Quantifying the Speed-Up from Non-Reversibility in MCMC Tempering Algorithms

Gareth O. Roberts and Jeffrey S. Rosenthal
University of Warwick University of Toronto

(January, 2025)

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms are extremely important for sampling from complicated high-dimensional densities, particularly in Bayesian Statistics (see e.g. [7] and the many references therein). Traditional MCMC algorithms like the Metropolis-Hastings algorithm [15, 12] are reversible. However, in recent years there has been a trend towards using versions which introduce "momentum" and hence are non-reversible in some sense, in an effort to avoid diffusive behaviour [16, 17, 8, 6, 4].

Many of the most challenging problems in sampling complex distributions come from multi-modality. In this context, the most successful approaches have been simulated and parallel tempering algorithms. These algorithms add auxiliary temperature variables to improve mixing between modes [27, 14, 9]. Parallel tempering (which proceeds with a particle at each of a collection of auxiliary temperatures) can be implemented by alternating even and odd index temperature swap proposals. The resulting algorithm is non-reversible (despite being constructed from reversible components) and can create an effect of momentum for each particle as it moves around the temperature space, thus increasing efficiency [18, 28, 25, 5, 26]. This idea was also combined with efficient parallel implementation to create general-purpose software [24].

In this paper, we provide a theoretical investigation of the extent to which such non-reversibility can improve the efficiency of tempering MCMC algorithms. In a certain diffusion-limit context, under strong assumptions, we prove that an optimally-scaled non-reversible MCMC sampler is indeed more efficient than the corresponding optimally-scaled reversible version, but the speed-up is only a modest 42%. This suggests that non-reversible MCMC is indeed worthwhile, but cannot hope to convert intractable algorithms into tractable ones.

To demonstrate this, we first study (Section 2) a simple Markov chain that can help model the reversible and non-reversible behaviour of tempering algorithms. We prove (Theorem 1) that even a non-reversible-style version of this chain still exhibits diffusive behaviour at appropriate scaling. We then consider (Section 3) rescaling space by a factor of ℓ, and

describe certain "optimal" ℓ values. We then apply (Section 4) this reasoning to tempering MCMC algorithms, and prove under the theoretical framework of [1, 21] that the reversible and non-reversible versions have different efficiency curves (Theorem 4) and optimal scaling values (Theorem 5), leading to the 42% improvement under optimality (Corollary 6).

2 A Double-Birth-Death Markov Chain

To study the effects of momentum on tempering, we first digress to study a simple double-birth-death Markov chain, which may be of independent interest.

Consider the following discrete-time countable-state-space Markov chain, which can be viewed as an infinite-size generalisation of the simple finite example studied in [8, 10].

This Markov chain has state space equal to $\mathbf{Z} \times \{+, -\}$, and transition probabilities given by $\mathbf{P}((i,+),(i+1,+)) = A$, $\mathbf{P}((i,+),(i-1,+)) = B$, $\mathbf{P}((i,+),(i,-)) = C$, $\mathbf{P}((i,-),(i-1,-)) = A$, $\mathbf{P}((i,-),(i+1,-)) = B$, $\mathbf{P}((i,-),(i,+)) = C$, where $A + B + C = 1$ are non-negative constants with $C > 0$. (See Figure 1.)

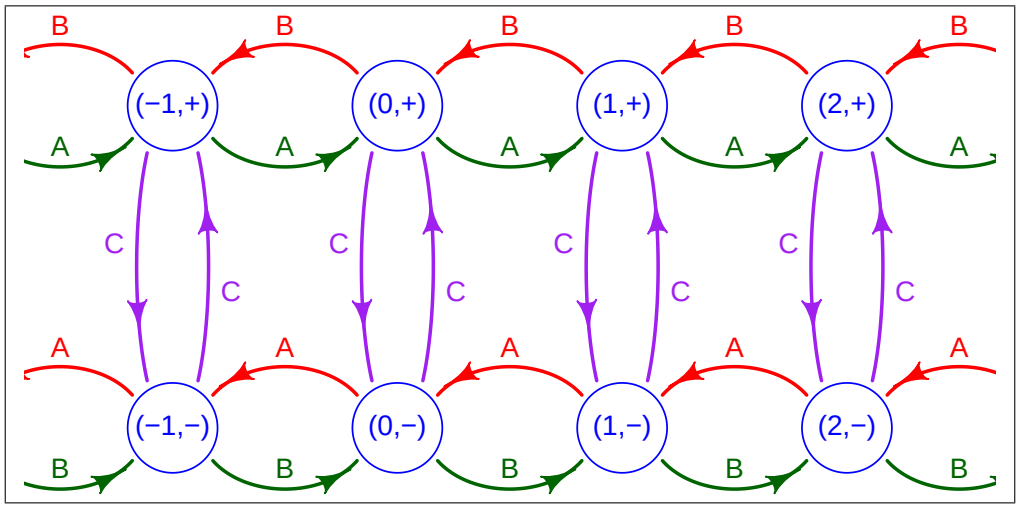

Figure 1: Diagram of the double-birth-death Markov chain.

This chain can be viewed as a "lifting" of a symmetric walk on $\mathbf{Z}$. That is, if states $(i,+)$ and $(i,-)$ are combined into a single state i for each $i \in \mathbf{Z}$, with the chain equally likely to be at $(i,+)$ or $(i,-)$, then this combined process is itself a Markov chain which has the symmetric transition probabilities $P(i,i+1) = P(i,i-1) = (A+B)/2$ and $P(i,i) = C$. However, the full non-combined chain has non-symmetric transitions whenever $A \neq B$.

Of course, if $A = B$, then this chain becomes a symmetric walk on both $\mathbf{Z} \times \{+\}$ and $\mathbf{Z} \times \{-\}$. By contrast, if $A > B$, then it has a positive bias on $\mathbf{Z} \times \{+\}$, and a negative bias on $\mathbf{Z} \times \{-\}$. Indeed, if $B = 0$, then it moves only positively on $\mathbf{Z} \times \{+\}$, and only negatively on $\mathbf{Z} \times \{-\}$.

Despite the non-reversible-seeming nature of this simple Markov chain, the following result (proved in Section 6) gives a diffusion limit of a rescaled version, with a full Functional Central Limit Theorem (i.e., Donsker's Invariance Principle). For notation, let the state of this chain at time n be given by (X_n, Y_n) where X_n is the horizontal integer and Y_n is the vertical $\pm$ value. Then we have:

Theorem 1. *Let $\{(X_n, Y_n)\}_{n=0}^{\infty}$ follow the Markov chain of Figure 1. Let $Z_{M,\cdot}$ be the random process defined by $Z_{M,t} := \frac{1}{\sqrt{M}} X_{\lfloor Mt \rfloor}$ for $t \geq 0$. Then as $M \to \infty$, the process $Z_{M,\cdot}$ converges weakly to Brownian motion with zero drift and with volatility parameter given by*

$$v = [(A - B)^2/C] + (A + B)\,. \tag{1}$$

In particular, for each fixed $t > 0$, as $M \to \infty$, the random variable $Z_{M,t} := \frac{1}{\sqrt{M}} X_{Mt}$ converges weakly to the $N(0, vt)$ distribution.

In the special case where $A = B$, this volatility becomes

$$v \;=\; [(A - B)^2/C] + (A + B) \;=\; [0] + (2A) \;=\; 2A\,. \tag{2}$$

In particular, if $C \searrow 0$ while $A = B \nearrow \frac{1}{2}$, then $v \to 1$ as for standard Brownian motion, exactly as it should. Or, in the special case where $B = 0$, this volatility becomes

$$v \;=\; [(A - B)^2/C] + (A + B) \;=\; [A^2/(1 - A)] + A \;=\; A/(1 - A)\,. \tag{3}$$

We note that the intention of such a chain, at least when $A \gg B$, is to provide a "momentum" whereby the chain moves to the right along the top row for long periods of time, and then to the left along the bottom row for long periods of time, thus sweeping and exploring large regions more efficiently than in the diffusive symmetric $A = B$ case. That is indeed the case, over modest time intervals. However, the chief observation of Theorem 1 is that over larger time intervals, the chain will reverse direction sufficiently that it will still exhibit diffusive behaviour, just on a larger time scale. By rescaling the chain appropriately, the diffusive behaviour can still be identified and quantified, and hence directly compared to the symmetric case, as we discuss below.

3 Rescaling Space and Transition Probabilities

The study of MCMC algorithms includes scaling questions, regarding how large their step sizes should be [19, 1]. For the Markov chain in Figure 1, this corresponds to expanding space by a constant factor of $\ell > 0$, i.e. regarding the adjacent points as being a distance ℓ apart rather than having unit distance.

In this context, the transition probabilities $A = \mathcal{A}(\ell)$ and $B = \mathcal{B}(\ell)$ and $C = \mathcal{C}(\ell)$ also become functions of ℓ (still summing to 1 for each ℓ). The value of v in Theorem 1 then becomes a corresponding function of ℓ, too, i.e.

$$v \;=\; v(\ell) \;=\; [(\mathcal{A}(\ell) - \mathcal{B}(\ell))^2/\mathcal{C}(\ell)] + (\mathcal{A}(\ell) + \mathcal{B}(\ell))\,.$$

Since distance is itself scaled by a factor of ℓ, it follows that the effective volatility of the rescaled process is now proportional to $\ell^2\, v(\ell)$, with $v(\ell)$ as above.

Now, it is known [21, Theorem 1] that limiting diffusions are most efficient in terms of minimising Monte Carlo variance when their volatility is largest. Hence, to make a MCMC algorithm most efficient, we need to maximise that effective volatility, henceforth referrred to as the efficiency function $\mathrm{eff}(\ell) := \ell^2\, v(\ell)$.

Of course, this maximisation depends on the functional dependence of $\mathcal{A}(\ell)$ and $\mathcal{B}(\ell)$, i.e. how the transition probabilities are affected by the spacing ℓ. However, there will often exist an optimal value $\ell_* > 0$ which maximises $\mathrm{eff}(\ell)$. For example, we have the following results.

Proposition 2. *If $\mathcal{A}(\ell)$ is positive, log-concave, C^1, and non-increasing, then there exists a unique $\ell_* > 0$ which maximises* $\mathrm{eff}(\ell)$.

Proof. Let $f(\ell) = \log \mathcal{A}(\ell)$, so that $\mathrm{eff}(\ell) = \ell^2 \mathcal{A}(\ell) = \ell^2 e^{f(\ell)}$. Then any stationary point of $\mathrm{eff}(\ell)$ must have $\mathrm{eff}'(\ell) = 0$, so that

$$2\ell e^{f(\ell)} + \ell^2 f'(\ell) e^{f(\ell)} \;=\; 0\,.$$

Eliminating the point $\ell = 0$ which is clearly a minimum, we need to satisfy

$$-f'(\ell) = 2/\ell\ .$$

Now, the right-hand side of this equation is strictly decreasing from ∞ to 0, and the left-hand side is non-decreasing from a finite value. Hence, since both functions are continuous, there must exist a unique stationary point $\ell_* > 0$. Since $\mathrm{eff}(\ell)$ is non-negative and $\mathrm{eff}(0) = 0$, it follows that ℓ_* is a maximum as required. □

For example, suppose that $\mathcal{A}(\ell) = 2\Phi(-c\ell/2)$ for the cumulative normal distribution function Φ, it is easy to check that $\mathcal{A}(\ell)$ is log-concave. We will see the relevance of this case in the next section.

Proposition 3. *If $\mathcal{B}(\ell) \equiv 0$, and $\mathcal{A}(\ell)$ is continuous, and $\lim_{\ell \searrow 0} \frac{\log[1-\mathcal{A}(\ell)]}{\log(\ell)} < 2$, and $\lim_{\ell \to \infty} \ell^2 \mathcal{A}(\ell) = 0$, then there exists $\ell_* \in (0, \infty)$ such that* $\mathrm{eff}(\ell)$ *is maximised at $\ell = \ell_*$.*

Proof. The assumptions imply that to first order as $\ell \searrow 0$, $1 - \mathcal{A}(\ell) = \ell^\eta$ for some $\eta < 2$, i.e. $\mathcal{A}(\ell) = 1 - \ell^\eta$. Then $\text{eff}(\ell) = \ell^2 \mathcal{A}(\ell)/(1 - \mathcal{A}(\ell)) = \ell^2 (1 - \ell^\eta)/\ell^\eta = \ell^{2-\eta} - \ell^2$. Since $\eta < 2$, this implies that $\text{eff}(\ell) > 0$ for all small positive ℓ. However, $\lim_{\ell \to \infty} \text{eff}(\ell) = \lim_{\ell \to \infty} \ell^2 \mathcal{A}(\ell)/(1 - \mathcal{A}(\ell)) \leq \lim_{\ell \to \infty} \ell^2 \mathcal{A}(\ell) = 0$ by assumption. Hence, by continuity and the Extreme Value Theorem, $\text{eff}(\ell)$ must take its maximum at some $\ell_* \in (0, \infty)$. □

However, the real value of these rescaling operations is to optimise MCMC algorithms like tempering, as we now discuss.

4 Application to Tempering Algorithms

Tempering algorithms, including simulated tempering [14] and parallel tempering [9], are now widely used to improve MCMC by allowing mixing between modes. They involve specifying a sequence of temperature values which increase from one (corresponding to the original "coldest" distribution) to some fixed large value (corresponding to a "hottest" distribution which facilitates easy mixing between modes).

Typically, we define inverse temperatures $0 \leq \beta_N < \beta_{N-1} < \ldots < \beta_1 < \beta_0 = 1$, and let $\pi_\beta(x) \propto [\pi(x)]^\beta$ be a power of the target density $\pi(x)$. Simulated Tempering (ST) augments the original state space with a one-dimensional component indicating the current inverse temperature level, thus creating a $(d+1)$-dimensional chain with target $\pi(\beta, x) \propto K(\beta)\pi(x)^\beta$, where ideally $K(\beta) = \left[\int_x \pi(x)^\beta \mathrm{d}x\right]^{-1}$ so that β has uniform marginal. By contrast, Parallel Tempering (PT) runs a chain on N copies of the state space, each at a different temperature, with target $\pi_N(x_0, x_1, \ldots, x_N) \propto \pi_{\beta_0}(x_0)\, \pi_{\beta_1}(x_1) \ldots \pi_{\beta_N}(x_N)$. Each algorithm attempts to use the hotter temperatures to help the chain move between modes, and thus better sample the original cold temperature target $\pi(x) = \pi_0(x)$.

For tempering algorithms to be useful, they have to move fairly efficiently between the extreme temperatures. In particular, the rate of temperature round-trips from coldest to hottest to coldest is often a good measure of a tempering algorithm's efficiency [28]. To study this, we use the theoretical framework developed in [1, 21]. This framework considers tempering within a single mode of the target distribution, such that spatial mixing is very easy and can be considered to happen immediately. (This corresponds to the "Efficient Local Exploration" (ELE) assumption in [28].) Furthermore, it assumes the same product i.i.d. structure as for theoretical efficiency study of random-walk Metropolis algorithms as in [19, 20].

It is known for tempering algorithms [1, 21], like for random-walk Metropolis (RWM) algorithms [19, 20], that under these strong assumptions, in the limit as the dimension

$d \to \infty$, proposed moves at scaling ℓ are accepted with asymptotic probability $2\,\Phi(-c\,\ell/2)$ for a specific problem-dependent constant $c > 0$ (here $c = \sqrt{I}$ in the notation of [1, 21]), where $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-u^2/2}\, du$ is the cumulative distribution fuction (cdf) of the standard normal distribution, with inverse function Φ^{-1}.

In the usual reversible version of tempering, the proposed moves would be to increase or decrease the temperature index by 1, with probability 1/2 each. In our context, this implies that $\mathcal{A}(\ell) = \mathcal{B}(\ell) = (1/2)[2\,\Phi(-c\,\ell/2)] = \Phi(-c\,\ell/2)$. Hence, the overall proposal acceptance rate then becomes $\mathrm{acc}(\ell) \equiv \mathcal{A}(\ell) + \mathcal{B}(\ell) = 2\,\Phi(-c\,\ell/2)$, as discussed in [1, 21].

By contrast, the non-reversible momentum version of tempering would always propose to increase the temperature index by 1 on $\mathbf{Z} \times \{+\}$, or decrease by 1 on $\mathbf{Z} \times \{-\}$). This corresponds to $\mathcal{B}(\ell) \equiv 0$ and $\mathrm{acc}(\ell) = \mathcal{A}(\ell) = (1)\,2\,\Phi(-c\,\ell/2) = 2\,\Phi(-c\,\ell/2)$.

We now derive various results about the relationship between efficiency and acceptance rate for the reversible and non-reversible versions of these tempering algorithms under these assumptions, as illustrated in Figure 2 (for the case $c = 1$). Note that we measure relative efficiency here in terms of the volatility of the limiting diffusion as justified by [21, Theorem 1] as discussed above. However, we will see in Section 5 below that simulated temperature round-trip rates do indeed follow these relative efficiency curves very closely.

Theorem 4. *Consider a tempering algorithm under the assumptions of [1, 21] as above. Then in the limit as the dimension $d \to \infty$, the efficiency measure* $\mathrm{eff}(\ell)$ *is related to the acceptance rate* $\mathrm{acc}(\ell)$ *as follows:*
(i) In the reversible case, $\mathrm{eff}(\ell) = \mathrm{acc}(\ell)\frac{4}{c^2}\left[\Phi^{-1}(\mathrm{acc}(\ell)/2)\right]^2$.
(ii) In the non-reversible case, $\mathrm{eff}(\ell) = \frac{\mathrm{acc}(\ell)}{1-\mathrm{acc}(\ell)}\frac{4}{c^2}\left[\Phi^{-1}(\mathrm{acc}(\ell)/2)\right]^2$.

Proof. In the reversible case, by (2), we have $v(\ell) = 2\,A(\ell) = \mathrm{acc}(\ell)$, whence $\mathrm{eff}(\ell) = \ell^2\, v(\ell) = \ell^2\,\mathrm{acc}(\ell)$.

In the non-reversible case, by (3), we have $v(\ell) = \frac{A(\ell)}{1-A(\ell)} = \frac{\mathrm{acc}(\ell)}{1-\mathrm{acc}(\ell)}$, whence $\mathrm{eff}(\ell) = \ell^2\, v(\ell) = \ell^2\,\frac{\mathrm{acc}(\ell)}{1-\mathrm{acc}(\ell)}$.

In either case, we have $\mathrm{acc}(\ell) = 2\,\Phi(-c\,\ell/2)$. Inverting this, $\ell = -\frac{2}{c}\,\Phi^{-1}(\mathrm{acc}(\ell)/2)$. Plugging this formula into the expressions for $\mathrm{eff}(\ell)$, the two formulae follow. □

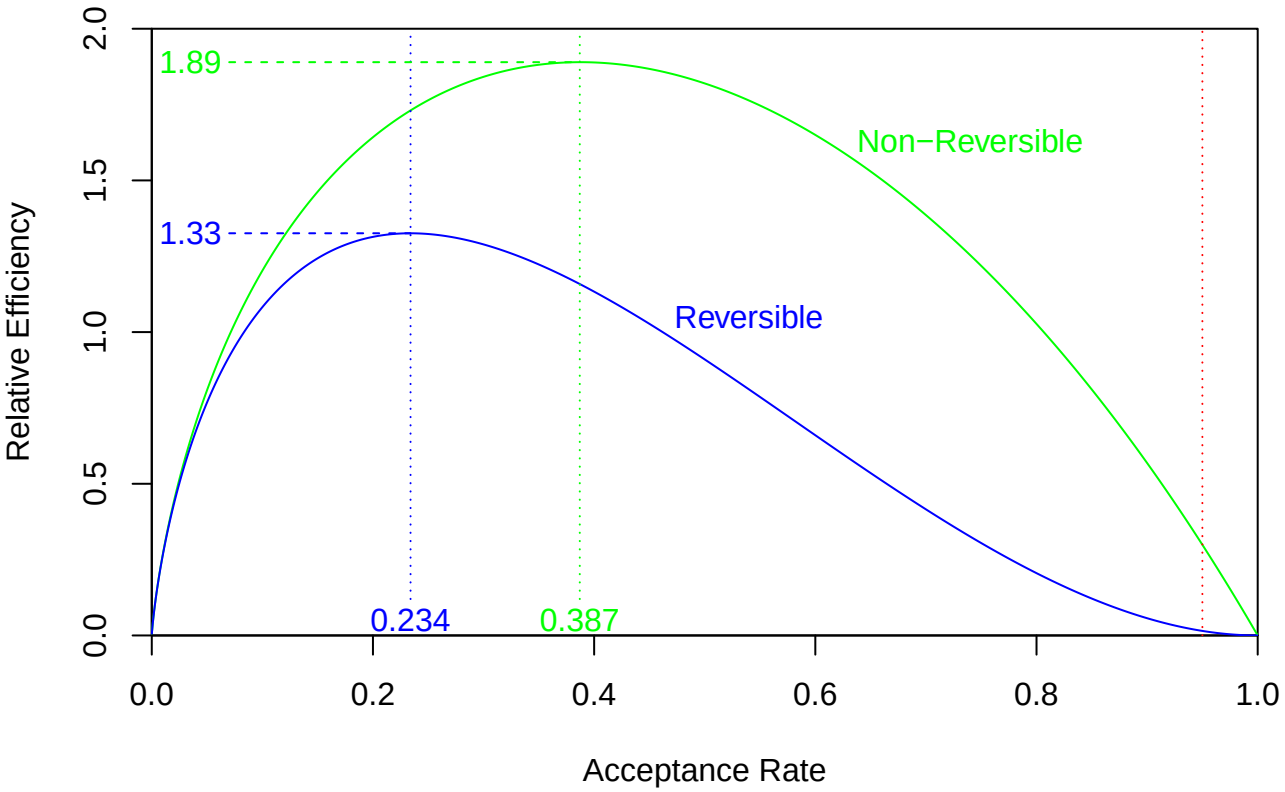

Figure 2: Efficiency curves for non-reversible (top, green) and reversible (bottom, blue) algorithms from Theorem 4 when c=1, with their optimal scaling values (dashed lines) from Theorem 5, including the 1.89/1.33 ratio from Corollary 6, and the infinity-tending ratio towards the right (red) from Corollary 7.

We can then maximise the efficiency curves from Theorem 4, as follows:

Theorem 5. *Consider the tempering algorithm described above, under the strong assumptions from [1, 21]. Then:*

(i) *[1, 21] In the reversible case where* $\mathcal{A}(\ell) = \mathcal{B}(\ell) = \Phi(-c\,\ell/2)$*, the efficiency fuction* $\mathrm{eff}(\ell)$ *is maximised (to three decimal points) by choosing* $\ell = \ell_* \doteq 2.38/c$*, whence* $\mathcal{A}(\ell_*) \doteq 0.117$ *and* $\mathrm{acc}(\ell_*) = 2\,\mathcal{A}(\ell_*) \doteq 0.234$ *and* $\mathrm{eff}(\ell_*) \doteq 1.33/c^2$*. By contrast:*

(ii) *In the non-reversible case where* $\mathcal{B}(\ell) \equiv 0$ *and* $\mathcal{A}(\ell) = 2\,\Phi(-c\,\ell/2)$*, the efficiency function* $\mathrm{eff}(\ell)$ *is maximised by choosing* $\ell = \ell_{**} \doteq 1.73/c$*, whence* $\mathrm{acc}(\ell_{**}) \doteq 0.387$ *and* $\mathrm{eff}(\ell_{**}) \doteq 1.89/c^2$.

Proof. In the reversible case, we need to choose ℓ to maximise

$$\mathrm{eff}(\ell) \;=\; \ell^2 v(\ell) \;=\; \ell^2\, 2\,\mathcal{A}(\ell) \;=\; 2\,\ell^2\,\Phi(-c\,\ell/2)\,.$$

Letting $s = c\ell/2$, this is equivalent to maximising $s^2\,\Phi(-s)$. Numerically, the latter is maxmised at $s = s_* \doteq 1.1906$, corresponding to $\ell = \ell_* = 2s_*/c \doteq 2.3812/c$, whence

$$\mathrm{eff}(\ell_*) \;=\; 2\ell_*^2\,\Phi(-s_*) \;\doteq\; 2(2.3812)^2 c^{-2}\,\Phi(-1.1906) \;\doteq\; 1.3257/c^2\,.$$

Then $\mathcal{A}(\ell_*) = \Phi(-s_*) \doteq \Phi(-1.1906) \doteq 0.117$. The corresponding optimal acceptance rate is thus $\mathcal{A}(\ell_*)+\mathcal{B}(\ell_*) = 2\,\mathcal{A}(\ell_*) \doteq 0.234$, just as for random-walk Metropolis algorithms [19, 20].

By contrast, in the non-reversible case, we need to choose ℓ to maximise

$$\text{eff}(\ell) = \ell^2 v(\ell) = \ell^2\,\mathcal{A}(\ell)/(1-\mathcal{A}(\ell)) = \ell^2\,2\,\Phi(-c\,\ell/2)/[1-2\,\Phi(-c\,\ell/2)]\,.$$

Letting $s = c\ell/2$, this is equivalent to maximising $s^2\,\Phi(-s)/[1-2\Phi(-s)]$. Numerically, the latter is maxmised at $s = s_{**} \doteq 0.8643$, corresponding to $\ell = \ell_{**} = 2s_{**}/c \doteq 1.7285/c$, whence

$$\begin{aligned}\text{eff}(\ell_{**}) \;&=\; \ell_{**}^2\,v(\ell_{**}) \;=\; \ell_{**}^2\,2\,\Phi(-c\,\ell_{**}/2)/[1-2\,\Phi(-c\,\ell_{**}/2)] \\ &\doteq\; (1.7285)^2\,2\,c^{-2}\,\Phi(-0.8643)/[1-\Phi(-0.8643)] \;\doteq\; 1.8896/c^2\,.\end{aligned}$$

Then $\text{acc}(\ell_{**}) = 2\,\mathcal{A}(\ell_{**}) \doteq 2\,\Phi(-c\ell_{**}/2) = 2\,\Phi(-c(1.7285)/2c) = 2\,\Phi(-0.8642) \doteq 0.387$, as claimed. □

Theorem 5 provides some practical guidance when running tempering algorithms. In the reversible case, the temperatures should be spaced so that the acceptance rate of adjacent moves or swaps is approximately 0.234 just like for random-walk Metropolis algorithms [19, 20], as derived in [1, 21]. By contrast, in the non-reversible case, the temperatures should be spaced so that the acceptance rate of adjacent moves or swaps is approximately 0.387.

Now, the ratio of optimal ℓ values is $1.73/2.38 \doteq 0.73$, corresponding to a 27% decrease in proposal scaling standard deviation for the non-reversible versus reversible case. More importantly, the ratio of optimal efficiency functions is $1.89/1.33 \doteq 1.42$, corresponding to a 42% increase in efficiency for the non-reversible versus reversible case. We record this formally as:

Corollary 6. *For the tempering algorithms as above, the maximum efficiency for the non-reversible algorithm is approximately 42% more efficient than the reversible algorithm.*

This corollary indicates that, when scaled with the optimal choice of parameter ℓ, the non-reversible case is indeed more efficient than the reversible case, but not massively so.

We also have:

Corollary 7. *Under the above assumptions, as $\ell \searrow 0$ (corresponding to smaller and smaller temperature spacings), the acceptance $\text{acc}(\ell) \nearrow 1$, and the ratio of efficiency of non-reversible tempering to reversible tempering converges to infinity.*

Proof. By Theorem 4, the ratio of efficiency measures for non-reversible versus reversible tempering is given by

$$\frac{\frac{\text{acc}(\ell)}{1-\text{acc}(\ell)}\frac{4}{c^2}\left[\Phi^{-1}(\text{acc}(\ell)/2)\right]^2}{\text{acc}(\ell)\frac{4}{c^2}\left[\Phi^{-1}(\text{acc}(\ell)/2)\right]^2} \;=\; \frac{1}{1-\text{acc}(\ell)} \;=\; \frac{1}{1-2\,\Phi(-c\,\ell/2)}\,.$$

As $\ell \searrow 0$, we have $\Phi(-c\,\ell/2) \nearrow 1/2$, so $\mathrm{acc}(\ell) \nearrow 1$, and this efficiency ratio converges to $+\infty$, as claimed. □

Corollary 7 indicates that the non-reversible algorithm becomes infinitely more efficient that the reversible algorithm as the proposal scaling becomes very small. This observation corresponds to the result of [28, Theorem 3] that, as the mesh size goes to zero and number of temperatures $N \to \infty$, the non-reversible algorithm achieves a higher-order roundtrip rate of $O(1)$, compared to the reversible algorithm rate of $O(1/N)$. However, when compared at their optimally scaled values, the 42% speed-up of Corollary 6 gives a more accurate measure of the relative improvement of using a non-reversible tempering algorithm.

5 Simulations

To test our theory, we performed a detailed computer simulation of both reversible and non-reversible tempering algorithms on a fixed target in $d = 100$ dimensions. We performed a total of 2×10^{10} tempering iterations on each of 20 different temperature spacing choices over the same temperature range, computed in parallel on the Digital Research Alliance of Canada (DRAC) high-speed compute servers. To conform to the above framework, we conducted the simulation on a single target mode, and counted the total number of round-trips of the temperature from coldest to hottest and back again. We used this count to compute a rate of round-trips per million iterations. The results are shown in Figure 3.

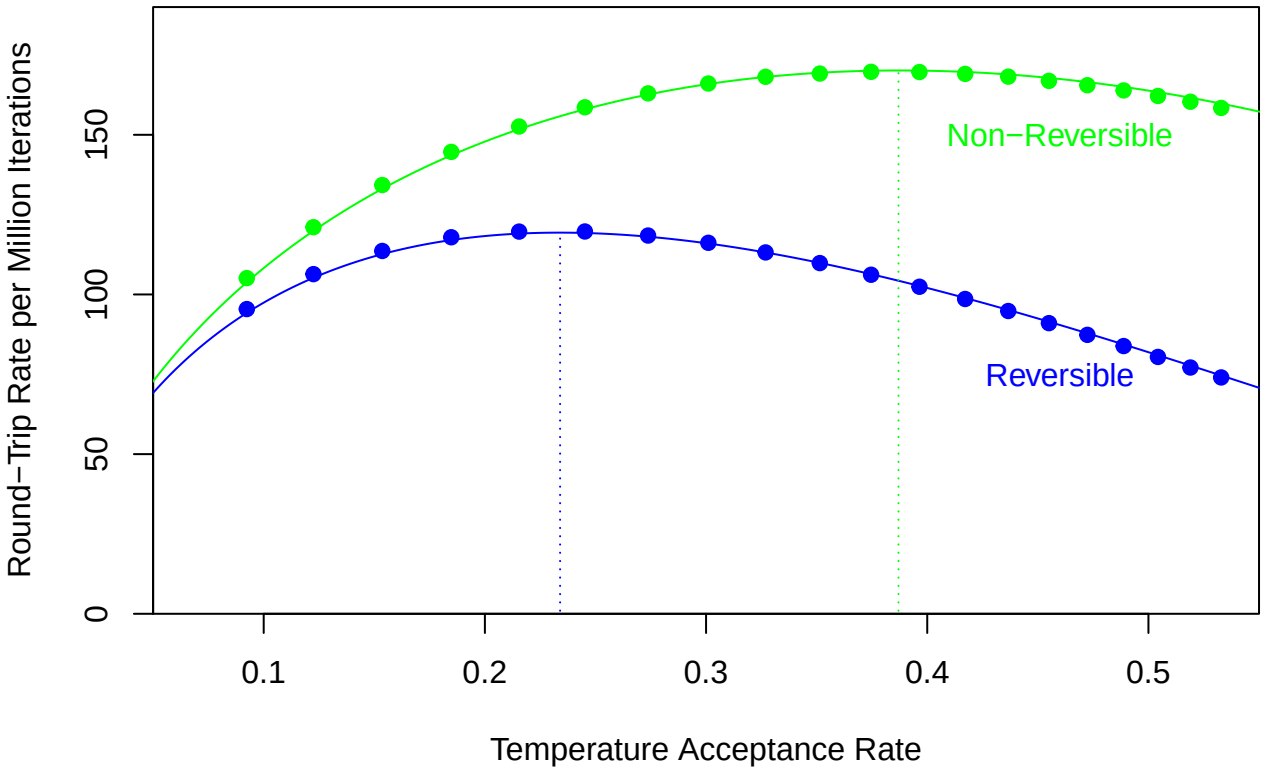

Figure 3: Simulated round-trip rates for non-reversible (top, green) and reversible (bottom, blue) tempering algorithms as a function of the temperature acceptance rate, together with the theoretical relative efficiency curves from Theorem 4 (scaled by a single appropriate constant multiplier), and their optimal scaling values (dashed lines) from Theorem 5, showing excellent agreement.

It is seen from Figure 3 that the simulated round-trip rates show excellent agreement with the theoretical relative efficiency curves from Theorem 4 (when scaled by a single appropriate constant multiplier, to convert the relative efficiency measure into the round-trip rate). This provides convincing evidence that our theoretical results about relative efficiency of different tempering MCMC algorithms, do indeed provide useful information about the practical information of these algorithms to achieve round-trips between the coldest and hottest temperatures.

6 Proof of Theorem 1

Finally, we now prove Theorem 1. For notation, let Geom(C) be the probability distribution with $\mathbf{P}(k) = (1-C)^k C$ for $k = 0, 1, 2, 3, \ldots$, with expectation $(1-C)/C$ and variance $(1-C)/C^2$. And, let μ be the probability distribution on ± 1 with $\mu(1) = A/(1-C)$ and $\mu(-1) = B/(1-C)$. (In particular, if $B = 0$, then $\mu(1) = 1$.)

Lemma 8. *Let $\{(X_n, Y_n)\}_{n=0}^{\infty}$ follow the Markov chain of Figure 1, with initial vertical value $Y_0 = +$. Then there are i.i.d. random variables $G_n, H_n \sim \mathrm{Geom}(C)$, and independent ± 1-valued random variables $E_n, F_n \sim \mu$, such that for all $n \geq 1$,*

$$X_{T_n} - X_{T_{n-1}} \;=\; \sum_{i=1}^{G_n} E_i - \sum_{i=1}^{H_n} F_i \,,$$

where $T_0 = 0$ and $T_n = \sum_{i=1}^{n} (G_i + H_i + 2)$.

Proof. Let G_1 be the last time just before we first move to $\mathbf{Z} \times \{-\}$. Then $G_1 \sim \mathrm{Geom}(C)$.

Next, let E_n be the increment in the x-direction from time $n-1$ to n, conditional on staying in $\mathbf{Z} \times \{+\}$, so $E_n \sim \mu$. Then the total displacement before hitting $\mathbf{Z} \times \{-\}$ is equal to $\sum_{n=1}^{G_1} E_n$. And, this takes G_1 steps, plus 1 step to move to $\mathbf{Z} \times \{-\}$.

Similarly, the total displacement after hitting $\mathbf{Z} \times \{-\}$ but before returning to $\mathbf{Z} \times \{+\}$ is then equal to $-\sum_{n=1}^{H_1} F_n$, for corresponding time $H_1 \sim \mathrm{Geom}(C)$ and independent increments $F_n \sim \mu$. And this takes H_1 steps, plus 1 step to move to $\mathbf{Z} \times \{-\}$.

It follows that $X_{T_1} := X_{G_1+H_1+2} = \sum_{i=1}^{G_1} E_i - \sum_{i=1}^{H_1} F_i$. Continuing in this way, counting the times when the chain moves from $\mathbf{Z} \times \{+\}$ to $\mathbf{Z} \times \{-\}$ and back, the result follows. $\square$

Lemma 9. *With T_n as in Lemma 8, for all $n \geq 1$ we have*

$$(i) \qquad \mathbf{E}[T_n - T_{n-1}] \;=\; 2/C \,,$$

and

$$(ii) \qquad \mathbf{E}[X_{T_n} - X_{T_{n-1}}] \;=\; 0 \,,$$

and

$$(iii) \qquad \mathrm{Var}[X_{T_n} - X_{T_{n-1}}] \;=\; 2[(A-B)^2/C^2] + 2[(1-C)/C] \,.$$

Proof. (i) With G_n and H_n as in Lemma 8, we compute that

$$\mathbf{E}[T_n - T_{n-1}] \;=\; \mathbf{E}[G_n + H_n + 2] \;=\; \mathbf{E}[G_n] + \mathbf{E}[H_n] + 2$$

$$=\; [(1/C) - 1] + [(1/C) - 1] + 2 \;=\; 2/C \,.$$

(ii) The quantities $\sum_{i=1}^{G_n} E_i$ and $\sum_{i=1}^{H_n} F_i$ have the same distribution, and hence the same mean, say m. Therefore,

$$\mathbf{E}[X_{T_n} - X_{T_{n-1}}] \;=\; \mathbf{E}\Big[\sum_{i=1}^{G_n} E_i - \sum_{i=1}^{H_n} F_i\Big] \;=\; m - m \;=\; 0 \,.$$

(iii) Let E_n and F_n be as in Lemma 8, and let $S = \sum_{n=1}^{G_1} E_n$ be the total distance traveled before first hitting $\mathbf{Z} \times \{-\}$.

Then $\mathbf{E}(E_n) = (A-B)/(1-C)$, and $\mathbf{E}(E_n^2) = 1$, so $\mathrm{Var}(E_n) = 1 - [(A-B)^2/(1-C)^2]$. Hence, using the formula for variance of a random-sized sum from Wald's identities,

$$\begin{aligned}
\mathrm{Var}(S) &= \mathbf{E}(G)\,\mathrm{Var}(E_1) + \mathrm{Var}(G)\,\mathbf{E}(E_1)^2 \\
&= [(1-C)/C]\,[1-(A-B)^2/(1-C)^2] + [(1-C)/C^2]\,[(A-B)/(1-C)]^2 \\
&= [A^2 - 2AB + B^2 + C - C^2]\,/\,C^2 \;=\; [(A-B)^2/C^2] + [(1-C)/C] \\
&= [(A-B)^2/C^2] + (A+B)/C\,. \qquad (4)
\end{aligned}$$

Now, (4) is the variance of a single piece, i.e. the part before moving to $\mathbf{Z} \times \{-\}$. Then $X_{T_n} - X_{T_{n-1}}$ is formed by combining two such pieces, of opposite sign. Hence, its variance is twice the value in (4), as claimed. □

Putting these lemmas together, we obtain our diffusion result:

Proof of Theorem 1. In the language of [13], following [22], the Markov chain described by Figure 1 has "regenerative increments" over the times $\{T_n\}$ specified in Lemma 8, with finite increment means and variances. Hence, for fixed $t > 0$, the process $W(M) := X_{\lfloor Mt \rfloor}$ has regenerative increments at times $\{T_n/t\}$. Then, it follows from [13, Theorem 1.4] that as $M \to \infty$ with $t > 0$ fixed, we have $W(M)/\sqrt{M} \equiv X_{\lfloor Mt \rfloor}/\sqrt{M} \to N(0, v)$, where the corresponding volatility parameter v is computed (using Lemma 9) to be:

$$v \;=\; \frac{\mathrm{Var}[X_{T_n} - X_{T_{n-1}}]}{\mathbf{E}[T_n - T_{n-1}]} \;=\; \frac{2[(A-B)^2/C^2] + 2[(A+B)/C]}{2/C} \;=\; [(A-B)^2/C] + (A+B)\,.$$

This proves the claim about convergence to $N(0, vt)$ for fixed $t > 0$. (Strictly speaking, Lemmas 8 and Lemma 9 assumed that the chain begins in a state with $Y_0 = +$, but clearly the initial Y_0 value will not matter in the $M \to \infty$ limit.) The extended claim about convergence of the entire process to Brownian motion then follows from e.g. looking at just the second component in [11, Theorem 5]. This completes the proof. □

Acknowledgements. We thank Nick Tawn, Fernando Zepeda, Hugo Queniat, Saifuddin Syed, Alexandre Bouchard-Côté, Trevor Campbell, Jeffrey Negrea, and Nikola Surjanovic for helpful discussions about tempering issues, the latter four at the Statistical Society of Canada annual conference in Newfoundland in June 2024. We thank Svante Janson for very useful guidance about invariance principles, and Nick Tawn and Fernando Zepeda for helpful conversations about non-reversible algorithms, and David Ledvinka for useful suggestions, and Duncan Murdoch for help with an R question. We thank Daniel Gruner and Ramses van Zon of the Digital Research Alliance of Canada (DRAC) for technical assistance with parallel

high-speed computing. We acknowledge financial support from UKRI grant EP/Y014650/1 as part of the ERC Synergy project OCEAN, by EPSRC grants Bayes for Health (R018561), CoSInES (R034710), PINCODE (EP/X028119/1), and EP/V009478/1, and from NSERC of Canada discovery grant RGPIN-2019-04142.

References

[1] Y. Atchadé, G.O. Roberts, and J.S. Rosenthal (2011), Towards Optimal Scaling of Metropolis-Coupled Markov Chain Monte Carlo. Stat. and Comput. **21(4)**, 555–568.

[2] J. Bierkens (2016), Non-reversible Metropolis-Hastings. Stat. Comput. **26**, 1213–1228.

[3] J. Bierkens and G.O. Roberts (2017), A piecewise deterministic scaling limit of lifted Metropolis-Hastings in the Curie-Weiss model. Ann. Appl. Prob. **27(2)**, 846–882.

[4] J. Bierkens, P. Fearnhead, and G.O. Roberts (2019), The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. Ann. Stat. **47(3)**, 1288–1320.

[5] M. Biron-Lattes, T. Campbell, and A. Bouchard-Cöté (2023), Automatic Regenerative Simulation via Non-Reversible Simulated Tempering. arXiv:2309.05578

[6] A. Bouchard-Côté, S.J. Vollmer, and A. Doucet (2018), The bouncy particle sampler: a nonreversible rejection-free Markov chain Monte Carlo method. J. Amer. Stat. Assoc. **113(522)**, 855–867.

[7] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds. (2011), Handbook of Markov chain Monte Carlo. Chapman & Hall, New York.

[8] P. Diaconis, S. Holmes, and R.M. Neal (2000), Analysis of a non-reversible Markov chain sampler. Ann. Appl. Prob. **10(3)**, 726–752.

[9] C.J. Geyer (1991), Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 156–163.

[10] C.J. Geyer and A. Mira (2000), On non-reversible Markov chains. In N. Madras (ed.), Fields Institute Communications, Volume **26**: Monte Carlo Methods, pp. 93–108. Providence, RI: American Mathematical Society.

[11] A. Gut and S. Janson (1983), The Limiting Behaviour of Certain Stopped Sums and Some Applications. Scand. J. Stat. **10(4)**, 281–292.

[12] W.K. Hastings (1970), Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109.

[13] S. Janson (2023), On a central limit theorem in renewal theory. Preprint. arXiv:2305.13229

[14] E. Marinari and G. Parisi (1992), Simulated tempering: a new Monte Carlo scheme. Europhys. Lett. **19**, 451–458.

[15] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091.

[16] R.M. Neal (1998), Suppressing Random Walks in Markov Chain Monte Carlo Using Ordered Overrelaxation. In: Jordan, M.I. (eds), Learning in Graphical Models, NATO ASI Series **89**. Springer, Dordrecht.

[17] R.M. Neal (2004), Improving Asymptotic Variance of MCMC Estimators: Non-reversible Chains are Better. Tech. Rep. **0406**, Dept. Statistics, University of Toronto.

[18] T. Okabe, M. Kawata, Y. Okamoto, and M. Mikami (2001). Replica-exchange Monte Carlo method for the isobaric-isothermal ensemble. Chemical Physics Letters **335(5–6)**, 435–439.

[19] G.O. Roberts, A. Gelman, and W.R. Gilks (1997), Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Prob. **7**, 110–120.

[20] G.O. Roberts and J.S. Rosenthal (2001), Optimal scaling for various Metropolis-Hastings algorithms. Stat. Sci. **16**, 351–367.

[21] G.O. Roberts and J.S. Rosenthal (2014), Minimising MCMC Variance via Diffusion Limits, with an Application to Simulated Tempering. Ann. Appl. Prob. **24(1)**, 131–149.

[22] R. Serfozo (2000), Basics of Applied Stochastic Processes. Springer-Verlag, Berlin.

[23] Y. Sun, F. Gomez, and J. Schmidhuber (2010), Improving the asymptotic performance of Markov Chain Monte-Carlo by inserting vortices. Adv. Neur. Inform. Proc. Syst. **23**, 2235–2243.

[24] N. Surjanovic, M. Biron-Lattes, P. Tiede, S. Syed, T. Campbell, A. Bouchard-Côté (2023), Pigeons.jl: Distributed sampling from intractable distributions. Preprint. arXiv:2308.09769

[25] N. Surjanovic, S. Syed, A. Bouchard-Cöté, and Trevor Campbell (2022), Parallel Tempering With a Variational Reference. Advances in Neural Information Processing Systems **35**.

[26] N. Surjanovic, S. Syed, A. Bouchard-Cöté, and Trevor Campbell (2024), Uniform Ergodicity of Parallel Tempering with Efficient Local Exploration. arXiv:2405.11384

[27] R.H. Swendsen and J.-S. Wang (1987), Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58**, 86–88.

[28] S. Syed, A. Bouchard-Cöté, G. Deligiannidis, and A. Doucet (2022), Non-reversible parallel tempering: A scalable highly parallel MCMC scheme. J. Roy. Stat. Soc. Ser. B **84(2)**, 321–350.

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