Adaptive schemes for piecewise deterministic Monte Carlo algorithms

Bertazzi, Andrea; Bierkens, Joris

DOI
10.3150/21-BEJ1423

Publication date
2022

Document Version
Final published version

Published in
Bernoulli

Bernoulli **28**(4), 2022, 2404–2430
https://doi.org/10.3150/21-BEJ1423

Adaptive schemes for piecewise deterministic Monte Carlo algorithms

ANDREA BERTAZZI[a] and JORIS BIERKENS[b]

Delft Institute of Applied Mathematics, TU Delft, Mekelweg 4, 2628 CD Delft, The Netherlands.
[a]*a.bertazzi@tudelft.nl,* [b]*joris.bierkens@tudelft.nl*

The Bouncy Particle sampler (BPS) and the Zig-Zag sampler (ZZS) are continuous time, non-reversible Monte Carlo methods based on piecewise deterministic Markov processes. Experiments show that the speed of convergence of these samplers can be affected by the shape of the target distribution, as for instance in the case of anisotropic targets. We propose an adaptive scheme that iteratively learns all or part of the covariance matrix of the target and takes advantage of the obtained information to modify the underlying process with the aim of increasing the speed of convergence. Moreover, we define an adaptive scheme that automatically tunes the refreshment rate of the BPS or ZZS. We prove ergodicity and a law of large numbers for all the proposed adaptive algorithms. Finally, we show the benefits of the adaptive samplers with several numerical simulations.

Keywords: Adaptive Markov process Monte Carlo; piecewise deterministic Markov processes; bouncy particle sampler; zig-zag sampler; ergodicity

1. Introduction

Piecewise Deterministic Markov processes (PDMP) have been recently used for Monte Carlo sampling as a continuous time alternative to Markov chain Monte Carlo (MCMC) methods. The Bouncy Particle sampler (BPS) and the Zig-Zag sampler (ZZS), introduced in [12] and [8] respectively, are primary examples of this new class of methods, with notable predecessors [24,25,27]. Both samplers are based on continuous time dynamics defined by piecewise linear, deterministic trajectories and changes in the velocity of the process at random times. The simplicity of these dynamics is such that the underlying processes can be simulated exactly in many relevant scenarios. A key aspect of these samplers is the non-reversibility of the underlying Markov process. This property has been observed to result in a lower asymptotic variance of Monte Carlo estimates for moments of the target density [2,7]. Moreover, these algorithms can be naturally modified to achieve asymptotically exact subsampling. In the Bayesian setting, this means that Piecewise Deterministic Monte Carlo (PDMC) algorithms need to access only a small portion of the data set at every iteration without introducing a bias. The reader is referred to [18,33] for an in depth description of the general methodology of PDMC algorithms.

Several papers have studied the convergence properties of the BPS and the ZZS in recent years. It was first observed in [12] that the BPS can fail to be ergodic unless a refreshment of the velocity vector is performed at random times. Exponential ergodicity of the BPS was proved in [16,17] for different distributions from which refreshments of the velocity vector can be drawn. Similarly, in [11] the ZZS was shown to converge exponentially to its invariant distribution under reasonable assumptions and without the need of velocity refreshments. Exponential convergence in the L^2 sense was established for both samplers in [1] using the hypocoercivity framework, and recently, using Poincaré inequalities in space-time, in [23]. A study of the scaling limits was conducted in [10], giving also a criterion to choose the refreshment rate of BPS. The asymptotic variance of these processes has been studied in e.g. [2,7]. It is also possible to design PDMC algorithms with non-linear trajectories, see e.g. [9,33].

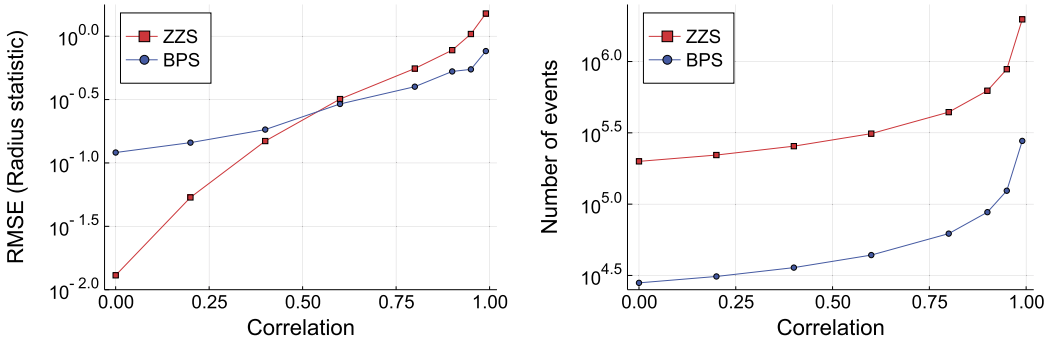


Figure 1. Root mean squared error (RMSE) for the radius statistic (left) and number of events (right) for 50-dimensional Gaussian targets for several values of the correlation between all components. The continuous time horizon is $T = 10^4$.

Although PDMC sampling methods offer some important benefits as mentioned above, computation remains expensive, which requires us to investigate possible performance improvements. In particular, a strong performance degradation is observed when the target distribution π is anisotropic. Figure 1 illustrates this phenomenon in the case of Gaussian targets as a function of the correlation between all components. The performance drop occurs due to a combination of decreasing accuracy of the estimates and increasing computational complexity of the algorithms, which is implied by the growing number of velocity change events. Our idea to improve this issue is to let the process learn (part of) the covariance matrix Σ_π and take advantage of it to enhance the mixing properties. The covariance estimate is used to linearly transform the target in such a way that it becomes more isotropic, i.e. with unitary covariance matrix. The standard samplers are then run targeting the transformed version of π, and the obtained sample is finally re-transformed to be approximately from π. The procedure is applied iteratively, and once a new estimate of Σ_π is computed, it is used to define the linear transformation of π. The estimate will eventually be close to the true covariance matrix and the process targets an isotropic version of π. This scheme can also be interpreted as an application of a linear transformation directly to the standard ZZS and BPS. The natural applications of this procedure are then targets with elliptical level curves, although performance improvements can be observed also for distributions that deviate from this class.

Furthermore, we address the problem of automatically choosing the rate of velocity refreshments. In [10] the authors consider a BPS with a specific target and derive that in the limit it is optimal to have a ratio of number of refreshments over number of total events of 0.7812. In this paper we use this criterion as a basis to define an adaptive algorithm that iteratively adjusts the refreshment rate to obtain the right ratio. This same adaptive scheme can be applied to the ZZS. Indeed it could be the case that adding velocity refreshments to the ZZ process leads to a faster convergence to the invariant measure, although it comes with a larger asymptotic variance. For an analysis of these two results we refer to [34] and [7] respectively.

Both the schemes we discussed take advantage of what the process has learned up until the current time to tune a parameter or to improve the performance. This idea is at the core of adaptive Markov chain Monte Carlo algorithms. For an introduction to this area we refer to [3,30], while standard results on convergence of these methods can be found among others in [4,19,20,29]. It is well known that adaptive MCMC algorithms can lose the right invariant measure if not applied with care (see for instance [29] and the examples therein). Therefore we study in depth the convergence properties of the proposed algorithms. To fit into the existing adaptive MCMC literature we let the adaptation happen at

fixed points in time. The main challenge consists of establishing a simultaneous geometric drift condition for a family of BPS's (see Lemma 4.5) and a simultaneous small set condition for a family of ZZ processes (see Lemma 4.1). The former result is obtained taking advantage of the Lyapunov function found in [17], while the latter is proved by extending on the one-dimensional case. The ergodicity and a law of large numbers for the proposed adaptive PDMC algorithms are then established in Theorem 3.8 and Theorem 3.10.

In Section 2 we introduce the adaptive schemes, while in Section 3 the theoretical aspects of the algorithms are studied. The skeleton of the proofs of the two main theorems can be found in Section 4, while all other proofs can be found in the supplement [6]. In Section 5 the adaptive BPS and ZZS are tested empirically on various Gaussian targets, on a Bayesian logistic regression problem with correlated data, and on a mixture of two Gaussian distributions. The details on the implementation of adaptive PDMC algorithms (with and without subsampling), as well as an alternative adaptive scheme for the refreshment rate, can be found in Appendix A in the supplement [6].

2. The adaptive schemes

We are interested in building adaptive strategies to make the ZZS and the BPS choose the refreshment rate themselves and/or converge faster to the target density. We begin with an introduction of the standard versions of both samplers, followed by a characterisation of the preconditioned processes in Section 2.2 and a discussion on the choice of the transformation matrix in Section 2.3. Finally, the adaptive algorithms are defined in Section 2.4.

2.1. The standard ZZS and BPS

Let the target density π be defined on $\mathcal{X} \subset \mathbb{R}^d$ as

$$\pi(\xi) = \frac{1}{Z} \exp\left(-U(\xi)\right),$$

where $U(\xi)$ is called potential or energy function, and $\xi \in \mathcal{X}$. Let us now define the standard ZZS with invariant measure π. Throughout the paper the position and velocity vectors at time t of the standard processes, both ZZS and BPS, are denoted respectively as $\Xi(t)$ and $\Theta(t)$. We distinguish between the Zig-Zag Sampler (i.e. a PDMC algorithm) and the Zig-Zag Process (i.e. the Markov process on which the algorithm is based). The Zig-Zag (ZZ) process is a Markov process $(\Xi(t), \Theta(t))_{t \geq 0}$ with state space $E = \mathbb{R}^d \times \{-1, +1\}^d$ that follows a linear trajectory in the position space with velocity Θ until one of its d-inhomogeneous Poisson clocks rings. When the i-th clock rings the velocity of the i-th coordinate switches sign. In mathematical terms, this means that the new velocity vector is $F_i\Theta$, where $F_i : \{-1, +1\}^d \to \{-1, +1\}^d$ is the flip operator defined as

$$(F_i\theta)_j = \begin{cases} \theta_j & \text{if } j \neq i, \\ -\theta_j & \text{if } j = i. \end{cases}$$

The rates of the Poisson clocks are defined for $i = 1, \ldots, d$ as

$$\lambda_i(\xi, \theta) = (\theta_i \partial_i U(\xi))_+ + \gamma_i(\xi, \theta),$$

where $(\theta_i \partial_i U(\xi))_+ = \max(0, \theta_i \partial_i U(\xi))$ and $\gamma_i(\xi, \theta) : E \to \mathbb{R}_+$ is called excess switching rate and is such that $\gamma_i(\xi, \theta) = \gamma_i(\xi, F_i\theta)$. It was shown in [8, Theorem 2.2] that this choice of switching rates ensures

that $\mu = \pi \times \text{Unif}(\{-1,+1\}^d)$ is the stationary distribution of the process. Moreover, the ZZ process is characterised by its infinitesimal generator

$$\mathcal{L}f(\xi,\theta) = \langle \theta, \nabla f(\xi,\theta)\rangle + \sum_{i=1}^{d} \lambda_i(\xi,\theta)(f(\xi,F_i\theta) - f(\xi,\theta)), \tag{1}$$

where function f should be in the domain of the generator $\mathcal{D}(\mathcal{L})$. The linear trajectories are represented in the first term, while the second term represents the event of a velocity flip.

Similarly, we denote the BPS as $(\Xi(t),\Theta(t))_{t\geq 0}$, but in this case the state space is $E = \mathbb{R}^d \times \mathbb{R}^d$ and so $\Theta(t) \in \mathbb{R}^d$. In contrast to the ZZS, the BPS has two Poisson clocks. The first one depends on the gradient of the energy function and has inhomogeneous rate $\lambda(\xi,\theta) = \langle \theta, \nabla U(\xi)\rangle_+ = \max(0, \langle \theta, \nabla U(\xi)\rangle)$. At event time the particle is reflected on the level curve of the potential U following an elastic bounce, and thus preserving the norm of the velocity vector. After a bounce the new velocity vector is given by

$$R(\xi)\theta = \theta - 2\frac{\langle \nabla U(\xi), \theta\rangle}{||\nabla U(\xi)||^2}\nabla U(\xi).$$

It was observed in [12] that the BPS needs refreshments of the velocity vector in order to be ergodic. This brings us to the second Poisson clock, which has rate $\lambda_r : E \to \mathbb{R}_+$. This is referred to as refreshment rate and should thus be strictly positive. When this clock rings, the velocity vector is refreshed by sampling from a distribution ψ. Possible choices are the Gaussian distribution $\psi = \mathcal{N}(0, \mathbb{1}_d)$, or $\psi = \text{Unif}(\mathbb{S}^{d-1})$, where $\mathbb{S}^{d-1}$ is the surface of the unit hypersphere. In the analysis that follows we focus on the former distribution. The infinitesimal generator of the BPS is defined for any $f \in \mathcal{D}(\mathcal{L})$ as

$$\begin{aligned}\mathcal{L}f(\xi,\theta) = \langle \theta, \nabla f(\xi,\theta)\rangle + \lambda(\xi,\theta)\big(f(\xi,R(\xi)\theta) - f(\xi,\theta)\big)\\ + \lambda_r(\xi,\theta)\int (f(\xi,\theta') - f(\xi,\theta))\psi(d\theta').\end{aligned}$$

The invariant measure of the BPS defined by the infinitesimal generator above is $\mu = \pi \times \psi$ as shown in [12, Proposition 1].

2.2. Applying a linear transformation to the ZZS and BPS

In this section we suppose a matrix $M \in \mathbb{R}^{d\times d}$ is given. We then wish to define a transformation scheme encoded by M, which we should think of being such that, for a suitable choice of M, it gives a "more isotropic" version of the target, and analyse its effects on the PDMC samplers.

The transformation scheme encoded by M consists of a linear transformation of the state space, which defines a new target distribution $\tilde{\pi}_M$ given by

$$\tilde{\pi}_M(\xi) := \frac{1}{\tilde{Z}_M}\exp(-\tilde{U}_M(\xi)),$$

with $\tilde{U}_M(\xi) = U(M\xi)$ and $\tilde{Z}_M = Z/|\det M|$. The idea is to apply the transformation to the target distribution π and simulate the standard PDMC sampler $(\Xi_t,\Theta_t)_{t\geq 0}$ with the resulting target $\tilde{\pi}_M$. Then the last thing to do is transform the obtained sample, which is approximately from $\tilde{\pi}_M$, by applying the inverse transformation. For this reason it is important that the matrix M is invertible, and thus that we can go from one state space to the other. This procedure is illustrated in Figure 2. An equivalent option

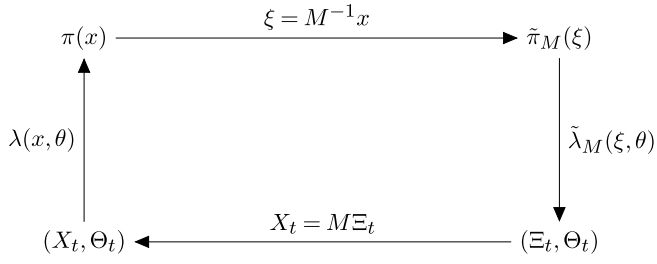


Figure 2. Transformation scheme.

is to simulate directly the process $(X_t,\Theta_t)_{t\geq 0}$ that results from the scheme in Figure 2. We will conveniently alternate between these two formulations when studying the ergodic properties of the samplers, while we will use the latter formulation for our experiments. The dynamics of process $(X_t,\Theta_t)_{t\geq 0}$ are studied in the remainder of this section.

Let us first focus on the case in which $(\Xi_t,\Theta_t)_{t\geq 0}$ is a standard ZZS with excess switching rate γ. In this case the switching rates are $\tilde{\lambda}_{M,i}(\xi,\theta) = (\theta_i\partial_i\tilde{U}_M(\xi))_+ + \gamma_i(\xi,\theta)$ for $i = 1,\dots,d$. Note that, unless stated otherwise, we will always use a tilde to indicate quantities related to the standard PDMC samplers with transformed target. In the next proposition we find the generator of the preconditioned ZZS. For a characterisation of the domain of the extended generator we refer to [15, Theorem 26.14].

Proposition 2.1. *Let $M \in \mathbb{R}^{d\times d}$ be an invertible matrix. Let $(\Xi(t),\Theta(t))_{t\geq 0}$ be a standard ZZS with target $\tilde{\pi}_M$ and excess switching rates $\gamma : E \to \mathbb{R}^d_+$. The process $(X(t),\Theta(t))_{t\geq 0} = (M\Xi(t),\Theta(t))_{t\geq 0}$ has extended generator $(\mathcal{L}_M,\mathcal{D}(\mathcal{L}_M))$ where for any $h \in \mathcal{D}(\mathcal{L}_M)$*

$$\mathcal{L}_M\, h(x,\theta) = \langle M\theta, \nabla_x h(x,\theta)\rangle + \sum_{i=1}^{d} \lambda_i^M(x,\theta)(h(x,F_i\theta) - h(x,\theta)), \tag{2}$$

in which for $i = 1,\dots,d$

$$\lambda_i^M(x,\theta) = \tilde{\lambda}_{M,i}(M^{-1}x,\theta) = (\theta_i\langle M_i, \nabla U(x)\rangle)_+ + \gamma_i(M^{-1}x,\theta), \tag{3}$$

where M_i denotes the i-th column of M.

Proof. All the proofs of this section can be found in Appendix B in the supplement [6]. □

Proposition 2.1 shows that the transformed process is again a PDMP with linear trajectories between jumps. The transformation affects the velocity of the process, which is now $v = M\theta$, and the switching intensities, which are as defined in (3). In particular, the available velocities for a fixed transformation M are in the following set:

$$V := \left\{v : v = M\theta,\, \theta \in \{-1,+1\}^d\right\}.$$

When a switch of the i-th velocity of the underlying standard ZZ process happens, the velocities of the transformed process change according to the operator $\bar{F}_i v = MF_i\theta = MF_i(M^{-1}v)$ for $i = 1,\dots,d$. Therefore, all components of the velocity are possibly affected by any single event. If M is diagonal the behaviour is more similar to the standard ZZ process and in particular we have that $\bar{F}_i \equiv F_i$. In the proposition below we check that $(X(t),\Theta(t))_{t\geq 0}$ targets the correct density function.

Proposition 2.2. *Consider the same setting of Proposition 2.1. Then, for any invertible $M \in \mathbb{R}^{d\times d}$, the modified ZZ process $(X(t),\Theta(t))_{t\geq 0}$ has invariant distribution $\mu = \pi \times \mathrm{Unif}(\{-1,+1\}^d)$.*

Let us now apply the same transformation scheme shown in Figure 2 to the BPS. In this case the switching rate of the standard BPS with target $\tilde{\pi}_M$ is $\tilde{\lambda}_M(\xi,\theta) = \langle \theta, \nabla \tilde{U}_M(\xi)\rangle_+$, while reflections on the level curves of $\tilde{U}_M$ are obtained by applying operator $\tilde{R}_M$. The following result, analogous to Proposition 2.1, studies the transformed process.

Proposition 2.3. *Let $M \in \mathbb{R}^{d\times d}$ be an invertible matrix. Let $(\Xi(t),\Theta(t))_{t\geq 0}$ be a standard BPS with target $\tilde{\pi}_M$ and refreshment rate $\lambda_r : E \to \mathbb{R}_+$. The process $(X(t),\Theta(t))_{t\geq 0} = (M\Xi(t),\Theta(t))_{t\geq 0}$ has extended generator $(\mathcal{L}_M, \mathcal{D}(\mathcal{L}_M))$ where for any $h \in \mathcal{D}(\mathcal{L}_M)$*

$$\begin{aligned}\mathcal{L}_M h(x,\theta) = \langle M\theta, \nabla_x h(x,\theta)\rangle + \lambda_M(x,\theta)\big(h(x,R_M(x)\theta) - h(x,\theta)\big) \\ + \lambda_r(M^{-1}x,\theta)\int (h(x,\theta') - h(x,\theta))\psi(d\theta'),\end{aligned} \tag{4}$$

where we defined

$$\lambda_M(x,\theta) = \tilde{\lambda}(M^{-1}x,\theta) = \langle M\theta, \nabla U(x)\rangle_+ \tag{5}$$

and

$$R_M(x)\theta = \tilde{R}_M(M^{-1}x)\theta = \theta - 2\frac{\langle M^T\nabla U(x),\theta\rangle}{||M^T\nabla U(x)||^2}M^T\nabla U(x). \tag{6}$$

Once again the true velocity of the process is $v = M\theta$. When a velocity refreshment takes place the new θ is sampled from $\psi = \mathcal{N}(0_d, \mathbb{1}_d)$, while the new velocity v is $v = M\theta \sim \mathcal{N}(0_d, MM^T)$. Observe also that the reflection rate is $\lambda_M(x,\theta) = \langle v, \nabla U(x)\rangle_+$ and thus preserves the same structure as in the standard BPS. It follows that the complexity of the simulation of event times remains unchanged. Finally consider the reflection operator in (6). This corresponds to a reflection in the opposite direction to the gradient in the transformed space, i.e. $\nabla_\xi \tilde{U}(\xi) = M^T\nabla U(x)$. After the bounce the process moves with velocity

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

This implies that $\langle v, \nabla U(x)\rangle = \langle M(R_M(x)\theta), \nabla U(x)\rangle = -\langle M\theta, \nabla U(x)\rangle = -\langle v, \nabla U(x)\rangle$.

Proposition 2.4. *Consider the same setting of Proposition 2.3. Then, for any invertible $M \in \mathbb{R}^{d\times d}$, the transformed BPS $(X(t),\Theta(t))_{t\geq 0}$ has invariant distribution $\mu = \pi \times \psi$.*

2.3. Choosing the transformation matrix

As explained above, we wish to transform the target to mitigate its anisotropies. To this end, some alternative choices of the transformation matrix M are the following:

a) $M = \sqrt{\mathrm{Cov}_\pi(X)}$: this transformation is such that the target $\tilde{\pi}_M$ has unitary covariance matrix. The downside of this choice is the additional $O(d^3)$ computations that are introduced by the calculation of the square root of the covariance;

b) M is a rotation matrix ($\det M = 1$) such that the transformed density has a certain angle. Although an interesting case, it is not clear whether there is an optimal angle that speeds up the convergence;
c) M is the diagonal matrix with $M_{ii} = \sqrt{\mathrm{Var}_\pi(X_i)}$ for any $1 \le i \le d$. This choice introduces $O(d)$ computations due to the square root of the variances, which is a negligible additional computational burden. However, correlations in the target are not picked up and only a rescaling of the axes is performed. The main advantage of this choice is that the scenario in which some components are explored quickly and others slowly is avoided.

Both the first and the third option can potentially change the expected number of switching events, and this could be an inconvenience in certain settings. It is not difficult to modify these transformations in such a way that the expected switching rate is enforced to be close to that of the original standard PDMC algorithm. For example, consider the transformed BPS with generator as in Proposition 2.3. Then in stationarity we have

$$\mathbb{E}_\mu(\langle M\Theta, \nabla U(X)\rangle_+) = \mathbb{E}_\pi \|M^T \nabla U(X)\|_2 \le \|M\|_2 \, \mathbb{E}_\pi \|\nabla U(X)\|_2.$$

Since the standard case corresponds to $M = \mathbb{I}_d$, we can normalise any M by dividing it by its Frobenius norm. Then the upper bound is the same for all such choices of M and the expected switching intensity will be close to the standard case. This does not make a difference from a computational point of view as it just amounts to reparametrization of the time parameter, but prevents unpredictable behaviour of the algorithm.

Naturally the options above are not available in practice as the covariance matrix is unknown. It is the goal of the next section to propose an adaptive scheme that overcomes this issue.

2.4. Adaptive PDMC algorithms

In the previous sections we defined the transformation scheme and we discussed the effect it has on the underlying process, together with different choices of the preconditioning matrix. We now describe how this idea can be applied in practice by designing an adaptive PDMC algorithm. Our general strategy is to simulate the process in continuous time and store the states of the process at discrete times. Then at predefined times the stored states are used to update the adaptation parameters. In addition to the adaptive preconditioner, we incorporate an adaptation of the refreshment rate, which makes its choice automatic.

Let us then define a family of Markov semigroups by $\mathcal{P} := \{(P^t_\Gamma)_{t\ge 0} : \Gamma \in \mathcal{Y}\}$, in which Γ is the adaptation parameter, $\mathcal{Y}$ is a compact space, and $(P^t_\Gamma)_{t\ge 0}$ is the semigroup of a modified ZZS or BPS. The modification is given by the adaptation parameter, which is then $\Gamma = (M, \lambda_r)$ for BPS and $\Gamma = (M, \gamma)$ for ZZS. Thus $\mathcal{Y}$ is a suitable compact space of preconditioners and refreshment rates/excess switching rates. Naturally, it is also possible to choose $\Gamma = M$ or $\Gamma = \lambda_r$ only. Now that we have defined a family of Markov processes, we define a rule that establishes how to choose a $P_\Gamma \in \mathcal{P}$ at every iteration. Let us begin by introducing a discretisation step Δt, which defines a discretisation of the time variable. At each time step $n \in \mathbb{N}$, which corresponds to continuous time $t = n\Delta t$, the adaptive scheme can update the parameter Γ_n based on the new information available, that is the new state of the process (X_n, Θ_n). This defines a random sequence $\{\Gamma_n\}_{n\ge 0}$. Once Γ_n is computed, the next state of the process is given by $(X_{n+1}, \Theta_{n+1}) \sim P^{\Delta t}_{\Gamma_n}((X_n, \Theta_n), \cdot)$. Then one updates the parameter, obtains the next state of the process, and so on.

The definition above defines the core ideas, which are written in pseudo-code form in Algorithm 1. A few issues remain to be clarified. A first question is how to simulate the PDMP semigroup of either

Algorithm 1 Adaptive PDMC sampler

Input: family of kernels $\mathcal{P} = \{P_\Gamma : \Gamma \in \mathcal{Y}\}$
Input: initial condition $(x,\theta) \in E$, $\Gamma_0 \in \mathcal{Y}$, set B, Δt, $\{p_n\}_{n\geq 0}$, number of steps N
Output: sequence of discrete samples $\{X_n, \Theta_n\}_{n=0}^N$
Initialise $n = 0$, $(X_0, \Theta_0) = (x,\theta)$, $Q_0 = \Gamma_0$
while $n \leq N$ **do**
    $(X_{n+1}, \Theta_{n+1}) \sim P_{\Gamma_n}^{\Delta t}((X_n, \Theta_n), \cdot)$
    $Q_{n+1} =$ **update**$(Q_n, (X_{n+1}, \Theta_{n+1}))$
    if $(X_{n+1}, \Theta_{n+1}) \in B$ **then**
        With probability p_n, set $\Gamma_{n+1} = Q_{n+1}$
        With probability $1 - p_n$, set $\Gamma_{n+1} = \Gamma_n$
    end if
    $n = n + 1$
end while

ZZS or BPS. Details on how the processes can be simulated in the case of a target with dominated Hessian can be found in Appendix A.1 in the supplement [6]. In a big data setting (large number of observations, moderate dimensionality of the problem) it can be beneficial to take advantage of subsampling techniques that can be implemented with PDMC algorithms. In Appendix A.2 in [6] details can be found on how to make use of subsampling in the context of the adaptive schemes here discussed. For further information on the general implementation of ZZS and BPS we refer to [8,12]. A second aspect of Algorithm 1 we focus on is the introduction of set B, and thus of the auxiliary sequence of random variables $(Q_n)_{n\geq 0}$. The idea is to update the adaptation parameter Γ_n only if $(X_n, \Theta_n) \in B$. This is useful from a theoretical point of view as it ensures that the process remains bounded in probability. Note that set B is defined by the user and can be chosen large. The auxiliary variable Q_n is updated even if the process is outside of B and then, as soon as the process enters B, or if it was already in B, we let $\Gamma_n = Q_n$. A third characteristic of Algorithm 1 is the sequence $\{p_n\}_{n\geq 0}$. This is a sequence for which $p_n \in [0,1]$ for all $n \in \mathbb{N}$ and $p_n \to 0$ as $n \to \infty$. The meaning is that at time step n we update the parameter Γ_n with probability p_n (assuming $(X_n, \Theta_n) \in B$), and with remaining probability $(1 - p_n)$ we set $\Gamma_n = \Gamma_{n-1}$. This choice is helpful when proving ergodicity of the adaptive scheme, as it enforces that the quantity of adaptation diminishes and eventually vanishes.

The function **update**$(Q_n, (X_{n+1}, \Theta_{n+1}))$ outputs the updated parameter given the new observation (X_{n+1}, Θ_{n+1}). As suggested in [3], the estimation of the covariance matrix can be done sequentially, or online, by applying

$$\begin{aligned} \hat{\mu}_{n+1} &= \hat{\mu}_n + r_{n+1}(X_{n+1} - \hat{\mu}_n), \\ \hat{\Sigma}_{n+1} &= \hat{\Sigma}_n + r_{n+1}((X_{n+1} - \hat{\mu}_n)(X_{n+1} - \hat{\mu}_n)^T - \hat{\Sigma}_n). \end{aligned} \tag{7}$$

Here $\{r_n\}_{n\geq 0}$ is a positive, decreasing sequence such that $r_n \to 0$ as $n \to \infty$. In our simulations we choose $r_n = 1/n$. Equation (7) is then used to define M_{n+1} such that $\hat{\Sigma}_{n+1} = M_{n+1}^T M_{n+1}$. The same principle can be used if one is not interested in estimating the full covariance matrix, but only the diagonal, or more generally only a subset of it. More advanced estimation techniques can be employed to preserve any existing conditional independence structure in the target, as discussed in [35]. We remark that $\hat{\Sigma}_{n+1}$ needs to be positive definite in order for M_{n+1} to be invertible as required. This property is achieved by choosing $\hat{\Sigma}_0 = \mathbf{1}_{d\times d}$, i.e. the identity matrix, and then observing that the second equation in (7) can be reformulated as

$$\hat{\Sigma}_{n+1} = (1 - r_{n+1})\hat{\Sigma}_n + r_{n+1}(X_{n+1} - \hat{\mu}_n)(X_{n+1} - \hat{\mu}_n)^T.$$

Indeed $(X_{n+1} - \hat{\mu}_n)(X_{n+1} - \hat{\mu}_n)^T$ is non-negative definite and by induction $\hat{\Sigma}_n$ is positive definite, and therefore $\hat{\Sigma}_{n+1}$ is itself positive definite. Moreover, in Section 3 we will see that to show ergodicity of the adaptive algorithms it is required that M_n lies in a compact space of positive definite matrices. Observe that positive definiteness follows from the fact that M_n is the square root of a positive definite matrix, while it is sufficient to set bounds on the norm of M_n in order to force it to be in a compact space. In particular we can impose that M_n is not updated if the norm of the new estimate is outside of a user chosen interval $[M_{\min}, M_{\max}]$. As $M_{\min}$ and $M_{\max}$ can be chosen arbitrarily small and large respectively, this condition is not restrictive in practice, although the choice of the cut-off value may influence convergence properties of the algorithm. Then refreshment rate of the BPS is assumed to be constant and is updated iteratively as follows. At time step n, n_{refl} reflections took place and thus we estimate the average reflection rate as $\overline{\lambda}_{\text{refl}}(n) = n_{\text{refl}}/(n\Delta t)$. Therefore, using the optimality criterion in [10] we have

$$\frac{\lambda_r^n}{\lambda_r^n + \overline{\lambda}_{\text{refl}}(n)} = 0.7812 \implies \lambda_r^n = \frac{0.7812}{0.2188}\,\overline{\lambda}_{\text{refl}}(n). \tag{8}$$

An alternative adaptive strategy for the refreshment can be found in Appendix A.3 in [6]. The scheme above can be applied to the excess switching rate of the ZZS. Although the analysis in [7] suggests that the best choice in terms of asymptotic variance is $\gamma \equiv 0$, adding some diffusivity could speed up the convergence to the target measure. In practice the user can select the wanted ratio of velocity switches over total events and proceed as above. However, a criterion to choose this ratio is currently unavailable for ZZS, and thus in this paper we limit ourselves to a theoretical study of this option.

Finally, we remark that in practice it is not reasonable to update the parameters at every iteration. The main reason for this is the computational cost of such an operation. In the most general case, the task of learning all components of Σ takes $O(d^2)$ operations, while the computation of its square root, which is needed to obtain the transformation matrix M, is an $O(d^3)$ operation. Therefore it is rather inconvenient to perform this at every time step. To avoid this issue it is sufficient to define the adaptive scheme such that adaptations happen every n_{adap} time steps, where n_{adap} is a user-defined integer. A possible choice is for instance $n_{\text{adap}} = 1000$. This modification is beneficial also because it allows the process to explore the target distribution before updating the parameters. Similarly, it is reasonable to update the refreshment rate based on the previous n_{adap} time steps, as in the long term this allows to stabilise around the wanted ratio. The covariance matrix can be updated as in (7) by simply processing the entire batch of n_{adap} data points one at a time.

3. Theoretical results

In the context of adaptive MCMC algorithms, convergence to the target density is usually proved with simultaneous drift conditions and small set conditions. In Section 3.1 we introduce the notation and the main existing theorems we make use of, and we extend these results to more general conditions in Theorem 3.5. In Section 3.2 we state Theorem 3.6, which shows ergodicity for an adaptive MCMC algorithm based on a continuous time process. In this result, the assumptions are formulated directly in continuous time. Finally, Theorems 3.8 and 3.10 in Section 3.3 show that the adaptive ZZS and the adaptive BPS discussed in Section 2.4 are ergodic and satisfy a weak law of large numbers under reasonable growth conditions on the potential.

3.1. Theory of adaptive MCMC

We denote the parameter that specifies the kernel as $\Gamma \in \mathcal{Y}$. At time step n a $\mathcal{Y}$-valued random variable Γ_n determines which transition kernel will be used to move to the next step. From here on each Markov

transition kernel P_Γ is assumed to define a Markov chain that has μ as stationary measure, and moreover it is aperiodic and irreducible. An adaptive MCMC algorithm is then said to be *ergodic* if

$$\lim_{n\to\infty} \|P(Z_n \in \cdot\, | z_0, \Gamma_0) - \mu(\cdot)\|_{\mathrm{TV}} = 0 \qquad \text{for all } z_0 \in E, \Gamma_0 \in \mathcal{Y}, \tag{9}$$

where $||\cdot||_{\mathrm{TV}}$ is the total variation distance, i.e. $\|\mu - \nu\|_{\mathrm{TV}} = \sup_{A\subset E} |\mu(A) - \nu(A)|$. A crucial quantity turns out to be the ε-convergence time function $M_\varepsilon : E \times \mathcal{Y} \to \mathbb{N}$, defined as

$$M_\varepsilon(z, \Gamma) = \inf\{n \geq 1 : \|P^n_\Gamma(z, \cdot) - \mu(\cdot)\|_{\mathrm{TV}} \leq \varepsilon\}.$$

The next theorem, proved in [29], is arguably the most important result for establishing ergodicity of adaptive MCMC methods.

Theorem 3.1 (Theorem 2 in [29]). *Consider an adaptive MCMC algorithm on a state space E with adaption parameter in a space $\mathcal{Y}$. Let μ be stationary for P_Γ for each $\Gamma \in \mathcal{Y}$. The adaptive algorithm is ergodic if the two following conditions hold:*

(a) *(Containment condition) For all $z_0 \in E, \Gamma_0 \in \mathcal{Y}$, $\varepsilon > 0$ the sequence $\{M_\varepsilon(Z_n, \Gamma_n)\}_{n=0}^\infty$ is bounded in probability given z_0, Γ_0;*
(b) *(Diminishing adaptation) The following limit holds in probability:*

$$\lim_{n\to\infty} \left(\sup_{z\in E} \|P_{\Gamma_{n+1}}(z, \cdot) - P_{\Gamma_n}(z, \cdot)\|_{\mathrm{TV}} \right) = 0. \tag{10}$$

The boundedness of $\{M_\varepsilon(Z_n, \Gamma_n)\}_{n=0}^\infty$ can be rephrased as for all $z_0 \in E, \Gamma_0 \in \mathcal{Y}$, $\delta > 0$, there exists $N \in \mathbb{N}$ such that $P(M_\varepsilon(Z_n, \Gamma_n) \leq N | z_0, \Gamma_0) \geq 1 - \delta$, for all $n \in \mathbb{N}$.

We are then interested in sufficient conditions that imply containment. A first case is the following, and was studied in [4].

Assumption 3.2 ([4]). The family $\{P_\Gamma : \Gamma \in \mathcal{Y}\}$ is *simultaneously geometrically ergodic* (SGE), i.e. there are $C \in \mathcal{B}(E)$, some integer $n_0 \geq 1$, a function $V : E \to [1, \infty)$, $\delta > 0$, $0 < \lambda < 1$, and $b < \infty$, such that $\sup_{z\in C} V(z) < \infty$, $\mu(V) < \infty$, and

(a) C is a uniform $(\nu_\Gamma, \delta, n_0)$-small set, i.e. for each Γ, there exists a probability measure $\nu_\Gamma(\cdot)$ on C such that $P^{n_0}_\Gamma(z, \cdot) \geq \delta \nu_\Gamma(\cdot)$ for all $z \in C$;
(b) (*simultaneous geometric drift condition*) $P_\Gamma V \leq \lambda V + b\mathbb{1}_C$ for all $\Gamma \in \mathcal{Y}$.

Then [4, Theorem 3] establishes that an SGE family satisfies the containment condition. In Section 3.3 we use this result to show that containment holds for the adaptive ZZS when the class of preconditioners is restricted to diagonal matrices.

In practice it is often hard to show that the family of Markov kernels is SGE, as it is not trivial to find a Lyapunov function that satisfies the simultaneous geometric drift condition. In [14] the authors introduced a way around this problem, although in a different context, and in [13] this was applied to adaptive MCMC. The fundamental idea is that it is possible to weaken the simultaneous drift condition by allowing adaptations only at time steps n at which the process Z_n is inside of a compact set B. This means that, defining an auxiliary random process $\{Q_n\}_{n\geq 1}$ that contains the current adaptation parameter independently of the position of Z_n, Γ_n is updated as

$$\Gamma_{n+1} = \begin{cases} \Gamma_n & \text{if } Z_{n+1} \notin B, \\ Q_{n+1} & \text{if } Z_{n+1} \in B. \end{cases} \tag{11}$$

This modification avoids unbounded detours of the process by sticking to the same ergodic kernel once the process exits a fixed compact set. The compact set can be chosen arbitrarily large, and therefore in most applications the process will not exit from it.

With this is mind we introduce the following sets of assumptions, which we show in Theorem 3.5 to be sufficient to enforce the containment condition.

Assumption 3.3. Let $\{P_\Gamma : \Gamma \in \mathcal{Y}\}$ be a family of discrete time Markov chains with state space E. There are $C \in \mathcal{B}(E)$, an integer $n_0 \geq 1$, a class of functions $\{V_\Gamma : E \to [1,\infty) : \Gamma \in \mathcal{Y}\}$, $\delta > 0$, $0 < \lambda < 1$, and $b < \infty$, such that $\sup_{z\in C,\, \Gamma\in\mathcal{Y}} V_\Gamma(z) < \infty$, $\mu(V_\Gamma) < \infty$, and

(a) C is a uniform $(\nu_\Gamma, \delta, n_0)$-small set, i.e. for each $\Gamma \in \mathcal{Y}$, there exists a probability measure $\nu_\Gamma(\cdot)$ on C such that $P_\Gamma^{n_0}(z,\cdot) \geq \delta\nu_\Gamma(\cdot)$ for all $z \in C$;
(b) for each $\Gamma \in \mathcal{Y}$, $z \in E$, $P_\Gamma V_\Gamma(z) \leq \lambda V_\Gamma(z) + b\mathbb{1}_C(z)$;
(c) the adaptation parameter is allowed to be updated only if the process is inside of a compact set B, as defined in (11).

Assumption 3.4. Let $\{P_\Gamma : \Gamma \in \mathcal{Y}\}$ be a family of discrete time Markov chains with state space E. There exist $\alpha, \lambda \in (0,1)$, $C_1 > 0$, $C_2 > 2C_1$, a class of functions $\{V_\Gamma : E \to [1,\infty) : \Gamma \in \mathcal{Y}\}$ with $\mu(V_\Gamma) < +\infty$, such that

(a) for each $\Gamma \in \mathcal{Y}$, for all $x, y \in E$ such that $V_\Gamma(x) + V_\Gamma(y) \leq C_2$ it holds that

$$\|P_\Gamma(x,\cdot) - P_\Gamma(y,\cdot)\|_{\mathrm{TV}} \leq 2(1-\alpha);$$

(b) for each $\Gamma \in \mathcal{Y}$ and for any $z \in E$, $P_\Gamma V_\Gamma(z) \leq \lambda V_\Gamma(z) + C_1(1-\lambda)$;
(c) the adaptation parameter is allowed to be updated only if the process is inside of a compact set B, as defined in (11).

Theorem 3.5. *Consider a family of discrete time Markov transition kernels $\{P_\Gamma : \Gamma \in \mathcal{Y}\}$. Assume that all kernels P_Γ are aperiodic, irreducible, and have stationary measure μ. Suppose the adaptive algorithm satisfies the diminishing adaptation condition, i.e. assumption (b) in Theorem 3.1, and let either Assumption 3.3 or Assumption 3.4 hold. Then the containment condition holds and the adaptive MCMC algorithm is ergodic.*

Proof. The proof can be found in Appendix B.2 in the supplement [6]. □

Remark. A weak law of large numbers (WLLN) for bounded and measurable functions follows immediately from containment and diminishing adaptation by Theorem 3.4 in [28]. Therefore under the conditions of Theorem 3.5 a WLLN holds.

3.2. Convergence properties of adaptive MCMC algorithms based on continuous time Markov processes

It could be the case, as it is in the present work, that one is interested in defining an adaptive scheme based on a family of continuous time Markov processes in continuous time. In this case a grid for the time variable needs to be introduced in order to indicate the times at which the adaptation occurs. In fact the adaptive chain only sees the process at times $m\Delta t$, where $\Delta t > 0$ is the step size and $m \in \mathbb{N}$. Although the resulting chain is in discrete time, it is in most cases easier to work directly with the continuous time process. The following result, which is analogous to Theorem 3.5, is helpful in this sense.

Theorem 3.6. *Consider a family of Markov processes with generators $\{\mathcal{L}_\Gamma : \Gamma \in \mathcal{Y}\}$, each being irreducible, aperiodic, and having μ as invariant measure. Consider a grid for the time variable with step Δt. Consider an adaptive scheme that at times $m\Delta t$, with $m \in \mathbb{N}$, chooses a process from the aforementioned family. Furthermore, suppose that the adaptive scheme satisfies the diminishing adaptation condition (10) for $P := P^{\Delta t}$. Finally assume one of the following two sets of conditions holds:*

1. *There exist a set $C \in \mathcal{B}(E)$, $t_0 > 0$, a class of functions $\{V_\Gamma : E \to [1,\infty) : \Gamma \in \mathcal{Y}\}$, $\delta > 0$, $A_1, A_2 > 0$, such that for each Γ, $\sup_{z\in C,\Gamma\in\mathcal{Y}} V_\Gamma(z) < \infty$, $\pi(V_\Gamma) < \infty$, and*
   (a) *for each $\Gamma \in \mathcal{Y}$ there exists a probability measure ν_Γ such that $P_\Gamma^{t_0}(z,\cdot) \geq \delta\nu_\Gamma(\cdot)$ for all $z \in C$;*
   (b) *for each $\Gamma \in \mathcal{Y}$ and $z \in E$ it holds that $\mathcal{L}_\Gamma V_\Gamma(z) \leq -A_1 V_\Gamma(z) + A_2 \mathbb{1}_C(z)$;*
   (c) *it holds that $\Delta t = mt_0$, for some $m \in \mathbb{N}$.*
2. *There exist $A_1, A_2 > 0$, $C_2 > 2A_2/A_1$, a class of functions $\{V_\Gamma : E \to [1,\infty) : \Gamma \in \mathcal{Y}\}$ with $\pi(V_\Gamma) < +\infty$, such that*
   (a) *for each $\Gamma \in \mathcal{Y}$, for all $x, y \in E$ such that $V_\Gamma(x) + V_\Gamma(y) \leq C_2$, there exists $\alpha, t_0 > 0$ such that*

$$\|P_\Gamma^{t_0}(x,\cdot) - P_\Gamma^{t_0}(y,\cdot)\|_{\mathrm{TV}} \leq 2(1-\alpha);$$

   (b) *for each $\Gamma \in \mathcal{Y}$ and for any $z \in E$, $\mathcal{L}_\Gamma V_\Gamma(z) \leq -A_1 V_\Gamma(z) + A_2$;*
   (c) *it holds that $\Delta t = t_0$.*

If the adaptation parameter is allowed to be updated only if the process is inside of a compact set B, as defined in (11), then the adaptive algorithm satisfies the containment condition and is thus ergodic.

Proof. The proof of this theorem can be found in Appendix B.3 of the supplement [6]. □

Remark. The restrictions on the step size can be milder than as stated in Theorem 3.6. For instance, if the minorisation condition (1a) of Theorem 3.6 holds for all $t \geq t_0$, then one is free to choose any step size $\Delta t > 0$. Furthermore, in both cases the assumption that the parameter can be updated only if the process is inside of a compact set at the adaptation time can be dropped when a simultaneous geometric drift condition holds (Assumption 3.2(b)).

3.3. Convergence properties of adaptive PDMC algorithms

Relying on Theorem 3.6, in this section we turn our attention to the ergodicity of adaptive PDMC algorithms. The proofs of the two theorems are postponed to Section 4. First, let us consider the adaptive ZZS. We assume the following conditions on the potential.

Assumption 3.7 (Growth Condition 3 in [11]). $U \in C^2(\mathbb{R}^d)$ and

$$\lim_{\|x\|\to\infty} \frac{\max(1, \|\nabla^2 U(x)\|)}{\|\nabla U(x)\|} = 0, \qquad \lim_{\|x\|\to\infty} \frac{\|\nabla U(x)\|}{U(x)} = 0.$$

Let us now state the ergodicity result for the adaptive ZZS.

Theorem 3.8. *Let $\mathcal{M}$ be a compact set of positive definite matrices and let Λ be a set of excess switching rates $\gamma : E \to \mathbb{R}_+^d$ for which there are $0 < \gamma_{\min} \leq \gamma_{\max} < \infty$ such that for all $\gamma \in \Lambda$*

$$\gamma_{\min} \leq \gamma(x,\theta) \leq \gamma_{\max} \qquad \text{for all } (x,\theta) \in E.$$

Let $\mathcal{P} = \{P_{M,\gamma} : M \in \mathcal{M}, \gamma \in \Lambda\}$ be a family of preconditioned Zig-Zag processes with generators defined by Equation (2). Suppose Assumption 3.7 holds and assume either of the following conditions holds:

a) $\mathcal{M} = \{M \in \mathbb{R}^{d\times d} : M_{ii} \in [a,b], M_{jk} = 0 \quad \text{for all } j \neq k\}$ *with* $b > a > 0$;
b) $\mathcal{M}$ *has no additional restrictions but adaptations are allowed only inside of a compact set* B,

and let Δt be the discretisation step. Then the containment condition holds. Moreover, if the adaptive strategy is as described in Section 2.4 and is such that $p_n \to 0$ as $n \to \infty$, then the diminishing adaptation condition holds and thus for $\mu = \pi \times \mathrm{Unif}(\Theta)$:

$$\lim_{n\to\infty} ||\mathbb{P}((X_n, \Theta_n) \in \cdot \,|x_0, \theta_0, \gamma_0) - \mu(\cdot)||_{TV} = 0 \qquad \text{for all } (x_0, \theta_0) \in E, \gamma_0 \in \Lambda. \tag{12}$$

Finally, for any bounded and measurable $f : \mathbb{R}^d \to \mathbb{R}$ a weak law of large numbers holds, i.e.

$$\frac{\sum_{n=1}^N f(X_n)}{N} \to \pi(f) \qquad \text{in probability}. \tag{13}$$

Remark. The time discretisation step Δt can be chosen freely and is not subject to constraints. Moreover, we remark that under condition (a) on $\mathcal{M}$ the adaptive algorithm is SGE, i.e. satisfies Assumption 3.2, while under condition (b) it satisfies the first set of conditions in Theorem 3.6. Thus if one is interested in learning only the diagonal elements of the covariance, then it is possible to take $B = \mathbb{R}^d$ and allow adaptations independently of the state of the process.

Remark. It was shown in [11] that the ZZS is geometrically ergodic under Assumption 3.7 also in the case $\gamma = 0$, whereas in Theorem 3.8 we require $\gamma(x,\theta) \geq \gamma_{\min} > 0$. This extra assumption is convenient when proving a simultaneous small set condition (see Lemma 4.1). Based on similar arguments as in [11], we expect the statement of Theorem 3.8 to remain valid even in the case $\gamma_{\min} = 0$. In practice, one is free to choose $\gamma_{\min}$ very small and thus this assumption does not represent a severe limitation.

Below we introduce a set of assumptions that is used to show ergodicity of the adaptive BPS. Here we limit our attention to the case of $\psi = \mathcal{N}(0, \mathbb{1}_d)$.

Assumption 3.9 (Assumptions A1, A2, and A7 in [17]). Let $U : \mathbb{R}^d \to [0,\infty)$ satisfy

(a) $U \in C^2(\mathbb{R}^d)$, and $x \to \|\nabla U(x)\|$ is integrable w.r.t. π;
(b) $\int_{\mathbb{R}^d} e^{-U(x)/2} dx < +\infty$ and $\lim_{\|x\|\to\infty} U(x) = +\infty$;
(c) There exists $\zeta \in (0,1)$ such that

$$\liminf_{\|x\|\to\infty} \frac{\|\nabla U(x)\|}{U^{1-\zeta}(x)} > 0, \qquad \limsup_{\|x\|\to\infty} \frac{\|\nabla U(x)\|}{U^{1-\zeta}(x)} < \infty,$$

and

$$\limsup_{\|x\|\to\infty} \frac{\|\nabla^2 U(x)\|}{U^{1-2\zeta}(x)} < \infty.$$

Theorem 3.10. *Let $\mathcal{M}$ be a compact set of positive definite matrices and $\Lambda_r = [\lambda_{\min}, \lambda_{\max}]$ for $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$. Let $\mathcal{P} = \{P_{M,\lambda_r} : M \in \mathcal{M}, \lambda_r \in \Lambda_r\}$ be a family of preconditioned BPS's as defined in Equation (4), where λ_r is the refreshment rate. Suppose Assumption 3.9 holds and let Δt be the*

discretisation step. Assume that $\psi = \mathcal{N}(0, \mathbb{1}_d)$. If adaptations are allowed only inside of a compact set as explained in Equation (11), then the containment condition holds. Furthermore, for the strategy discussed in Section 2.4 the diminishing adaptation holds as long as $p_n \to 0$ as $n \to \infty$. Thus the ABPS is ergodic in the sense of Equation (12) and satisfies a WLLN of the form (13) for any bounded and measurable $f : \mathbb{R}^d \to \mathbb{R}$.

Remark. Proving ergodicity of the adaptive BPS with refreshments from $\mathbb{S}^{d-1}$, i.e. the unit sphere centred at the origin, is more challenging due to the more involved drift condition proved in [16]. In particular, it is not straightforward to convert it into a simultaneous drift condition as required in assumption 2(b) of Theorem 3.6.

4. Proofs of the main theorems

4.1. Proof of Theorem 3.8

In order to prove the theorem we show that, for suitable families of preconditioners, either Assumption 3.2 holds for the family of discretised ZZ processes, or condition (1) in Theorem 3.6 holds for the family of continuous time ZZ processes. In the next two sections we state auxiliary results, while in Section 4.1.3 we assemble them to show the theorem. Proofs of the auxiliary results can be found in Appendix B4 in the supplement to this paper [6].

4.1.1. Minorisation condition for the ZZS

The following lemma shows that a simultaneous small set condition holds for the family of ZZ processes. The strategy of the proof is to reduce the d-dimensional minorisation condition to 1-dimensional conditions for every component of the process. Then we can take advantage of Lemma B.2 in Appendix B.4.1 in the supplement [6], which establishes that a simultaneous minorisation condition holds for a 1-dimensional ZZ process as long as lower and upper bounds for the switching rates are available.

Lemma 4.1. *Let $U \in C^1$. Consider the family of d-dimensional Zig-Zag processes with generators $\{\mathcal{L}_{M,\gamma} : M \in \mathcal{M}, \gamma \in \Lambda\}$, in which $\mathcal{M}$ is a compact set of positive definite matrices and Λ is a set of switching rates $\gamma : E \to \mathbb{R}^d_+$. Assume that there are $\gamma_{\min}, \gamma_{\max}$ such that for all $\gamma \in \Lambda$*

$$0 < \gamma_{\min} \leq \gamma(x,\theta) \leq \gamma_{\max} < \infty \qquad \text{for all } (x,\theta) \in E.$$

Then for any set of the form $C = D \times V$, where $D \subset \mathbb{R}^d$ is a compact set and $V \subseteq \Theta$, there exists $t_0 > 0$ such that for any $t \geq t_0$ there are $\delta > 0$, and probability measures $\{\nu_M\}_{M \in \mathcal{M}}$ on E such that

$$P^t_{M,\gamma}((x,\theta),\cdot) \geq \delta \nu_M(\cdot) \qquad \text{for all } (x,\theta) \in C.$$

In particular, t_0 and δ do not depend neither on M nor on γ.

Proof. The proof can be found in Appendix B.4.1 in the supplement [6]. □

4.1.2. Drift conditions for the Zig-Zag process

If we restrict our attention to the class of diagonal matrices with positive, bounded entries, then the Lyapunov function in [11, Lemma 11] satisfies also a simultaneous drift condition. This is shown in the following lemma.

Lemma 4.2. *Let Assumption 3.7 hold. Consider the family of linearly transformed Zig-Zag processes with generators* $\{\mathcal{L}_{M,\gamma} : M \in \mathcal{M}, \gamma \in \Lambda\}$*, where*

$$\mathcal{M} = \{M \in \mathbb{R}^{d\times d} : M_{ii} \in [V^i_{\min}, V^i_{\max}],\ M_{jk} = 0 \text{ for all } j \neq k\}, \tag{14}$$

with $V_{\max} \geq V^i_{\max} \geq V^i_{\min} \geq V_{\min} > 0$ *for each* $i = 1, \ldots, d$*, and where* Λ *is a set of excess switching rates* $\gamma : E \to \mathbb{R}^d_+$ *such that for all* $\gamma \in \Lambda$ *it holds that*

$$\gamma_i(x,\theta) \leq \gamma_{\max} \qquad \text{for all } (x,\theta) \in E,\ i = 1, \ldots, d. \tag{15}$$

Let $\delta > 0$ *and* $\alpha > 0$ *be such that* $0 < (\delta\gamma_{\max})/V_{\min} < \alpha < 1$ *and define* $\phi(s) = \frac{1}{2}\text{sign}(s)\ln(1 + \delta|s|)$*. Then the function*

$$V(x,\theta) = \exp\left(\alpha U(x) + \sum_{i=1}^{d} \phi(\theta_i \partial_i U(x))\right) \tag{16}$$

is a simultaneous Lyapunov function for the family of ZZ processes, that is there exist $A_1 > 0$*,* $A_2 > 0$*, a compact set* $C \subset E$ *such that*

$$\mathcal{L}_{M,\gamma} V(x,\theta) \leq -A_1 V(x,\theta) + A_2 \mathbb{1}_C(x,\theta) \qquad \text{for all } (x,\theta) \in E,\ M \in \mathcal{M},\ \gamma \in \Lambda,$$

where A_1*,* A_2*,* C *do not depend neither on* M *nor on* γ *(but depend on* $\mathcal{M}$ *and* Λ*).*

Proof. The proof can be found in Appendix B.4.2 in the supplement [6]. □

If we wish to consider a more general class of positive-definite matrices, then we have to settle for the following, weaker result.

Lemma 4.3. *Let Assumption 3.7 hold. Consider a family of linearly transformed Zig-Zag processes with generators* $\{\mathcal{L}_{M,\gamma} : M \in \mathcal{M}, \gamma \in \Lambda\}$*, where* $\mathcal{M} \subset \mathbb{R}^{d\times d}$ *is a compact space of positive definite matrices and* Λ *is a space of excess switching rates* $\gamma : E \to \mathbb{R}_+$ *such that (15) is satisfied for some* $\gamma_{\max}$*. Let* $\delta > 0$ *and* $\alpha > 0$ *be such that* $0 < \delta\gamma_{\max} < \alpha < 1$*. Define for each* $M \in \mathcal{M}$ *the function*

$$V_M(x,\theta) = \exp\left(\alpha U(x) + \sum_{i=1}^{d} \phi(\theta_i \langle M_i, \nabla U(x)\rangle)\right), \tag{17}$$

where M_i *denotes the* i*-th column of* M *and* $\phi : \mathbb{R} \to \mathbb{R}$ *was defined in Lemma 4.2. Then there are* $A_1 > 0$*,* $A_2 > 0$*, and a compact set* $C \subset E$ *such that for all* $M \in \mathcal{M}$ *the following simultaneous drift condition holds:*

$$\mathcal{L}_{M,\gamma} V_M(x,\theta) \leq -A_1 V_M(x,\theta) + A_2 \mathbb{1}_C(x,\theta) \qquad \text{for all } (x,\theta) \in E.$$

In particular A_1*,* A_2*,* C *do not depend neither on* M *nor on* γ *(but depend on* $\mathcal{M}$ *and* Λ*).*

Proof. The proof can be found in Appendix B.4.2 in the supplement [6]. □

4.1.3. Finalising the proof of Theorem 3.8

Let us first consider the case of diagonal preconditioners. Let $\Delta t > 0$ be the discretisation step. Then by Lemma 4.1 for any set of the form $C = D \times V$, with D compact and $V \subseteq \Theta$, there exist $\delta > 0$, $n_0 := \inf\{n \in \mathbb{N} : n \geq \frac{t_0}{\Delta t}\}$, $\nu_M(\cdot)$ such that C is a uniform (ν_M, n_0, δ)-small set for the family of discretised processes. Observe that no conditions on Δt are required. Moreover, a simultaneous drift condition holds by Lemma 4.2 combined with Lemma B.3 in the supplement [6] for any Δt. The condition $\mu(V) < \infty$ is satisfied by definition of the Lyapunov function V, and in fact it also holds that $\sup_{(x,\theta)\in C} V(x,\theta) < \infty$ by continuity of V in x. Therefore all the conditions of Assumption 3.2 are verified, which means the family is simultaneously geometrically ergodic and by Theorem 3 in [4] the containment condition is satisfied.

In the case of a non-diagonal transformation matrix, parts (a)-(c) of condition (1) in Theorem 3.6 are verified by Lemmas 4.1 and 4.3. It also holds that $\sup_{\{(x,\theta)\in C, M\in\mathcal{M}\}} V_M(x,\theta) < \infty$ for (small) sets $C = D \times V$, with D compact and $V \subseteq \Theta$, because of continuity of each V_M in x and M, together with the fact that D and $\mathcal{M}$ are compact spaces (see Lemma 4.3 for the definition of $\{V_M\}_{M\in\mathcal{M}}$). Moreover $\mu(V_M) = \tilde{\mu}_M(\tilde{V}_M) < \infty$, where $\tilde{\mu}_M = \tilde{\pi}_M \otimes \mathrm{Unif}(\Theta)$ and $\tilde{V}_M$ is a Lyapunov function for a standard ZZ process with invariant measure $\tilde{\mu}_M$. Theorem 3.6 ensures that the containment condition holds true with no restriction on Δt.

Proposition 4.6 implies that, under the assumption that $p_n \to 0$ as $n \to \infty$, the adaptive strategy satisfies diminishing adaptation. Therefore ergodicity follows from diminishing adaptation and containment.

4.2. Proof of Theorem 3.10

In the next two sections we show respectively that condition (2) in Theorem 3.6 is verified for the family of preconditioned BPS and/or of BPS with refreshment rates in a compact set. In Section 4.2.3 we use these auxiliary results to show the theorem.

4.2.1. Simultaneous coupling inequality

The next lemma states that a simultaneous coupling inequality is satisfied for the BPS with adaptive preconditioner and/or adaptive refreshment rate. The proof is based on the proof of Lemma 12 in [17], which shows a coupling inequality result for the standard BPS.

Lemma 4.4. *Let condition (a) in Assumption 3.9 hold for the energy function U. Consider the family of BP processes $\{P^t_{M,\lambda_r} : M \in \mathcal{M}, \lambda_r \in \Lambda_r\}$, where $\mathcal{M}$ is a compact space of non-singular preconditioning matrices and $\Lambda_r = [\lambda_r^{\min}, \lambda_r^{\max}]$ is the set of refreshment rates, for some $0 < \lambda_r^{\min} \leq \lambda_r^{\max} < \infty$. Then for any compact set $K \subset \{(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^d : \|x\| + \|\theta\| \leq R\}$, with $R \geq 0$, there exists $\alpha > 0$ such that for all (x,θ), $(\tilde{x},\tilde{\theta}) \in K$, for all $t > 0$, and for all $M \in \mathcal{M}$ and $\lambda_r \in \Lambda_r$*

$$\|P^t_{M,\lambda_r}((x,\theta),\cdot) - P^t_{M,\lambda_r}((\tilde{x},\tilde{\theta}),\cdot)\|_{\mathrm{TV}} \leq 2(1-\alpha).$$

In particular α is independent of M and λ_r.

Proof. The proof can be found in Appendix B.5.1 in the supplement [6]. □

4.2.2. Drift condition for the BPS

The second condition we need is uniformity of the constants in the drift condition for the family of preconditioned BPS and/or for the family of BPS with different refreshment rate. To this end, we go through the proof of Lemma 7 from [17] to show that this is indeed the case.

Lemma 4.5. *Consider a family of BP processes with generators* $\{\mathcal{L}_{M,\lambda_r} : M \in \mathcal{M}, \lambda_r \in \Lambda_r\}$, *where* $\mathcal{M}$ *is a compact space of non-singular matrices that act as preconditioners,* λ_r *is the refreshment rate and* $\Lambda_r = [\lambda_{min}, \lambda_{max}]$ *for some* $0 < \lambda_{min} \leq \lambda_{max} < \infty$. *Let Assumption 3.9 hold and let* $\psi = \mathcal{N}(0, \mathbb{1}_d)$. *Then there are* $A_1, A_2 > 0$ *and a class of functions* $\{V_{M,\lambda_r} : M \in \mathcal{M}, \lambda_r \in \Lambda_r\}$ *such that for each* $M \in \mathcal{M}$ *and* $\lambda_r \in \Lambda_r$ *it holds that*

$$\mathcal{L}_{M,\lambda_r} V_{M,\lambda_r}(x,\theta) \leq -A_1 V_{M,\lambda_r}(x,\theta) + A_2 \qquad \text{for all } (x,\theta) \in \mathbb{R}^d \times \mathbb{R}^d,$$

where in particular A_1, A_2 *do not depend on* M.

Proof. The proof can be found in Appendix B.5.2 in the supplement [6]. □

4.2.3. Finalising the proof of Theorem 3.10

Let $\Delta t > 0$ be a discretisation step. Lemma 4.5 gives the drift condition, that is condition (b) in Theorem 3.6. Then Lemma 4.4 implies that a coupling inequality holds for any compact set. Sets of the form $V_{M,\lambda_r}(x,\theta) + V_{M,\lambda_r}(x,\theta) \leq C_2$ are compact by definition of the class of Lyapunov functions $\{V_{M,\lambda_r} : M \in \mathcal{M}, \lambda_r \in \Lambda_r\}$. We are in particular free to choose the constant C_2 as large as we wish. Note also that the coupling inequality holds for all $t > 0$, hence there are no constraints on the choice of Δt. Moreover $\mu(V_{M,\lambda_r}) = \tilde{\mu}_M(\tilde{V}_{M,\lambda_R}) < \infty$ for all $M \in \mathcal{M}$ and $\lambda_r \in \Lambda_r$. Here $\tilde{V}_{M,\lambda_R}$ is the Lyapunov function of a standard BPS with refreshment rate λ_r and target $\tilde{\mu}_M = \tilde{\pi}_M \times \Psi$. The containment condition is thus verified as all conditions in part (2) of Theorem 3.6 hold. Proposition 4.6 implies the diminishing adaptation condition, and thus ergodicity follows.

4.3. Proving the diminishing adaptation condition

A key part of Theorem 3.1 is condition (b), i.e. the diminishing adaptation condition. For the adaptive scheme described in Section 2.4 the condition can be easily shown to be true as the adaptation happens with diminishing probability.

Proposition 4.6. *Consider the adaptive schemes in Section 2. In particular, assume that* $\{p_n\}_{n\geq 0}$, *i.e. the sequence of probabilities of updating the adaptation parameters, is such that* $p_n \to 0$ *as* $n \to \infty$. *Then the diminishing adaptation holds for any* $t \geq 0$.

Proof. Consider for example the adaptive BPS. Observe that $M_{n+1} = M_n$ and $\lambda_r^{n+1} = \lambda_r^n$ with probability $1 - p_{n+1}$ and thus

$$\|P^{\Delta t}_{M_{n+1},\lambda_r^{n+1}}((x,\theta),\cdot) - P^{\Delta t}_{M_n,\lambda_r^n}((x,\theta),\cdot)\|_{\mathrm{TV}} \leq 2p_{n+1} \to 0 \qquad \text{as } n \to \infty.$$

Thus the diminishing adaption holds. The same reasoning works for the adaptive ZZS. □

5. Numerical experiments

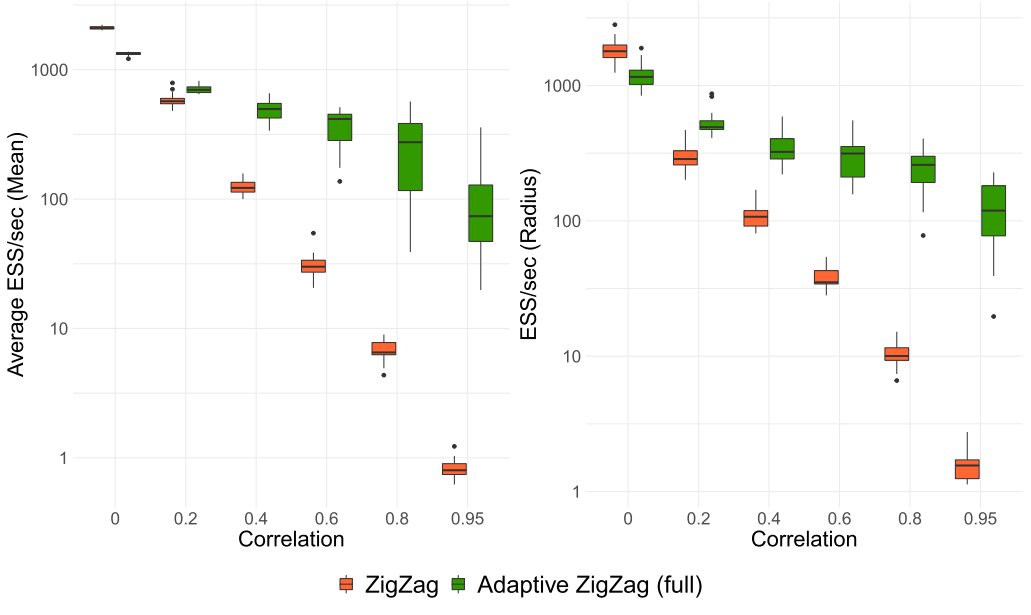
In this section we test the empirical performance of the adaptive schemes we defined in Section 2.4. All experiments are implemented in Julia and the corresponding codes can be found at https://github.com/andreabertazzi/Adaptive_PDMC_samplers. Let us state some settings that hold for all experiments below. The time horizon is set to $T = 10^5$ for all processes. This is large enough for the adaptive PDMC samplers to learn and take advantage of the covariance structure. When considered fixed, the refreshment rate of the BPS is taken to be $\lambda_r = 1$. The excess switching rate for ZZS is set to 0 in all experiments. The discretisation step is chosen to be $\Delta t = 0.5$, which in our experiments turns out to be a good choice for a wide range of targets. Moreover, we set adaptation times to be every $t_{\text{adap}} = 2000$ continuous time units. The probability of adapting decays as $O(\log \log n)$. Finally, no normalisation in the sense discussed at the end of Section 2.3 is employed. The performance measures we consider are the *effective sample size per second* (ESS/sec) for the mean and for the radius statistic $t(x) = \sum_{i=1}^{d} x_i^2$. In Sections 5.1 and 5.2 these are computed in continuous time as discussed in [8] by estimating the asymptotic variance with the batch means method, and the variance of the Monte Carlo estimate on the continuous time trajectories of the processes. On the other hand, in Section 5.3 we compute the mean squared error (MSE) in discrete time for the sample mean and radius statistic and take advantage of the fact that in the large time horizon regime the MSE is approximately given by the asymptotic variance of the observable divided by the number of generated samples. This alternative way to compute the ESS avoids poor convergence of the batch means method in the multimodal case. Finally, in all settings we repeat the same task 20 times and report all the results in box-plots.

5.1. Multidimensional Gaussian target

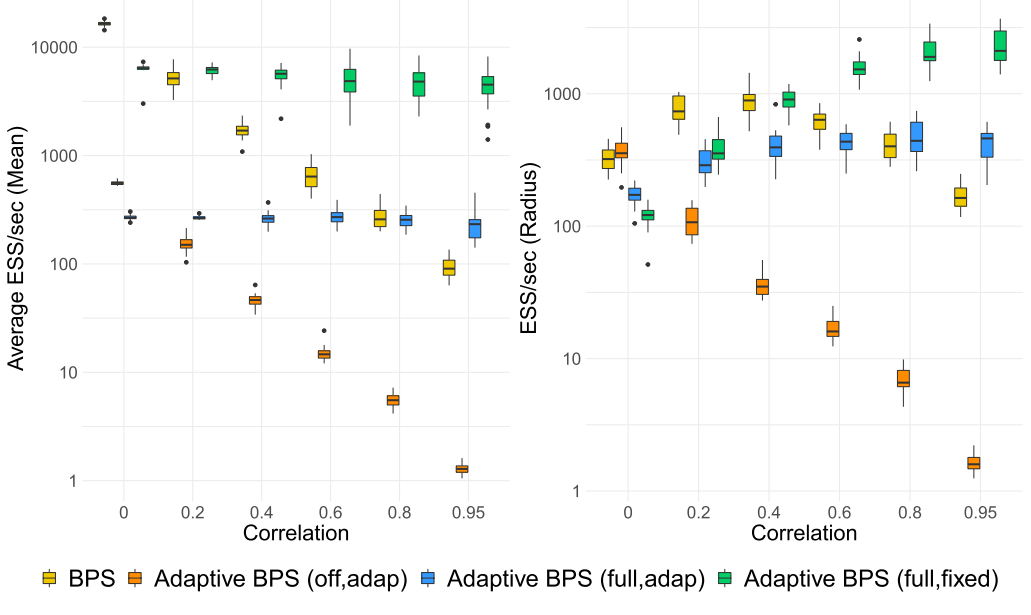
In this section we focus on two different kinds of multivariate Gaussian target distributions. The first one, denoted by **MG1**, has unitary variances and correlation ρ between all components. Denoting the covariance matrix by Σ, this means that $\Sigma_{ii} = 1$ for each $i = 1, \dots, d$ and $\Sigma_{ij} = \rho$ for all $i \neq j$. We study how the adaptive PDMC algorithms compare to their non-adaptive counterparts for different values of ρ and different dimensionalities. In this setting we focus on adaptive algorithms that estimate the full covariance matrix. The second Gaussian target we consider has variances $0.5, 1, 5, 10, 15$ repeated depending on the dimension, together with a milder correlation between components. This setting is denoted as **MG2** and is useful to compare all kinds of adaptive algorithms we introduced.

*5.1.1. **MG1** target*

In the first experiment we consider a 50-dimensional **MG1** target for different values of ρ. In Figure 3 the average ESS/sec and the ESS/sec for the radius statistic are shown. As expected, the performance of the ZZS is degrading as the correlation increases. This behaviour is for the most part caused by the very large number of events that have to be simulated for very narrow targets. The adaptive ZZS successfully improves over this inconvenience and is stable with respect to the increasing correlation. The standard BPS with fixed refreshment rate shows a decaying average ESS/sec, while the ESS/sec for the radius statistic appears to increase as ρ grows up until $\rho = 0.4$ and then becomes smaller. This behaviour is likely due to the fact that the choice $\lambda_r = 1$ is more suited for the estimation of the radius in case of a more concentrated target rather than for a standard Gaussian. A similar behaviour is shown by the adaptive BPS's. Overall we notice a marked improvement for the BPS with adaptive preconditioner and fixed λ_r. Choosing to adapt only the refreshment turns out to be a detrimental decision when the target is correlated. Indeed the optimality criterion derived [10] assumes a standard Gaussian target.



(a) Comparison between the ZZS and the ZZS with adaptive preconditioner learning the **full** covariance matrix.



(b) Comparison between the BPS and various alternative adaptive BPS's. **Adaptive BPS (full,adap)** denotes the BPS that learns the entire covariance matrix (**full**) and with adaptive refreshment rate (**adap**). Adaptation of the preconditioner can be turned **off**, and similarly the refreshment rate can be **fixed**.

Figure 3. Results as a function of the correlation for **MG1** targets.

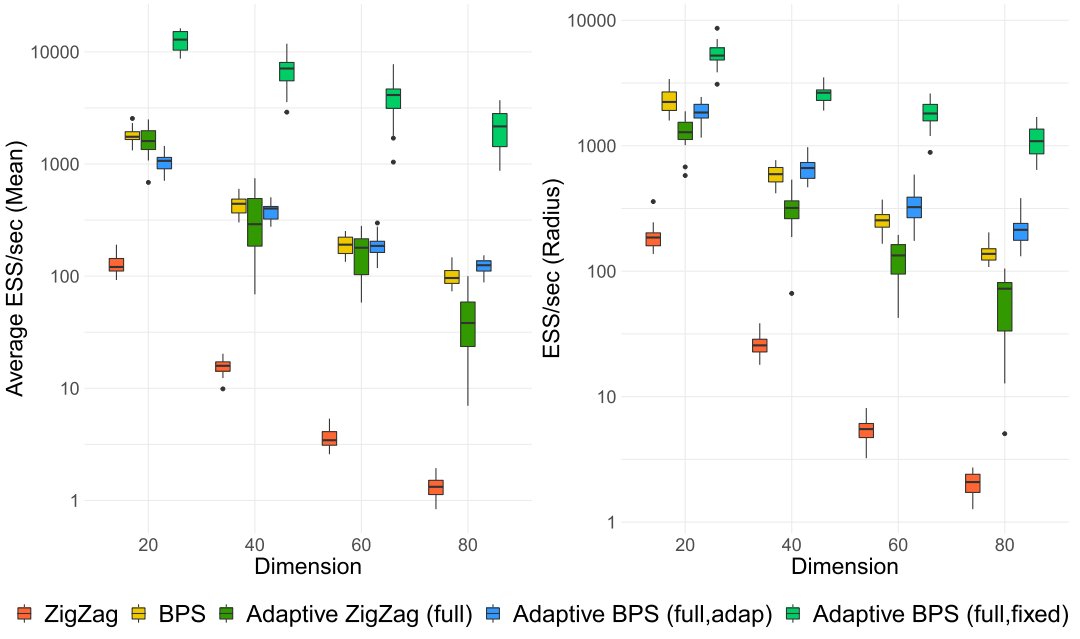
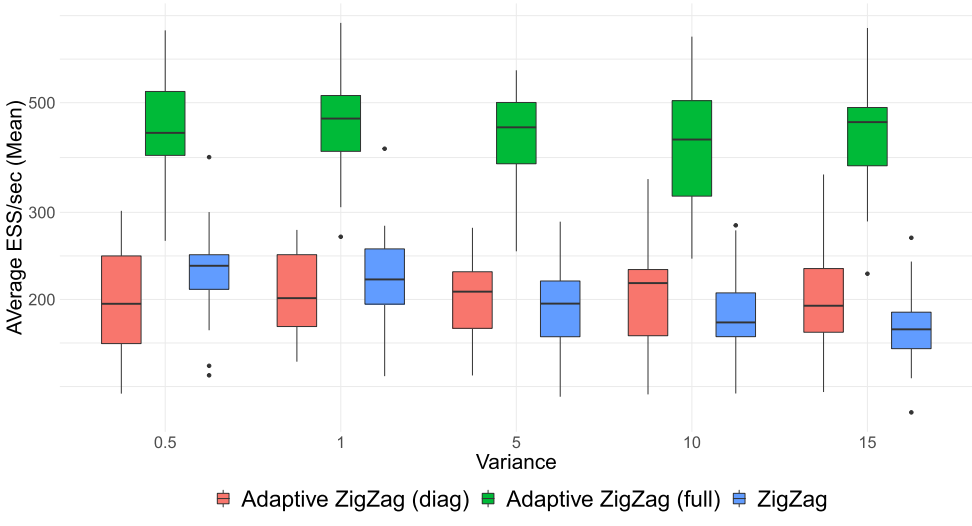


Figure 4. **MG1** target with $\rho = 0.8$ and different dimensionalities.

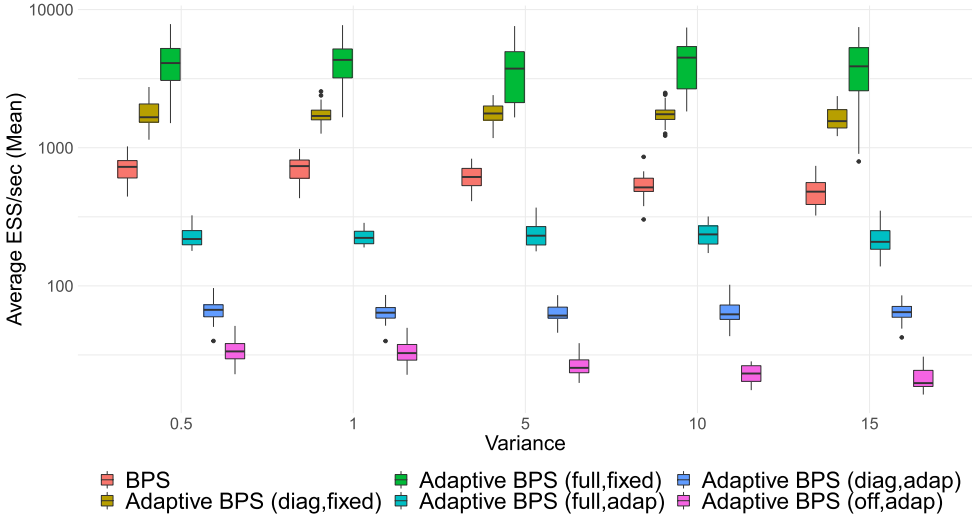
In Figure 4 we study how the adaptive schemes compare to the standard ones for an **MG1** target with correlation $\rho = 0.8$ and increasing dimensionalities of target. The plots show that when the target is strongly correlated the effect of the adaptation shows no sign of diminishing. It also seems clear that the sampler of choice in this case should be the adaptive BPS with fixed, rather than adaptive, refreshment rate. This could be due to the fact that, when the refreshment rate is updated adaptively and the target is anisotropic, a too large λ_r is chosen at first due to the high number of reflections, thus slowing down the estimation of the covariance matrix. As suggested by a reviewer, one could avoid this issue by keeping the refreshment rate fixed until the estimate of the covariance stabilises, and only then starting to learn the optimal λ_r. It is worth pointing out that the performance of the BPS with adaptive preconditioner and refreshment improves compared to the BPS as the dimension increases. This is according to the theoretical results in [10], which are indeed obtained in the high dimensional limit. Therefore we expect that for a large d it is reasonable to apply both the transformation scheme and the tuning of λ_r.

*5.1.2. **MG2** target*

Let us now consider a 50-dimensional **MG2** target with a mild correlation set to $\rho = 0.3$. Figure 5 shows the results for several adaptive PDMC samplers. The adaptive algorithms that learn the entire covariance matrix show the largest gain in terms of ESS/sec. For the BPS the choice of learning only the variance of each component of the target seems interesting, also in view of larger dimensions. As in the previous section, we observe that the adaptation of the refreshment rate seems to have a bad effect for anisotropic targets.



(a) Results of the **MG2** experiment for the ZZS. The option **diag** refers to the adaptive algorithm that learns only the diagonal of the covariance matrix.



(b) Results of the **MG2** experiment for the BPS.

Figure 5. Comparison of several samplers in the context of Section 5.1.2.

5.2. Logistic regression with correlated data

The next numerical experiment we consider is a Bayesian logistic regression task. In this setting for $j = 1, \ldots, n_{\text{obs}}$ a binary output value $y_j \in \{0, 1\}$ has distribution

$$\mathbb{P}(Y_j = 1|\beta) = \frac{1}{1 + \exp(-\beta^T x_j)},$$

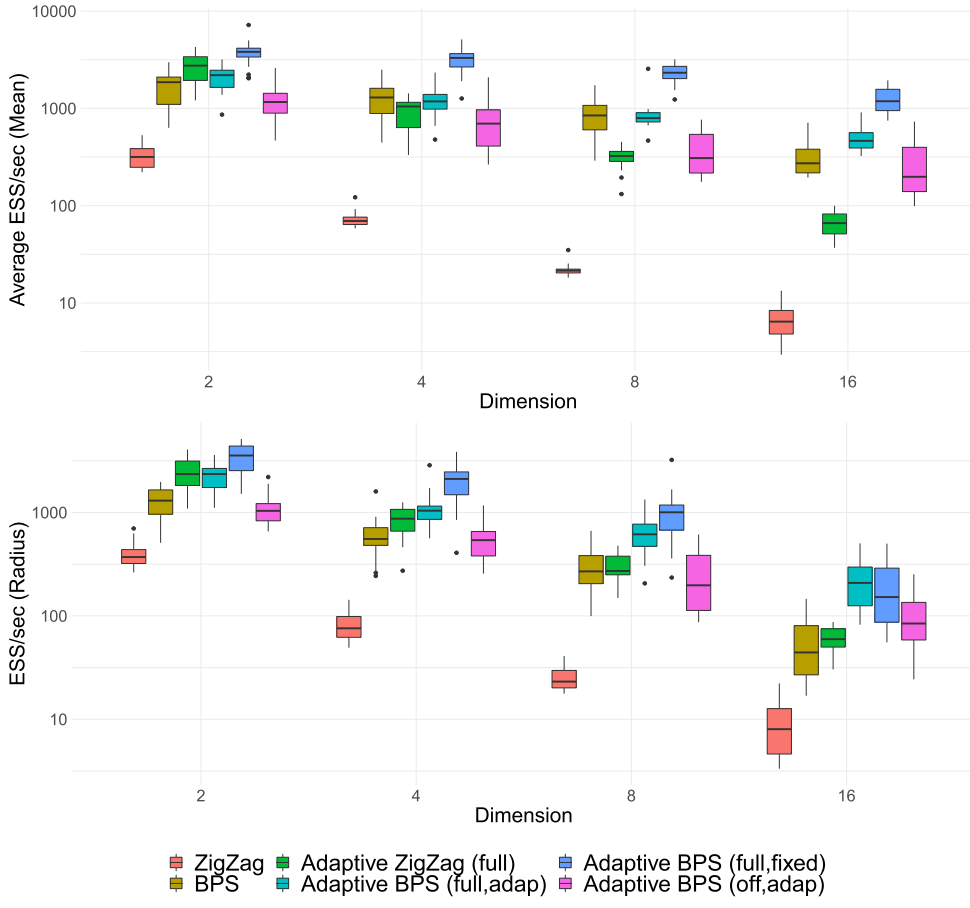


Figure 6. Logistic regression task of Section 5.2.

where $\{x_j\}_{j=1}^{n_{\text{obs}}}$ are known covariates, and $\beta \in \mathbb{R}^d$ is an unknown parameter. We take a flat prior and thus obtain the posterior

$$\pi(\beta|\{y_j\}_{j=1}^{n_{\text{obs}}}) \propto \prod_{j=1}^{n_{\text{obs}}} \frac{\exp(-y_j \beta^T x_j)}{1+\exp(-\beta^T x_j)}.$$

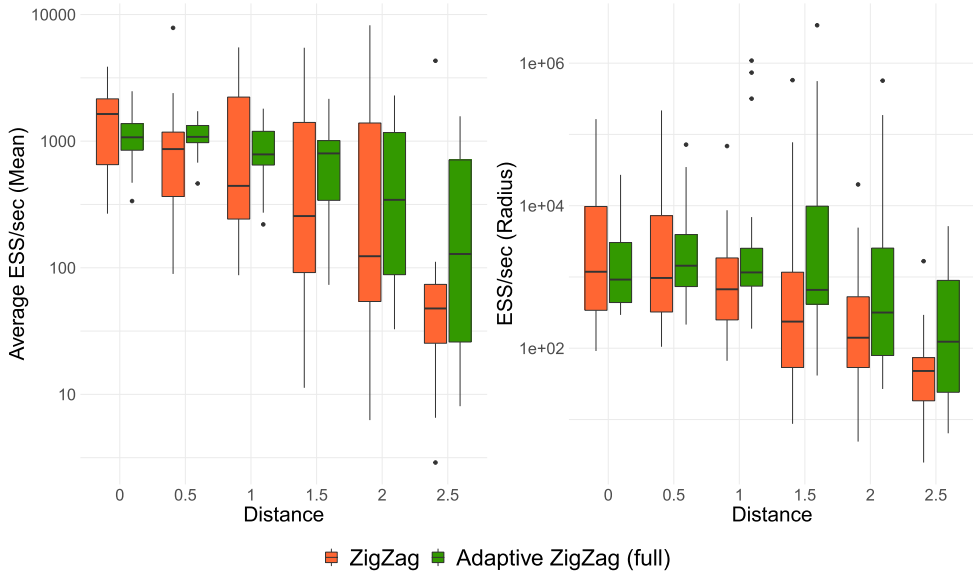
We force correlation between some components of the parameter by taking, for $j = 1,\ldots,n_{\text{obs}}$ and $i = 1,\ldots,d$, $(x_j)_i = 1+\varepsilon N_{ji}$, where $N_{ji} \sim \mathcal{N}(0,1)$ and $\varepsilon = 0.1$. The results of the experiment are reported in Figure 6, in which the samplers are tested with targets as above with $d = 2,4,8,16$ and $n_{\text{obs}} = 1000$. The adaptation of the refreshment rate follows the alternative scheme discussed in Appendix A.3 in the supplement [6]. This scheme seems more stable as the refreshment rate is update gradually and cannot jump immediately to very large or small values. Although the dimensionality is small and the correlation is limited to a subset of the coordinates, we observe that the adaptive schemes outperform their standard counterparts.

5.3. Mixture of Gaussian distributions

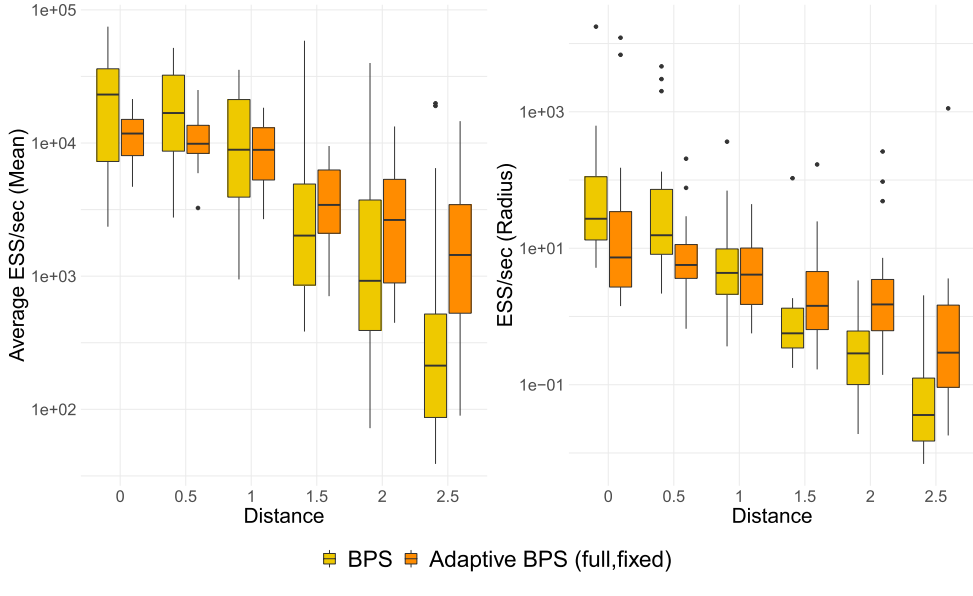
Consider a mixture of two 30-dimensional Gaussian distributions $\mathcal{N}(0_d, \Sigma)$ and $\mathcal{N}(\mu, \Sigma)$, both with weight $\frac{1}{2}$. Here we take Σ with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.25$ for $j \neq i$. Moreover we take $\mu = a \times (1, \dots, 1)$, where a is a parameter that determines the distance between the two means. We investigate the performance of the adaptive schemes compared to standard ones as a function of the parameter a. The details on the implementation of this experiment can be found in Appendix A.4 in the supplement [6]. In this experiment we test the robustness of the algorithm in a case in which the target distribution is cigar shaped, but multimodal. The time horizon for ZZS is 10^5, while for BPS it is 4×10^5. Figure 7 shows the results for the adaptive samplers in this setting. We observe that as the distance between the means increases the performance of the adaptive samplers improves over the standard ones. We remark that for small values of the distance a one cannot expect improvements of the adaptive algorithms as the correlation is small and thus the target is not very anisotropic.

6. Discussion

In this paper we proposed adaptive schemes to overcome two of the current issues with the BPS and ZZS. We have shown that the refreshment rate and the excess switching rate can be tuned on the fly as long as the updates or the probabilities of updating decrease to 0. With this approach the user does not have to worry about tuning the refreshment rate. A current limitation is that more theory or experiments are needed to determine a criterion that works well with anisotropic targets. In addition to this, we have proposed a way to make the PDMC samplers learn and take advantage of the covariance structure of the target. The theoretical results stated in Theorems 3.8 and 3.10 ensure that the adaptive samplers are ergodic and thus converge to the correct measure π. It is challenging to prove theoretical statements regarding performance improvements of the adaptive schemes over the standard PDMC samplers. However, the numerical experiments we conducted suggest that our adaptive algorithms can lead to a significant performance improvement when there are strong anisotropies. An alternative approach could be to use an optimisation algorithm to obtain an estimate of the covariance matrix by computing the Hessian of the target at its point of maximum. However, this optimisation step can be expensive and moreover for realistic problems it is not a given that this estimator is a good approximation of the posterior covariance. In particular in our theory we do not assume log convexity of the target, and also we do not assume that we are in the large sample regime where a Bernstein-von Mises theorem holds. In addition, the adaptive schemes discussed in this paper can be applied to the Boomerang sampler [9]. This would result in elliptical dynamics that are adapted to resemble the (unimodal) target at hand. Naturally, the adaptive algorithms should be run with a time horizon that is large enough to benefit from the adaptation. Two other important settings are the discretisation step and the time between two adaptations. Based on our experience with the experiments, we suggest $\Delta t = O(10^{-1})$ and $t_{\text{adaps}} = O(10^3)$. For concentrated targets, as for instance posteriors when there is a very large number of data points, it is suggested to choose both values small. We remark that in very high dimensional settings it may be unfeasible to let the samplers learn the full covariance matrix, as the computation of M entails calculating the square root of the empirical covariance matrix. In such cases we suggest either learning only the diagonal elements or blocks of the covariance. We remark that the adaptive PDMC algorithms with subsampling are applicable in the setting of tall data, that is when data-set is made of a large number of observations, but with a moderate dimensionality. In such settings subsampling can be shown empirically to result in an improved efficiency (in terms of ESS per second); a result which is backed by a heuristic argument, based on posterior contraction, i.e., the Bernstein-Von Mises theorem; see [8] for details. More research is necessary to understand in which situations subsampling can



(a) Results for the ZZS with time horizon $T = 10^5$.



(b) Results for the BPS with time horizon $T = 4 \times 10^5$.

Figure 7. Numerical results for a mixture of two Gaussian distributions as described in Section 5.3.

lead to improved efficiency, and in particular if improved efficiency is possible in cases for which the Bernstein-von Mises theorem does not apply; see also [5,22] for a critical discussion of subsampling methods.

A question that one could naturally ask is how applicable this transformation scheme is in case of a multimodal target. The answer depends on the specific target at hand, but one can design a target as a mixture of Gaussian distributions for which applying the transformation scheme would not speed up the convergence of the sampler. However, when the target is multimodal it is possible for instance to use the adaptive PDMC samplers together with the framework proposed in [28]. In this framework, the adaptive PDMC samplers would be beneficial since the regions around each mode would be explored more efficiently by taking advantage of the covariance structure of the specific mode.

Finally, we remark that the idea of learning the covariance structure of the target on the fly could be applied to obtain adaptive versions of the Hamiltonian Monte Carlo (HMC) algorithm [26] and of the Metropolis Adjusted Langevin Algorithm (MALA) [32]. In particular both the HMC algorithm and the MALA are sensitive to correlation in the target and can thus benefit from a suitable preconditioner, as argued respectively in Section 4.1 of [26] and in [31]. Moreover, the preconditioner could be chosen to take advantage of the geometry of the target, as proposed in [21] for HMC and MALA. The preconditioner could be estimated adaptively with an appropriate adaptation strategy, together with similar ideas presented in this manuscript.

Acknowledgements

This work is part of the research programme 'Zigzagging through computational barriers' with project number 016.Vidi.189.043, which is financed by the Dutch Research Council (NWO). We acknowledge helpful discussions with Gareth Roberts. We are thankful to the Associate Editor and the two Referees for their comments, which led to several improvements in the manuscript.

Supplementary Material

Supplement to "Adaptive schemes for piecewise deterministic Monte Carlo algorithms" (DOI: 10.3150/21-BEJ1423SUPP; .pdf). In the supplement [6] we give details on the implementation of the numerical experiments and we prove all the theoretical results stated in the paper.

References

[1] Andrieu, C., Durmus, A., Nüsken, N. and Roussel, J. (2021). Hypocoercivity of piecewise deterministic Markov process-Monte Carlo. *Ann. Appl. Probab.* **31** 2478–2517. MR4332703 https://doi.org/10.1214/20-aap1653

[2] Andrieu, C. and Livingstone, S. (2021). Peskun-Tierney ordering for Markovian Monte Carlo: Beyond the reversible scenario. *Ann. Statist.* **49** 1958–1981. MR4319237 https://doi.org/10.1214/20-aos2008

[3] Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. MR2461882 https://doi.org/10.1007/s11222-008-9110-y

[4] Bai, Y., Roberts, G.O. and Rosenthal, J.S. (2011). On the containment condition for adaptive Markov chain Monte Carlo algorithms. *Adv. Appl. Stat.* **21** 1–54. MR2849670

[5] Bardenet, R., Doucet, A. and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *J. Mach. Learn. Res.* **18** 47. MR3670492

[6] Bertazzi, A. Bierkens, J. (2022). Supplement to "Adaptive schemes for piecewise deterministic Monte Carlo algorithms." https://doi.org/10.3150/21-BEJ1423SUPP

[7] Bierkens, J. and Duncan, A. (2017). Limit theorems for the zig-zag process. *Adv. in Appl. Probab.* **49** 791–825. MR3694318 https://doi.org/10.1017/apr.2017.22

[8] Bierkens, J., Fearnhead, P. and Roberts, G. (2019). The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *Ann. Statist.* **47** 1288–1320. MR3911113 https://doi.org/10.1214/18-AOS1715

[9] Bierkens, J., Grazzi, S., Kamatani, K. and Roberts, G. (2020). The boomerang sampler. In *Proceedings of the 37th International Conference on Machine Learning* (H.D. III and A. Singh, eds.). *Proceedings of Machine Learning Research* **119** 908–918. PMLR.
[10] Bierkens, J., Kamatani, K. and Roberts, G.O. (2018). High-dimensional scaling limits of piecewise deterministic sampling algorithms. Available at arXiv:1807.11358.
[11] Bierkens, J., Roberts, G.O. and Zitt, P.-A. (2019). Ergodicity of the zigzag process. *Ann. Appl. Probab.* **29** 2266–2301. MR3983339 https://doi.org/10.1214/18-AAP1453
[12] Bouchard-Côté, A., Vollmer, S.J. and Doucet, A. (2018). The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *J. Amer. Statist. Assoc.* **113** 855–867. MR3832232 https://doi.org/10.1080/01621459.2017.1294075
[13] Chimisov, C., Latuszynski, K. and Roberts, G. (2018). Adapting The Gibbs Sampler. Available at arXiv:1801.09299.
[14] Craiu, R.V., Gray, L., Łatuszyński, K., Madras, N., Roberts, G.O. and Rosenthal, J.S. (2015). Stability of adversarial Markov chains, with an application to adaptive MCMC algorithms. *Ann. Appl. Probab.* **25** 3592–3623. MR3404645 https://doi.org/10.1214/14-AAP1083
[15] Davis, M.H.A. (1993). *Markov Models and Optimization*. *Monographs on Statistics and Applied Probability* **49**. London: CRC Press. MR1283589 https://doi.org/10.1007/978-1-4899-4483-2
[16] Deligiannidis, G., Bouchard-Côté, A. and Doucet, A. (2019). Exponential ergodicity of the bouncy particle sampler. *Ann. Statist.* **47** 1268–1287. MR3911112 https://doi.org/10.1214/18-AOS1714
[17] Durmus, A., Guillin, A. and Monmarché, P. (2020). Geometric ergodicity of the bouncy particle sampler. *Ann. Appl. Probab.* **30** 2069–2098. MR4149523 https://doi.org/10.1214/19-AAP1552
[18] Fearnhead, P., Bierkens, J., Pollock, M. and Roberts, G.O. (2018). Piecewise deterministic Markov processes for continuous-time Monte Carlo. *Statist. Sci.* **33** 386–412. MR3843382 https://doi.org/10.1214/18-STS648
[19] Fort, G., Moulines, E. and Priouret, P. (2011). Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.* **39** 3262–3289. MR3012408 https://doi.org/10.1214/11-AOS938
[20] Fort, G., Moulines, E., Priouret, P. and Vandekerkhove, P. (2014). A central limit theorem for adaptive and interacting Markov chains. *Bernoulli* **20** 457–485. MR3178506 https://doi.org/10.3150/12-BEJ493
[21] Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 123–214. MR2814492 https://doi.org/10.1111/j.1467-9868.2010.00765.x
[22] Johndrow, J.E., Pillai, N.S. and Smith, A. (2020). No free lunch for approximate MCMC. ArXiv preprint. Available at arXiv:2010.12514.
[23] Lu, J. and Wang, L. (2020). On explicit L^2-convergence rate estimate for piecewise deterministic Markov processes. Available at arXiv:2007.14927.
[24] Michel, M., Kapfer, S.C. and Krauth, W. (2014). Generalized event-chain Monte Carlo: Constructing rejection-free global-balance algorithms from infinitesimal steps. *J. Chem. Phys.* **140** 054116. https://doi.org/10.1063/1.4863991
[25] Monmarché, P. (2016). Piecewise deterministic simulated annealing. *ALEA Lat. Am. J. Probab. Math. Stat.* **13** 357–398. MR3487077
[26] Neal, R.M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. Boca Raton, FL: CRC Press. MR2858447
[27] Peters, E.A.J.F. and De With, G. (2012). Rejection-free Monte Carlo sampling for general potentials. *Phys. Rev. E, Stat. Nonlin. Soft Matter Phys.* **85** 1–5. https://doi.org/10.1103/PhysRevE.85.026703
[28] Pompe, E., Holmes, C. and Łatuszyński, K. (2020). A framework for adaptive MCMC targeting multimodal distributions. *Ann. Statist.* **48** 2930–2952. MR4152629 https://doi.org/10.1214/19-AOS1916
[29] Roberts, G.O. and Rosenthal, J.S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.* **44** 458–475. MR2340211 https://doi.org/10.1239/jap/1183667414
[30] Roberts, G.O. and Rosenthal, J.S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.* **18** 349–367. MR2749836 https://doi.org/10.1198/jcgs.2009.06134
[31] Roberts, G.O. and Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.* **4** 337–357. MR2002247 https://doi.org/10.1023/A:1023562417138
[32] Roberts, G.O. and Tweedie, R.L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2** 341–363. MR1440273 https://doi.org/10.2307/3318418

[33] Vanetti, P., Bouchard-Côté, A., Deligiannidis, G. and Doucet, A. (2017). Piecewise-Deterministic Markov Chain Monte Carlo. Available at arXiv:1707.05296.

[34] Vialaret, M. and Maire, F. (2020). On the convergence time of some non-reversible Markov chain Monte Carlo methods. *Methodol. Comput. Appl. Probab.* **22** 1349–1387. MR4129137 https://doi.org/10.1007/s11009-019-09766-w

[35] Wallin, J. and Bolin, D. (2018). Efficient adaptive MCMC through precision estimation. *J. Comput. Graph. Statist.* **27** 887–897. MR3890878 https://doi.org/10.1080/10618600.2018.1459303

Received December 2020 and revised September 2021