IMA Journal of Numerical Analysis (2022) Page 1 of 30
doi:10.1093/imanum/drnxxx

Nonreversible Langevin Samplers: Splitting Schemes, Analysis and Implementation

A.B. DUNCAN [†], G.A. PAVLIOTIS [‡]
Department of Mathematics, Imperial College London
London SW7 2AZ, UK

AND K.C. ZYGALAKIS [§]
School of Mathematics, University of Edinburgh,
Edinburgh, EH9 3FD, UK

[Received on 16 August 2022]

For a given target density π on $\mathbb{R}^d$, there exist infinitely many diffusion processes that are ergodic with respect to π and that can be used in order to sample from this distribution. As observed in a number of papers Lelièvre *et al.* (2013); Duncan *et al.* (2016); Rey-Bellet & Spiliopoulos (2015a,b) samplers based on nonreversible diffusion processes can significantly outperform their reversible counterparts both in terms of reducing the asymptotic variance as well in increasing the rate of convergence to equilibrium. In this paper, we take advantage of this observation in order to construct efficient sampling algorithms based on the Lie-Trotter decomposition of a nonreversible diffusion process into reversible and nonreversible components. We show that samplers based on this scheme can significantly outperform standard MCMC methods, at the cost of introducing some controlled bias. In particular, we prove that numerical integrators constructed according to this decomposition are geometrically ergodic. Moreover we characterize fully their asymptotic bias and variance by analysing the solution of a discrete Poisson equation, and show that the samplers inherit the good mixing properties of the underlying nonreversible diffusion. This is illustrated further with a number of numerical examples ranging from highly correlated low dimensional distributions, to logistic regression problems in high dimensions as well as inference for spatial models.

Keywords: Langevin equation, non-reversible, Markov chain Monte Carlo

1. Introduction

Consider the problem of computing expectations with respect to a probability distribution with smooth density $\pi(x)$, known only up to the normalization constant, i.e. we wish to evaluate

$$\pi(f) = \int_{\mathbb{R}^d} f(x)\pi(x)\,dx. \tag{1.1}$$

For high dimensional distributions, deterministic techniques are no longer tractable. On the other hand, probabilistic methods do not suffer the same curse of dimensionality and thus are often the method of choice. One such approach is *Markov Chain Monte Carlo* (MCMC) which is based on the construction of a Markov process on $\mathbb{R}^d$ whose unique invariant distribution is $\pi(x)$. Due to their simplicity

[†] Email: a.duncan@imperial.ac.uk)
[‡] Email: g.pavliotis@imperial.ac.uk
[§] Email: k.zygalakis@ed.ac.uk

and wide applicability, Markov chains based on Metropolis-Hastings (MH) transition kernels Hastings (1970); Metropolis *et al.* (1953) and their numerous variants remain the most widely used scheme for sampling from a general target probability distribution, despite having been introduced over 60 years ago. As there are infinitely many Markov processes which are ergodic with respect to a given target distribution π, a natural question is whether a Markov process can be chosen which is more efficient, in terms of accelerating convergence to equilibrium and improving mixing. Metropolized schemes are reversible Markov chains by construction. It is a well documented fact that nonreversible chains convergence to equilibrium faster than reversible ones Neal (2004); Diaconis *et al.* (2000); Mira & Geyer (2000) and have a smaller asymptotic variance. Various MCMC schemes have been proposed which are based on the general idea of breaking reversibility by introducing an augmented target measure on an extended state space, along with dynamics which is invariant with respect to the augmented target measure. For discrete state spaces, the lifting method Diaconis *et al.* (2000); Hukushima & Sakai (2013); Turitsyn *et al.* (2011) is one such approach, where the Markov chain is "lifted" from the state space E to $E\times\{1,-1\}$. The transition probabilities in each copy of E are modified by introducing transitions between the copies to preserve the invariant distribution but now promote the sampler to generate long trajectories. For continuous state spaces, analogous approaches involve augmenting the state space with a velocity/momentum variable and constructing Makovian dynamics which are able to mix more rapidly in the augmented state space. Such methods include Hybrid Monte Carlo (HMC) methods, inspired by Hamiltonian dynamics. While the standard construction of HMC Duane *et al.* (1987); Neal (2011) is reversible, it is straightforward to construct dynamics based on the Generalized HMC scheme Horowitz (1991) which will not be reversible, see also Ottobre *et al.* (2016) and more recently Ma *et al.* (2016).

Deferring issues of simulation until later, another candidate Markov process for sampling from the distribution π is the diffusion process $(X_t)_{t\geqslant 0}$ defined by the following Itô stochastic differential equation (SDE):

$$dX_t = b(X_t)\,dt + \sqrt{2}\,dW_t, \tag{1.2}$$

where W_t is a standard $\mathbb{R}^d$–valued Brownian motion and $b:\mathbb{R}^d\to\mathbb{R}^d$ is a smooth vector field which satisfies

$$b(x) = \nabla\log\pi(x) + \gamma(x), \quad \nabla\cdot(\pi(x)\gamma(x)) = 0, \tag{1.3}$$

for some smooth vector field γ on $\mathbb{R}^d$ satisfying some mild assumptions (c.f. Proposition 2.2). It is a well known fact that the process X_t is reversible if and only if the vector field γ vanishes, $\gamma = 0$, see (Pavliotis, 2014, Ch. 4).

By the Birkhoff ergodic theorem,

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T f(X_s)\,ds = \mathbb{E}_\pi[f] =: \pi(f), \quad f\in L^1(\pi),$$

and thus one can use

$$\pi_T(f) := \frac{1}{T}\int_0^T f(X_s)\,ds$$

as an estimator for $\pi(f)$, for T sufficiently large. A natural way to measure the efficiency of such estimator is the mean square error (MSE) given by

$$\mathrm{MSE}(T) := \mathbb{E}|\pi_T(f)-\pi(f)|^2. \tag{1.4}$$

Under appropriate conditions on X_t and f, the estimator $\pi_T(f)$ will satisfy a *central limit theorem*, i.e.

$$\lim_{T\to+\infty}\sqrt{T}\,(\pi_T(f)-\pi(f)) = \mathscr{N}(0, 2\sigma^2(f)), \tag{1.5}$$

where $\sigma^2(f)$ is the *asymptotic variance* of the estimator $\pi_T(f)$ which can be expressed by

$$\sigma^2(f) := \langle \phi, (-\mathscr{L})\phi \rangle_\pi, \tag{1.6}$$

where $\mathscr{L}$ is the infinitesimal generator of (1.2) and ϕ is the mean zero solution of the following Poisson equation on $\mathbb{R}^d$,

$$-\mathscr{L}\phi = f - \pi(f). \tag{1.7}$$

The mean square error MSE (1.4) can be naturally decomposed it in terms of *bias* $\mu_T(f)$ and *variance* $\sigma_T^2(f)$ as follows

$$\mathbb{E}|\pi_T(f) - \pi(f)|^2 = (\mathbb{E}\pi_T(f) - \pi(f))^2 + \mathbb{E}(\pi_T(f) - \mathbb{E}\pi_T(f))^2 = (\mu_T(f))^2 + \sigma_T^2(f).$$

For large T, the variance satisfies $\sigma_T^2(f) \simeq T^{-1}\sigma^2(f)$, while $(\mu_T(f))^2 = o(T^{-1})$. Since $\gamma(x)$ is not uniquely defined in (1.3), i.e. there are infinitely many solutions to the partial differential equation $\nabla \cdot (\gamma\pi) = 0$, a natural question is how it should be chosen to ensure that for a given time T, the MSE in (1.4) is as small as possible. This can be achieved in two manners, the first by maximising the rate of convergence to equilibrium of (1.2) as was considered in Lelièvre *et al.* (2013); Wu *et al.* (2014). In general, constructing a nonreversible flow γ by which to maximise the rate of convergence in $L^2(\pi)$ is challenging, even for Gaussian target measures. An alternative is to choose $\gamma(x)$ in such a way so as to reduce the asymptotic variance $\sigma^2(f)$ Duncan *et al.* (2016). It should be emphasised that the optimal choice will be different for each case, and will depend specifically on the observable f. In particular in Duncan *et al.* (2016); Rey-Bellet & Spiliopoulos (2015a,b), it was shown that the choice $\gamma(x) = 0$, which corresponds to using reversible dynamics, gives the maximum value of asymptotic variance for a given choice of diffusion tensor. More precisely, introducing a nonreversible perturbation will never decrease the performance of an estimator based on Langevin dynamics, both in terms of convergence to equilibrium and asymptotic variance.

In general (1.2) cannot be simulated exactly, and one typically resorts to a discretisation of the SDE, denoted by $\widehat{X}_n^{\Delta t}$, in order to approximate $\pi(f)$. In particular, the following ergodic average is used

$$\widehat{\pi}_T^{\Delta t}(f) := \frac{1}{N}\sum_{k=0}^{N} f(\widehat{X}_k^{\Delta t}), \quad N\Delta t = T. \tag{1.8}$$

Extra caution has to be taken in order to ensure that the above quantity converges in the limit of $T \to \infty$ since even if (1.2) is ergodic (or even exponentially ergodic), this will not necessarily be the case for its numerical discretisation Roberts & Stramer (2002); Stramer & Tweedie (1999a,b). In addition, even when the numerical discretization is ergodic and thus

$$\lim_{T\to\infty} \widehat{\pi}_T^{\Delta t}(f) = \widehat{\pi}^{\Delta t}(f) = \int_{\mathbb{R}^d} f(x)\widehat{\pi}^{\Delta t}(x)dx, \tag{1.9}$$

it is not true in general that $\widehat{\pi}^{\Delta t} = \pi$, since the underlying numerical discretization introduces bias in the estimation of $\pi(f)$ (see Talay & Tubaro (1990); Abdulle *et al.* (2014, 2015)). However, as discussed in Section 4.3, this bias tends to 0 as $\Delta t \to 0$ under appropriate conditions on the numerical integrator. One way to eliminate such bias is through Metropolization Smith & Roberts (1993); Tierney (1994), *i.e.* the introduction of an accept-reject step that ensures that the corresponding Markov chain is ergodic with respect to the target distribution π. However, such bias elimination might not be advantageous in practice since the Metropolised chain will be reversible by construction, thus eliminating any benefit introduced by the nonreversible perturbation γ. When computing expectations of distributions with expensive likelihoods, it might be too costly to sample a long Markov chain trajectory. If an appropriate nonreversible Langevin dynamics (1.2) can be introduced which does give rise to a dramatic reduction in asymptotic variance, then it might be advantageous to permit a controlled amount of bias in exchange for needing to sample far less. This bias-variance tradeoff, in the context of numerical discretisations of (1.2) is the subject of study of this paper.

In recent years, several Langevin-type sampling schemes have been proposed that are different from the standard overdamped Langevin dynamics and for which it is possible to prove that they have better properties, in the sense that they converge faster to the target distribution and that the asymptotic variance is smaller. A partial list of such modified Langevin samplers is presented in Duncan *et al.* (2017). It is important to note, however, that it is not a priori clear that the discretized diffusion will inherit the advantageous properties of the continuous time process. Therefore, great care has to be taken in order to discretize the modified Langevin dynamics in a way that preserves that optimal properties of the SDE. The main goal of this paper is to address this issue for the class of nonreversible Langevin dynamics that were introduced in Hwang *et al.* (2005) and analyzed in e.g. Lelièvre *et al.* (2013); Hwang *et al.* (2015); Rey-Bellet & Spiliopoulos (2015a,b). In particular, we present a complete analysis of the performance of splitting schemes for simulating nonreversible Langevin SDEs that are ergodic with respect to a given target distribution.

In particular, we will consider discretizations based on a Lie-Trotter splitting between the reversible and the nonreversible part of the dynamics. More specifically, we consider integrators of the form

$$\widehat{X}_{n+1}^{\Delta t} = \Theta_{\Delta t} \circ \Phi_{\Delta t}(\widehat{X}_n^{\Delta t}), \tag{1.10}$$

where $\Phi_{\Delta t}(x)$ is a integrator that approximates the flow map corresponding to the deterministic dynamics

$$\frac{dx_t}{dt} = \gamma(x_t), \tag{1.11}$$

and $\Theta_{\Delta t}(x)$ which approximates the reversible dynamics

$$dx_t = \nabla \log \pi(x_t) dt + \sqrt{2} dW_t. \tag{1.12}$$

In this paper we shall focus on the specific case when the reversible dynamics is simulated using a Metropolized scheme, while the nonreversible dynamics are simulated using a high-order ODE integrator. We mention here that this splitting idea has also been used recently in Poncet (2017) to construct a non-reversible sampler with no bias. This however, comes with the cost of having to solve (1.11) using an implicit integrator. Furthermore, in Futami *et al.* (2020, 2021) a non-Metropolised ensemble version that discretises directly the non-reversible dynamics was proposed and studied for its non-asymptotic convergence properties.

The choice of $\Phi_{\Delta t}, \Theta_{\Delta t}$ has a fundamental influence on the bias, asymptotic variance and stability of the resulting sampler. In particular, if one chooses $\Phi_{\Delta t}$ to be a Metropolised integrator Bou-Rabee & Hairer (2012) then, similarly to the result in Abdulle *et al.* (2015), the order of convergence of the deterministic integrator $\Theta_{\Delta t}$ provides a lower bound for the difference between expectations with respect to $\widehat{\pi}^{\Delta t}$ and π. However, this is not the case for the numerical asymptotic variance $\widehat{\sigma}_{\Delta t}^2(f)$, since even though we can show that it is a perturbation of $\sigma^2(f)$ the difference will depend crucially on the choice of $\Theta_{\Delta t}$. These results are important as they allow to choose the correct combination of dynamics and numerical scheme that drastically reduces the computational cost required to achieve a given tolerance of error.

In summary, the main contributions of this paper are:

1. proving geometric ergodicity for the Markov chain given by (1.10) for a variety of different numerical integrators applied to the reversible part;

2. a complete characterisation of the asymptotic bias of (1.10);

3. showing that, by completely characterising the asymptotic variance, numerical integrators of the type (1.10) inherit the asymptotic variance benefits of the non reversible SDE (1.2);

4. exhibiting the potential of using nonreversible integrators for sampling as illustrated from a number of different numerical experiments on inference for spatial models as well as real data sets.

We mention here that the proof of the geometric ergodicity uses the approach described in Meyn & Tweedie (1993a), while the characterisation of the asymptotic bias uses the framework developed in Abdulle *et al.* (2014). Additionally, the characterisation of the asymptotic variance relies heavily on the analysis of the discrete Poisson equation associated with the splitting scheme. A similar analysis was carried out in Mijatović & Vogrinc (2018) and has also recently been used to analyse the asymptotic variance of random walk Metropolis chains Mijatović & Vogrinc (2019).

The rest of the paper is organised as follows. In Section 2 we describe some known theoretical results for the SDE (1.2) which are necessary for the development of this paper. In Section 3 we identify sufficient conditions to guarantee geometric ergodicity of the Lie-Trotter splitting scheme (1.10) on $\mathbb{R}^d$. In Section 4 we study the asymptotic properties of a class of numerical integrators for (1.2) for which the Lie-Trotter scheme is a special case. In particular we derive perturbative expansions for the asymptotic bias and variance. In Section 5 we apply these results to characterise the asymptotic bias and variance of the Lie-Trotter scheme on the bounded domain $\mathbb{T}^d$. To demonstrate the efficacy of the irreversible schemes, in Section 6 we present a number of numerical experiments on inference for spatial models as well as on Bayesian logistic regression. Finally, a discussion of the results presented in this paper and potential future research directions can be found in Section 7.

2. Properties of Overdamped Langevin Diffusions

In this section we discuss different known theoretical results that are useful for understanding the main results of the paper. We start by listing the assumptions we shall make on π and the SDE (1.2) to ensure ergodicity.

Assumption 2.1

1. The measure π possesses a positive smooth density $\pi(x) > 0$, known up to a normalizing constant, such that $\pi \in L^1(\mathbb{R}^d)$.

2. The drift vector $b : \mathbb{R}^d \to \mathbb{R}^d$ of (1.2) is smooth and satisfies (1.3) with $\gamma : \mathbb{R}^d \to \mathbb{R}^d$ being a smooth vector field with components in $L^1(\pi)$.

The following result provides necessary and sufficient conditions on the coefficients of (1.2) to ensure that X_t possesses a unique stationary distribution π.

PROPOSITION 2.2 Suppose that Assumptions 2.1 hold. Then the diffusion process X_t defined by (1.2) possesses a strongly continuous semigroup $(P_t)_{t\geqslant 0}$ on $L^2(\pi)$ defined by

$$P_t f(x) = \mathbb{E}[f(X_t)\,|X_0 = x]. \tag{2.1}$$

The associated infinitesimal generator is an an extension of

$$\mathscr{L} = \frac{1}{\pi}\nabla\cdot(\pi\nabla\cdot) + \gamma\cdot\nabla \tag{2.2}$$

with core $C_c^\infty(\mathbb{R}^d)$. Moreover, P_t has unique invariant distribution π. Conversely, given a diffusion process of the form (1.2) which is invariant with respect to π, then the drift b necessarily satisfies (1.3).

Proof. The first part of this result is a direct application of (Lorenzi & Bertoldi, 2006, Thm 8.1.26). The converse implication can be checked using integration by parts. □

While many choices for γ are possible (see Ma *et al.* (2015) for a more complete recipe) a natural family of vector fields is given by $\gamma(x) = J\nabla\Phi(\pi(x))$, where Φ is a smooth function satisfying $\nabla\Phi(\pi(\cdot)) \in L^1(\pi)$ and J is $d\times d$ skew-symmetric matrix. We shall focus specifically on the following three choices:

1. If π satisfies $\int_{\mathbb{R}^d}|\nabla\log\pi(x)|\pi(dx) < \infty$, then the vector field

$$\gamma(x) = J\nabla\log\pi(x), \quad J = -J^\top, \tag{2.3}$$

satisfies condition (1.3). This was the choice which was studied in Duncan *et al.* (2016).

2. If $\int_{\mathbb{R}^d}|\nabla\log\pi(x)|\pi^{1+\alpha}(dx) < \infty$ for some $\alpha > 0$ then another natural choice for the vector field is given by

$$\gamma(x) = J\nabla\pi^\alpha(x), \quad J = -J^\top. \tag{2.4}$$

Although (2.4) introduces an additional tuning parameter α, one might prefer this choice as it coincides with the intuition that when far away from the modes the sampler should move towards the modes as quickly as possible, and should only undergo these deterministic meanders in regions of high probability.

3. Let $\Psi : \mathbb{R}\to\mathbb{R}$ be a smooth, compactly supported function. Then

$$\gamma(x) = J\nabla\log\pi(x)\Psi(\pi(x)), \quad J = -J^\top, \tag{2.5}$$

will always satisfy (1.3). Moreover, if π has compact level sets, then γ will also be compactly supported on $\mathbb{R}^d$.

Applying the results detailed in Glynn & Meyn (1996); Meyn & Tweedie (1993b), we shall assume that the process X_t possesses a Lyapunov function, which is sufficient to ensure the exponential ergodicity of X_t, as detailed in the subsequent proposition.

Assumption 2.3 (Foster–Lyapunov Criterion) There exists a function $V : \mathbb{R}^d \to \mathbb{R}$ and constants $c > 0$ and $b \in \mathbb{R}$ such that

$$\mathscr{L}V(x) \leqslant -cV(x) + b\mathbf{1}_C, \text{ and } V(x) \geqslant 1, \quad x \in \mathbb{R}^d, \tag{2.6}$$

where $\mathbf{1}_C$ is the indicator function over a *petite set*.

For the definition of a petite set we refer the reader to Meyn & Tweedie (1993a). For the generator $\mathscr{L}$ corresponding to the process (1.2) compact sets are always petite. The exponential ergodicity of X_t follows from the following proposition (see also Mattingly *et al.* (2002); Meyn & Tweedie (1993a)).

PROPOSITION 2.4 Suppose that Assumption 2.3 holds, then there exist constants $C > 0$ and $\lambda > 0$ such that:

$$|P_t f(x) - \pi(f)| \leqslant CV(x)e^{-\lambda t}, \quad x \in \mathbb{R}^d, \tag{2.7}$$

for all f satisfying $|f| \leqslant V$.

Moreover, the Foster-Lyapunov criterion also provides a sufficient condition for the Poisson equation (1.7) to be well-posed, and thus for the central limit theorem (1.5) to hold.

PROPOSITION 2.5 (Glynn & Meyn, 1996, Theorem 4.3) Suppose that Assumption 2.3 holds and that $\pi(U^2) < \infty$, then for any function f such that $|f| \leqslant U$ and for any initial distribution, the central limit theorem (1.5) holds, i.e. $\sqrt{T}(\pi_T(f) - \pi(f))$ converges weakly to a $\mathscr{N}(0, 2\sigma^2(f))$–distributed random variable, with

$$\sigma^2(f) = \int_{\mathbb{R}^d} \phi(x)(-\mathscr{L})\phi(x)\pi(x)\,dx,$$

where ϕ is the unique mean zero solution to the Poisson equation (1.7). Moreover the solution ϕ can be expressed as

$$\phi = \int_0^\infty [P_t f - \pi(f)]\,dt.$$

The following lemma provides a sufficient condition on π for (1.2) to possess a Lyapunov function. It is a slight generalisation of a similar result from Roberts & Tweedie (1996), extended to also apply in the case of nonreversible diffusion processes.

LEMMA 2.1 (Roberts & Tweedie, 1996, Theorem 2.3) Consider the process X_t defined by (1.2) with drift coefficient b satisfying (1.3) . Suppose that π is bounded, there exists $0 < \delta < 1$ such that,

$$\liminf_{|x|\to\infty}\left((1-\delta)|\nabla \log \pi(x)|^2 + \Delta \log \pi(x)\right) > 0, \tag{2.8}$$

and the vector field γ satisfies

$$\nabla \cdot \gamma(x) = 0, \quad x \in \mathbb{R}^d. \tag{2.9}$$

Then the Foster–Lyapunov criterion holds for (1.2) with $U(x) = \pi^{-\delta}(x)$ and moreover $\pi(U) < \infty$.

REMARK 2.1 Note that when $\gamma(x) = J\nabla\Phi(\pi(x))$ equation (2.9) is automatically satisfied. Hence the choices of choices of γ specified by (2.3), (2.4) and (2.5) all satisfy (2.9).

3. Stochastic Stability of the splitting scheme on $\mathbb{R}^d$

In this section we identify sufficient conditions under which the Lie-Trotter scheme on $\mathbb{R}^d$ is geometrically ergodic with respect to an invariant distribution $\widehat{\pi}^{\Delta t}$ which will be a perturbation of π. In general, a discretization of the ergodic diffusion process (1.2) need not be ergodic, geometric or otherwise, see for example Roberts & Tweedie (1996) . For the splitting scheme we shall show that provided the approximate nonreversible flow $\Phi_{\Delta t}$ is sufficiently weak away from the origin, the process (1.10) will inherit the geometric ergodicity from the reversible dynamics. We follow Meyn and Tweedie Meyn & Tweedie (1993a) to demonstrate geometric ergodicity of $\left(\widehat{X}_n^{\Delta t}\right)_{n\in\mathbb{N}}$. Consider the reversible process defined by

$$Z_{n+1}^{\Delta t} = \Theta_{\Delta t} Z_n^{\Delta t}, \tag{3.1}$$

and $\widetilde{P}_{\Delta t}$ be the corresponding transition semigroup. We shall assume that the reversible dynamics are a Metropolis-Hastings chain, with proposal kernel $q_{\Delta t}(\cdot|x)$. More specifically, given $x \in \mathbb{R}^d$, $\Theta_{\Delta t}(x)$ is constructed as follows

1. Sample $y \sim q_{\Delta t}(\cdot\,|x)$.
2. With probability
$$\alpha(x,y) = \min\left(1, \frac{\pi(y)q_{\Delta t}(x|y)}{\pi(x)q_{\Delta t}(y|x)}\right), \tag{3.2}$$
set $\Theta_{\Delta t}x := y$ otherwise $\Theta_{\Delta t}x := x$.

It is well known that the target distribution π is invariant under the map $\Theta_{\Delta t}$ Metropolis *et al.* (1953); Hastings (1970). In this paper, we shall focus on two specific proposals, namely the Langevin proposal

$$q_{\Delta t}(\cdot\,|x) = x + \Delta t \nabla \log \pi(x) + \sqrt{2\Delta t}g, \tag{3.3}$$

and the random walk proposal

$$q_{\Delta t}(\cdot\,|x) = x + \sqrt{2\Delta t}g, \tag{3.4}$$

where g is a standard d-dimensional Gaussian random variable. The resulting scheme is known as *Metropolis-Adjusted Langevin Algorithm* (MALA) when proposal (3.3) is used, and *Random Walk Metropolis Hastings* (RWMH) when (3.4) is used. Denote by $\widehat{P}_{\Delta t}(x,\cdot)$ and $\widetilde{P}_{\Delta t}(x,\cdot)$ the transition distribution functions of the splitting scheme (1.10) and (3.1) respectively. Then clearly,

$$\widehat{P}_{\Delta t}f(x,A) = (\widetilde{P}_{\Delta t}f)(\Phi_{\Delta t}(x),A), \quad A \in \mathscr{B}(\mathbb{R}^d).$$

Following the approach of Mengersen & Tweedie (1996) we first show that (1.10) is a π-irreducible, aperiodic Markov chain. Moreover, we will show that all compact sets are small, i.e. for every compact set C, there exists a $\delta > 0$ and $n > 0$ such that

$$\widehat{P}^n_{\Delta t}(x,\cdot) \geqslant \delta \nu(\cdot), \quad x \in C.$$

Finally, we will show that if a Foster-Lyapunov condition holds for the reversible dynamics $\widetilde{P}_{\Delta t}$, then it also holds for $\widehat{P}_{\Delta t}$. To this end, we shall make the following assumptions.

Assumption 3.1 For Δt sufficiently small, we assume that

1 The reversible chain (3.1) satisfies a Foster-Lyapunov condition, i.e. there exists a continuous function $V \geqslant 1$, a compact set $C \subset \mathbb{R}^d$ and constants $\lambda \in (0,1)$ and $b \geqslant 0$ such that
$$\widetilde{P}_{\Delta t}V(x) \leqslant \lambda V(x) + b\mathbf{1}_C(x), \quad x \in \mathbb{R}^d. \tag{3.5}$$

2 The nonreversible flow map $\Phi_{\Delta t}$ satisfies the following condition,
$$\lim \sup_{|x|\to\infty} \frac{V(\Phi_{\Delta t}(x)) - V(x)}{V(x)} < \frac{1}{\lambda} - 1. \tag{3.6}$$

3 The preimage $\Phi^{-1}_{\Delta t}(C)$ is bounded.

The main theorem of this section establishes the geometric ergodicity of (1.10).

THEOREM 3.2 Suppose that Assumptions 3.1 hold, and that π and $q_{\Delta t}(y|x)$ are positive and continuous for all $x, y \in \mathbb{R}^d$. Then for Δt sufficiently small, the process $\widehat{X}^{\Delta t}_n$ is geometrically ergodic, i.e. there exists $\rho \in (0,1)$ and $K > 0$ such that

$$\sup_{|g|\leqslant V} \left| \int_{\mathbb{R}^d} g(y) \left(\widehat{P}^n_{\Delta t}(x,y) - \hat{\pi}_{\Delta t}(y)\right) dy \right| \leqslant KV(x)\rho^n, \quad n \in \mathbb{N}.$$

Proof. The proof of this theorem can be found in the supplementary material. □ The following result is an application of Theorem 3.2 for the Random Walk proposal (3.4).

COROLLARY 3.1 (Geometric Ergodicity of Lie-Trotter scheme with RWMH dynamics) Consider the Lie-Trotter splitting scheme $\hat{X}_n^{\Delta t}$ where the reversible dynamics (1.12) are simulated using a RWMH scheme with proposal defined by (3.4). Suppose that the conditions on π and $q_{\Delta t}$ specified in (Roberts *et al.*, 1998, Theorem 3.2) hold and moreover that

$$\lim_{|x|\to\infty} \left(|\Phi_{\Delta t}(x)| - |x|\right) = 0, \tag{3.7}$$

for Δt sufficiently small. Then $\hat{X}_n^{\Delta t}$ is geometrically ergodic.

An almost identical result holds for the MALA proposal (3.3).

COROLLARY 3.2 (Geometric Ergodicity of Lie-Trotter scheme with MALA dynamics) Consider the Lie-Trotter splitting scheme $\widehat{X}_n^{\Delta t}$ where the reversible dynamics (1.12) are simulated using a MALA scheme with proposal defined by (3.3). Suppose that the conditions on π and $q_{\Delta t}$ specified in (Roberts & Tweedie, 1996, Theorem 4.1) hold and moreover that (3.7) holds for Δt sufficiently small. Then $\widehat{X}_n^{\Delta t}$ is geometrically ergodic.

In particular, suppose that $\lim_{|x|\to\infty}\pi(x) \to 0$, and that, given $\alpha > 0$, there exist positive constants α', K_1 and K_2 such that

$$|\nabla\pi^{\alpha}(x)| \leqslant K_1\pi^{\alpha'}(x), \quad |\nabla\nabla\pi^{\alpha}(x)|_{max} \leqslant K_2, \quad x \in \mathbb{R}^d, \tag{3.8}$$

where $|\cdot|_{max}$ denotes the max norm. If $\gamma = J\nabla\pi^{\alpha}$ for J antisymmetric, then condition (3.7) will hold if $\Phi_{\Delta t}(x)$ is simulated using an explicit Euler or Runge-Kutta scheme. A similar result holds for γ given by (2.5).

4. Asymptotic Bias and Variance Estimates for general integrators

In this section we consider the asymptotic behaviour of the estimator (1.8) for $\pi(f)$, obtained for a general numerical scheme $(\widehat{X}_k^{\Delta t})_{k\geqslant 0}$. In particular, we shall derive estimates for the asymptotic bias and asymptotic variance of the estimator $\widehat{\pi}^{\Delta t}(f)$. For simplicity we shall focus on the case where the domain is $\mathbb{T}^d$, i.e. the unit hypercube with periodic boundary conditions. As in Mattingly *et al.* (2010) this set-up greatly simplifies the derivation of expressions for bias and variance, particularly since remainder terms arising from Taylor expansions can be easily controlled. We expect that extending these results to unbounded domains should be possible by following analogous approaches in Kopec (2014). Throughout this section, we shall assume that the numerical integrator $\widehat{X}_k^{\Delta t}$ is ergodic, with unique invariant distribution $\widehat{\pi}^{\Delta t}$.

4.1 *Notation*

We first introduce the notation which will be used in this section and the remainder of the paper. Given a probability measure μ on $(\mathbb{T}^d, \mathscr{B}(\mathbb{T}^d))$ define $L^2(\mu)$ to be the Hilbert space of square integrable functions on $\mathbb{T}^d$, equipped with inner product $\langle\cdot,\cdot\rangle_\mu$ and norm $\|\cdot\|_{L^2(\pi)}$. The subspace $L_0^2(\mu)$ of $L^2(\mu)$ is defined to be

$$L_0^2(\mu) = \{f \in L^2(\mu) : \mu(f) = 0\}, \tag{4.1}$$

We define $L^\infty(\mu)$ (also denoted by $L^\infty(\mathbb{T}^d)$) to be the Banach space of essentially bounded functions on $\mathbb{T}^d$ equipped with norm $\|\cdot\|_{L^\infty(\mathbb{T}^d)}$. The subspace $L^\infty_0(\mu)$ of $L^\infty(\mu)$ is defined analogously to (4.1). Finally, given a (signed) measure ν on $(\mathbb{T}^d, \mathscr{B}(\mathbb{T}^d))$ we denote the total variation norm of ν by $\|\nu\|_{TV}$.

4.2 *Backward error analysis for ODEs*

Backward error analysis is a powerful tool for the analysis of numerical integrators for differential equations Sanz-Serna & Calvo (1994); Leimkuhler & Reich (2004); Hairer *et al.* (2006). In particular, it is the main ingredient for the proof of the good energy conservation (without drift) of symplectic Runge-Kutta methods when applied to deterministic Hamiltonian systems over exponentially long time intervals Hairer *et al.* (2006). In our context it is useful to characterize the infinitesimal generator of the numerical flow $\Phi_{\Delta t}$ approximating the solution of the ODE (1.11). Indeed, given a consistent integrator $z_{n+1} = \Phi_{\Delta t}(z_n)$ for the ODE

$$\frac{dz(t)}{dt} = f(z(t)), \tag{4.2}$$

the idea of backward error analysis is to search for a modified differential equation written as a formal series in powers of the stepsize Δt,

$$\frac{d\widetilde{z}}{dt} = f(\widetilde{z}) + \Delta t f_1(\widetilde{z}) + \Delta t^2 f_2(\widetilde{z}) + \ldots, \quad \widetilde{z}(0) = z_0 \tag{4.3}$$

such that (formally) $z_n = \widetilde{z}(t_n)$, where $t_n = n\Delta t$ (in the above differential equation, we omit the time variable for brevity). The numerical solution can thus be interpreted as a higher order approximation of the exact solution of a modified ODE. For all reasonable integrators, the vector fields f_j can be constructed inductively Leimkuhler & Reich (2004); Hairer *et al.* (2006), starting from $f_0 = f$. In general, the series in (4.3) will diverge for nonlinear systems, and thus needs to be truncated. We thus consider the truncated modified ODE at order s

$$\frac{d\widetilde{z}}{dt} = f(\widetilde{z}) + \Delta t f_1(\widetilde{z}) + \Delta t^2 f_2(\widetilde{z}) + \ldots + \Delta t^s f_s(\widetilde{z}), \quad \widetilde{z}(0) = z_0. \tag{4.4}$$

One can then show that $z_n = \widetilde{z}(t_n) + \mathscr{O}(\Delta t^{s+1})$ for $\Delta t \to 0$ for bounded times $t_n = n\Delta t \leqslant T$. We note that the flow $\widetilde{\Phi}_{\Delta t}(z)$ of the modified differential equation (4.4) satisfies

$$\phi \circ \widetilde{\Phi}_{\Delta t} = \left(\sum_{k=0}^{M} \frac{\Delta t^k \widetilde{\mathscr{L}}_D^k}{k!}\right)\phi + \mathscr{O}(\Delta t^{M+1}), \qquad \widetilde{\mathscr{L}}_D = F_0 + \Delta t F_1 + \Delta t^2 F_2 + \ldots + \Delta t^s F_s, \tag{4.5}$$

for all $M \geqslant 0$, and smooth test functions ϕ, and where $F_j\phi = f_j \cdot \nabla\phi,\ j = 0, \ldots, s$ and $f_0 = f$[1].

4.3 *Asymptotic bias of numerical integrators*

The aim of this subsection is to describe the conditions on a numerical integrator for (1.2) which are sufficient for the numerical invariant distribution $\widehat{\pi}^{\Delta t}$ to approximate π to order r in the weak sense.

[1] For all Δt small enough, the sum in (4.5) can be shown to converge for $M \to \infty$ in the case of analytic vector fields f_j (and analytic test functions ϕ), which permits to remove the $\mathscr{O}$ remainder.

These conditions relate directly to the expansion of one-step numerical expectations in powers of Δt. In particular, denote by $\widehat{P}_{\Delta t}$ the transition semigroup associated with $\widehat{X}^{\Delta t}$, i.e.

$$\widehat{P}_{\Delta t} f := \mathbb{E}\left[f(\widehat{X}_1^{\Delta t})|X_0 = x\right]$$

and assume that the following expansion holds

$$\widehat{P}_{\Delta t} f = f + \Delta t A_0 f + \ldots + \Delta t^k A_{k-1} f + \Delta t^{k+1} A_k f + \Delta t^q Q_{f,\Delta t}, \quad q > k+1, \tag{4.6}$$

where $A_i, i = 0, 1, \cdots, k$ are linear differential operators with coefficients depending smoothly on $\pi(x)$, its derivatives, and the choice of the numerical integrator. In addition $Q_{f,\Delta t}$ is a smooth remainder term depending both on f and Δt while being uniformly bounded with respect to Δt. The following theorem provides sufficient conditions for expectations with respect to $\widehat{\pi}^{\Delta t}$ to approximate expectations with respect to π to order r.

THEOREM 4.1 Consider equation (1.2) solved by a numerical scheme which is ergodic with respect to some probability measure $\widehat{\pi}^{\Delta t}$ and such that the one step transition semigroup satisfies (4.6) with

$$A_j^* \pi = 0, \quad \text{for} \quad j = 1, \cdots, r-1, \tag{4.7}$$

where $q > r$. Then one obtains

$$\int_{\mathbb{T}^d} f(x) \widehat{\pi}^{\Delta t}(dx) = \int_{\mathbb{T}^d} f(x)\pi(dx) + \Delta t^r \int_{\mathbb{T}^d} A_r(-\mathscr{L})^{-1}(f - \pi(f))\pi(dx) + \Delta t^q R_{f,\Delta t}, \tag{4.8}$$

where the remainder term $R_{f,\Delta t}$ is uniformly bounded with respect to Δt, for Δt sufficiently small.

Proof. The proof can be found in (Abdulle *et al.*, 2014, Theorem 3.1). □

REMARK 4.1 Integrators $\widehat{X}_n^{\Delta t}$ which have weak error order r will automatically satisfy condition (4.7) for $j = 0, \ldots, r-1$. However, the converse is not necessarily true, see Abdulle *et al.* (2014) for further discussion.

Since the ergodic average (1.8) satisfies $\widehat{\pi}_{N\Delta t}^{\Delta t}(f) \to \widehat{\pi}^{\Delta t}(f) = \int_{\mathbb{T}^d} f(x)\pi(x)\,dx$ as $N \to \infty$, it follows immediately from Theorem 4.1 that, for sufficiently small Δt,

$$\lim_{N\to\infty} \widehat{\pi}_{N\Delta t}(f) = \pi(f) + \Delta t^r \int_{\mathbb{T}^d} A_r(-\mathscr{L})^{-1}(f - \pi(f))\pi(dx) + o(\Delta t^r),$$

provided (4.7) holds.

4.4 *Asymptotic variance of numerical integrators*

The aim of this subsection is to derive a perturbation expansion in the small timestep regime for the asymptotic variance of an arbitrary ergodic numerical integrator for the dynamics (1.2). To this end, we consider a diffusion X_t for which the central limit theorem (1.5) holds. Moreover, we shall make the following assumption, which implies that the corresponding numerical scheme $\widehat{X}_k^{\Delta t}$ converges to equilibrium exponentially fast in $L^\infty(\mathbb{T}^d)$, with rate which is uniform with respect to Δt.

Assumption 4.2 There exist constants $C > 0$ and $\lambda > 0$ independent of Δt such that, for Δt sufficiently small,

$$\left\|\widehat{P}_{\Delta t}^k f - \widehat{\pi}^{\Delta t}(f)\right\|_{L^\infty(\mathbb{T}^d)} \leqslant Ce^{-\lambda k \Delta t} \left\|f - \widehat{\pi}^{\Delta t}(f)\right\|_{L^\infty(\mathbb{T}^d)}, \quad f \in L^\infty(\mathbb{T}^d).$$

REMARK 4.2 This condition is nontrivial to verify in general. For the specific case of the Lie-Trotter integrator (1.10), when the reversible component of the dynamics is integrated either using MALA or random walk proposals, it is shown in Theorem 5.3 that Assumption 4.2 holds.

Given an observable $f \in C^{\infty}(\mathbb{T}^d)$ we consider $\widehat{\pi}_T^{\Delta t}$ as in (1.8). We define the rescaled asymptotic variance of the estimator $\widehat{\pi}_T^{\Delta t}$ as follows

$$\widehat{\sigma}^2_{\Delta t}(f) = \Delta t \lim_{N\to\infty} N \mathrm{Var}_{\widehat{\pi}^{\Delta t}}\left[\frac{1}{N}\sum_{k=0}^{N-1} f(\widehat{X}_k^{\Delta t})\right]. \tag{4.9}$$

Here we rescale the asymptotic variance with Δt, to guarantee a well–defined limit when $\Delta t \to 0$. Assumption 4.2 implies that there exists a constant $K > 0$, independent of Δt such that

$$\left\|\left[\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right]^{-1}\right\|_{L_0^{\infty}(\widehat{\pi}^{\Delta t})} < K, \tag{4.10}$$

for Δt sufficiently small. In particular, we can express (4.9) as

$$\widehat{\sigma}^2_{\Delta t}(f) = 2\Delta t\left\langle \left(f - \widehat{\pi}^{\Delta t}(f)\right), \left(I - \widehat{P}_{\Delta t}\right)^{-1}\left(f - \widehat{\pi}^{\Delta t}(f)\right)\right\rangle_{\widehat{\pi}^{\Delta t}} - \Delta t \mathrm{Var}_{\widehat{\pi}^{\Delta t}}[f]. \tag{4.11}$$

It should be clear from (4.11) that there will be two contributions to the error between $\widehat{\sigma}^2_{\Delta t}(f)$ and $\sigma^2(f)$: one arising from the order of weak convergence of the numerical method, and one from the time discreteness of the process $\widehat{X}_k^{\Delta t}$. Indeed, even when one considers the exact discrete time dynamics defined by

$$X_n^{\Delta t} = X(n\Delta t), \quad n \in \mathbb{N},$$

the error between the corresponding asymptotic variance $\sigma^2_{\Delta t}(f)$ and $\sigma^2(f)$ will be non-zero, despite the fact that both discrete and continuous time Markov processes have the same invariant distribution. To isolate the different sources of error, we present first Proposition 4.3 which quantifies the effect of the time-discreteness on the asymptotic variance. In Theorem 4.4 we then quantify the error between the asymptotic variances $\sigma^2_{\Delta t}(f)$ and $\widehat{\sigma}^2_{\Delta t}(f)$ of $X_n^{\Delta t}$ and $\widehat{X}_n^{\Delta t}$, respectively.

PROPOSITION 4.3 For all $f \in C^{\infty}(\mathbb{T}^d)$ there exists a smooth function R_f such that for Δt sufficiently small,

$$\sigma^2_{\Delta t}(f) = \sigma^2(f) - 2\Delta t \mathrm{Var}_{\pi}[f] + \frac{\Delta t^2}{6}\langle(-\mathscr{L})(f-\pi(f)), f - \pi(f)\rangle_{\pi} + \Delta t^2 R_f$$

where R_f is bounded, independent of Δt.

Proof. The proof can be found in Section .1. □

Define the operator $M_{\Delta t}$ to be the projector onto functions with mean zero with respect to $\widehat{\pi}^{\Delta t}$, i.e.

$$M_{\Delta t}\phi(x) = \phi(x) - \int_{\mathbb{T}^d}\phi(y)\widehat{\pi}^{\Delta t}(y)\,dy.$$

The following theorem characterises the difference between the asymptotic variance arising from the exact discrete time dynamics $X_n^{\Delta t}$ and the numerical integrator $\widehat{X}_n^{\Delta t}$.

THEOREM 4.4 Suppose that, for some $k \in \mathbb{N}$, $k \geqslant 1$, there exist operators $A_0, \ldots, A_k$ on $C^\infty(\mathbb{T}^d)$, bounded uniformly with respect to Δt, where $A_i = \frac{\mathscr{L}^{i+1}}{(i+1)!}, i = 0, \cdots, k-1$ and such that for all $\psi \in C^\infty(\mathbb{T}^d)$ the semigroup $\widehat{P}_{\Delta t}$ satisfies (4.6). Suppose that the corresponding invariant distribution $\widehat{\pi}^{\Delta t}$ satisfies

$$\int_{\mathbb{T}^d} \psi(x)\widehat{\pi}^{\Delta t}(x)\,dx = \int_{\mathbb{T}^d} \psi(x)\pi(x)\,dx + \Delta t^r R_\psi,$$

where $r > k$ and R_ψ is a smooth remainder term, uniformly bounded with respect to Δt. Moreover, suppose that $\widehat{P}_{\Delta t}$ satisfies (4.10). Then for all $f, g \in C^\infty(\mathbb{T}^d)$ such that $\pi(f) = \pi(g) = 0$, we have the expansion

$$\left\langle g, \left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f \right\rangle_\pi = \left\langle M_{\Delta t} g, \left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t} f \right\rangle_{\widehat{\pi}^{\Delta t}} + \Delta t^k R_1(f,g) + o(\Delta t^k), \tag{4.12}$$

where

$$R_1(f,g) = \left\langle \left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t}\left(\frac{\mathscr{L}^{k+1}}{(k+1)!} - A_k\right)\left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f, M_{\Delta t} g \right\rangle_{\widehat{\pi}^{\Delta t}}. \tag{4.13}$$

In particular

$$\widehat{\sigma}^2_{\Delta t}(f) = \sigma^2_{\Delta t}(f) + 2\Delta t^k R_1(f,f) + o(\Delta t^k). \tag{4.14}$$

Moreover, we can write the remainder term as

$$R_1(f,g) = \left\langle (-\mathscr{L})^{-1}\left(\frac{\mathscr{L}^{k+1}}{(k+1)!} - M_0 A_k\right)(-\mathscr{L})^{-1} f, g \right\rangle_\pi + o(\Delta t^k), \tag{4.15}$$

where $M_0\psi = \psi - \int_{\mathbb{T}^d} \psi(y)\pi(y)\,dy$.

Proof. The proof can be found in Section .1. □

To complete this analysis we shall consider the asymptotic variance arising from a perturbed diffusion process $\widetilde{X}_t$ having infinitesimal generator $\widetilde{\mathscr{L}_{\Delta t}}$ such that, for Δt sufficiently small

$$\widetilde{\mathscr{L}_{\Delta t}} f = \mathscr{L} f + \Delta t^k \mathscr{L}_k f + \Delta t^{q-1} R_f, \quad f \in C^\infty(\mathbb{T}^d), \tag{4.16}$$

where $q > k+1$. We shall also assume that $(\widetilde{L}_{\Delta t})^{-1}$ is bounded in $L^\infty_0(\widehat{\pi}^{\Delta t})$ uniformly with respect to Δt. More specifically there exists $K > 0$, independent of Δt such that

$$\left\| \left(-\widetilde{\mathscr{L}_{\Delta t}}\right)^{-1} \right\|_{L^\infty_0(\widehat{\pi}^{\Delta t})} < K, \tag{4.17}$$

for Δt sufficiently small. The following result characterises the influence of this perturbation on the asymptotic variance for small Δt. For numerical approximations of X_t for which a modified SDE Zygalakis (2011) is known, the following result combined with Proposition 4.3 provide a convenient means of obtaining an expression for the asymptotic variance $\widetilde{\sigma}^2_{\Delta t}(f)$ of the numerical scheme in terms of $\sigma^2(f)$.

PROPOSITION 4.5 Consider a diffusion process $\widetilde{X}_t$ on $\mathbb{T}^d$ with smooth coefficients and generator $\widetilde{\mathscr{L}_{\Delta t}}$ which satisfies (4.16) and (4.17). Suppose that $\widetilde{X}_t$ has unique invariant distribution $\widehat{\pi}^{\Delta t}$ which satisfies

$$\int \psi(x)\widehat{\pi}^{\Delta t}(x)\,dx = \int \psi(x)\pi(x)\,dx + \Delta t^r R_\psi, \tag{4.18}$$

where $r > k$, and R_ψ is a smooth remainder term, uniformly bounded with respect to Δt. Then for all $f \in C^\infty(\mathbb{T}^d)$ with $\pi(f) = 0$,

$$\widetilde{\sigma}^2_{\Delta t}(f) = \sigma^2_{\Delta t}(f) + 2\Delta t^k R_f + o(\Delta t^k). \tag{4.19}$$

where

$$R_f = \left\langle \left(-\widetilde{\mathscr{L}_{\Delta t}}\right)^{-1} M_{\Delta t}(-\mathscr{L}_k)\,(-\mathscr{L})^{-1} f, M_{\Delta t} f \right\rangle_{\widehat{\pi}^{\Delta t}}. \tag{4.20}$$

Moreover, we can express the remainder term as

$$R_f = \left\langle (-\mathscr{L})^{-1} M_0(-\mathscr{L}_k)\,(-\mathscr{L})^{-1} f, f \right\rangle_\pi + o(\Delta t^k), \tag{4.21}$$

where $M_0\psi = \psi - \int_{\mathbb{T}^d} \psi(y)\,dy$.

Proof. The result follows from an argument similar to that of Theorem 4.4. □

5. Asymptotic Bias and Variance Estimates for the splitting scheme

In this section we derive asymptotic bias and variance estimates for the Lie-Trotter splitting scheme (1.10) on $\mathbb{T}^d$ by applying the general results derived in Section 4. In Section 5.1 we apply Theorem 4.1 to obtain an asymptotic bias estimate for the splitting scheme. In particular, we find that when an unbiased method is used for the reversible part of the dynamics, then the order of the bias of the splitting scheme depends only on the properties of the deterministic integrator applied to the nonreversible part of the dynamics. Furthermore, in Section 5.2 we obtain estimates for the asymptotic variance, in the particular case where a Metropolized integrator is used to integrate the reversible part of the dynamics. These estimates confirm the soundness of the spitting approach as they imply that for Δt sufficiently small, the numerical asymptotic variance mimics the good properties of the asymptotic variance of the exact dynamics.

5.1 *Asymptotic bias of the splitting scheme*

We now consider the Lie-Trotter scheme (1.10) on $\mathbb{T}^d$. In this section we obtain estimates for the asymptotic bias of the scheme by applying Theorem 4.1.

THEOREM 5.1 Suppose that π is invariant by $\Theta_{\Delta t}$, the integrator used for the reversible dynamics, and that that the deterministic flow $\Phi_{\Delta t}$ satisfies a modified backward equation of the form (4.3) where the vector fields f_j satisfy

$$\nabla\cdot(f_j(x)\pi(x)) = 0, \quad j = 1,\dots,r-1. \tag{5.1}$$

Then, assuming ergodicity, the Lie-Trotter splitting (1.10) has order r of accuracy for the invariant measure. More precisely, for all $\phi \in C^2(\mathbb{T}^d)$ and Δt sufficiently small

$$\int_{\mathbb{T}^d} \phi(x)\widehat{\pi}^{\Delta t}(dx) = \int_{\mathbb{T}^d} \phi(x)\pi(dx) + \Delta t^r C_{r,\phi} + \Delta t^{r+1} R_{\phi,\Delta t}, \tag{5.2}$$

where $C_{r,\phi}$ and $R_{\phi,\Delta t}$ are uniformly bounded and

$$C_{r,\phi} = \left\langle f_r, (-\mathscr{L})^{-1}(\phi - \pi(\phi))\right\rangle_\pi.$$

REMARK 5.1 From standard elliptic energy estimates, the remainder term $C_{r,\phi}$ in (5.2) satisfies the a priori bound

$$|C_{r,\phi}| \leqslant 2\rho^{-1}\|f_r\|_{L^2(\pi)}\|\phi\|_{L^2(\pi)},$$

where ρ is the $L^2(\pi)$ Poincare constant.

Theorem 5.1 follows from a direct application of Theorem 4.1 and is proved in Section **??**. Suppose that the nonreversible dynamics is determined by (1.11) where $\gamma(x) = \beta\widetilde{\gamma}(x)$, for $\beta \in \mathbb{R}$ and for some smooth vector field $\widetilde{\gamma}$. If $\Psi_{\Delta t}$ is an integrator for the flow with error order r, then it is straightforward to show that $\Psi_{\Delta t}$ will satisfy a modified backward equation of the form (4.3) where the vector fields f_j satisfy the scaling $f_j = |\beta|^{j+1}\widetilde{f}_j$, with $\|\widetilde{f}_j\|_{L^2(\pi)} \sim O(1)$ for $j = 0,\dots,r-1$. It follows that if the conditions of Theorem 5.1 hold, then the leading order term of the bias is of the form $C\Delta t^r|\beta|^{r+1}$, where C is independent of Δt and β. This estimate provides a rule of thumb for choosing the magnitude of the nonreversible perturbation β. Clearly, this should be as large as possible while maintaining a given tolerance ε for the bias. To this end, for $\Delta t \ll 1$, β must satisfy

$$|\beta| \asymp \varepsilon^{\frac{1}{r+1}}\Delta t^{-\frac{r}{r+1}}.$$

In particular, assuming that $|\beta| \asymp \Delta t^{-\kappa}$ where $\kappa \in \mathbb{R}$, we obtain an upper bound

$$\kappa \leqslant -\frac{1}{r+1}\frac{\log\varepsilon}{\log\Delta t} + \frac{r}{r+1}. \tag{5.3}$$

For $\varepsilon \asymp \Delta t$, this rule suggests that β should have been chosen to be $O(1)$ with respect to Δt if a first order integrator is used to simulate the nonreversible dynamics. Employing a higher order integrator however, permits larger values of $|\beta|$, in particular $|\beta| \asymp \Delta t^{-0.6}$ for a fourth order scheme as considered in the examples of Section 6. We emphasise that unless we have explicit control on the growth of the remainder term in (4.5) as a function of β, then (5.3) is only heuristic. Moreover, we are assuming that the integrator $\Psi_{\Delta t}$ is stable for this parameter regime. In general though, if this heuristic choice of β was leading to unstable integration, then one would either use a smaller β (which would though lead to reduced benefits in terms of asymptotic variance) or could instead use an appropriate stiff numerical integrator.

5.2 *Asymptotic variance of the splitting scheme*

In this section we characterise the asymptotic variance of the splitting scheme (1.10). As we are working under the assumption that Δt is small, we see that the asymptotic variance will agree, to leading order with the asymptotic variance of the continuous underlying dynamics, with any distinctions arising as second order behaviour. We show that, unlike the bias estimates obtained in Theorem 5.1 these higher-order error terms will also depend on the choice of integrator for the reversible dynamics $\Theta_{\Delta t}$. For clarity, we shall focus specifically on the case where $\Theta_{\Delta t}$ is simulated using MALA. We again shall assume that the integrator $\Phi_{\Delta t}$ for the nonreversible flow satisfies the following expansion

$$\Phi_{\Delta t}\phi = \phi + \Delta t\mathscr{A}_1\phi + \Delta t^2\mathscr{A}_2\phi + \Delta t^3 R_\phi, \quad \phi \in C^\infty(\mathbb{T}^d),$$

where $\mathscr{A}_1 = \gamma(x)\cdot\nabla$ is the antisymmetric part of $\mathscr{L}$ in $L^2(\pi)$ and $R_\phi \in C^\infty(\mathbb{T}^d)$ is bounded independently of Δt.

Proposition **??** in the Appendix implies that the reversible integrator $\Theta_{\Delta t}$ satisfies the following perturbation expansion

$$\Theta_{\Delta t}\phi = \phi + \Delta t \mathscr{G}_1 \phi + \Delta t^2 \mathscr{G}_2 \phi + \Delta t^{5/2} R_\phi, \quad \phi \in C^\infty(\mathbb{T}^d), \tag{5.4}$$

where $\mathscr{G}_1 = \mathscr{S}$ is the symmetric part of $\mathscr{L}$ in $L^2(\pi)$, $\mathscr{G}_2$ is given by (**??**), and R_ϕ is a smooth remainder term bounded independently with respect to Δt. The following theorem characterises the asymptotic variance of the Lie-Trotter splitting scheme (1.10) for this choice of reversible dynamics. It is a direct application of Theorem 4.4 and is proved in Section **??**.

THEOREM 5.2 Consider the Lie-Trotter splitting scheme defined by (1.10) where $\Theta_{\Delta t}$ is integrated using MALA and suppose that the nonreversible dynamics preserves the invariant distribution up to order 2. Then for all $f \in C^\infty(\mathbb{T}^d)$ we have

$$\begin{aligned}\widehat{\sigma}^2_{\Delta t}(f) &= \sigma^2(f) - 2\Delta t \operatorname{Var}_\pi[f] \\ &\quad + \Delta t \left\langle (-\mathscr{L})^{-1}(\mathscr{L}^2 - 2\left(\mathscr{A}_2 + \mathscr{G}_1\mathscr{A}_1 + \mathscr{G}_2\right)(-\mathscr{L})^{-1}(f - \pi(f), f - \pi(f)\right\rangle_\pi + o(\Delta t).\end{aligned}$$

If moreover, the nonreversible dynamics is integrated using a second order scheme then the $O(\Delta t)$ term can be written as

$$\left\langle (-\mathscr{L})^{-1}\left((\mathscr{S}^2 - 2\mathscr{G}_2) + [\mathscr{S}, \mathscr{A}]\right)(-\mathscr{L})^{-1}(f - \pi(f), f - \pi(f)\right\rangle_\pi,$$

where $\mathscr{S}$ and $\mathscr{A}$ are the symmetric and antisymmetric parts of $\mathscr{L}$ in $L^2(\pi)$, respectively.

From the point of view of tuning the nonreversible Langevin sampler defined by (1.10) the main conclusion of Theorem 5.2 is that, for Δt sufficiently small, the asymptotic variance of (1.10) is, to leading order, equal to the asymptotic varaince of the exact dynamics (1.2). In particular, given an observable f, this result implies that a choice of flow γ which reduces the variance of a sampler based on (1.2) will have a similarly beneficial effect on (1.10). One can thus leverage the theory detailed in Duncan *et al.* (2016) and Lelièvre *et al.* (2013) to design efficient samplers for a given target distribution π and observable f.

5.3 *Uniform rate of convergence to equilibrium for the Splitting Scheme*

In this section we shall show that Assumption 4.2 holds when the reversible dynamics is simulated using a Metropolis-Hastings scheme using MALA. To establish this, it is sufficient to show that a uniform minorization condition holds. More specifically, there exists Δt^* and $\widetilde{\alpha} > 0$ and a probability measure ν such that for any bounded measurable non-negative function f and $x \in \mathbb{T}^d$,

$$P_{\Delta t}^{\lceil T/\Delta t \rceil} f(x) \geqslant \widetilde{\alpha} \int_{\mathbb{T}^d} f \, d\nu, \tag{5.5}$$

where $0 < \Delta t \leqslant \Delta t^*$. This approach will follow very closely (Fathi & Stoltz, 2015, Sec. 4.4), and we shall only illustrate the slightly different set-up of the proof here.

THEOREM 5.3 Consider the Markov chain $\widehat{X}^n_{\Delta t}$ defined by (1.10) where the reversible dynamics $\Theta_{\Delta t}$ are simulated using a Metropolis-Hastings scheme with MALA (3.3). Then, for Δt sufficiently small, the uniform minorisation condition (5.5) holds, and as a result, Assumption 4.2 holds for $\widehat{X}^n_{\Delta t}$.

Proof. It is straightforward from the construction of the Lie-Trotter process (1.10) that we can write

$$\widehat{X}_n^{\Delta t} = \widehat{X}_0^{\Delta t} + \mathscr{G}_n + \mathscr{F}_n, \tag{5.6}$$

where

$$\mathscr{G}_n = \sqrt{2\Delta t}\sum_{k=0}^{n-1} \mathbf{1}\left[u_k \leqslant \alpha\left(\Phi_{\Delta t}\left(\widehat{X}_k^{\Delta t}\right), \Psi_{\Delta t}\left(\Phi_{\Delta t}\left(\widehat{X}_k^{\Delta t}\right), g_k\right)\right)\right] g_k,$$

and

$$\mathscr{F}_n = -\Delta t\sum_{k=0}^{n-1} \mathbf{1}\left[u_k \leqslant \alpha\left(\Phi_{\Delta t}\left(\widehat{X}_k^{\Delta t}\right), \Psi_{\Delta t}\left(\Phi_{\Delta t}\left(\widehat{X}_k^{\Delta t}\right), g_k\right)\right)\right] \nabla U\left(\Psi_{\Delta t}\left(\Phi_{\Delta t}\left(\widehat{X}_k^{\Delta t}\right), g_k\right)\right),$$

where $(u_k)_{k=0}^{n-1}$ are i.i.d $U[0,1]$ distributed random variables, $(g_k)_{k=0}^{n-1}$ are i.i.d $\mathscr{N}(0,I)$ distributed random variables, where α is the acceptance probability and $\Psi_{\Delta t}$ is the proposal function, i.e.

$$\Psi_{\Delta t}(x,g) = x + \Delta t \nabla U(x) + \sqrt{2\Delta t} g.$$

We introduce the decomposition $\mathscr{G}_n = \widetilde{\mathscr{G}}_n + \widehat{G}_n$ where

$$\widetilde{G}_n = \sqrt{2\Delta t}\sum_{k=0}^{n-1} \mathbf{1}\left[u_k \leqslant 1\right] g_k, \tag{5.7}$$

and

$$\widehat{G}_n = \sqrt{2\Delta t}\sum_{k=0}^{n-1} \left(\mathbf{1}\left[u_k \leqslant \alpha\left(\Phi_{\Delta t}\left(\widehat{X}_k^{\Delta t}\right), \Psi_{\Delta t}\left(\Phi_{\Delta t}\left(\widehat{X}_k^{\Delta t}\right), g_k\right)\right)\right] - \mathbf{1}[u_k \leqslant 1]\right) g_k. \tag{5.8}$$

Following (Fathi & Stoltz, 2015, Sec 4.4), one decomposes each random variable in the summand into a drift plus a martingale increment term, i.e.

$$\left(\mathbf{1}\left[u_k \leqslant \alpha\left(\Phi_{\Delta t}\left(\widehat{X}_k^{\Delta t}\right), \Psi_{\Delta t}\left(\Phi_{\Delta t}\left(\widehat{X}_k^{\Delta t}\right), g_k\right)\right)\right] - \mathbf{1}[u_k \leqslant 1]\right) g_k = D(\widehat{X}_k^{\Delta t}) + M_k,$$

where M_k is a martingale adapted to the filtration of $\widehat{X}_k^{\Delta t}$. We obtain

$$D(x) = \mathbb{E}_{g\sim\mathscr{N}(0,I)}\left[\left(\alpha\left(\Phi_{\Delta t}(x), \Psi_{\Delta t}\left(\Phi_{\Delta t}(x), g\right)\right) - 1\right) g\right]. \tag{5.9}$$

It follows from (**??**) that there exists a constant C independent of Δt such that

$$|D(x)| \leqslant C\Delta t^{3/2}, \tag{5.10}$$

for Δt sufficiently small. Thus, it follows that

$$\Delta t^{1/2}\sum_{k=0}^{n-1} D(\widehat{X}_k^{\Delta t}) \leqslant C\Delta t. \tag{5.11}$$

Similarly one can show that

$$\mathbb{E}\left[|M_k|^2 \,\middle|\, \widehat{X}_k^{\Delta t}\right] \leqslant C'\Delta t^{1/2},$$

so that by Chebyschev's inequality, for $n \leqslant \lceil T/\Delta t \rceil$,

$$\mathbb{P}\left[\left|\widehat{\mathscr{G}_n} - \sqrt{2\Delta t}\sum_{k=0}^{n-1} D(\widehat{X}_k^{\Delta t})\right| \geqslant \frac{1}{2}\right] \leqslant C''\Delta t^{1/2}, \tag{5.12}$$

for some constant C'' independent of Δt. Applying (5.11) and choosing Δt sufficiently small we obtain

$$\mathbb{P}\left[\left|\widehat{\mathscr{G}_n}\right| \geqslant 1\right] \leqslant \widehat{C}\Delta t \leqslant \frac{1}{2},$$

where $\widehat{C}$ is a constant independent of Δt. The remainder of the argument involves controlling the magnitude of $\mathscr{F}_n$ and the distribution of $\widetilde{\mathscr{G}_n}$ to obtain the minorisation condition (5.5) and follows identically to (Fathi & Stoltz, 2015, Sec 4.4). □

6. Numerical experiments

In this section, we perform a number of different numerical investigations that illustrate the superiority of the nonreversible Langevin samplers over standard Metropolis-Hastings algorithms for a fixed computational budget. In particular, we define computational cost here in terms of number of density evaluations which is the dominating cost in high dimensions. To this end we ensure that every comparison is made for the same computational cost, *i.e.*, same number of density evaluations.

As a first numerical example we consider the expectation of an observable with respect to the following two dimensional distribution

$$\pi(x) \propto \exp\left(-\frac{x_1^2}{100} - (x_2 + bx_1^2 - 100b)^2\right), \tag{6.1}$$

where $x = (x_1, x_2)$. The parameter $b > 0$ controls the degree of warpedness, and is chosen to be $b = 0.05$. Our objective is to estimate $\pi(f)$ when $f(x) = |x|^2$. The nonreversible flow γ is chosen as follows:

$$\gamma(x) = J\nabla \log \pi(x), \quad J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

In Figure 1, we plot characteristic trajectories of MALA as well its nonreversible counterpart (for $\beta = 25$) starting from the initial point $x = (15, 2)$. The figure suggests superior mixing of the nonreversible samplers, which improves further with increasing β values. In Figure 2 the mean-square error is plotted as a function of stepsize for different values of flow strength β. The reversible part of the Lie-Trotter scheme is simulated using MALA and RWMH in Figures 2a, and 2b , respectively. The "exact" value of $\pi(f)$ used to compute the MSE is obtained via adaptive Gaussian quadrature, accurate up to 10^{-10}. In accordance with the results of Theorems 5.1 and 5.2, the MSE is a tradeoff between bias and variance. For a fixed computational budget as Δt decreases, the bias arising from the discretisation of the nonreversible flow decreases. However, the variance simultaneously increases as the total simulated time $T = N\Delta t$ is reduced. This competion between bias and variance suggest an optimal choice of timestep Δt which minimises the MSE. This tradeoff is further exacerbated when β is increased. Nevertheless, for an appropriate choice of β the MSE can be up to an order of magnitude lower than that of MALA, at the same computational cost.

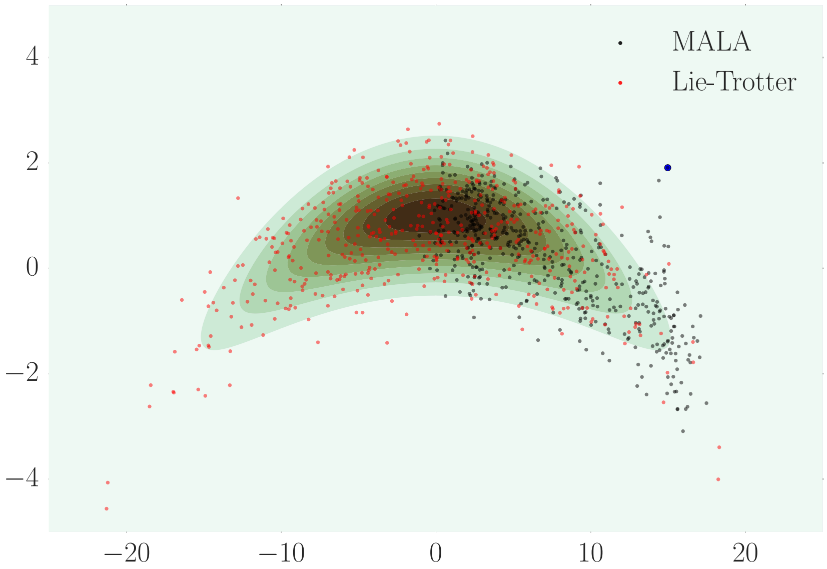
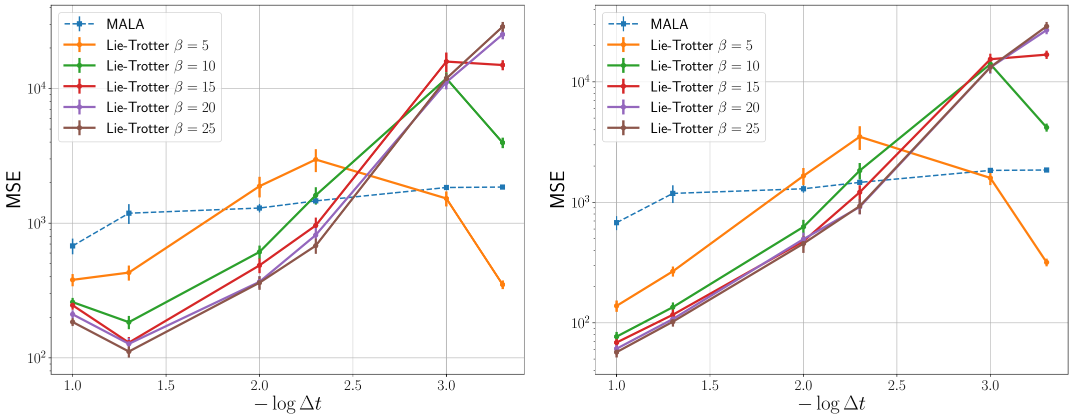


FIG. 1. Typical trajectories for MALA and Lie-Trotter splitting scheme applied to the warped Gaussian distribution (6.1), with computational budget of 3200 density evaluations. Both schemes started from $x = (15, 2)$ depicted by a blue dot.



(a) MALA for the reversible part

(b) RMWH rule for the reversible part

FIG. 2. Comparison of the MSE between MALA and different nonreversible samplers applied to the warped Gaussian distribution (6.1). The computational budget is set to $N = 3.5 \cdot 10^3$ density evaluations, and 4^{th} order Runge-Kutta method is used for the nonreversible component.

6.1 *Logistic Regression*

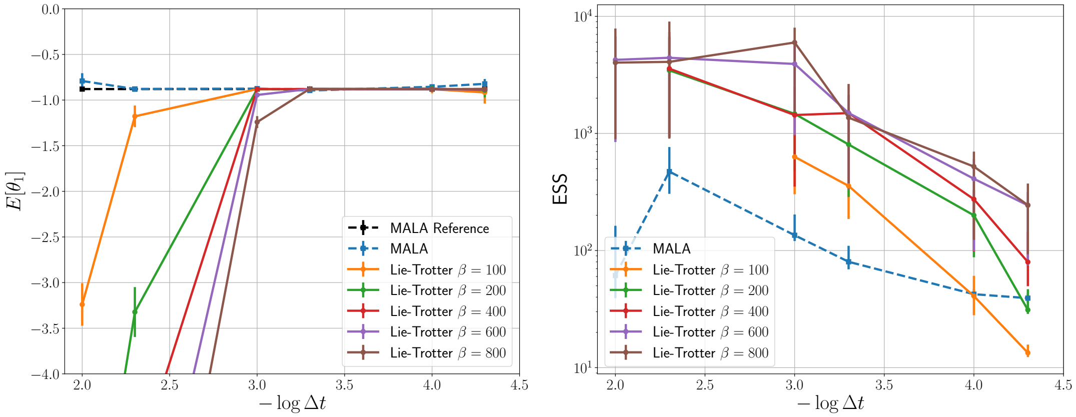
Let X be a $m \times d$ design matrix comprising m samples with d covariates and a binary response variable $Y \in \{-1, 1\}^m$. A Bayesian logistic regression model of the binary response is obtained by the introduction of the regression coefficient $\theta \in \mathbb{R}^d$. For the sake of exposition, we shall assume a Gaussian prior of θ, *i. e.*, $\theta \sim \mathcal{N}(0,\Sigma)$. The posterior distribution $\pi(\theta|X,Y)$ is given by

$$\pi(\theta|(X,Y)) \propto \exp\left(\sum_{i=1}^{m} Y_i\theta^T X_i - \log\left(1+e^{\theta^T X_i}\right) - \frac{1}{2}\theta^T\Sigma^{-1}\theta\right). \tag{6.2}$$

In Figure 3 we investigate the use of the Lie Trotter sampler applied to this problem for the Pima Indians[2] dataset obtained from the UCI machine learning repository. The skew symmetric matrix J is chosen by generating a random permutation $\sigma(1),\ldots,\sigma(d)$ and setting

$$J_{\sigma(i),\sigma(i+1)} = 1 \text{ and } J_{\sigma(i+1),\sigma(i)} = -1,$$

for $i = 1,\ldots,d-1$, and zero elsewhere. In Figure 3a we plot the first estimator $\widehat{\pi}_T^{\Delta t}(\theta_1)$ with 95% confidence intervals for different values of β and stepsize. Each point in the plot costs $3.5\cdot 10^3$ density evaluations. To provide a comparison against the truth, an optimally tuned MALA scheme was integrated over 10^7 timesteps. In Figure 3b we plot the effective sample size (ESS) of the Lie-Trotter scheme for different values of β and Δt. The markers denote the median value of the ESS with the markers denoting the 5% and 95% percentiles. We note however that there would typically be a very small number of observables for which the nonreversible scheme offers no advantage. This agrees with the theory detailed in Duncan *et al.* (2016) which characterises the minimum attainable variance reduction in terms of the projection of the observable f on the nullspace of the operator $J\nabla V(x)\cdot\nabla$. As J is chosen randomly, there will always been a number of observables which are close to this subspace, and thus the nonreversible dynamics offer no advantage. One possible remedy around this is to periodically resample the nonreversible matrix J, but we do not investigate this here.



(a) $\mathbb{E}(Y_{1,1})$ vs Step-size

(b) ESS vs Step-size

FIG. 3. Confidence interval for an estimator of $Y_{1,1}$ and ESS for estimators for $\pi(\theta_i)$, $i = 1,\ldots,9$ for logistic regression of the Pima Indians data set. Each data point in these plots is set to $3.5\cdot 10^3$ density evaluations. The results are compared to an optimally tuned MALA simulation run for 10^7 density evaluations.

[2] Here $m = 768, d = 9$.

6.2 *Spatial model*

We now consider a high dimensional target distribution related to inference for a log-Gaussian Cox point process previously considered in Møller *et al.* (1998). In particular, given the location of 126 Scots pine saplings in a natural forest in Finland, we wish to infer the average intensity of a corresponding Poisson point process. Following Christensen *et al.* (2005), we consider a discretised version of the model where the spatial region is discretised to a 64×64 regular grid. For each i, j $X_{i,j}$ is the random variable counting the number of observations in the (i, j)-cell ,and hence the dimension of the problem is $d = 64^2 = 4096$. The observations are assumed to be generated by a Poisson point process with unobserved intensity $\Lambda_{i,j}, i, j = 1, \cdots, 64$. Given the $\Lambda_{i,j}$ the random variables $X_{i,j}$ are assumed to be conditionally independent with Poisson distributed mean $m\Lambda_{i,j}$, where $m = 1/4096$ is the area of a single cell. We impose a log-Gaussian prior on $\Lambda_{i,j}$, more specifically

$$\Lambda_{i,j} = \exp(Y_{i,j}),$$

where $Y = (Y_{i,j}, i, j = 1, \cdots 64) \sim \mathscr{N}(\mu\mathbf{1}, \Sigma)$ where

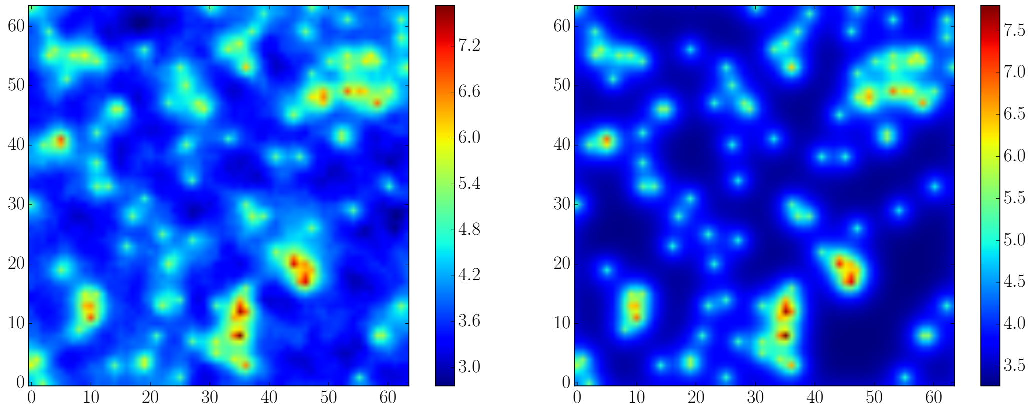
$$\Sigma_{i,j,i',j'} = \sigma^2 \left[-\frac{-\{(i-i')^2 + (j-j')^2\}^{1/2}}{64\beta} \right], \quad i, j, i', j' = 1 \cdots, 64.$$

The posterior distribution is thus given by

$$f(y|x) \propto \prod_{i,j=1}^{64} \exp\{(x_{i,j} y_{i,j}) - m \exp(y_{i,j})\} \exp\{-0.5(y - \mu\mathbf{1})^T \Sigma^{-1} (y - \mu\mathbf{1})\}.$$

Due to the poor scaling of the posterior distribution in Christensen *et al.* (2005) a reparametrization of y is introduced to improve the mixing of the Metropolis-Hastings scheme. This procedure is expensive with a computational cost of $\mathscr{O}(d^3)$. However, in the case of the nonreversible samplers, the nonreversible perturbation compensates for the poor scaling, thus rendering this reparametrisation unnecessary.

In Figure 4 we plot an estimator of $\mathbb{E}(\Lambda \,|\, x)$ using MALA and its nonreversible counterpart respectively. For this computation the skew-symmetric matrix J was generated randomly as in the logistic regression example. Due to the large number of variables, for any given random choice of J, there would be a small number of variables for which the nonreversible scheme does not offer significant advantage over MALA, as described in Duncan *et al.* (2016). To better understand the effect of the nonreversible flow on an average covariate, we generate 10 independent random skew-symmetric matrices, and compute the average ESS over J. The results are presented in Figure 5. In Figure 5c a histogram of the ESS over all variables is plotted for both MALA and the splitting scheme for specific choices of Δt and β. We observe that the ESS for the nonreversible scheme is orders of magnitude better than MALA. To illustrate the dependence of ESS on timestep, similarly to the case of logistic regression, in Figure 5b we plot the median ESS for different choices of timestep. It is clear that increasing β and Δt as much as possible increases the ESS. However, this comes at the cost of increasing bias as can be observed in Figure 5a. Nonetheless, it is evident that the nonreversible sampler significantly outperforms the MALA scheme.



(a) Average intensity estimated using MALA

(b) Average intensity estimated using Lie-Trotter scheme

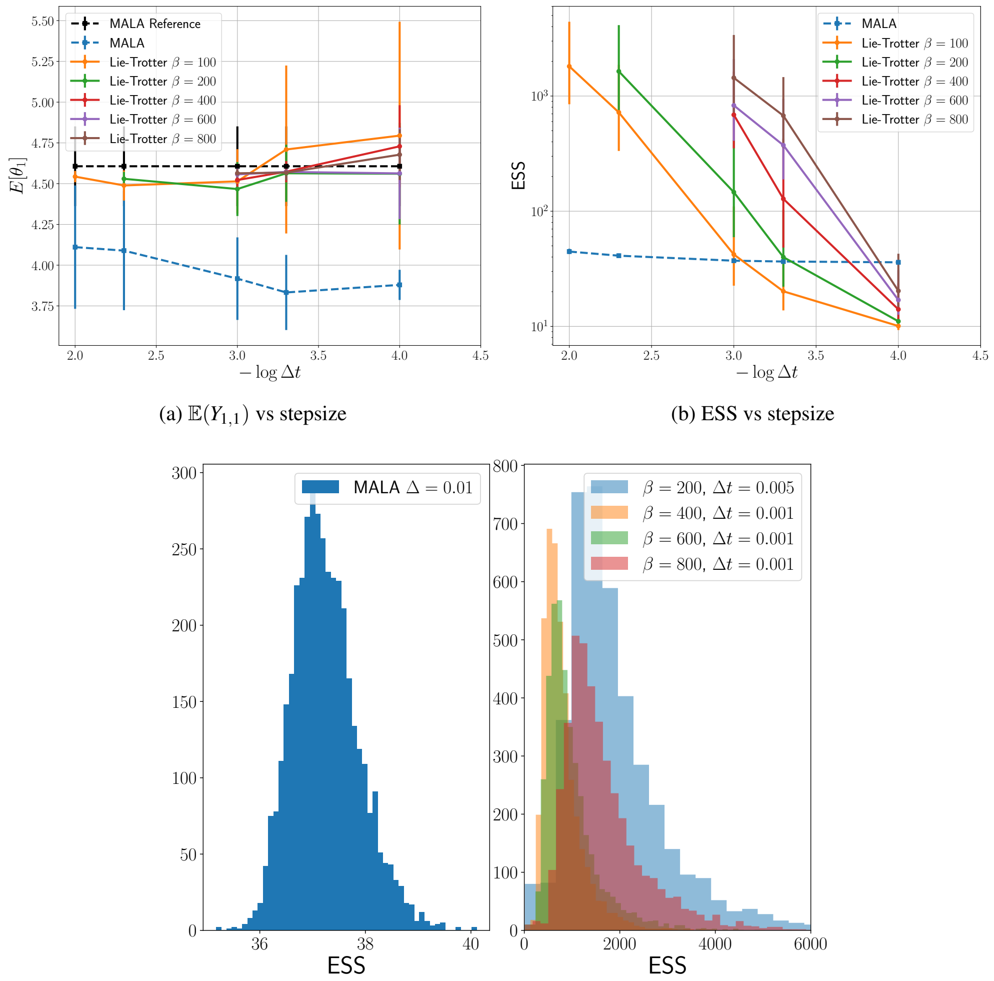
FIG. 4. $\mathbb{E}(Y_{i,j})$ estimated using different schemes. The computational budget is set to $N = 3.5 \cdot 10^3$ gradient evaluations.

7. Discussion

In this paper sampling methods based on nonreversible diffusions have been proposed and evaluated on a range of different inference problems. The development of these methods is an attempt to improve on existing MCMC methodology in the case of target densities that might be of high dimension and exhibit strong correlations. The key idea behind these samplers is the exploitation of the irreversibility of an underlying diffusion process, which leads to reduced asymptotic variance. This becomes possible through a careful discretisation of the underlying SDE that introduces a controllable bias, but more importantly mimics the reduced asymptotic variance of the nonreversible diffusion.

From a practical point of view, the careful balancing of the bias and variance achieved by the nonreversible samplers leads to much more efficient sampling than MALA. In particular, across all our experiments we observe improvements of two orders of magnitude in terms of effective sample size. Moreover, all our comparisons are being made on the basis of the same number of density evaluations used in the nonreversible samplers and MALA. Furthermore, in the case of the log- Gaussian Cox model the nonreversible samplers are able to achieve this dramatic improvement in terms of the ESS without the need of an expensive $\mathscr{O}(d^3)$ reparametrisation, which is also the computational bottleneck in high dimensions for more sophisticated sampling algorithms such as MMALA Girolami & Calderhead (2011). We mention also here that a new class of methods Titsias & Papaspiliopoulos (2018) shares this improved performance without the need for such an expensive rescaling.

There exist a number of different directions that one could extend this work. In particular, when dealing with the nonreversible part of the dynamics further computational benefits may be achieved with the use of adaptive integration. Furthermore, one could replace the Metropolis-Hasting scheme used for simulating the reversible part of the dynamics by appropriate numerical schemes Abdulle *et al.* (2014) that preserve the invariant measure to high order. In this situation one would expected the results of our analysis to still hold which is important as the corresponding nonreversible samplers would allow for greater flexibility in the presence of big data, where traditional MCMC methods might become prohibitively expensive.



(a) $\mathbb{E}(Y_{1,1})$ vs stepsize

(b) ESS vs stepsize

(c) Histogram of ESS

FIG. 5. Results for the inference of the log-Gaussian Cox process. The computational budget is set to $N = 3.5 \cdot 10^3$ density evaluations. A reference MALA simulation run for 10^7 density evaluations is provided for comparison.

Acknowledgements

The authors would like to thank Mark Girolami, Andrew Stuart, Gabriel Stoltz and Ioannis Papastathopoulos for useful discussions about the paper.

REFERENCES

ABDULLE, A., VILMART, G. & ZYGALAKIS, K. C. (2014) High order numerical approximation of the invariant measure of ergodic SDEs. *SIAM J. Numer. Anal.*, **52**, 1600–1622.

ABDULLE, A., VILMART, G. & ZYGALAKIS, K. C. (2015) Long time accuracy of Lie–Trotter splitting methods for langevin dynamics. *SIAM Journal on Numerical Analysis*, **53**, 1–16.

BOU-RABEE, N. & HAIRER, M. (2012) Nonasymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis*.

CHRISTENSEN, O. F., ROBERTS, G. O. & ROSENTHAL, J. S. (2005) Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 253–268.

DIACONIS, P., HOLMES, S. & NEAL, R. M. (2000) Analysis of a nonreversible Markov chain sampler. *The Annals of Applied Probability*, **10**, 726–752.

DUANE, S., KENNEDY, A. D., PENDLETON, B. J. & ROWETH, D. (1987) Hybrid Monte Carlo. *Physics letters B*, **195**, 216–222.

DUNCAN, A. B., LELIÈVRE, T. & PAVLIOTIS, G. A. (2016) Variance reduction using nonreversible Langevin samplers. *Journal of Statistical Physics*, **163**, 457–491.

DUNCAN, A. B., NÜSKEN, N. & PAVLIOTIS, G. A. (2017) Using perturbed underdamped langevin dynamics to efficiently sample from probability distributions. *Journal of Statistical Physics*, **169**, 1098–1131.

FATHI, M. & STOLTZ, G. (2015) Improving dynamical properties of metropolized discretizations of overdamped Langevin dynamics. *Numerische Mathematik*, 1–58.

FUTAMI, F., SATO, I. & SUGIYAMA, M. (2020) Accelerating the diffusion-based ensemble sampling by non-reversible dynamics. *Proceedings of the 37th International Conference on Machine Learning* (H. D. III & A. Singh eds). Proceedings of Machine Learning Research, vol. 119. PMLR, pp. 3337–3347.

FUTAMI, F., IWATA, T., UEDA, N. & SATO, I. (2021) Accelerated diffusion-based sampling by the non-reversible dynamics with skew-symmetric matrices. *Entropy*, **23**.

GAO, X., GURBUZBALABAN, M. & ZHU, L. (2020) Breaking reversibility accelerates langevin dynamics for non-convex optimization. *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin eds), vol. 33. Curran Associates, Inc., pp. 17850–17862.

GIROLAMI, M. & CALDERHEAD, B. (2011) Riemann Manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 123–214.

GLYNN, P. W. & MEYN, S. P. (1996) A Liapounov bound for solutions of the Poisson equation. *The Annals of Probability*, **24**, 916–931.

HAIRER, E., LUBICH, C. & WANNER, G. (2006) *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer Series in Computational Mathematics 31, second edn. Berlin: Springer-Verlag.

HASTINGS, W. K. (1970) Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

HOROWITZ, A. M. (1991) A generalized guided Monte Carlo algorithm. *Physics Letters B*, **268**, 247–252.

HUKUSHIMA, K. & SAKAI, Y. (2013) An irreversible Markov-chain Monte Carlo method with skew detailed balance conditions. *Journal of Physics: Conference Series*, vol. 473. IOP Publishing, IOP Publishing, p. 012012.

HWANG, C. R., HWANG-MA, S. Y., SHEU, S. J. *et al.* (2005) Accelerating diffusions. *The Annals of Applied Probability*, **15**, 1433–1444.

HWANG, C.-R., NORMAND, R. & WU, S.-J. (2015) Variance reduction for diffusions. *Stochastic Processes and their Applications*, **125**, 3522–3540.

KOPEC, M. (2014) Weak backward error analysis for overdamped Langevin processes. *IMA Journal of Numerical Analysis*, dru016.

LEIMKUHLER, B., MATTHEWS, C. & STOLTZ, G. (2013) The computation of averages from equilibrium langevin molecular dynamics. *IMA J. Numer. Anal.*

LEIMKUHLER, B. & REICH, S. (2004) *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics 14. Cambridge: Cambridge University Press.

LELIÈVRE, T., NIER, F. & PAVLIOTIS, G. A. (2013) Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *Journal of Statistical Physics*, **152**, 237–274.

LORENZI, L. & BERTOLDI, M. (2006) *Analytical methods for Markov semigroups*. CRC Press.

MA, Y.-A., CHEN, T. & FOX, E. (2015) A complete recipe for stochastic gradient MCMC. *Advances in Neural Information Processing Systems*. *Advances in Neural Information Processing Systems*., pp. 2899–2907.

MA, Y.-A., CHEN, T., WU, L. & FOX, E. B. (2016) A unifying framework for devising efficient and irreversible mcmc samplers. *arXiv preprint arXiv:1608.05973*.

MATTINGLY, J. C., STUART, A. M. & HIGHAM, D. J. (2002) Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, **101**, 185–232.

MATTINGLY, J. C., STUART, A. M. & TRETYAKOV, M. V. (2010) Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM Journal on Numerical Analysis*, **48**, 552–577.

MENGERSEN, K. L. & TWEEDIE, R. L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, **24**, 101–121.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. (1953) Equation of state calculations by fast computing machines. *The journal of chemical physics*, **21**, 1087–1092.

MEYN, S. P. & TWEEDIE, R. L. (1993a) Stability of Markovian processes III: Foster-Lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, **25**, 518–548.

MEYN, S. P. & TWEEDIE, R. L. (1993b) A survey of Foster-Lyapunov techniques for general state space Markov processes. *Proceedings of the Workshop on Stochastic Stability and Stochastic Stabilization, Metz, France*. Citeseer, Citeseer.

MIJATOVIĆ, A. & VOGRINC, J. (2018) On the poisson equation for metropolis–hastings chains. *Bernoulli*, **24**, 2401–2428.

MIJATOVIĆ, A. & VOGRINC, J. (2019) Asymptotic variance for random walk metropolis chains in high dimensions: logarithmic growth via the poisson equation. *Advances in Applied Probability*, **51**, 994–1026.

MIRA, A. & GEYER, C. J. (2000) On non-reversible markov chains. *Monte Carlo Methods, Fields Institute/AMS*, 95–110.

MØLLER, J., SYVERSVEEN, A. R. & WAAGEPETERSEN, R. P. (1998) Log Gaussian Cox processes. *Scandinavian journal of statistics*, **25**, 451–482.

NEAL, R. M. (2004) Improving asymptotic variance of MCMC estimators: Nonreversible chains are better. *arXiv preprint math/0407281*.

NEAL, R. M. (2011) MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. Jones & X.-L. Meng eds). Boca Raton: CRC press, pp. 113–162.

OTTOBRE, M., PILLAI, N. S., PINSKI, F. J. & STUART, A. M. (2016) A function space HMC algorithm with second order Langevin diffusion limit. *Bernoulli*, **22**, 60–106.

PAVLIOTIS, G. A. (2014) *Stochastic processes and applications*. Texts in Applied Mathematics, vol. 60. Springer, New York, pp. xiv+339. Diffusion processes, the Fokker-Planck and Langevin equations.

PONCET, R. (2017).

REY-BELLET, L. & SPILIOPOULOS, K. (2015a) Irreversible Langevin samplers and variance reduction: a large deviations approach. *Nonlinearity*, **28**, 2081.

REY-BELLET, L. & SPILIOPOULOS, K. (2015b) Variance reduction for irreversible langevin samplers and diffusion on graphs. *Electron. Commun. Probab.*, **20**, 16 pp.

ROBERTS, G. O., ROSENTHAL, J. S. & SCHWARTZ, P. O. (1998) Convergence properties of perturbed Markov chains. *Journal of Applied Probability*, **35**, 1–11.

ROBERTS, G. O. & STRAMER, O. (2002) Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, **4**, 337–357.

ROBERTS, G. O. & TWEEDIE, R. L. (1996) Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, **2**, 341–363.

SANZ-SERNA, J. M. & CALVO, M. P. (1994) *Numerical Hamiltonian problems*. Applied Mathematics and Mathematical Computation, vol. 7. Chapman & Hall, London, pp. xii+207.

SMITH, A. F. M. & ROBERTS, G. O. (1993) Bayesian computation via the Gibbs sampler and related markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–23.

STRAMER, O. & TWEEDIE, R. L. (1999a) Langevin-type models I: Diffusions with given stationary distributions and their discretizations. *Methodology and Computing in Applied Probability*, **1**, 283–306.

STRAMER, O. & TWEEDIE, R. L. (1999b) Langevin-type models II: self-targeting candidates for mcmc algorithms. *Methodology and Computing in Applied Probability*, **1**, 307–328.

TALAY, D. & TUBARO, L. (1990) Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, **8**, 483–509.

TIERNEY, L. (1994) Markov chains for exploring posterior distributions. *the Annals of Statistics*, 1701–1728.

TITSIAS, M. K. & PAPASPILIOPOULOS, O. (2018) Auxiliary gradient-based sampling algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**, 749–767.

TURITSYN, K. S., CHERTKOV, M. & VUCELJA, M. (2011) Irreversible Monte Carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, **240**, 410–414.

WU, S.-J., HWANG, C.-R. & CHU, M. T. (2014) Attaining the optimal gaussian diffusion acceleration. *Journal of Statistical Physics*, **155**, 571–590.

ZYGALAKIS, K. C. (2011) On the existence and the applications of modified equations for stochastic differential equations. *SIAM J. Sci. Comput.*, **33**, 102–130.

Appendix A. Proofs of the main results

In this section we prove the main results of the paper relating to the asymptotic bias and variance of the splitting method. The proof of the geometric ergodicity of the splitting scheme (1.10) on $\mathbb{R}^d$ is relatively standard, and therefore deferred to the supplementary material.

.1 *Asymptotic variance of numerical integrators*

Here we prove Proposition 4.3 and Theorem 4.4 which characterises the error in the asymptotic variance for an arbitrary numerical integrator.

Proof of Proposition 4.3. It follows from the maximum principle that the operator $(-\mathscr{L})^{-1}$ is bounded on $L^\infty_0(\pi)$. Given a smooth $\psi \in L^\infty(\mathbb{T}^d)$, we first show that $\Delta t(I - P_{\Delta t})^{-1}\psi$ and its derivatives with respect to x are bounded, uniformly with respect to Δt. To this end, we Taylor expand $((I - P_{\Delta t})/\Delta t)\psi$ with respect to Δt around 0, yielding

$$\frac{(I-P_{\Delta t})}{\Delta t}\psi = (-\mathscr{L})\psi + \frac{\Delta t}{2}(-\mathscr{L})^2\psi + \frac{\Delta t^2}{6}(-\mathscr{L})^3\psi + \frac{\Delta t^3}{24}(-\mathscr{L})^4 P_s\psi,$$

for some $s \in [0, \Delta t]$. Comparing coefficients of equal powers, it follows that

$$\left((-\mathscr{L})^{-1} - \frac{\Delta t}{2}I + \frac{\Delta t^2}{12}(-\mathscr{L})\right)\frac{(I-P_{\Delta t})}{\Delta t}\psi = (I + \Delta t^3 R)\psi,$$

where $R\psi = (-\mathscr{L})^3\frac{1}{24}P_s\psi - \frac{\Delta t}{48}(-\mathscr{L})^4 P_s\psi + \frac{1}{24}(-\mathscr{L})^3\psi + \frac{\Delta t}{72}(-\mathscr{L})^4\psi + \frac{\Delta t^2}{288}(-\mathscr{L})^5 P_s\psi$. The operator R is bounded on $L^\infty_0(\mathbb{T}^d)$ and so, for Δt sufficiently small, $(I + \Delta t^3 R)$ is invertible. It follows that

$$\left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} = (I + \Delta t^3 R)^{-1}\left((-\mathscr{L})^{-1} - \frac{\Delta t}{2}I + \frac{\Delta t^2}{12}(-\mathscr{L})\right).$$

Provided that ψ is smooth with bounded derivatives on $\mathbb{T}^d$ then one can show that, for $0 \leqslant s \leqslant \Delta t$:

$$\|\nabla_x P_s\psi\|_{L^\infty(\mathbb{T}^d)} \leqslant C(1 + \|\nabla V\|_{L^\infty(\mathbb{T}^d)} + \|\nabla\nabla V\|_{L^\infty(\mathbb{T}^d)})(\|\psi\|_{L^\infty(\mathbb{T}^d)} + \|\nabla\psi\|_{L^\infty(\mathbb{T}^d)})\Delta t,$$

for some constant C independent of Δt which implies that

$$\|\nabla_x R\psi\| \leqslant C\left(1 + \sup_{i=0,\ldots,10}\|\nabla^i V\|_{L^\infty(\mathbb{T}^d)}\right)\sup_{i=0,\ldots,9}\|\nabla^i\psi\|_{L^\infty(\mathbb{T}^d)}\Delta t.$$

These together with standard energy estimates for solutions of elliptic PDEs on $\mathbb{T}^d$ yield that $\left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1}$ has a bounded first derivative, uniformly with respect to Δt. Similar arguments yield that, provided that V are ψ are smooth on $\mathbb{T}^d$, then $\left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1}\psi$ has all derivatives bounded, uniformly with respect to Δt. Since R is a bounded operator on $L^2(\pi)$, there exists δ such that $\Delta t^3\|R\|_{L^2(\pi)} \leqslant 2$, for all $\Delta t \leqslant \delta$. It follows that $(I + \Delta t^3 R)^{-1}\psi = \psi + \Delta t^3 r_\psi$ where r_ψ is bounded in L^2, uniformly with respect to $\min(\Delta t, \delta)$, so that

$$\left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1}\psi = (-\mathscr{L})^{-1}\psi - \frac{\Delta t}{2}\psi + \frac{\Delta t^2}{12}(-\mathscr{L})\psi + \Delta t^3 r_\psi. \tag{A.1}$$

Let $f \in C^\infty(\mathbb{T}^d)$, then similar to (4.9), the asymptotic variance of the estimator $\Delta t N^{-1}\sum_{n=0}^{N-1} f(X_n^{\Delta t})$ for the discretized exact process is given by

$$\sigma^2_{\Delta t}(f) = 2\left\langle \left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1}(f-\pi(f)), f-\pi(f)\right\rangle_\pi - \Delta t\,\mathrm{Var}_\pi[f].$$

By (A.1) it follows that

$$\begin{aligned}\sigma^2_{\Delta t}(f) &= 2\left\langle (-\mathscr{L})^{-1}(f-\pi(f)), f-\pi(f)\right\rangle_\pi - \Delta t\left\langle (f-\pi(f)), f-\pi(f)\right\rangle_\pi - \Delta t\,\mathrm{Var}_\pi[f]\\ &\quad + \frac{\Delta t^2}{6}\left\langle (-\mathscr{L})(f-\pi(f)), f-\pi(f)\right\rangle_\pi + \Delta t^3 R_f\\ &= \sigma^2(f) - 2\Delta t\,\mathrm{Var}_\pi[f] + \frac{\Delta t^2}{6}\left\langle (-\mathscr{L})(f-\pi(f)), f-\pi(f)\right\rangle_\pi + \Delta t^3 R_f,\end{aligned}$$

as required. □

Proof of Theorem 4.4. The proof of this result follows closely that of (Leimkuhler *et al.*, 2013, Theorem 2.9). To this end, given $f, g \in C^\infty(\mathbb{T}^d)$ such that $\pi(f) = \pi(g) = 0$, consider $\left\langle \left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f, g\right\rangle_\pi$. Since $\left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f$ has mean zero with respect to π, then

$$\begin{aligned}\left\langle \left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f, g\right\rangle_\pi &= \left\langle \left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f, M_{\Delta t} g\right\rangle_\pi\\ &= \left\langle \left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f, M_{\Delta t} g\right\rangle_{\widehat{\pi}^{\Delta t}} + \Delta t^r R_{f,g},\end{aligned}$$

for a smooth remainder term $R_{f,g}$ bounded uniformly with respect to Δt. Using the expansion (4.6) for the semigroup $\widehat{P}_{\Delta t}$:

$$\begin{aligned}&\left\langle \left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f, M_{\Delta t} g\right\rangle_{\widehat{\pi}^{\Delta t}} = \left\langle \left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t}\left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)\left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f, M_{\Delta t} g\right\rangle_{\widehat{\pi}^{\Delta t}}\\ &= \left\langle \left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t}\left(\frac{I-P_{\Delta t}}{\Delta t} + \Delta t^k\left(\frac{\mathscr{L}^{k+1}}{(k+1)!} - A_k\right)\right)\left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f, M_{\Delta t} g\right\rangle_{\widehat{\pi}^{\Delta t}}\\ &\quad + \Delta t^{q-1}\left\langle \left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t} R_f, M_{\Delta t} g\right\rangle_{\widehat{\pi}^{\Delta t}}\\ &= \left\langle \left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t} f, M_{\Delta t} g\right\rangle_{\widehat{\pi}^{\Delta t}}\\ &\quad + \Delta t^k\left\langle \left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t}\left(\frac{\mathscr{L}^{k+1}}{(k+1)!} - A_k\right)\left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f, M_{\Delta t} g\right\rangle_{\widehat{\pi}^{\Delta t}}\\ &\quad + \Delta t^{q-1}\left\langle \left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t} R_f, M_{\Delta t} g\right\rangle_{\widehat{\pi}^{\Delta t}},\end{aligned} \tag{A.2}$$

where R_f is a smooth function depending on f, bounded uniformly with respect to Δt. By Assumption 4.2, the coefficients of the Δt^k and Δt^{q-1} terms are bounded uniformly with respect to Δt. Equation (4.12) then follows immediately, and thus (4.14). Noting that $M_{\Delta t} = M_{\Delta t} M_0$ then by applying (A.2) with

$$f = \left(\frac{\mathscr{L}^{k+1}}{(k+1)!} - M_0 A_k \right) \left(\frac{I - P_{\Delta t}}{\Delta t} \right)^{-1} f, \quad \text{and} \quad g = g,$$

we obtain

$$\begin{aligned} R_1(f,g) &= \left\langle \left(\frac{I - \widehat{P}_{\Delta t}}{\Delta t} \right)^{-1} M_{\Delta t} \left(\frac{\mathscr{L}^{k+1}}{(k+1)!} - M_0 A_k \right) \left(\frac{I - P_{\Delta t}}{\Delta t} \right)^{-1} f, M_{\Delta t} g \right\rangle_{\widehat{\pi}^{\Delta t}} \\ &= \left\langle \left(\frac{I - P_{\Delta t}}{\Delta t} \right)^{-1} \left(\frac{\mathscr{L}^{k+1}}{(k+1)!} - M_0 A_k \right) \left(\frac{I - P_{\Delta t}}{\Delta t} \right)^{-1} f, g \right\rangle_{\pi} + \Delta t^{q-1} R_2(f,g), \end{aligned}$$

for some smooth, uniformly bounded remainder term R_2. Note that, as detailed in the proof of Proposition 4.3, this choice of f is smooth, with all derivatives bounded uniformly with respect to Δt. We now apply (A.1) to the discrete generator $\Delta t^{-1}(I - P_{\Delta t})$ to obtain

$$\begin{aligned} \left\langle \left(\frac{I - P_{\Delta t}}{\Delta t} \right)^{-1} \left(\frac{\mathscr{L}^{k+1}}{(k+1)!} - M_0 A_k \right) \left(\frac{I - P_{\Delta t}}{\Delta t} \right)^{-1} f, g \right\rangle_{\pi} &= \left\langle (-\mathscr{L})^{-1} \left(\frac{\mathscr{L}^{k+1}}{(k+1)!} - M_0 A_k \right) (-\mathscr{L})^{-1} f, g \right\rangle_{\pi} \\ &\quad + \Delta t R_3(f,g), \end{aligned}$$

for a smooth bounded remainder term R_3, from which (4.15) follows. □