Monte Carlo Markov Chains Algorithms and Particle Methods: Non-reversible Metropolis-Hastings - Joris Bierkens

Benjamin Cohen

July 6, 2024

Abstract

Metropolis-Hastings (MH) serves as a cornerstone in Markov Chain Monte Carlo (MCMC) methods, yet its reversibility may hinder efficiency in exploring complex state spaces. In this work, we introduce an extension of the MH algorithm, termed Non-Reversible Metropolis-Hastings (NRMH), to generate non-reversible Markov chains. The key innovation lies in the modification of the acceptance probability, leveraging the concept of the vorticity matrix. The resulting Markov chain exhibits non-reversibility, offering potential advantages in sampling efficiency. Drawing from the literature on asymptotic variance, large deviations theory, and mixing time, we highlight the favorable properties of non-reversible Markov chains in these contexts. Specifically, we adapt results from the large deviations theory to elucidate the advantageous properties of NRMH. To demonstrate the applicability of NRMH in continuous settings, we develop the necessary theoretical framework and apply it to Gaussian distributions in three and nine dimensions. Empirical evaluations, including auto-correlation analysis and estimation of asymptotic variance, reveal substantial improvements in NRMH over MH with identical step sizes. This study underscores the promising prospects of non-reversible MCMC methods in enhancing sampling efficiency and exploring high-dimensional spaces.

1 Introduction

The Metropolis-Hastings (MH) algorithm, pioneered by Metropolis et al. (1953) [1] and extended by Hastings (1970) [2], stands as a foundational method in the realm of Markov Chain Monte Carlo (MCMC) techniques. Its applications span various domains of mathematics, notably Bayesian inference and statistical mechanics (Diaconis and Saloff-Coste, 1998 [3]; Diaconis, 2008; Levin et al., 2009 [4]). Central to the efficacy of MH in solving computational problems is the efficiency of the Markov chain it generates. However, the conventional MH algorithm generates reversible chains, adhering to detailed balance principles. While this reversibility ensures convergence to the desired invariant distribution, it may constrain exploration efficiency, particularly in complex or high-dimensional spaces.

Non-reversible Markov chains offer a potential solution to this limitation, exhibiting improved mixing behavior and asymptotic variance (Suwa and Todo, 2010 [5]; Turitsyn et al., 2011 [6]; Vucelja, 2014 [7]). Both experimental and theoretical evidence supports the advantages of non-reversibility in certain cases (Diaconis et al., 2000; Neal, 2004; Sun et al., 2010; Chen and Hwang, 2013; Rey-Bellet and Spiliopoulos, 2014).

There are two primary strategies for constructing non-reversible chains from reversible ones: one involves augmenting the state space, while the other introduces non-reversibility directly (Diaconis et al., 2000; Neal, 2004; Turitsyn et al., 2011; Vucelja, 2014; Sun et al., 2010 [3, 6, 7, 8]). In this paper, we focus on the latter approach, particularly applicable in continuous spaces, where lifting the state space may not be feasible. While altering transition probabilities directly may pose computational challenges, it offers a solution for constructing non-reversible chains without expanding the state space.

To address these challenges, we propose an extension of the MH algorithm termed 'Non-Reversible Metropolis-Hastings' (NRMH). NRMH enables non-reversible transitions by modifying the acceptance ratio, a concept elaborated in Section 2. Initially developed for discrete state spaces for pedagogical clarity, we demonstrate how adjusting the acceptance probability imbues the resulting chain with specified 'vorticity', ensuring non-reversibility. Remarkably, any Markov chain satisfying a symmetric structure condition can be realized through NRMH, underscoring the algorithm's generality.

Theoretical advantages of finite state space non-reversible chains, including improved asymptotic variance and large deviation estimates, are briefly discussed in Section 3, drawing upon previous research (Sun et al., 2010; Rey-Bellet and Spiliopoulos, 2014). While the construction of non-reversible chains in continuous state spaces has been elusive, NRMH offers a promising avenue in this direction, as explored in Section 4. In section 5, we present experimental validations, including a non-reversible adaptation of the Metropolis Adjusted Langevin Algorithm (MALA) for Gaussian multivariate target distributions.

Finally, we conclude with insights into future research directions in Section 6.

Before delving further, we establish essential notation. We denote finite and infinite-dimensional vectors and matrices, employ indicators for sets, and specify norms and spectral properties consistently throughout the paper. Such notation facilitates clarity and coherence in subsequent discussions.

2 Metropolis-Hastings Generalized to Obtain Non-Reversible Chains

In order to extend the Metropolis-Hastings (MH) algorithm to generate non-reversible Markov chains, we introduce the notion of a vorticity matrix in this section. This concept is pivotal for developing a non-reversible version of the classical MH algorithm.

2.1 Non-Reversible Markov Chains and Vorticity

Consider a Markov chain with transition matrix $P = (P(x, y))$ defined over a finite or countable state space S. A distribution on S is represented by a vector with positive elements in $L^1(S)$, and it does not need to be normalized. A distribution π is considered invariant for P if $\pi P = \pi$, and it satisfies the detailed balance condition if $\text{diag}(\pi)P = P^T\text{diag}(\pi)$, implying reversibility. A chain not adhering to detailed balance is termed non-reversible. We introduce the vorticity matrix V, given by

$$V(x, y) = \pi(x)P(x, y) - \pi(y)P(y, x) \tag{1}$$

which is essentially the skew-symmetric part of P. Key observations include:

1. V is skew-symmetric, i.e., $V = -V^T$;
2. π satisfies detailed balance with respect to P if and only if $V = 0$;
3. π is invariant for P if and only if $V\mathbf{1} = 0$.

Thus, a matrix V which is skew-symmetric and satisfies $V1 = 0$ is termed a vorticity matrix. It is crucial in constructing non-reversible chains from reversible ones.

Remark 2.2 introduces a method for creating a non-reversible chain P from a reversible one K and a vorticity matrix V. However, implementing this technique might pose computational challenges, particularly in scenarios with uncountable state spaces. Essentially, P is formed by combining K and V according to the formula $P(x, y) = K(x, y) + \frac{1}{2\pi(x)}V(x, y)$, ensuring that P remains a probability matrix with non-negative entries. Transitioning from state x requires computing the values of $K(x, \cdot)$ and $V(x, \cdot)$ to ascertain the transition probabilities. Despite its computational demands, this method becomes indispensable in situations involving uncountable state spaces.

2.2 Metropolis-Hastings

In the Metropolis-Hastings (MH) algorithm, we construct a Markov chain P_0 with a predefined invariant distribution π. To simplify matters, we'll assume that $\pi(x) > 0$ for all states x in our set S. This construction involves utilizing another Markov chain Q, which satisfies a special symmetry condition: whenever there's no probability of transitioning from state x to state y in Q, there's also no probability of transitioning from y to x:

$$Q(y, x) = 0 \longleftrightarrow Q(x, y) = 0 \tag{2}$$

The Hastings ratio $R_0(x, y)$ is a key concept in this algorithm. It's defined as the ratio of probabilities for transitioning from x to y compared to transitioning from y to x, under the distributions

π and Q. We use this ratio to compute acceptance probabilities $A_0(x, y)$, which determine whether proposed transitions are accepted or rejected:

$$R_0(x,y) = \begin{cases} \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, & \text{for all } x, y \in S \text{ for which } \pi(x)Q(x,y) \neq 0, \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

The transition probabilities $P_0(x, y)$ for the Markov chain P_0 are then calculated based on Q and the acceptance probabilities $A_0(x, y)$:

$$A_0(x,y) := min(1, R_0(x,y)) \tag{4}$$

$$P_0(x,y) = \begin{cases} Q(x,y)A_0(x,y), & x \neq y, \\ Q(x,x) + \sum_{z \neq x} Q(x,z)(1 - A_0(x,z)), & x = y. \end{cases} \tag{5}$$

It's important to note that the chain P_0 is designed such that π remains its invariant distribution.

One noteworthy observation is that the acceptance ratio $R_0(x, y)$ being less than or equal to 1 implies that the reverse acceptance ratio $R_0(y, x)$ is greater than or equal to 1, a pattern that recurs in subsequent analyses.

2.3 Constructing Non-Reversible Chains with Metropolis-Hastings

This subsection presents a framework for extending the Metropolis-Hastings (MH) algorithm to generate non-reversible Markov chains. Given a vorticity matrix V and a transition matrix Q for a Markov chain satisfying the symmetric structure condition, we define the non-reversible Hastings ratio R_V and derive acceptance probabilities A_V akin to MH. To ensure non-negative acceptance probabilities, V is required to satisfy a specific constraint. Transition probabilities P_V are subsequently computed based on Q and A_V. The resulting chain P_V exhibits π as its invariant distribution, with V serving as its vorticity matrix, as proven in **Lemma 2.4** and **Theorem 2.5**.

Now, we extend this framework to construct Markov chains that are generally non-reversible. Let $V \in \mathbb{R}^{n \times n}$ be a vorticity matrix, and Q be the transition matrix of a Markov chain, satisfying the symmetric structure condition (2). Here, $\pi : S \to (0, \infty)$ represents a distribution that may not be normalized and has solely positive entries.

We define the non-reversible Hastings ratio R_V between states x and y as:

$$R_V(x,y) = \begin{cases} \frac{V(x,y)+\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, & \text{if } \pi(x)Q(x,y) \neq 0, \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

Similar to MH, the acceptance probabilities A_V are defined as:

$$A_V(x,y) = \min(1, R_V(x,y)). \tag{7}$$

Since entries of V can be negative, we explicitly constrain the vorticity matrix V to adhere to:

$$V(x,y) \geq -\pi(y)Q(y,x) \text{ for all } x, y \in S. \tag{8}$$

Note that Equation (8) implies, via skew-symmetry of V, that $-\pi(y)Q(y,x) \leq V(x,y) \leq \pi(x)Q(x,y)$ for all $x, y \in S$. In particular, V should have zeroes wherever Q has zeroes, in accordance with the symmetric structure condition (2).

Transition probabilities $P_V(x, y)$, akin to Metropolis-Hastings, are defined as:

$$P_V(x,y) = \begin{cases} Q(x,y)A_V(x,y), & x \neq y, \\ Q(x,x) + \sum_{z \neq x} Q(x,z)(1 - A_V(x,z)), & x = y. \end{cases} \tag{9}$$

To ascertain that the resulting Markov chain possesses π as its invariant density, we need to verify that V, π, and P_V satisfy Equation (1). This is established akin to the Metropolis-Hastings framework:

Lemma 2.3. Suppose V is a vorticity matrix, Q is a transition probability matrix satisfying (2), π is a non-zero distribution, and (8) holds. Then $V(y, x) > 1$ if and only if $V(x, y) < 1$ for any $x, y \in S$ where $Q(x, y) \neq 0$.

Using the aforementioned lemma, it is straightforward to demonstrate that V is indeed the vorticity matrix of (P_V, π).

Lemma 2.4. Let Q be a Markov chain, V be a vorticity matrix, and π be a distribution on S, satisfying (2) and (8). Define P_V through (6), (7), and (9). Then Equation (1) holds for (P_V, π), i.e.,

$$V(x, y) = \pi(x)P_V(x, y) - \pi(y)P_V(y, x), \text{ for all } x, y \in S. \tag{10}$$

Theorem 2.5. Given a Markov chain Q, a vorticity matrix V, and a distribution π on S that is everywhere positive, satisfying (2) and (8), and P_V defined through (6), (7), and (9), then P_V has π as its invariant distribution and V as its vorticity matrix.

Remark 2.6. We term a combination (Q, V, π) that satisfies (2) and (8) as compatible. While checking (8) requires conditions on π, we can relax these constraints to obtain only a lower bound for π. In the proof of Theorem 4.2, the analogous condition for continuous spaces is examined as an example.

Remark 2.7. When we have a compatible combination of proposal chain Q, vorticity matrix V, and target distribution π, NRMH exhibits similar advantageous properties as MH, where only local information is necessary: Q, π, and V need to be evaluated solely at the current and proposed states, and no normalization of π is mandated. In Section 4, this becomes particularly crucial when NRMH is applied to problems in continuous state space.

2.4 General Observations on NRMH

This section concludes with general observations on NRMH. Notably, **Proposition 2.8** highlights the versatility of the approach by demonstrating its ability to reconstruct any Markov chain. **Remark 2.9** emphasizes conditions necessary for irreducibility, and **Proposition 2.10** establishes a connection between different approaches to constructing non-reversible chains.

3 Advantages of Non-Reversible Markov Chains (NMRC) in Finite State Spaces

NMRC leads to computational advantages over reversible chains. We will evoke some of these advantages as they pertain to finite state space Markov chains. It remains open to prove Sections 3.1 and 3.2 in a generic way to uncountable and continuous state spaces.

3.1 Asymptotic Variance

We take a Markov chain P on state space S with invariant probability distribution μ ande define a function $f : S \to \mathbb{R}$.

f satisfies a Central Limit Theorem (CLT) if there exists a $\sigma_f^2 < \infty$ called asymptotic variance such that the normalized sum $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [f(X_i) - \mu(f)]$ converges weakly to a $N(0, \sigma_f^2)$ distribution.

In a finite state space S and an irreducible P, a CLT is satisfied for any $f : S \to \mathbb{R}$ [see, e.g., Roberts and Rosenthal (2004)].

[9] states that for an irreducible reversible Markov chain with invariant probability distribution μ and a non-zero vorticity matrix V, if $P = K + \frac{1}{2}\text{diag}(\mu)^{-1}V$ is the transition matrix of an irreducible Markov chain, then for any $f : S \to \mathbb{R}$, the asymptotic variance of f with respect to P, denoted by $\sigma_{f,P}^2$, is less than or equal to the asymptotic variance of f to K, denoted by $\sigma_{f,K}^2$. Moreover, there exists an f such that $\sigma_{f,P}^2 < \sigma_{f,K}^2$. In the end, adding non reversibility reduces asymptotic variance.

3.2 Large Deviations

[10] noted that non-reversible diffusions on compact manifolds exhibit favorable properties in terms of large deviations of the occupation measure from the invariant distribution.

Along this thought, for finite state spaces: Assuming S is finite, we can transform a discrete-time Markov chain on S into a continuous-time chain by making transitions after random waiting times

with independent Exp(λ) distributions. The discrete chain with transition matrix P then transforms into a continuous-time chain with generator $G(x,y) = \begin{cases} \lambda P(x,y) & \text{if } x \neq y \\ -\lambda \sum_{z \neq x} P(x,z) & \text{if } x = y \end{cases}$.

The occupation measure of the resulting Markov process is defined as $L_t = \frac{1}{t}\int_0^t \delta_{X_s} ds$. If G is irreducible, the occupation measure satisfies a large deviation principle with rate function $I_G(\mu) = \sup_{u>0}\left\{-\sum_{x\in S}\frac{\mu(x)}{u(x)}(Gu)(x)\right\}$, where $\mu \in P(S)$, and $P(S)$ is the set of probability distributions on S. This rate function quantifies the probability of a large deviation of the occupation measure from the invariant distribution for large t.

Proposition 3.2 states that deviations for non-reversible continuous-time chains from the invariant distribution are asymptotically less likely than for the corresponding reversible chain. If G admits a decomposition of the form $G(x,y) = K(x,y) + \frac{1}{2}\pi(x)V(x,y)$, where $\pi(x)K(x,y) = \pi(y)K(y,x)$ and $V(x,y) = -V(y,x)$ for all $x \neq y$, then $I_G(\mu) \geq I_K(\mu)$ for all $\mu \in P(S)$, and the inequality is strict if $\|V\| \neq 0$.

3.3 Mixing Time and Spectral Gap

While non-reversibility in a Markov chain can have favorable effects on mixing time and spectral gap, there are no general results in this direction. For theoretical results, readers are referred to [3, 4, 8, 6, 7], where non-reversibility appears to improve sampling, especially in the case of sampling from a multimodal distribution.

4 Non-Reversible Metropolis-Hastings in Euclidean Space

This portion delves into broadening the scope of the non-reversible Metropolis-Hastings algorithm to encompass an Euclidean state space. The discourse provided here focuses on a particular instance within the broader context of measurable spaces.

4.1 General Setting

Consider a scenario where we have a Markov transition kernel Q operating on $\mathbb{R}^n$, characterized by a density function $q(x,y)$ with respecto to the Lebesgue measure. This is denoted as $Q(x, \mathrm{d}y) = q(x,y)\mathrm{d}y$. Our goal is to sample from a distribution on $\mathbb{R}^n$ with a Lebesgue density π, without the prerequisite that $\int_{\mathbb{R}^n}\pi(x)\mathrm{d}x = 1$. Let $V : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be Lebesgue measurable, satisfying the following conditions:

1. $V(x,y) = -V(y,x)$ for all $x, y \in \mathbb{R}^n$,
2. $\int_{A\times\mathbb{R}^n} V(x,y)\mathrm{d}x\mathrm{d}y = 0$ for all $A \in \mathcal{B}(\mathbb{R}^n)$, where $\mathcal{B}(\mathbb{R}^n)$ denotes the Borel σ-algebra generated by open sets in $\mathbb{R}^n$.

Additionally, suppose:

1. $V(x,y) = 0$ whenever $\pi(x)q(x,y) = 0$,
2. $\pi(x)q(x,y) = 0$ if and only if $\pi(y)q(y,x) = 0$ for all $x, y \in \mathbb{R}^n$,
3. $V(x,y) + \pi(y)q(y,x) \geq 0$ whenever $\pi(x)q(x,y) = 0$.

Now, let's define the Hastings ratio $R(x,y)$, acceptance probabilities $A(x,y)$, and transition kernel $P(x,B)$ as follows:

$$
\begin{aligned}
R(x,y) &:= \frac{V(x,y) + \pi(y)q(y,x)}{\pi(x)q(x,y)}, \\
A(x,y) &:= \min(1, R(x,y)), \\
P(x,B) &:= \int_B A(x,y)q(x,y)\mathrm{d}y + \mathbb{I}_{x\in B}\left(1 - \int_{\mathbb{R}^n} A(x,y)q(x,y)\mathrm{d}y\right).
\end{aligned}
$$

Theorem 4.1: Under the aforementioned conditions, P constitutes a Markov transition kernel with an invariant density π.

Similar to the discrete state space setup, V is referred to as the vorticity density of (P, π). If V equals zero on a set of positive Lebesgue measure, then P is non-reversible, indicating the existence of sets $B_1, B_2 \subset \mathbb{R}^n$ for which:

$$\int_{B_1} \int_{B_2} \pi(x) P(x, \mathrm{d}y) \mathrm{d}x \neq \int_{B_2} \int_{B_1} \pi(x) P(x, \mathrm{d}y) \mathrm{d}x.$$

4.2 Langevin Diffusions for Sampling in Euclidean Space

Exploring non-reversible sampling methods in Euclidean space is an area that has received limited attention. In this overview, we explore using Langevin diffusions and non-reversible Metropolis-Hastings methods for generating samples from desired density functions.

Assume we have a smooth target density function π. One common method for sampling from π is using the Langevin diffusion process, which maintains π as its invariant density.

$$\mathrm{d}X(t) = \nabla(\log \pi)(X(t))\mathrm{d}t + \sqrt{2}\mathrm{d}W(t), \quad t \geq 0, \tag{11}$$

where W represents a standard Brownian motion in $\mathbb{R}^n$. To discretize Langevin diffusion for sampling, we use the Euler-Maruyama method, creating a discrete-time Markov chain. However, choosing an appropriate step size (h) is crucial, as small h values lead to slow convergence, while large h values introduce significant discrepancies between continuous and discrete versions. To address this, we often use Euler-Maruyama discretization as a proposal for the Metropolis-Hastings algorithm, leading to methods like the Metropolis Adjusted Langevin Algorithm (MALA).

$$X_{k+1} \sim \mathcal{N}(X_k + h\nabla(\log \pi)(X_n), 2h), \quad k = 0, 1, 2, \ldots, \tag{12}$$

Non-reversible diffusions, characterized by skew-symmetric matrices (S), also maintain π as their invariant density. These diffusions offer potential benefits over reversible Langevin diffusions, especially for multivariate Gaussian target distributions. They tend to reduce deviations of the empirical distribution from the invariant distribution.

$$\mathrm{d}X(t) = -(I + S)\nabla(\log \pi)(X(t))\mathrm{d}t + \sqrt{2}\mathrm{d}W(t), \quad t \geq 0, \tag{13}$$

When discretizing non-reversible diffusions, we must address discretization errors through a Metropolis-Hastings accept/reject step. However, traditional Metropolis-Hastings produces reversible chains, compromising the advantages of non-reversibility. Implementing non-reversible Metropolis-Hastings aims to preserve these advantages, which we'll exemplify in the context of multivariate Gaussian distributions.

4.3 Non-Reversible Metropolis-Hastings for Sampling Multivariate Gaussian Distributions

Consider a centered normal distribution with a positive definite covariance matrix V as the target distribution. In this scenario, the Langevin diffusion transforms into the Ornstein-Uhlenbeck process.

$$\mathrm{d}X(t) = -V^{-1}X(t)\mathrm{d}t + \sqrt{2}\mathrm{d}W(t), \quad t \geq 0, \tag{14}$$

where $(W(t))$ is an n-dimensional standard Brownian motion. Hwang et al. (1993) demonstrated that incorporating a 'non-reversible' component in the drift, represented as $-SV^{-1}$, where S is skew-symmetric, can enhance the convergence of the sample covariance. Consequently, we will examine the Ornstein-Uhlenbeck process with an adjusted drift.

$$\mathrm{d}X(t) = BX(t)\mathrm{d}t + \sqrt{2}\mathrm{d}W(t), \tag{15}$$

where $B := -(I+S)V^{-1}$ with S skew-symmetric. For any choice of skew-symmetric S, this diffusion keeps π invariant. The convergence to equilibrium of the diffusion is governed by the spectral bound, $s(B) := \max\{\mathrm{Re}\lambda : \lambda \in \sigma(B)\}$. More specifically,

$$\mathrm{Cov}(X(t)) = 2\int_0^t e^{Bs}e^{Bs^\top}\mathrm{d}s \to 2\int_0^\infty e^{Bs}e^{Bs^\top}\mathrm{d}s = V, \quad (\text{as } t \to \infty), \tag{16}$$

with rate of convergence

$$\ln\|\frac{\int_t^\infty e^{Bs}e^{Bs^\top}\mathrm{d}s}{t}\| \to 2s(B). \tag{17}$$

Furthermore, for any choice of S, it holds that $s(B) \leq s(-V^{-1})$. This implies that incorporating a non-reversible term enhances the convergence speed towards equilibrium. Lelièvre et al. (2013) demonstrated the possibility of selecting S optimally, resulting in $s(B) = -\mathrm{tr}(V^{-1})/n$. This optimal choice of S ensures that the convergence of the chain is primarily determined by the average of the eigenvalues, contrasting with the reversible scenario where the 'worst' eigenvalue governs the convergence speed.

We will now apply the theory outlined in Section 4.1 to discretize the non-reversible Ornstein-Uhlenbeck process in time. To fulfill condition (4.1.3) later, we need flexibility in both the drift multiplier B and the diffusivity. We consider the time discretization of equation (15) with a positive step size h:

$$X_{k+1} = X_k + hBX_k + \sqrt{2h\sigma}Z_{k+1}, \tag{18}$$

Moreover, (Z_k) are i.i.d. standard normal, and $\sigma > 0$. When $\sigma = 1$, this corresponds to the standard Euler-Maruyama discretization. We define the transition kernel of the Euler-Maruyama discretization as our proposal distribution $Q(x, \mathrm{d}y) = q(x,y)\mathrm{d}y$. Our first step is to determine the vorticity density V of Q.

Let $r(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$ denote the spectral radius of a square matrix A. Assuming $r(I + hB) < 1$, the invariant probability distribution of Q is a centered normal distribution with covariance R, where R is the unique positive definite matrix solution to the discrete-time Lyapunov equation. Denote by ρ the density of the invariant probability distribution of Q. We define $f(x,y) = \rho(x)q(x,y)$. The vorticity density of the proposal chain is

$$V(x,y) := f(x,y) - f(y,x). \tag{19}$$

As the target density, we have

$$\pi(x) = \frac{((2\pi)^n \det V)^{-1/2}}{\exp\left(\frac{1}{2}x^\top V^{-1}x\right)}, \quad x \in \mathbb{R}^n. \tag{20}$$

V satisfies 4.1.1 and 4.1.2 since π and q are non-degenerate. The same statements hold trivially for scalar multiples of V.

Verification of 4.1.3 requires more effort. We provide a sufficient condition:

Theorem 4.2 *Define constants* $0 < C_1 \leq C_2$ *by*

$$C_1 = \left\|V^{-1/2}(I+S)V^{-1}(I-S)V^{1/2}\right\| \quad \text{and} \tag{21}$$

$$C_2 = \left\|V^{-1/2}(I+S)V^{-1/2}\right\|^2 \|V\|. \tag{22}$$

Suppose $c > 0$, $h > 0$, and $\sigma > 0$ satisfy

$$h < \frac{2}{C_2}, \quad \sigma^2 \leq \frac{2 - hC_2}{2 - h(C_2 - C_1)}, \quad \text{and} \quad c \leq \sigma^n. \tag{23}$$

Then $V(x, y) := cV(x, y)$, with V as constructed above, satisfies 4.1.3 , and is therefore a vorticity density compatible with the proposal distribution $Q(x, \mathrm{d}y) \sim \mathcal{N}((I + hB)x, 2h\sigma^2 I)$ and invariant distribution $\mathcal{N}(0, V)$.

Remark 4.3 *How ought to one select c, h, and σ? It appears sensible to select σ^2 rise to the maximal permitted esteem in so that the deviation from the Euler-Maruyama discretization (which has $\sigma = 1$) is negligible; i.e., let*

$$\sigma = \sigma(h) = \sqrt{\frac{2 - hC_2}{2 - h(C_2 - C_1)}}. \tag{24}$$

To maximize the non-reversibility effects in the acceptance probability, one should choose c as large as possible, i.e., $c = \sigma n$. A heuristic estimate for the scaling of the expected step size is given by the step size h times the multiplicative factor in the vorticity, $c = \sigma n$ in the acceptance probability. Note that $h\sigma n = 0$ for $h = 0$ and $h = 2/C_2$. Maximization of $h\sigma n$ with respect to h yields

$$h = \frac{2}{C_2} + \frac{(n + 2)C_1}{2C_2(C_2 - C_1)} - \frac{\sqrt{(n - 2)^2 C_1^2 + 8nC_1C_2}}{2C_2(C_2 - C_1)}. \tag{25}$$

as long as $C_1 < C_2$ (which is the case in which S and V do not commute). This expression satisfies the condition $h < 2/C_2$. A first-order Taylor approximation around $1/n$ yields the simplified expression

$$h = \frac{4}{(n + 2)C_2} < \frac{2}{C_2}. \tag{26}$$

The corresponding value of $\sigma^2(h)$ is, to first order, equal to $\sigma^2(h) \approx 1 - 2C_1/(n + 2)C_2$. In case $C_1 = C_2$, the optimal value of h is given by $h = \frac{4}{(n+2)C_2}$, with $\sigma^2(h) = 1 - \frac{2}{n+2}$.

5 Numerical experiments

Here, we conduct two experiments illustrating the approach above. For the Markov chain realization $(X_1, ..., X_P)$ obtained, we estimate the decorrelation by considering the empirical auto-correlation function (EACF) $r_i(k)$ defined by

$$r_i(k) = \frac{1}{P - k} \sum_{p=1}^{P-k} \left(\frac{(X_{i,p} - \mu_i)(X_{i,p+k} - \mu_i)}{\sigma_i^2} \right), \tag{27}$$

where i ranges over the coordinates $i = 1, ..., n$, and where $\mu_i = \frac{1}{P} \sum_{p=1}^{P} X_{i,p}$ represents the empirical average of the i-th coordinate. A fast-decaying EACF indicates that the samples generated by the Markov chain are quickly decorrelating.

5.1 Three-dimensional example

In this example, from Hwang et al. (1993), we take as target covariance structure V a diagonal matrix with $\mathrm{diag}(V) = (1, 1, 1/4)$. The optimal nonlinear drift is obtained by letting

$$S = \begin{pmatrix} 0 & \sqrt{3} & 1 \\ -\sqrt{3} & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix}.$$

We choose the parameter values following Remark 4.3, resulting in $c = 0.5333$, $h = 0.0334$, $\sigma = 0.8109$. The performance of NRMH is compared to MH with identical step-size h, and reversible proposal distribution $Q(x, dy) \sim N((I - hV^{-1})x, 2hI)$. In Fig. 1,2,3 the EACFs for this 3-dimensional example are plotted. Here we see that NRMH helps to decrease the autocorrelations of the slowly decorrelating components in MH (here, the first two components). It achieves this without increasing autocorrelations of components that are already quickly decorrelating (here, the third component).

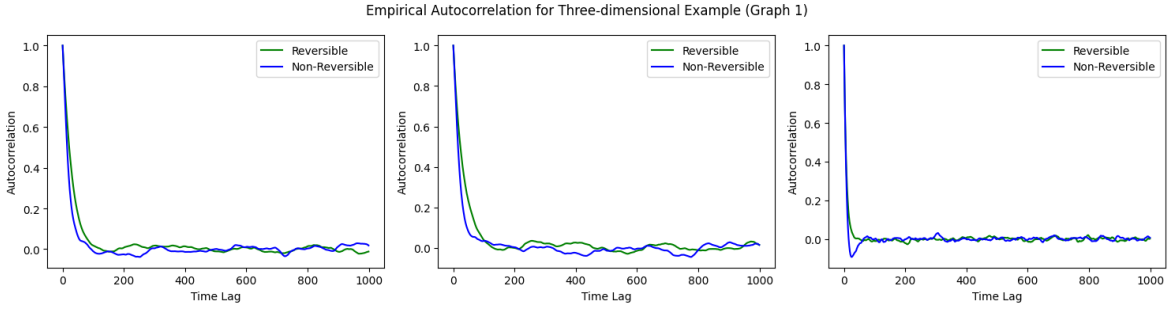

Figure 1: Empirical Auto-correlation for the Three-dimensional Example to time-lag of 1 000

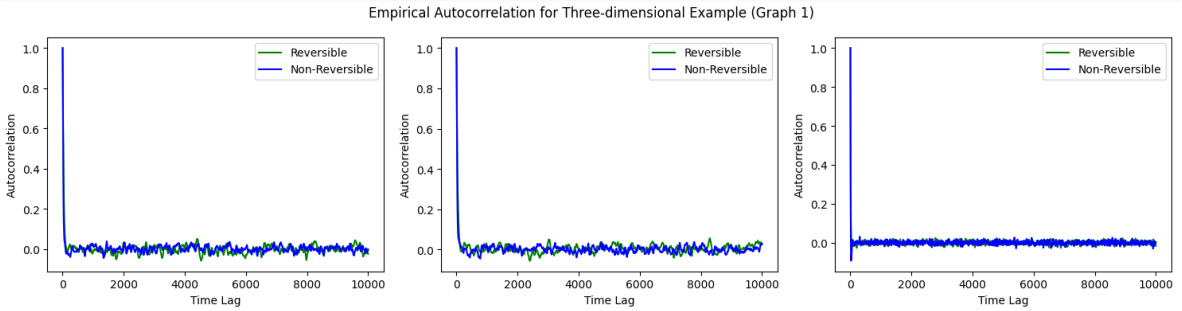

Figure 2: Empirical Auto-correlation for the Three-dimensional Example to time-lag of 10 000

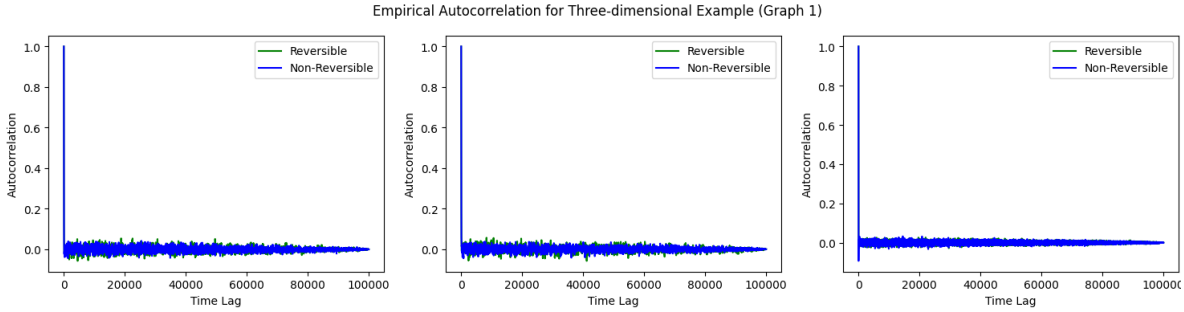

Figure 3: Empirical Auto-correlation for the Three-dimensional Example to time-lag of 100 000

5.2 Nine-dimensional example

In this instance, we constructed a diagonal covariance matrix V with random diagonal elements. Employing the method outlined by [14], we computed an optimal non-reversible drift $B = -(I + S)V^{-1}$. A comparison of the spectral bounds between reversible and non-reversible dynamics showcased significant enhancements with non-reversible dynamics. Introducing non-reversibility led to a harmonizing effect on the autocorrelation of different components, thereby improving the convergence of slower-decorrelating coordinates. For this specific case, selecting suitable values for c, h, and σ as per Remark

Component	1	2	3	4	5	6	7	8	9
NRMH	599.96	661.17	40.80	572.26	159.05	27.35	230.41	401.98	718.64
MH	1315.3	1522.2	47.156	1473.3	876.46	28.316	204.05	708.83	1578.2

Table 1: Estimated asymptotic variances for the 9-dimensional example of Sect. 4.4.2, based on an MCMC trajectory of 10^7 steps

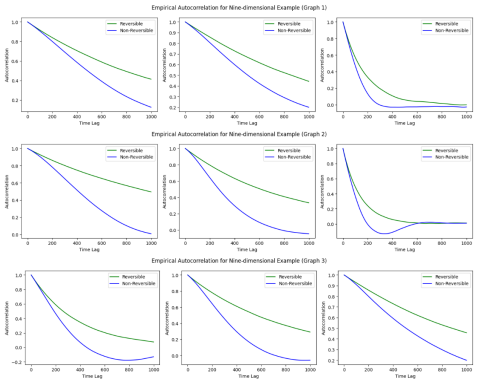

Figure 4: Empirical Auto-correlation for the Nine-dimensional Example to time-lag of 1 000

4.3 yielded noteworthy improvements. Assessing the acceptance ratios based on 10^7 proposed transitions and estimating the asymptotic variance using the batch means method provided insights into the efficacy of the approach. However, it's important to acknowledge that the concept of asymptotic variance hinges on the Central Limit Theorem (CLT), which may not be strictly applicable in this context, as discussed by Roberts and Rosenthal (2004).

The random diagonal matrix V was generated with elements:

$$\text{diag}(V) = (0.8147, 0.9058, 0.1270, 0.9134, 0.6324, 0.0975, 0.2785, 0.5469, 0.9575).$$

Using the algorithm described in Lelièvre et al. (2013), an optimal non-reversible drift $B = -(I + S)V^{-1}$ can be computed. For reversible dynamics, we have $s(-V^{-1}) = -1.0444$, while for the optimal non-reversible dynamics, $s(B) = -\text{tr}V^{-1}/n = -3.2891$. The following equation illustrates this comparison:

$$s(-V^{-1}) = -1.0444,$$
$$s(B) = -\frac{1}{n}\text{tr}(V^{-1}) = -3.2891.$$

One can see the typical effect of adding non-reversibility in the EACFs for this 9-dimensional example plotted in Fig. 4,5,6. In this case, choosing c, h, and σ as in Remark 4.3 results in values $c = 0.4313$, $h = 7.0822 \times 10^{-4}$, and $\sigma = 0.9108$. Evaluating the acceptance ratios based on 10^7 proposed transitions and estimating the asymptotic variance using the batch means method offered insights into the efficacy of the approach. However, it's important to acknowledge that the concept of asymptotic variance hinges on the Central Limit Theorem (CLT), which may not be strictly applicable in this context, as discussed in [11].

Method	Acceptance ratio
NRMH	0.7383
MH	0.9343

Table 2: Acceptance ratios for 9-dimensional example of Sect. 4.4.2, for a sample path consisting of 10^7 proposals, and with a step size $h = 7.0822 \times 10^{-4}$

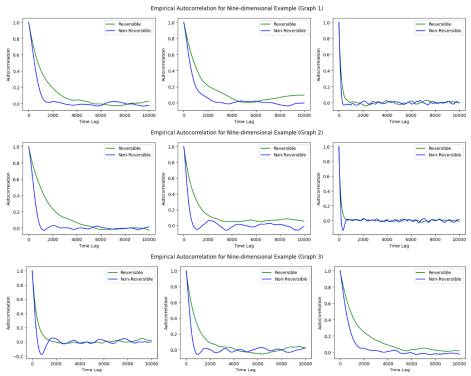

Figure 5: Empirical Auto-correlation for the Nine-dimensional Example to time-lag of 10 000

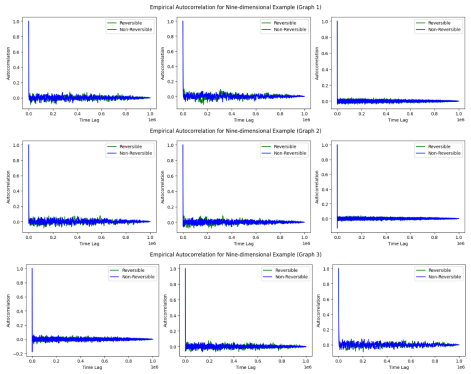

Figure 6: Empirical Auto-correlation for the Nine-dimensional Example to time-lag of 1 000 000

6 Discussion

The effectiveness of non-reversible Markov chains in MCMC can yield substantial improvements in terms of both asymptotic variance reduction and enhanced mixing properties, as noted in the referenced paper. NRMH introduces a valuable addition to the MCMC toolkit, harnessing these advantages. Specifically, before this study, there was a gap in understanding how to construct non-reversible Markov chains for MCMC sampling in continuous state spaces, especially when considering the need for correction steps in time discretization of diffusions.

Leveraging the theoretical framework presented in Section 4, NRMH can be applied to general distributions on $\mathbb{R}^n$ as follows: Suppose a target density function π admits a Gaussian distribution $\mathcal{N}(0, V)$ with density function π_0 on $\mathbb{R}^n$, satisfying $\kappa\pi_0 \leq \pi$ on $\mathbb{R}^n$ for some $\kappa > 0$. If V serves as a suitable vorticity density for sampling from $\mathcal{N}(0, V)$, with proposal density $q(x, y)$, then for $\tilde{V} := \kappa V$, it follows that

$$\tilde{V}(x, y) + \pi(x)q(x, y) = \kappa V(x, y) + \pi(x)q(x, y) \geq \kappa \left(V(x, y) + \pi_0(x)q(x, y)\right) \geq 0,$$

thus satisfying condition 4.1.3 for this choice of π, $\tilde{V}$, and q, allowing the application of Theorem 4.1. The determination of such a suitable V is described in Section 4.3.

The approach outlined in Section 4 represents the initial exploration of the NRMH framework in continuous spaces. As discussed earlier, ensuring the non-negativity of quantities is crucial for utilizing the framework, which can pose challenges. Efforts to relax the stringent conditions of Theorem 4.2 are underway, particularly addressing the issue of excessively small step sizes h that impede NRMH's progress. It remains an open question whether NRMH in continuous spaces can achieve the same level of performance as optimally tuned Metropolis Adjusted Langevin Algorithm (MALA).

The theoretical discourse in Section 3 and the empirical investigation in Section 5 vividly illustrate the potential efficiency gains afforded by employing non-reversible Metropolis-Hastings. Building upon these promising findings, future endeavors aim to extend these results to more diverse settings. The

practical implementation of NRMH hinges on identifying vorticity structures compatible with proposal chains, presenting both a promising and challenging avenue for research.

Analyzing non-reversible Markov chains poses significant challenges, primarily due to the loss of self-adjointness. Without self-adjointness, connecting spectral theory to mixing properties becomes notably more intricate. A promising approach to elucidate the benefits of non-reversible sampling lies in studying Cesàro averages, as discussed in [4] and the results on large deviations in Section 3.2. Although the qualitative results in Section 3 demonstrate superior asymptotic variance or large deviation properties of non-reversible chains, quantifying the extent of improvement remains an open question. Additionally, investigating whether these results can be extended to countable and uncountable state spaces, and under what conditions the resulting chains are geometrically ergodic and/or satisfy a Central Limit Theorem, warrants further exploration.

References

[1] 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. https://doi.org/10.1063/1.1699114

[2] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**(1), 97-109. https://doi.org/10.1093/biomet/57.1.97

[3] Diaconis, P., & Saloff-Coste, L. (1998). What do we know about the Metropolis algorithm? *Journal of Computational and Graphical Statistics*, **7**(1), 1-34. https://doi.org/10.1080/10618600.1998.10474787

[4] Levin, D. A., Peres, Y., & Wilmer, E. L. (2009). *Markov Chains and Mixing Times*. American Mathematical Society.

[5] Suwa, H., & Todo, S. (2010). Markov chain Monte Carlo method without detailed balance. *Physical Review Letters*, **105**(12), 120603. https://doi.org/10.1103/PhysRevLett.105.120603

[6] Turitsyn, K., Chertkov, M., & Vucelja, M. (2011). Irreversible Monte Carlo algorithms for efficient sampling. *Physical Review Letters*, **107**(7), 070601. https://doi.org/10.1103/PhysRevLett.107.070601

[7] Vucelja, M. (2014). Non-reversible Markov chains in Monte Carlo sampling, random combinatorial structures and scalable algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, **2014**(6), P06008. https://doi.org/10.1088/1742-5468/2014/06/P06008

[8] Sun, X., Palacios, J. L., & Schütte, C. (2010). A non-reversible Markov chain sampler with a prescribed equilibrium distribution. *Journal of Chemical Physics*, **133**(13), 134101. https://doi.org/10.1063/1.3490199

[9] Chen, J., & Hwang, C. R. (2013). Eigenvalues and eigenvectors of a random matrix: A non-Hermitian extension of the circular law. *The Annals of Probability*, **41**(6), 4011-4071. https://doi.org/10.1214/12-AOP762

[10] Rey-Bellet, L., & Spiliopoulos, K. (2014). Irreversible Langevin samplers: Variable step size and convergence rates. *Journal of Applied Probability*, **51**(1), 101-118. https://doi.org/10.1239/jap/1395771427

[11] Roberts, G. O., & Rosenthal, J. S. (1996). Quantitative bounds for convergence rates of continuous-time Markov processes. *Electronic Communications in Probability*, **1**(4), 1-12. https://doi.org/10.1214/ECP.v1-978

[12] Roberts, G. O., & Rosenthal, J. S. (1998). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, **13**(4), 335-346. https://doi.org/10.1214/ss/1028905931

[13] Hwang, C. R., Sheu, S. J., & Sun, X. (1993). Optimal properties of nonreversible Markov processes with spectral gap. *Journal of Functional Analysis*, **113**(2), 451-472. https://doi.org/10.1006/jfan.1993.1059

[14] Lelièvre, T., Rousset, M., & Stoltz, G. (2013). *Free Energy Computations: A Mathematical Perspective*. Imperial College Press.