



Automatic Zig-Zag sampling in practice

Alice Corbella^{1,2} · Simon E. F. Spencer¹ · Gareth O. Roberts^{1,3}

Received: 24 June 2022 / Accepted: 30 August 2022
© The Author(s) 2022

Abstract

Novel Monte Carlo methods to generate samples from a target distribution, such as a posterior from a Bayesian analysis, have rapidly expanded in the past decade. Algorithms based on Piecewise Deterministic Markov Processes (PDMPs), non-reversible continuous-time processes, are developing into their own research branch, thanks their important properties (e.g., super-efficiency). Nevertheless, practice has not caught up with the theory in this field, and the use of PDMPs to solve applied problems is not widespread. This might be due, firstly, to several implementational challenges that PDMP-based samplers present with and, secondly, to the lack of papers that showcase the methods and implementations in applied settings. Here, we address both these issues using one of the most promising PDMPs, the Zig-Zag sampler, as an archetypal example. After an explanation of the key elements of the Zig-Zag sampler, its implementation challenges are exposed and addressed. Specifically, the formulation of an algorithm that draws samples from a target distribution of interest is provided. Notably, the only requirement of the algorithm is a closed-form differentiable function to evaluate the log-target density of interest, and, unlike previous implementations, no further information on the target is needed. The performance of the algorithm is evaluated against canonical Hamiltonian Monte Carlo, and it is proven to be competitive, in simulation and real-data settings. Lastly, we demonstrate that the super-efficiency property, i.e. the ability to draw one independent sample at a lesser cost than evaluating the likelihood of all the data, can be obtained in practice.

Keywords Automatic inference · Piecewise deterministic Markov processes · Zig-Zag sampler · Gradient-based MCMC · Super-efficiency

1 Introduction

Applications of Bayesian inference have proliferated immensely in the most disparate fields during the recent decades. The diffusion of Bayesian methods in several scientific communities owns its credit, among other things, to advances in software that allow one to draw samples from a posterior distribution $p(\theta|y)$ of interest. The availability of programs such as BUGS (Gilks et al. 1994) and JAGS

(Plummer 2003), made standard Markov chain Monte Carlo (MCMC) algorithms such as the Metropolis-within-Gibbs sampler available to the community and used in many applications.

In parallel to this proliferation of applications, the methodology behind MCMC also expanded: recent research focused on the exploitation of the gradient of the target density to explore the space in a more efficient manner. Early examples include the Metropolis-adjusted Langevin algorithm (MALA) (Roberts and Rosenthal 1998; Roberts and Tweedie 1996); and the Metropolis-adjusted Langevin algorithm (MALA) algorithm (Neal et al. 2011); these algorithms showed the practical gain in efficiency of exploiting information from the gradient. HMC gained popularity in the 2010s thanks to the software Stan (Carpenter et al. 2017), which has an embedded Automatic Differentiation tool that allows to draw samples from a target distribution, needing only the functional form of its probability density function (pdf).

More recently, algorithms based on PDMPs (Fearnhead et al. 2018) have been proposed and showed great poten-

✉ Alice Corbella
alice.corbella@warwick.ac.uk

Simon E. F. Spencer
s.e.f.spencer@warwick.ac.uk

Gareth O. Roberts
g.o.roberts@warwick.ac.uk

¹ Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

² MRC Biostatistics Unit, University of Cambridge, Cambridge CB2 0SR, UK

³ The Alan Turing Institute, London NW1 2DB, UK

tial (Bierkens et al. 2018; Bouchard-Côté et al. 2018) thanks to their continuous-time behaviour and to convenient properties such as super-efficiency. Nevertheless, their use is not yet widespread, and very few papers use PDMP-based algorithms to address Bayesian estimation problems (Chevallier et al. 2020; Koskela 2022). Even fewer papers attempt to implement PDMP-based algorithm in a general form (Bertazzi et al. 2021; Pagani et al. 2022), unfortunately they don't retain exactness.

This paper intends to help the practice to catch up with the advances in the theory in three ways: (i) it provides a lay explanation of the implementation of PDMP algorithms, and specifically of the Zig-Zag sampler, making PDMPs available to a wider audience, both in terms of comprehension of the method and possibility of its applications; (ii) it addresses some of the obstacles that prevent the use in practice of the algorithms for a general target density of interest, in particular, the availability of explicit form of the gradient of the target density and a bounding constant for it; and (iii) it provides examples of the use of PDMP algorithms for real-data analyses.

Section 2 introduces PDMPs in their general form and gives an example of a PDMP-based algorithm: the canonical Zig-Zag sampler. This algorithm is used as a reference through the manuscript as its simple formulation makes illustration of many aspects of PDMPs clear and as it was shown to outperform other PDMP-based algorithms in some applied settings (Chevallier et al. 2020). Section 3 addresses the main implementation problems of the Zig-Zag sampler and provides the formulation of an algorithm that requires as input only a function that evaluates a differentiable target density at a specific point: the *Automatic Zig-Zag* sampler. In Sect. 4 the performance of this algorithm is evaluated against a competitive gradient-based scheme. Section 5 provides examples of analyses carried out using Automatic Zig-Zag sampling. Section 6 illustrates how super efficiency can be achieved in the context provided of the Automatic Zig-Zag sampler. Discussion and conclusions follow in Sect. 7.

2 Background: the Zig-Zag sampler

A PDMP is a continuous-time stochastic process denoted by Z_t , which, in between random times, evolves according to deterministic dynamics. Values z_t of the process can, for now, be thought of as d -dimensional vectors with elements $z_t^{(i)}$ for $i = 1, \dots, d$. A PDMP can be defined through specifying the following three components (Fearnhead et al. 2018):

- (i) a deterministic dynamic describing the change of the process over time which can be specified through an ordinary differential equation,

$$\frac{dz_t^{(i)}}{dt} = \Phi_i(z_t) \quad \text{for } i = 1, \dots, t. \tag{1}$$

hence the state of the process at time $t+s$ can be computed as a deterministic function of the state of the process at time t and the elapsed time s : $z_{t+s} = \Psi(z_t, s)$;

- (ii) random switching times which happen with rate dependent on the current state of the process $\lambda(z_t)$; and
- (iii) a transition kernel $q(\cdot|z_t)$ that determines the distribution of events that take place at the switching times and depends, again, on the current state of the process z_t .

The various PDMP-based algorithms differ among themselves in one or more of these specifics.

2.1 Definition

The Zig-Zag sampler is based on the simulation of a PDMP composed of two, distinguishable, elements: a location $X \in \mathbb{R}^d$ and a velocity V . The velocity can be thought of as an auxiliary variable defined on the space $\mathbb{V} = \{-1, +1\}^d$; the location instead, is typically the main component of interest: the sampler is constructed so that X has stationary distribution with density $\pi(x)$ (e.g. a posterior density). Crucial to the definition of the Zig-Zag sampler is that the target density could be written as $\pi(x) \propto e^{-U(x)}$, where $U(x)$ is sometimes called the *potential*.

Concerning the deterministic dynamics (i), the vector of velocities v is assumed to be constant between switching times, with each dimension of x increasing or decreasing at the same rate, so that Eq. (1) is effectively:

$$\frac{dx_t^{(i)}}{dt} = v_t^{(i)} \tag{2}$$

for $i = 1, \dots, d$. Given a starting state of the process (x_s, v_s) , the velocity then switches according to (ii) the minimum of d non-homogeneous Poisson processes (NHPPs) with rates

$$\lambda^{(i)}(t; x_s, v_s) = \max \left\{ v_s^{(i)} \frac{\partial}{\partial x^{(i)}} U(x_t), 0 \right\}. \tag{3}$$

for $i = 1, 2, \dots, d$, with $x_t = x_s + v_s \cdot t$ from (i). The intuition behind this formulation of the rate $\lambda(\cdot)$ is similar to that of many other gradient-based scheme: if the value of the potential is growing, the chains is moving away from where the mass concentrates, and hence the direction changes.

Lastly, (iii) the transition kernel $q(\cdot|z_t)$ is defined by the flipping operator $F_m(\cdot)$ that inverts the sign of the m -th dimension of the velocity, where m denotes the dimension of the earliest event in the realizations of the NHPPs.

$$F_m(v^{(i)}) = \begin{cases} -v^{(i)} & \text{for } i = m \\ v^{(i)} & \text{for } i \neq m \end{cases} \tag{4}$$

Bierkens et al 2019 (Bierkens et al. 2019) proved that a Zig-Zag process, under mild regularity conditions, converges to the invariant distribution of interest X with density $\pi(x)$.

To obtain the earliest realization of the d NHPPs with rates (3) it is possible sample from a one-dimensional inhomogeneous Poisson process with rate:

$$\lambda(t; \mathbf{x}_s, \mathbf{v}_s) = \sum_{i=1}^d \lambda^{(i)}(t; \mathbf{x}_s, \mathbf{v}_s). \tag{5}$$

The dimension in which the switch takes place is the realization of a Multinomial random variable (rv) with probabilities:

$$p_i = \frac{\lambda^{(i)}(t; \mathbf{x}_s, \mathbf{v}_s)}{\lambda(t; \mathbf{x}_s, \mathbf{v}_s)} \tag{6}$$

for $i = 1, 2, \dots, d$.

An illustration of the first steps of the simulation of a Canonical Zig-Zag process is reported in Figure 1.

2.2 Implementation

The practical implementation of the algorithm requires sampling from an NHPP with rate $\lambda(t)$, where arguments \mathbf{x}_s and \mathbf{v}_s are omitted since they are constant between switching times. As summarised by Lewis and Shedler (1979), this can be done either via time-scale transformation, finding τ such that:

$$\int_0^\tau \lambda(t) dt = u \tag{7}$$

given u sampled from an $\text{Exp}(1)$; or via thinning, i.e. (i) finding a constant upper bound $\bar{\lambda}$ such that $\bar{\lambda} \geq \lambda(t)$, either globally $\forall t$ or in some interval $[a, b]$, (ii) sampling a candidate point τ^* from an homogeneous Poisson process (HPP) with rate $\bar{\lambda}$ and (iii) accepting the candidate point with probability $\frac{\lambda(\tau^*)}{\bar{\lambda}}$. These sampling techniques are illustrated in Fig. 2.

Analytically determining the point τ that satisfies Eq. (7) is often impossible, above all due to the maximum contained in (3). Solving Eq. (7) numerically is often more expensive than finding a suitable upper bound $\bar{\lambda}$ and simulating the process via thinning: while the latter requires only a limited, wisely chosen, number of evaluations of the objective function, numerical integration implies a discretization of the domain t and the evaluation of the function at these numerous discrete points. Hence, here the thinning method is used to simulate an NHPP.

Using the ingredients of Eqs. (2), (3), and (4), it is possible to obtain the positions and the velocity of the process at each switching time $t_k: \{\mathbf{x}_{t_k}, \mathbf{v}_{t_k}\}_{k=1}^K$. These are called the *skeleton points* of the sampled distribution. The value of the process at each time t between two skeleton points can then

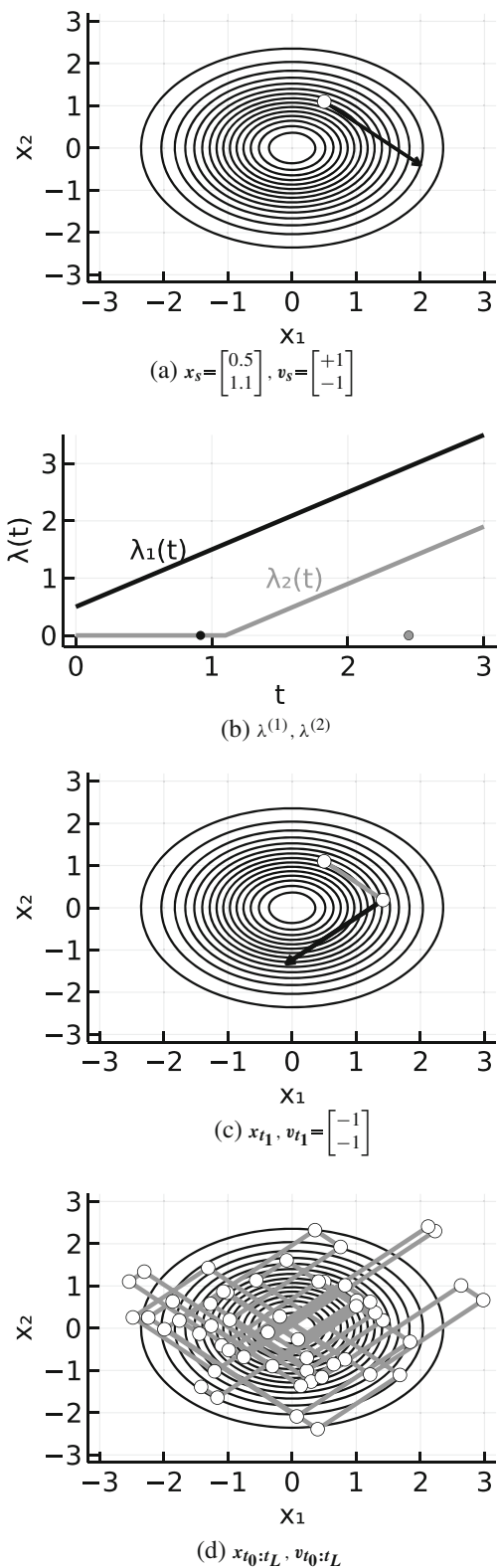


Fig. 1 Simulation of a Zig-Zag process targeting a bivariate independent standard normal distribution . **a** Initial location and velocity; **b** time-varying rate and samples from the NHPP for dimension 1 (black) and 2 (grey); **c** first switching time; **d** first 50 switching times (white) and continuous-time sample (grey)

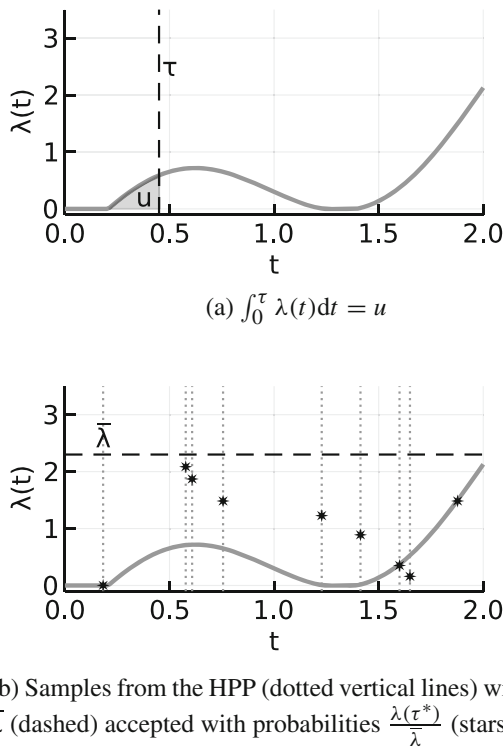


Fig. 2 Sampling mechanisms for a NHPP: **a** time-scale transformation and **b** thinning method

be obtained using the deterministic dynamics of Equation (2) which results in:

$$\mathbf{x}_t = \mathbf{x}_{t_k} + \mathbf{v}_{t_k}(t - t_k) \quad \text{for } t \in [t_k, t_{k+1}].$$

The pseudo-code of the Zig-Zag sampler with thinning when a global upper bound $\bar{\lambda}$ is known, is reported in Algorithm 1.

2.3 Beyond canonical Zig-Zag sampling

The Canonical Zig-Zag algorithm is not the only example of the use of PDMPs to sample from a target density of interest $\pi(\mathbf{x})$. The basic algorithm can be changed and extended in a number of ways to improve its performance on specific targets; moreover, different deterministic dynamics and switching rates/kernels can be used to formulate other PDMP-based algorithms (see, for example Bouchard-Côté et al. (2018), Wu and Robert (2020), and Bierkens et al. (2020)). Nevertheless, the focus of this paper is on the Canonical Zig-Zag algorithm to provide a simple example where our methods are applicable.

Algorithm 1 Canonical Zig-Zag algorithm with thinning with known global bound $\bar{\lambda}$.

```

Require: Initial velocity  $\mathbf{v}_0$ ; initial location  $\mathbf{x}_0$ ; total number of skeleton
points to be sampled  $K$ ; functions to compute the dimension-wise
and global switching rate  $\lambda(\cdot)$  and  $\lambda^{(i)}(\cdot)$ ; an upper bound for the
global rate  $\bar{\lambda}$ .
Ensure: Time, location and velocity at  $K$  skeleton points:
 $\{t_k, \mathbf{x}_k, \mathbf{v}_k\}_{k=1}^K$ 
1:  $t_0 = 0$  #set starting time
2:  $T \leftarrow t_0, \mathbf{X} \leftarrow \mathbf{x}_0, \mathbf{V} \leftarrow \mathbf{v}_0$  #set current state of the process
3: while  $k \leq K$  do
4:  $\tau^* \sim \text{Exp}(\bar{\lambda})$  #propose switching time
5:  $\lambda(\tau^*) = \sum_{i=1}^d \lambda^{(i)}(\tau^*; \mathbf{x}, \mathbf{v})$  # compute the global rate at  $\tau^*$ 
6:  $u \sim \text{Ber}(\frac{\lambda(\tau^*)}{\bar{\lambda}})$  # accept/reject  $\tau^*$ 
7: if  $u = 1$  then
8:  $T \leftarrow T + \tau^*$  #progress time
9:  $\mathbf{X} \leftarrow \mathbf{X} + \mathbf{V}\tau^*$  #progress location
10:  $m \sim \text{Multinom}\left(1 : d; \left\{\frac{\lambda^{(i)}(\tau^*; \mathbf{x}, \mathbf{v})}{\lambda(\tau^*)}\right\}_{i=1}^d\right)$  #sample component
11:  $\mathbf{V} \leftarrow F_m(\mathbf{V})$  #flip velocity of dim  $m$ 
12:  $t_k \leftarrow T, \mathbf{v}_k \leftarrow \mathbf{V}, \mathbf{x}_k \leftarrow \mathbf{X}$  #save skeleton point
13:  $k = k + 1$ 
14: else
15:  $T \leftarrow T + \tau^*$  #progress time
16:  $\mathbf{X} \leftarrow \mathbf{X} + \mathbf{V}\tau^*$  #progress location
17:  $\mathbf{V} \leftarrow \mathbf{V}$  #retain velocity
18: end if
19: end while
    
```

2.3.1 Non-canonical Zig-Zag sampling algorithms

The switching rate in Eq. (3) could be further extended by adding an excess switching rate $\gamma^{(i)}(\mathbf{x}_t, \mathbf{v}_t)$ such that

$$\begin{cases} \gamma^{(i)}(\mathbf{x}_t, \mathbf{v}_t) \geq 0 \\ \gamma^{(i)}(\mathbf{x}_t, \mathbf{v}_t) = \gamma^{(i)}(\mathbf{x}_t, F_i(\mathbf{v}_t)). \end{cases} \quad (8)$$

leading to switching rate:

$$\lambda^{(i)}(t; \mathbf{x}_s, \mathbf{v}_s) = \max \left\{ v_s^{(i)} \frac{\partial}{\partial x^{(i)}} U(\mathbf{x}_t), 0 \right\} + \gamma^{(i)}(\mathbf{x}_t, \mathbf{v}_t) \quad (9)$$

for $i = 1, \dots, d$.

This simple modification, discussed in Bierkens et al. (2019), allows the process to still converge to the correct target distribution (Bierkens et al. 2019) and slightly increases the event rate, generating extra switching times in addition to those driven by the potential $U(\mathbf{x})$. These switches are often called *refreshments* and, while in principle adding excessive refreshments will impoverish the mixing of the process (Andrieu and Livingstone 2021), many interesting constructs such as the Zig-Zag with subsampling, can be built by considering refreshment switches.

Other extensions have been formulated, in order to improve the performance of the Zig-Zag sampler on specific distributions/applications (e.g. heavy tailed distributions, highly correlated distributions, variable selection problems, etc.). One of these extension proposed the addition of moves beyond the flipping operator or the extension of the velocity domain beyond $\mathbb{V} = \{-1; +1\}^d$ (see for example Chevallier et al. (2020)). Vasdekis and Roberts (2021) proposed the use of a function $S(\mathbf{x}_t)$ that allows the acceleration of the process according to its position (e.g. speeding up in the tails). In a recent work (Bertazzi and Bierkens 2020), an adaptive version of the Zig-Zag sampler and other PDMP algorithms was proposed, whereby the velocity is changed so that the performance of the algorithm would be equal to that of the canonical Zig-Zag sampler on an isotropic Gaussian distribution. This was proven to substantially improve efficiency.

3 Automatic Zig-Zag sampling

This section describes some methods to allow the *automatic* use of the Zig-Zag process. Here *automatic* means that the only input needed is a differentiable functional form for the potential $U(\mathbf{x}) = -\log(\pi(\mathbf{x})) + c$, where $\pi(\mathbf{x})$ is the target density. Note that this goal, not only implies that manual differentiation of $U(\mathbf{x})$ should not be needed prior to start the analysis, but also that the algorithm should be run (i.e. produce a sample from the PDMP) without relying on any external information about properties of the density such as its concavity or bounds.

3.1 Automatic differentiation

AD is a set of techniques that, given a function $f(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^m$, allows the evaluation of $f^{(i)}(\mathbf{x}_0)$, the derivative of f for a specific point $\mathbf{x}_0 \in \mathbb{R}^n$ w.r.t dimension $i = 1, \dots, n$ (Baydin et al. 2018). Notably, Automatic differentiation, not only provides an exact solution, but also it tends to be efficient: following the *Cheap Gradient Principle*, the computational cost of computing the gradient of a scalar-valued function is nearly the same (often within a factor of 5) as that of simply computing the function itself (Griewank and Walther 2008).

The basis of Automatic Zig-Zag sampling is in computing the rate at Eq. (3) via AD for the point $\mathbf{x}_t = \mathbf{x}_s + \mathbf{v}_s \cdot t$ whenever needed; Algorithm 1 follows identically as before.

Since AD does not introduce any numerical approximation, all results proven for the Zig-Zag sampler (e.g. the main convergence statements of Bierkens et al. (2019)) hold for the Automatic Zig-Zag sampler.

3.2 Rate bounds

In the practical implementation of the Automatic Zig-Zag sampler, the main challenge is to find an upper bound for the

global rate $\lambda(t)$ of the NHPP. While a global or local upper bound to the gradient of $U(\mathbf{x})$ might be known for many distributions of interest, we are looking for a general method that could bound, at least locally, any closed-form density of interest.

Constant upper bounds are used here and should be found under the consideration that if the upper bound is too large, then a large amount of computational effort is wasted in sampling candidate skeleton points (and evaluating $\lambda^{(i)}(t)$) that are then rejected. Therefore, the upper bound should be as close as possible to the time-varying rate $\lambda(t)$. Hence, a pragmatic approach is chosen: the rate bound is defined locally (i.e. specific to the current location and velocity of the process) to be the maximum of the global rate in an interval of size t_{MAX} :

$$\bar{\lambda}(t_{MAX}, \mathbf{x}_s, \mathbf{v}_s) = \max_{t \in [0, t_{MAX}]} \{\lambda(t; \mathbf{x}_s, \mathbf{v}_s)\} \tag{10}$$

which, for brevity is denoted by $\bar{\lambda}$, dropping the notation of the local dependence. If no events are sampled in the NHPP in the interval $[0, t_{MAX}]$, then the Zig-Zag process jumps straight to $\mathbf{z}_{s+t_{MAX}} = (\mathbf{x}_s + \mathbf{v}_s \cdot t_{MAX}, \mathbf{v}_s)$ without any further evaluations of the rates. The rate bound is then re-evaluated for the next interval and sampling continues. Values of t_{MAX} are further discussed in Sect. 3.3.

Since $\lambda(t)$ consists of a blackbox and there is no explicit form of the rate function, finding an analytical maximum is impossible. Among the numerical optimization methods, gradient- and Hessian-free methods are particularly attractive since they are highly efficient and robust for univariate optimization problems, such as this one.

3.2.1 Brent’s optimization method

Similarly to other univariate optimization methods, the goal of this routine is to obtain the minimum of an objective function $f : \mathbb{R}^1 \rightarrow \mathbb{R}^1$ (if the maximum is needed, as in this case, the optimization routine is run on $-f$ instead). Brent’s method (Vetterling et al. 1992) combines inverse parabolic interpolation with Golden Section search (Kiefer 1953).

Parabolic interpolation starts from three points $(a, f(a))$, $(b, f(b))$, $(c, f(c))$ such that $a < b < c$, $f(b) \leq f(a)$ and $f(b) \leq f(c)$, and finds the abscissa x of the vertex of a parabola interpolating the three points via the formula:

$$x = b - \frac{1}{2} \frac{(b-a)^2[f(b)-f(c)] - (b-c)^2[f(b)-f(a)]}{(b-a)[f(b)-f(c)] - (b-c)[f(b)-f(a)]} \tag{11}$$

Substituting the highest point among $(a, f(a))$, $(b, f(b))$, $(c, f(c))$ with $(x, f(x))$ and iterating this formula, until a

fixed tolerance is reached, should approach the minimum of the function f .

The Golden Section search brackets the minimum of $f(x)$ with intervals that are chosen to respect the golden ratio $\frac{1+\sqrt{5}}{2}$, so that their width can be reduced most efficiently.

The Brent method combines these two methods by keeping track of 6 points:

- a / b lowest/highest point of the interval bracketing the minimum
- x best candidate minimum point found so far
- v point with the second least value found so far
- w value of v at the previous iteration
- u point of the most recent evaluation of f

The optimization scheme is as follows:

1. Propose a new point x^* by parabolic interpolation with Eq. (11) on $(x, f(x))$, $(v, f(v))$ and $(w, f(w))$
2. **if** the new point lies in the bracketing interval: $a \leq x^* \leq b$
and convergence is obtained by steps that are increasingly smaller $|f(x) - f(x^*)| \leq 0.5|f(v) - f(w)|$ accept the new proposed point and uprate the bracketing interval to either (a, x) or (x, b)
3. otherwise update the bracketing interval by Golden Search.

These steps are iterated until some tolerance is reached.

Note that the Golden Search method is slow and highly reliable, while polynomial interpolation is much quicker but is founded on the assumption that the function has an approximately parabolic behaviour. Brent’s method would then be at worst as slow as Golden Search method.

3.2.2 Modification for Zig-Zag

In the application considered here, Brent’s optimization method is used to solve Equation (??) and obtain a maximum. In this context, a few considerations can be made:

- (i) If the distribution considered is unimodal, the rates (??) will be often monotonic;
- (ii) If t_{MAX} is chosen to be smaller than the distance to the nearest mode, even in the case of a multimodal distribution, the rates would be mostly monotonic in the optimization interval $[0, t_{MAX}]$;
- (iii) If the function to be maximised is monotonic in the interval $[0, t_{MAX}]$, the maximum is either at 0 or at t_{MAX} .

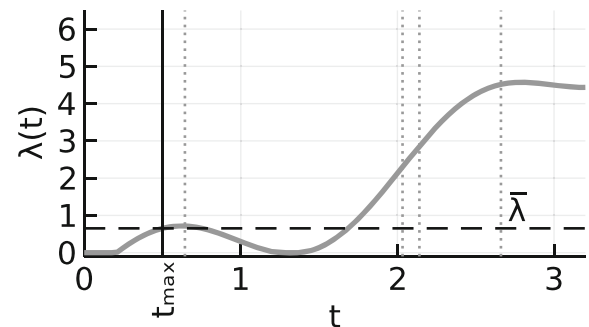
Given these considerations, Brent’s method can be modified and computations can be shortened after some tests

for monotonicity. For this reason, a modification to Brent’s method is proposed: after the first iteration is carried out, a check is run to assess if any of the two limits of the bracketing interval are unchanged. If so, then a second check is performed to confirm that the rate function approaches the end of the interval from below, by evaluating $\lambda(t; \mathbf{x}_s, \mathbf{v}_s)$ a distance ε from the end, for some small $\varepsilon > 0$. If this is the case, the rate is assumed to be monotonic in $[0, t_{MAX}]$ and the value of the rate at the selected limit is taken as upper bound $\bar{\lambda}$; alternatively Brent’s algorithm is run until convergence to the resulting maximum x and set $\bar{\lambda} = x$.

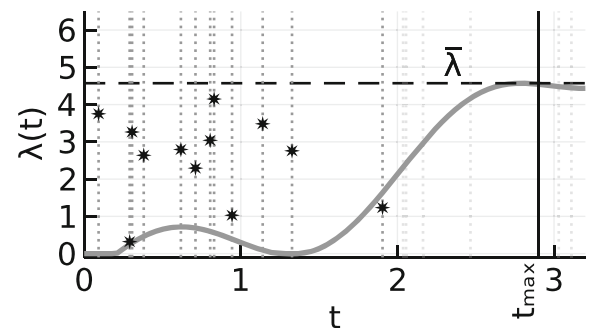
3.3 Tuning of t_{MAX}

With Eq. (11), a parameter t_{MAX} is introduced into the Automatic Zig-Zag algorithm. This is effectively a tuning parameter, with $\bar{\lambda}$ being more or less local according to the magnitude of t_{MAX} .

When t_{MAX} is small, $\bar{\lambda}$ would be very local, with $\lambda(t; \mathbf{x}_s, \mathbf{v}_s)$ varying little in the interval, the rate should be smaller, hence the HPP proposal events should be more rare, making it more likely for the PDMP to reach t_{MAX} without any switch; every time this happens, another optimization step needs to be run



(a) Small t_{max}



(b) Large t_{max}

Fig. 3 Tuning of t_{MAX} . **a** A deterministic step encouraged by the choice of a small value. **b** Many proposals of HPP prior to acceptance, caused by a very large value of the local bound

to obtain a new bound $\bar{\lambda}$. On the other hand, if t_{MAX} is very large, HPP events are likely to be proposed more often, and for all the proposed times the rate $\lambda(s)$ has to be evaluated. An illustration of this tuning criterion can be found in Fig. 3.

Algorithm 2 Automatic Zig-Zag algorithm with thinning and local optimization.

```

Require: Initial velocity  $\mathbf{v}_0$ ; initial location  $\mathbf{x}_0$ ; total number of skeleton
points to be sampled  $K$ ; functions to compute the dimension-wise
and global switching rate  $\lambda(\cdot)$  and  $\lambda^{(i)}(\cdot)$ ;  $t_{MAX}$  the tuning parameter.
Ensure: Time, location and velocity at  $K$  skeleton points:
 $\{t_k, \mathbf{x}_k, \mathbf{v}_k\}_{k=1}^K$ 
1:  $t_0 = 0, k = 1$  #set starting time and skeleton count
2:  $T \leftarrow t_0, X \leftarrow \mathbf{x}_0, V \leftarrow \mathbf{v}_0$  #set current state of the process
3:  $\bar{\lambda} = \max_{t \in [0, t_{MAX}]} \{\lambda(t; X, V)\}$  #compute upper bound
4:  $\tau^* \sim \text{Exp}(\bar{\lambda})$  #propose switching time
5:  $\tau^{opt} \leftarrow \tau^*$  #track time from last optimization
6: while  $k \leq K$  do
7:  $u = 0$  #set acceptance to 0 until next proposal
8: while  $\tau^{opt} \leq t_{MAX}$  and  $u = 0$  do
9:  $\lambda(\tau^{opt}) = \sum_{i=1}^d \lambda^{(i)}(\tau^{opt}; X, V)$  # compute the global rate at
 $\tau^{opt}$ 
10:  $u \sim \text{Ber}(\frac{\lambda(\tau^{opt})}{\bar{\lambda}})$  # accept/reject  $\tau^{opt}$ 
11: if  $u = 1$  then
12:  $m \sim \text{Multinom}\left(1 : d; \left\{\frac{\lambda^{(i)}(\tau^{opt}; X, V)}{\lambda(\tau^*)}\right\}_{i=1}^d\right)$  #sample
component to switch
13:  $T \leftarrow T + \tau^{opt}$  #progress time
14:  $X \leftarrow X + V\tau^{opt}$  #progress location
15:  $V \leftarrow F_m(V)$  #flip velocity of dimension  $m$ 
16:  $t_k \leftarrow T, \mathbf{v}_k \leftarrow V, \mathbf{x}_k \leftarrow X$  #save skeleton point
17:  $k = k + 1$  #increase skeleton count
18:  $\bar{\lambda} = \max_{t \in [0, t_{MAX}]} \{\lambda(t; X, V)\}$  #compute new upper
bound from switch
19:  $\tau^* \sim \text{Exp}(\bar{\lambda})$  #propose switching time
20:  $\tau^{opt} \leftarrow \tau^*$  #reset time from last optimization
21: else
22:  $\tau^* \sim \text{Exp}(\bar{\lambda})$  #sample new time increment
23:  $\tau^{opt} \leftarrow \tau^{opt} + \tau^*$  #compute new switching proposal
24: end if
25: end while
26: if  $\tau^{opt} > t_{MAX}$  and  $u = 0$  then #if the horizon is reached with
no switch
27:  $T \leftarrow T + t_{MAX}$  #progress time deterministically until horizon
28:  $X \leftarrow X + Vt_{MAX}$  #progress location deterministically until
horizon
29:  $V \leftarrow V$  #retain velocity
30:  $\bar{\lambda} = \max_{t \in [0, t_{MAX}]} \{\lambda(t; X, V)\}$  #compute new upper bound
from new location
31:  $\tau^* \sim \text{Exp}(\bar{\lambda})$  #propose switching time
32:  $\tau^{opt} \leftarrow \tau^*$  #reset time from last optimization
33: end if
34: end while
    
```

The optimal t_{MAX} is chosen by minimizing the number of evaluations of the rate $\lambda(s)$ per switching time, which includes both the evaluations within the optimization algorithm and the computation of the acceptance probabilities. This can be done via some preliminary runs of the algorithm.

The pseudo-code of the Automatic Zig-Zag sampling taking as input a value of t_{MAX} is reported in Algorithm 2.

4 Performance evaluation

This section investigates the performance of the Automatic Zig-Zag sampler. The performance is tested on some bivariate distributions starting from an uncorrelated bivariate normal and exploring increasingly-more-challenging features. Main results are reported in Sect. 4.3 and an exhaustive description of each simulation is reported in Online Resource 1.

4.1 Performance metrics

Performance is evaluated according to two criteria: efficiency and robustness.

4.1.1 Efficiency

To measure efficiency, the Effective Sample Size (ESS) of the sample drawn with the two algorithms is compared; the samplers are run given a specific budget. The computational budget c is defined as the total number of evaluations of the gradient of the minus-log density of the target distribution $(\nabla U(\mathbf{x}))$.

For the Automatic Zig-Zag algorithm, the number of gradient evaluations required to produce each skeleton point comprises, for skeleton point k : C_k^{OPT} , the number of evaluations of the switching rate during the optimization routine to find the bound $\bar{\lambda}$; and C_k^{TPP} , the number of proposed times for the thinned Poisson process. The number of evaluations over all the sampled skeleton is:

$$C^{ZZ} = \sum_{k=1}^K \{C_k^{OPT} + C_k^{TPP}\} \tag{12}$$

and therefore, the sampler stops at the smallest K such that $C^{ZZ} \geq c$.

For a canonical HMC algorithm that performs L leapfrog steps per iteration and K iterations, the number of evaluation of the gradient is:

$$C^{HMC} = (L + 1) \times K. \tag{13}$$

Hence the sampler is run for $K = \frac{c}{L+1}$ steps.

The Automatic-Zig-Zag efficiency is computed using the ESS for continuous-time trajectories presented in Bierkens et al. (2019) (Supplementary Information S.2) for the function $h(\mathbf{x}) = x_i$ for all the i coordinates. Similarly, the batch-means approach for ESS calculated from discrete-time samples is used to evaluate the efficiency of the runs of the HMC algorithm. To summarise the results in ESS across multiple dimensions, it is useful to compare the dimension with smallest ESS (Median ESS over 100 independent chains)

since this dimension mixes more slowly and hence constrain the chain to an overall slower mixing.

4.1.2 Robustness

The other aspect examined to assess the performance of the Automatic Zig-Zag sampler was whether or not the algorithm was robust with respect to particular features of the distribution (e.g. heavy or light tails, multimodality).

In particular, the ability of a tuned algorithm to properly explore the target distribution was investigated, even when starting from location far away from the mode. This was conducted mainly graphically and robustness was assessed qualitatively.

4.2 Simulation set up

The Automatic Zig-Zag algorithm is compared with the Canonical HMC algorithm (for a description of the latter see Sect. 3 of Neal et al. (2011) or Section S1 of Online Resource 1 of this paper). The HMC algorithm is said to be *canonical* when, in the velocity-position framework similar to the one defined above, the velocity is sampled from an independent multivariate Normal distribution. This is a rigid structure, compared to other versions of the HMC algorithm that choose a velocity distribution optimally with respect to the target density. Similarly, the version of Zig-Zag sampler used here is the canonical Zig-Zag, which employs constant velocities in $\{-1, +1\}^d$, with no attempt to choose an optimal velocity that matches to the target distribution.

Both algorithms are tuned before the comparison via preliminary runs. More specifically, t_{MAX} is chosen according to the criterion explored in Sect. 3.3, while the choice of the tuning parameters of the HMC (i.e. the total integration time $L \times \varepsilon$ and of the number of leapfrog steps L) is known to be a troublesome task (Sherlock et al. 2021). The procedure adopted here for tuning includes many graphical assessments and is reported in Section S1 of Online Resource 1.

4.3 Results

The results of the efficiency analysis on various forms of Bivariate Gaussian distribution are reported in Table 1. The algorithms were tested on an isotropic Gaussian distribution (IsoG2); on a bivariate Gaussian distribution where the two components had the same scale and high correlation $\rho = 0.9$ (CorG2); on a bivariate Gaussian distribution with independent components with very different scales $\sigma_1^2 = 1, \sigma_2^2 = 100$ (DscG2); and on a bimodal distribution, a mixture of Gaussians (BimodG2).

The two algorithms performed very similarly on IsoG2 (with a ESS less than 20% larger when HMC was used), HMC proved to be 4 to 5 times more efficient than Automatic Zig-

Table 1 Smallest ESS (Median) obtained with the Automatic Zig-Zag algorithm and HMC algorithm given a pre-specified budget on bivariate Gaussian distributions and heavy-/ light-tailed distributions

Target	Sampler	min Me ESS
IsoG2	ZZ	1723
	HMC	2049
CorG2	ZZ	317
	HMC	1419
DscG2	ZZ	261
	HMC	43
BimodG2	ZZ	185
	HMC	727
LT2	ZZ	1311
	HMC	2820
HT2	ZZ	85
	HMC	182

Zag sampling on CorG2. Conversely, Zig-Zag sampling was 6 to 7 times more efficient than HMC on DscG2. Despite the intrinsic advantage of HMC, which is built to perform excellently on Gaussian targets, the observed comparable efficiency shows that the Automatic Zig-Zag sampling is competitive.

With respect to robustness on these Gaussian targets, both algorithms performed well: the chains started in the mode reached the tails with adequate frequency and the chains initiated in the tails quickly converged towards the mode and continued to explore the target distribution.

The performance was then tested against an heavy-tailed bivariate target (HT2) and a light-tailed bivariate target (LT2). The former is assumed to be distributed according to a bivariate Student-T with 2 degrees of freedom and the latter is assumed to have density $p(x) \propto e^{-\sum_{i=1}^d x_i^4/4}$ for $d = 2$. HMC was twice as efficient as Automatic Zig-Zag on HT2, whilst on LT2, HMC was almost two times more efficient than Automatic Zig-Zag.

The Automatic Zig-Zag algorithm however, proved to be more robust to these two examples providing consistent exploration of the tails in HT2 and fast convergence towards the mode when starting in the tails for both HT2 and LT2. These are reported graphically in Fig. 4 where multiple chains starting from a grid of values in the tails of the distribution were run for a limited number of iterations/skeleton points. In Figs. 4a and c the rapid convergence towards the mode of the Zig-Zag algorithm can be appreciated. Conversely, the HMC chains struggled to move towards the mode of the heavy tailed distribution (Figure 4b) and did not move at all on the light-tailed distribution (Figure 4d): the gradient in these locations suggested proposals far off in the opposite tail which were then never accepted.

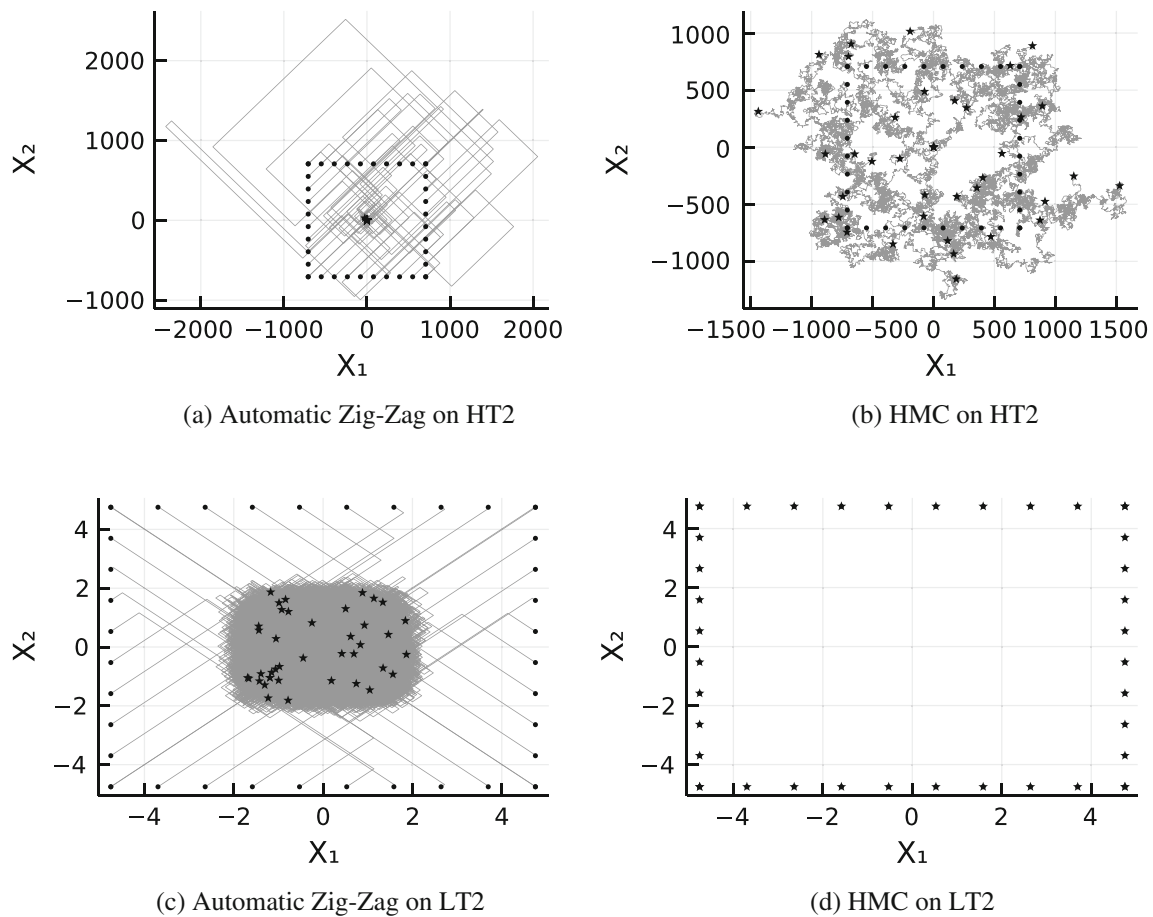


Fig. 4 Robustness of the algorithms on heavy- and light-tailed distributions: the two algorithms are tuned a priori and the chains are initiated at a grid of values in the tail of the distribution (dots). The Zig-Zag algo-

rithm stops at 1000 skeleton points and HMC stops at 1000 iterations (final sample denoted by a star). When the chains don't move, such as in panel (d), the first and last sample overlay

Comprehensive results from the simulation study, including illustrations of the optimality of the tuning of the Zig-Zag algorithm, are reported in Section S3 of Online Resource 1.

5 Real data applications

In this section, some examples of the application of Automatic Zig-Zag sampling to real data analyses are proposed. The first is an example of a non-linear regression model from a Bayesian Methods textbook (Carlin and Louis 2008); and the second example is a parametric survival model.

5.1 A textbook example

We reproduce the analysis of (Carlin and Louis 2008, page 176), which analyses data on dugongs (sea cows), considering a non-linear growth model to relate their length in meters

(Y_j) to their age in years (z_j). The model assumed is:

$$Y_j = \alpha - \beta\gamma^{z_j} + \varepsilon_j \quad \text{for } j = 1, \dots, J \tag{14}$$

with normally distributed errors $\varepsilon_j \stackrel{iid}{\sim} N(0, \sigma^2)$.

The parameters are $\alpha > 0, \beta > 0, 0 \leq \gamma \leq 1, \sigma > 0$; the parameters are explored on the following transformed space:

$$\begin{aligned} x_1 &= \log(\alpha) \\ x_2 &= \log(\beta) \\ x_3 &= \log\left(\frac{\gamma}{1-\gamma}\right) \\ x_4 &= \log(\sigma). \end{aligned} \tag{15}$$

The priors are assumed flat on their original domain except for γ which has a Beta(7, 7/3) prior. This model presents some challenges in that this parametrization favours correlation in the posterior distribution and different scales for the parameters.

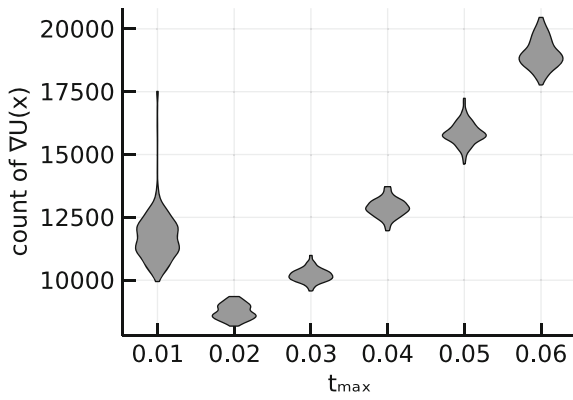


Fig. 5 Selection of t_{MAX} via preliminary runs. Violin plots of the total number of gradient evaluations form 100 samples of 1000 skeleton points at given values of t_{MAX}

The selection of an appropriate t_{MAX} was done via preliminary runs, as described in Sect. 3, that guided the choice of an efficient value at $t_{MAX} = 0.02$ (see Figure 5).

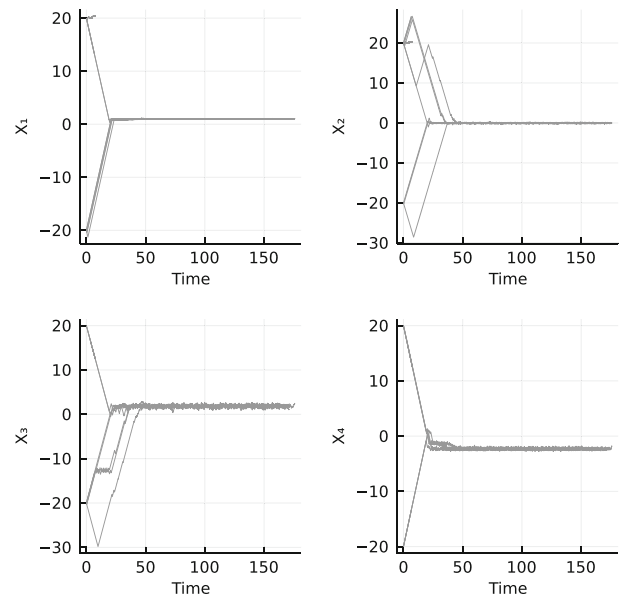
The comparative results against HMC showed the same pattern observed in Sect. 4: HMC was slightly faster than Zig-Zag in exploring the space, leading to an increased ESS given a limited budget. Zig-Zag however was much more robust to the choice of initial values: it was able to reach the mass of the distribution very quickly. Conversely HMC often remained stuck in initial values (or in other values away from the mode), struggling to reach convergence (Figure 6). This behaviour was also observed when more elaborate adaptations of HMC were used, such as the Non U-Turn Sampler (Hoffman and Gelman 2014) implemented in the software Stan (Carpenter et al. 2017).

5.2 Parametric survival model

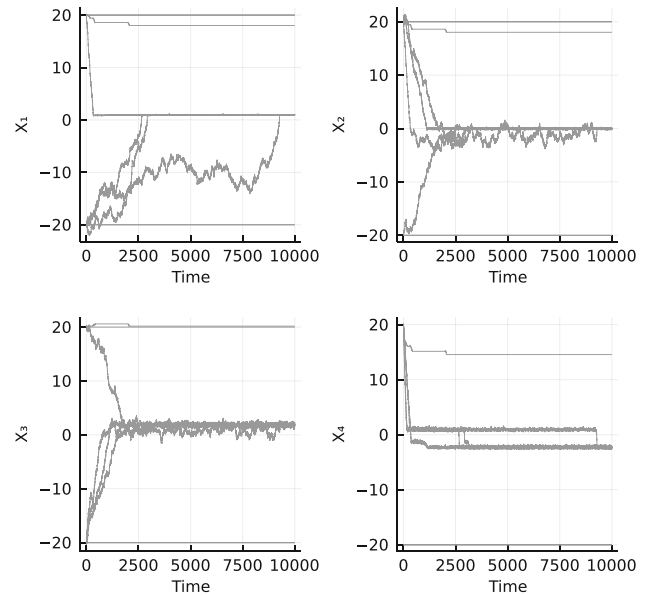
Automatic Zig-Zag was tested on the inference of a Bayesian parametric survival regression model fitted to a sample of individuals from a large synthetic database (Health Data Insight CiC, n.d.). The whole data are described below and a model was fitted initially to a sample of 500 individuals. The dataset is analysed in full in the next section, where automatic super-efficiency is explored.

5.2.1 Data

The dataset comprises information on 2,200,626 synthetic patients and their 2,371,281 synthetic tumours, including the time of each cancer diagnosis, the time/type of final event observed (i.e. time of death if dead or censoring time if alive), basic demographics of the patients and on their tumour history (e.g. time of surgery if surgically addressed, therapy type and timings).



(a) Automatic Zig-Zag skeleton



(b) HMC samples

Fig. 6 Skeleton plot **a** and HMC chains **b** for the 4 parameters when initiated in location not close to the mode: while Zig-Zag rapidly converges to the true value, HMC takes much longer to reach convergence or, in some case does not move

A parametric survival regression model (Jackson 2016) was fitted to these data in order to explain the survival-time from first tumour diagnosis with few individual-specific covariates. Note that the results reported here should not be interpreted as real, not only because the data used are synthetic, but also because the effects estimated here should be corrected for other covariates which were not included in this analysis and are known to affect and confound survival

from diagnosis. Other simplifying assumption were made, including uninformative missingness, uninformative loss to follow-up and no left censoring. Thanks to the high completeness of the dataset only 2,565 patients were excluded due to missing at least one key variable (i.e. date/type of final outcome).

The final dataset analysed consisted of: a set of times t_j from diagnosis of the first tumour to either death or censoring; a set of event type c_j , with $c_j = 1$ for death and $c_i = 0$ for (administrative) censoring; and a set of covariates z_j^1, \dots, z_j^g for $j = 1, 2, \dots, J$, with $J = 2,198,061$ individuals.

5.2.2 Model

A Weibull model was assumed, i.e. the time to death has probability density function:

$$f(t; \mu, \alpha) = \frac{\alpha}{\mu} \left(\frac{t}{\mu}\right)^{\alpha-1} e^{-\left(\frac{t}{\mu}\right)^\alpha} \tag{16}$$

and survival function:

$$S(t; \mu, \alpha) = e^{-\left(\frac{t}{\mu}\right)^\alpha} \tag{17}$$

so that the overall likelihood of the vectors of outcomes $\mathbf{t} = t_1, t_2, \dots, t_J$ and $\mathbf{c} = c_1, c_2, \dots, c_J$ is:

$$\ell(\mathbf{t}, \mathbf{c}|\mu, \alpha) = \prod_{j=1}^J \mathbb{I}_{c_j=1} f(t; \mu, \alpha) + \mathbb{I}_{c_j=0} S(t; \mu, \alpha) \tag{18}$$

The scale parameter μ was related to the covariate of interest z^1, \dots, z^g via log link:

$$\log(\mu_j) = \beta_0 + \beta_1 z_j^1 + \dots + \beta_g z_j^g \tag{19}$$

Let z_j^1 be the age at diagnosis of patient j , and z_j^2 be the discrete variable identifying the spreading status of the cancer: if $z_j^2 = 0$, the cancer of patient j haven't spread to other sites (i.e. it is in stage 2 or smaller) if $z_j^2 = 1$, the cancer of patient j is likely to have spread to other sites (i.e. it is in stage 3 or greater).

In the Zig-Zag notation, the location vector \mathbf{X} was then composed by all the parameters of the model:

$$\mathbf{X} = (\log(\alpha), \beta_0, \beta_1, \beta_2). \tag{20}$$

5.2.3 Results for 500 individuals

A randomly selected subset of $J = 500$ individuals was initially analysed.

In this model, the parameter space is slightly unbalanced: the first component ($\log(\alpha)$) highly affects the shape of the

potential, constraining all the other components, hence the MCMC is doomed to mix slowly overall. This ill-behaviour is a combination of two aspects explored in the simulations of Sect. 4: the components of \mathbf{X} have different scales and are highly correlated.

The Zig-Zag sampler performed satisfactory in exploring this challenging target distribution: it was shown to be more robust than a properly-tuned HMC (results reported in Section S4 of Online Resource 1). Moreover, the Zig-Zag sampler was shown to be more efficient than HMC, achieving systematically higher ESS on all dimensions as reported in Figure 7.

These results come from the analysis of a small subset of the population but, as more data are included, the evaluation of the likelihood and its gradient becomes more and more expensive, and the overall exploration of the space is slower. This motivates the need to exploit super-efficiency which is described in Sect. 6 in a general context. Results from the analysis of the full dataset using our super-efficient Zig-Zag sampler are presented in Sect. 6.4.

6 Automatic super-efficiency

One of the most appealing properties of the Zig-Zag algorithm, and of PDMPs more generally, is super-efficiency. An algorithm is defined to be *super-efficient* if it “is able to generate independent samples from the target distribution at a higher efficiency than if we would draw independently

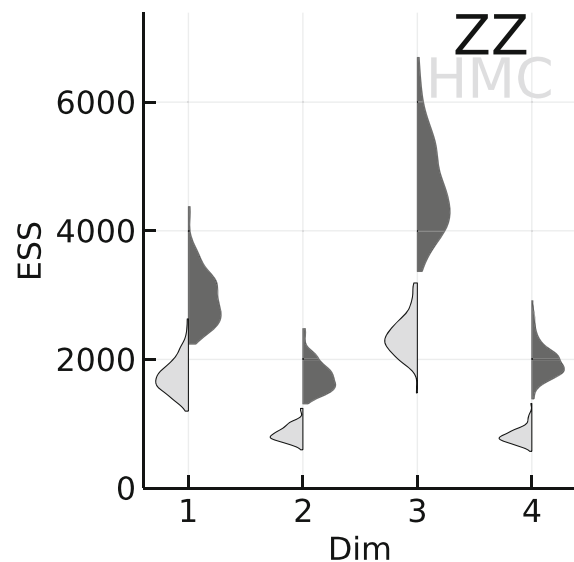


Fig. 7 Violin plot of the ESS of each dimension obtained on 100 simulated chains of HMC and Automatic Zig-Zag given a pre-sepcified budget of 200000 gradient evaluations

from the target distribution at the cost of evaluating all data” (Bierkens et al. 2019).

6.1 Subsampling

Super-efficiency can be obtained if the potential $U(\mathbf{x})$ takes a particular form. Specifically, consider $U(\mathbf{x})$ for which $\partial_i U(\mathbf{x}) = \frac{\partial U(\mathbf{x})}{\partial x_i}$ admits representation:

$$\partial_i U(\mathbf{x}) = \frac{1}{J} \sum_{j=1}^J E_i^j(\mathbf{x}) \tag{21}$$

for $i = 1, \dots, d$. This representation is available, for example, when the target density can be factorised in a series of J components (e.g. a sum of J observation-specific likelihoods of independent and identically distributed (iid) observations).

With representation (21), the following steps allow the construction of an algorithm to sample from the correct target distribution.

1. Define a dimension-specific collection of switching rates (with $i = 1, \dots, d$ indexing the dimension), where each element of the collection can be thought of as the observation-specific factor of the potential (with $j = 1, \dots, J$ indexing the observation):

$$m_i^j(t) := \max \left\{ v_i E_i^j(\mathbf{x}(t)), 0 \right\} \tag{22}$$

for $i = 1, \dots, d; j = 1, \dots, J$.

2. Find a collection-specific function $M_i(t)$ which bounds all the rate of a specific dimension i :

$$m_i^j(t) \leq M_i(t) \quad \text{for all } j = 1, \dots, J$$

for $i = 1, \dots, d$. This bound can vary over time t or be constant, i.e. $M_i(t) = c_i$.

3. Sample the first event time from d homogeneous Poisson processes: $\tau_i \sim PP(M_i(t))$ and take:

$$\begin{aligned} \tau &= \min \{ \tau_1, \tau_2, \dots, \tau_d \} \\ i_0 &= \operatorname{argmin} \{ \tau_1, \tau_2, \dots, \tau_d \}. \end{aligned}$$

4. Sample an index of the observations:

$$j_0 \sim \text{Uniform}(1, 2, \dots, J).$$

5. Accept the switch for dimension i_0 with probability $m_{i_0}^{j_0}(\tau) / M_{i_0}(\tau)$.

The process of using only one observation (or, any other unbiased estimator of $\partial_i U(\mathbf{x})$ in (21) which uses less than J computations) is called *subsampling*. Subsampling as

described above (i.e. when only one observation is used) allows to reduce computational complexity of the algorithm by a factor $O(J)$. This result has been proven in Bierkens et al. (2019) and a few considerations were drawn: the resulting chain mixes more slowly than a chain obtained with the non-subsampling algorithm; nevertheless, control variates can be used to further improve the efficiency of the Zig-Zag with subsampling.

A straightforward way to extend the methods presented in Sect. 3 is to allow the input to be directly the observation-specific density E_j^i , with the formulation of a generic potential which depends on the observation index j .

6.1.1 Challenges

To properly implement subsampling, a collection-specific upper bound $M_i(t)$ (or a constant bound c_i) must be available, but in a generic example it may not be possible to find a bound analytically. With the introduction of an automatic method, all the functional information on the derivatives of the potential is lost.

To address this issue, a constant-local approach is again adopted: it would be sufficient to find a value c_i for given starting values $(\mathbf{v}_s, \mathbf{x}_s)$ within an horizon of length t_{MAX} : ($t \in [0, t_{\text{MAX}}]$), so for a specific dimension i the bounding rate would be:

$$m_i^j(t) \leq c_i \quad \forall j = 1, 2, \dots, J \text{ for } t \in (0, t_{\text{MAX}}). \tag{23}$$

If this approach is taken, c_i refers specifically to the starting values $(\mathbf{v}_s, \mathbf{x}_s)$ and a new c_i should be considered whenever a switch or a deterministic move is made. Even within this horizon $[0, t_{\text{MAX}}]$, however, finding a maximum by evaluating and maximizing all the J observation-specific rates and then comparing them would be counter-productive: all the gain of super-efficiency would be lost in this optimization step. A super-efficient method to overcome this challenge is proposed below.

6.2 Bounding unknown rates

The main idea of our proposal to find an efficient estimate \hat{c}_i of c_i is to consider only a small sample of size q of the available switching rates, maximise them to obtain a sample of rate-specific maxima/bounds and finally apply extreme-value theory methods to infer the population maximum across all the rates.

Given a local starting point $(\mathbf{v}_s, \mathbf{x}_s)$ and within a given horizon of length t_{MAX} , an estimate \hat{c}_i of c_i is obtained with the following steps:

1. select a sample \mathcal{Q} of size q from the J rates available in the collection;

- run a numerical optimization algorithm (e.g. our version of Brent’s method) to obtain rate-specific maxima of the $q \times d$ dimension-specific sampled rates:

$$\bar{\lambda}_i^j = \max_{t \in (0, t_{MAX})} m_i^j(t);$$

for $j \in \mathcal{Q}$, for $i = 1, \dots, d$;

- for each dimension $i = 1, \dots, d$, use the q values of $\bar{\lambda}_i^j$ to fit a Generalised Pareto Distribution (GPD) and obtain estimated parameters $\hat{\xi}_i, \hat{\sigma}_i$ of the GPD;
- use the parameters to predict c_i in a *return value* perspective: $\hat{c}_i = q_{\frac{1}{J}}^{(i)}$, with $q_{\frac{1}{J}}^{(i)}$ the $1 - 1/J$ th quantile of the extreme value distribution with parameters $\hat{\xi}_i, \hat{\sigma}_i$; for each dimension $i = 1, \dots, d$.

The estimated \hat{c}_i can be then considered as the *population bound*: the value that would be the maximum (the only one at or above its value) if we had a sample of size J .

More detail on the results used from extreme-values is reported in Online Resource 1, Section S5. If the acceptance step of the subsampling algorithm shows that \hat{c}_i is found not to bound some rates, than a new set of rates \mathcal{Q} is drawn and steps 2 to 4 above are run again.

6.3 Practical considerations

The method proposed in Sect. 6.2 still retains the *automatic* flavour of the algorithms proposed here but allows to exploit what is thought of as the most-promising property of Zig-Zag samplers and other PDMPs. In implementing this idea in practice, however, a few choices must be made.

Firstly, one should decide on the *level of super efficiency* desired: one of many iid observations already provides a unbiased estimate for the rate, but it might be better to include more, say h , observations in order to have a more-representative sample of the population. The larger h , the more homogenous the subsample-specific rates are. As a consequence the process mixes better as the subsample-specific rates resemble better the population rate. When one, or very few observations contribute to each subsample-specific rate, the process will switch often reflecting the heterogeneity across them.

Likewise, q , i.e. the number of rates that are selected for the estimation of the bounds, highly affects the quality of the estimator \hat{c}_i , which, if underestimated, could lead to the samples from the ZZ sampler being overdispersed with respect to the target distribution. A robustness factor $r \geq 1$ is introduced so that the upper bound is effectively larger than the predicted return value by the Generalised Pareto: $\hat{c}_i = r \times q_{\frac{1}{J}}^{(i)}$.

These quantities: the number of observations per rate, q , and r , should be considered tuning parameters and chosen on

a case-by-case basis via preliminary analysis as exemplified in the following section. For example, finding that that the rates exceed their estimated bound \hat{c} often, suggests that r might have to be increased.

Lastly, note that, while q rates are needed to infer \hat{c} , the optimization routine on each of these rates could be paral-
lised: the q maxima $\bar{\lambda}_i^j$ can be computed independently, enabling even higher efficiency.

6.4 Parametric survival model on big data

In this section we fit the Parametric survival model of Eqs. (16)–(19) to the total population of $J = 2, 198, 061$ individuals.

As a starting point, we attempted the more computationally-expensive approach of using the standard Automatic Zig-Zag algorithm, whose results are reported in blue in Figure 8. To obtain such a skeleton (composed by 5000 switching times), circa 63,000 gradient evaluations were made, each of which is a computation of order $J \approx 2$ million. The overall clock time elapsed was 4 hours, after careful tuning of t_{MAX} .

6.4.1 Subsampling setup

We ran the Automatic Zig-Zag with subsampling, where at each iteration the rate was approximated by considering a potential that accounts for fewer than J observations.

When we implemented the most drastic subsampling, using only one observation and we approximated the rate by $m_i^j(t)$ as defined in (22), many problem arose. The switching rates were very heterogeneous in the observation that was (sub-)sampled which meant that it was difficult to estimate appropriate bounding constants \hat{c}_i that were valid for

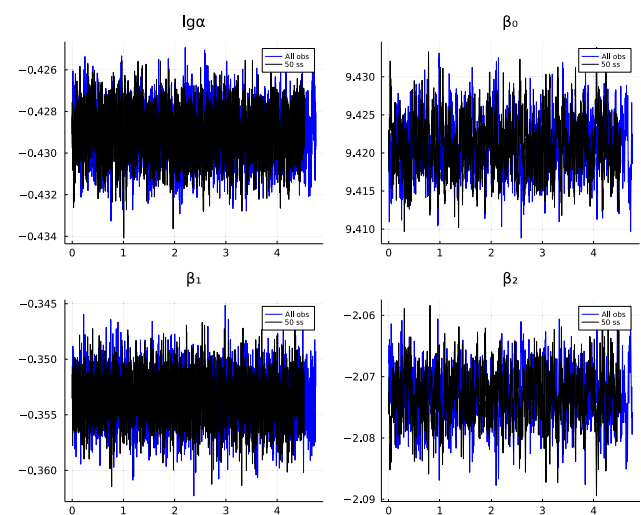


Fig. 8 Skeleton of the Zig-Zagprocess without subsampling (black) and with subsampling (blue)

the whole population; for the same reason mixing was very slowly.

We therefore decided to use subsamples of size $h > 1$ to approximate the rates. Let S_l be a sample of size h of indexes, drawn without replacement from $\{1, 2, \dots, J\}$. The rates are generated using estimates for the potential of the type:

$$E_i^l(x) = \sum_{j \in S_l} \frac{E_i^j}{h}.$$

A subsample size of $h = 20$ gave very satisfactory results with robust estimates of \hat{c} and good mixing.

The estimates of \hat{c} were obtained by evaluating $q = 1000$ other rates, whose maxima were used to fit a GPD; the robustness factor was set to $r = 2$.

6.4.2 Results

The resulting 5000 switching-point skeleton appears to have mixed well and converged to the same distribution as its expensive, non-subsampling counterpart (see Figure 8).

The overall clock-time elapsed is circa 30 minutes, hence the gain from applying the subsampling techniques is tangible: our implementation of the subsampling technique was 7 to 8 times faster than the standard method. The algorithm was run without any parallelization in the estimation of the local upper bounds, hence an even shorter computation time could be achieved. Moreover, because the implementation of the Automatic Zig-Zag with subsampling runs substantially faster, a more precise estimate of the optimal t_{MAX} can be produced from the same computational budget. For the full data it was almost impossible to accurately tune t_{MAX} , given the long computation time, and our initial guess led to a run-time of 24 hours, which was reduced to 4 hours only after using the t_{MAX} obtained from the pilot runs of the sub-sampling algorithm.

While the choice of the level of subsampling h was done by trial and error, it is a straightforward process that, thanks to the speed of the algorithm, can be performed a priori. Notably, even if the Automatic Zig-Zag with subsampling requires a more accurate tuning of the parameters h , r and q , it still retains the automatic properties that the original algorithm has, since no further information on the shape or properties of the target distribution were used.

7 Discussion

The theory behind PDMPs is developing quickly and forming a substantive body of results that make PDMP-based algorithms extremely promising. Little work exists on the use of these algorithms to address applied problems, with notable exceptions including: variable selection problems (Chevalier et al. 2020), inference of diffusion bridges (Bierkens

et al. 2021), and inference of phylogenetic trees (Koskela 2022). These applications develop bespoke versions of the Zig-Zag sampler, and other PDMP-based algorithms, and demonstrate their usefulness and efficiency within the specific applications considered.

Generalisations of PDMP algorithms that make them applicable in any context are even more rare: the simulation of a PDMP is strictly constrained by the availability of adequate upper bounds of the switching rate or by closed-form solutions to the integral of the rate for the time-scale transformation. To our knowledge, there are only two papers that provide a general tool to draw samples using PDMPs requiring only the evaluation of the gradient of the target density. The Numeric Zig-Zag (NuZZ) (Pagani et al. 2022) uses numerical integration to simulate the next switching event by time-scale transformation. The numeric integrator requires the evaluation of the rate $\lambda(t)$ for a grid of values for t (from 7 to 14 points), and it is computed at each iteration of a root-finding method that derives the switching time (τ in Equation 7). While there might be cases when the NuZZ is the most efficient solution, we have found that its numerical routine requires more evaluations of $\lambda(t)$ per switching point compared to our algorithm, whose optimization method resulted extremely efficient, requiring often only 4 evaluations of $\lambda(t)$; the appropriate tuning of t_{MAX} keeps the total number of Poisson Process proposals for thinning small and, in the best cases, around 1. Lastly, the NuZZ is, differently from ours, an approximated algorithm, whose error diminishes as the number of points used for the numerical integrator increases.

Another simulation scheme for PDMPs is proposed in Bertazzi et al. (2021), which solves the same problem by exploiting Euler approximations of the switching rate, abandoning once again exactness for the sake of generalizability. Similarly to the NuZZ, approximation schemes require the evaluation of $\lambda(t)$ for a grid of values, jeopardizing efficiency.

Our work instead welcomes an intensive use of modern AD techniques, which allow the exploration of any target whose (minus log) density is differentiable. Rates computed via AD are matched with a numeric optimization method that allows the quick computation of a local upper bound to sample the switching time via thinning. The resulting Automatic Zig-Zag sampler provides a robust and general way to sample from any distribution with differentiable log-density without the need of any further information. We tested Automatic Zig-Zag and showed it to be competitive with HMC: although HMC is often apparently more efficient, we found it to be considerably less robust when more challenging situations are presented and when starting values are far from the support of the target distribution. On most of the real-data scenarios presented, the Automatic Zig-Zag sampler was shown

to be superior to HMC providing robust results with a simpler tuning process.

In addition to automation of the differentiation and upper bound calculation, the Automatic Zig-Zag sampler has been further extended to benefit from super-efficiency, the most appealing property of PDMP samplers. The power of super-efficiency in this context has been demonstrated in practice on the analysis of a large dataset.

Automatic Zig-Zag presents only a few limitations, the first of which is the use of a numeric method to determine a local upper-bound on the switching rate. As most of the available optimization methods, Brent’s optimization (and our modified method) does not guarantee convergence to a global maximum in the interval considered. Nevertheless, we have found that in practice the method is robust and it rarely fails on the type of functions that need to be bounded in the Zig-Zag algorithms and, given its low computational burden, we were able to introduce further checks to prevent avoidable errors in the computation of the upper bound. Moreover, the tuning parameter t_{MAX} , i.e. the width of the interval over which the optimization is run, can be reduced to decrease the probability of optimization failure. A similar consideration applies to the method presented in Section 6 which lacks guarantees that the estimator \hat{c} would bound all the rates. Nevertheless, we again introduce checks and parameters that can make automatic super-efficiency more robust. Another limitation of the work presented here is that it contemplates only smooth densities on unbounded domain. The general question of the behaviour of PDMPs on piecewise-smooth and bounded densities is addressed in Chevallier et al. (2021), however, the results presented were derived using the knowledge of the discontinuities in smoothness and on the bound, hence they are not applicable in a general context.

Another possible improvement to the Automatic Zig-Zag sampler is the adaptation of the velocity space to the target density considered, similarly to Bertazzi and Bierkens (2020). This would improve the general performance of the algorithm, not only in the aspects described by Bertazzi and Bierkens (2020), but also it should lead to a choice of t_{MAX} that is homogeneously optimal for all dimensions. Progress in this direction is the focus of our future work.

Lastly, while a supplementary code of this paper is provided and contains useful functions to understand and replicate our methods, a full package that implements the Automatic Zig-Zagsampler for Bayesian analyses is being developed to make this method usable by practitioners in all settings.

The availability of a continuous-time algorithm that provides samples from a desired target requiring only a functional form for its (minus log) density opens several possibilities for probabilistic programming languages, substantially advancing the current state of the art. In this paper we have made contributions which facilitate the use of PDMP

methods on a substantially expanded family of targets, and we hope that our work can therefore greatly expand the wide applicability of PDMPs.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11222-022-10142-x>.

Acknowledgements This work has been funded by EPSRC grant EP/R018561/1, New Approaches to Bayesian Data Science: Tackling Challenges from the Health Sciences. GOR is further supported by The Alan Turing Institute. The authors are grateful to professor Jonathan Tawn for his insights and comments on the extreme-values methods used in Sect. 6.

Data Availability Data and model for the case study in Sect. 5 is available in Carlin and Louis (2008), (page 43). The case-study in Sects. 6 and 5 uses artificial data from the Simulacrum, a synthetic dataset developed by Health Data Insight CiC derived from anonymous cancer data provided by the National Cancer Registration and Analysis Service, which is part of Public Health England Health Data Insight CiC (n.d.) and is available to download from <https://simulacrum.healthdatainsight.org.uk/using-the-simulacrum/requesting-data/>.

Declarations

Ethical Approval Data analysed in Sects. 6 and 5 are artificial hence no Ethical approval or Consent to participate is needed. All code used to generate results and plots is available online at <https://github.com/alicecorbella/ZZpaper>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

Andrieu, C., Livingstone, S.: Peskun-tierney ordering for Markovian Monte Carlo: Beyond the reversible scenario. *Ann. Stat.* **49**(4), 1958–1981 (2021)

Baydin, A.G., Pearlmutter, B.A., Radul, A.A., Siskind, J.M.: Automatic differentiation in machine learning a survey. *J. Mach. Learn. Res.* **18**, 1–43 (2018)

Bertazzi, A., Bierkens, J.: Adaptive schemes for piecewise deterministic Monte Carlo algorithms. (2020). arXiv preprint [arXiv:2012.13924](https://arxiv.org/abs/2012.13924)

Bertazzi, A., Bierkens, J., Dobson, P.: Approximations of Piecewise Deterministic Markov Processes and their convergence properties. (2021). arXiv preprint [arXiv:2109.11827](https://arxiv.org/abs/2109.11827)

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). <https://doi.org/10.1214/18-AOS1715>

- Bierkens, J., Grazzi, S., Kamatani, K., Roberts, G.: The Boomerang sampler. International conference on machine learning 908–918, (2020) ([arXiv:2006.13777](https://arxiv.org/abs/2006.13777))
- 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). <https://doi.org/10.1007/s11222-021-10008-8>
- Bierkens, J., Roberts, G.O., Zitt, P.-A.: Ergodicity of the Zig-Zag process. *Ann. Appl. Probab.* **29**(4), 2266–2301 (2019). <https://doi.org/10.1214/18-AAP1453>
- Bierkens, Joris, Bouchard-Côté, Alexandre, Doucet, Arnaud, Duncan, Andrew B., Fearnhead, Paul, Lienart, Thibaut, Roberts, Gareth, Vollmer, Sebastian J.: Piecewise deterministic Markov processes for scalable Monte Carlo on restricted domains. *Stat. Probab. Lett.* **136**, 148–154 (2018). <https://doi.org/10.1016/j.spl.2018.02.021>
- 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). <https://doi.org/10.1080/01621459.2017.1294075>
- Carlin, B.P., Louis, T.A.: Bayesian methods for data analysis. CRC Press, US (2008). <https://doi.org/10.1201/b14884>
- Carpenter, Bob, Gelman, Andrew, Hoffman, Matthew D., Lee, Daniel, Goodrich, Ben, Betancourt, Michael, Brubaker, Marcus, Guo, Jiqiang, Li, Peter, Riddell, Allen: Stan: A probabilistic programming language. *J. Stat. Software* **76**(1), 1–32 (2017). <https://doi.org/10.18637/jss.v076.i01>
- Chevallier, A., Fearnhead, P., Sutton, M.: Reversible Jump PDMP Samplers for Variable Selection. (2020). [arXiv preprint arXiv:2010.11771](https://arxiv.org/abs/2010.11771)
- Chevallier, A., Power, S., Wang, A.Q., Fearnhead, P.: PDMP Monte Carlo methods for piecewise-smooth densities. (2021). [arXiv preprint arXiv:2111.05859](https://arxiv.org/abs/2111.05859)
- Fearnhead, P., Bierkens, J., Pollock, M., Roberts, G.O.: Piecewise deterministic Markov processes for continuous-time Monte Carlo. *Stat. Sci.* **33**(3), 386–412 (2018). <https://doi.org/10.1214/18-STS648>
- Gilks, W.R., Thomas, A., Spiegelhalter, D.J.: A language and program for complex Bayesian modelling. *J. Royal Stat. Soc. Ser. D (The Statistician)* **43**(1), 169–177 (1994). <https://doi.org/10.2307/2348941>
- Griewank, A., Walther, A.: Evaluating derivatives: principles and techniques of algorithmic differentiation. SIAM, New Delhi (2008)
- Health Data Insight CiC (n.d.). Simulacrum. Artificial patient-like cancer data to help researchers gain insight. Retrieved 01-12-2022, from <https://simulacrum.healthdatainsight.org.uk>
- Hoffman, M.D., Gelman, A., et al.: The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**(1), 1593–1623 (2014)
- Jackson, C.H.: Flexsurv: a platform for parametric survival modeling in R. *J. Stat. Softw.* (2016). <https://doi.org/10.18637/jss.v070.i08>
- Kiefer, J.: Sequential minimax search for a maximum. *Proceedings of the American mathematical society* **4**(3), 502–506 (1953)
- Koskela, J.: Zig-Zag sampling for discrete structures and non-reversible phylogenetic MCMC. *J. Comput. Gr. Stat. (just-accepted)*, 1–000 (2022). <https://doi.org/10.1080/10618600.2022.2032722>
- Lewis, P.W., Shedler, G.S.: Simulation of nonhomogeneous Poisson processes by thinning. *Naval Res. Logist. Quart.* **26**(3), 403–413 (1979). <https://doi.org/10.1002/nav.3800260304>
- Neal, R.M., et al.: MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, **2**(11), 2 (2011). ([arXiv:1206.1901](https://arxiv.org/abs/1206.1901))
- Pagani, F., Chevallier, A., Power, S., House, T., Cotter, S.: NuZZ: numerical Zig-Zag sampling for general models. (2022). [arXiv preprint arXiv:2003.03636](https://arxiv.org/abs/2003.03636)
- Plummer, M., et al.: JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing **124**, 1–10 (2003)
- Roberts, G.O., Rosenthal, J.S.: Optimal scaling of discrete approximations to Langevin diffusions. *J. Royal Stat. Soc. Ser. B (Statistical Methodology)* **60**(1), 255–268 (1998). <https://doi.org/10.1111/1467-9868.00123>
- Roberts, G.O., Tweedie, R.L.: Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**, 341–363 (1996). <https://doi.org/10.2307/3318418>
- Sherlock, C., Urbas, S., Ludkin, M.: Apogee to Apogee Path Sampler. (2021). [arXiv preprint arXiv:2112.08187](https://arxiv.org/abs/2112.08187)
- Vasdekis, G., Roberts, G.O.: Speed Up Zig-Zag. (2021). [arXiv preprint arXiv:2103.16620](https://arxiv.org/abs/2103.16620)
- Vetterling, W.T., Press, W.H., Teukolsky, S.A., Flannery, B.P.: Numerical recipes: example book C (The Art of Scientific Computing). Press Syndicate of the University of Cambridge. (Section 10.3: Parabolic Interpolation and Brent’s Method in One Dimension) (1992)
- Wu, C., Robert, C.P.: Coordinate sampler: a non-reversible Gibbs-like MCMC sampler. *Statistics and Computing* **30**(3), 721–730 (2020). <https://doi.org/10.1007/s11222-019-09913-w>

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.