}$$
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
$$(X(\tau_{k+1}), V(\tau_{k+1})) \sim \mathcal{Q}(\varphi(\Delta_k, X(\tau_k), V(\tau_k)), \cdot).$$

ith coordinate is determined by the inhomogeneous rate

$$\lambda_{i,\text{refl}}(x, v) = \mathbb{1}_{i \in \alpha(x,v)}(v_i \partial_i \Psi(x))^+. \tag{2.5}$$

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 et al. (2019b), the Zig-Zag sampler does not require refreshment events in general to be ergodic.

2.3 Theoretical aspects of the Sticky Zig-Zag sampler

A theoretical analysis of the sticky Zig-Zag sampler is given in "Appendix A.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 et al. (2019a). 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 (2.2) starting in (x, v) and τ is the natural topology on E. Define the operator $\mathcal{A}$ with domain

$$\mathcal{D}(\mathcal{A}) = \{f \in \mathcal{M}(E) : t \mapsto f(\varphi(t, x, v))\tau$$
-absolutely continuous $\forall (x, v)$ and

$$\forall i : \lim_{t \downarrow 0} f(x[i : 0^+ + t], \cdot) = f(x[i : 0^+], \cdot),$$
$$\lim_{t \downarrow 0} f(x[i : 0^- - t], \cdot) = f(x[i : 0^-], \cdot)\}$$

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

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

Proposition 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 1. □

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 A.6, Appendix 1 for details.

Theorem 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*

$$\mu(\mathrm{d}x, \mathrm{d}v) = \frac{1}{C} \sum_{u \in \mathcal{V}} \exp(-\Psi(x)) \prod_{i=1}^{d} \left(\mathrm{d}x_i + \frac{1}{\kappa_i} \left(\mathbb{1}_{v_i > 0} \, \delta_{0^-}(\mathrm{d}x_i) + \mathbb{1}_{v_i < 0} \, \delta_{0^+}(\mathrm{d}x_i) \right) \delta_u(\mathrm{d}v) \right), \tag{2.6}$$

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 1. We then prove that the process is Feller and strong Markov ("Appendix A.2" and "Appendix A.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 A.5. □

Theorem 2.3 *Suppose Ψ satisfies Assumption A.8. Then the sticky Zig-Zag process is ergodic and μ is its unique stationary measure.*

Proof See Appendix 1. □

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 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 A.7" shows that the expected time of the process to return to the null model is

$$\frac{1 - \mu(\{\mathbf{0}\})}{d\kappa \mu(\{\mathbf{0}\})}. \tag{2.7}$$

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 et al. 2019a) 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 (1.2). To that end, assume the following decomposition:

$$\partial_{x_i} \Psi(x) = \left(\sum_{j=1}^{N_i} S(x, i, j) \right), \quad \forall x \in \overline{\mathbb{R}}^d, \quad i = 1, 2, \ldots, d, \tag{2.8}$$

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

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

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

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

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

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

which is specified by the user and such that Poisson times with inhomogeneous rate $\tau \sim \text{Poiss}(s \to \overline{\lambda}_i(s, x, v))$ can be simulated (see "Appendix D.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 Sect. 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 2.5 *The Sticky Zig-Zag with subsampling has a unique stationary measure given by Eq.* (2.6).

The proof of Proposition 2.5 follows with a similar argument made in the proof of Bierkens et al. (2019a, 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 pre-computed). 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 et al. (2019a, Section 5) and Bierkens et al. (2020, Section 3) for large N and when choosing x^* to be the mode of the posterior density.

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 (Eq. (2.1) 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 et al. (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 Sect. 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 Tables 1 and 2. The technical details of this section are given in "Appendix D".

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 $\mathbb{R}^d$:

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

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 Eq. (1.2) are given by

$$B_j(\alpha) = \frac{\mu(\mathcal{M}_{\alpha \cup \{j\}})}{\mu(\mathcal{M}_{\alpha \setminus \{j\}})} = \frac{\kappa_j \int_{\mathbb{R}^{|\alpha \cup \{j\}|}} \exp(-\Psi(y)) \mathrm{d}x_{\alpha \cup \{j\}}}{\int_{\mathbb{R}^{|\alpha \setminus \{j\}|}} \exp(-\Psi(z)) \mathrm{d}x_{\alpha \setminus \{j\}}}, \tag{2.1}$$

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

1. Update α by choosing randomly $j \sim \text{Unif}(\{1, 2, \ldots, d\})$ 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 1, we give an analytical expressions for the right hand-side of Eq. (2.1) 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 et al. 2013 for details).

Table 1 Computational scaling of the Sticky Zig-Zag algorithm and the Gibbs sampler for variable selection for p and sample size N

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

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

Table 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)

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

The results are for targets which are supported in every model and for targets supported on a single sequence of nested sub-models

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 D.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 Sect. 3.1, benefits when the conditional dependence structure of the target is sparse. In "Appendix D.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 $\alpha(t) := \{i : x_i(t) \neq 0\}$ 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 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 Sect. 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 Sect. 4 (Figs. 5 and 8).

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 D.4". Table 2 summarizes the scaling results (in terms of dimensions d) derived in the two cases considered.

4 Examples

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

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.

4.2 *(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.

4.3 *(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 *(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 D.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:

$$c \mapsto \mathcal{E}_{\mathrm{s}}(c) := \sum_{i=1}^{d} (p_i^{\mathrm{s}}(c) - \overline{p}_i)^2, \tag{2.1}$$

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 16 GB DDR4 RAM. Pre-processing time and memory allocation of both algorithms are comparable.

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

$$\begin{aligned} \mathrm{d}U_i(s) = -\lambda U_i(s)\mathrm{d}s + \sum_{j \neq i} x_{i,j}(U_j(s) \\ - U_i(s))\mathrm{d}s + \sigma \mathrm{d}W_i(s), \quad 1 \leq i \leq p \end{aligned} \tag{2.2}$$

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 $x = \{x_{i,j} : i \neq j\} \in \mathbb{R}^{p^2-p}$ 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 et al. (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). 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 Eq. (2.2) 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

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

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

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

$$-\frac{1}{2\sigma^2}\int_0^T \|C(x)Y(s)\|^2 \mathrm{d}s. \tag{2.3}$$

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 2000, Ch. IV, Sect. 47) and time integrals. We assume x to be sparse which is incorporated by choosing a spike-and-slab prior for x as in Eq. (1.1). The posterior measure is of the form of (1.2) with κ and $\Psi(x)$ as in (1.3). 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 $j_1, j_2 \sim \mathrm{Unif}(\{1, 2, \ldots, d\}\backslash i)$ 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 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 Fig. 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 s computing time).

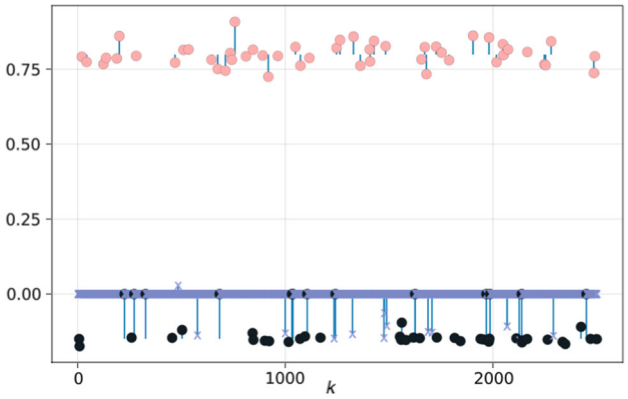

Fig. 2 Posterior median estimate of x_k (where k can be identified with (i, j)) versus k computed using the Sticky Zig-Zag sampler. Thin vertical lines indicate distance to the truth. True zeros are plotted with the symbol $\times$, others are plotted as points. With $p = 50$ agents, the dimension of the problem is $d = 2450$

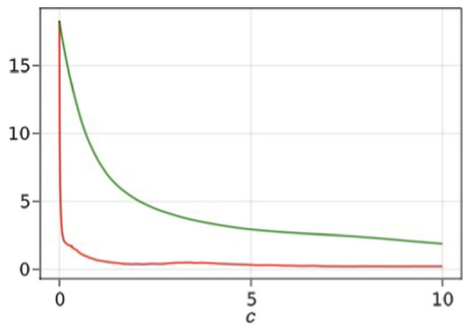

Fig. 3 Squared error of the marginal inclusion probabilities (Eq. 2.1) $c \to \mathcal{E}_{\text{zig-zag}}(c)$ (red) and $c \to \mathcal{E}_{\text{gibbs}}(c)$(green) where c represent the computing time in seconds. With $p = 20$ agents the dimension of the problem is $p(p-1)/2 = 380$

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

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

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

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, 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 D.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 E.2" for details. We set the final clock of the Sticky Zig-Zag sampler to 500. Results from running the sampler are summarized in Fig. 4.

In Fig. 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 1, rightmost column.

In Fig. 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.

4.3 Logistic regression

Suppose $\{0, 1\} \ni Y_i \mid x \sim \text{Ber}(\psi(x^T a_i))$ 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 (1.1) with zeromean Gaussian slabs and (large) variance σ_0^2. Then the posterior can be written as in Eq. (1.2), with Ψ and κ as in Equation (1.3).

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 $z_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(0.1)$ 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 E.1. We chose $w_1 = w_2 = \cdots = w_d = 0.1$ and $\sigma_0^2 = 10^2$ and ran the 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 7 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 et al. 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

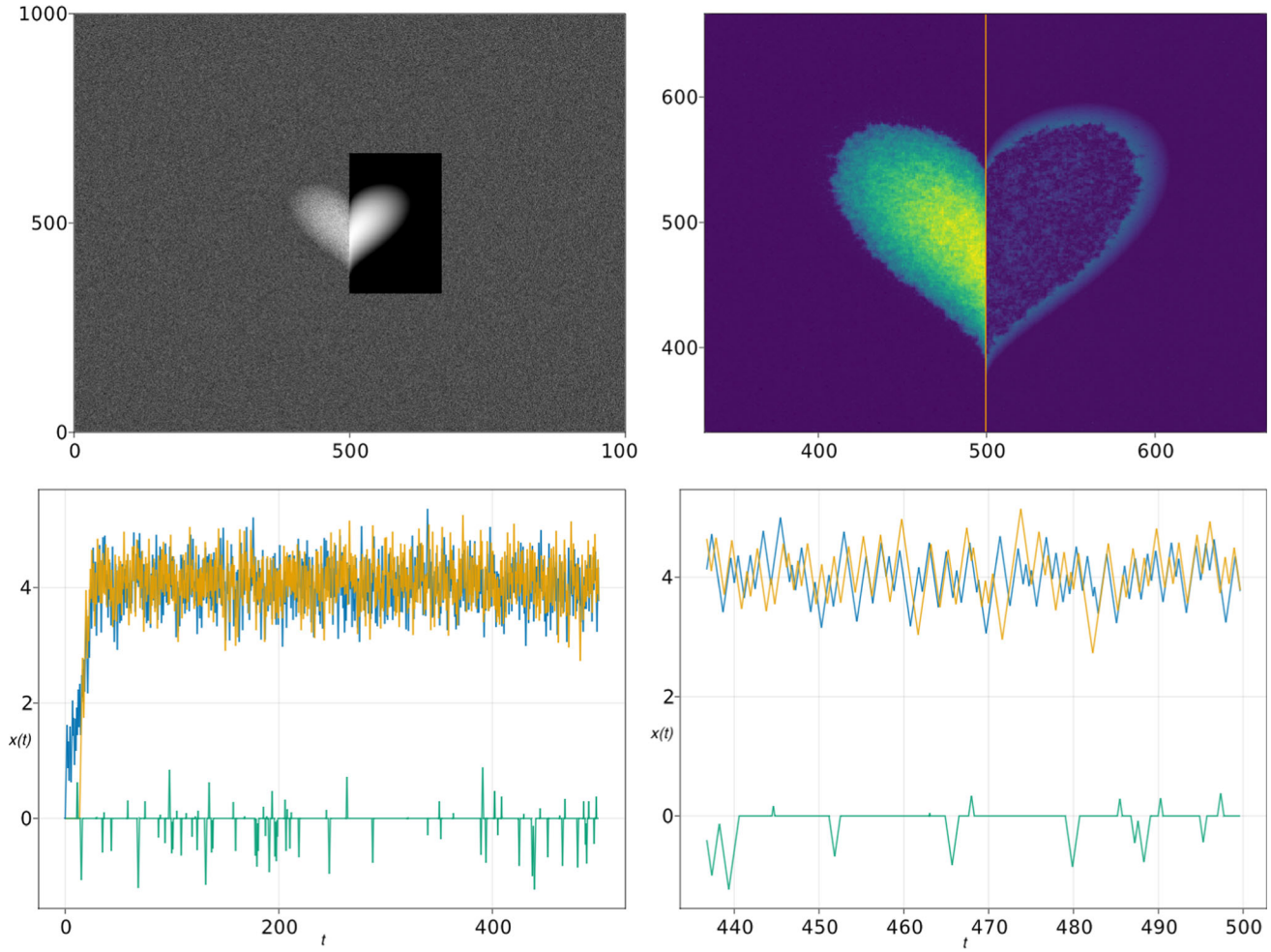

Fig. 4 Top-left: observed 1000×1000 image of a heart corrupted with white noise, with part of the ground truth inset. Top-right, left half: posterior mean estimated from the trace of the Sticky Zig-Zag sampler (detail). Top-right, right half: mirror image showing the absolute error between the posterior 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

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 8 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 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.

4.4 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 et al. (2021), and reference therein, for details.

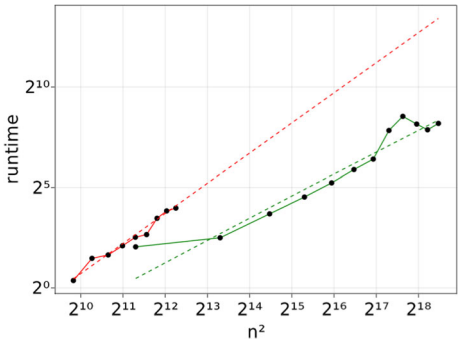

Fig. 5 Runtime comparison of the Sticky Zig-Zag sampler (green) and the Gibbs sampler (red) for the example in Subsection 4.2. The horizontal axis displays the dimension of the problem, which is n^2. The vertical axis shows runtime in seconds. The runtime is evaluated at $n^2 = 50^2, 100^2, \ldots, 600^2$ 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

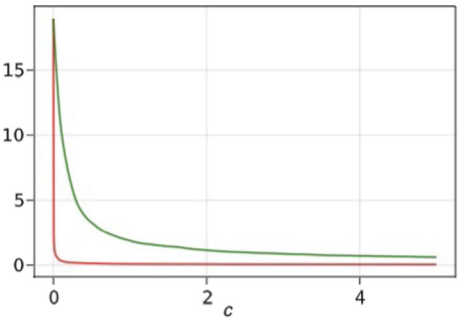

Fig. 6 Squared error of the marginal inclusion probabilities (Eq. 2.1) $c \to \mathcal{E}_{\text{zig-zag}}(c)$ (red) and $t \to \mathcal{E}_{\text{gibbs}}(c)$ (green) where c represent the computational time in seconds; right-panel: zoom-in near 0. Here the dimension of the problem is $n^2 = 400$

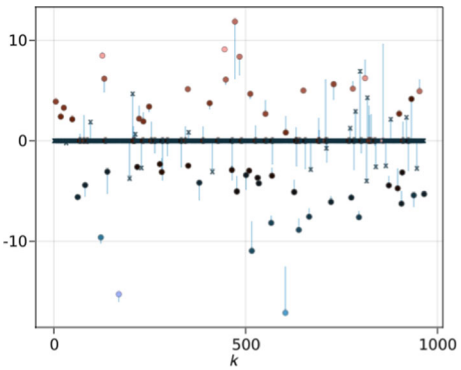

Fig. 7 Results for the logistic regression coefficients derived with the Sticky Zig-Zag sampler with subsampling. Description as in caption of Figure 2. The dimension of this problem is $d = 965$

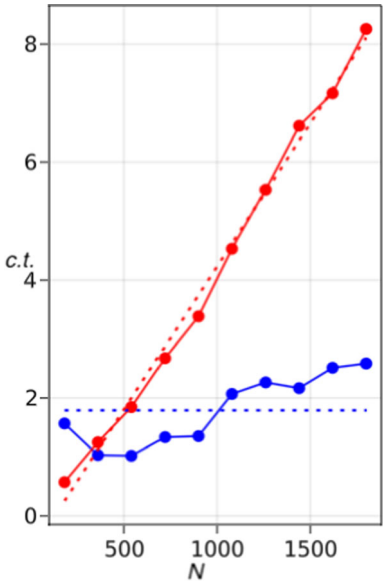

Fig. 8 Logistic regression example: computing time in seconds versus number of observations. Solid red line: Gibbs samplers with 10^3 iterations. Solid blue line: Sticky Zig-Zag samplers with subsampling ran for 10^3 time units. The dashed lines correspond to the scaling results displayed in Table 1. Here, the dimension of the problem is fixed to $d = 9$

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

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

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 Eq. (1.2), as the diagonal

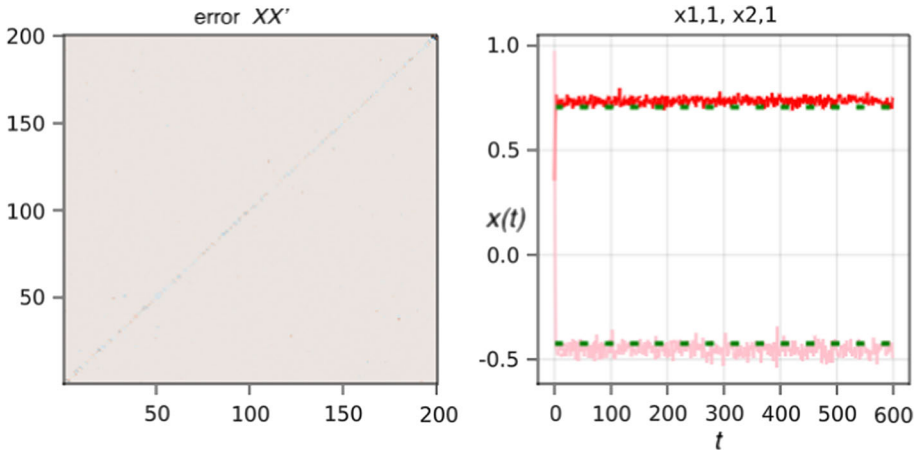

Fig. 9 Left: error between the true precision matrix and the precision matrix obtained with the estimated posterior mean of the lower-triangular matrix (colour gradient between white (no error) and black (maximum error)). Right: traces of two non-zero coefficients ($x_{1,1}$ in red and $x_{2,1}$ in pink) 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 $p(p+1)/2 = 20\,100$

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 E.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 $1 \le j \le i \le p$ 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 Fig. 9 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 Fig. 9 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 Eq. (2.1).

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 Eq. (2.5), 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 D.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.

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

$$\dot{x} = v, \quad \dot{v} = -\nabla\Psi(x) \tag{2.1}$$

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 Sect. 2, such process can be made sticky with Eq. (1.2) as its stationary measure.

Unfortunately, in most cases, the dynamics in (2.1) 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.

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 submodels 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.

Acknowledgements this work is part of the research programme *Bayesian inference for high dimensional processes* with project number 613.009.034c, which is (partly) financed by the Dutch Research Council (NWO) under the *Stochastics—Theoretical and Applied Research* (STAR) grant. J. Bierkens acknowledges support by the NWO for the research project *Zig-zagging through computational barriers* with project number 016.Vidi.189.043.

A. Details of the Sticky Zig-Zag sampler

A.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 $\mathfrak{F}_i = \{(x, v) \in E : x_i = 0^-,\ v_i > 0 \text{ or } x_i = 0^+,\ v_i < 0\}$ 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 Sect. 2. Only at a later stage we recover the definition of E.

Define the set

$$K = \left\{ \to\circ, \circ\to, \leftarrow\circ, \circ\leftarrow, \overset{\leftarrow}{\circ}, \overset{\rightarrow}{\circ} \right\}$$

and

$$|K| = \{\circ, \leftarrow\circ\to, \to\circ\leftarrow\}$$

(note that $|K|$ does not denote the cardinality of the set K). Define the functions $k : \mathbb{R} \times \mathbb{R} \to K$ and $|k| : \mathbb{R} \times \mathbb{R} \to |K|$ by

(x, v)	$k(x, v)$	at (x, v) the process is…	$\lvert k\rvert(x, v)$
$x > 0, v > 0$	$\circ\to$	…moving away from 0 with positive velocity	$\leftarrow\circ\to$
$x < 0, v < 0$	$\leftarrow\circ$	…moving away from 0 with negative velocity	$\leftarrow\circ\to$
$x > 0, v < 0$	$\circ\leftarrow$	…moving toward 0 with negative velocity	$\to\circ\leftarrow$
$x < 0, v > 0$	$\to\circ$	…moving toward 0 with positive velocity	$\to\circ\leftarrow$
$x = 0, v > 0$	$\overset{\rightarrow}{\circ}$	…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 $k : \overline{\mathbb{R}}^d \times \mathcal{V} \to K^d$ and $|k| : \overline{\mathbb{R}}^d \times \mathcal{V} \to |K|^d$ by applying the map k and $|k|$ coordinatewise.

For each $\ell \in K^d$ define

$$\widetilde{E}^\circ_\ell = \{(\ell, x, v) : k(x, v) = \ell\}$$

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 $i = 1, 2, \ldots, d$ 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:

$$\begin{aligned}\widetilde{E} &= \bigcup_\ell \widetilde{E}_\ell, \quad \widetilde{E}_\ell \\ &= \widetilde{E}_\ell^\circ \cup \{(\ell, x, v) \in \partial\widetilde{E}_\ell^\circ : x_i = 0 \Rightarrow |\ell_i| \neq \rightarrow\circ\leftarrow \text{ for all } i\}.\end{aligned}$$

(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

$$\begin{aligned}\Gamma &= \bigcup_\ell \Gamma_\ell \subset \bigcup_\ell \partial\widetilde{E}_\ell^\circ, \\ \Gamma_\ell &= \left\{(\ell, x, v) \in \partial\widetilde{E}_\ell^\circ, \exists i : x_i = 0, |\ell_i| = \rightarrow\circ\leftarrow\right\},\end{aligned}$$

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 A.1 *A bijection $\iota : \widetilde{E} \to E$ is given by*

$$\iota((\ell, \widetilde{x}, v)) = (x, v)$$

where

$$x_i = \begin{cases} 0^+ \,(0^-) & \ell_i = \overleftarrow{\circ} \;\; (\ell_i = \overrightarrow{\circ}) \\ 0^+ \,(0^-) & \ell_i = \circ\rightarrow \;\; (\ell_i = \leftarrow\circ), \; \widetilde{x}_i = 0 \\ \widetilde{x}_i & \textit{otherwise.} \end{cases}$$

Proof Recall that $\alpha(x, v) := \{i \in \{1, 2, \ldots, d\} : (x, v) \notin \mathfrak{F}_i\}$ 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 $e = (\ell, \widetilde{x}, v) \in \widetilde{E}\backslash\widetilde{E}^\circ$ 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 $\widetilde{\phi}_\ell : [0, \infty) \times \widetilde{E}_\ell^\circ \to \widetilde{E}_\ell^\circ$ 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 $x_i + v_i t(\mathbb{1}_{|\ell_i|\neq\circ}), i = 1, 2, \ldots, d$ 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

$$\mathcal{Q}^{\mathrm{s}}((\ell, x, v), \cdot) = \frac{\sum_i \lambda_i^{\mathrm{s}}(\ell, x, v)\delta_{(\ell[i:\,\ell_i'],x,v)}}{\sum_i \lambda_i^{\mathrm{s}}(i, x, v)}$$

with $\ell_i' = \circ\rightarrow$ if $\ell_i = \overrightarrow{\circ}$ 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 $\lambda : \widetilde{E} \to \mathbb{R}^+$

$$\lambda(\ell, x, v) = \sum_{i=1}^d \lambda_i^{\mathrm{r}}(\ell, x, v) + \lambda_i^{\mathrm{s}}(i, x, v)$$

and a Markov kernel $\mathcal{Q}\colon (\widetilde{E}\cup\Gamma, \mathcal{B}(\widetilde{E}\cup\Gamma)) \to [0,1]$ by

$$\mathcal{Q}((\ell,x,v),.) = \begin{cases} \frac{\sum_i \lambda_i^{\mathrm{r}}(\ell,x,v)}{\lambda(\ell,x,v)}\mathcal{Q}^{\mathrm{r}}((\ell,x,v),.) \\ \quad + \frac{\sum_i \lambda_i^{\mathrm{s}}(\ell,x,v)}{\lambda(\ell,x,v)}\mathcal{Q}^{\mathrm{s}}((\ell,x,v),.) & (\ell,x,v)\in\widetilde{E}, \\ \mathcal{Q}^{\mathrm{f}}((\ell,x,v),.) & (\ell,x,v)\in\Gamma. \end{cases}$$

Proposition A.2 *$\mathfrak{X}, \lambda, \mathcal{Q}$ 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 $[0, \varepsilon(\ell,x,v))$, for some $\varepsilon > 0$, 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.

A.2 Strong Markov property

Proposition A.3 *(Part of Theorem 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

$$\widetilde{\mathcal{A}}f(\ell,x,v) = \mathfrak{X}_\ell f(\ell,x,v) + \lambda(\ell,x,v) \\ \int_{\widetilde{E}} (f(\ell',x',v') - f(\ell,x,v))\mathcal{Q}(\ell,x,v,\mathrm{d}(\ell,x,v)).$$

The strong Markov property of $(\widetilde{Z}_t)$ follows by Davis (1993), Theorem 25.5. Denote by $(\widetilde{P}_t)_{t\ge0}$ the Markov transition semigroup of $(\widetilde{Z}_t)$ and let $(P_t)_{t\ge0}$ be a family of probability kernels on E and such that for any bounded measurable function $f\colon E\to\mathbb{R}$ and any $t\ge0$,

$$\widetilde{P}_t(f\circ\iota) = (P_t f)\circ\iota.$$

Then $(P_t)_{t\ge0}$ is the Markov transition semigroup of the process $Z_t = (\iota(\widetilde{Z}_t))$. By Rogers and Williams (2000), 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. □

A.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

$$\widetilde{G}\psi(t,\ell,x,v) = E[f(z(t))\mathbb{1}_{t<T_1} + \psi(t-T_1,z(t))\mathbb{1}_{t\ge T_1}].$$

We have that

$$\widetilde{G}\psi(t,\ell,x,v) = f(\widetilde{\varphi}(t,\ell,x,v))\times\mathcal{T} \tag{A.1}$$

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 A.4 *(Part of Theorem 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 (A.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

$$\begin{aligned}\widetilde{G}_n \psi_\ell(t, x, v) = E_x \big[& f(\widetilde{\varphi}_\ell(t, x, v)) \mathbb{1}_{t < T_n} \\ & + \psi(t - T_n, \widetilde{\varphi}_\ell(t, x, v)) \mathbb{1}_{t \geq T_n} \big],\end{aligned}$$

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 A.5 The proof of the Feller and Markov property follow similarly for the Bouncy Particle and the Boomerang sampler.

A.5 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:

$$\begin{aligned}\mathcal{D}(\mathcal{A}) = \{ & f \in \mathcal{M}(E);\ t \to f(\varphi(t, x, v))\tau \\ & \text{-absolutely continuous } \forall (x, v); \\ & \lim_{t \to 0} f(x[i: 0^+ + t], v) = f(x[i: 0^+], v); \\ & \lim_{t \to 0} f(x[i: 0^- - t], v) = f(x[i: 0^-], v)\}.\end{aligned}$$

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

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

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 $C_b(E) = \{f \in C(E): f \text{ is bounded}\}$ 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 A.6 *We have*

$$\begin{aligned}D = \Big\{ & f \in C_c^1(E): v_i \kappa_i (f(T_i(x, v)) - f(x, v)) \\ & = v_i \partial_i f(x, v) + \lambda_i(x, v)(f(x, v[i: -v_i])) \\ & \quad - f(x, v)), (x, v) \in \mathfrak{F}_i \text{ for all } i = 1, \ldots, d\}.\end{aligned}$$

For $f \in D$, $\mathcal{A}f = \mathcal{L}f$, where $\mathcal{L}f = \sum_{i=1}^d \mathcal{L}_i f$ with

$$\begin{aligned}\mathcal{L}_i f(x, v) = & v_i \partial_{x_i} f(x, v) + \lambda_i(x, v) \\ & (f(x, v[i: -v_i]) - f(x, v)).\end{aligned}$$

Proposition A.7 *(Proposition 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

$$\begin{aligned}& f(\ell(t), x(t), v(t)) - f(\ell, x, v) \\ & + \int_0^t \mathcal{A}f(\ell(s), x(s), v(s)) \mathrm{d}s, \quad \forall (\ell, x, v) \in \widetilde{E}\end{aligned}$$

is a local martingale (Davis 1993, Section 24) on $\widetilde{E}$, then $f \circ \iota: f \in \mathcal{D}(\widetilde{\mathcal{A}})$ 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

$$\mathcal{L}f(x, v) := \lim_{t \downarrow 0} \frac{\mathrm{E}[f(X_t, V_t) \mid X_0 = x, V_0 = v] - f(x, v)}{t}$$

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 $f \in D$:

$$\int \mathcal{L} f \mathrm{d}\mu = 0.$$

A.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 Sect. 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 Proposition (A.6)). 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

$$\begin{aligned}
\int \mathcal{L}_i f \mathrm{d}\mu = &\sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} \left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-}\right)\right.\\
&\left.\left(a_i \partial_{x_i} f_i^+ + \lambda_i^+(f_i^- - f_i^+)\right)\exp(-\Psi(x))\mathrm{d}x_i \prod_{j\neq i}\rho(\mathrm{d}x_j, v_j)\right)\\
&+ \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} \left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-}\right)\right.\\
&\left(-a_i \partial_{x_i} f_i^- + \lambda_i^-(f_i^+ - f_i^-)\right)\\
&\left.\exp(-\Psi(x))\mathrm{d}x_i \prod_{j\neq i}\rho(\mathrm{d}x_j, v_j)\right)\\
&+ \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} a_i \left(f_i^+(0^+) - f_i^+(0^-)\right)\right.\\
&\left.\exp\left(-\Psi(x[i\colon 0])\right)\prod_{j\neq i}\rho(\mathrm{d}x_j, v_j)\right)\\
&+ \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} -a_i\right.\\
&\left.\left(f_i^-(0^-) - f_i^-(0^+)\right)\exp(-\Psi(x[i\colon 0]))\right.\\
&\left.\prod_{j\neq i}\rho(\mathrm{d}x_j, v_j)\right).
\end{aligned}$$

By integration by parts we have that $\left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-}\right)\left(\partial_{x_i} f(x, v)\right.$ $\left.\exp(-\Psi(x))\right)\mathrm{d}x_i$ is equal to

$$\begin{aligned}
&\left(f(x[i\colon 0^-], v) - f(x[i\colon 0^+], v)\right)\\
&\exp(-\Psi(x[i\colon 0])) + \left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-}\right)\\
&\left(\partial_i \Psi(x) f(x, v)\exp(-\Psi(x))\right)\mathrm{d}x_i
\end{aligned}$$

so that $\int \mathcal{L}_i f \mathrm{d}\mu$ is equal to

$$\begin{aligned}
&\sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} \left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-}\right)\right.\\
&\left.\left(a_i \partial_{x_i}\Psi(x) + \lambda_i^+ - \lambda_i^-\right) f_i^- \exp(-\Psi(x))\mathrm{d}x_i \prod_{j\neq i}\rho(\mathrm{d}x_j, v_j)\right)\\
&+ \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} \left(\int_{0^+}^{\infty} + \int_{-\infty}^{0^-}\right)\right.\\
&\left.\left(-a_i \partial_{x_i}\Psi(x) + \lambda_i^- - \lambda_i^+\right) f_i^+ \exp(-\Psi(x))\mathrm{d}x_i \prod_{j\neq i}\rho(\mathrm{d}x_j, v_j)\right)\\
&+ \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} a_i \left(f_i^+(0^+) - f_i^+(0^-)\right)\right.\\
&\left.\exp(-\Psi(x[i\colon 0]))\prod_{j\neq i}\rho(\mathrm{d}x_j, v_j)\right)\\
&+ \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} -a_i \left(f_i^-(0^-) - f_i^-(0^+)\right)\right.\\
&\left.\exp(-\Psi(x[i\colon 0]))\prod_{j\neq i}\rho(\mathrm{d}x_j, v_j)\right)\\
&+ \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} a_i \left(f_i^+(0^-) - f_i^+(0^+)\right)\right.\\
&\left.\exp(-\Psi(x[i\colon 0]))\prod_{j\neq i}\rho(\mathrm{d}x_j, v_j)\right)\\
&+ \sum_{v\in\mathcal{V}^{-i}} \left(\int_{\mathbb{R}^{d-1}} -a_i \left(f_i^-(0^+) - f_i^-(0^-)\right)\right.\\
&\left.\exp(-\Psi(x[i\colon 0]))\prod_{j\neq i}\rho(\mathrm{d}x_j, v_j)\right) = 0,
\end{aligned}$$

where we used that $-v_i\partial_i\Psi(x) + \lambda_i(x, v) - \lambda_i(x, F_i(v)) = 0,\ \forall (x, v) \in E$.

A.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 et al. 2019b), we

start by making similar assumptions on Ψ as appearing in that paper.

Assumption A.8 (*Assumptions of* Bierkens et al. 2019b, *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 $\alpha \subset \{1, 2, \ldots, d\}$, we define the sub-space $\mathcal{M}_\alpha = \{x \in \mathbb{R}^d : x_i = 0,\ i \notin \alpha\}$ 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 A.9 *Suppose Ψ satisfies Assumption A.8. Then for every set $\alpha \subset \{1, 2, \ldots, d\}$, $(Z_t^{(\alpha)})_{t\geq 0}$ 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 A.8 holds on $\mathbb{R}^d$, then it holds on any the sub-space $\mathcal{M}_\alpha$, $\alpha \subset \{1, 2, \ldots, d\}$, for functions $x \mapsto \Psi(x)$, $x \in \mathcal{M}_\alpha$. Proposition A.9 follows from the ergodic theorem of ordinary Zig-Zag processes (Bierkens et al. 2019b, 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 $\bar{\delta}_0(\mathrm{d}x, \mathrm{d}v) = \bigotimes_{i=1}^d (\delta_{0^+,-1}(\mathrm{d}x_i, \mathrm{d}v_i) + \delta_{0^-,+1}(\mathrm{d}x_i, \mathrm{d}v_i))$, the set $\mathfrak{S} = \cap_{i=1}^d \mathfrak{F}_i$ and the first hitting time $\tau_A = \inf\{t > 0 : Z_t \in A\}$, where $Z_t = (X_t, V_t)$ is the sticky Zig-Zag process.

Proposition A.10 (Harris recurrence) *Suppose Ψ satisfies Assumption A.8. Then for any initial state $Z_0 = z_0 \in E$, we have that $\mathbb{P}(\tau_{\mathfrak{S}} < \infty) = 1$.*

Proof Let $x_0 \in \mathcal{M}_\alpha$ for an arbitrary $\alpha \subset \{1, 2, \ldots, d\}$. Denote the random time of the first stuck coordinate x_i, $i \in \alpha^c$ leaving zero by $T_1 \sim \mathrm{Exp}(\sum_{j\in\alpha^c} \kappa_j) > 0$. 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 sub-space $\mathcal{M}_\alpha$. 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 A.9, 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 $\mathcal{M}_{\alpha_2}, \mathcal{M}_{\alpha_3}, \ldots, \mathcal{M}_{\alpha_{|\alpha|-1}}$, 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 $s_0 = ((0^-, \ldots, 0^-), (+1, \ldots, +1))$. Next, we show that $\mathbb{P}(\tau_S < \infty) = 1$. Let $\tau_{\mathfrak{S}}$ be the hitting time to the set $\mathfrak{S}$ of the sticky Zig-Zag $Z(t)_{t>0}$. Denote by $\beta := \{i : Z_i(\tau_{\mathfrak{S}} \neq [s_0]_i)\} \subset \{1, 2, \ldots, d\}$ the set of indices for which the coordinate 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 $\Delta T \sim \mathrm{Exp}(\sum_{j=1}^d \kappa_j)$. 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, $\Delta T_1 \sim \mathrm{Exp}(\sum_{j\neq i} \kappa_j)$ (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 A.9, 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 $\mathbb{P}(\tau_S < \infty) = 1$. □

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 any skeleton $Z^{(\Delta)} = (Z(0)), Z(\Delta), Z(2\Delta), \ldots)$ (with $\Delta > 0$) is irreducible relative to the measure $\bar{\delta}_0$ as the process, once it has reached the null model, it will stay there for a random time $\Delta T \sim \mathrm{Exp}(\sum_{j=1}^d \kappa_j)$ and $\mathbb{P}(\Delta T > \Delta) > 0$.

A.7 Recurrence time of the sticky Zig-Zag to 0

The recurrent time to the point $\mathbf{0} = (0, 0, \ldots, 0)$ 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

$$\frac{(\kappa d)^{-1}}{\mu(\{\mathbf{0}\})} = \frac{\mathbb{E}[\tau_0]}{1 - \mu(\{\mathbf{0}\})}.$$

B. Other sticky PDMP samplers

Here we extend the results presented in Sect. 2.3 for two other Sticky PDMP samplers: the sticky version of the Bouncy particle sampler (Bouchard-Côté et al. 2018) and the Boomerang sampler (Bierkens et al. 2020), the latter having Hamiltonian deterministic dynamics invariant to a prescribed Gaussian

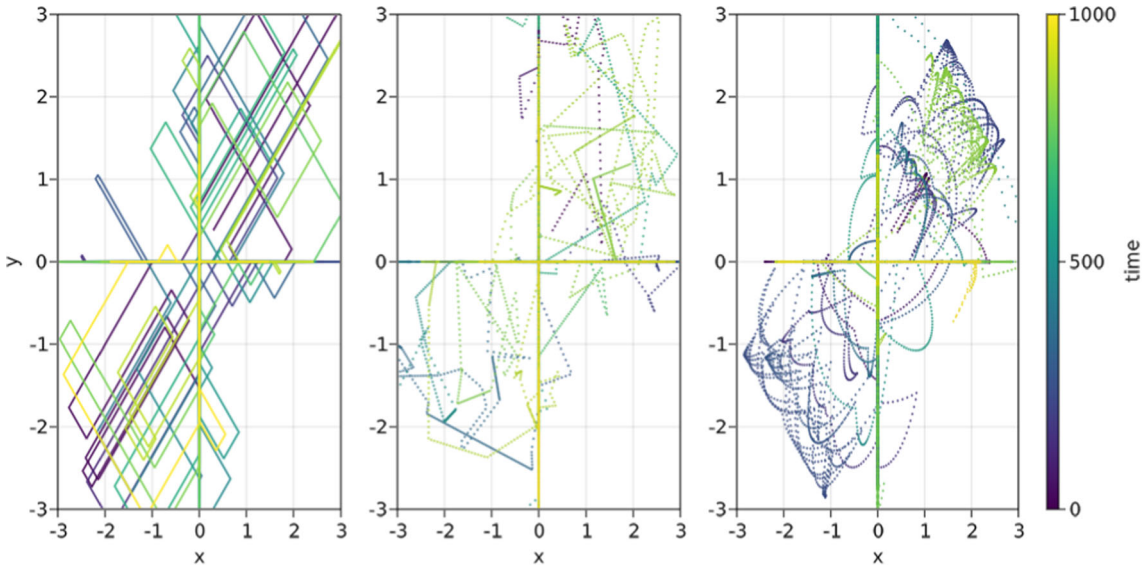

Fig. 10 $(x\text{-}y)$ phase portraits, of 3 different sticky PDMP samplers targeting the measure of Eq. (1.2) with $\exp(-\Psi)$ being a mixture of two bivariate Gaussian densities centered respectively in the first and the third quadrant of the $x\text{-}y$ axes. Left: Sticky Zig-Zag sampler. Middle: 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

measure. To visually assess the difference in sample paths, we show in Fig. 10 a typical realization of the Sticky Zig-Zag sampler, Sticky Bouncy particle sampler and Sticky Boomerang sampler.

B.2 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 Eq. (2.1). 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 et al. 2018). At refreshment times, each coordinate renews its velocity component independently according to the following refreshment rule

$$v_i' \sim \begin{cases} Z_i & (x, v) \notin \mathfrak{F}_i, \\ \text{sign}(v_i)|Z_i| & (x, v) \in \mathfrak{F}_i, \end{cases} \tag{B.1}$$

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 et al. (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 $f : E \to \mathbb{R}$ in the extended domain of the generator. Here, $\rho(y)$ is the standard normal density function evaluated at y.

Proposition B.1 *The d-dimensional sticky Bouncy Particle sampler is invariant to the measure*

$$\mu(\mathrm{d}x, \mathrm{d}v) = \frac{1}{C}\rho(v)\mathrm{d}v \exp(-\Psi(x)) \prod_{i=1}^{d} \left(\mathrm{d}x_i + \frac{1}{\kappa_i}\left(\mathbb{1}_{v_i>0}\delta_{0^-}(\mathrm{d}x_i) + \mathbb{1}_{v_i<0}\delta_{0^+}(\mathrm{d}x_i)\right)\right) \quad \text{(B.2)}$$

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

$$\begin{aligned} \|R_\Psi(x,v)v\|^2 &= \|v\|^2_{\alpha^c} + \|R_\Psi(x,v)v\|^2_\alpha \\ &= \|v\|^2_{\alpha^c} + \|v\|^2_\alpha = \|v\|^2 \end{aligned}$$

so, $\rho(R^A_\Psi(x)v) = \rho(v)$ ($\rho(x)$ here denotes the standard Gaussian density evaluated at x). Furthermore λ satisfies

$$-\langle v, \nabla\Psi(x)\rangle_\alpha + \lambda(x,v) - \lambda(x, R_\Psi(x,v)v) = 0, \quad \forall (x,v) \in E. \quad \text{(B.3)}$$

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 $f_i(y) = f(x[i:y], v)$, $Rf(x,v) = f(x, R_\Psi(x,v)v)$ 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 $\Delta_i f(x,v) := f(x[i:0^+], v) - f(x[i:0^-], v))$. We have this preliminary result:

$$\begin{aligned} &\int \sum_{i=1}^{d} \mathcal{G}_i f \mathrm{d}\mu = \frac{1}{C} \\ &\sum_i \int \left(\mathcal{G}_i f \exp(-\Psi(x))(\mathrm{d}x_i + \frac{1}{\kappa_i}\delta_0(\mathrm{d}x_i))\right) \\ &\prod_{j\neq i}\left(\mathrm{d}x_j + \frac{1}{\kappa_j}\delta_0(\mathrm{d}x_j, v_j)\right)\rho(v)\mathrm{d}v \\ &= \frac{1}{C}\sum_i \int \\ &\left(v_i\partial_{x_i} f \exp(-\Psi(x))\mathrm{d}x_i + v_i \Delta_i f \exp(-\Psi(x))\delta_0(\mathrm{d}x_i)\right) \\ &\prod_{j\neq i}\left(\mathrm{d}x_j + \frac{1}{\kappa_j}\delta_0(\mathrm{d}x_j, v_j)\right)\rho(v)\mathrm{d}v \quad \text{(B.4)} \\ &= \frac{1}{C}\sum_i \int \left(v_i\partial_{x_i}\Psi(x) f(x,v)\exp(-\Psi(x))\mathrm{d}x_i\right) \\ &\prod_{j\neq i}\left(\mathrm{d}x_j + \frac{1}{\kappa_j}\delta_0(\mathrm{d}x_j, v_j)\right)\rho(v)\mathrm{d}v \quad \text{(B.5)} \\ &= \frac{1}{C}\sum_{A\subset\{1,\ldots,d\}} \\ &\left(\sum_{i\in A}\left(\int v_i\partial_{x_i}\Psi(x) f(x,v)\exp(-\Psi(x))\mathrm{d}x_A\right)\right. \\ &\left.\prod_{j\in A^c,}\frac{1}{\kappa_j}\delta_0(\mathrm{d}x_j, v_j)\right) \\ &= \frac{1}{C}\sum_{A\subset\{1\ldots,d\}}\int \langle v, \nabla\Psi(x[A^c:0])\rangle_A f(x[A^c:0],v) \\ &\exp(-\Psi(x[A^c:0]))\mathrm{d}x_A \prod_{j\in A^c}\frac{1}{\kappa_j}\rho(v)\mathrm{d}v \quad \text{(B.6)} \end{aligned}$$

Here from (B.4) to (B.5) we used integration by parts in the two half planes $(\infty, 0^+]$ and $[0^-, -\infty)$. For the equivalence of (B.5) to (B.6) 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

$$\begin{aligned} &\int \lambda_{\text{ref}} \int (f(x,w) - f(x,v))\varrho(w)\mathrm{d}w\mathrm{d}\mu \\ &= \frac{1}{C}\sum_{A\subset\{1,2,\ldots,d\}} \lambda_{ref} \\ &\int (f(x,w) - f(x,v))\exp(-\Psi(x))\mathrm{d}x_A \\ &\quad \times \prod_{i\in A^c}\frac{1}{\kappa_i}\delta_{0^-}(\mathrm{d}x_i) \\ &\mathbb{1}_{v_i>0}\mathbb{1}_{w_i>0}2^{|A^c|}\rho(v)\rho(w)\mathrm{d}v\mathrm{d}w \\ &\quad + \frac{1}{C}\sum_{A\subset\{1,2,\ldots,d\}} \lambda_{\text{ref}} \\ &\int (f(x,w) - f(x,v))\exp(-\Psi(x))\mathrm{d}x_A \end{aligned}$$

$$\times \prod_{i\in A^c} \frac{1}{\kappa_i}\delta_{0^+}(\mathrm{d}x_i)\mathbb{1}_{v_i<0}\mathbb{1}_{w_i<0}2^{|A^c|}$$
$$\rho(v)\rho(w)\mathrm{d}v\mathrm{d}w,$$

which is equal to 0 by symmetry between v and w. Then

$$\begin{aligned}
&\int \mathcal{L}f\mathrm{d}\mu = \frac{1}{C}\sum_{A\subset\{1\ldots,d\}}\int \langle v,\\
&\nabla\Psi(x[A^c:0])\rangle_A \exp(-\Psi(x[A^c:0]))\\
&f(x[A^c:0],v)\mathrm{d}x_A \prod_{j\in A^c}\frac{1}{\kappa_j}\rho(v)\mathrm{d}v\\
&\quad + \int (\lambda(x,R_\Psi(x,v)) - \lambda(x,v))f(x,v)\mu(\mathrm{d}x,\mathrm{d}v)\\
&= \frac{1}{C}\sum_{A\subset\{1\ldots,d\}}\int \langle v,\\
&\nabla\Psi(x[A^c:0])\rangle_A \exp(-\Psi(x[A^c:0]))f(x[A^c:0],v)\\
&\mathrm{d}x_A \prod_{j\in A^c}\frac{1}{\kappa_j}\rho(v)\mathrm{d}v \qquad \text{(B.7)}\\
&+\frac{1}{C}\sum_{A\subset\{1,\ldots,d\}}\int \big(\lambda(x[A^c:0],R_\Psi v) - \lambda(x[A^c:0],v)\big)\\
&f(x[A^c:0],v)\exp(-\Psi(x[A^c:0]))\mathrm{d}x_A\\
&\times \prod_{j\in A^c}\frac{1}{\kappa_j}\rho(v)\mathrm{d}v\\
&= 0, \qquad \text{(B.8)}
\end{aligned}$$

where in Eqs. (B.7)–(B.8) we used a change of variable $v' = R_\Psi(x,v)v$ and property (B.3). □

Remark B.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 $A \sqcup F = \{1,\ldots,d\}$ 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

$$(\mathcal{F}(x,v,w))_i = \begin{cases} 0^- & \text{if } x_i = 0^+, \quad v_i < 0, \quad w_i > 0,\\ 0^+ & \text{if } x_i = 0^-, \quad v_i > 0, \quad w_i < 0,\\ x_i & \text{else}\end{cases}$$

leaves the target measure μ invariant.

The transition kernels given in Remark B.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 B.1, leave μ invariant. For example, the preconditioned Crank–Nicolson scheme Cotter et al. (2013) falls withing this setting.

B.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

$$\begin{aligned}(x_i(t),v_i(t)) &= (\cos(t)x_i(0) + \sin(t)v_i(0),\\ &\quad -x_i(0)\sin(t) + \cos(t)v_i(0)),\end{aligned} \qquad \text{(B.9)}$$

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 (Eq. (1.1)) where $\{\pi_i\}_{i=1}^d$ are centered Gaussian densities with variance σ_i^2. Then a natural choice is $\Sigma = \mathrm{Diag}(\sigma_1^2,\sigma_2^2,\ldots,\sigma_n^2)$.

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

$$\begin{aligned}&(R_U(x,v)v)_i\\ &= \begin{cases} v_i & i\notin\alpha\\ v_i - 2\frac{\langle\nabla U(x),v\rangle_\alpha}{\|\nabla\Sigma^{1/2}U(x))\|_\alpha^2}\langle\Sigma_{[i,:]},\nabla U(x)\rangle_\alpha & \text{else}\end{cases}\end{aligned}$$

and refreshes the velocity at exponentially distributed times with rate equal to λ_{ref} according to the rule given in Eq. (B.1).

Proposition B.3 *The d-dimensional sticky Boomerang sampler is invariant to the measure in Eq. (B.2).*

Proof The extended generator of the sticky d-dimensional Boomerang process is given by

$$\begin{aligned}\mathcal{A}f(x,v) &= \sum_{i=1}^d \mathcal{G}_i f(x,v) + \lambda(x,v)(f(x,R_U(x,v)v)\\ &\quad - f(x,v)) + \lambda_{\mathrm{ref}}\int (f(x,w) - f(x,v))\,\varrho_{x,v}(w)\mathrm{d}w\end{aligned}$$

and

$$\mathcal{G}_i f(x,v) = \begin{cases} |v_i|\kappa_i\,(f(T_i(x,v)) - f(x,v)) & (x,v)\in\mathfrak{F}_i\\ v_i\partial_{x_i}f(x,v) + x_i\partial_{v_i}f(x,v) & \text{else,}\end{cases}$$

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 $f : E \to \mathbb{R}$ 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 Eq. (B.9)) 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 B.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 $\{\tau_j : j \in \alpha(x, v)\}$ with $\tau_j \sim \text{Poiss}(t \to \lambda_j(\varphi(t, x, v))$ 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 $v \to v[i : -v_i]$ where $i = \text{argmin}\{\tau_j : j \in \alpha(x, v)\}$. 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. 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 et al. (2020) which, similarly to us, addresses variable selection problems using PDMP samplers.

The approach taken in Chevallier et al. (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 et al. (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.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.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.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 et al. (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 $\{z_t : 0 < t < T\}$ 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 Fig. 11, 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

$$\prod_{i=1}^{d} \left(\mathrm{d}x_i + c \sum_{j \in \mathbb{N}} \delta_{j*0.01}(\mathrm{d}x_i)\right), \tag{C.1}$$

with $c = 1.0$. In Fig. 11, 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 Fig. 11, 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.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 equal to 0 (Fig. 12) and 0.99 (Fig. 13) relative to the reference measure of Eq. (C.1) 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 $x_2 = -x_1$).

D. Details of Section 3

D.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

$$X \mid (Y = y) \sim \mathcal{N}(\mu_{x|y}, \Gamma_x^{-1}) \tag{D.1}$$

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

$$\int \phi([x, y]; \mu, \Gamma^{-1})\mathrm{d}x (2\pi)^{\frac{d_x - d}{2}} |\Gamma|^{\frac{1}{2}} |\Gamma_x|^{-\frac{1}{2}} \exp\left(\frac{1}{2}\mu'_{x|y}\Gamma_x\mu_{x|y} - \frac{1}{2}[-\mu_x, y - \mu_y]'\Gamma[-\mu_x, y - \mu_y]\right) \tag{D.2}$$

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 Eq. (2.1) 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 Eq. (1.2) are given by

$$\begin{aligned} B_j(\alpha) &= \frac{\mu(\mathcal{M}_{\alpha \cup \{j\}})}{\mu(\mathcal{M}_{\alpha \setminus \{j\}})} \\ &= \frac{\kappa_i \int_{\mathbb{R}^{|\alpha \cup \{j\}|}} \exp(-\Psi(y))\mathrm{d}x_{\alpha \cup \{j\}}}{\int_{\mathbb{R}^{|\alpha \setminus \{j\}|}} \exp(-\Psi(z))\mathrm{d}x_{\alpha \setminus \{j\}}}, \end{aligned} \tag{D.3}$$

where $y = \{x \in \mathbb{R}^d : x_i = 0,\ i \notin (\alpha \cup \{j\})\}$, $z = \{x \in \mathbb{R}^d : x_i = 0,\ i \notin (\alpha \setminus \{j\})\}$. 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 Eqs. D.1 and D.2 we have that the right hand side of Eq. (D.3) 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 Eq. D.1, the random variable at step 2 of the Gibbs sampler presented in Sect. 3.1 can be simulated as $X_\alpha | (X_{\alpha^c} = \mathbf{0}) \sim \mathcal{N}(\mu_{x_\alpha | x_{\alpha^c} = \mathbf{0}}, \Gamma_{x_\alpha})$.

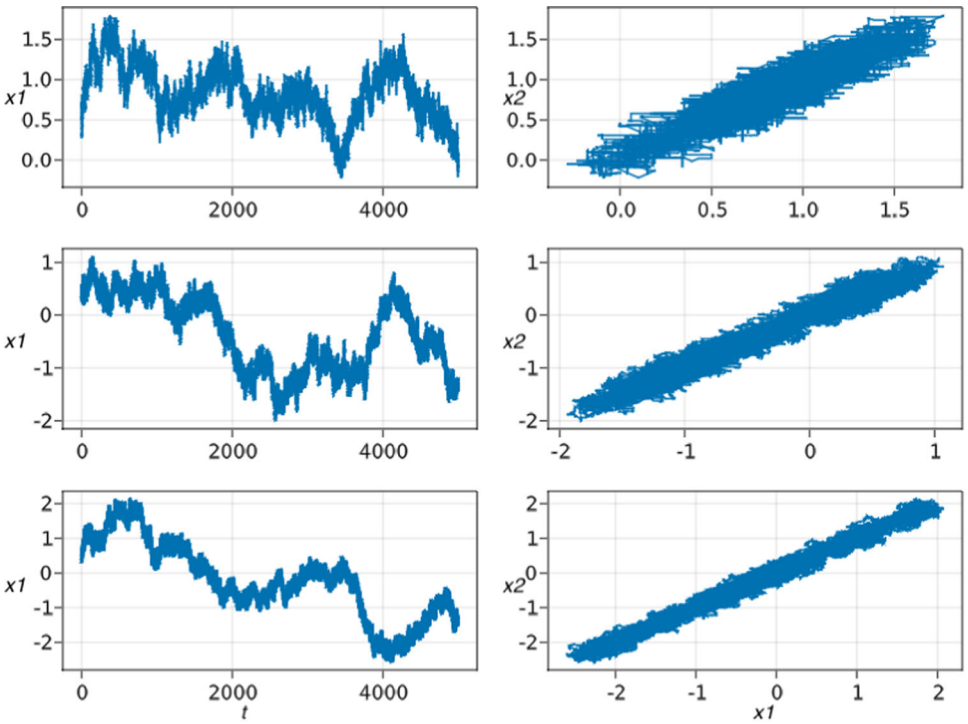

Fig. 11 x_1 trace plots (left) and x_1-x_2 phase portraits (right) of the Sticky Zig-Zag samplers with final clock $T = 50^3$ with p equal to 0.1 (top), 0.5 (center), 1.0 (bottom). The target measure has a Gaussian density with pairwise correlation equal to 0.99 relative to the reference measure of Eq. (C.1). 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 $p = 1$

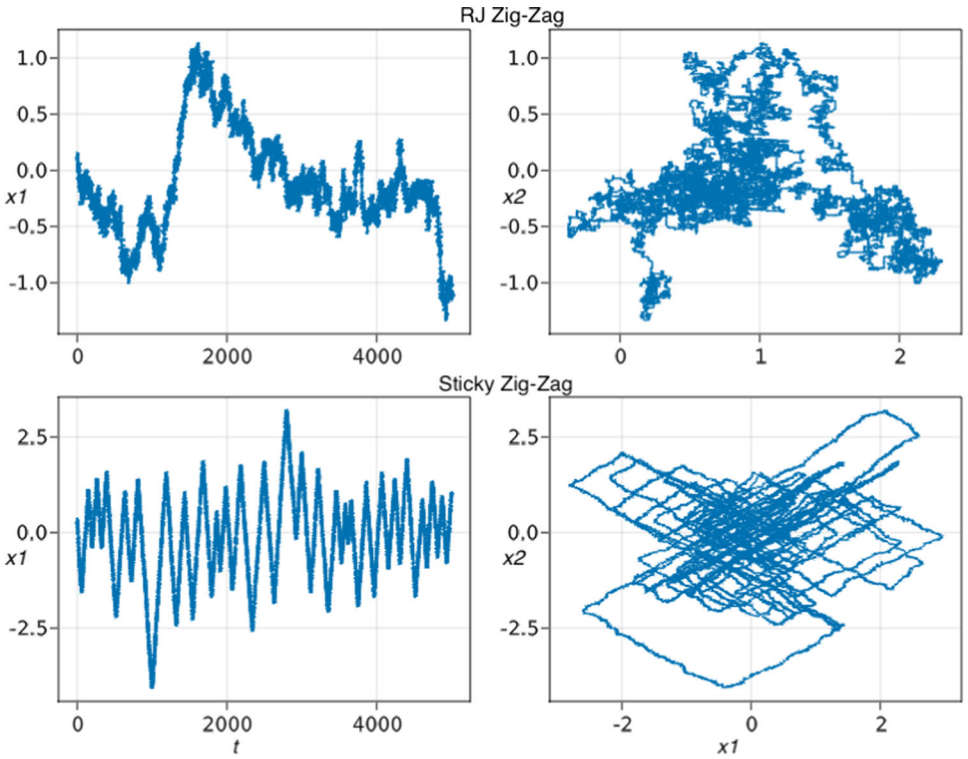

Fig. 12 Comparison between RJ Zig-Zag samplers (first row) and Sticky Zig-Zag samplers (second row) targeting a 20 dimensional measure with Gaussian density with pairwise correlation equal to 0.0 and relative to the reference measure in Eq. (C.1). Column 1: trace plot of the first coordinate. Column 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

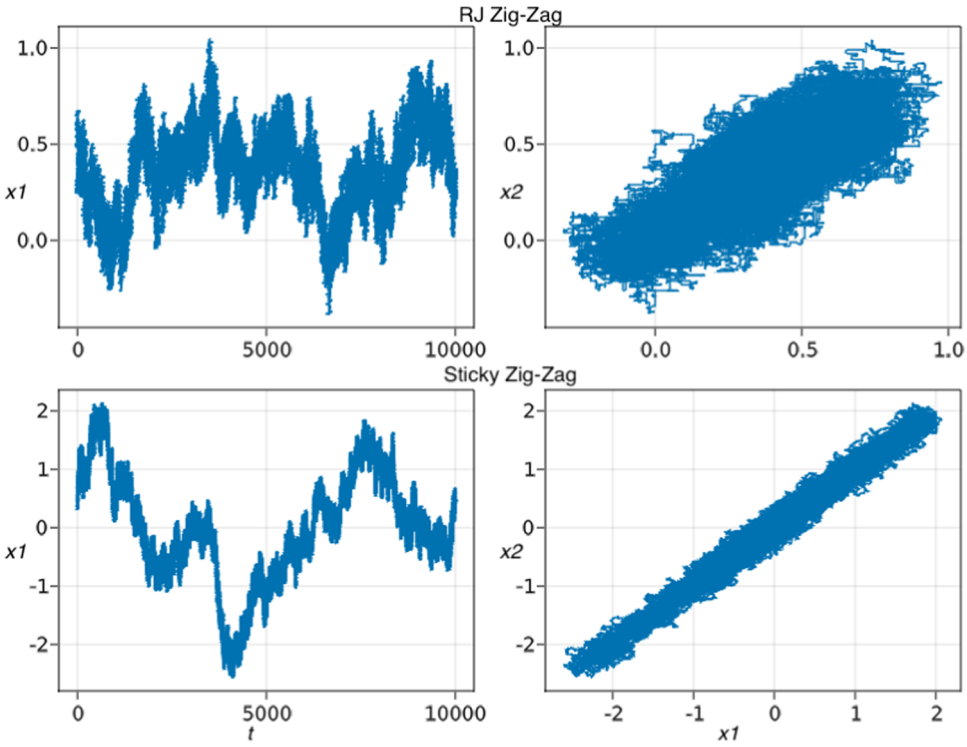

Fig. 13 Same description as in Fig. 12, except now for a Gaussian measure with pairwise correlation equal to 0.99. By looking for example at the symmetry along the axis $x_2 = -x_1$ and the ranges of the coordinates, it is clear that the Sticky Zig-Zag outperforms the RJ Zig-Zag

D.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 Sect. 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 Sect. 2.4.

D.2.3 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 (2.4). A sample from this distribution can be found by solving for τ' in the equation

$$\int_0^{\tau'} \lambda(\varphi(s, x, v))\mathrm{d}s = t, \quad t = \mathrm{Exp}(1). \tag{D.4}$$

We then write that $\tau' \sim \mathrm{Poiss}(\lambda(\varphi(\cdot, x, v))$. When it is not possible to find the root of Eq. (D.4) 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 $\Delta = \Delta(x, v) > 0$

$$\overline{\lambda}(t, x, v) \geq \lambda(\varphi(t, x, v)), \quad \Delta \geq t \geq 0, \tag{D.5}$$

for which this is possible and use the thinning scheme: Let $\tau' \sim \mathrm{Poiss}(\bar{\lambda}(\cdot, x, v))$; if $\tau' > \Delta$ then the proposed time is rejected and a new time has to be drawn as $\tau' \sim \mathrm{Poiss}(\bar{\lambda}(\cdot, \phi(\Delta, x, v)))$. 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 et al. (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 $\tau_i \sim \mathrm{Poiss}(\lambda_i(\varphi(\cdot, x, v))$ for $i = 1, 2, \ldots, d$. 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.

D.2.2 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 $f_i : \mathbb{R}^+ \times \mathbb{R}^{|\overline{A}_i|} \to \mathbb{R}^+$ with $\overline{A}_i \subset \{1, 2, \dots, d\}$. Given an initial position (x, v) and random times $\tau_j \sim \text{Poiss}(t \to \overline{\lambda}_j(t, x, v))$, for $i \in \alpha$, denote by $i = \text{argmin}_{j\in\alpha(x,v)}\ \tau_j$ 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 $\{\overline{\lambda}_j : j \in \alpha, \overline{A}_j \not\ni i\}$ 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 $\{\tau_j : i \in \overline{A}_j\}$ have to be renewed. However this is not always the case, as there are applications, such as the one in Sect. 4.3, 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.

D.2.3 Fully local implementation:

Consider now the decomposition of $\partial_{x_i}\Psi$, $i = 1, 2, \dots, d$ given in Eq. (2.8) 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 $f_{i,j} : \mathbb{R}^{|\widetilde{A}_{i,j}|} \to \mathbb{R}$ 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 $i = \text{argmin}(\tau_j : j \in \alpha(x, v))$ 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 $J \sim \text{Unif}(\{1, 2, \dots, N_i\})$. Hence, it is only required to update the position of the coordinates with index in $\widetilde{A}_{i,J}\backslash\alpha^c(x, v)$. 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 D.2.2) and, to this end, needs to update the position of the coordinates with index $\{k \in \overline{A}_j\backslash\alpha^c(x, v) : i \in \overline{A}_j\}$;
- 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 D.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 $\{x_i : i \in \alpha\}$.

D.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.

D.3.1 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

$$\lambda(x, v) = \sum_{i \in \alpha} \lambda_i(x, v) + \sum_{i \in \alpha^c} |v_i| \kappa_i \tag{D.6}$$

where, as before, $\alpha = \{i : x_i \neq 0\}$. In the following analysis, we drop the dependence on (x, v) and we assume that the size of $\alpha(t) := \{i : x_i(t) \neq 0\}$ 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 D.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 Eq. (D.6). 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 D.2.2). 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 (Sect. 2.4) in ideal scenarios (Bierkens et al. 2019a). Table 1 in Sect. 3 summarizes the overall scaling complexity of the Sticky Zig-Zag algorithm for the quantities p and N.

D.3.2 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 Sect. 4.2, the complexity of this operation is $\mathcal{O}(p^{3/2})$. This is followed by computing sufficient statistics in step 2 of Sect. 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 1 in Sect. 3 and validated by the numerical experiments of Sect. 4 (Figs. 5 and 8).

D.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 2 in Sect. 3 summarizes the scaling results derived in the two cases considered above.

E Details of Section 4

E.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 et al. (2019a). Given a posterior density of the form of Eq. (1.2) 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 Sect. 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 $J \sim \mathrm{Unif}(\Gamma_k)$, 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:

$$\begin{aligned} \partial_{x_i}(S(x, k, j)) &= \frac{A_{[j,k]} A_{[j,i]} e^{\langle A_{[j,:]}, x\rangle}}{\left(1 + e^{\langle A_{[j,:]}, x\rangle}\right)^2} \\ &\leq \frac{1}{4} A_{[j,k]} A_{[j,i]}, \end{aligned}$$

which means that for $i = 1, 2, \ldots, d$

$$\begin{aligned} &|S(x, i, j) - S(x', i, j)| \leq C_i \|x - x'\|_p, \quad p \geq 1, \\ &\quad j \in \Gamma_i,\ x, x' \in \mathbb{R}^d, \end{aligned}$$

with

$$C_k = \frac{1}{4} \max_{j=1,\ldots,N} |A_{[j,k]}| \, \|A_{j,:}\|_2.$$

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 $i \in \alpha(x, v)$:

$$\begin{aligned} &\widetilde{\lambda}_i(x(t), v(t)) \\ &= \left(v_i \left(\partial_{x_i} U(x') + \sigma^{-2} x_i(t) + |\Gamma_i|(S(x(t), i, j) - S(x', i, j)) \right) \right)^+ \\ &\leq (v_i(\partial_{x_i} U(x') + \sigma^{-2}(x_i + v_i t)))^+ + |v_i||\Gamma_i| \\ &\left(|S(x(t), i, j) - S(x, i, j)| + |S(x, i, j) - S(x', i, j)|\right) \\ &\leq (v_i(\partial_{x_i} U(x') + \sigma^{-2}(x_i + v_i t))^+ \\ &+ |v_i||\Gamma_i| C_i \left(t\|v\|_p + \|x - x'\|_p\right). \end{aligned}$$

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.

E.2 Spatially structured sparsity

For this application, we use the thinning scheme as presented in Appendix D.2.3. The bounding rates are of the form

$$\bar{\lambda}_i(t, x(t_0), v(t_0)) = (c + v_i(t_0)\partial_{x_i}\Psi(x(t_0)))^+ \tag{E.1}$$

for $t \in [0, \Delta]$ with $\Delta = 1/c$. To see this, define the Lipschitz growth bound $L_{x,v,\Delta}$ as

$$\begin{aligned} &P\Big(\sup_{0<t<\Delta} \frac{1}{t} |V_i(t)\partial_{x_i}\Psi(X(t))| \leq L_{x,\Delta} \\ &\quad | \, X(0) = 0, V(0) = v) = 1, \quad i = 1, 2, \ldots, d, \end{aligned}$$

which gives an explicit expression for c in Eq. (E.1) as

$$c - L_\Delta \Delta = 0 \Rightarrow \Delta = 1/c,$$

such that the inequality (D.5) 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 Sect. 4.2. With this given choice, in the simulations of Sect. 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 D.2.2 (with sets $\overline{A_i} = i$ for all i) in conjunction with the sparse algorithm as in Remark D.1.

E.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

$$\partial_{x_{i,j}}\Psi(x) = (YY')_{(i,:)}X_{(:,j)} + \gamma_{i,j}(x_{i,j}-c_{i,j}) - \mathbf{1}_{(i=j)}\left(\frac{N}{x_{i,j}}\right). \tag{E.2}$$

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

$$\lambda_{i,i}(\phi(t,x,v) \leq \bar{\lambda}_{i,i}(t,x,v) + \bar{\bar{\lambda}}_{i,i}(t,x,v), \quad t>0, \forall i.$$

Here $\bar{\lambda}_{i,i}(t,x,v) = (v_{i,i}(YY'_{i,:}(X_{:,i}+vt)+\gamma_{i,i}(x_{i,i}+vt-c_{i,i})))^+$ and $\bar{\bar{\lambda}}_{i,i}(t,x,v) = \left(-v_{i,i}\frac{N}{x_{i,i}+v_{i,i}t}\right)$ and a Poisson time form the bounding rate is simulated as $\min(\tau_1,\tau_2)$ where $\tau_1 \sim \text{Poiss}(s\to\bar{\lambda}_{i,i}(s,x,v))$ and $\tau_2 \sim \text{Poiss}(s\to\bar{\bar{\lambda}}_{i,i}(s,x,v))$.

References

Andrieu, C., Livingstone, S.:. Peskun–Tierney ordering for Markov chain and process Monte Carlo: beyond the reversible scenario (2019). arXiv: 1906.06197

Bento, J., Ibrahimi, M., Montanari, A.: Learning networks of stochastic differential equations (2010). arXiv: 1011.0415

Bierkens, J., Fearnhead, P., Roberts, G.: The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. Ann. Stat. **47**(3), 1288–1320 (2019)

Bierkens, J., Grazzi, S., Kamatani, K., Roberts, G.: The boomerang sampler. In: International Conference on Machine Learning, PMLR, pp. 908–918 (2020)

Bierkens, J., Grazzi, S., van der Meulen, F., Schauer, M.: A piecewise deterministic Monte Carlo method for diffusion bridges. Stat. Comput. **31**(3), 1–21 (2021)

Bierkens, J., Roberts, G.O., Zitt, P.-A.: Ergodicity of the zigzag process. Ann. Appl. Probab. **29**(4), 2266–2301 (2019)

Bouchard-Côté, A., Vollmer, S.J., Doucet, A.: The bouncy particle sampler: a nonreversible rejection-free Markov chain Monte Carlo method. J. Am. Stat. Assoc. **113**(522), 855–867 (2018)

Chevallier, A., Fearnhead, P., Sutton, M.: Reversible jump PDMP samplers for variable selection (2020). arXiv: 2010.11771

Cotter, S.L., Roberts, G.O., Stuart, A.M., White, D.: MCMC methods for functions: modifying old algorithms to make them faster. Stat. Sci. **28**, 424–446 (2013)

Davis, M.H.A.: Markov models and optimization. In: Monographs on Statistics and Applied Probability, vol. 49. Chapman & Hall, London (1993)

Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. Phys. Lett. B **195**(2), 216–222 (1987)

George, E.I., McCulloch, R.E.: Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88**(423), 881–889 (1993)

Grazzi, S., Schauer, M.: Boid animation. https://youtu.be/O1VoURPwVLI (2021)

Green, P.J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**, 711–732 (1995)

Green, P.J., Hastie, D.I.: Reversible jump MCMC. Genetics **155**(3), 1391–1403 (2009)

Griffin, J.E., Brown, P.J.: Bayesian global-local shrinkage methods for regularisation in the high dimension linear model. Chemom. Intell. Lab. Syst. **210**, 104255 (2021)

Guan, Y., Stephens, M.: Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann. Appl. Stat. **5**(3), 1780–1815 (2011)

Ishwaran, H., Rao, J.S.: Spike and slab variable selection: frequentist and Bayesian strategies. Ann. Stat. **33**(2), 730–773 (2005)

JuliaCon: 2020 by Jesse Bettencourt. JuliaCon 2020—Boids: Dancing with friends and enemies. https://www.youtube.com/watch?v=8gS6wejsGsY (2020)

Liang, X., Livingstone, S., Griffin, J.: Adaptive random neighbourhood informed Markov chain Monte Carlo for high-dimensional Bayesian variable Selection. arXiv:2110.11747 (2021)

Liggett, T.M.: Continuous time Markov processes. In: Graduate Studies in Mathematics, vol. 113. American Mathematical Society, Providence, RI (2010)

Meyn, S.P., Tweedie, R.L.: Stability of Markovian processes II: continuous-time processes and sampled chains. Adv. Appl. Probab. **25**(3), 487–517 (1993)

Mitchell, T.J., Beauchamp, J.J.: Bayesian variable selection in linear regression. J. Am. Stat. Assoc. **83**(404), 1023–1032 (1988)

Neal, R.M., et al.: MCMC using Hamiltonian dynamics. Handb. Markov Chain Monte Carlo **2**(11), 2 (2011)

Polson, N.G., Scott, J.G., Windle, J.: Bayesian inference for logistic models using Pòlya- Gamma latent variables. J. Am. Stat. Assoc. **108**(504), 1339–1349 (2013)

Ray, K., Szabo, B., Clara, G.:Spike and slab variational Bayes for high dimensional logistic regression (2020). arXiv: 2010.11665

Reynolds, C. W.: Flocks, herds and schools: a distributed behavioral model. In: Association for Computing Machinery (1987)

Rogers, L.C.G., Williams, D.: Diffusions, Markov Processes and Martingales: Volume 2, Itô calculus. vol. 2. Cambridge University Press (2000)

Rogers, L., Williams, D.: Diffusions, Markov processes, and martingales: foundations. In: Cambridge Mathematical Library, vol. 1. Cambridge University Press (2000)

Schauer, M., Grazzi, S.: mschauer/ZigZagBoomerang.jl: v0.6.0. Version v0.6.0. https://doi.org/10.5281/zenodo.4601534 (2021)

Shi, W., Ghosal, S., Martin, R.: Bayesian estimation of sparse precision matrices in the presence of Gaussian measurement error. Electron. J. Stat. **15**(2), 4545–4579 (2021)

Sutton, M., Fearnhead, P.: Concave-convex PDMP-based sampling. arXiv:2112.12897 (2021)

Tibshirani, R., et al.: Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. **67**(1), 91–108 (2005)

Zanella, G., Roberts, G.: Scalable importance tempering and Bayesian variable selection. J. R. Stat. Soc. Ser. B Stat. Methodol. **81**(3), 489–517 (2019)