Nonreversible Langevin Samplers: Splitting Schemes, Analysis and Implementation

A.B. Duncan
School of Mathematical and Physical Sciences
University of Sussex
Falmer
and
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

August 9, 2017

Abstract

For a given target density π on $\mathbb{R}^d$, there exist infinitely many diffusion processes that 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 (2015*a*,*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 sampler inherits 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.

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 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, i.e. they satisfy *detailed balance*. 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 $(X_t)_{t\geq 0}$ 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

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

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

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

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

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

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

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

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 (2015*a*,*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 (1999*a*,*b*). In addition, even when the numerical discretization is ergodic and thus

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

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)). 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 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 higher-order quadrature 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.

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 $\Phi_{\Delta t}$ 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 $\widehat{\sigma}^2_{\Delta t}(f)$, 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 (1993*b*), 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 Mijatovic & Vogrinc (2015) and has also recently been used to analyse the asymptotic variance of random walk Metropolis chains Mijatović & Vogrinc (2017).

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 $\mathbb{T}^d$. In Section 6, we focus on the case where the target distribution is Gaussian and study analytically the trade-off between the asymptotic bias and asymptotic variance in this case. To demonstrate the efficacy of these schemes, in Section 7 we present a number of numerical experiments on inference for spatial models as well as on Bayesian logistic regression. Proofs of the main results of this paper are deferred to Section 8 as well as the Appendices. Finally, a discussion of the results presented in this paper and potential future research directions can be found in Section 9.

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 $(P_t)_{t\geq 0}$ 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

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

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

$$\gamma(x) = J\nabla \log \pi(x)\Psi(\pi(x)), \quad J = -J^\top, \text{ and } \beta \in \mathbb{R}, \tag{2.5}$$

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 (1993*c*), 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*

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

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 (1993*b*). For the generator $\mathcal{L}$ 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 (1993*b*)).

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

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

for all f *satisfying* $|f| \le V$*.*

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. *Suppose that Assumption 2.3 holds and that* $\pi(U^2) < \infty$*, then for any function* f *such that* $|f| \le U$*, the central limit theorem (1.5) holds, i.e.* $\sqrt{T}(\pi_T(f) - \pi(f))$ *converges weakly to a* $\mathcal{N}(0, 2\sigma^2(f))$*–distributed random variable, with*

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

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

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

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 2.6. *(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 2.7. *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 non-reversible 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 (1993*b*) 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
$$\alpha(x, y) = \min\left(1, \frac{\pi(y)q_{\Delta t}(x|y)}{\pi(x)q_{\Delta t}(y|x)}\right),$$
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

$$q_{\Delta t}(\cdot\,|x) = x + \Delta t \nabla \log \pi(x) + \sqrt{2\Delta t}g, \tag{3.2}$$

and the random walk proposal

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

where g is a standard d-dimensional Gaussian random variable. The resulting scheme is known as *Metropolis-Adjusted Langevin Algorithm* (MALA) when proposal (3.2) is used, and *Random Walk Metropolis Hastings* (RWMH) when (3.3) 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,

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

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

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

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 $V \geq 1$, a compact set $C \subset \mathbb{R}^d$ and constants $\lambda \in (0, 1)$ and $b \geq 0$ such that*

$$\widetilde{P}_{\Delta t} V(x) \leq \lambda V(x) + b \mathbf{1}_C(x), \quad x \in \mathbb{R}^d. \tag{3.4}$$

2 *The nonreversible flow map $\Phi_{\Delta t}$ satisfies the following condition,*

$$\lim \sup_{|x| \to \infty} \frac{V(\Phi_{\Delta t}(x)) - V(x)}{V(x)} < \frac{1}{\lambda} - 1. \tag{3.5}$$

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*

$$\sup_{|g| \leq V} \left| \int_{\mathbb{R}^d} g(y) \left(P_{\Delta t}(x, y) - \pi(y) \right) dy \right| \leq K V(x) \rho^n, \quad n \in \mathbb{N}.$$

The following result is an application of Theorem 3.2 for the Random Walk proposal (3.3).

Corollary 3.3 (Geometric Ergodicity of Lie-Trotter scheme with RWMH dynamics)**.** *Consider the Lie-Trotter splitting scheme $\hat{X}^{\Delta t}_n$ where the reversible dynamics (1.12) are simulated using a RWMH scheme with proposal defined by (3.3). Suppose that the conditions on π and $q_{\Delta t}$ specified in (Roberts et al. 1998, Theorem 3.2) hold and moreover that*

$$\lim_{|x| \to \infty} (|\Phi_{\Delta t}(x)| - |x|) = 0, \tag{3.6}$$

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

An almost identical result holds for the MALA proposal (3.2).

Corollary 3.4 (Geometric Ergodicity of Lie-Trotter scheme with MALA dynamics)**.** *Consider the Lie-Trotter splitting scheme $\widehat{X}^{\Delta t}_n$ where the reversible dynamics (1.12) are simulated using a MALA scheme with proposal defined by (3.2). Suppose that the conditions on π and $q_{\Delta t}$ specified in (Roberts & Tweedie 1996, Theorem 4.1) hold and moreover that* (3.6) *holds for Δt sufficiently small. Then $\widehat{X}^{\Delta t}_n$ 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

$$|\nabla\pi^{\alpha}(x)| \le K_1\pi^{\alpha'}(x), \quad |\nabla\nabla\pi^{\alpha}(x)|_{max} \le K_2, \quad x \in \mathbb{R}^d, \tag{3.7}$$

where $|\cdot|_{max}$ denotes the max norm. If $\gamma = J\nabla\pi^{\alpha}$ for J antisymmetric, then condition (3.6) 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 $(\widehat{X}_k^{\Delta t})_{k\ge 0}$. 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 $(\mathbb{T}^d, \mathcal{B}(\mathbb{T}^d))$ 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 $L^2_0(\mu)$ of $L^2(\mu)$ is defined to be

$$L^2_0(\mu) = \{f \in L^2(\mu) \,:\, \mu(f) = 0\}, \tag{4.1}$$

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 $(\mathbb{T}^d, \mathcal{B}(\mathbb{T}^d))$ 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 $z_n = \widetilde{z}(t_n) + \mathcal{O}(\Delta t^{s+1})$ for $\Delta t \to 0$ for bounded times $t_n = n\Delta t \leq T$. We note that the flow $\widetilde{\Phi}_{\Delta t}(z)$ of the modified differential equation (4.4) satisfies

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

for all $M \geq 0$, and smooth test functions ϕ, and where $F_j \phi = f_j \cdot \nabla \phi,\ j = 0, \ldots, s$ and $f_0 = f$. Note that the $\mathcal{O}(\Delta t^{M+1})$ terms in (4.5) are independent of $\Delta t \to 0$ but depend on M, s and ϕ[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. 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.

[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 $\mathcal{O}$ remainder.

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*

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

where $q > r$*. Then one obtains*

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

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). □

Remark 4.2. *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.*

An immediate corollary of Theorem (4.1) is that, if (4.7) holds, then for Δt sufficiently small, the estimator $\widehat{\pi}_T$ given by (1.8) satisfies

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

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 4.3. *There exist constants* $C > 0$ *and* $\lambda > 0$ *independent of* Δt *such that, for* Δt *sufficiently small,*

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

Remark 4.4. *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 B.1 that Assumption 4.3 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

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

Here we rescale the asymptotic variance with Δt, to guarantee a well–defined limit when $\Delta t \to 0$. Assumption 4.3 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 4.5 which quantifies the effect of the time-discreteness on the asymptotic variance. In Theorem 4.6 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 4.5. *For all $\phi \in C^\infty(\mathbb{T}^d)$, such that $\pi(\phi) = 0$ there exists a smooth function R_ϕ such that for Δt sufficiently small,*

$$\left(\frac{I - P_{\Delta t}}{\Delta t}\right)^{-1} \phi(x) = (-\mathcal{L})^{-1} \phi(x) + \frac{\Delta t}{2} \left(\frac{I - P_{\Delta t}}{\Delta t}\right)^{-1} (-\mathcal{L})\phi(x) - \frac{\Delta t^2}{6} \left(\frac{I - P_{\Delta t}}{\Delta t}\right)^{-1} (-\mathcal{L})^2 \phi + \Delta t^3 R_\phi, \tag{4.12}$$

where R_ϕ is bounded, independent of Δt. In particular, for $f \in C^\infty(\mathbb{T}^d)$,

$$\sigma^2_{\Delta t}(f) = \sigma^2(f) + \frac{\Delta t^2}{6} \left\langle (-\mathcal{L})\left(f - \pi(f)\right), f - \pi(f) \right\rangle_\pi + o(\Delta t^2).$$

Proof. The proof can be found in Section 8.2. □

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 4.6. *Suppose that, for some $k \in \mathbb{N}$, $k \geq 1$, there exist operators $A_0, \dots, A_k$ on $C^\infty(\mathbb{T}^d)$, bounded uniformly with respect to Δt, where $A_i = \frac{\mathcal{L}^{i+1}}{(i+1)!}, i = 0, \cdots, k-1$ 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*

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

where

$$R_1(f,g) = \left\langle \left(\frac{I - \widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t} \left(\frac{\mathcal{L}^{k+1}}{(k+1)!} - A_k\right) \left(\frac{I - P_{\Delta t}}{\Delta t}\right)^{-1} f, M_{\Delta t} g \right\rangle_{\widehat{\pi}^{\Delta t}}. \tag{4.14}$$

In particular

$$\widehat{\sigma}^2_{\Delta t}(f) = \sigma^2_{\Delta t}(f) + 2\Delta t^k R_1(f,f) + o(\Delta t^k). \tag{4.15}$$

Moreover, we can write the remainder term as

$$R_1(f,g) = \left\langle (-\mathcal{L})^{-1} \left(\frac{\mathcal{L}^{k+1}}{(k+1)!} - M_0 A_k\right) (-\mathcal{L})^{-1} f, g \right\rangle_\pi + o(\Delta t^k), \tag{4.16}$$

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

Proof. The proof can be found in Section 8.2. □

Remark 4.7. *It is interesting to note that contrary to the case of the asymptotic bias in Theorem 4.1, the order of error for the discrete asymptotic variance in Theorem 4.6 depends crucially on the order of the weak convergence of the underlying numerical integrator. Furthermore, we see from equation* (4.15) *that if the weak order of the integrator is higher than two ($k > 2$) then the leading order error term between $\widehat{\sigma}^2_{\Delta t}(f)$ and the asymptotic variance of the continuous process $\sigma^2(f)$ equals to the leading order term of difference between $\sigma^2_{\Delta t}(f)$ and $\sigma^2(f)$.*

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

$$\widetilde{\mathcal{L}}_{\Delta t} f = \mathcal{L} f + \Delta t^k \mathcal{L}_k f + \Delta t^{q-1} R_f, \quad f \in C^\infty(\mathbb{T}^d), \tag{4.17}$$

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

$$\left\| \left(-\widetilde{\mathcal{L}}_{\Delta t}\right)^{-1} \right\|_{L^\infty_0(\widehat{\pi}^{\Delta t})} < K, \tag{4.18}$$

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 4.5 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 4.8. *Consider a diffusion process $\widetilde{X}_t$ on $\mathbb{T}^d$ with smooth coefficients and generator $\widetilde{\mathcal{L}}_{\Delta t}$ which satisfies (4.17) and (4.18). Suppose that $\widetilde{X}_t$ has unique invariant distribution $\widehat{\pi}^{\Delta t}$ which satisfies*

$$\int \psi(x)\widehat{\pi}^{\Delta t}(x)\,dx = \int \psi(x)\pi(x)\,dx + \Delta t^r R_\psi, \tag{4.19}$$

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 $\pi(f) = 0$,

$$\widetilde{\sigma}^2_{\Delta t}(f) = \sigma^2_{\Delta t}(f) + 2\Delta t^k R_f + o(\Delta t^k). \tag{4.20}$$

where

$$R_f = \left\langle \left(-\widetilde{\mathcal{L}}_{\Delta t}\right)^{-1} M_{\Delta t}(-\mathcal{L}_k)\,(-\mathcal{L})^{-1} f, M_{\Delta t} f \right\rangle_{\widehat{\pi}^{\Delta t}}. \tag{4.21}$$

Moreover, we can express the remainder term as

$$R_f = \left\langle (-\mathcal{L})^{-1} M_0(-\mathcal{L}_k)\,(-\mathcal{L})^{-1} f, f \right\rangle_\pi + o(\Delta t^k), \tag{4.22}$$

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

Proof. The result follows from an argument similar to that of Theorem 4.6. □

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 the integrator $\Theta_{\Delta t}$ used for the reversible dynamics is invariant with respect to π 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*

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

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

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

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

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

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 8.3. 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 $j = 0, \ldots, r-1$. 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

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

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

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

For $\epsilon \asymp \Delta t$, 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 7. 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 practice, the stiffness of the ODE (1.11) would impose additional constraints on β.

5.2 Asymptotic variance of the splitting scheme

In this section we characterise the asymptotic variance of the splitting scheme (1.10). Unlike the bias estimates obtained in Theorem 5.1 the resulting variance estimates will depend on the choice of integrator for the reversible dynamics $\Theta_{\Delta t}$. 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

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

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

Proposition A.1 in the Appendix implies that the reversible integrator $\Theta_{\Delta t}$ satisfies the following perturbation expansion

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

where $\mathcal{G}_1 = \mathcal{S}$ is the symmetric part of $\mathcal{L}$ in $L^2(\pi)$, $\mathcal{G}_2$ is given by (A.2), 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 4.6 and is proved in Section 8.4.

Theorem 5.3. *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*

$$\widehat{\sigma}^2_{\Delta t}(f) = \sigma^2(f) + \Delta t \left\langle (-\mathcal{L})^{-1}(\mathcal{L}^2 - 2\left(\mathcal{A}_2 + \mathcal{G}_1\mathcal{A}_1 + \mathcal{G}_2\right)(-\mathcal{L})^{-1}(f - \pi(f), f - \pi(f)\right\rangle_\pi + o(\Delta t).$$

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

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

where $\mathcal{S}$ and $\mathcal{A}$ are the symmetric and antisymmetric parts of $\mathcal{L}$ 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 5.3 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.

6 Gaussian target distributions

In Sections 5.1 and 5.2, the asymptotic bias and variance for estimators based on Lie-Trotter splitting scheme (1.10) were characterised in terms of stepsize Δt and magnitude of the nonreversible perturbation β. This detailed analysis was however restricted to the case of $\mathbb{T}^d$–valued diffusions, as a similar analysis for $\mathbb{R}^d$ would be significantly more involved (see for example Kopec (2014)). To demonstrate that analogous expressions for the asymptotic variance and bias can be derived in the $\mathbb{R}^d$ case, in this section we consider the class of linear SDEs given by

$$dX_t = -AX_t dt + dW_t \tag{6.1}$$

where $X_t \in \mathbb{R}^d$, W_t is a standard d-dimensional Brownian motion. In the case where $-A$ is stable the dynamics generated by (6.1) are ergodic with respect to $\mathcal{N}(0, \Sigma_\infty)$ where Σ_∞ satisfies the Lyapunov equation Gardiner (1985):

$$A\Sigma_\infty + \Sigma_\infty A^T = I. \tag{6.2}$$

We shall consider a vector field γ satisfying (1.3) which is given by

$$\gamma(z) = \beta J A z,$$

where J is a skew symmetric matrix, and β is a free parameter. Hence (1.2) becomes

$$dX_t = -(I - \beta J) A X_t \, dt + dW_t. \tag{6.3}$$

The fact that equation (6.3) is linear implies that is amenable to very detailed analysis, as for certain classes of numerical schemes, one can find another linear SDE that the numerical method solves exactly in the weak sense. We explain this idea further in Section 6.1, while in Section 6.2 we extend the formula for the asymptotic variance from Duncan et al. (2016) to linear diffusions with a general positive definite diffusion tensor. This allows the use of the modified equation analysis presented in Section 6.1 not just to study the infinite time bias of numerical schemes applied to (6.3), but also the asymptotic variance. This is discussed further in Section 6.3 in the context of a simple two dimensional example. In particular, using the results from the previous two sections, we are able to calculate analytically the expressions for the asymptotic bias and variance of our numerical integrator, which allows in turn for an analytic expression for the mean square error of the integrator at some time T. We then use this expression to explore the bias-variance tradeoff as a function of the properties of the integrator. This provides valuable insights, that we later use in Section 7 when studying more complicated numerical examples.

6.1 Exact modified equation

Consider a one step method applied to (6.3)

$$\widehat{X}^{\Delta t}_{n+1} = B(\Delta t)\widehat{X}^{\Delta t}_n + f(\Delta t, \omega), \quad \widehat{X}^{\Delta t}_0 = x_0, \tag{6.4}$$

where $f(\Delta t, \omega)$ is the flow map for the noise process and $B(\Delta t) \in \mathbb{R}^{d\times d}$ satisfies $B(0) = I$. For an Euler-Maruyama discretisation of (6.3),

$$\begin{aligned} B(\Delta t) &= I - \Delta t (I - \beta J) A, \\ f(\Delta t, \omega) &= \sqrt{\Delta t}\xi, \end{aligned}$$

where $\xi \in \mathbb{R}^d$ satisfies $\xi \sim \mathcal{N}(0, I)$. The fact that (6.4) remains linear implies that the solution $\widehat{X}^{\Delta t}_n$ remains Gaussian at all times, assuming a deterministic initial condition x_0. This implies Zygalakis (2011), that the numerical solution (6.4) satisfies exactly in the weak sense at all times the following stochastic differential equation

$$d\widetilde{X}_t = -\widetilde{B}\widetilde{X}_t + \widetilde{\Sigma}^{1/2} dW_t \tag{6.5}$$

where $\widetilde{B} \in \mathbb{R}^{d\times d}$ and $\widetilde{\Sigma} \in \mathbb{R}^{d\times d}_{sym}$ are defined by

$$\widetilde{B} = -\frac{\log(B(\Delta t))}{\Delta t}, \tag{6.6a}$$

$$B(\Delta t)\widetilde{\Sigma}B(\Delta t)^T - \widetilde{\Sigma} = \widetilde{B}Z + Z\widetilde{B}^T, \tag{6.6b}$$

where $Z = \mathbb{E}(ff^T)$. For sufficiently small Δt one can show that (6.5) is ergodic with respect to $\mathcal{N}(0, \widetilde{\Gamma})$ where $\widetilde{\Gamma}$ satisfies the following Lyapunov equation

$$\widetilde{B}\widetilde{\Gamma} + \widetilde{\Gamma}^T \widetilde{B} = \widetilde{\Sigma}. \tag{6.7}$$

Thus, by solving this equation we can obtain an expression for the invariant measure that the numerical scheme is ergodic with respect to, and hence have an explicit expression for the asymptotic bias of the numerical method.

6.2 Asymptotic variance

By extending the results from Duncan et al. (2016) one can calculate the asymptotic variance for (6.5). In particular if we consider the SDE (6.1) our objective is to derive an explicit expression for the asymptotic variance $\sigma^2(f)$ of

$$I_t = \frac{1}{t}\int_0^t f(\widetilde{X}_s)\,ds,$$

where f is a function of the form

$$f(x) = x \cdot Mx + L \cdot x + K, \tag{6.8}$$

for some $M \in \mathbb{R}^{d\times d}_{sym}$, $L \in \mathbb{R}^d$ and $K \in \mathbb{R}$. To do so we make use of the following result which is a slight generalisation of (Duncan et al. 2016, Proposition 1).

Proposition 6.1. *Consider the linear diffusion defined by the SDE,*

$$dX_t = -AX_t\,dt + \sigma\,dW_t, \tag{6.9}$$

where W_t is a m-dimensional Brownian motion, $\sigma \in \mathbb{R}^{d\times m}$ such that $\Sigma = \sigma\sigma^\top$ is positive definite and $-A$ is stable. Then for the observable of the form (6.8) *the asymptotic variance $\sigma^2(f)$ of the estimator $\pi_t(f)$ for $\pi(f)$ is given by*

$$\sigma^2(f) = 4\mathit{Tr}\left[Q_\infty\left(\int_0^\infty e^{-A^\top t}Me^{-At}\,dt\right)Q_\infty M\right] + 2L\cdot A^{-1}Q_\infty L,$$

where Q_∞ is the covariance of the invariant distribution of X_t, given by the solution of the Lyapunov equation

$$AQ_\infty + Q_\infty A^\top = \Sigma.$$

The proof of this result is deferred to Appendix C.

6.3 Example

We now consider the linear diffusion (6.3) where

$$A = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix},$$

for which we know that the stationary covariance satisfies

$$Q_\infty = \begin{pmatrix} \frac{1}{2\alpha} & 0 \\ 0 & \frac{1}{2\alpha} \end{pmatrix}.$$

We now study the properties of integrators where the $\Phi_{\Delta t}$ and $\Theta_{\Delta t}$ in (1.10) are given by

$$\Phi_{\Delta t}(z) = \left(I + \Delta t\beta JA + \frac{\Delta t^2}{2}(\beta JA)^2 + \cdots + \frac{\Delta t^p}{p!}(\beta JA)^p\right)z, \tag{6.10a}$$

$$\Theta_{\Delta t}(z) = e^{-A\Delta t}z + \sigma_{\Delta t}\xi, \tag{6.10b}$$

where

$$\sigma_{\Delta t}\sigma_{\Delta t}^T = \int_0^{\Delta t} e^{-A(\Delta t-s)}\Sigma e^{-A^T(\Delta t-s)}ds = \frac{1}{2\alpha}\left[1 - e^{-2\alpha\Delta t}\right]I.$$

More precisely we solve the reversible part of the dynamics exactly, while we apply a Taylor-based method of order p to the nonreversible part of the dynamics. We note here that the exact solution of the reversible part of the dynamics is only possible because the dynamics are linear. A further consequence of the linearity of the dynamics is that it is possible to conserve the invariant measure for the reversible part by using the θ method with $\theta = \frac{1}{2}$, see Abdulle et al. (2014) . Hence we will also consider the integrator $\widetilde{\Theta}_{\Delta t}(z)$ given by

$$\widetilde{\Theta}_{\Delta t}(z) = \left(I + \frac{\Delta t}{2}A\right)^{-1}\left[\left(I - \frac{\Delta t}{2}A\right)z + \sqrt{\Delta t}\xi\right] \tag{6.11}$$

The other interesting feature of (6.11) is that even though not exact like (6.10), when metropolised, proposals generated from by (6.11) will be accepted almost surely. For nonlinear problems, the reversible dynamics cannot be integrated exactly, and it is impossible to construct an exact solution like (6.10b) or a solution like (6.11) that conserves the invariant measure, and hence always gets accepted in a Metropolis step. Hence one would replace these integrators with one that conserves the invariant measure by introducing a Metropolisation step, and Theorem 5.1 would still hold.

Study of the invariant measure bias. We now study the properties of the numerical invariant measure using (6.6). We use Mathematica to symbolically calculate the solutions to (6.6) and then obtain an expression for the numerical invariant measure, when a first and a second order numerical method is used to solve the nonreversible part of the diffusion by solving the corresponding Lyapunov equation (6.7). We note here that in order for this equation to make sense one needs $\widetilde{B}$ to be positive definite, which clearly gives rise to time-step restrictions as a function of α, β. We now present in Tables 1 and 2 exact expressions for the first element of the covariance matrix of the numerical invariant measure based on the Lie-Trotter splitting (1.10), for different ordering of the splitting and different choices of integrators for the reversible and nonreversible part. Furthermore, in Figure 1a we plot the 2-norm of the difference between the covariance matrix of the numerical method and the true covariance matrix Σ_∞ when the nonreversible part is solved first and when the θ-method with $\theta = 1/2$ is used for the reversible part[2]. As we can see the order of convergence is always odd. This was also observed in Abdulle et al. (2015) and it relates with the fact that for the deterministic methods used here, the coefficient f_p in Theorem 5.1 is always zero when p is even, hence giving the extra order of convergence observed in Figure 1a. Additionally in Figure 1b we plot the asymptotic bias of Δt when a numerical integrator of order 1 is used to solve the nonreversible part for different values of β. As we can see, the larger the value of β the larger the asymptotic bias.

Study of the asymptotic variance. We now study the properties of the asymptotic variance using (6.6). In particular the idea is that since our numerical solution satisfies exactly in the weak sense the corresponding modified equation then it is enough to look at Proposition 6.1 where A and σ are now replaced with the modified coefficients (6.6). Similarly to the case of the invariant measure bias we use Mathematica to symbolically calculate the solutions to (6.6) and then obtain an expression for the

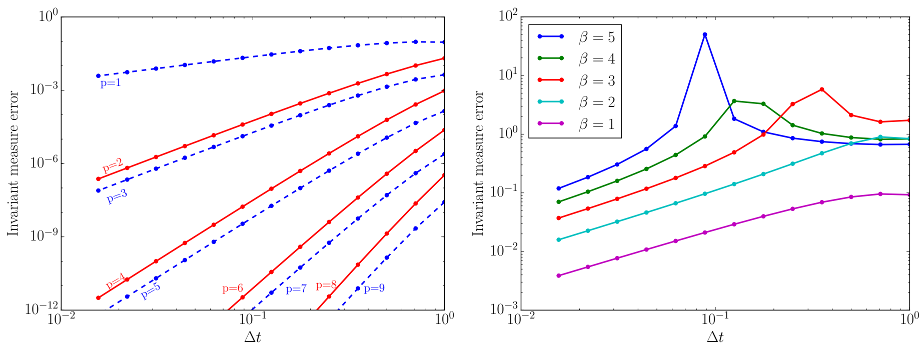
[2]We have not included any of the other possible combinations of ordering of splitting and numerical integrators for the reversible part as the results are qualitatively the same

	Reversible first	Non reversible first
$p=1$	$\frac{(1-e^{-2\alpha\Delta t})(1+\alpha^2\beta^2\Delta t^2)}{2\alpha[1-e^{-2\alpha\Delta t}(1+\alpha^2\beta^2\Delta t^2)]}$	$\frac{1-e^{-2\alpha\Delta t}}{2\alpha[1-e^{-2\alpha\Delta t}(1+\alpha^2\beta^2\Delta t^2)]}$
$p=2$	$\frac{(1-e^{-2\alpha\Delta t})(\alpha^4\beta^4\Delta t^4+4)}{2\alpha(e^{-2\alpha\Delta t}(\alpha^4\beta^4\Delta t^4+4)-4)}$	$\frac{2(1-e^{-2\alpha\Delta t})}{\alpha[4-e^{-2\alpha\Delta t}(4+\alpha^4\beta^4\Delta t^4)]}$

Table 1: First component of numerical invariant covariance when the reversible part is solved exactly

	Reversible first	Non reversible first
$p=1$	$\frac{4+4\alpha^2\beta^2\Delta t^2}{8\alpha-4\alpha^2\beta^2\Delta t+4\alpha^3\beta^2\Delta t^2-\alpha^4\beta^2\Delta t^3}$	$\frac{4}{8\alpha-4\alpha^2\beta^2\Delta t+4\alpha^3\beta^2\Delta t^2-\alpha^4\beta^2\Delta t^3}$
$p=2$	$\frac{4(4+\alpha^4\beta^4\Delta t^4)}{\alpha(32-\alpha^3\beta^4\Delta t^3(2-\alpha\Delta t)^2)}$	$\frac{16}{\alpha(32-\alpha^3\beta^4\Delta t^3(2-\alpha\Delta t)^2)}$

Table 2: First component of numerical invariant covariance when the reversible part is solved by θ-method for $\theta = \frac{1}{2}$.



(a) Error when using nonreversible integrator of order $p = 1, 2, 3, \ldots, 9$

(b) Errors for different values of β.

Figure 1: Accuracy of the numerical invariant measure (covariance matrix error) of the Lie-Trotter splitting scheme, for different orders p of deterministic integrator, and for different values of β respectively. A first order scheme is used in the second figure.

asymptotic variance [3], when a first and a second order numerical method is used to solve the nonreversible part of the diffusion. In particular, we take $K = 0$, $L = 0$ and

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

in (6.8), to find that when the reversible part is solved exactly (or with the θ-method) that for $p = 1$, we have

$$\widetilde{\sigma}^2_{\Delta t}(f) = \frac{\beta^2 + 2}{4a^3\left(\beta^2 + 1\right)} + \frac{\left(3\beta^6 + 7\beta^4 + 6\beta^2\right)\Delta t}{8a^2\left(\beta^2 + 1\right)^2} + O\left(\Delta t^2\right),$$

[3]This again only makes sense if $\widetilde{B}$ is positive definite which gives rise to restrictions to Δt as a function of α, β

independently of the ordering of the splitting, while for $p = 2$ we have independently of the order of the splitting

$$\widetilde{\sigma}^2_{\Delta t}(f) = \frac{\beta^2 + 2}{4a^3 \left(\beta^2 + 1\right)} - \frac{\beta^4 \Delta t^2}{12a \left(\beta^2 + 1\right)^2} + O\left(\Delta t^3\right),$$

when the reversible part is solved exactly, while when solved with the θ method

$$\widetilde{\sigma}^2_{\Delta t}(f) = \frac{\beta^2 + 2}{4a^3 \left(\beta^2 + 1\right)} - \frac{\left(5\beta^4 + \beta^2 + 2\right) \Delta t^2}{48 \left(a \left(\beta^2 + 1\right)^2\right)} + O\left(\Delta t^3\right),$$

again independently of the order of the splitting. We note here that these results agree with Proposition 4.8, since for $p = 1$ the leading order perturbation in terms of the continuous time variance is $\mathcal{O}(\Delta t)$ while for $p = 2$ is $\mathcal{O}(\Delta t^2)$.

Mean Square Error. Having obtained analytical expressions for the asymptotic bias of the invariant measure as well as for the asymptotic variance of the corresponding numerical schemes, we combine them in order to study the mean square error. More precisely, decomposing the MSE into bias and variance,

$$\mathbb{E}|\widehat{\pi}_T(f) - \pi(f)|^2 = (\mathbb{E}\widehat{\pi}_T(f) - \pi(f))^2 + \mathbb{E}(\widehat{\pi}_T(f) - \mathbb{E}\widehat{\pi}_T(f)) = \widehat{\mu}_T^2 + \widehat{\sigma}_T^2,$$

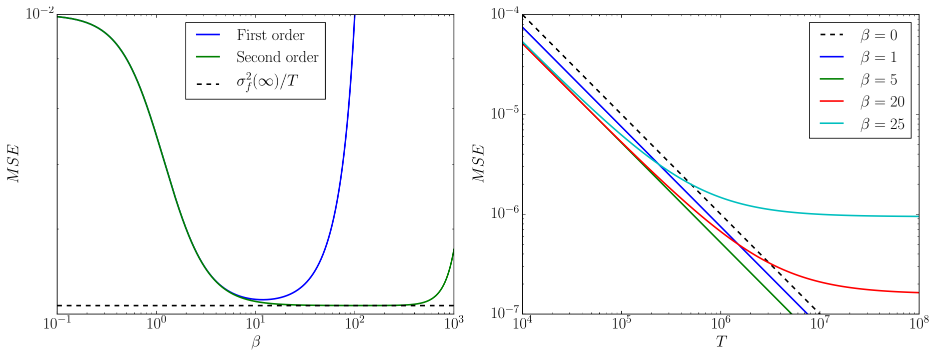
we approximate $\widehat{\mu}_T$ by the invariant measure bias, while on the other hand

$$\widehat{\sigma}_T^2 \simeq \frac{\widehat{\sigma}^2(f)}{T}.$$

We now plot in Figure 2a the MSE when a first and a second order numerical method is used to solve the nonreversible part and the reversible part is solved exactly. In particular, we choose our timestep $\Delta t = 10^{-4}, \alpha = 1, T = 10^3$ and we study the influence of β on the MSE. As can be seen in both cases there is a range of values of the parameter β for which the MSE is reduced almost to the theoretical minimum $\sigma_f^2(\infty)/T$ attainable using this choice of dynamics Duncan et al. (2016), where $\sigma_f^2(\infty) = \sigma_f^2(0)/2 = \frac{1}{4}$. This range of parameters is wider in the case of case of the higher order integrator since the asymptotic bias is $\mathcal{O}(\Delta t^3)$ and will not dominate the MSE for a wider range of β values. However, one would expect that as the simulation time T is increased this benefit will be lost, as the corresponding reduction in variance becomes less significant. This tradeoff is indeed illustrated further in Figure 2b. Here, the MSE is plotted as a function of time when a second order integrator for the nonreversible part is used and $\Delta t = 10^{-2}$. We see that while increasing β leads to smaller MSE initially, eventually the additional bias (as also observed in Figure 1b) will be the dominant contribution to the MSE, and thus for large enough T the exact scheme (i.e. $\beta = 0$) will always have better MSE.

7 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.



(a) MSE for fixed time.

(b) MSE for varying time.

Figure 2: MSE for different methods applied the nonreversible part as a function of β and T.

7.1 Warped Gaussian distribution

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

$$\pi(x) \propto \exp\left(-\frac{x_1^2}{100} - (x_2 + bx_1^2 - 100b)^2\right), \tag{7.1}$$

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 3, 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 4 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, RWMH and Barker rule in Figures 4a, 4b and 4c, 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 5.3, 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.

7.2 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

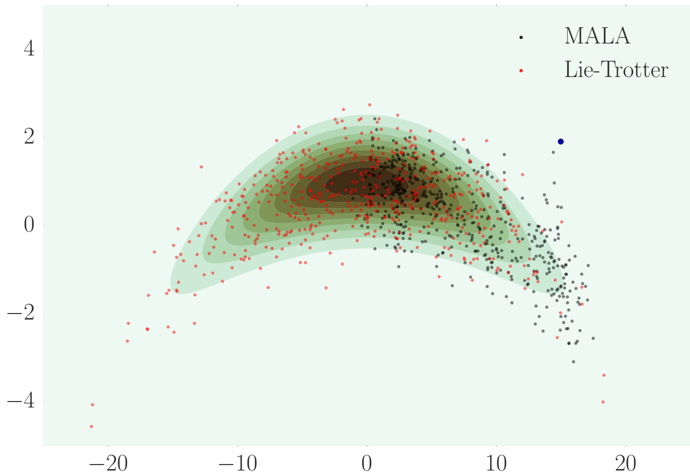
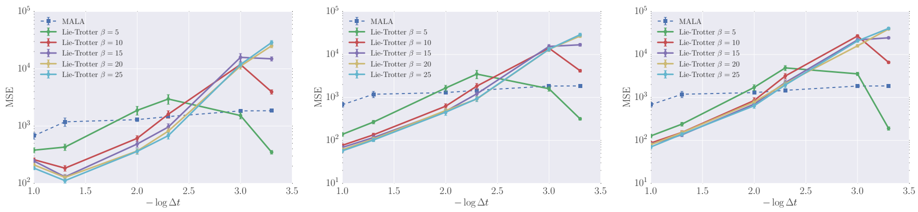


Figure 3: Typical trajectories for MALA and Lie-Trotter splitting scheme applied to the warped Gaussian distribution (7.1), with computational budget of 3200 density evaluations. Both schemes started from $x = (15, 2)$ depicted by a blue dot.



(a) MALA for the reversible part

(b) RMWH rule for the reversible part

(c) Barker Scheme for the reversible part

Figure 4: Comparison of the MSE between MALA and different nonreversible samplers applied to the warped Gaussian distribution (7.1). 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.

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

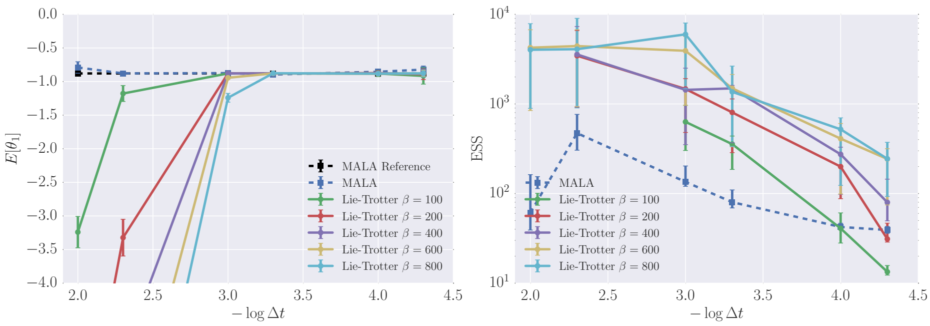
$$\pi(\theta|(X, Y)) \propto \exp\left(\sum_{i=1}^{m} Y_i \theta^T X_i - \log\left(1 + e^{\theta^T X_i}\right) - \frac{1}{2}\theta^T \Sigma^{-1} \theta\right). \qquad (7.2)$$

In Figure 5 we investigate the use of the Lie Trotter sampler applied to this problem for the Pima Indians[4] dataset obtained from the UCI machine learning repository. The skew symmetric matrix J is chosen by generating a random permutation $\sigma(1), \dots, \sigma(d)$ and setting

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

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

for $i = 1, \dots, d-1$, and zero elsewhere. In Figure 5a 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 5b 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

Figure 5: Confidence interval for an estimator of $Y_{1,1}$ and ESS for estimators for $\pi(\theta_i)$, $i = 1, \dots, 9$ 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.

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

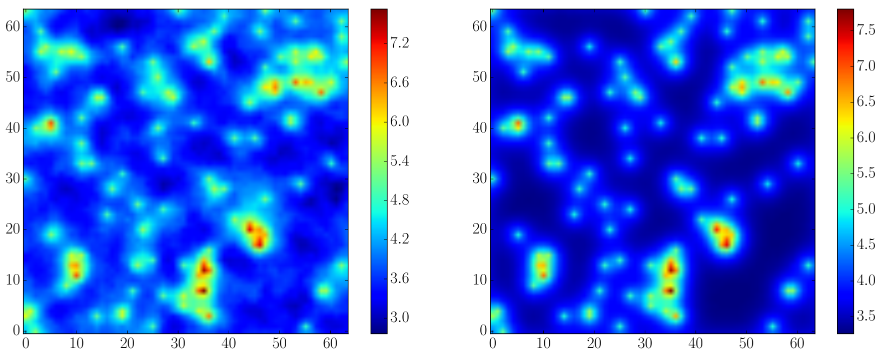
$$\Sigma_{i,j,i',j'} = \sigma^2 \left[-\frac{-\{(i-i')^2 + (j-j')^2\}^{1/2}}{64\beta} \right], \quad i, j, i', j' = 1 \cdots, 64.$$

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 $\mathcal{O}(d^3)$. However, in the case of the nonreversible samplers, the nonreversible perturbation compensates for the poor scaling, thus rendering this reparametrisation unnecessary.

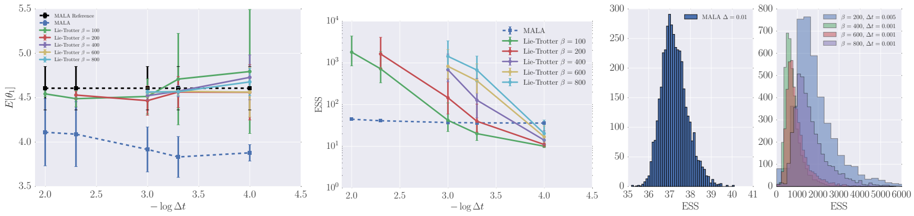
In Figure 6 we plot an estimator of $\mathbb{E}(\Lambda \mid x)$ using MALA and its nonreversible counterpart respectively. For this computation the skew-symmetric matrix J was generated randomly as in the logistic regression example. Due to the large number of covariates, for any given random choice of J, there would be a small number of covariates 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 7. In Figure 7c a histogram of the ESS over all covariates 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 7b 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 7a. 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

Figure 6: $\mathbb{E}(Y_{i,j})$ estimated using different schemes. The computational budget is set to $N = 3.5 \cdot 10^3$ gradient evaluations.



(a) $\mathbb{E}(Y_{1,1})$ vs stepsize (b) ESS vs stepsize (c) Histogram of ESS

Figure 7: 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.

8 Proofs of the main results

In this section we prove the main results of the paper. In particular, in Section 8.1 we prove the geometric ergodicity of the splitting scheme (1.10), while in Sections 8.3 and 8.4 we prove the results related to the asymptotic bias and variance of the splitting method.

8.1 Ergodicity of the splitting scheme

Here we prove the theorems and corollaries stated in Section 3.

Proof of Theorem 3.2. We verify the criteria for geometric ergodicity formulated in Chapters 15 and 16 of Meyn & Tweedie (1993*a*).

1. We show that $P_{\Delta t}(x, \cdot)$ is π-irreducible. Let $A \subset \mathbb{R}^d$ such that $\pi(A) > 0$, then

$$\begin{aligned} P_{\Delta t}(x, A) = &\int_A q_{\Delta t}(y \,|\, \Phi_{\Delta t}(y)) \alpha(\Phi_{\Delta t}(x), y)\, dy \\ &+ \mathbf{1}_A(x) \int_{\mathbb{R}^d} q_{\Delta t}(z \,|\, \Phi_{\Delta t}(x))(1 - \alpha(z, \Phi_{\Delta t}(x)))\, dz > 0, \end{aligned}$$

   which implies that $P_{\Delta t}$ is π-irreducible.

2. We now show that every compact set C of positive measure is small. To this end, let C be such a set and B a measurable subset of C. Then $D = C \cup \Phi_{\Delta t}(C)$ is also a compact set of positive measure. Since the target density π and proposal $q_{\Delta t}(y|x)$ are positive and continuous for all x, y, applying (Mengersen & Tweedie 1996, Lemma 1.2) implies that there exists $\eta > 0$ such that

$$\widetilde{P}_{\Delta t}(x, B) \geq \eta \pi(B), \quad B \subset D, x \in D.$$

   In particular,

$$\widehat{P}_{\Delta t}(x, B) = \widetilde{P}_{\Delta t}(\Phi_{\Delta t}(x), B) \geq \eta \pi(B), \quad B \subset C, x \in C, \tag{8.1}$$

   so that C is small. Aperiodicity of the chain follows immediately from (8.1).

3. To complete the proof we show that $P_{\Delta t}$ satisfies a Foster-Lyapunov condition for the Lyapunov function V. To this end using (3.4), given $x \in \mathbb{R}^d$:

$$\begin{aligned}
\widehat{P}_{\Delta t} V(x) &\le \lambda V(\Phi_{\Delta t}(x)) + b\mathbf{1}_C(\Phi_{\Delta t}(x)) \\
&\le \lambda V(\Phi_{\Delta t}(x)) + b\mathbf{1}_{\Phi^{-1}_{\Delta t}(C)}(x) \\
&\le \lambda V(x) + \lambda\left(V(\Phi_{\Delta t}(x)) - V(x)\right) + b\mathbf{1}_{\Phi^{-1}_{\Delta t}(C)}(x).
\end{aligned}$$

By Assumptions 3.1(2) and 3.1(3), there exists a compact set $D \subset \mathbb{R}^d$ and $0 < c < 1$ such that $\Phi^{-1}_{\Delta t}(C) \subset D$ and moreover

$$\lambda V(x) + \lambda\left(V(\Phi_{\Delta t}(x)) - V(x)\right) \le cV(x), \quad x \in \mathbb{R}^d \setminus D,$$

which implies that

$$\widehat{P}_{\Delta t} V(x) \le cV(x) + b\mathbf{1}_D(x),$$

as required.

□

Proof of Corollaries 3.3 and 3.4. By (Roberts & Tweedie 1996, Theorem 3.2), the RWMH chain satisfies a Foster drift condition for the Lyapunov function $V(x) = \pi^{-1/2}(x)$, so that (A2) of Assumptions 3.1 holds. In particular, for $U(x) = \log \pi(x)$ we have

$$\frac{V(\Phi_{\Delta t}(x)) - V(x)}{V(x)} = e^{(U(\Phi_{\Delta t}(x)) - U(x))/2} - 1,$$

so that, by (3.6) given $\Delta t > 0$, there exists $M > 0$ such that $(V(\Psi_{\Delta t}(x)) - V(x))/V(x) < \delta$ for all $|x| \ge M$. Thus, Assumption 3.1(2) follows immediately. Moreover, we note that (3.6) implies that there exists $K > 0$ such that

$$|\Phi_{\Delta t}(x)| - K \le |x| \le |\Phi_{\Delta t}(x)| + K, \quad x \in \mathbb{R}^d,$$

from which 3.1(3) follows immediately. Hence, the conditions of Theorem 3.2 all hold, and thus the process $\widehat{X}^{\Delta t}_n$ is geometrically ergodic.

In the MALA case, provided that the conditions of (Roberts & Tweedie 1996, Theorem 4.1) hold, then the chain satisfies a Foster-Lyapunov condition for $V(x) = e^{s|x|}$ for $s > 0$ sufficiently small. If we consider

$$\frac{V(\Phi_{\Delta t}(x)) - V(x)}{V(x)} = e^{s(|\Phi(x)| - |x|)} \le e^{s(|\Phi(x)| - |x|)},$$

then Assumption 3.1(2) follows immediately.

Suppose now that $\gamma = J\nabla\pi^{\alpha}$, where $J = -J^{\top}$ and $\alpha > 0$. Suppose $\Phi_{\Delta t}(x)$ is an explicit Runge-Kutta discretisation of the nonreversible dynamics having s stages. Then we can write

$$\Phi_{\Delta t}(x) = x + h\sum_{i=1}^{s} b_i k_i(x), \tag{8.2}$$

where

$$
\begin{aligned}
k_1(x) &= \gamma(x) \\
k_2(x) &= \gamma\left(x + hw_{2,1}k_1(x)\right) \\
k_3(x) &= \gamma\left(x + h(w_{3,1}k_1(x) + w_{3,2}k_2(x))\right) \\
&\vdots \\
k_s(x) &= \gamma\left(x + h\sum_{i=1}^{s-1} w_{s,i}k_i(x)\right),
\end{aligned}
$$

where $(w_{i,j})$ is the Runge-Kutta matrix associated with the discretisation. By (3.7) there exist positive constants α', K' and K_1 such that

$$k_1(x) \le |\gamma(x)| \le K'|\nabla\pi^{\alpha}(x)| \le K_1\pi^{\alpha'}(x).$$

Suppose now that there exists constants $K_2, \ldots, K_{i-1}$ such that

$$|k_j(x)| \le K_j\pi^{\alpha'}(x), \quad x \in \mathbb{R}^d, \quad j = 1, \ldots, i-1.$$

By (3.7) the matrix $\nabla\gamma = (\partial_{x_i}\gamma_j(x))_{i,j}$ has bounded components in $\mathbb{R}^d$ and so applying the mean value theorem to every component of γ, it follows that

$$|k_i(x)| \le |\gamma(x)| + h\left(\sup_{x\in\mathbb{R}^d}|\nabla\gamma(x)|_{max}\right)\sum_{j=1}^{i-1}|w_{i,j}||k_j(x)| \le K_i\pi^{\alpha'}(x),$$

for some constant K_i. It follows by induction that $|\Phi_{\Delta t}(x) - x| \le K\pi^{\alpha'}(x)$, for all $x \in \mathbb{R}^d$, which implies (3.6). The corresponding result for γ given by (2.5) follows similarly. □

8.2 Asymptotic variance of numerical integrators

Here we prove Proposition 4.5 and Theorem 4.6 which characterises the error in the asymptotic variance for an arbitrary numerical integrator

Proof of Proposition 4.5. It follows from the maximum principle that the operator $(-\mathcal{L})^{-1}$ is bounded on $L_0^\infty(\pi)$. Similarly the operator $\Delta t(I - P_{\Delta t})^{-1}$ is bounded on $L_0^\infty(\pi)$, uniformly with respect to Δt.

Let $\psi \in C^\infty(\mathbb{T}^d)$ with $\pi(\psi) = 0$. There exists R_ψ, smooth and bounded uniformly with respect to Δt such that

$$\left(\frac{I - P_{\Delta t}}{\Delta t}\right)\psi = -\mathcal{L}\psi - \frac{\Delta t}{2}\mathcal{L}^2\psi - \frac{\Delta t^2}{6}\mathcal{L}^3\psi + \Delta t^3 R_\psi, \tag{8.3}$$

provided that Δt is sufficiently small. Hence using (8.3) we obtain

$$
\begin{aligned}
(-\mathcal{L})^{-1}\psi &= \left(\frac{I - P_{\Delta t}}{\Delta t}\right)^{-1}\left(\frac{I - P_{\Delta t}}{\Delta t}\right)(-\mathcal{L})^{-1}\psi \\
&= \left(\frac{I - P_{\Delta t}}{\Delta t}\right)^{-1}\psi + \frac{\Delta t}{2}\left(\frac{I - P_{\Delta t}}{\Delta t}\right)^{-1}\mathcal{L}\psi + \frac{\Delta t^2}{6}\left(\frac{I - P_{\Delta t}}{\Delta t}\right)^{-1}\mathcal{L}^2\psi \\
&\quad + \Delta t^3\left(\frac{I - P_{\Delta t}}{\Delta t}\right)^{-1}R_{(-\mathcal{L})^{-1}\psi}.
\end{aligned}
\tag{8.4}
$$

Since both sides of (8.3) have mean zero and $\pi(\mathcal{L}^i\psi) = 0$ for $i \geq 0$, it follows that $\pi(R_{(-\mathcal{L})^{-1}\psi}) = 0$. Thus the remainder term in (8.4) is well-defined and uniformly bounded with respect to Δt. Now 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 (8.4) it follows that

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

where R_f is a remainder term depending on f. Since f is smooth, we can iteratively apply (8.4) to the second term and third terms on the RHS obtaining

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

as required. □

Proof of Theorem 4.6. 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 $\widehat{P}_{\Delta t}$:

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

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

$$+ \Delta t^{q-1} \left\langle \left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t} R_f, M_{\Delta t} g \right\rangle_{\widehat{\pi}^{\Delta t}}$$

$$= \left\langle \left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t} f, M_{\Delta t} g \right\rangle_{\widehat{\pi}^{\Delta t}}$$

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

$$+ \Delta t^{q-1} \left\langle \left(\frac{I-\widehat{P}_{\Delta t}}{\Delta t}\right)^{-1} M_{\Delta t} R_f, M_{\Delta t} g \right\rangle_{\widehat{\pi}^{\Delta t}}, \tag{8.5}$$

where R_f is a smooth function depending on f, bounded uniformly with respect to Δt. By Assumption 4.3, the coefficients of the Δt^k and Δt^{q-1} terms are bounded uniformly with respect to Δt. Equation (4.13) then follows immediately, and thus (4.15). Noting that $M_{\Delta t} = M_{\Delta t} M_0$ then by applying (8.5) with

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

we obtain

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

$$= \left\langle \left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} \left(\frac{\mathcal{L}^{k+1}}{(k+1)!} - M_0 A_k\right) \left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f, g \right\rangle_{\pi} + \Delta t^{q-1} R_2(f,g),$$

for some smooth, uniformly bounded remainder term R_2. We now apply (4.12) to the discrete generator $\Delta t^{-1}(I - P_{\Delta t})$ to obtain

$$\left\langle \left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} \left(\frac{\mathcal{L}^{k+1}}{(k+1)!} - M_0 A_k\right) \left(\frac{I-P_{\Delta t}}{\Delta t}\right)^{-1} f, g \right\rangle_{\pi} = \left\langle (-\mathcal{L})^{-1} \left(\frac{\mathcal{L}^{k+1}}{(k+1)!} - M_0 A_k\right) (-\mathcal{L})^{-1} f, g \right\rangle_{\pi}$$

$$+ \Delta t R_3(f,g),$$

for a smooth bounded remainder term R_3, from which (4.16) follows.

□

8.3 Asymptotic bias of the splitting scheme

Here we prove the results from Section 5.1

Proof of Theorem 5.1. Assume that the transition semigroup associated with $\widehat{X}^{\Delta t}$ satisfies the expansion (4.6). In order to prove the first part of Theorem 5.1 it is enough to show that

$$A_j^*\pi = 0 \quad \text{for} \quad j = 1, \cdots r-1, \qquad A_r^*\pi = \operatorname{div}(f_r\pi). \tag{8.6}$$

The result then follows immediately from Theorem 4.1 using the identity

$$\int_{\mathbb{T}^d} A_r\psi(z)\pi(z)dz = -\int_{\mathbb{T}^d} \psi(z)\operatorname{div}(f_r(z)\pi(z))dz. \tag{8.7}$$

We now start with the calculation of A_j. In particular, given $\phi \in C^\infty(\mathbb{T}^d)$ and $x \in \mathbb{R}^N$, using the semigroup property of the Markov process we have

$$\mathbb{E}\left[\phi\left(\widehat{X}_1^{\Delta t}\right)|\widehat{X}_0^{\Delta t} = x\right] = \mathbb{E}\left[\phi\left(\Phi_{\Delta t}\circ\Theta_{\Delta t}(x)\right)\right] = e^{\Delta t\mathcal{L}_{S,num}}(\phi\circ\Phi_{\Delta t})(x), \tag{8.8}$$

where $e^{\Delta t\mathcal{L}_{S,num}}\phi$ denotes the numerical flow generated by the numerical method applied to the reversible part of the dynamics (1.12). We next recall the generator (4.5) of the truncated modified equation (4.4) of the integrator $\Phi_{\Delta t}$,

$$\widetilde{\mathcal{L}}_D\phi = F_0 + \Delta t F_1\phi + \ldots + \Delta t^r F_r\phi + \Delta t^{r+1}R_\phi,$$

where R_ϕ is a smooth remainder term bounded uniformly with respect to Δt and where we define the differential operators $F_j\phi = f_j\cdot\nabla\phi$ (with $f_0 = f$). We then have

$$\begin{aligned}
\mathbb{E}\left[\phi\left(\widehat{X}_1^{\Delta t}\right)|\widehat{X}_0^{\Delta t} = x\right] &= \left(\sum_{k=0}^{r}\frac{\Delta t^k\mathcal{L}_{S,num}^k}{k!}\right)\left(\sum_{k=0}^{r}\frac{\Delta t^k\widetilde{\mathcal{L}}_D^k}{k!}\right)\phi(x) + \Delta t^{r+2}R'_\phi \\
&= \phi(x) + \Delta t\mathcal{L}\phi(x) + \sum_{k=1}^{r}\Delta t^{k+1}A_k\phi(x) + \Delta t^{r+2}R'_\phi,
\end{aligned}$$

for a smooth remainder term R'_ϕ and where

$$A_k = \sum_{j=0}^{k+1}\mathcal{L}_{S,num}^{k+1-j}\Big(\sum_{\substack{1\le i\le j\\ n_1+n_2+\cdots+n_i=j-i}}\frac{1}{i!(k+1-j)!}F_{n_1}\cdots F_{n_i}\Big),$$

where the second sum above is over integers $n_1,\ldots,n_i \geq 0$ and is equal to the identity I when $j = 0$. We obtain for all $k \geq 1$,

$$A_k^*\pi = \sum_{j=0}^{k+1}\Big(\sum_{\substack{1\le i\le j\\ n_1+n_2+\cdots+n_i=j-i}}\frac{1}{i!(k+1-j)!}F_{n_i}^*\cdots F_{n_1}^*\Big)(\mathcal{L}_{S,num}^*)^{k+1-j}\pi.$$

Now since the integrator applied to the reversible part preserves the invariant measure we have $\mathcal{L}_{S,num}^*\pi = 0$ which together with $F_i^*\pi = 0,\ i = 1,\cdots r-1$ implies that for $k \leq r$, the only possibly non-zero term in the above sum is obtained for $j = r+1, k = r, i = 1$, i.e., $F_r^*\pi = \operatorname{div}(f_r\pi)$. Hence, we deduce (8.6) which permits to conclude the proof. □

8.4 Asymptotic variance of the splitting scheme

Here we prove the results from Section 5.2.

Proof of Theorem 5.3. Clearly, Assumption 4.3 holds immediately from Theorem B.1 in the Appendix. Consider the one step semigroup $\widehat{P}_{\Delta t} = \Theta_{\Delta t}\widehat{\Phi}_{\Delta t}$ be the one-step semigroup corresponding to the Lie-Trotter splitting scheme (1.10), where $\Theta_{\Delta t}$ is the one-step semigroup integrated by MALA. By Proposition A.1 one obtains

$$\widehat{P}_{\Delta t}\phi = \phi + \Delta t A_0\phi + \Delta t^2 A_1\phi + \Delta t^{5/2}R_\phi,$$

where

$$\begin{aligned} A_0 &= \mathcal{A}_1 + \mathcal{G}_1 = \mathcal{L}, \\ A_1 &= \mathcal{A}_2 + \mathcal{G}_1\mathcal{A}_1 + \mathcal{G}_2, \end{aligned}$$

and where R_ϕ is a smooth remainder term, bounded uniformly with respect to Δt. Since the integrator $\widehat{\Phi}_{\Delta t}$ is assumed to preserve the invariant distribution up to order 2, and $\widehat{\Theta}_{\Delta t}$ preserves π it follows that

$$\pi\left(\left(\mathcal{A}_2 + \mathcal{G}_1\mathcal{A}_1 + \mathcal{G}_2\right)\phi\right) = 0, \quad \phi \in C^\infty(\mathbb{T}^d).$$

Applying Theorem 4.6, it follows that for $f \in C^\infty(\mathbb{T}^d)$,

$$\widehat{\sigma}^2_{\Delta t}(f) = \sigma^2_\Delta(f) + \Delta t R_f + o(\Delta t),$$

where

$$R_f = 2\left\langle(-\mathcal{L})^{-1}(\mathcal{L}^2/2 - \left(\mathcal{A}_2 + \mathcal{G}_1\mathcal{A}_1 + \mathcal{G}_2\right)(-\mathcal{L})^{-1}(f-\pi(f), f-\pi(f)\right\rangle_\pi.$$

Finally, invoking Theorem 4.1 we obtain

$$\widehat{\sigma}^2_{\Delta t}(f) = \sigma^2(f) + \Delta t R_f + o(\Delta t),$$

as required.

□

9 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 $\mathcal{O}(d^3)$ reparametrisation, which is also the computational bottleneck in high dimensions for more sophisticated sampling algorithms such as MMALA Girolami & Calderhead (2011).

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.

Acknowledgements

G.A. Pavliotis is supported by the Engineering and Physical Sciences Research Council of the UK through Grants Nos. EP/P031587, EP/L020564, EP/L024926 and EP/L025159. A. B. Duncan acknowledges the EPSRC for support under EP/J009636/1, EP/L020564/1, EP/K009788/2. K. C. Zygalakis was supported by a grant from the Simons Foundation and by the Alan Turing Institute under the EPSRC grant EP/N510129/1. Part of this work was done during the authors' stay at the Newton Institute for the program Stochastic Dynamical Systems in Biology: Numerical Methods and Applications. The authors would like to thank Mark Girolami, Andrew Stuart, Gabriel Stoltz and Ioannis Papastathopoulos for useful discussions about the paper.

A Expansions for the One-Step Semigroups of the Reversible Dynamics

In this section we present the expansion of the one-step transition semigroups for MALA, which is directly obtained from Fathi et al. (2015), Lelièvre & Stoltz (2016). We shall define $U = -\log \pi$ to be the potential corresponding to the positive target density π.

Proposition A.1. *(Fathi et al. 2015, Lemma 1) Let $P_{\Delta t}$ denote the one-step transition semigroup corresponding to the MALA scheme, then for all smooth $\psi : \mathbb{T}^d \to \mathbb{R}$:*

$$(P_{\Delta t} - I)\psi(x) = \Delta t \mathcal{G}_1 \psi(x) + \Delta t^2 \mathcal{G}_2 \psi + \Delta t^{5/2} r(x),$$

where

$$\mathcal{G}_1 \psi = -\nabla U \cdot \nabla \psi + \Delta \psi, \tag{A.1}$$

and

$$\mathcal{G}_2\psi(x) = \frac{1}{2}\nabla U(x)\cdot\nabla\nabla\psi(x)\nabla U(x) - \frac{1}{3}\nabla U(x)\cdot\nabla\Delta\psi(x) + 2\Delta^2\psi - \sqrt{2}\int (1\wedge T(x,g))\,\langle\nabla\psi(x), g\rangle \frac{e^{-|g|^2/2}}{\sqrt{(2\pi)^d}}\,dg, \tag{A.2}$$

and where $|r| \le C$*, uniformly in* $0 < \Delta t \le 1$*. Moreover, for* Δt *sufficiently small, the acceptance probability of the proposal* $y = x - \nabla U(x)\Delta t + \sqrt{2\Delta t}g$*, where* $g \sim \mathcal{N}(0, I)$ *and* x *satisfies*

$$\alpha(x, x - \nabla U(x)\Delta t + \sqrt{2\Delta t}g) = 1 - \Delta t^{3/2}\left[G_{1/2}(x,g)\right]_+ + \Delta t^2 r(x,g), \tag{A.3}$$

where

$$G_{1/2}(x,g)) = -\frac{1}{3\sqrt{2}}\nabla\nabla\nabla U(x) : g^{\otimes 3} + \frac{1}{\sqrt{2}}\nabla U(x)\cdot\nabla\nabla U(x) g,$$

and $[a]_+ = \max(0,a)$.

B Exponential Ergodicity the Splitting Scheme

In this section we shall show that Assumption 4.3 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 $x \in \mathbb{T}^d$,

$$P_{\Delta t}^{\lceil T/\Delta t\rceil} f(x) \ge \widetilde{\alpha}\int_{\mathbb{T}^d} f\,d\nu, \tag{B.1}$$

where $0 < \Delta t \le \Delta t^*$.

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 B.1. *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* (3.2)*. Then, for* Δt *sufficiently small, the uniform minorisation condition* (B.1) *holds, and as a result, Assumption 4.3 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

$$\widehat{X}^{\Delta t}_n = \widehat{X}^{\Delta t}_0 + \mathcal{G}_n + \mathcal{F}_n, \tag{B.2}$$

where

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

and

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

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 $\mathcal{N}(0,I)$ 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 $\mathcal{G}_n = \widetilde{\mathcal{G}}_n + \widehat{G}_n$ where

$$\widetilde{G}_n = \sqrt{2\Delta t} \sum_{k=0}^{n-1} \mathbf{1}\left[u_k \le 1\right] g_k, \tag{B.3}$$

and

$$\widehat{G}_n = \sqrt{2\Delta t} \sum_{k=0}^{n-1} \left(\mathbf{1}\left[u_k \le \alpha \left(\Phi_{\Delta t}\left(\widehat{X}_k^{\Delta t}\right), \Psi_{\Delta t}\left(\Phi_{\Delta t}\left(\widehat{X}_k^{\Delta t}\right), g_k\right)\right)\right] - \mathbf{1}[u_k \le 1] \right) g_k. \tag{B.4}$$

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.

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

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

$$D(x) = \mathbb{E}_{g\sim\mathcal{N}(0,I)}\left[\left(\alpha\left(\Phi_{\Delta t}(x), \Psi_{\Delta t}(\Phi_{\Delta t}(x), g)\right) - 1\right) g\right]. \tag{B.5}$$

It follows from (A.3) that there exists a constant C independent of Δt such that

$$|D(x)| \le C \Delta t^{3/2}, \tag{B.6}$$

for Δt sufficiently small. Thus, it follows that

$$\Delta t^{1/2} \sum_{k=0}^{n-1} D(\widehat{X}_k^{\Delta t}) \le C \Delta t. \tag{B.7}$$

Similarly one can show that

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

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

$$\mathbb{P}\left[\left| \widehat{\mathcal{G}}_n - \sqrt{2\Delta t} \sum_{k=0}^{n-1} D(\widehat{X}_k^{\Delta t}) \right| \ge \frac{1}{2} \right] \le C'' \Delta t^{1/2}, \tag{B.8}$$

for some constant C'' independent of Δt. Applying (B.7) and choosing Δt sufficiently small we obtain

$$\mathbb{P}\left[\left| \widehat{\mathcal{G}}_n \right| \ge 1 \right] \le \widehat{C} \Delta t \le \frac{1}{2},$$

where $\widehat{C}$ is a constant independent of Δt. The remainder of the argument involves controlling the magnitude of $\mathcal{F}_n$ and the distribution of $\widetilde{\mathcal{G}}_n$ to obtain the minorisation condition (B.1) and follows identically to (Fathi & Stoltz 2015, Sec 4.4). □

C Analysis for Gaussian Distributions

In this section we provide a proof of Proposition 6.1. Taking a different approach to Duncan et al. (2016), we shall obtain this result via the Green-Kubo formula, i.e.

$$\sigma^2(f) = 2\int_0^\infty \langle P_t f - \pi(f), f - \pi(f)\rangle_\pi \, dt, \tag{C.1}$$

where P_t is the semigroup corresponding to the linear diffusion (6.9). We note that for $l > 0$, the process X_t satisfies the Foster-Lyapunov condition (2.6) with Lyapunov function $V_l(x) = 1 + |x|^{2l}$. In particular, by Proposition 2.5, a CLT for the estimator $\pi_T(f) = T^{-1}\int_0^T f(X_t)\, dt$ will hold for all observables f having algebraic growth, and moreover (C.1) is well defined and finite.

Proof of Proposition 6.1. Provided that $-A$ is stable, the process X_t defined by (6.9) is ergodic with unique invariant distribution $\mathcal{N}(0, Q_\infty)$. The stationary covariance Q_∞ can be explicitly writtien as $\lim_{t\to\infty} Q_t$, where

$$Q_t = \int_0^t e^{-As}\Sigma e^{-A^\top s}\, ds$$

We first prove the result for $\Sigma = I$, then obtaining the general case via a simple linear transformation.

It is well known that the semigroup $(P_t)_{t\geq 0}$ is given by this formula (Lorenzi & Bertoldi 2006, Lemma 9.3.6),

$$(P_t f)(t) = \mathbb{E}[f(X_t) \,|\, X_0 = x] = \int f(e^{-At}x + Q_t^{1/2}z)\rho(z)\, dz, \tag{C.2}$$

for all $f \in L^p(\pi)$, $p \in [1, \infty)$ where ρ is the density of a standard Gaussian in $\mathbb{R}^d$. In particular, given $f_1(x) = x \cdot Mx$, with $M \in \mathbb{R}^{d\times d}_{sym}$:

$$P_t\left(f_1 - \pi(f_1)\right)(x) = x \cdot \left[e^{-A^\top t} M e^{-At}\right] x - \mathrm{Tr}\left((Q_\infty - Q_t)M\right).$$

Similarly, given $f_2(x) = L\cdot x$, then $\pi(f_2) = 0$ and $P_t f_2(x) = x\cdot e^{-A^\top t}L$. Considering the Poisson equation

$$-Ax \cdot \nabla\phi(x) + \Delta\phi(x) = f(x) - \pi(f),$$

where $f = f_1 + f_2$ we can write the unique, mean-zero solution as

$$\phi(x) = x \cdot \left[\int_0^\infty e^{-A^\top t} M e^{-At}\, dt\right] x - \mathrm{Tr}\left[Q_\infty \int_0^\infty e^{-A^\top t} M e^{-At}\, dt\right] + L \cdot A^{-1}x,$$

which is finite by the stability of $-A$. From (C.1) it follows that

$$\sigma^2(f) = 4\mathrm{Tr}\left[Q_\infty \Pi Q_\infty M\right] + 2L \cdot A^{-1} Q_\infty L, \tag{C.3}$$

where

$$\Pi = \int_0^\infty e^{-A^\top t} M e^{-At}\, dt.$$

The case when $\Sigma \neq I$, can be recovered immediately by applying the transformation $\phi(x) = \psi(\Sigma^{-1/2}x)$, thus reducing it to the previous case, from which we obtain the same identity. □

Remark C.1. *Note that there is no impediment to deriving the asymptotic variance for observables involving higher powers, e.g. a third order tensor of the form $\sum_{i,j,k} K_{i,j,k} x_i x_j x_k$, but we only provide the result up to second order for the sake of clarity. A more general approach would potentially be possible by considering the decomposition of an observable f with respect to the eigenbasis of the Ornstein Uhlenbeck operator $\mathcal{L}$, which can be shown to be Hermite polynomials Metafune et al. (2002).*

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**(4), 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), 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**(2), 253–268.

Diaconis, P., Holmes, S. & Neal, R. M. (2000), 'Analysis of a nonreversible Markov chain sampler', *The Annals of Applied Probability* **10**(3), 726–752.

Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. (1987), 'Hybrid Monte Carlo', *Physics letters B* **195**(2), 216–222.

Duncan, A. B., Lelièvre, T. & Pavliotis, G. A. (2016), 'Variance reduction using nonreversible Langevin samplers', *Journal of Statistical Physics* **163**(3), 457–491.

Fathi, M., Homman, A.-A. & Stoltz, G. (2015), 'Error analysis of the transport properties of metropolized schemes', *ESAIM: Proceedings and Surveys* **48**, 341–363.

Fathi, M. & Stoltz, G. (2015), 'Improving dynamical properties of metropolized discretizations of overdamped Langevin dynamics', *Numerische Mathematik* pp. 1–58.

Gardiner, C. W. (1985), *Handbook of stochastic methods*, second edn, Springer-Verlag, Berlin.

Girolami, M. & Calderhead, B. (2011), 'Riemann Manifold Langevin and Hamiltonian Monte Carlo methods', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(2), 123–214.

Glynn, P. W. & Meyn, S. P. (1996), 'A Liapounov bound for solutions of the Poisson equation', *The Annals of Probability* **24**(2), 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, Springer-Verlag, Berlin.

Hastings, W. K. (1970), ‘Monte carlo sampling methods using markov chains and their applications’, *Biometrika* **57**(1), 97–109.

Horowitz, A. M. (1991), ‘A generalized guided Monte Carlo algorithm’, *Physics Letters B* **268**(2), 247–252.

Hukushima, K. & Sakai, Y. (2013), An irreversible Markov-chain Monte Carlo method with skew detailed balance conditions, *in* ‘Journal of Physics: Conference Series’, Vol. 473, IOP Publishing, p. 012012.

Kopec, M. (2014), ‘Weak backward error analysis for overdamped Langevin processes’, *IMA Journal of Numerical Analysis* p. 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 University Press, Cambridge.

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**(2), 237–274.

Lelièvre, T. & Stoltz, G. (2016), ‘Partial differential equations and stochastic methods in molecular dynamics’, *Acta Numerica* **25**, 681–880.

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, *in* ‘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**(2), 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**(2), 552–577.

Mengersen, K. L. & Tweedie, R. L. (1996), ‘Rates of convergence of the Hastings and Metropolis algorithms’, *The Annals of Statistics* **24**(1), 101–121.

Metafune, G., Pallara, D. & Priola, E. (2002), ‘Spectrum of Ornstein-Uhlenbeck operators in L_p spaces with respect to invariant measures’, *Journal of Functional Analysis* **196**(1), 40–60.

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**(6), 1087–1092.

Meyn, S. P. & Tweedie, R. L. (1993*a*), 'Markov chains and stochastic stability. communication and control engineering series', *Springer-Verlag London Ltd., London* **1**, 993.

Meyn, S. P. & Tweedie, R. L. (1993*b*), 'Stability of Markovian processes III: Foster-Lyapunov criteria for continuous-time processes', *Advances in Applied Probability* **25**(3), 518–548.

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

Mijatovic, A. & Vogrinc, J. (2015), 'On the poisson equation for metropolis-hastings chains', *arXiv preprint arXiv:1511.07464* .

Mijatović, A. & Vogrinc, J. (2017), 'Asymptotic variance for random walk metropolis chains in high dimensions: logarithmic growth via the poisson equation', *arXiv preprint arXiv:1707.08510* .

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

Møller, J., Syversveen, A. R. & Waagepetersen, R. P. (1998), 'Log Gaussian Cox processes', *Scandinavian journal of statistics* **25**(3), 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, *in* S. Brooks, A. Gelman, G. Jones & X.-L. Meng, eds, 'Handbook of Markov Chain Monte Carlo', CRC press, Boca Raton, 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**(1), 60–106.

Pavliotis, G. A. (2014), *Stochastic processes and applications*, Vol. 60 of *Texts in Applied Mathematics*, Springer, New York. Diffusion processes, the Fokker-Planck and Langevin equations.

Poncet, R. (2017), 'Generalized and hybrid metropolis-hastings overdamped langevin algorithms'.

Rey-Bellet, L. & Spiliopoulos, K. (2015*a*), 'Irreversible Langevin samplers and variance reduction: a large deviations approach', *Nonlinearity* **28**(7), 2081.

Rey-Bellet, L. & Spiliopoulos, K. (2015*b*), '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), 1–11.

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

Roberts, G. O. & Tweedie, R. L. (1996), 'Exponential convergence of Langevin distributions and their discrete approximations', *Bernoulli* **2**(4), 341–363.

Sanz-Serna, J. M. & Calvo, M. P. (1994), *Numerical Hamiltonian problems*, Vol. 7 of *Applied Mathematics and Mathematical Computation*, Chapman & Hall, London.

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)* pp. 3–23.

Stramer, O. & Tweedie, R. L. (1999*a*), 'Langevin-type models I: Diffusions with given stationary distributions and their discretizations', *Methodology and Computing in Applied Probability* **1**(3), 283–306.

Stramer, O. & Tweedie, R. L. (1999*b*), 'Langevin-type models II: self-targeting candidates for mcmc algorithms', *Methodology and Computing in Applied Probability* **1**(3), 307–328.

Talay, D. & Tubaro, L. (1990), 'Expansion of the global error for numerical schemes solving stochastic differential equations', *Stochastic Anal. Appl.* **8**(4), 483–509.

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

Turitsyn, K. S., Chertkov, M. & Vucelja, M. (2011), 'Irreversible Monte Carlo algorithms for efficient sampling', *Physica D: Nonlinear Phenomena* **240**(4), 410–414.

Wu, S.-J., Hwang, C.-R. & Chu, M. T. (2014), 'Attaining the optimal gaussian diffusion acceleration', *Journal of Statistical Physics* **155**(3), 571–590.

Zygalakis, K. C. (2011), 'On the existence and the applications of modified equations for stochastic differential equations', *SIAM J. Sci. Comput.* **33**(1), 102–130.