Non-reversible Metropolis-Hastings

Joris Bierkens*

February 25, 2014

Abstract

The Metropolis-Hastings algorithm is extended to ‘non-reversible Metropolis Hastings’ so that it can generate non-reversible Markov chains, using the notion of a vorticity matrix. It is universal, in the sense that any finite Markov chain (reversible or non-reversible) satisfying a very mild condition on its directed graph structure, can be generated by the new approach. The method is based on the novel notion of the ‘vorticity matrix’ of a Markov chain. It can be shown that any chain obtained from Metropolis-Hastings, given a suitable vorticity matrix, can be altered quite elegantly into a non-reversible chain that performs at least as well as the original chain in terms of asymptotic variance. This can be done without loss of the practical properties that make Metropolis-Hastings successful. A numerical experiment on the $n$-cycle illustrates the promise of this approach. Mixing times on the cycle typically reduce significantly as compared to classical Metropolis-Hastings, and the relative effectiveness increases as a function of the problem size. The non-reversible nature of the resulting chains helps especially to overcome multi-modality in the target distribution.

AMS Subject Classification: 65C40, 60J20

Keywords and keyphrases: Markov Chain Monte Carlo, MCMC, Metropolis-Hastings, detailed balance, reversible Markov processes, non-reversible Markov processes, mixing time, spectral gap, asymptotic variance, irreversible Markov chains

1 Introduction

The Metropolis-Hastings algorithm [MRR 53 Has70] is a Markov chain Monte-Carlo method of profound importance to many fields of mathematics such as Bayesian inference and statistical mechanics [DSC98, Dia08, LPW09]. The applicability of Metropolis-Hastings to a particular computational problem depends on the mixing properties of the Markov chain that is generated by the algorithm. The chains generated by the classical Metropolis-Hastings algorithm are reversible, or, in other words, satisfy detailed balance; in fact, this reversibility is essential in showing that the resulting chains have the right invariant probability distribution.

However, non-reversible Markov chains have better mixing properties. This can be shown experimentally in special cases [ST10, TCV11], theoretically in special cases [DHN00, Nea04], but in fact, also in general [SGS10]. In the latter paper it is shown that the asymptotic variance of non-reversible chains is in general better than the asymptotic variance of the ‘corresponding’ reversible chain. There exist two basic approaches to constructing non-reversible chains from reversible chains: one can ‘lift’ the Markov chain to a larger state space [DHN00, Nea04, TCV11], or one can introduce non-reversibility without altering the state space [SGS10]. In this document attention is confined to a novel method in the second category, for which non-reversible Metropolis-Hastings would be a suitable name. Other noteworthy publications on non-reversible Markov chains are [Wil99, GM00]. In this paper we propose a general method of the second type, i.e. there will be no need to augment the state space. We restrict our attention to finite Markov chains; non-reversible

* Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour. The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 270327 (CompLACS)
Markov chains on general spaces will not be discussed in this paper, but it may be noted that the main result of this paper, discussed in Section 2, generalizes in a straightforward way to Markov chains on general state spaces.

In this paper Metropolis-Hastings is extended to ‘non-reversible Metropolis-Hastings’. The main idea of this paper is discussed in Section 2. It is shown how Metropolis-Hastings, and in particular the acceptance probability of Metropolis-Hastings, can be adjusted so that the resulting chain in non-reversible Metropolis-Hastings has a specified ‘vorticity’, and therefore, will be non-reversible. Any Markov chain satisfying a symmetric structure condition can be constructed by non-reversible Metropolis-Hastings, which establishes the universality of the algorithm. In Section 3 it is shown that non-reversible Metropolis-Hastings can be seen as ‘adding’ non-reversibility to a reversible Markov chain, and therefore, by application of a result of [SGS10], has improved performance with respect to asymptotic variance, compared to classical Metropolis-Hastings. For the convenience of the reader two practical approaches of applying non-reversible Metropolis-Hastings are discussed in Section 4, along with some results that are aimed at understanding what a vorticity matrix is and how it can perhaps be constructed. For the particular example of a multi-modal distribution on the n-cycle, illustrate the application and efficiency of non-reversible Metropolis-Hastings in Section 5. Finally, conclusions and directions of further research are discussed in Section 6.

2 Metropolis-Hastings generalized to non-reversible chains

As a preliminary to non-reversible Metropolis-Hastings, we require the notion of vorticity matrix, which is introduced in Section 2.1. The classical Metropolis-Hastings algorithm, discussed in Section 2.2, is extended using the notion of vorticity matrix to a non-reversible version in Section 2.3.

2.1 Vorticity matrices

Let \( P = P(x, y) \) denote a matrix of transition probabilities of a Markov chain on a finite state space \( S = \{1, \ldots, n\} \). A distribution on \( S \) is a vector with positive elements in \( \mathbb{R}^n \), and is not necessarily normalized, i.e. it is not necessarily the case that \( \sum_{x \in S} \pi(x) = 1 \). If \( \pi \) is a distribution such that \( \sum_{x \in S} \pi(x) = 1 \), then we call \( \pi \) a probability distribution. We will always assume that \( \pi(x) > 0 \) for all \( x \in S \). A (probability) distribution \( \pi \) on \( S \) is said to be an invariant (probability) distribution of \( P \) if \( \pi^T P = \pi^T \), i.e. \( \sum_{x \in S} \pi(x) P(x, y) = \pi(y) \) for all \( y \in S \). A distribution \( \pi \) on \( S \) is said to satisfy the detailed balance condition with respect to \( P \) if \( \text{diag}(\pi) P = P^T \text{diag}(\pi) \), i.e. \( \pi(x) P(x, y) = P(y, x) \pi(y) \) for all \( x, y \in S \). If there exists a distribution \( \pi \) on \( S \) that satisfies detailed balance with respect to \( P \), then \( P \) is said to be reversible. As is well known, and straightforward to check, if \( \pi \) satisfies detailed balance with respect to \( P \), then \( \pi \) is invariant for \( P \).

To investigate possible generalizations of the detailed balance condition, let us look at the matrix \( \Gamma \) defined, for some distribution \( \pi \), by

\[
\Gamma(x, y) = \pi(x) P(x, y) - \pi(y) P(y, x), \quad x, y \in S,
\]

or in matrix notation

\[
\Gamma = \text{diag}(\pi) P - P^T \text{diag}(\pi).
\]

The following simple results are worth observing.

**Lemma 2.1.** Let \( P \) be the transition matrix of a Markov chain on \( S \) and let \( \pi \) be a distribution on \( S \). Let \( \Gamma \) be defined by (1).

(i) \( \Gamma \) is skew-symmetric, i.e. \( \Gamma = -\Gamma^T \);

(ii) \( \pi \) satisfies detailed balance with respect to \( P \) if and only \( \Gamma = 0 \);

(iii) \( \pi \) is invariant for \( P \) if and only if \( \Gamma \mathbb{1} = 0 \).
Proof. (i) and (ii) are immediate. As for (iii),

\[(\Gamma \mathbb{1})(x) = \sum_{y \in S} [\pi(x)P(x, y) - \pi(y)P(y, x)] = \pi(x) - \sum_{y \in S} \pi(y)P(y, x),\]

which is zero for all \(x\) if and only if \(\pi\) is invariant for \(P\).

In light of Lemma 2.1, a matrix \(\Gamma \in \mathbb{R}^{n \times n}\) which is skew-symmetric and satisfies \(\Gamma \mathbb{1} = 0\) is called a vorticity matrix. If \(\Gamma\) is related by (1) to a Markov chain \(P\) with invariant distribution \(\pi\), it is called the vorticity of \(P\) and \(\pi\). This notion can be considered to be essentially synonymous to non-reversibility although it has a specific definition. It will be a key ingredient in the construction of a non-reversible version of Metropolis-Hastings.

Remark 2.2 (Bibliographical note). The application of matrices with the properties of Lemma 2.1 to Markov chains was previously considered in [SGS10].

2.2 Classical reversible Metropolis-Hastings

In the Metropolis-Hastings algorithm a reversible Markov chain \(P_0\) with a given invariant distribution \(\pi\) is constructed. We will assume, mainly for simplicity, throughout this paper that \(\pi(x) > 0\) for all \(x \in S\). As an ingredient for the construction of \(P_0\), a Markov chain \(Q\) is used, with the property that

\[Q(y, x) = 0\text{ whenever } Q(x, y) = 0, \quad x, y \in S.\]

In other words, whenever a transition from \(x\) to \(y\) has positive probability, the reverse probability also has positive probability. Condition (2) can be seen as a symmetric structure condition. The Hastings ratio \(R_0(x, y)\) is defined as

\[R_0(x, y) := \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}, \quad \text{for all } x, y \in S \text{ for which } \pi(x)Q(x, y) \neq 0.\]

The value of \(R_0\) at other values of \(x, y\) has no significance. With this definition of \(R_0\), acceptance probabilities are defined as

\[A_0(x, y) := \min (1, R_0(x, y)),\]

and transition probabilities \(P_0(x, y)\) are defined by

\[P_0(x, y) := \begin{cases} Q(x, y)A_0(x, y), & x \neq y, \\ Q(x, x) + \sum_{z \neq x} Q(x, z)(1 - A_0(x, z)), & x = y. \end{cases}\]

It is a straightforward exercise to show that the chain \(P_0\) has \(\pi\) as its invariant distribution. An important step is the observation that \(R_0(x, y) \leq 1\) if and only if \(R_0(y, x) \geq 1\), which will be a recurring phenomenon in the sequel.

2.3 Generalization to non-reversible Metropolis-Hastings

We will now discuss how this framework can be extended to construct Markov chains that are, in general, non-reversible. Let \(\Gamma \in \mathbb{R}^{n \times n}\) be a vorticity matrix satisfying \(\Gamma \mathbb{1} = 0\), and let \(Q\) be the transition matrix of a Markov chain, satisfying (2). Again, \(\pi : S \to (0, \infty)\) is some distribution that is not necessarily normalized, but that has only positive entries.

We define for \(x, y \in S\) the non-reversible Hastings ratio as

\[R_\Gamma(x, y) := \frac{\Gamma(x, y) + \pi(y)Q(y, x)}{\pi(x)Q(x, y)}, \quad \text{if } \pi(x)Q(x, y) \neq 0,\]

and let, as before, the acceptance probabilities \(A_\Gamma\) be

\[A_\Gamma(x, y) := \min (1, R_\Gamma(x, y)).\]
Entries of $\Gamma$ can be negative. In order to avoid the situation that $A_\Gamma$ becomes negative, we will explicitly constrain vorticity matrix $\Gamma$ to satisfy
\[
\Gamma(x, y) \geq -\pi(y)Q(y, x) \quad \text{for all } x, y \in S.
\] (8)

Equation (8) implies, by skew-symmetry of $\Gamma$, that
\[-\pi(y)Q(y, x) \leq \Gamma(x, y) \leq \pi(x)Q(x, y) \quad \text{for all } x, y \in S.
\]

In particular, by the symmetric structure condition (2), $\Gamma$ should have zeroes wherever $Q$ has zeroes. As with Metropolis-Hastings, the transition probabilities $P_\Gamma(x, y)$ are defined by
\[
P_\Gamma(x, y) := \begin{cases} 
Q(x, y)A_\Gamma(x, y), & x \neq y, \\
Q(x, x) + \sum_{z \neq x} Q(x, z)(1 - A_\Gamma(x, z)), & x = y.
\end{cases}
\] (9)

Indeed $P_\Gamma$ is a matrix of transition probabilities. For $\Gamma = 0$, $A_\Gamma$ and therefore $P_\Gamma$ reduce to $A_0$ and $P_0$, so that the chosen notation is consistent.

In order to check that the proposed Markov chain has $\pi$ as its invariant density, we need to verify that $\Gamma$, $\pi$ and $P_\Gamma$ are related through (1). As a crucial step, we employ the following lemma, in analogy with Metropolis-Hastings.

**Lemma 2.3.** Let $\Gamma$ be a vorticity matrix, $Q$ a matrix of transition probabilities satisfying (2), $\pi$ a distribution that is nowhere zero, such that (3) holds. Let $R_\Gamma$ be as above. Then $R_\Gamma(x, y) \geq 1$ whenever $R_\Gamma(x, y) \leq 1$, for any $x, y \in S$ for which $Q(x, y) \neq 0$.

**Proof.** Suppose $R_\Gamma(x, y) \leq 1$, i.e. $\Gamma(x, y) + \pi(y)Q(y, x) \leq \pi(x)Q(x, y)$. Then
\[
\pi(x)Q(x, y) + \Gamma(y, x) = \pi(x)Q(x, y) - \Gamma(x, y) \geq \pi(y)Q(y, x),
\]
so that $R_\Gamma(x, y) \geq 1$, using that $\Gamma$ is skew-symmetric.

Using the previous lemma, it is now straightforward to show that $\Gamma$ is the vorticity matrix of $(P_\Gamma, \pi)$.

**Lemma 2.4.** Let $Q$ be a Markov chain, $\Gamma$ a vorticity matrix, and $\pi$ a distribution on $S$, such that (2) and (3) are satisfied. Let $P_\Gamma$ be defined through (6), (7) and (9). Then (1) holds for $(P_\Gamma, \pi)$, i.e.
\[
\Gamma(x, y) = \pi(x)P_\Gamma(x, y) - \pi(y)P_\Gamma(y, x), \quad \text{for all } x, y \in S.
\]

**Proof.** If $x, y$ are such that $R_\Gamma(x, y) < 1$. Then $A_\Gamma(x, y) = R_\Gamma(x, y)$, and $A_\Gamma(y, x) = 1$. Therefore
\[
\pi(x)P_\Gamma(x, y) = \pi(x)Q(x, y)A_\Gamma(x, y) = \pi(x)Q(x, y) \left( \frac{\Gamma(x, y) + \pi(y)Q(y, x)}{\pi(x)Q(x, y)} \right)
\]
\[
= \Gamma(x, y) + \pi(y)Q(y, x) = \Gamma(x, y) + \pi(y)A_\Gamma(y, x)Q(y, x) = \Gamma(x, y) + \pi(y)P_\Gamma(x, y),
\]
using Lemma 2.3 (i). so that (1) is satisfied for such $x$ and $y$. The case where $R_\Gamma(x, y) \geq 1$ is analogous, using Lemma 2.3 (ii).

Finally, since by assumption $\Gamma 1 = 0$, Lemma 2.1 (ii) gives that $\pi$ is invariant for $P_\Gamma$. We have obtained our main result.

**Theorem 2.5.** Let $Q$ be a Markov chain, $\Gamma$ a vorticity matrix, and $\pi$ a distribution on $S$, such that (2) and (3) are satisfied. Let $P_\Gamma$ be defined through (6), (7) and (9). Then $P_\Gamma$ has $\pi$ as invariant distribution and $\Gamma$ as its vorticity matrix.

The following trivial observation serves to indicate the generality of this approach. It asserts that every Markov chain may be build, in a trivial way, by the described procedure.
Proposition 2.6. Let $P$ be a Markov chain with invariant distribution $\pi$ and corresponding vorticity $\Gamma$, satisfying (2). Then, for $Q = P$ and $P_T$ as defined through (6), (7), and (9), we have $P_T = Q = P$.

Proof. It suffices to note, that by $Q = P$ and (1), $A_T(x, y) = 1$ for all $x, y \in S$, $x \neq y$.

Remark 2.7. If, for some pair $(x, y) \in S \times S$, (8) holds with equality, the transition probability $P_T(x, y) = 0$ even when $Q(x, y) \neq 0$. Therefore irreducibility of $Q$ does not imply irreducibility of $P_T$, unless we impose the stronger condition:

$$\Gamma(x, y) > -\pi(y)Q(y, x) \quad \text{for all } x, y \in S. \quad (10)$$

The application of these results in practice will be discussed in Section 4. First we discuss a theoretical advantage of non-reversible Metropolis-Hastings.

3 Non-reversible chains constructed from reversible chains improve asymptotic variance

In this section a second method of constructing non-reversible chains is discussed (Section 3.1). This conceptually simple approach consists of adding vorticity to a reversible chain that already has the desired invariant distribution. In Section 3.2 it is explained how non-reversible Metropolis-Hastings is essentially equivalent to the construction of Section 3.1. The advantage of the second approach is that it allows, using a result of [SGS10], a detailed analysis of the asymptotic variance, which is a performance measure of the MCMC method. Because of its importance, their analysis is repeated in some more detail in Section 3.3. In particular, it is shown that adding vorticity to a reversible chain always improves the asymptotic variance. Translating this result back to non-reversible Metropolis-Hastings gives us a way of incorporating non-reversibility in Metropolis-Hastings with guaranteed improved performance; this observation will play a key role in ‘Approach (ii)’ of Section 4.

3.1 Adding vorticity to a reversible chain

Suppose $K$ is a reversible chain with respect to the distribution $\pi$ and let $\Gamma$ be a given vorticity matrix. Define

$$P = K + \frac{1}{2}\text{diag}(\pi)^{-1}\Gamma. \quad (11)$$

Then

$$P\mathbb{I} = K\mathbb{I} + \frac{1}{2}\text{diag}(\pi)^{-1}\Gamma\mathbb{I} = K\mathbb{I} = \mathbb{I},$$

and

$$\pi^T P = \pi^T K + \frac{1}{2}\pi^T \text{diag}(\pi)^{-1}\Gamma\mathbb{I} = \pi^T K + 0 = \pi^T.$$

Hence $\Gamma$ is the vorticity matrix corresponding to $\pi$ and $P$. Indeed,

$$\text{diag}(\pi)P - P^T \text{diag}(\pi) = \text{diag}(\pi)K + \frac{1}{2}\Gamma - K^T \text{diag}(\pi) - \frac{1}{2}\Gamma^T = \Gamma,$$

since $K$ satisfies detailed balance and $\Gamma$ is skew-symmetric. In order for (11) to have non-negative entries, and to save the irreducibility structure of $K$, we require that

$$\Gamma(x, y) > -2\pi(x)K(x, y), \quad \text{for all } x, y \in S,$$

which by symmetry of $\text{diag}(\pi)K$ is equal to the constraint

$$\Gamma(x, y) > -2\pi(y)K(y, x), \quad \text{for all } x, y \in S. \quad (12)$$

This condition should be compared to (10). We see that $P$ is a stochastic matrix, and $\pi$ is invariant with respect to $P$. We see that in the construction (11), we have more freedom in choosing $\Gamma$, as the bound for $\Gamma$ is a factor two larger.

Essentially what we have done, is constructing $\text{diag}(\pi)P$ from its symmetric part $\text{diag}(\pi)K$ and its anti-symmetric part, $\frac{1}{2}\Gamma$. 

5
Remark 3.1. This construction can not be extended in a straightforward way to the case where the invariant distribution of the reversible chain $K$ is not proportional to $\pi$.

Remark 3.2. This construction was observed earlier in [SGSI0].

### 3.2 Equivalence with non-reversible Metropolis-Hastings

Let $H$ be an irreducible Markov chain satisfying the symmetric structure condition \([\text{2}]\). Let $\pi$ be a given density that is nowhere zero. Let $K$ be the (reversible) Metropolis-Hastings chain defined by $H$ and $\pi$, i.e. for $x, y \in S$, $x \neq y$,

$$K(x, y) := \min \left( 1, \frac{\pi(y)H(y, x)}{\pi(x)H(x, y)} \right) H(x, y),$$

(13)

and with remaining values $K(x, x)$ as usual determined by the requirement that $K$ defines a Markov chain. Let $\Gamma$ be a vorticity matrix, and define

$$P(x, y) = K(x, y) + \frac{1}{2\pi(x)} \Gamma(x, y), \quad x, y \in S.$$  

(14)

**Lemma 3.3.** Let $P$ be as constructed above. The following are equivalent:

(i) $P$ defines a valid transition matrix;

(ii) The following requirement on $\Gamma$, $\pi$ and $H$ is satisfied:

$$|\Gamma(x, y)| \leq 2 \min(\pi(x)H(x, y), \pi(y)H(y, x)) \quad \text{for all } x, y \in S.$$  

(15)

If (i) and (ii) hold, then $P$ has invariant distribution $\pi$ and vorticity matrix $\Gamma$.

**Proof.** If (i) holds, it is clear that $\pi$ is invariant and $\Gamma$ is its vorticity matrix (by reversibility of $K$).

(i) $\Rightarrow$ (ii): In case $x = y$, then $\Gamma(x, y) = 0$. Suppose $x \neq y$. When $H(x, y) = 0$, then $K(x, y) = 0$. Requiring that the entries of $P$ are non-negative, we therefore necessarily have $\Gamma(x, y) \geq 0$, and by skew-symmetry, $\Gamma(y, x) \leq 0$. Interchanging $x$ and $y$ gives $\Gamma(x, y) = 0$.

Now suppose $H(x, y) \neq 0$. Again requiring that the entries of $P$ are non-negative, we have that

$$K(x, y) + \frac{1}{2\pi(x)} \Gamma(x, y) \geq 0,$$

For $H(x, y) \neq 0$, by the definition of $K(x, y)$, this translates in the requirement that

$$\min\left(1, \frac{\pi(y)H(y, x)}{\pi(x)H(x, y)} \right) H(x, y) + \frac{1}{2\pi(x)} \Gamma(x, y) \geq 0,$$

so that (after multiplying by $2\pi(x)$ and rearranging)

$$\Gamma(x, y) \geq -2 \min(\pi(x)H(x, y), \pi(y)H(y, x)).$$

By skew-symmetry of $\Gamma$, it follows that

$$\Gamma(y, x) = -\Gamma(x, y) \leq 2 \min(\pi(x)H(x, y), \pi(y)H(y, x)).$$

Since these inequalities hold for all $x$ and $y$, for which $x \neq y$ and $H(x, y) \neq 0$, interchanging $x$ and $y$ gives

$$\Gamma(x, y) \leq 2 \min(\pi(x)H(x, y), \pi(y)H(y, x)),$$

which establishes (15).

(ii) $\Rightarrow$ (i): The non-negativity of the entries of $P$ follows directly by reversal of the above argument. Furthermore

$$\sum_{y \in S} P(x, y) = \sum_{y \in S} \left( K(x, y) + \frac{1}{2\pi(x)} \Gamma(x, y) \right) = \sum_{y \in S} K(x, y) = 1,$$

so that $P$ is a valid transition matrix. $\square$
Proposition 3.4. Let $Q$ be irreducible, and suppose $Q$ satisfies the symmetric structure condition $(\mathcal{S})$. Furthermore suppose the non-negativity condition $(\mathcal{N})$ is satisfied. Define $H(x, y) = Q(x, y) - \frac{1}{\pi(x)} \Gamma(x, y)$, $x, y \in S$. Let $P$ be as constructed by $(\mathcal{R})$, $(\mathcal{Q})$ from $\Pi$, $\pi$ and $\Gamma$. Then $P$ coincides with $P_1$ as constructed in Section 2.3 by $(\mathcal{R})$, $(\mathcal{Q})$, and $(\mathcal{S})$.

Proof. We compute, for $x \neq y$, with $Q(x, y) \neq 0$,

$$\pi(x)P(x, y) = \pi(x) \min \left(1, \frac{\Gamma(x, y) + \pi(y)Q(y, x)}{\pi(x)Q(x, y)}\right)Q(x, y)$$

$$= \min (\pi(x)Q(x, y), \Gamma(x, y) + \pi(y)Q(y, x))$$

$$= \min (\pi(x)H(x, y) + \frac{1}{2}\Gamma(x, y), \Gamma(x, y) + \pi(y)H(y, x) + \frac{1}{2}\Gamma(y, x))$$

$$= \min (\pi(x)H(x, y), \pi(y)H(y, x)) + \frac{1}{2}\Gamma(x, y)$$

$$= \pi(x) \min \left(1, \frac{\pi(y)H(y, x)}{\pi(x)H(x, y)}\right)H(x, y) + \frac{1}{2}\Gamma(x, y) = \pi(x)P(x, y).$$

Both $P$ and $P_1$ have zeroes on the off-diagonal entries for which $Q$ has zeroes. Since both $P$ and $P_1$ represent transition probabilities, this also fixes there diagonal elements, which concludes the proof. \qed

Remark 3.5. Conversely, the construction of $P$ as a sum of a reversible matrix $K$ and a vorticity component $\frac{1}{2} \text{diag}(\pi)^{-1} \Gamma$ can be considered to be a special case of non-reversible Metropolis-Hastings (as discussed in Section 2.3). This is an immediate consequence of Proposition 2.6.

3.3 Analysis of asymptotic variance of non-reversible chains

The main result of this section is based on [SGS10]. In Appendix A a self-contained version of the proof is provided.

Let $K$ be a transition matrix of a reversible Markov chain with invariant probability distribution $\mu$, and let $\Gamma$ be a vorticity matrix. Let $P = K + \frac{1}{2} \text{diag}(\mu)^{-1} \Gamma$ and suppose $P$ is the transition matrix of an irreducible Markov chain. As observed before, $P$ has $\mu$ as its invariant probability distribution. Let $f : S \to \mathbb{R}$ and write $\bar{f} = f - \mu(f)$. We consider the partial sums of $f(X_t^n)$,

$$S_N(x) := \sum_{k=1}^{N} f(X_t^k) - N\mu(f) = \sum_{k=1}^{N} \bar{f}(X_t^k), \quad N \in \mathbb{N},$$

where the chain $X_t^n$ has starting position $X_0^n = x \in S$. By the ergodic theorem, $\frac{1}{N}S_N(x) \to \mu(f) := \sum_{i=1}^{N} \mu(i)f(i)$ for any $x \in S$.

We will use the notation

$$(\phi, \psi)_\mu = \sum_{i=1}^{N} \mu(i)\phi(i)\psi(i), \quad \phi, \psi : S \to \mathbb{R}.$$ 

The Dirichlet form $\mathcal{E}$ corresponding to $\mu$ is defined to be the quadratic form

$$\mathcal{E}(\phi, \psi) = (\phi, \psi)_\mu - (P\phi, P\psi)_\mu, \quad \phi, \psi : S \to \mathbb{R}.$$ 

The following result is known as the Central Limit Theorem for Markov chains [MT93, Theorem 17.4.4].

Proposition 3.6. Suppose $P$ is irreducible. Let $g$ be a solution of the Poisson equation for $P$ and $f$, i.e. $(I - P)g = f - \mu(f)$, and let $\sigma^2 := \mathcal{E}(g, g)$. Then $\sqrt{N\sigma^2}S_N(x)$ converges in distribution to the standard normal distribution, as $N \to \infty$. In particular, $\lim_{N \to \infty} N\mathbb{E} \left(\frac{1}{N}S_N(x)\right)^2 = \sigma^2$ for all initial conditions $x \in S$. 

7
We call $\sigma_f^2$ the asymptotic variance of $f$ with respect to $P$. Let $\sigma_{\Gamma f}^2$ denote the asymptotic variance of $f$ with respect to $P$, and $\sigma_{0f}^2$ the asymptotic variance of $f$ with respect to $K$.

The following theorem may be found in Appendix A as Corollary A.5.

**Theorem 3.7.** Suppose $\Gamma$ is a vorticity matrix that is not identically zero, and such that $P$ as defined above is a valid transition matrix. Then for all $f : S \to \mathbb{R}$, we have $\sigma_{\Gamma f}^2 \leq \sigma_{0f}^2$, and there exists an $f : S \to \mathbb{R}$ such that $\sigma_{\Gamma f}^2 < \sigma_{0f}^2$.

In Section 4.2 ("Approach (ii)") we discuss the application of this result to the construction of non-reversible Metropolis-Hastings schemes with a guaranteed improvement of asymptotic variance.

4 User's guide to non-reversible Metropolis-Hastings

In this section we discuss the application of the theory of the previous chapters to sampling problems. In a typical problem we are given a target distribution $\pi : S \to [0, \infty)$ that is not necessarily normalized. To apply non-reversible Metropolis-Hastings, we have to combine two elements:

- an irreducible Markov chain on $S$ with transition matrix $Q$ satisfying the symmetric structure condition (2);
- a vorticity matrix $\Gamma$ that is consistent with $Q$, in the sense that (10) is satisfied.

There are two different approaches to finding a compatible combination of $Q$ and $\Gamma$:

(i) Start with a non-reversible irreducible proposal chain $Q$, of which we know a invariant distribution (which does not need to be normalized).

(ii) Start with any proposal chain $H$ and vorticity matrix $\Gamma_0$ and combine these in a theoretically 'optimal' way.

The two approaches will be discussed in Sections 4.1 and 4.2, respectively. Approach (ii) may be considered the most favourable approach. This is because this second approach uses the analysis of Section 3.3 to have a guaranteed improvement of asymptotic variance. Also, in a numerical example, discussed in Section 5, approach (ii) seems to outperform approach (i) in many cases (but not all!) with respect to mixing time and spectral gap.

As an important tool required for method (ii), but essentially for both methods, it is useful to understand, construct and manipulate vorticity matrices. Therefore we discuss the structure of the set of vorticity matrices in more detail in Section 4.3.

4.1 Approach (i): Using the vorticity of a proposal chain with known invariant distribution

Suppose $Q$ is an irreducible Markov chain, satisfying the symmetric structure condition (2) with invariant distribution $\rho$. Then $\rho$ is everywhere non-zero by irreducibility of $Q$. Let $\Gamma_Q(x,y) := \rho(x)Q(x,y) - \rho(y)Q(y,x)$, $x,y \in S$, and define $\Gamma := \alpha \Gamma_Q$, with $\alpha \geq 0$. In other words, up to a scalar multiple we are using the vorticity matrix $\Gamma_Q$ of $Q$ in the construction of $P\Gamma$. By definition $\Gamma$ is skew-symmetric and satisfies $\Gamma \mathds{1} = 0$. Furthermore if $Q$ is reversible, $\Gamma = 0$. Condition (8) translates into the requirement

$$0 \leq \Gamma(x,y) + \pi(y)Q(y,x) = \alpha \rho(x)Q(x,y) + [\pi(y) - \alpha \rho(y)]Q(y,x), \quad x,y \in S.$$  

The reader is reminded that condition (8) guarantees that the resulting chain $P$ constructed by the non-reversible Metropolis-Hastings scheme is a proper transition matrix, and the slightly stronger condition (10) ensures that $P$ inherits the irreducibility of $Q$.
We see that a sufficient condition in order for (S) to hold is
\[ 0 \leq \alpha \leq \min_{y \in S} \left( \frac{\pi(y)}{\rho(y)} \right). \]  
(16)

We may obtain a better bound by using the ratio
\[ \theta := \max \left\{ \frac{\rho(x)Q(x, y)}{\rho(y)Q(y, x)} : \rho(y)Q(y, x) \neq 0 \right\} \geq 1, \]  
(17)
where equality holds only in case \((Q, \rho)\) is reversible. Then
\[ \Gamma(x, y) + \pi(y)Q(y, x) = \alpha \rho(x)Q(x, y) - \alpha \rho(y)Q(y, x) + \pi(y)Q(y, x) \]
\[ \geq \left[ -\alpha \left( 1 - \frac{1}{\theta} \right) \rho(y) + \pi(y) \right] Q(y, x), \]
so that a sufficient condition for (S) to hold is
\[ 0 \leq \alpha \leq \alpha_{\text{max}} := \left( \frac{\theta}{\theta - 1} \right) \min_{y \in S} \left( \frac{\pi(y)}{\rho(y)} \right) \]  
(18)
for \(\theta > 1\). The case where \(\theta = 1\) corresponds to reversibility of \((Q, \rho)\); in this case \(\Gamma_Q = 0\) so that the choice of \(\alpha\) is of no significance.

Remark 4.1. If the proposal chain \(Q\) is reversible, then by definition \(\rho(x)Q(x, y) = \rho(y)Q(y, x)\) for some distribution \(\rho\) that is everywhere non-zero. Choosing \(\Gamma\) to be the vorticity of the proposal chain would provide a reversible Metropolis-Hastings chain, since \(\Gamma = 0\) in this case.

4.2 Approach (ii): Guaranteed improvement of asymptotic variance

Corollary [A,5] provides us with a practical approach to the construction of non-reversible Metropolis-Hastings schemes which are guaranteed to improve asymptotic variance.

Again we wish to sample from some invariant distribution \(\pi\) (not necessarily normalized). Suppose we have at our disposal an irreducible chain with transition matrix \(H\), satisfying symmetric structure condition (2), and a vorticity matrix \(\Gamma_0\). Suppose furthermore that (15) is satisfied for \(\Gamma = \eta \Gamma_0\), for some \(\eta \in \mathbb{R}^+, \eta \neq 0\). In particular, it is required that \(\Gamma(x, y) = 0\) whenever \(H(x, y) = 0\) (or equivalently, by symmetric structure, \(H(y, x) = 0\)). In words: there is no vorticity possible along transitions that have zero probability under \(H\). Then \(Q\), defined by
\[ Q(x, y) = H(x, y) + \frac{\eta}{2\pi(x)} \Gamma_0(x, y) \]  
(19)
is a valid transition matrix.

We will take \(Q\) and \(\Gamma = \eta \Gamma_0\) as our proposal chain and proposal vorticity for non-reversible Metropolis-Hastings, respectively. By Proposition 3.4, the resulting transition \(P_\Gamma\) coincides with the chain obtained by adding skew symmetric part \(\frac{1}{2\pi(x)} \Gamma(x, y)\) to the reversible Metropolis-Hastings chain with matrix \(K(x, y)\), as given in (13). By Corollary [A,5], this construction is guaranteed to have improved asymptotic variance relative to reversible Metropolis-Hastings, with proposal chain \(H\). Furthermore (23) tells us that, in order to have the best asymptotic variance, we should choose \(|\eta|\) as large as possible. The maximal value of \(\eta\) is determined by (15), i.e.
\[ \eta_{\text{max}} = \sup \{ \eta > 0 : \eta |\Gamma_0(x, y)| \leq 2 \min (\pi(x)H(x, y), \pi(y)H(y, x)) \text{ for all } x, y \in S \}. \]  
(20)

Remark 4.2. Of course the two approaches may lead to identical results, depending on the choice of initial data. Throughout the discussion in this and the previous sections, the use of the symbols \(Q, \pi\) and \(\Gamma\) is consistent. In Approach (i) we start with \(Q\), and from this the matrix \(H\) of Approach (ii) may be computed as \(H(x, y) = Q(x, y) - \frac{\eta}{2\pi(x)} \Gamma(x, y)\). Conversely, \(Q\) as defined in Approach (ii) from \(H, \pi, \eta\) and \(\Gamma_0\), by (19), will have some invariant distribution \(\rho\).
4.3 Remarks on vorticity matrices

Identifying a suitable vorticity matrix $\Gamma$ corresponding to a given proposal chain may in general pose a challenge. In this section we collect some results of independent interest that improve our understanding of vorticity matrix, as well as mention simple ways of constructing vorticity matrices. The results of Section 4.3 are not used in the remainder of this paper.

A cycle in a graph $G = (V, E)$ with $n$ vertices may be represented by a matrix $\Gamma$ having the following properties:

(i) $\Gamma = -\Gamma^T$.

(ii) $\Gamma \mathbb{1} = 0$, i.e. $\sum^n_{j=1} \Gamma(i,j) = 0$ for all $i = 1, \ldots, n$.

(iii) $\Gamma(i, j) = 0$ whenever $A(i, j) = 0$, where $A$ is the adjacency matrix of $G$.

(iv) $\Gamma \neq 0$.

This is made precise by the following Proposition.

**Proposition 4.3.** Let $G$ be a graph over $n$ vertices and let $\Gamma$ satisfy the conditions (i)-(iv) above. Then there exists a cycle, i.e. a non-empty sub-graph $C = (V_C, E_C)$ with $V_C = \{x_0, x_1, \ldots, x_k\}$ and $E_C = \{x_0x_1, x_1x_2, \ldots, x_kx_0\}$, such that for every $x_i, x_j \in E_C$, $\Gamma(x_i, x_j) > 0$. Conversely, if there exists a non-empty cycle $C = (V_C, E_C)$, then there exists a matrix $\Gamma$ satisfying (i)-(iv), and $\Gamma(i, j) > 0$ for $i, j \in E_C$.

**Proof.** "⇒": Since $\Gamma \neq 0$ and $\Gamma$ is skew-symmetric, there exists a pair $(i_0, i_1)$, $i_1 \neq i_0$, such that $\Gamma(i_0, i_1) > 0$. Since $\Gamma$ is skew-symmetric, $\Gamma(i_1, i_0) < 0$, and because rows sum to zero, there must be a positive element on row $i_1$. Suppose this is at position $(i_1, i_2)$. Again $i_2 \neq i_1$. We may repeat this procedure until we encounter a node $i_k$ that we already obtained. If this vertex is $i_0$, we are done. If this vertex is $i_l = i_k$ for some $0 < l < k - 1$ (note $l = k$ is impossible by skew-symmetry), we obtain a cycle $\{x_{i_1}, x_{i_1+1}, \ldots, x_k\}$ with the required properties by removing vertices $i_0, \ldots, i_{l-1}$.

"⇐": Let the entries of $\Gamma(i, j) = 1$ and $\Gamma(j, i) = -1$ whenever there is a edge between $i$ and $j$ in the directed cycle $x_0, x_1, \ldots, x_k$, and $\Gamma(i, j) = 0$ otherwise. For any $i$, $\sum^n_{j=1} \Gamma(i, j) = 2|\{\text{edges out of } i\} - |\{\text{edges into } i\}| = 0$, so that (ii) is satisfied. The other conditions (i), (iii), (iv) are clearly verified. \( \square \)

**Theorem 4.4.** Let $V_n$ denote the set of $n \times n$ vorticity matrices, i.e.

$$V_n = \{ \Gamma \in \mathbb{R}^{n \times n} : \Gamma = -\Gamma^T, \Gamma \mathbb{1} = 0 \}.$$  

(i) $V_n$ is a linear space of dimension $\frac{1}{2}(n - 1)(n - 2)$.

(ii) Let $\{u_1, \ldots, u_{k-1}\}$ denote a linear basis of $\mathbb{1}^\perp$ in $\mathbb{R}^n$. Then

$$V_n = \text{span} \left\{ \Gamma_{ij} : \Gamma_{ij} = u_i u_j^T - u_j u_i^T, i \neq j \right\},$$

and $\Gamma_{ij}$ and $\Gamma_{kl}$, with $i \neq j$ and $k \neq l$, are linearly independent if and only if the unordered pair $(i, j)$ is not equal to the unordered pair $(k, l)$.

**Proof.** See [SGS10] Proposition 2]. \( \square \)

**Remark 4.5.** (Construction from two Markov chains). Let $P$, $Q$ be two Markov chains, both with invariant distribution $\pi$ and vorticity matrices $\Gamma_P$ and $\Gamma_Q$, respectively. Then $R := PQ$ defines the transition matrix function of a Markov chain that also has $\pi$ as its invariant distribution. We compute

$$\Gamma_R := \text{diag}(\pi)PQ - Q^TP^T \text{diag}(\pi)$$

$$= (\text{diag}(\pi)P - P^T \text{diag}(\pi))Q + P^T \text{diag}(\pi)Q - Q^TP^T \text{diag}(\pi)$$

$$= \Gamma_P Q + P^T (\text{diag}(\pi)Q - Q^T \text{diag}(\pi)) + P^T Q^T \text{diag}(\pi) - Q^T P^T \text{diag}(\pi)$$

$$= \Gamma_P Q + P^T \Gamma_Q + [P^T, Q^T] \text{diag}(\pi),$$
\[
\cdot, \cdot\]
denotes the usual matrix commutator, \([A, B] = AB - BA\).

In case \(P\) and \(Q\) are both reversible, then \(\Gamma_P = \Gamma_Q = 0\), and \(\Gamma_R\) may be simplified to

\[
\Gamma_R = [P^T, Q^T] \text{diag}(\pi) = P^T Q^T \text{diag}(\pi) - Q^T P^T \text{diag}(\pi)
= P^T \text{diag}(\pi)Q - Q^T \text{diag}(\pi)P = \text{diag}(\pi)(PQ - QP) = \text{diag}(\pi)[P, Q].
\]

**Remark 4.6 (Vorticity as the skew-symmetric part of a doubly stochastic matrix).** Let \(M \in \mathbb{R}^{n \times n}\) satisfy

\[
M = v, \quad M^T = v,
\]

for some \(v \in \mathbb{R}^n\). Then \(\Gamma := M - M^T\) is skew-symmetric and \(\Gamma = (M - M^T) = v - v = 0\).

## 5 Numerical experiment: multi-modal distributions on the \(n\)-cycle

In this section we apply the theoretical observations of the previous sections to the example of a multi-modal distribution on the \(n\)-cycle. The experimental set-up is discussed in Section 5.1 and the corresponding results in Section 5.2. A concise discussion of these results may be found in Section 5.3.

### 5.1 Experimental set-up

Let \(S = \{1, \ldots, n\}\) with \(n > 2\) odd (to avoid issues of periodicity). As a ‘goal’ distribution consider

\[
\pi(x) = \exp \left( -\beta \sin^2 \left( \frac{M \pi x}{n} \right) \right), \quad x \in \{1, \ldots, n\}.
\]  \hspace{1cm} (21)

Here \(M = 0, 1, 2, \ldots\) is a parameter indicating the number of modes in the distribution, and \(\beta \geq 0\) is the ‘inverse temperature’ or steepness of the distribution.

For \(n = 49\), plots of \(\pi\) for some values of \(\beta\) and \(M\) are provided in Figure 1.

We should determine which proposal chain \(Q\) and which vorticity matrix \(\Gamma\) we shall use in the non-reversible Metropolis-Hastings scheme. Together, \(\pi\), \(\Gamma\) and \(Q\) will completely determine the \(P\) by the construction described in Section 2.3. In order to determine suitable \(Q\) and \(\Gamma\) we will apply the two approaches discussed in Section 4.

As reversible proposal chain \(H\) we take

\[
H(i, j) = \begin{cases} 
\frac{1}{2} & \text{if } j - i = \pm 1 \text{ mod } n, \\
0 & \text{otherwise}.
\end{cases}
\]

The linear space of compatible vorticity matrices is one dimensional and spanned by the matrix

\[
\Gamma_0(i, j) = \begin{cases} 
1 & \text{if } j - i = 1 \text{ mod } n, \\
-1 & \text{if } j - i = -1 \text{ mod } n, \\
0 & \text{otherwise}.
\end{cases}
\]

We now consider the two approaches discussed in Section 4.

- **Approach (i):** As proposal chain \(Q\) for non-reversible Metropolis-Hastings we take \(Q = H + \gamma \Gamma_0\), so \(\Gamma Q = \gamma \Gamma_0\). In words, the chain \(Q\) moves with probability \(\frac{1}{2}(1 + \gamma)\) from node \(i\) to \(i + 1\) for \(i < n\), and then from \(n\) back to 1, and with probability \(\frac{1}{2}(1 - \gamma)\) in the reverse direction. Clearly \(Q\) has the uniform distribution as its invariant distribution, and the chain is reversible if and only if \(\gamma = 0\). We see that \(\gamma\) is a ‘measure of non-reversibility’ of \(Q\), or equivalently of the vorticity of \(Q\). In order for the ‘symmetric structure’ requirement, (2), to hold, we require \(\gamma < 1\).

As discussed in Section 4.1, we use \(Q\) as proposal chain and \(\Gamma = \alpha \Gamma Q\) as vorticity matrix for non-reversible Metropolis-Hastings, with

\[
\alpha = \alpha(c_1, \gamma) = c_1 \times \alpha_{\max}(\gamma),
\]
for $0 < \gamma < 1$ and $0 \leq c_1 < 1$ as motivated by Section 4.1 and where $\alpha_{\text{max}}$ is to be determined. If $\gamma = 0$, the value of $\alpha$ has no significance since the base chain $Q$ is reversible and thus $\Gamma_Q = 0$.

Since the (flat) invariant distribution $\rho$ of $Q$ used in the following construction does not need to be normalized, we choose $\rho := 1$ for simplicity. From (17) we compute $\theta = \max \left\{ \frac{\rho(x)Q(x,y)\rho(y)Q(y,x)}{\rho(y)Q(y,x)} : \rho(y)Q(y,x) \neq 0 \right\} = \max \left( \frac{1+\gamma}{1-\gamma}, \frac{1-\gamma}{1+\gamma} \right) = \frac{1+\gamma}{1-\gamma}, 0 \leq \gamma < 1$. Then (18) gives

$$\alpha_{\text{max}}(\gamma) = \left( \frac{\theta}{\theta - 1} \right) \min_i \left( \frac{\pi(i)}{\rho(i)} \right) = \left( \frac{1+\gamma}{2\gamma} \right) \left( \min_i \pi(i) \right), \quad 0 < \gamma < 1.$$

The case $\gamma = 0$ corresponds to the case in which the proposal chain $Q$ is reversible. The case $c_1 = 0$ corresponds to the case in which the resulting non-reversible Metropolis-Hastings chain is in fact reversible.

- Approach (ii): As in Section 4.2 we take as non-reversible Metropolis-Hastings proposal chain $Q = H + \frac{1}{2} \eta \text{diag}(\pi)^{-1} \Gamma_0$, where $\eta = c_2 \eta_{\text{max}}$ with $0 \leq c_2 < 1$ and, from (20),

$$\eta_{\text{max}} = \min_{x \in \mathcal{S}} \{ \pi(x)H(x,y) : H(x,y) \neq 0 \} = \min_{i \in \mathcal{S}} \pi(i).$$

We further take $\Gamma = c_2 \eta_{\text{max}} \Gamma_0$ as vorticity matrix in the non-reversible Metropolis-Hastings algorithm. The case $c_2$ corresponds to the case in which both the proposal chain and the resulting non-reversible Metropolis-Hastings chain are reversible.
The conditions $0 \leq \gamma < 1$ and $0 \leq c_1 < 1$ (for Approach (i)) en $0 \leq c_2 \leq 1$ (for Approach (ii)) ensure that the resulting chain $P$ is well-defined (in particular, that condition 5 is satisfied) and that it is irreducible.

To measure the performance of the different chains $P$ that we generate using this scheme we consider two closely related quantities:

- The **spectral gap** $\sigma := 1 - |\lambda_2|$. Here the eigenvalues of $P$, $\lambda_1, \ldots, \lambda_n$, are listed in order of descending absolute value and with multiplicities according to algebraic multiplicity. A large spectral gap indicates fast mixing;
- The $\varepsilon$-**mixing time** $t_{\text{mix}} := t_{\text{mix}} := \inf\{t \in \mathbb{N} : \max_{x \in S} \|P^t(x, \cdot) - \mu\|_{\text{TV}} < \varepsilon\}$. Here $\mu(x) = \frac{\pi(x)}{\sum_{y \in S} \pi(y)}$ is the normalization of $\pi$ and $\|\mu - \nu\|_{\text{TV}} := \frac{1}{2} \sum_{y \in S} |\mu(y) - \nu(y)|$ is the total variation distance of the discrete probability distributions $\mu$ and $\nu$. The choice of $\varepsilon$ is not of fundamental importance; we let $\varepsilon = \frac{1}{4}$. See [LPW09, Section 4.5].

We fix $n = 49$, choosing an odd value of $n$ to avoid problems of periodicity for the case $\gamma = 0$. The choice of $c_1$, $\gamma$ (Approach (i) and $c_2$ (Approach (ii)) influence spectral gap and mixing time. In our experiments we have optimized over discrete values of these parameters. We make further observations on this dependence in Section 6 below.

### 5.2 Experimental results

A natural question is for what shapes of invariant distributions $\pi$ we can expect the largest speed-up in sampling. To test this numerically, we varied parameters $\beta$ and $M$ in [21]. For each value of $(\beta, M)$ we optimized $c_1$, $\gamma$ and $c_2$ over a discrete set of values with respect to (a) spectral gap and (b) mixing time. The results of the two approaches are plotted in Figure 2 and for some values of $\beta$ and $M$ the results are displayed in Table 1.

In Figure 3 the dependence of the spectral gap and the mixing time on $n$ is displayed for different values of $\beta$, and $M = 2$.

**Remark 5.1.** The obtained optimal location of $(c, \gamma)$ for the case of a flat distribution ($\beta = 0$ and/or $M = 2$), as given in Table 1 ($c^*, \gamma^*) = (0.22, 0.11)$ is a true minimizer, in the sense that the values of $t_{\text{mix}}(c, \gamma)$ and $\sigma(c, \gamma)$ as functions of $c$ and $\gamma$ are certainly not flat.

### 5.3 Discussion of the experimental results

The following observations can be made from the numerical experiments. These are listed more or less in order of their importance.

- For a flat distribution ($M = 0$), a significant gain (of approximately factor 4) in both spectral gap and mixing time is already obtained. With just one mode ($M = 1$), there is some improvement, but relatively small. In this case, ordinary Metropolis-Hastings seems adequate. The largest improvement is found for two modes ($M = 2$). In this regime, at steepness $\beta = 4$, dramatic gains in mixing time and spectral gap are obtained. For a larger number of modes ($M > 2$), there is significant improvement using non-reversible Metropolis-Hastings, although with decreasing effectiveness as a function of $M$ (Figure 2).
- For a large range of values of $\beta$ ($0 \leq \beta \leq 3$), the multi-modal cases have better mixing times with non-reversible Metropolis-Hastings than the uni-modal chains have with reversible Metropolis-Hastings (Figure 2 (b), (d), (f)).
- The relative efficiency gain of non-reversible Metropolis-Hastings is increasing in the problem size $n$. E.g. for $\beta = 3$, $n = 5$, $\sigma_2^*/\sigma_{\text{MH}} = 2.5$, while for $\beta = 3$, $n = 257$, $\sigma_2^*/\sigma_{\text{MH}} = 20.0$. Similarly, for $\beta = 3$, $n = 5$, $t_{\text{mix},1}/t_{\text{mix, MH}} = 1.6$, while for $\beta = 3$, $n = 257$, $t_{\text{mix},1}/t_{\text{mix, MH}} = 9.4$ (Figure 3).
Table 1: Comparison of spectral gap and mixing time of non-reversible Metropolis-Hastings versus ordinary (reversible) Metropolis-Hastings for different choices of $\pi$ in terms of steepness ($\beta$) and number of modes ($M$). In the columns are displayed the optimal spectral gaps $\sigma_1^*, \sigma_2^*$ for approaches (i) and (ii), and for the ordinary Metropolis-Hastings spectral gap $\sigma_{MH}$, and similarly for the mixing time. Optimization in the variables $c_1, \gamma$ (Approach (i)) and $c_2$ (Approach (ii) is performed over 50 evenly spaces values between 0.00 and $1 - 10^{-4}$.

| $M$ | $\beta$ | $\sigma_1^*$ | $\sigma_2^*$ | $\sigma_{MH}$ | $t_{mix,1}^{*}$ | $t_{mix,2}^{*}$ | $t_{mix,MH}$ |
|-----|---------|-------------|-------------|--------------|---------------|---------------|-------------|
| 0   | 0       | 0.00814     | 0.02005     | 0.00205      | 116           | 450           | 456         |
| 1   | 1       | 0.00918     | 0.00948     | 0.00832      | 126           | 120           | 148         |
| 2   | 2       | 0.0132      | 0.0147      | 0.00907      | 115           | 92            | 164         |
| 3   | 3       | 0.0157      | 0.0184      | 0.0104       | 119           | 95            | 166         |
| 4   | 4       | 0.0177      | 0.0205      | 0.0122       | 120           | 100           | 159         |
| 2   | 1       | 0.00897     | 0.00919     | 0.00465      | 110           | 106           | 179         |
| 3   | 2       | 0.0129      | 0.0141      | 0.00248      | 109           | 83            | 310         |
| 4   | 3       | 0.0164      | 0.0185      | 0.00125      | 146           | 108           | 585         |
| 5   | 4       | 0.0104      | 0.00703     | 0.000598     | 237           | 176           | 1189        |
| 3   | 1       | 0.00848     | 0.00870     | 0.00633      | 120           | 116           | 164         |
| 4   | 2       | 0.0107      | 0.0125      | 0.00375      | 107           | 91            | 275         |
| 5   | 3       | 0.00868     | 0.0119      | 0.00195      | 145           | 109           | 517         |
| 6   | 4       | 0.00493     | 0.00592     | 0.000943     | 246           | 188           | 1055        |
| 4   | 1       | 0.00795     | 0.00817     | 0.00649      | 121           | 117           | 141         |
| 5   | 2       | 0.00901     | 0.0109      | 0.00403      | 117           | 97            | 224         |
| 6   | 3       | 0.00672     | 0.00902     | 0.00215      | 149           | 114           | 415         |
| 7   | 4       | 0.00377     | 0.00468     | 0.00105      | 276           | 227           | 847         |
| 5   | 1       | 0.00746     | 0.00768     | 0.00648      | 130           | 126           | 145         |
| 6   | 2       | 0.00782     | 0.00955     | 0.00407      | 129           | 107           | 225         |
| 7   | 3       | 0.00560     | 0.00737     | 0.00219      | 184           | 142           | 404         |
| 8   | 4       | 0.00312     | 0.00386     | 0.00107      | 310           | 251           | 810         |

- The absolute difference in spectral gap is largest for moderately steep distributions ($\beta = 4$). For higher values of $\beta$ the difference becomes smaller (Figure 2 (a), (c), (e), Table 1).
- Overall, Approaches (i) and (ii) have very similar performance, with in most cases an improved performance of Approach (ii). Only for values of $\beta \ll 1$, Approach (i) seems to outperform Approach (ii) (Figure 2).
- For Approaches (i) and (ii), the optimal values of $c_1$ and $c_2$ (respectively) are as close to 1 as possible (within the discrete set of choices) for values of $\beta \geq 2$. For $\beta < 2$ lower values of $c_1$ and $c_2$ are optimal.
- To maximize spectral gap and/or minimize mixing time, the optimal value of $\gamma$ in approach (i) depends on the problem parameters $\beta$ and $M$. The optimal value of $\gamma$ seems to be assumed for values strictly less than 1, which are increasing in $\beta$.

6 Discussion

The theoretical discussion of Section 3 and the numerical experiment of Section 5 illustrate a great gain by employing non-reversible Metropolis-Hastings. This encourages one to apply these results to other settings. The practical application of non-reversible Metropolis-Hastings depends on the identification of suitable vorticity matrices $\Gamma$ that are compatible with proposal chains $Q$, and establishing these in practical examples provides a promising direction of research.

We are currently working on the application of non-reversible Metropolis-Hastings to spin systems. The challenge there is to identify suitable vorticity structures that are compatible with a
proposal chain that performs single bit flips, and to investigate which of these structures provide the best MCMC performance. There is much that can be said, and is still being researched, about this example. Therefore we choose to postpone discussion of this example to a second paper.

As another application of non-reversible Metropolis-Hastings one can think of continuous spaces. It is straightforward to transfer the principle of non-reversible Metropolis-Hastings to general measurable spaces. As an application, one can think of the following. Suppose we wish to sample from a given distribution by means of some Langevin equation. In practical sampling algorithms, time is discretized, which introduces errors into the sampling method [RT96]. This should be corrected by performing an Metropolis-Hastings accept/reject step, but this destroys the possibly advantageous non-reversible nature of the Markov diffusion process. Here non-reversible Metropolis-Hastings may well be advantageous. This promising idea will be subject of future research.

Theoretical results on convergence to equilibrium of reversible chains are numerous. For non-reversible chains this topic is still a great challenge. One point in this context that deserves attention is the following. One can in principle apply the theory of Dirichlet forms as discussed e.g. in [Fil91] and reduce the problem of estimating mixing time of a non-reversible chain to estimating mixing time of the reversible chain $\hat{P} P$, where $\hat{P}$ is the time reversal of $P$. However, it is our experience that the obtained estimates do not capture the improvements that can be obtained by employing non-reversible chains. As far as I am aware, it remains a challenging open problem to provide, even for simple examples, estimates on mixing times that capture the difference between reversible and non-reversible chains.

Acknowledgements

I thankfully acknowledge valuable discussions with my colleagues Prof. Hilbert J. Kappen and Dr. Kevin Sharp, both at Radboud Universiteit Nijmegen, that have had substantial influence on the direction of research in this work.

A Proof of Theorem 3.7

In this Section we provide the proof of Theorem 3.7 based on the analysis in [SGS10]. Let $L^2(\mu)$ denote the linear space of functions $h : S \to \mathbb{R}$ and let $L^2_0(\mu)$ denote the linear subspace of functions $h : S \to \mathbb{R}$ satisfying $\mu(h) = 0$, where both spaces are endowed with the inner product $(\cdot, \cdot)_\mu$ as defined above. For $h : S \to \mathbb{R}$, write $\overline{h} = h - \mu(h) \in L^2_0(S)$ and $\text{var}(h) := \sum_{x \in S} \pi(x)(h(x) - \mu(h))^2 = (\overline{h}, \overline{h})_\mu$. The following lemma collects a few facts about the solution $g$ of the Poisson equation, $(I - P)g = \overline{f}$, and $\mathcal{E}(g, g)$. We will denote linear operators on $L^2_0(\mu)$ by calligraphic script, e.g. $\mathcal{A}, \mathcal{G}$, to distinguish these from matrices, denoted by Roman script, e.g. $K, P$. This distinction will be of importance when considering inverses and adjoints. The adjoint operator of $\mathcal{A}$ in $L^2_0(\mu)$ will be denoted by $\mathcal{A}^*$.

Lemma A.1. Suppose $P$ is irreducible. Then

(i) The linear mapping $\mathcal{A} : h \mapsto (I - P)h$ is invertible as a mapping from $L^2_0(\mu)$ into itself.

(ii) For $f : S \to \mathbb{R}$, let $g := \mathcal{G}\overline{f}$, so that $g$ is a solution of the Poisson equation for $P$ and $f$. Then

$$\mathcal{E}(g, g) = 2(\overline{f}, \mathcal{G}\overline{f})_\mu - ||\overline{f}||_\mu^2. \quad (22)$$

Proof. (i) $(I - P) : L^2_0(S) \to L^2_0(S)$. By irreducibility of $P$, the only functions $h : S \to \mathbb{R}$ satisfying $(I - P)h = 0$ are the constant functions, so that $I - P$ is injective and therefore invertible.
(ii) \( g \in L_0^2(\mu) \) and \( Pg = g - \mathcal{T} \). Therefore
\[
\mathcal{E}(g,g) = (g,g) - (Pg,Pg)_\mu = (g,g)_\mu - (g - \mathcal{T},g - \mathcal{T})_\mu \\
= 2(\mathcal{T},g)_\mu - (\mathcal{T},\mathcal{T})_\mu = 2(\mathcal{T},\mathcal{G}\mathcal{T})_\mu - ||\mathcal{T}||_\mu^2.
\]

From Lemma A.1 (iii), we see that the effect of the choice of \( P \) is, for fixed distribution \( \mu \), fully explained by the \( L_0^2(\mu) \)-self-adjoint part of \( \mathcal{G} \). The following lemma of [Mat92] will be helpful, as observed by [SGS10].

**Lemma A.2.** Let \( V \) be a finite dimensional inner product space and let \( \mathcal{A} : V \to V \) be linear and have self-adjoint part \( \mathcal{H} = H[\mathcal{A}] = \frac{1}{2}(\mathcal{A} + \mathcal{A}^*) \) and skew-adjoint part \( \mathcal{S} = S[\mathcal{A}] = \frac{1}{2}(\mathcal{A} - \mathcal{A}^*) \). Suppose \( \mathcal{H} \) is positive definite. Then \( \mathcal{A} \) is invertible and \( \mathcal{A}^{-1} \) has positive definite self-adjoint part
\[
H[\mathcal{A}^{-1}] = \frac{1}{2}(\mathcal{A}^{-1} + \mathcal{A}^{-*}) = (\mathcal{H} + \mathcal{S}^*\mathcal{H}^{-1}\mathcal{S})^{-1}.
\]

For \( \mathcal{A} \) as defined in [A.1] we have for the skew-adjoint and self-adjoint parts \( \mathcal{S} \) and \( \mathcal{H} \),
\[
\mathcal{S} f(x) = \frac{1}{2\mu(x)} \sum_{y \in S} \Gamma(x,y)f(y), \quad \mathcal{H} f(x) = \sum_{y \in S} (I - K)(x,y)f(y), \quad x \in S.
\]
The adjoint of \( \mathcal{S} \) in \( L_0^2(\mu) \) is then
\[
\mathcal{S}^* f(x) = -\mathcal{S} f(x) = \frac{1}{2\mu(x)} \sum_{y \in S} \Gamma(y,x)f(y).
\]

As an immediate corollary of Lemma A.2 we have the following result.

**Lemma A.3.** The \( L_0^2(\mu) \)-self-adjoint part of \( \mathcal{G} \) is
\[
H[\mathcal{G}] = (\mathcal{H} + \mathcal{S}^*\mathcal{H}^{-1}\mathcal{S})^{-1}.
\]

We can now describe the main result of [SGS10] regarding asymptotic variance of non-reversible Markov chains.

**Theorem A.4.** Let \( K \) be a transition matrix of a reversible Markov chain with invariant probability distribution \( \mu \). Let \( \Gamma_1 \) and \( \Gamma_2 \) be vorticity matrices such that \( P_i = K + \frac{1}{2} \text{diag}(\mu)^{-1}\Gamma_i, \) \( i = 1,2, \) are irreducible Markov transition matrices. The following are equivalent.

(i) For all \( f : S \to \mathbb{R} \), we have \( \sigma_i^2 f \leq \sigma_2^2 f \), where \( \sigma_i^2 f, i = 1,2 \), denote the asymptotic variances of \( f \) with respect to \( P_i \), \( i = 1,2 \).

(ii)
\[
\Gamma_1^T(\text{diag}(\mu) - \text{diag}(\mu)K + \mu\mu^T)^{-1}\Gamma_1 \succeq \Gamma_2^T(\text{diag}(\mu) - \text{diag}(\mu)K + \mu\mu^T)^{-1}\Gamma_2 \tag{23}
\]
where \( \succeq \) denotes the partial ordering with respect to the cone of positive definite matrices.

Furthermore, strict inequality holds in (i) for some \( f \) if and only if the left and right hand sides of (23) are not equal.

**Proof.** By the Central Limit Theorem, Proposition 3.6 the asymptotic variances \( \sigma_i^2 f \) are given by \( \mathcal{E}(\mu_i,\mu_i) \), where \( \mu_i = G_i(f - \mu f) \) is the solution of the Poisson equation for \( f \) and \( P_i \). Using Lemma A.1 (ii), we find that the dependence of \( \mathcal{E}(\mu_i,\mu_i) \) on \( P_i \) is equal to the first term in (22), 2(\( \mathcal{S},\mathcal{G}\mathcal{S} \))_\mu, which depends only on the \( \mu \)-self-adjoint part \( H[\mathcal{G}] \) of \( \mathcal{G} \). Therefore \( \sigma_1^2 f \leq \sigma_2^2 f \) for all \( f : S \to \mathbb{R} \) if and only if \( H[\mathcal{G}_1] \preceq H[\mathcal{G}_2] \). By Lemma A.3 this is equivalent to \( (\mathcal{H} + \mathcal{S}_1^*\mathcal{H}^{-1}\mathcal{S}_1)^{-1} \succeq (\mathcal{H} + \mathcal{S}_2^*\mathcal{H}^{-1}\mathcal{S}_2)^{-1} \). From linear algebra we know that this is equivalent to \( \mathcal{H} + \mathcal{S}_2^*\mathcal{H}^{-1}\mathcal{S}_2 \preceq \mathcal{H} + \mathcal{S}_1^*\mathcal{H}^{-1}\mathcal{S}_1 \), or equivalently \( \mathcal{S}_2^*\mathcal{H}^{-1}\mathcal{S}_2 \preceq \mathcal{S}_1^*\mathcal{H}^{-1}\mathcal{S}_1 \). Now on \( L_0^2(\mu) \), the linear operator \( \mathcal{H}^{-1} \)
allows the matrix representation \((I - K + 1_\mu^T)^{-1}\), and \(S^\ast\) and \(S^\ast\) have matrix representations
\[
\frac{1}{2} \text{diag}(\mu)^{-1} \Gamma \quad \text{and} \quad \frac{1}{2} \text{diag}(\mu)^{-1} \Gamma^T,
\]
respectively, resulting in (23) on \(L^2(\mu)\). Since \(\Gamma / BD = 0\), it follows that (23) holds on
\(L^2(\mu) = L^2_0(\mu) \oplus \text{span} 1\) and only if it holds on \(L^2_0(\mu)\).

From the above argument we see immediately that, if for some \(h : S \to \mathbb{R}\),
\[
(h, S_2^\ast \mathcal{H}^{-1} S_2 h)_{\mu} < (h, S_1^\ast \mathcal{H}^{-1} S_1 h)_{\mu},
\]
then there exists an \(f : S \to \mathbb{R}\) so that \(\sigma^2_{1,f} < \sigma^2_{2,f}\).

\textbf{Corollary A.5.} Suppose \(\Gamma_1\) is a vorticity matrix that is not identically zero, and such that \(P_i\) as defined in Theorem A.4 is a valid transition matrix. Suppose furthermore that \(\Gamma_2 = 0\). Then for all \(f : S \to \mathbb{R}\), we have \(\sigma^2_{1,f} \leq \sigma^2_{2,f}\), and there exists an \(f\) such that \(\sigma^2_{1,f} < \sigma^2_{2,f}\).

\textbf{References}

[DHN00] Persi Diaconis, Susan Holmes, and RM Neal. Analysis of a nonreversible Markov chain sampler. \textit{Annals of Applied Probability}, (June 1997), 2000.

[Dia08] Persi Diaconis. The Markov chain Monte Carlo revolution. \textit{Bulletin of the American Mathematical Society}, 46(2):179–205, November 2008.

[DSC98] P Diaconis and L Saloff-Coste. What do we know about the Metropolis algorithm? \textit{Journal of Computer and System Sciences}, 57:20–36, 1998.

[Fil91] James Allen Fill. Eigenvalue Bounds on Convergence to Stationarity for Nonreversible Markov Chains, with an Application to the Exclusion Process. \textit{The Annals of Applied Probability}, 1(1):62–87, February 1991.

[GM00] Charles J Geyer and Antonietta Mira. On non-reversible Markov chains. In \textit{Monte Carlo methods (Toronto, ON, 1998)}, volume 26 of \textit{Fields Inst. Commun.}, pages 95–110. Amer. Math. Soc., Providence, RI, 2000.

[Has70] WK Hastings. Monte Carlo sampling methods using Markov chains and their applications. \textit{Biometrika}, 57(1):97–109, 1970.

[LPW09] DA Levin, Yuval Peres, and EL Wilmer. \textit{Markov chains and mixing times}. 2009.

[Mat92] R Mathias. Matrices with positive definite Hermitian part: Inequalities and linear systems. \textit{SIAM journal on matrix analysis and applications}, pages 1–17, 1992.

[MRR+53] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. \textit{The Journal of Chemical Physics}, 21(6):1087, 1953.

[MT93] SSP Meyn and RL Tweedie. \textit{Markov chains and stochastic stability}. 1993.

[Nea04] Radford M. Neal. Improving Asymptotic Variance of MCMC Estimators: Non-reversible Chains are Better. Technical report, No. 0406, Department of Statistics, University of Toronto, July 2004.

[RT96] Gareth O Roberts and Richard L Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. \textit{Bernoulli}, 2(4):pp. 341–363, 1996.

[SGS10] Yi Sun, Faustino Gomez, and Juergen Schmidhuber. Improving the Asymptotic Performance of Markov Chain Monte-Carlo by Inserting Vortices. In J Lafferty, C K I Williams, J Shawe-Taylor, R S Zemel, and A Culotta, editors, \textit{Advances in Neural Information Processing Systems 23}, pages 2235–2243. 2010.
[ST10] Hidemaro Suwa and Synge Todo. Markov Chain Monte Carlo Method without Detailed Balance. *Physical review letters*, 120603(September):1–4, 2010.

[TCV11] Konstantin S. Turitsyn, Michael Chertkov, and Marija Vucelja. Irreversible Monte Carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240(4-5):410–414, February 2011.

[Wil99] EL Wilmer. *Exact rates of convergence for some simple non-reversible Markov chains*. PhD thesis, Harvard University, 1999.
Figure 2: Comparison of the two approaches to non-reversible Metropolis-Hastings and classical Metropolis-Hastings for different choices of steepness (β) and number of modes (M). The spectral gap (a) and mixing time (b) are displayed for Metropolis-Hastings (c = 0, γ = 0) and for optimized non-reversible Metropolis-Hastings. For ‘Approach (i)’, close to optimal values of $c_1$ and $\gamma$ are obtained by optimization over $50 \times 50$ discrete values. For ‘Approach (ii)’, a close to optimal value of $c_2$ is obtained by optimization over 50 discrete values. These optimizations are performed with respect to spectral gap (left) and with respect to mixing time (right). The $M = 0$ case, i.e. a flat distribution, is identical to the $\beta = 0$ case.
Figure 3: Dependence of spectral gap and mixing time on $n$. Here the number of modes $M = 2$. Results are displayed for $n = 5, 9, 17, 33, 65, 129, 257$. 