Non-Reversible Parallel Tempering: an Embarrassingly Parallel MCMC Scheme

Saifuddin Syed[*], Alexandre Bouchard-Côté[*], George Deligiannidis[†], Arnaud Doucet[†]

May 7, 2019

Abstract

Parallel tempering (PT) methods are a popular class of Markov chain Monte Carlo schemes used to explore complex high-dimensional probability distributions. These algorithms can be highly effective but their performance is contingent on the selection of a suitable annealing schedule.

In this work, we provide a new perspective on PT algorithms and their tuning, based on two main insights. First, we identify and formalize a sharp divide in the behaviour and performance of reversible versus non-reversible PT methods. Second, we analyze the behaviour of PT algorithms using a novel asymptotic regime in which the number of parallel compute cores goes to infinity. Based on this approach we show that a class of non-reversible PT methods dominates its reversible counterpart and identify distinct scaling limits for the non-reversible and reversible schemes, the former being a piecewise-deterministic Markov process (PDMP) and the latter a diffusion. In particular, we identify a class of non-reversible PT algorithms which is provably scalable to massive parallel implementation, in contrast to reversible PT algorithms, which are known to collapse in the massive parallel regime. We then bring these theoretical tools to bear on the development of novel methodologies. We develop an adaptive non-reversible PT scheme which estimates the event rate of the limiting PDMP and uses this estimated rate to approximate the optimal annealing schedule.

We provide a wide range of numerical examples supporting and extending our theoretical and methodological contributions. Our adaptive non-reversible PT method outperforms experimentally state-of-the-art PT methods in terms of taking less time to adapt, as well as providing better target approximations. Our scheme has no tuning parameters and appears in our simulations robust to violations of the theoretical assumption used to carry out our analysis. The method is implemented in an open source probabilistic programming available at https://github.com/UBC-Stat-ML/blangSDK.

[*]Department of Statistics, University of British Columbia, Canada.
[†]Department of Statistics, University of Oxford, UK.

1 Introduction

Problem formulation. Markov Chain Monte Carlo (MCMC) methods are widely used to approximate expectations with respect to a probability distribution with density $\pi(x)$ known up to a normalizing constant, i.e., $\pi(x) = \gamma(x)/\mathcal{Z}$ where γ can be evaluated pointwise but the normalizing constant $\mathcal{Z}$ is unknown. Approximating such expectations is of central importance in the vast majority of modern Bayesian analysis scenarios as well as frequentist models with complex random effects. In both cases, $\gamma(x)$ can be written as a likelihood $L(x)$ times a prior $\pi_0(x)$, and the distribution of interest is a posterior distribution over a variable $x \in \mathcal{X}$. When the posterior distribution has multiple well-separated modes, highly varying curvature or when one is interested in sampling over combinatorial spaces, standard MCMC algorithms such as Metropolis–Hastings, slice sampling and Hamiltonian Monte Carlo can perform very poorly. This work is motivated by the need for practical methods for these difficult sampling problems and a natural direction is to use multiple cores and/or to distribute the computation.

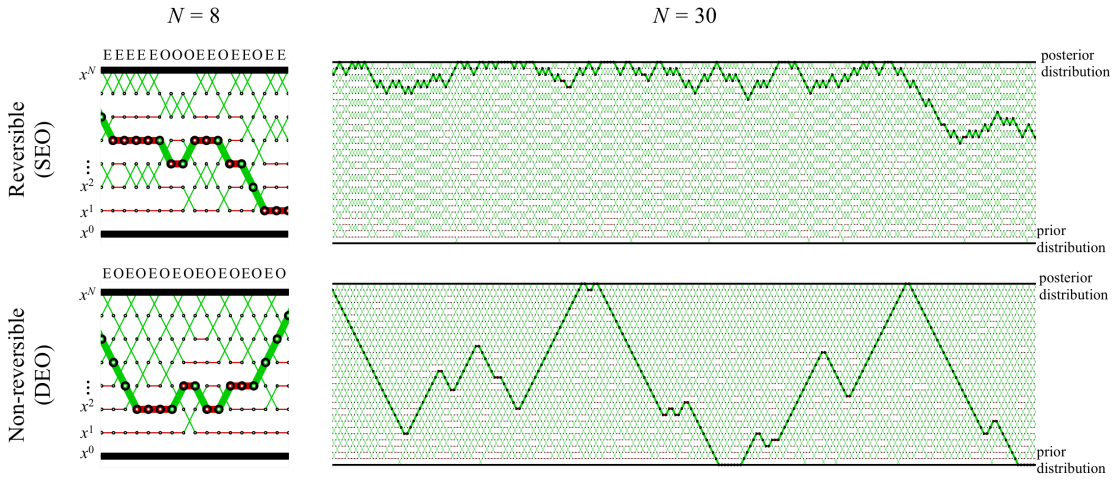


Figure 1: Stochastic Even-Odd swaps (SEO, top row) and Deterministic Even-Odd swaps (DEO, bottom row) for PT, ran with $N = 8$ chains (left column) and $N = 30$ chains (right column) on a Bayesian change-point detection model applied to text message data [DP15]. *Even swap* moves (rows labelled 'E') propose to exchange states at chains with an even index i with the corresponding state in chain $i+1$. Each such swap is independently accepted (green oblique edges) or rejected (red horizontal edges) according to a Metropolis–Hastings step. *Odd swap* moves (rows labelled 'O') propose between odd index i and $i+1$. The only difference between DEO and SEO is the way these moves are composed: in SEO, they are selected at random, while in DEO, the two are deterministically alternated. For both SEO and DEO, exploration kernels are used between each swap round (not shown). This sequence of moves forms $N+1$ annealing parameter trajectories (paths formed by the red and green edges) in the space of annealing parameters $[0, 1]$. We show one such paths B_n in bold as a visual aid. Here for simplicity we use equally spaced annealing parameters. From this figure it is evident that this choice is suboptimal: notice that most swaps between the prior and chain $\beta = 1/N$ are rejected. This is corrected by adaptive tuning (Section 5.4).

Background: Parallel Tempering (PT). One popular approach for multi-core/distributed exploration of complex distributions is Parallel Tempering (PT) which was introduced indepen-

dently in statistics [Gey91] and physics [HN96]; see also [SW86] for an earlier related proposal. Since its inception, PT remains to this day the go-to "workhorse" MCMC method to sample from complex multi-modal target distributions arising in physics, chemistry, biology, statistics, and machine learning; see, e.g., [DLCB14, CRI10, ED05, AFGL05, PS03, CL08]. A recent empirical benchmark [BHH+17] shows PT methods consistently outperform other state-of-the-art sampling methods in practice.

To sample from the distribution of interest π, PT introduces a sequence of auxiliary *tempered* or *annealed* probability distributions with densities $\pi^{(\beta)}(x) \propto L(x)^{\beta}\pi_0(x)$. The auxiliary distributions are parameterized by an *annealing schedule*, which consists of an increasing sequence of annealing parameters $0 = \beta_0 < \beta_1 < \cdots < \beta_N = 1$. This bridge of auxiliary distributions is used to progressively transform samples from the prior ($\beta = 0$), for which it is often possible to obtain independent samples, into samples from the posterior distribution ($\beta = 1$), for which only poorly mixing MCMC kernels may be available.

More precisely PT algorithms are based on Markov chains in which the states are $(N+1)$-tuples, $\mathbf{x} = (x^0, x^1, x^2, \ldots, x^N) \in \mathcal{X}^{N+1}$. The augmented MCMC sampler is designed so that its stationary distribution is given by $\boldsymbol{\pi}(\mathbf{x}) = \prod_{i=0}^{N} \pi^{(\beta_i)}(x^i)$. At each iteration n, PT proceeds by applying in parallel $N+1$ MCMC kernels targeting $\pi^{(\beta_i)}$ for $i = 0, ..., N$. We call these model-specific kernels the *exploration kernels*. The chains closer to the prior chain (i.e. those with annealing parameter β close to zero) can traverse regions of low probability mass under π while the chain at $\beta = 1$ ensures that asymptotically we obtain samples from the distribution of interest. Frequent communication between the chains at the two ends of the spectrum is therefore critical for good performance, and achieved by proposing to swap the states of chains at adjacent annealing parameters. These proposals are accepted or rejected according to a Metropolis mechanism inducing a random permutation of the $N+1$ components of $\mathbf{x}$.

Background: Tuning PT. The effectiveness of PT is determined by how quickly the swapping scheme can transfer information from the prior chain to the posterior chain. There have been many proposals made to improve this information transfer by adjusting the annealing schedule; see, e.g., [KK05, ARR11, MMV13] or adaptively reducing annealing parameters; see, e.g., [LM16]. These proposals are useful but do not address a crucial limitation of PT, illustrated in the top row of Figure 1: in standard PT algorithms, each annealing parameter trajectory (shown in bold in Figure 1 and formally defined in Section 2.3) exhibits a diffusive behaviour, hence we can expect that when N is large it takes roughly $O(N^2)$ swap attempts for a state at $\beta_0 = 0$ to reach $\beta_N = 1$ [DHN00]. The user thus faces a trade-off. If N is too large, the acceptance probabilities of the swap moves are high but it takes a time of order $O(N^2)$ for a state at $\beta = 0$ to reach $\beta = 1$. If N is too low, the acceptance probabilities of swap moves deteriorate resulting in poor mixing between the different chains. Informally, even in a multi-core or distributed setting, for N large, the $O(N)$ gains in being able to harness more cores do not offset the $O(N^2)$ cost of the diffusion (see Figures 4, and Section 3.4 where we formalize this argument). As a consequence,

the general consensus is that the temperatures should be chosen to allow for about a 20–40% acceptance rate to maximize the square jump distance travelled per swap in the space of annealing parameters $[0, 1]$ [KK05, LDMT09, ARR11]. Previous work has shown that adding more chains past this threshold actually deteriorates the performance of PT and there have even been attempts to adaptability reduce the number of additional chains [LM16]. This is a lost opportunity, since PT is otherwise suitable to implementation on multi-core or distributed architectures.

Overview of our contributions. The literature on optimal PT tuning strategies has so far implicitly assumed that the algorithm was reversible and/or serial. Our first contribution is a rigorous, non-asymptotic result showing that a popular non-reversible PT algorithm introduced in the physics literature, Deterministic Even-Odd swap (DEO) [OKOM01], is guaranteed to outperform its reversible counterpart, which we call Stochastic Even-Odd swap (SEO); see Figure 1 for an informal introduction to DEO and SEO. This result holds under an *efficient local exploration* condition that we argue is a reasonable model for scenarios of practical interest. The notion of optimality we analyze, the *round trip rate*, is closely aligned to the running time of practical PT algorithms; see Section 3.

Our second contribution is the asymptotic analysis of the round trip rate in which the number of parallel chains and cores is taken to infinity (Sections 4-5). This novel asymptotic regime is highly relevant to modern computational architectures such as GPUs and distributed computing, and yields several additional results both theoretical and practical:

1. In particular, we show in Section 4 that in the non-reversible regime (DEO) for challenging sampling problems one should use at least as many chains as the number of cores available. This contrasts with the reversible algorithm SEO, where adding more chains, even in a multi-core setup, is eventually detrimental. In other words, for this non-reversible PT algorithm, the optimal tuning recommendations are qualitatively different compared to reversible PT algorithms.

2. While adding more parallel cores to the task improves the performance of non-reversible PT, we show formally that there is a diminishing return in doing so for large N. We quantify this diminishing return using both non-asymptotic bounds as well as an asymptotic analysis letting both the dimension of the problem and the number of chains go to infinity (Section 4.1 and 4.2 respectively).

3. In Section 5 we analyze optimal annealing schedules using our high parallelism asymptotics. We then develop a novel adaptive scheme (Procedure 3), which is both experimentally effective and simple to implement.

Our third contribution is a novel analysis of the scaling limit for the annealing parameter trajectories as the number of parallel chains goes to infinity (Section 6). We show that non-reversible PT scales weakly to a Piecewise Deterministic Markov Process (PDMP) under realistic

conditions, contrasting with the diffusive limit we obtain for reversible PT. This offers intuition explaining the fundamental differences observed between the round trip rates for non-reversible and reversible PT as discussed in Section 3.4 and 4.2. The rate parameter of the limiting PDMP is intimately connected to our adaptive scheme and provides more intuition on its behaviour.

Finally in Section 7, we present a variety of experiments to validate and extend our theoretical analysis. We compare the performance of our non-reversible scheme with other state-of-the-art adaptive PT methods. We also provide empirical evidence that our adaptive scheme is robust to situations where a simplifying assumption used to carry out our theoretical analysis is violated.

Literature review. Previous theoretical studies analyzed the asymptotic behaviour of standard PT based on a target consisting of a product of independent components of increasing dimension [ARR11], or an increased swap frequency relative to a continuous time sampling process [DLPD12]. We instead let the number of cores available in a massively parallel setup go to infinity. One advantage of our approach is that, in contrast to these previous analyses, we do not need to make assumptions on the structure of neither the target distribution (such as [ARR11] where they assume the target distribution is a product of d independent and identical distributions and d is large) nor the exploration kernels (such as [DLPD12], where the exploration kernel is assumed to be driven by a class of stochastic differential equations).

The DEO algorithm was proposed in [OKOM01]. This algorithm was presumably devised on algorithmic grounds (it performs the maximum number of swap attempts in parallel) since no theoretical justification was provided and the non-reversibility of the scheme was not mentioned. The arguments given in [LDMT09] to explain the superiority of DEO communication over various PT algorithms rely on an erroneous assumption, namely a diffusive scaling limit. We show in this work that the scaling limit of non-reversible PT is actually not diffusive as the number of parallel chains goes to infinity. In particular, [LDMT09] still recommends to stop adding chains after a target acceptance rate is achieved.

Another related PT algorithm is the Lifted Parallel Tempering algorithm (LPT), described in [Wu17]; see [SH16] for a closely related idea developed in the context of simulated tempering, and also [SBN13] for an earlier attempt to build a non-reversible PT scheme which was later shown *not* to be invariant with respect to the distribution of interest [ZC14]. These strategies are based on a common recipe to design non-reversible sampling algorithms, which consists in expanding the state space to include a "lifting" parameter that allows for a more systematic exploration of the state space [CLP99, DHN00, TCV11, Vuc16]. We will show here that both LPT and DEO are actually closely related in that the marginal behaviour of individual chains under DEO is in fact distributionally equivalent to the one LPT chain. In a multi-core/distributed context, the DEO scheme therefore dominates LPT by having up to $N/2$ swaps per iteration whereas LPT only performs one.

2 Setup and notation

2.1 Annealed distributions

Henceforth we will assume that the probability distributions π and π_0 on $\mathcal{X}$ admit strictly positive densities with respect to a common dominating measure $\mathrm{d}x$. We will also denote these densities somewhat abusively by π and π_0. It will be useful to define $V_0(x) = -\log \pi_0(x)$ and $V(x) = -\log L(x)$. Using this notation, the *annealed distribution* at an annealing parameter β is given by

$$\pi^{(\beta)}(x) = \frac{L(x)^\beta \pi_0(x)}{\mathcal{Z}(\beta)} = \frac{e^{-\beta V(x) - V_0(x)}}{\mathcal{Z}(\beta)}, \tag{1}$$

where $\mathcal{Z}(\beta) = \int_{\mathcal{X}} L(x)^\beta \pi_0(x) \mathrm{d}x$ is the corresponding normalizing constant.

We denote an *annealing schedule* by $0 = \beta_0 < \beta_1 < \cdots < \beta_N = 1$, and in our asymptotic analysis we will view it as a partition $\mathcal{P} = \{\beta_0, \ldots, \beta_N\}$ of $[0, 1]$ with mesh-size $\|\mathcal{P}\| = \sup_i \{\beta_i - \beta_{i-1}\}$. Given an annealing schedule $\mathcal{P}$ we define $\boldsymbol{\pi}(\mathbf{x})$, the joint distribution on the augmented space $\mathcal{X}^{N+1}$, by

$$\boldsymbol{\pi}(\mathbf{x}) = \prod_{i=0}^{N} \pi^{(\beta_i)}(x^i). \tag{2}$$

2.2 Parallel tempering

In this section, we define formally the Markov kernels corresponding to the reversible (SEO) and non-reversible (DEO) PT algorithms described informally in the introduction and in Figure 1. We provide pseudo-code for the overall algorithm in Procedure 1.

The two phases of Parallel Tempering. For both SEO and DEO, the overall Markov kernel $\mathbf{K}_n^{\mathrm{PT}}$ describing the algorithm is obtained by the composition of a exploration kernel $\mathbf{K}^{\mathrm{expl}}$ and a communication kernel $\mathbf{K}_n^{\mathrm{comm}}$,

$$\mathbf{K}_n^{\mathrm{PT}} = \mathbf{K}_n^{\mathrm{comm}} \mathbf{K}^{\mathrm{expl}}, \tag{3}$$

where $\mathbf{K}_n^{\mathrm{comm}} \mathbf{K}^{\mathrm{expl}}$ denotes the alternation of $\mathbf{K}^{\mathrm{expl}}$ followed by $\mathbf{K}_n^{\mathrm{comm}}$, i.e. for any two transition kernels $\mathbf{K}_1$ and $\mathbf{K}_2$, $(\mathbf{K}_1 \mathbf{K}_2)(\mathbf{x}, A) = \int \mathbf{K}_1(\mathbf{x}, \mathrm{d}\mathbf{x}') \mathbf{K}_2(\mathbf{x}', A)$. The difference between SEO and DEO is in the communication phase, namely $\mathbf{K}_n^{\mathrm{comm}} = \mathbf{K}^{\mathrm{SEO}}$ in the former case and $\mathbf{K}_n^{\mathrm{comm}} = \mathbf{K}_n^{\mathrm{DEO}}$ in the latter. Both communication kernels are detailed further.

The exploration kernels. These are defined in the same way for both SEO and DEO. They are also model specific, so we assume we are given one $\pi^{(\beta_i)}$-invariant kernel $K^{(\beta_i)}$ for each annealing parameter $\beta_0, \beta_1, \ldots, \beta_N$. These can be based on Hamiltonian Monte Carlo, Metropolis–Hastings, Gibbs Sampling, Slice Sampling, etc. The exploration kernel of the prior chain can often be taken to be π_0, i.e. $K^{(0)}(x, A_0) = \pi_0(A_0)$. We construct the overall exploration kernel by applying the

annealing parameter specific kernels to each component independently from each other:

$$\mathbf{K}^{\text{expl}}(\mathbf{x}, A_0 \times A_1 \times \dots A_N) = \prod_{i=0}^{N} K^{(\beta_i)}(x^i, A_i). \tag{4}$$

Swap kernels. Before defining the communication scheme, it will be useful to first construct its fundamental building block, the swap kernel $\mathbf{K}^{(i,j)}$. A swap kernel is a Metropolis–Hastings move with a deterministic proposal which consists of permuting two coordinates of a state vector. The proposed state is denoted

$$\mathbf{x}^{(i,j)} = (x^0, x^1, \dots, x^{i-1}, x^j, x^{i+1}, \dots, x^{j-1}, x^i, x^{j+1}, \dots, \dots, x^N). \tag{5}$$

The Metropolis–Hastings acceptance ratio of this proposal is given by

$$\alpha^{(i,j)}(\mathbf{x}) = \min\left\{1, \frac{\boldsymbol{\pi}\left(\mathbf{x}^{(i,j)}\right)}{\boldsymbol{\pi}(\mathbf{x})}\right\} \tag{6}$$

$$= \exp\left(\min\{0, (\beta_j - \beta_i)(V(x^j) - V(x^i))\}\right). \tag{7}$$

Let $\mathbf{K}^{(i,j)}$ denote the Metropolis-Hastings kernel corresponding to this update:

$$\mathbf{K}^{(i,j)}(\mathbf{x}, A) = \left(1 - \alpha^{(i,j)}(\mathbf{x})\right) \delta_{\mathbf{x}}(A) + \alpha^{(i,j)}(\mathbf{x}) \delta_{\mathbf{x}^{(i,j)}}(A), \tag{8}$$

where δ_x denotes the Dirac delta.

The Odd and Even kernels. These kernels are maximal groups of swap moves such that members of the group do not interfere with each other. See Figure 1 for an illustration. We first define the even and odd indices:

$$E = \{i : 0 \le i < N, i \text{ is even}\}, \tag{9}$$

$$O = \{i : 0 \le i < N, i \text{ is odd}\}. \tag{10}$$

The corresponding even and odd kernels $\mathbf{K}^{\text{even}}$ and $\mathbf{K}^{\text{odd}}$ are then given by

$$\mathbf{K}^{\text{even}} = \prod_{i \in E} \mathbf{K}^{(i,i+1)}, \quad \mathbf{K}^{\text{odd}} = \prod_{i \in O} \mathbf{K}^{(i,i+1)}. \tag{11}$$

The communication kernel for SEO and DEO. For SEO, the kernel $\mathbf{K}_n^{\text{comm}} = \mathbf{K}^{\text{SEO}}$ is given by a mixture of the odd and even kernels in equal proportion:

$$\mathbf{K}^{\text{SEO}} = \frac{1}{2}\mathbf{K}^{\text{odd}} + \frac{1}{2}\mathbf{K}^{\text{even}}. \tag{12}$$

For DEO, the kernel $\mathbf{K}_n^{\text{comm}} = \mathbf{K}_n^{\text{DEO}}$ is given by a deterministic alternation between odd and even

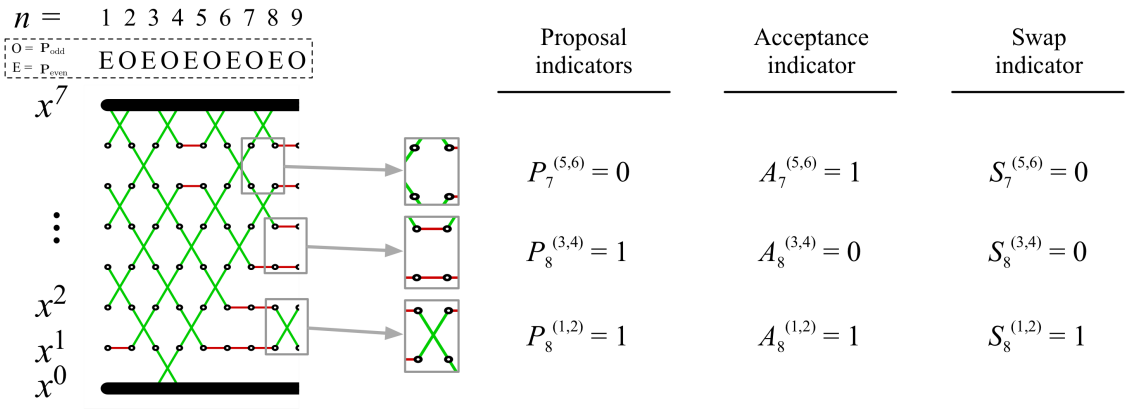


Figure 2: Illustration of the proposal, acceptance and swap indicators on a non-reversible realization.

kernels. This is encoded by the following time heterogeneous kernel

$$\mathbf{K}_n^{\mathrm{DEO}} = \begin{cases} \mathbf{K}^{\mathrm{even}} & \text{if } n \text{ is even,} \\ \mathbf{K}^{\mathrm{odd}} & \text{if } n \text{ is odd.} \end{cases} \tag{13}$$

Proposal and swap indicators. In our theoretical analysis it will be useful to re-express the exploration kernels in the following equivalent fashion. Let

$$\mathbf{P}_n = \left(P_n^{(0,1)}, P_n^{(1,2)}, \dots, P_n^{(N-1,N)} \right), \tag{14}$$

where $P_n^{(i,j)}$ denotes an indicator that a swap is proposed (attempted) between chains i and j at iteration n. In DEO, $\mathbf{P}_n$ is deterministic, i.e. $\mathbf{P}_n = \mathbf{P}_{\mathrm{even}} = (1, 0, 1, \dots)$ for even n and $\mathbf{P}_n = \mathbf{P}_{\mathrm{odd}} = (0, 1, 0, \dots)$ for odd n. In SEO, $\mathbf{P}_n \sim \mathrm{Unif}\{\mathbf{P}_{\mathrm{even}}, \mathbf{P}_{\mathrm{odd}}\}$. We also set $P_n^{(i+1,i)} = P_n^{(i,i+1)}$. To avoid having too many subscripts, we use the same random variables for SEO and DEO but differentiate their behaviour by using two different probability measures $\mathbb{P}_{\mathrm{SEO}}$ and $\mathbb{P}_{\mathrm{DEO}}$ with associated expectation operators $\mathbb{E}_{\mathrm{SEO}}$ and $\mathbb{E}_{\mathrm{DEO}}$. We use $\mathbb{P}$ and $\mathbb{E}$ for statements that hold for both algorithms.

The swap proposals are then defined from the proposal indicators as

$$S_n^{(i,j)} = P_n^{(i,j)} A_n^{(i,j)}, \tag{15}$$

where $A_n^{(i,j)} | \mathbf{X}_n \sim \mathrm{Bern}(\alpha^{(i,j)}(\mathbf{X}_n))$ are acceptance indicator variables (see Figure 2). The equivalence between $\mathbf{K}^{\mathrm{expl}}$ and this representation is given by

$$\mathbf{X}_{n+1} | \mathbf{X}_n \sim \mathbf{K}^{\mathrm{expl}}(\mathbf{X}_n, \cdot) \Longleftrightarrow X_{n+1}^i = \begin{cases} X_n^{(i+1)} & \text{if } S_n^{(i,i+1)} = 1, \\ X_n^{(i-1)} & \text{if } S_n^{(i,i-1)} = 1, \\ X_n^i & \text{otherwise,} \end{cases} \quad \text{for all } i \in 0, ..., N. \tag{16}$$

Permutation augmentation. Another useful construction is to add a permutation to the state space to keep track of the cumulative effect of the swaps. The augmented state space becomes $\mathcal{X}^{N+1} \times \mathrm{Perm}([N])$, where $\mathrm{Perm}([N])$ denotes the group of bijections of $\{0, 1, 2, \ldots, N\}$. A first instance where this construction is useful is in the context of PT algorithms distributed over several compute nodes or machines. In this context a key implementation point is that instead of having pairs of machines exchanging states when a swap is accepted (which could be detrimental due to network latency and lower throughput), the machines should exchange annealing parameters. If $0, 1, 2, \ldots, N$ are indices for $N+1$ machines, then the permutation $\sigma_n \in \mathrm{Perm}([N])$ at iteration n encodes that machine i is currently responsible for annealing parameter $\beta_{\sigma_n(i)}$. Formally, the swap kernel in the augmented space is defined as:

$$\mathbf{K}^{(i,j)}(\bar{\mathbf{x}}, A \times \{\sigma'\}) = \left(1 - \alpha^{(i,j)}(\mathbf{x})\right) \delta_{\mathbf{x}}(A) \mathbb{I}[\sigma' = \sigma] + \alpha^{(i,j)}(\mathbf{x}) \delta_{\mathbf{x}^{(i,j)}}(A) \mathbb{I}[\sigma' = (i\ j) \circ \sigma], \tag{17}$$

where $\bar{\mathbf{x}} = (\mathbf{x}, \sigma)$ denotes an augmented state, $(i\ j) \in \mathrm{Perm}([N])$ denotes a transposition (swap) between i and j, and $(i\ j) \circ \sigma$ is the composition of σ followed by the swap $(i\ j)$. We abuse notation here and denote the kernel in the augmented space with the same symbol. The exploration kernel does not cause swaps, so in the permutation-augmented space we set it to

$$\mathbf{K}^{\mathrm{expl}}((\mathbf{x}, \sigma), A_0 \times A_1 \times \ldots A_N \times \{\sigma'\}) = \mathbb{I}[\sigma' = \sigma] \mathbf{K}^{\mathrm{expl}}(\mathbf{x}, A_0 \times A_1 \times \ldots A_N),$$

with a similar abuse of notation.

Invariant distribution. The kernels $\mathbf{K}^{\mathrm{expl}}, \mathbf{K}^{\mathrm{SEO}}, \mathbf{K}_n^{\mathrm{DEO}}$ and hence $\mathbf{K}_n^{\mathrm{PT}}$ are all invariant with respect to

$$\bar{\boldsymbol{\pi}}(\mathbf{x}, \sigma) = \frac{1}{(N+1)!} \boldsymbol{\pi}(\mathbf{x}). \tag{18}$$

See Appendix A for details.

Non-reversibility of DEO. Written as an homogeneous Markov chain, DEO takes the form $\mathbf{K}^{\mathrm{even}} \mathbf{K}^{\mathrm{expl}} \mathbf{K}^{\mathrm{odd}} \mathbf{K}^{\mathrm{expl}}$. If $N > 1$, this kernel is in general non-reversible (it satisfies global balance but not detailed balance). Examples violating detailed balance can be constructed using a uniform likelihood, $L(x) \propto 1$, in which case $\pi = \pi_0$ and swaps are systematically accepted. Non-reversibility will be explored more deeply in Section 3.3.

Reversibility of SEO. Let us assume that the exploration kernel can be decomposed as $\mathbf{K}^{\mathrm{expl}} = \mathbf{K}^{1/2} \mathbf{K}^{1/2}$. This is a reasonable assumption: often $\mathbf{K}^{\mathrm{expl}}$ is itself a composition of n_{expl} exploration passes, we are just assuming here that n_{expl} is even. In this case, it is reasonable to analyse the kernel $\mathbf{K}^{1/2} \mathbf{K}^{\mathrm{SEO}} \mathbf{K}^{1/2}$. Since this kernel is palindromic, if $\mathbf{K}^{1/2}$ is reversible then the palindrome is also reversible. We will refer to the PT algorithm with SEO and DEO communication as *reversible PT* and *non-reversible PT* respectively even when $\mathbf{K}^{\mathrm{expl}}$ is non-reversible.

Procedure 1 Non-reversible PT (number of scans n, annealing schedule $\mathcal{P}$)

$\hat{r}^{(i,i+1)} \leftarrow 0$ for all $i \in \{0, 1, \ldots, N-1\}$ ▷ Swap rejection statistics used in 5.4 to adapt
$\mathbf{x} \leftarrow \mathbf{x}_0$ ▷ Initialize chain
for t **in** 1, 2, …, n **do**
  for k **in** $1, 2, \ldots, n_{\text{expl}}$ **do**
    $\mathbf{x}' \sim \mathbf{K}^{\text{expl}}(\mathbf{x}, \cdot)$ ▷ Exploration phase (embarrassingly parallel)
    $\mathbf{x} \leftarrow \mathbf{x}'$
  if t is even **then** ▷ Non-reversibility inducing alternation
    $S \leftarrow E$ ▷ Equation (9)
  else
    $S \leftarrow O$ ▷ Equation (10)
  for i **in** $0, \ldots, N-1$ **do** ▷ Communication phase (embarrassingly parallel)
    $\alpha \leftarrow \alpha^{(i,i+1)}(\mathbf{x})$
    $\hat{r}^{(i,i+1)} \leftarrow \hat{r}^{(i,i+1)} + (1 - \alpha)$
    $A \sim \text{Bern}(\alpha)$
    if $i \in S$ **and** $A = 1$ **then**
      $(x^i, x^{i+1}) \leftarrow (x^{i+1}, x^i)$ ▷ Equation (7).
  $\mathbf{x}_t \leftarrow \mathbf{x}$
$\hat{r}^{(i,i+1)} \leftarrow \hat{r}^{(i,i+1)}/n$ for all $i \in \{0, 1, \ldots, N-1\}$ ▷ Equation (57)
return $(\mathbf{x}_1, \ldots, \mathbf{x}_n), (\hat{r}^{(0,1)}, \ldots, \hat{r}^{(N-1,N)})$

2.3 Annealing trajectories and the index process

Chains and replicas. We will refer to the sequences X_n^i and $X_n^{\sigma_n(i)}$, as the i-th *chain* and *replica* respectively. The i-th chain tracks the sequence of states with annealing parameter β_i and the replica tracks the sequence of states on machine i.

Annealing trajectories. Closely related to the replica, we define the *annealing trajectory* for index i by $B_n^i = \beta_{\sigma_n(i)}$. As discussed in the last section, index i can be interpreted as a machine in a distributed context. We will use the notation B_n when i is unimportant. The annealing trajectory tracks the sequence of annealing parameters that machine i is responsible of, as a function of the iteration index n. The concept is best understood visually: refer to the bold piecewise linear path in Figure 1. We shall see in Section 3 that annealing trajectories encode the impact of the communication kernel in PT algorithms, and will allow us to illuminate fundamental differences between reversible and non-reversible PT.

Index process. To analyse the annealing trajectory $B_n^i = \beta_{\sigma_n(i)}$, it will be equivalent and easier to study the sequence $I_n^i = \sigma_n(i)$. For the remainder of this section we will introduce an alternative recursive construction to give intuition on the dynamics of I_n^i. This recursion forms the basis for the analysis in the rest of the paper.

We define the *index process* for machine i as $Y_n^i = (I_n^i, \varepsilon_n^i) \in \{0, \ldots, N\} \times \{-1, 1\}$ and use the notation $Y_n = (I_n, \varepsilon_n)$ when i is unimportant. Initialize $I_0 = i$ and $\varepsilon_0 = 1$ if $P_0^{(i,i+1)} = 1$ and

$\varepsilon_0 = -1$ otherwise. For $n > 0$, we have

$$I_{n+1} = \begin{cases} I_n + \varepsilon_n & \text{if } S_n^{(I_n, I_n+\varepsilon_n)} = 1, \\ I_n & \text{otherwise,} \end{cases} \tag{19}$$

and,

$$\varepsilon_{n+1} = \begin{cases} 1 & \text{if } P_n^{(I_{n+1}, I_{n+1}+1)} = 1, \\ -1 & \text{otherwise.} \end{cases} \tag{20}$$

The variables ε_n represent the direction B_n proposed at iteration n. For SEO communication, the variables ε_n are independent and identically distributed and equal to 1 or -1 with equal probability, and consequentially the annealing trajectories exhibit a random walk behaviour. In contrast for DEO communication, we have $\varepsilon_{n+1} = \varepsilon_n$ so long as the proposal involving replica i was accepted and $\varepsilon_{n+1} = -\varepsilon_n$ otherwise. Therefore annealing trajectories for non-reversible PT have a persistence in one direction and only change when a swap involving replica i is rejected or if the boundary is reached. The qualitative differences between the two regimes can be seen in Figure 1.

3 Non-asymptotic analysis of PT

In this section, we motivate a formal notion of computational efficiency for SEO and DEO, the round trip rate, denoted τ_{SEO} and τ_{SEO} for the two algorithms, and provide conditions under which non-reversible DEO is guaranteed to perform better than its reversible counterpart, $\tau_{\text{SEO}} \leq \tau_{\text{DEO}}$.

3.1 Model of compute time

We start with a definition of what we model as one unit of compute time: throughout the paper, we assume a massively parallel computational setup, and hence that sampling once from each of the kernels $\mathbf{K}^{\text{expl}}$, $\mathbf{K}_n^{\text{DEO}}$ and $\mathbf{K}^{\text{SEO}}$ takes one unit of time, independently of the number of chains $N + 1$.

This assumption is realistic in both GPU and parallel computing scenarios, since the communication cost in PT only involves pairs of neighbours, and moreover does not increase with the dimensionality of the problem (as explained when we introduced the permutation augmentation in Section 2.2). In particular, all simulations considered in this work involve at most an order of tens to hundreds of chains (see for example Fig 16 for an example with up to 640 chains), so they are within reach of current commodity hardware: for example GPUs used in modern scientific applications often have roughly 1,000—5,000 cores as of 2019.

We also assume that the number of MCMC iterations will still dominate the number of parallel core available, i.e. $n \gg N$. This is reasonable since the focus of this paper is in challenging sampling problems.

Empirical studies on multi-core and distributed implementation of PT are numerous [ADHR04, MB12, FFT+14]. However, despite its practical relevance, we are not aware of previous theoretical

work investigating this computational model for PT. From now on, all analysis and recommendations in this paper assume a parallelized or distributed setup.

3.2 Performance metrics for PT methods

The standard notion of computational efficiency of MCMC schemes is the effective sample size (ESS) per compute time [Fle08]. However, for PT methods, since the ESS per compute time depends on the details of the exploration kernels $K^{(\beta_i)}$, alternatives have been developed in the literature. These alternative metrics allow us to give a representative analysis of PT as a "meta-algorithm" without reference to the specifics of the exploration kernels. In this section we motivate and describe the *round trip rate*, one such PT performance metric popular in the PT literature [KTHT06,LDMT09]. The notion of round trip rate seems a priori somewhat disconnected to ESS per compute time, so we first introduce a more intuitive notion, the *restart rate*, and then show that the restart and round trip rates are essentially equivalent.

Tempered restarts. Our definition of a tempered restart is motivated by situations where the prior chain ($\beta = 0$) provides one independent sample at each iteration. In this context, notice that each sample from the prior chain will either "succeed" in getting propagated to the posterior chain ($\beta = 1$), or "fail" and come back to the prior chain. The number of tempered restarts $\mathcal{T}_n$ is defined as the number of distinct independent samples generated by the prior chain which are successfully propagated, via communication and exploration steps, to the posterior chain during the first n iterations. This notion of optimality is not the full picture since intermediate chains also perform exploration, but nonetheless captures the essence of difficult multi-modal problems where only an independent initialization combined with successive exploration and communication steps can reach distinct modes in reasonable computational time. Informally, a tempered restart can be thought of as a sampling equivalent to what is known in optimization as a random restart. We define the restart rate as $\tau_{\text{restart}} = \lim_{n\to\infty} \mathbb{E}[\mathcal{T}_n]/n$. Note also that each tempered restart is carried by one of the $N+1$ annealing trajectories, so we can decompose the tempered restarts as $\mathcal{T}_n = \mathcal{T}_n^0 + \mathcal{T}_n^1 + \cdots + \mathcal{T}_n^N$.

Round trips. Next, we say a *round trip* has occurred for replica i when the annealing trajectory for replica i successfully increases from $\beta = 0$ to $\beta = 1$ and back to $\beta = 0$. Formally, we recursively define $T^i_{\downarrow,0} = \inf\{n : (I^i_n, \varepsilon^i_n) = (0, -1)\}$ and for $k \geq 1$,

$$T^i_{\uparrow,k} = \inf\{n > T^i_{\downarrow,k-1} : (I^i_n, \varepsilon^i_n) = (N, 1)\}, \tag{21}$$

$$T^i_{\downarrow,k} = \inf\{n > T^i_{\uparrow,k} : (I^i_n, \varepsilon^i_n) = (0, -1)\}. \tag{22}$$

We say the k-th round trip for replica i occurred at iteration $T^i_{\downarrow,k}$. Let $\mathcal{R}^i_n$ denote the number of round trips for replica i in the first n iterations, and $\mathcal{R}_n = \mathcal{R}^0_n + \mathcal{R}^1_n + \cdots + \mathcal{R}^N_n$ be the total number of round trips. We define the round trip rate as $\tau_{\text{round trip}} = \lim_{n\to\infty} \mathbb{E}[\mathcal{R}_n]/n$.

This metric is commonly used in the literature to compare the effectiveness of various PT algorithms [KTHT06, LDMT09].[1]

Equivalence: each restart, except possibly for the last one, coincides with a round trip in one of the annealing trajectories. Hence, $\mathcal{T}_n^i \leq \mathcal{R}_n^i \leq \mathcal{T}_n^i + 1$, so $\mathcal{T}_n \leq \mathcal{R}_n \leq \mathcal{T}_n + N$, and $\tau = \tau_{\text{round trip}} = \tau_{\text{restart}}$.

Alternative PT performance metrics. Another performance metric used in the PT literature is the *expected square jump distance* (ESJD) [ARR11, KK05], defined as

$$\text{ESJD} = \mathbb{E}\left[(\beta_{I+1} - \beta_I)^2 \, \alpha^{(I,I+1)}(\mathbf{X})\right], \tag{23}$$

where $I \sim \text{Unif}\{0, 1, 2, \ldots, N\}$ and $\mathbf{X} \sim \boldsymbol{\pi}$. While this criterion is useful within the context of reversible PT for selecting the optimal number of parallel chains, it is too coarse to compare reversible against non-reversible PT methods. Indeed, for any given annealing schedule, the EJSD for DOE and SEO are identical. More generally, the metric is less directly aligned to the quantity practitioners care about, which is the restart rate.

The work of [CS11] proposes to use the relaxation time of the process I_n. However, in our context this ignores the special structure of the chain at $\beta = 0$, which is an independent sequence of random variables distributed according to π_0.

3.3 Index process as a Markov chain

The analysis of the round trip times is in general intractable because the index process Y_n is not Markovian. This is because simulating a transition depends on the acceptance indicators $A_n^{(i,i+1)}$ (see Equation (15)), the distributions of which themselves depend on the full state configuration $\mathbf{X}$. If we further assume that the sequence $\mathbf{X}_n$ is stationary and the exploration kernel is "locally efficient," as defined below, we obtain that the index process Y_n is actually Markovian, and this will allow us to analytically compute round trip rates for both SEO and DEO communication schemes. We formally outline these assumptions below.

Stationarity. We assume $\mathbf{X}_0 \sim \boldsymbol{\pi}$ and thus $\mathbf{X}_n \sim \boldsymbol{\pi}$ for all n as the kernel $\mathbf{K}_n^{\text{PT}}$ is $\boldsymbol{\pi}$-invariant. An important observation that follows from assuming the stationarity regime is that the marginal behaviour of the communication scheme only depends on the distribution of the state $\mathbf{X}_n$ via $N+1$ univariate distributions, namely the $N+1$ distributions of the chain-specific energies $V^{(i)} = V\left(X^{(i)}\right)$, $i \in \{0, 1, 2, \ldots, N\}$. To see why, note that if $\mathbf{X}_n \sim \boldsymbol{\pi}$, then by the definition of the stationary distribution, and Equation (7), the random variables $V^{(i)}$ are independent, and

$$\alpha^{(i,i+1)}(\mathbf{X}) = \exp\left(\min\left\{0, (\beta_{i+1} - \beta_i)\left(V^{(i+1)} - V^{(i)}\right)\right\}\right). \tag{24}$$

Remarkably, this observation allows us to build a theoretical analysis of PT which makes no as-

[1] In [KTHT06, LDMT09] the round trip rate per annealing trajectory was optimized, i.e. $\tau_{\text{round trip}}/(N+1)$.

sumption on the nature of the state space $\mathcal{X}$. In contrast, previous work such as [ARR11] assumed a product space $\mathcal{X} = \mathcal{X}_0^d$ for large d.

Efficient Local Exploration (ELE). Let V and V' denote the negative log-likelihood before and after an exploration step for any chain i, $V = V(X)$, $V' = V(X')$ for $X \sim \pi^{(\beta_i)}$, $X'|X \sim K^{(\beta_i)}(X, \cdot)$. The ELE assumption posits that V and V' are independent.

This condition is more reasonable than it may appear at first glance and it is weaker than assuming that X and X' are independent as typically done in the literature [ARR11, RR14]. Consider for example a scenario where we seek to explore the posterior distribution of a mixture model with symmetries induced by label switching. In such cases, being able to design exploration kernels such as V and V' are approximately independent can be understood as being able to efficiently visit a neighbourhood of one of the local maxima. In contrast, being able to sample X' independently from X would defy the need for using PT in the first place.

These two assumptions are assumed to hold throughout the paper. The assumptions are not expected to be exactly satisfied in real problems. However, they provide the foundations of a model for PT algorithms. We validate the predictions made by the model in Section 7.2 and empirically show robustness in performance even when the ELE assumption is violated. We also provide an heuristic argument to explain why our theoretical results appear robust to ELE violations for non-reversible PT. Moreover, the model is used to make algorithm optimization choices such as picking annealing parameters, and even if a slightly suboptimal PT algorithm is used, this PT algorithm still exactly targets the distribution of interest. Previous work on analyzing PT has also made modelling assumptions that are not expected to hold in practice but yield useful guidelines.

Markov transition kernel for the index process. Under ELE, we can express the acceptance indicators as *independent* Bernoulli random variables $A_n^{(i,i+1)} \sim \text{Bern}(s^{(i,i+1)})$. The constant $s^{(i,i+1)}$ is given by the expectation of Equation (24),

$$s^{(i,i+1)} = \mathbb{E}\left[\alpha^{(i,i+1)}(\mathbf{X})\right] = \mathbb{E}\left[\exp\left(\min\left\{0, (\beta_{i+1} - \beta_i)\left(V^{(i+1)} - V^{(i)}\right)\right\}\right)\right], \tag{25}$$

where the expectation is over two independent random variables $V^{(i)}, V^{(i+1)}$, satisfying $V^{(i)} \stackrel{d}{=} V(X^{(\beta_i)})$ for $X^{(\beta_i)} \sim \pi^{(\beta_i)}$. From this, we obtain that $Y_n = (I_n, \varepsilon_n)$ is Markovian under ELE by mirroring the construction in Section 2.3.

For SEO, initialize $I_0 = i$ and $\varepsilon_0 \sim \text{Unif}\{-1, 1\}$. Define the Markov transition kernel, $Y_{n+1}|Y_n \sim P_{\text{SEO}}(Y_n, \cdot)$ via chain rule in two steps. In the first step, simulate

$$I_{n+1}|Y_n = (i, \varepsilon) \sim \begin{cases} (i + \varepsilon) \wedge N \vee 0 & \text{with probability } s^{(i,i+\varepsilon)}, \\ i & \text{otherwise,} \end{cases} \tag{26}$$

where the expression "$\wedge N \vee 0$" enforces the annealing parameter boundaries. In the second step, independently sample $\varepsilon_{n+1} \sim \text{Unif}\{-1, +1\}$.

Similarly for DEO, initialize $I_0 = i$ and $\varepsilon_0 = 1$ if i is even and -1 otherwise. Analogous to the SEO construction, we define $Y_{n+1}|Y_n \sim P_{\text{DEO}}(Y_n, \cdot)$ via chain rule in two steps. We first update $I_{n+1}|Y_n = (i, \varepsilon)$ same as (26), but in the the second step we apply the deterministic update,

$$\varepsilon_{n+1} = \begin{cases} \varepsilon & \text{if } I_{n+1} = i + \varepsilon, \\ -\varepsilon & \text{otherwise.} \end{cases} \tag{27}$$

The lifted property of non-reversible PT. By inspection, we have for $y, y' \in \{0, \dots, N\} \times \{-1, 1\}$

$$P_{\text{SEO}}(y, y') = P_{\text{SEO}}(y', y), \tag{28}$$

implying that P_{SEO} is reversible with respect to the uniform stationary distribution. In fact, since ε_n are independent and identically distributed, I_n by itself is a reversible Markov process for SEO. The index process for SEO has been analysed in this context by [NH07], where a master equation for I_n was heuristically assumed to hold. However this approach does not provide a good approximation to the DEO case since, even if one assumes ELE, the process I_n is *not* Markovian in contrast to the index process $Y_n = (I_n, \varepsilon_n)$. However, contrary to the SEO case, the index process does not satisfy the detailed balance condition (28) but the following skew-detailed balance condition,

$$P_{\text{DEO}}(y, y') = P_{\text{DEO}}(R(y'), R(y)), \tag{29}$$

where $R(i, \varepsilon) = (i, -\varepsilon)$. This implies that the index process for DEO falls within the generalized Metropolis–Hastings framework outlined in [LSR10, Wu17], and is non-reversible with respect to the uniform distribution.

Reversibility necessitates that the constructed MCMC chain must be allowed to backtrack its movements, which leads to inefficient exploration of the state space. As a consequence, non-reversibility is typically a favourable property for MCMC chains. One popular approach to design non-reversible MCMC samplers is to enlarge the state space with a "lifting parameter" which breaks reversibility and forces persistency in exploration [CLP99, DHN00].

We can interpret the index process $Y_n = (I_n, \varepsilon_n)$ for DEO communication as a "lifted" version of the index process for SEO with lifting parameter ε_n. In DEO communication, I_n travels in the direction of ε_n and only reverses direction when I_n reaches a boundary or when a swap rejection occurs. This idea was first explored by [Wu17] in the context of parallel tempering.

Consequentially, this "lifted property" built into DEO trajectories explains the qualitatively different behaviour between SEO and DEO. In Section 3.4 we will formally show that DEO annealing trajectories perform round trips in $O(N)$ PT iterations whereas SEO annealing trajectories require instead $O(N^2)$ PT iterations. We will also show in Section 6 that the scaling behaviour of the index process for reversible and non-reversible PT are qualitatively different, in particular for non-reversible PT, the index process is non-diffusive in contrast to its reversible counterpart.

3.4 Non-asymptotic domination of non-reversible PT

Based on our two assumptions, we will be able to compute explicit formulae for the round trip rates PT with DEO and SEO communication (and hence restart rates). Using the fact that the index process is Markovian, we can rewrite the round trip rate via an expected hitting time of I_n, and then provide analytic expressions for the expected hitting times of index process Y_n for both DEO and SEO communication based on its transition probabilities. This yields that non-reversible PT outperforms reversible PT for any annealing schedule.

Computation of the round trip rate. We defined our notion of optimality τ using an asymptotic expression in the number of iteration n. Our first goal is to obtain an analytic and non-asymptotic expression for τ for a given annealing schedule $\mathcal{P}$. As we will show shortly, the choice of schedule $\mathcal{P}$ enters in the said analytic expression in terms of a *schedule inefficiency* defined as a sum of rejection odds:

$$E(\mathcal{P}) = \sum_{i=1}^{N} \frac{r^{(i-1,i)}}{1 - r^{(i-1,i)}}, \tag{30}$$

where $r^{(i-1,i)} = 1 - s^{(i-1,i)}$ is the probability of rejecting a swap.

To achieve our first goal, we note that for each $i = 0, \dots, N$, $\mathcal{R}_n^i$ is delayed renewal processes with inter-arrival times $T_k^i = T_{\downarrow,k}^i - T_{\downarrow,k-1}^i$ for $k \geq 1$ and $i = 0, \dots, N$. In particular, T_k^i are independent and identically distributed with common distribution T. The key renewal theorem then implies

$$\tau = \sum_{i=0}^{N} \lim_{n\to\infty} \frac{\mathbb{E}[\mathcal{R}_n^i]}{n} = \frac{N+1}{\mathbb{E}[T]}. \tag{31}$$

The following proposition gives us analytical expressions for the expected round trip times of PT with SEO and DEO communication respectively in terms of the schedule inefficiency. The proof can be found in Appendix B.

Proposition 3.1. *For any annealing schedule* $\mathcal{P} = \{\beta_0, \dots, \beta_N\}$,

$$\mathbb{E}_{\text{SEO}}[T] = 2(N+1)N + 2(N+1)E(\mathcal{P}), \tag{32}$$

$$\mathbb{E}_{\text{DEO}}[T] = 2(N+1) + 2(N+1)E(\mathcal{P}). \tag{33}$$

The first term of (33) is the minimum number of swaps needed for a round trip to occur with no rejections. In contrast, the first term in (32) is the expected number of steps needed for a simple random walk on $\mathcal{P}$ of size $N+1$ to make a round trip. The second term of (32) and (33) represents the expected impact of rejected swaps on the round trip times.

Intuitively, $\mathbb{E}[T]$ can be interpreted as the expected number of scans required before the first replica achieves a round trip. Therefore we should Proposition 3.1 implies we need $O(N)$ scans for

non-reversible PT before the first round trip occur. This is in contrast to the $O(N^2)$ scans required for reversible PT.

Corollary 3.2. *For any annealing schedule $\mathcal{P}$ we have*

$$\tau_{\mathrm{SEO}}(\mathcal{P}) := \frac{N+1}{\mathbb{E}_{\mathrm{SEO}}[T]} = \frac{1}{2N + 2E(\mathcal{P})}, \tag{34}$$

$$\tau_{\mathrm{DEO}}(\mathcal{P}) := \frac{N+1}{\mathbb{E}_{\mathrm{DEO}}[T]} = \frac{1}{2 + 2E(\mathcal{P})}, \tag{35}$$

so $\tau_{\mathrm{SEO}}(\mathcal{P}) \leq \tau_{\mathrm{DEO}}(\mathcal{P})$.

4 Asymptotic analysis of PT

While the main result from the previous section ranks the performance of DEO communication relative to SEO communication, it does not provide insight on the absolute performances of these schemes, because of the inefficiency term $E(\mathcal{P})$.

Overview. In this section, we provide asymptotic estimates of $E(\mathcal{P})$ as $\|\mathcal{P}\| \to 0$. The main result in this section is that the round trip rate $\tau_{\mathrm{SEO}}(\mathcal{P})$ of the reversible PT decays to zero. This in contrast to the non-reversible PT, where $\tau_{\mathrm{DEO}}(\mathcal{P})$ *asymptotically increases* (as defined below) to a positive constant $\bar{\tau}$. Moreover, we provide a characterization of $\bar{\tau}$ in terms of a "communication barrier", Λ, measuring the deviance of π from π_0. We show both $\bar{\tau}$ and Λ can be estimated from the MCMC trace in Section 5 and can be used as the basis of schedule adaptation schemes.

4.1 The communication barrier

We begin by analyzing the behaviour of the PT swaps as $\|\mathcal{P}\|$ goes to zero. In order to do so, we define the swap and rejection functions $s, r : [0,1]^2 \to [0,1]$ respectively as,

$$s(\beta, \beta') = \mathbb{E}\left[\exp\left(\min\{0, (\beta - \beta')(V^{(\beta)} - V^{(\beta')})\}\right)\right], \tag{36}$$

$$r(\beta, \beta') = 1 - s(\beta, \beta'), \tag{37}$$

where $V^{(\beta)} \overset{d}{=} V(X^{(\beta)})$ for $X^{(\beta)} \sim \pi^{(\beta)}$ are independent. The quantities $s(\beta, \beta')$ and $r(\beta, \beta')$ are symmetric in their arguments and represent the probability of swapping and rejection occurring between β and β' under the ELE assumption. Note that $s^{(i-1,i)} = 1 - r^{(i-1,i)} = s(\beta_{i-1}, \beta_i)$.

Local communication barrier. To take the limit as $\|\mathcal{P}\| \to 0$, it will be useful to understand the behaviour of $s(\beta, \beta')$ when $\beta \approx \beta'$. The key quantity that drives this asymptotic regime is given by a function $\lambda : [0,1] \to \mathbb{R}_+$ defined as

$$\lambda(\beta) = \frac{1}{2}\mathbb{E}\left[|V_1^{(\beta)} - V_2^{(\beta)}|\right], \tag{38}$$

where $V_1^{(\beta)}, V_2^{(\beta)}$ are independent random variables with common distribution $V^{(\beta)}$. We will use the following estimate for $s(\beta, \beta')$ derived in the context of the design of a different class of tempering models used in the physics literature called incomplete beta function laws [PPC04].

Theorem 4.1. *[PPC04] For $\beta \leq \beta'$, let $\bar{\beta} = \frac{\beta+\beta'}{2}$ and $\delta = \beta' - \beta$. Suppose V^3 is integrable with respect to π_0 and π then we have,*

$$s(\beta, \beta') = 1 - \delta\lambda(\bar{\beta}) + O(\delta^3), \tag{39}$$

$$r(\beta, \beta') = \delta\lambda(\bar{\beta}) + O(\delta^3). \tag{40}$$

Theorem 4.1 shows that λ encodes up to third order the behaviour of s and r as the annealing parameter difference between the chains goes to 0. Since $r(\beta, \beta) = 0$, Theorem 4.1 implies that $\lambda(\beta)$ can be expressed equivalently as the instantaneous rate of rejection of a proposed swap at annealing parameter β,

$$\lambda(\beta) = \lim_{\delta \to 0^+} \frac{r(\beta, \beta + \delta) - r(\beta, \beta)}{\delta}. \tag{41}$$

Note that $r(\beta, \beta')$ is small when $\pi^{(\beta)} \approx \pi^{(\beta')}$, which combined with Theorem 4.1 and (41) implies $\lambda(\beta)$ measures how sensitive $\pi^{(\beta)}$ is to perturbation in β.

Replica with annealing trajectory B_n will have very little difficulty accepting swaps when $\lambda(B_n)$ is small and will suffer from high rejection rates in regions when $\lambda(B_n)$ is large. Since chains communicate only when swaps are successful, $\lambda(\beta)$ measures the difficulty of communication at β.

Global communication barrier. When $\beta < \beta'$, $\delta\lambda(\bar{\beta})$ is the Riemann sum for $\int_\beta^{\beta'} \lambda(b)db$ with a single rectangle. If $\lambda \in C^2([0,1])$, then standard midpoint rule error estimates yield

$$\left| \int_\beta^{\beta'} \lambda(b)\mathrm{d}b - \delta\lambda(\bar{\beta}) \right| \leq \frac{1}{12} \left\| \frac{\mathrm{d}^2\lambda}{\mathrm{d}\beta^2} \right\|_\infty \delta^3. \tag{42}$$

Proposition 4.2 implies that the regularity of λ is controlled by the moments of V with respect to π and π_0.

Proposition 4.2. *If V^k is integrable with respect to π_0 and π, then $\lambda \in C^{k-1}([0,1])$.*

By applying Proposition 4.2, we can substitute (42) into Theorem 4.1, to obtain the following corollary.

Corollary 4.3. *If V^3 is integrable with respect to π and π_0, we have*

$$s(\beta, \beta') = 1 - \int_\beta^{\beta'} \lambda(b)\mathrm{d}b + O(\delta^3), \tag{43}$$

$$r(\beta, \beta') = \int_\beta^{\beta'} \lambda(b)\mathrm{d}b + O(\delta^3). \tag{44}$$

Motivated by Corollary 4.3 we will henceforth assume that V^3 is integrable with respect to π_0 and π and define $\Lambda : [0,1] \to \mathbb{R}_+$ by

$$\Lambda(\beta) = \int_0^\beta \lambda(\beta')\mathrm{d}\beta'. \tag{45}$$

We denote $\Lambda = \Lambda(1)$ as the *global communication barrier*.

Remark 4.4. Notice that $\Lambda \geq 0$ with equality if and only if $\lambda(\beta) = 0$ for all $\beta \in [0,1]$. It can be easily verified from (38) that $\lambda = 0$ if and only if $V^{(\beta)}$ is constant $\pi^{(\beta)}$-a.s. for all $\beta \in [0,1]$ which happens precisely when $\pi_0 = \pi$. So Λ defines a natural symmetric divergence and measures the difficulty of communication between π_0 and π.

High-dimensional scaling of communication barrier. We now determine the asymptotic behaviour of λ and Λ when the dimension of $\mathcal{X}$ is large. To make the analysis tractable, we assume that $\pi_d(x) = \prod_{i=1}^d \pi(x_i)$ as in [ARR11,RR14]. This provides a model for weakly dependent high-dimensional distributions.

The corresponding tempered distributions are thus given by

$$\pi_d^{(\beta)}(x) = \prod_{i=1}^d \pi^{(\beta)}(x_i) \propto \exp\left(-\beta \sum_{i=1}^d V(x_i) - \sum_{i=1}^d V_0(x_i)\right). \tag{46}$$

Let λ_d and Λ_d be the local and global communication barriers for π_d respectively. It follows from Proposition 4.5 that λ_d and Λ_d increase at a rate of $O(d^{1/2})$ as $d \to \infty$.

Proposition 4.5 (High Dimensional Scaling)**.** *Define* $\sigma^2(\beta) = \mathrm{Var}(V^{(\beta)})$*, then for all* $\beta \in [0,1]$*, we have as* $d \to \infty$*,*

$$\lambda_d(\beta) \sim \sqrt{\frac{d}{\pi}}\sigma(\beta) \tag{47}$$

and,

$$\Lambda_d \sim \sqrt{\frac{d}{\pi}}\int_0^1 \sigma(\beta)\mathrm{d}\beta. \tag{48}$$

Remark 4.6. We emphasize that we make only this structural assumption on the state space and distribution for Proposition 4.5. All the other results presented in this work are agnostic to the structure of the $\mathcal{X}$ and π.

4.2 Asymptotic analysis of round trip rate

Suppose $\mathcal{P}_N$ is a sequence of annealing schedules of size $N+1$ such that $\mathcal{P}_N \subset \mathcal{P}_{N+1}$. By Corollary 3.2 we can asymptotically characterize the behaviour of the round trip rate as $\|\mathcal{P}_N\| \to 0$ through

the schedule inefficiency $E(\mathcal{P}_N)$.

Asymptotic increasing sequence. In this section we will use the following two definitions: first, we write $a_n \lesssim b_n$ as $n \to \infty$ if and only if there is c_n such that $a_n \leq c_n$ and $c_n \sim b_n$ as $n \to \infty$. Second, we say a_n is *asymptotically decreasing* (respectively *asymptotically increasing*) if $a_{n+1} \lesssim a_n$ (respectively $a_n \lesssim a_{n+1}$).

Proposition 4.7. *If $\|\mathcal{P}_N\| \to 0$, then $E(\mathcal{P}_N)$ asymptotically decreases to Λ at a rate of $O(\|\mathcal{P}_N\|)$.*

A consequence of Proposition 4.7 and Corollary 3.2, we obtain the following key result.

Corollary 4.8. *Suppose $\|\mathcal{P}_N\| \to 0$ as $N \to \infty$.*

(a) The round trip rate τ_{SEO} goes to zero:

$$\tau_{SEO}(\mathcal{P}_N) \sim \frac{1}{2N} \to 0. \tag{49}$$

(b) The round trip rate τ_{DEO} asymptotically increases at a $O(\|\mathcal{P}_N\|)$ rate to the following upper bound:

$$\tau_{DEO}(\mathcal{P}_N) \to \bar{\tau} = \frac{1}{2 + 2\Lambda} > 0. \tag{50}$$

By Remark 4.4, Λ is large when π_0 deviates significantly from π, therefore we expect a higher round trip rate when the prior is chosen to be a good approximation to the target. Since Λ is problem specific, this identifies a limitation of PT present even in its non-reversible flavour, namely that adding more cores to the task will never be harmful, but does have a diminishing return. The bound $\bar{\tau} = (2+2\Lambda)^{-1}$ could be very small for complex problems. Moreover, it is independent of the choice of annealing schedule, hence this bound cannot be improved by the algorithmic optimizations described in Section 5. Thankfully, the more classical asymptotic perspective in Proposition 4.5 shows that Λ is expected to grow as the square root rate of the dimensionality d in a certain special cases where the state space is a product space, $\mathcal{X}^d$. Hence we expect that weakly dependent high dimensional problems will have a moderate Λ and $\bar{\tau}$ is expected to decrease at a $O(d^{-1/2})$ rate.

5 Tuning non-reversible PT algorithms

Context. So far, in addition to showing the superiority of the non-reversible communication scheme DEO, we have established that in the massively parallel regime, non-reversible PT will benefit from utilizing at least as many cores as available. Moreover, by Equation (50), asymptotically, the choice of annealing schedule $\mathcal{P}$ does not matter as long as its mesh size goes to zero. However, given a finite number of available cores, there are still gains to be made by optimizing the annealing schedule. In this section, we introduce a novel approach to this optimization problem,

which relies on the communication barrier λ.

Section overview. We first show that, under reasonable assumptions, the optimal annealing schedule maximizing the round trip rate is obtained by having a constant rejection rate between chains. This "equi-acceptance" result is not surprising given that other theoretical frameworks and notions of efficiency also obtained recommendations involving equal acceptance rate between chains [ARR11, KTHT06, LDMT09, Kof02, PPC04]. However implementing this equi-acceptance recommendation in practice is non-trivial. Previous work relied on Robbins-Monro schemes [ARR11, MMV13], which introduce several tuning parameters. Our second result in this section is an easy to implement scheme to achieve equi-acceptance, based on the communication barrier λ. The third result in this section is to show that this function $\lambda : [0, 1] \to \mathbb{R}$ can be easily estimated from the MCMC output, hence creating an end-to-end method for non-reversible PT tuning.

Relation to previous work. We reiterate an important difference in the non-reversible PT tuning process compared to previous work. In the existing literature, focusing on reversible and/or serial computation, deciding the number of chains N was done as part of the tuning process. Here, in the context of difficult sampling problem we instead assume that the number of chains is taken to be as large as possible and hence determined by the characteristics of a massively parallel architecture. Given this N, we build an equi-acceptance annealing schedule.

5.1 Optimal round trip rate

In this section we show that for a fixed large number of chains $N > \Lambda$, having equal swap acceptance probabilities maximizes the following optimization program over annealing schedules $\mathcal{P}$:

$$\begin{aligned} &\text{maximize:} && \tau_{\mathrm{DEO}}(\mathcal{P}) \\ &\text{subject to:} && |\mathcal{P}| = N + 1. \end{aligned} \tag{51}$$

To approach this optimization, we first use Corollary 3.2 to rewrite the maximization of the round trip rate $\tau_{\mathrm{DEO}}(\mathcal{P})$ into a minimization of the schedule inefficiency, $E(\mathcal{P})$. Recall that $E(\mathcal{P})$ is defined in Equation (30) as the sum of rejection odds $r^{(i-1,i)}/(1 - r^{(i-1,i)})$ over the pair of chains $(i-1, i)$. Hence, we can rewrite the optimization objective in terms of the variables $r_i = r^{(i-1,i)}$. To get a tractable approximate characterization of the feasible region of $r_1, r_2, \ldots, r_N$, we use Corollary 4.3, which implies that for all schedules $\mathcal{P}$ we have for $r_i = r(\beta_{i-1}, \beta_i)$,

$$\sum_{i=1}^{N} r_i = \Lambda + O\left(N\|\mathcal{P}\|^3\right). \tag{52}$$

Therefore assuming $\|\mathcal{P}\|$ is small enough to ignore the error term in (52), finding $\mathcal{P}_{\text{optimal}}$ is approximately equivalent to solving the following optimization problem:

$$\begin{aligned} \text{minimize:} \quad & \sum_{i=1}^{N} \frac{r_i}{1-r_i} \\ \text{subject to:} \quad & \sum_{i=1}^{N} r_i = \Lambda, \\ & r_i \geq 0. \end{aligned} \tag{53}$$

This can be solved using Lagrange multipliers and leads to a solution where r_i^* is constant in i. We denote this constant by r^*.

5.2 Optimal annealing schedule

The previous section established that for a fixed $N > \Lambda$, we should target an equi-acceptance annealing schedule. However, algorithmically we need to perform the optimization over the actual annealing parameters β_i in order to be able to run the PT simulation. Assuming we know λ for now (and we show how to estimate it in the next section), the idea is that to obtain the optimal schedule $\mathcal{P}_{\text{optimal}} = \{\beta_0^*, \ldots, \beta_N^*\}$, we partition the interval $[0, 1]$ such that the area under the curve λ between successive β_i^* and β_{i+1}^* is constant and equal to r^*.

Computing β_k^* from communication barrier. To formalize this intuition, recall that for the optimal schedule $\mathcal{P}_{\text{optimal}}$ of size $N+1$, we have $r_i = r^*$ for all i which by (52) satisfies,

$$r^* = \frac{\Lambda}{N} + O(\|\mathcal{P}\|^3). \tag{54}$$

By Corollary 4.3 we have for all $i = 0, \ldots, N$,

$$r^* = \int_{\beta_{i-1}^*}^{\beta_i^*} \lambda(\beta) \mathrm{d}\beta + O(\|\mathcal{P}\|^3). \tag{55}$$

If we equate (54) and (55) while ignoring the $O(\|\mathcal{P}\|^3)$ error terms and sum from $i = 0, \ldots, k$ we get,

$$\Lambda(\beta_k^*) \approx \Lambda \frac{k}{N}. \tag{56}$$

The problem of numerically solving Equation (56) for β_k^* is similar to that of finding the k/N quantiles corresponding to a random variable with CDF $F_\lambda(\beta) = \Lambda(\beta)/\Lambda$. This suggests we want to pick $\mathcal{P}_{\text{optimal}}$ with density proportional to λ.

5.3 Estimation of the communication barrier

Computing $\lambda(\beta)$ or $\Lambda(\beta)$ exactly via (38) is in general intractable. In this section, we present a simple Monte Carlo approximation.

Setup: assume we have access to a collection of samples, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, coming from a non-reversible PT scheme based on an arbitrary annealing schedule $\mathcal{P}_N$ of size $N+1$ (see Procedure 1). These samples may come from a short pilot run, or, as described in the next section, from the previous iteration of an adaptive scheme.

Estimation of optimal round trip rate $\bar{\tau}$. Let $\beta_i \in \mathcal{P}$. Under ELE, we have the following Monte Carlo estimates for the rejection rates:

$$\hat{r}^{(i-1,i)} = \frac{1}{n}\sum_{k=1}^{n} \alpha^{(i-1,i)}(\mathbf{X}_k) = r^{(i-1,i)} + O_p(n^{-1/2}). \tag{57}$$

Next, we apply i times Corollary 4.3 on the pairs $(\beta_0, \beta_1), (\beta_1, \beta_2), \ldots, (\beta_{i-1}, \beta_i)$ and sum Equation (44), obtaining:

$$\sum_{j=1}^{i} r^{(j-1,j)} = \sum_{j=1}^{i} \left(\int_{\beta_{j-1}}^{\beta_j} \lambda(b)\mathrm{d}b + O(\|\mathcal{P}\|^3 \right) = \Lambda(\beta_i) + O(N\|\mathcal{P}\|^3). \tag{58}$$

This motivates the following approximation for $\Lambda(\beta_i)$,

$$\hat{\Lambda}(\beta_i) = \sum_{j=1}^{i} \hat{r}^{(j-1,j)}, \tag{59}$$

which assuming ELE has an error of $O_p(\sqrt{N/n} + N\|\mathcal{P}\|^3)$. In particular, we also arrive at a consistent estimator $\hat{\tau}$ for the optimal round trip rate $\bar{\tau}$,

$$\hat{\tau} = \frac{1}{2 + 2\hat{\Lambda}}, \tag{60}$$

where $\hat{\Lambda} = \hat{\Lambda}(1)$. In particular $\hat{\tau}$ allows us to diagnose if a low round trip is due to design choices for PT, or due to π, π_0 itself. We can compare the empirically observed round trip rate against $\hat{\tau}$ to determine how far our implementation deviates from optimal performance.

Estimation of $\Lambda(\beta)$ and $\lambda(\beta)$ via interpolation. Given the estimates $\Lambda(\beta_0), \ldots, \Lambda(\beta_N)$ obtained above, we estimate the function $\Lambda(\beta)$ via interpolation, with the constraint that the interpolated function should be monotone increasing (since $\lambda(\beta) \geq 0$). Specifically, we use the Fritsch-Carlson monotone cubic spline method [FC80]. We denote the monotone interpolation by $\hat{\Lambda}(\beta)$. More sophisticated interpolation methods could be used, for example method taking the Monte Carlo

standard error into account.

While we only use $\Lambda(\beta)$ in our adaptation procedure, it is still useful to estimate $\lambda(\beta)$ for visualization purpose. We do this by taking the derivative of our interpolation, $\hat{\lambda}(\beta) = \hat{\Lambda}'(\beta)$, which is just a piecewise quadratic function.

5.4 Adaptive algorithm

Updating. The ideas described in this section so far are summarized in Procedure 2, which given rejection statistics collected for a schedule provide an updated schedule.

Procedure 2 UpdateSchedule(swap rejection estimates $\{\hat{r}^{(i-1,i)}\}$, previous schedule $\mathcal{P}$)

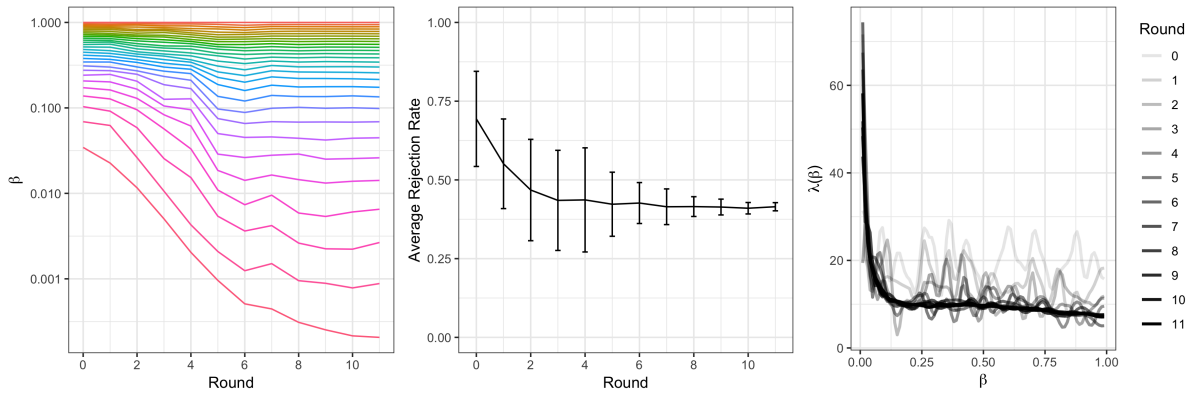
$N \leftarrow |\mathcal{P}| - 1$
For each $\beta_i \in \mathcal{P}$, compute $\hat{\Lambda}(\beta_i)$ ▷ Equation (59)
$S \leftarrow \{(\beta_0, \hat{\Lambda}(\beta_0)), (\beta_1, \hat{\Lambda}(\beta_1)), \ldots, (\beta_N, \hat{\Lambda}(\beta_N))\}$
Compute a monotone increasing interpolation $\hat{\Lambda}(\cdot)$ of the points S ▷ e.g. using [FC80]
$\hat{\Lambda} \leftarrow \hat{\Lambda}(1)$
for k in $1, 2, ..., N-1$ **do**
    Find β_k^* such that $\hat{\Lambda}(\beta_k^*) = \hat{\Lambda}\frac{k}{N}$ using e.g. bisection.
return $\mathcal{P}^* = (0, \beta_1^*, \beta_2^*, \ldots, \beta_{N-1}^*, 1)$

Adapting. Next, we push this idea a bit further in Procedure 3, which iteratively refines the annealing schedule. By construction, the procedure guarantees that the second half of the samples of the chain at $\beta = 1$ provide a consistent estimate of expectations under π (PT algorithms are ergodic under much weaker conditions, such as ergodicity of the exploration kernels). We show in Figure 3 a visualization of the execution of the adaptive algorithm, Procedure 3, on a real dataset.

Procedure 3 is qualitatively very different from existing adaptive PT algorithms such as [ARR11, MMV13, LM16]. We do not suggest a continuous state-dependent adaptation, instead, we recommend using only the second half of the samples produced by Procedure 3, which by construction follow an homogeneous chain. This allows us to circumvent the hurdles that arise in practice when doing continuous adaptation. Procedure 3 experimentally out-performs existing adaptive methods in terms of round trip rates and effective sample size per second, as discussed in Section 7.3.

Procedure 3 Non-reversible PT with adaptation

$N + 1 \leftarrow$ number of cores available
$\mathcal{P} \leftarrow$ initial annealing schedule of size $N + 1$ (e.g. uniform)
$n \leftarrow 2$
for round **in** 1, 2, ..., number of rounds requested **do**
    $\{\hat{r}^{(i-1,i)}\} \leftarrow \text{DEO}(n, \mathcal{P})$ ▷ Procedure 1
    $\mathcal{P} \leftarrow$UpdateSchedule($\{\hat{r}^{(i-1,i)}\}, \mathcal{P}$) ▷ Procedure 2
    $n \leftarrow 2n$ ▷ Rounds use an exponentially increasing number of scans



(a) (b) (c)

Figure 3: Visualization of our adaptive non-reversible PT algorithm ran on a hierarchical Bayesian model applied to the historical failure rates of 5 667 launches for 367 types of rockets with. This was done with $N = 30$ chains and 11 adaptive rounds, the last one consisting of 5 000 scans, the penultimate of 2 500, etc. and estimated $\hat{\Lambda} = 12.03$. (a) Progression of the adaptive annealing schedule (colours index parallel chains, y-axis, the values β_k for each adaptation round, in log scale). (b) Progression of the average empirical rejection rates $\{\hat{r}^{(i-1,i)}\}_{i=1}^{N}$ with their sample standard deviation. Notice the average is fairly consistent, but as the adaptive rounds increase, the rejection rates converge to the average as desired. (c) Progression of the estimated $\hat{\lambda}(\beta)$ evolution with adaption rounds.

6 Scaling limit of annealing trajectories

Suppose Y_n is the index process for annealing schedule $\mathcal{P}_N$ of size $N+1$ and meshsize $\|\mathcal{P}_N\|$ taking values in $\{0, \dots, N\} \times \{-1, 1\}$. Figure 4 suggests that Y_n behaves qualitatively different as N increases for both reversible and non-reversible PT. The goal of this section is investigate these differences and classify the scaling limits for the index process. We will show that such limits exist under the stationary and ELE assumptions specified in Section 3.3. As $\|\mathcal{P}_N\| \to 0$, we will show for reversible PT, the index process weakly converges to a diffusion independent of π, π_0 and sequence of schedules $\mathcal{P}_N$. In contrast, the index process for non-reversible PT does not have a diffusive limit (contrary to [LDMT09]) but rather scales to a Piecewise Deterministic Markov Process (PDMP) controlled by λ, and the choice of the annealing schedule.

Schedule generating function. Suppose $G \in C^1([0,1])$ is an increasing function satisfying $G(0) = 0$ and $G(1) = 1$. We say that G is a *schedule generator* for $\mathcal{P} = \{\beta_0, \dots, \beta_N\}$ if $\mathcal{P} = G(\mathcal{P}_{\text{uniform}})$, or equivalently

$$\beta_i = G\left(\frac{i}{N}\right). \tag{61}$$

We will now assume without loss of generality that the sequence of schedules $\mathcal{P}_N$ are generated by some common G. In particular the mean value theorem implies $\|\mathcal{P}_N\| = O(N^{-1})$ as $N \to \infty$. This is not as strict of a requirement as it seems since most annealing schedules commonly used fall within this framework:

- The uniform schedule $\mathcal{P}_{\text{uniform}} = \{0, 1/N, \ldots, 1\}$ is generated by $G(w) = w$.
- The optimal schedule $\mathcal{P}_{\text{optimal}} = \{\beta_0^*, \ldots, \beta_N^*\}$ derived in Section 5.2 is generated by $G(w) = F_\lambda^{-1}(w)$, where $F_\lambda(\beta) = \Lambda(\beta)/\Lambda$.
- If $\pi_0(x) \propto \pi(x)^\gamma$ for some $\gamma \in (0,1)$, and $L(x) \propto \pi(x)^{1-\gamma}$ then some simple algebraic manipulation shows that $G(w) = \frac{\gamma^{1-w}-\gamma}{1-\gamma}$ corresponds to the geometric schedule commonly used by practitioners.

6.1 Scaled index process

To establish scaling limit for $Y_n = (I_n, \varepsilon_n)$ it will be convenient to work in a continuous time setting. To do this, we suppose the times that PT iterations occur according to a Poisson process $\{M(\cdot)\}$ with mean μ_N. The number of PT iterations that occur by time $t \geq 0$, satisfies $M(t) \sim \text{Poisson}(\mu_N t)$. We define the *scaled index process* by $Z^N(t) = (W^N(t), \varepsilon^N(t))$ where $W^N(t) = I_{M(t)}/N$ and $\varepsilon^N(t) = \varepsilon_{M(t)}$. For convenience, we will denote $\beta_w = G(w)$ and use $z = (w, \varepsilon) \in [0,1] \times \{-1, 1\}$ to be a *scaled index*.

The process Z^N takes values on the discrete set $\mathcal{P}_{\text{uniform}} \times \{-1, 1\}$ and is only well-defined when $Z^N(0) = z_0 \in \mathcal{P}_{\text{uniform}} \times \{-1, 1\}$. To establish convergence it is useful to extend it to a process Z^N which can be initialized at any $z_0 \in [0,1] \times \{-1, 1\}$. Suppose $Z^N(0) = z_0 \in [0,1] \times \{-1, 1\}$, and $T_1, T_2, \ldots$ are the iteration times generated by the Poisson process M. We construct $Z^N(t)$ as follows: define $Z^N(t) = z_n$ for $t \in [T_n, T_{n+1})$ and update $z_{n+1}|z_n$ via a transition kernel which depends on the communication scheme. We determine this transition kernel mirroring the construction from Section 3.4.

Before we do this it will be useful to define the backward and forward shift operators $\Phi_-^N, \Phi_+^N : [0,1] \to [0,1]$ by,

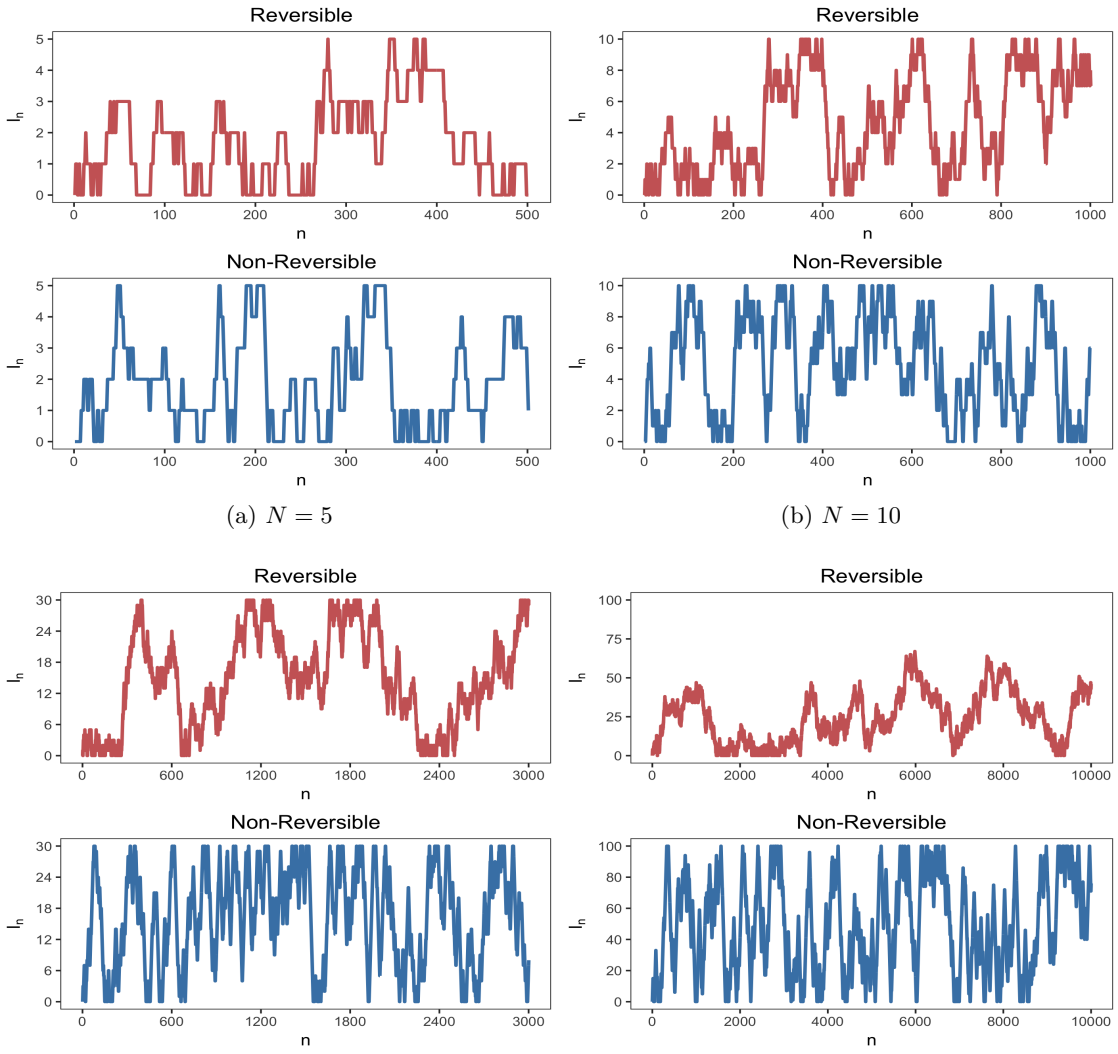
$$\Phi_-^N(w) = \begin{cases} w - \frac{1}{N} & w \in \left[\frac{1}{N}, 1\right], \\ \frac{1}{N} - w & w \in \left[0, \frac{1}{N}\right), \end{cases} \tag{62}$$

and similarly,

$$\Phi_+^N(w) = \begin{cases} w + \frac{1}{N} & w \in \left[0, 1 - \frac{1}{N}\right], \\ 1 - \left(\frac{1}{N} - (1-w)\right) & w \in \left(1 - \frac{1}{N}, 1\right]. \end{cases} \tag{63}$$

Intuitively $\Phi_\varepsilon^N(w)$ represents the location in $[0,1]$ after w moves a distance $\frac{1}{N}$ in the direction of ε with a reflection at 0 and 1.

Scaled index process for reversible PT: Under the SEO communication scheme, if $z_n = (w_n, \varepsilon_n) \in \mathcal{P}_{\text{uniform}} \times \{-1, 1\}$, then we have $w_{n+1} = \Phi_{\varepsilon_n}^N(w_n)$ if a swap successfully occurred and



(a) $N = 5$ (b) $N = 10$

(c) $N = 30$ (d) $N = 100$

Figure 4: Panels (a)-(d) compare sample trajectories of the index process for a Gaussian model with $\Lambda = 5$ and $\mathcal{P}_{\text{optimal}}$ for both reversible and non-reversible PT. We compare the trajectories over a period of $n = 100N$ scans. When $N = 5, 10, 30, 100$ there are $1, 3, 2, 0$ total number of round trips respectively made by the the reversible trajectories. The non-reversible trajectories make in contrast $4, 6, 8, 9$ round trips in the same number of iterations. These simulations are in agreement with the theoretical result in Equation (50): the estimate $\bar{\tau} = (2+2\Lambda)^{-1}$ derived from Section 5.3 suggests we should expect on average of $100\bar{\tau} \approx 8.33$ round trips when N is large for non-reversible PT.

$w_{n+1} = w_n$ otherwise. In both cases, $\varepsilon_{n+1} \sim \text{Unif}\{-1, +1\}$. Since $\Phi_\varepsilon^N(w)$ is not only well-defined for $w \in \mathcal{P}_{\text{uniform}}$ but for $w \in [0,1]$, we naturally extend this construction to any $w \in [0,1]$.

Formally, we generate $(w_{n+1}, \varepsilon_{n+1})$ in two steps. In the first step we simulate,

$$w_{n+1}|w_n, \varepsilon_n \sim \begin{cases} \Phi_{\varepsilon_n}^N(w_n) & \text{with probability } s(\beta_{w_n}, \beta_{\Phi_{\varepsilon_n}^N(w_n)}), \\ w_n & \text{otherwise.} \end{cases} \tag{64}$$

In the second step we simulate $\varepsilon_{n+1} \sim \text{Unif}\{-1, +1\}$. This defines a continuous time Markov pure jump process $W^N \in D(\mathbb{R}_+, [0,1])$[2] with jumps occurring according to an exponential of rate μ_N and is well defined when initialized at any state $w_0 \in [0,1]$.

Scaled index process for Non-reversible PT: Before defining the transition kernel for the scaled index process under DEO communication, it will be convenient to define the propagation function $\Phi^N : [0,1] \times \{-1,1\} \to [0,1] \times \{-1,1\}$ for $z = (w, \varepsilon)$,

$$\Phi^N(z) = \begin{cases} (\Phi_\varepsilon^N(w), \varepsilon) & \text{if } \Phi_\varepsilon^N(w) = w + \frac{\varepsilon}{N}, \\ (\Phi_\varepsilon^N(w), -\varepsilon) & \text{otherwise,} \end{cases} \tag{65}$$

and similarly the rejection function $R : [0,1] \times \{-1,1\} \to [0,1] \times \{-1,1\}$,

$$R(z) = (w, -\varepsilon). \tag{66}$$

Under the DEO scheme, if $z_n = (w_n, \varepsilon_n) \in \mathcal{P}_{\text{uniform}} \times \{-1,1\}$, then we have $z_{n+1} = \Phi^N(z_n)$ when a swap successfully occurs and $z_{n+1} = R(z_n)$ otherwise. Since $\Phi^N(z)$ and $R(z)$ are well-defined for all of $z \in [0,1] \times \{-1,1\}$, we naturally extend this construction to any $z \in [0,1] \times \{-1,1\}$.

Formally, we generate z_{n+1} according to the transition kernel,

$$z_{n+1}|z_n \sim \begin{cases} \Phi^N(z_n) & \text{with probability } s(\beta_{w_n}, \beta_{\Phi_{\varepsilon_n}^N(w_n)}), \\ R(z_n) & \text{otherwise.} \end{cases} \tag{67}$$

This defines a continuous time Markov pure jump process $Z^N \in D(\mathbb{R}_+, [0,1] \times \{-1,1\})$ with jumps occurring according to an exponential of rate μ_N which is well defined when initialized at any $z_0 \in [0,1] \times \{-1,1\}$.

6.2 Scaling limit of scaled index process

We will now characterize the generators of W^N and Z^N and identify their scaling limits as N is taken to infinity by establishing of their infinitesimal generators.

[2]Given a metric space $(\mathcal{S}, d)$, we define $C(\mathbb{R}_+, \mathcal{S})$ and $D(\mathbb{R}_+, \mathcal{S})$ to be set of functions $f : \mathbb{R}_+ \to \mathcal{S}$ that are continuous and càdlàg respectively.

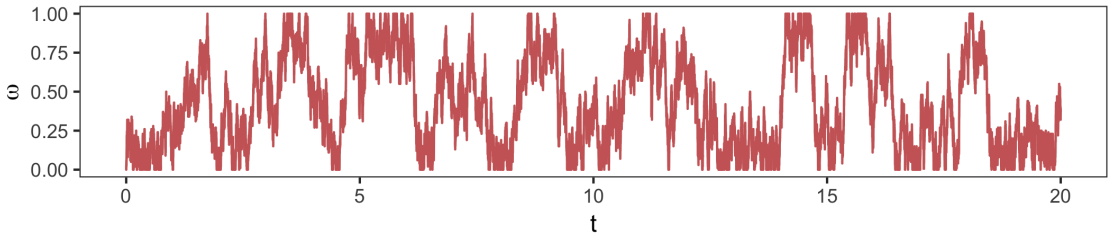


Figure 5: Sample trajectory of W.

Scaling limit for Reversible PT: By Proposition 17.2 in [Kal97] the infinitesimal generator for W^N with SEO communication is

$$\mathcal{L}_{W^N} f(w) = \frac{\mu_N}{2} \sum_{\varepsilon \in \{-,+\}} \left(f(\Phi_\varepsilon^N(w)) - f(w) \right) s(\beta_w, \beta_{\Phi_\varepsilon^N(w)}), \tag{68}$$

where the domain $\mathcal{D}(\mathcal{L}_{W^N})$ is given by the set of functions such that $\mathcal{L}_{W^N} f$ is continuous. Since Φ_+^N, Φ_-^N are continuous, we have $\mathcal{D}(\mathcal{L}_{W^N}) = C([0,1])$.

Define $W \in C(\mathbb{R}_+, [0,1])$ to be the diffusion on $[0,1]$ with generator

$$\mathcal{L}_W f(w) = \frac{1}{2} \frac{\mathrm{d}^2 f}{\mathrm{d}w^2}, \tag{69}$$

where the domain $\mathcal{D}(\mathcal{L}_W)$ consisting of $f \in C^2([0,1])$ such that $f'(0) = f'(1) = 0$. W is a Brownian motion on $[0,1]$ with reflective boundary conditions admitting the uniform distribution Unif$([0,1])$ as stationary distribution.

Theorem 6.1. *Suppose $\mu_N = N^2$ and $W^N(0)$ converges weakly to $W(0)$, then W^N converges weakly to W in $D(\mathbb{R}_+, [0,1])$.*

Theorem 6.1 tells us that for a index process for reversible PT scales to a Brownian motion on $[0,1]$ with reflecting boundary conditions if we speed the scans by factor for $O(N^2)$. Note that this limit W is independent of π_0, π and partition generator G.

Scaling limit for non-reversible PT: Analogously to the reversible case, under DEO communication, the infinitesimal generator for Z^N is

$$\mathcal{L}_{Z^N} f(z) = \mu_N \left(f(\Phi^N(z)) - f(z) \right) s(\beta_w, \beta_{\Phi_\varepsilon^N(w)}) + \mu_N \left(f(R(z)) - f(z) \right) r(\beta_w, \beta_{\Phi_\varepsilon^N(w)}), \tag{70}$$

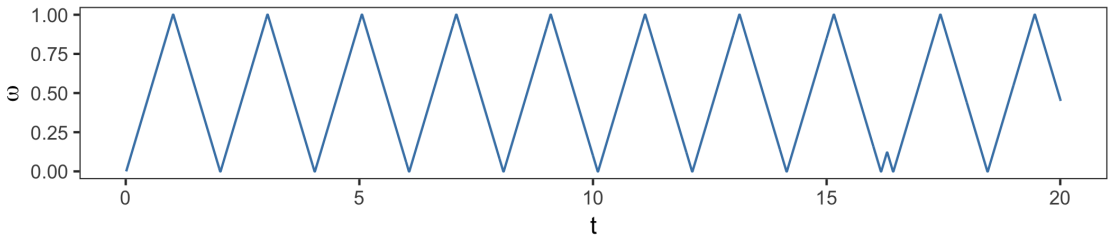
where $z = (w, \varepsilon)$ and $\mathcal{D}(\mathcal{L}_{Z^N})$ is given by the set of functions f such that $\mathcal{L}_{Z^N} f$ is continuous. Since Φ^N has discontinuities at $(\frac{1}{N}, -1)$ and $(1 - \frac{1}{N}, 1)$, we can verify that $f \in \mathcal{D}(\mathcal{L}_{Z^N})$ if and only if $f(w_0, -1) = f(w_0, 1)$ for $w_0 \in 0, 1$.

Define $Z \in C(\mathbb{R}_+, [0,1] \times \{-1,1\})$ to be the PDMP on $[0,1] \times \{-1,1\}$ with generator

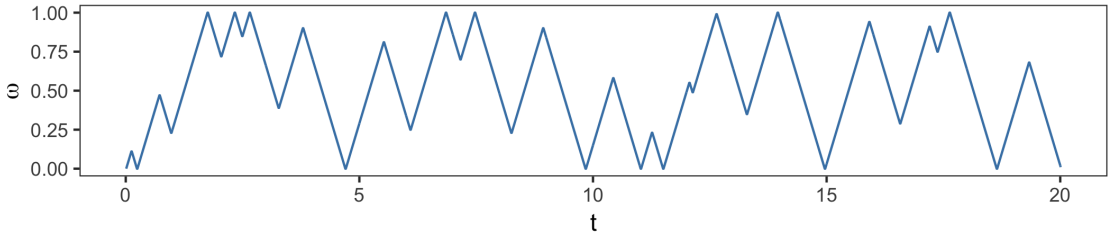
$$\mathcal{L}_Z f(z) = \varepsilon \frac{\partial f}{\partial w}(z) + \lambda(\beta_w) G'(w) \left(f(R(z)) - f(z) \right), \tag{71}$$

where $f \in \mathcal{D}(\mathcal{L}_Z)$ is the set of functions $f \in C^1([0,1] \times \{-1,1\})$ such that $f(w_0, -1) = f(w_0, 1)$ and $\frac{\partial f}{\partial w}(w_0, -1) = -\frac{\partial f}{\partial w}(w_0, 1)$ for $w_0 \in \{0,1\}$. Intuitively we have $Z(t) = (W(t), \varepsilon(t))$ is a PDMP on $[0,1] \times \{-1,1\}$ where $W(t)$ moves in $[0,1]$ with velocity $\varepsilon(t)$. The sign of $\varepsilon(t)$ is reversed at rate $\lambda(\beta_{W(t)}) G'(W(t))$ or when one hits a boundary; see [BBCD+18] for a discussion of PDMP on restricted domains.

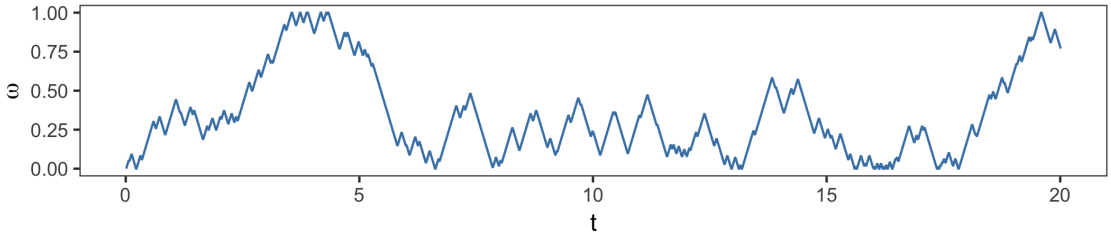
Note that when $G = G_{\text{optimal}} = F_\lambda^{-1}$ we have $\lambda(\beta_w) G'(w) = \Lambda$ for all $w \in [0,1]$. So for the optimal schedule, $\varepsilon(t)$ changes direction at constant rate Λ. Figure 6 shows sample trajectories for various values of Λ. When Λ is small, there are little to no changes in $\varepsilon(t)$ in contrast to when Λ is large.



(a) $\Lambda = 0.1$



(b) $\Lambda = 1$



(c) $\Lambda = 10$

Figure 6: (a)-(c) shows sample trajectories of $W(t)$ where $Z(t) = (W(t), \varepsilon(t))$ under an optimal schedule generated by G_{optimal} with $\Lambda = 0.1, 1, 10$ respectively.

Theorem 6.2. *Suppose $\mu_N = N$ and $Z^N(0)$ converges weakly to $Z(0)$, then Z^N converges weakly*

to Z in $D(\mathbb{R}_+, [0,1] \times \{-1,1\})$. Moreover, the stationary distribution of Z is $\mathrm{Unif}([0,1] \times \{-1,1\})$.

Theorem 6.2 shows that the scaling limit corresponding to the non-reversible index process is not a diffusion. Unlike reversible PT, the scaling limit depends on both the model through λ and on the schedule through G.

7 Experiments

We organize this section into three subsections. In the first subsection, we check the predictions made by our theory on simple models, selected so that analytical calculations are possible while still capturing aspects of more interesting models (hence these are "models of models," or meta-models). In the second subsection, we look at the effect of violating the ELE assumption. Finally, we compare the performance of our non-reversible scheme with other parallel tempering methods.

Reproducibility. To make our adaptive non-reversible method easy to use we implemented it as an inference engine in the open source probabilistic programming language (PPL) Blang `https://github.com/UBC-Stat-ML/blangSDK`. A full description of the models used in the paper are available at `https://github.com/UBC-Stat-ML/blangDemos`, see in particular `https://github.com/UBC-Stat-ML/blangDemos/blob/master/src/main/resources/demos/models.csv` for a list of command line options and data paths used for each model. All methods use the same exploration kernels, namely slice sampling with exponential doubling followed by shrinking [Nea03]. Scripts documenting replication of our experiments are available at `https://github.com/UBC-Stat-ML/ptbenchmark`.

Multi-core implementation. We use lightweight threads [Fri15] to parallelize both the exploration and communication phases, as shown in Procedure 1. We use the algorithm of [LSS12] as implemented in [SL13] to allow each PT chain to have its own random stream. This technique avoids any blocking across threads and hence makes the inner loop of our algorithm truly embarrassingly parallel in N. Moreover, the method of [LSS12] combined with the fact that we fix random seeds means that the numerical value output by the algorithm is not affected by the number of threads used. Increasing the number of threads simply makes the algorithm run faster. In all experiments unless noted otherwise we use the maximum numbers of threads available in the host machine, by default an Intel i5 2.7 GHz (which supports 8 threads via hyper-threading) except for Section 7.3 where we use an Amazon EC2 instance of type `c4.8xlarge`, which is backed by a 2.9 GHz Intel Xeon E5-2666 v3 Processor (20 threads).

7.1 Tractable meta-models

Example 7.1 (Discrete multi-modal problem)**.** Consider a discrete state space $\mathcal{X} = \{0, \dots, 2k\}$, and let $1_{\mathrm{Even}} : \Omega \to \{0,1\}$ denote the indicator function for even numbers. Define $\pi(x) \propto a^{1_{\mathrm{Even}}(x)}$ for $a > 1$ and $\pi_0(x) \propto 1$ with $V(x) = -1_{\mathrm{Even}}(x) \log a$. The distribution π has $k+1$ modes located

where x is even with low probability "barriers" located at x odd. The parameter a controls the relative mass put on the modes. Therefore we have

$$\pi^{(\beta)}(x) = \frac{a^{\beta 1_{\text{Even}}(x)}}{Z(\beta)}, \tag{72}$$

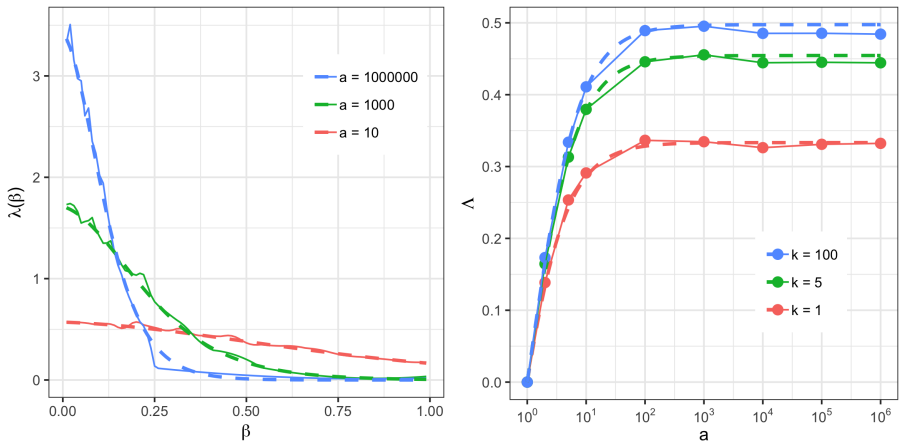
where $Z(\beta) = k + (k+1)a^\beta$. A simple computation using (38) shows that the local communication barrier is,

$$\lambda(\beta) = \frac{k(k+1)a^\beta \log a}{(k+(k+1)a^\beta)^2} \xrightarrow[k\to\infty]{} \frac{a^\beta \log a}{(1+a^\beta)^2}. \tag{73}$$

By integrating we obtain the global communication barrier between π and π_0,

$$\Lambda = \frac{k(k+1)(a-1)}{(2k+1)(k+(k+1)a)} \xrightarrow[k\to\infty]{} \frac{a-1}{2(a+1)}. \tag{74}$$

It can be seen that for all $a > 1$, $\lambda(\beta)$ is decreasing in β as seen in Figure 7 and Λ is increasing in a and k. Therefore, one should expect to see an increase in the intensity of rejection as the relative modes of π become more "peaked" and when the number of modes increases.



(a) (b)

Figure 7: (a) The local communication barrier for $k = 5$ and various values of a. (b) The global communication barrier as a function of a for various k. In (a) the solid line is the approximation $\hat{\lambda}(\beta)$ (respectively $\hat{\Lambda}$ in (b)), resulting from Procedure 3 ($N = 20$, $n = 10\,000$ scans) and the dotted line is the analytic expression in (73) (respectively, (74)).

Example 7.2 (Gaussian)**.** Suppose $\pi \sim N(0, \tau^{-1}\mathbb{I}_d)$, and $\pi_0 \sim N(0, \tau_0^{-1}\mathbb{I}_d)$ with $\tau_0 < \tau$. It can be shown that $\pi^{(\beta)} \sim N(0, \tau_\beta^{-1}\mathbb{I}_d)$ where $\tau_\beta = (1-\beta)\tau_0 + \beta\tau$. Theorem 1 in [PPC04] implies the following closed form expression for λ in the Gaussian case:

$$\lambda(\beta) = \frac{2^{1-d}(\tau - \tau_0)}{B\left(\frac{d}{2}, \frac{d}{2}\right)\tau_\beta}, \tag{75}$$

where $B(a,b)$ is the Beta function. Moreover, for $\beta < \beta'$ the swap function satisfies

$$s(\beta,\beta') = 2F_{\frac{d}{2},\frac{d}{2}}\left(\frac{\beta}{\beta+\beta'}\right), \tag{76}$$

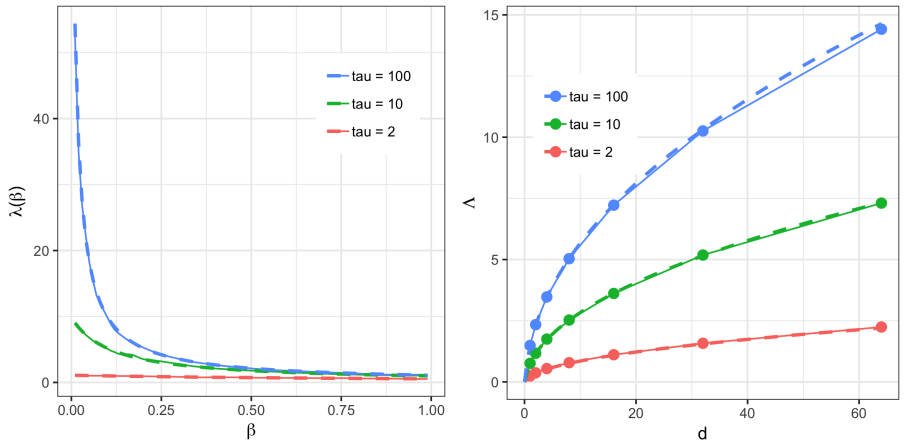
where $F_{a,b}(x)$ is the CDF of a beta distribution with shape parameters a,b. By integrating λ we get the global communication barrier is,

$$\Lambda(\beta) = \frac{2^{1-d}}{B\left(\frac{d}{2},\frac{d}{2}\right)} \log\left(\frac{\tau_\beta}{\tau_0}\right). \tag{77}$$

As $d \to \infty$, we have

$$\Lambda \sim \sqrt{\frac{d}{2\pi}} \log\left(\frac{\tau}{\tau_0}\right), \tag{78}$$

which is consistent with Proposition 4.5. We see from Figure 8 that the empirical approximation of λ, Λ from Procedure 3 are consistent with (75),(77).

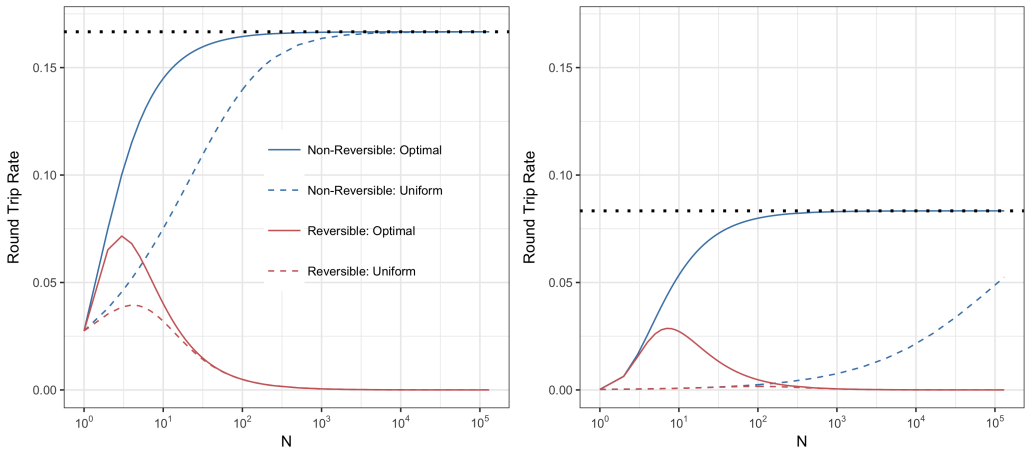


(a) (b)

Figure 8: Analysis of the Gaussian model where $\tau_0 = 1$ for various d, τ. (a) The local communication barrier for $d = 8$ and various values of τ. (b) The global communication barrier as a function of d for various τ. In (a) the solid line is the approximation $\hat{\lambda}(\beta)$ (respectively $\hat{\Lambda}$ in (b)), resulting from Procedure 3 ($N = 60$ and $n = 10\,000$ scans) and the dotted line is the analytic expression in (75) (respectively, (77)).

To determine the optimal annealing schedule $\mathcal{P}_{\text{optimal}} = \{\beta_0^*, \ldots, \beta_N^*\}$, we substituting (77) into in $\Lambda(\beta_k^*) = \Lambda\frac{k}{N}$ and solve for β_k^* as discuss in Section 5.2. This implies the optimal schedule satisfies,

$$\tau_{\beta_k^*} = \tau_{\beta_{k-1}^*}\left(\frac{\tau}{\tau_0}\right)^{\frac{1}{N}}. \tag{79}$$



(a) $\Lambda = 2$ (b) $\Lambda = 5$

Figure 9: (a) Figures (a) and (b) show the round trip rate of a Gaussian model with $\Lambda = 2$ and $\Lambda = 5$ as a function of N. We compare the round trip rates with a uniform schedule (dashed) to the optimal schedule (solid) for both DEO (blue) and REO (red). The dotted horizontal line represents $\bar{\tau}$.

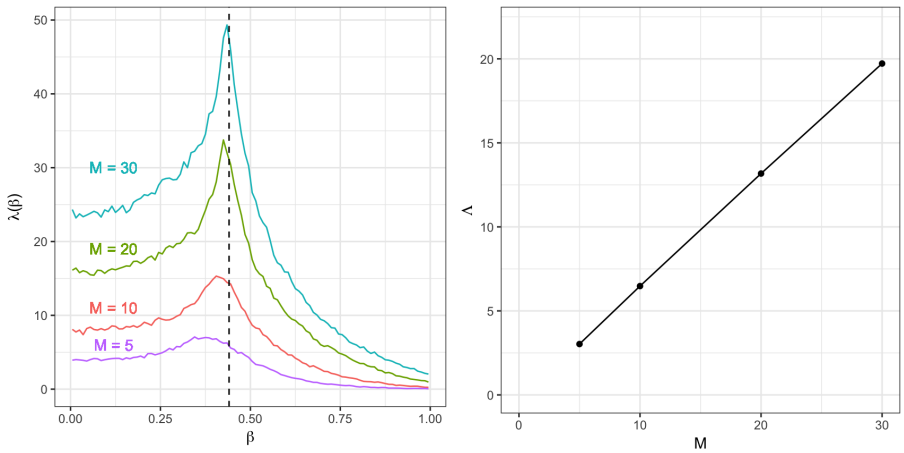
By substituting in $\tau_{\beta_k^*} = \tau_0 + \beta_k^*(\tau - \tau_0)$, we get $\beta_k^* = G(\frac{i}{N})$ where,

$$G(w) = \frac{\left(\frac{\tau_0}{\tau}\right)^{1-w} - \frac{\tau_0}{\tau}}{1 - \frac{\tau_0}{\tau}}. \tag{80}$$

This is the same spacing obtained (based on a different theoretical approach) in [ARR11] and [PPC04] for the Gaussian model (with a small notation change). As described in the next section, in general it is not possible to get analytical expressions for optimal schedules, in which case we resort to Procedure 3. Moreover, we remind the reader that non-reversible allow for annealing schedules containing more chains compared to reversible methods.

Figure 9 compares the theoretical round trip rate for the Gaussian model using the uniform and optimal schedule, this was computed by using Corollory 3.2 and substituting in the exact rejection rates computed from (76). Notice that, for both reversible and non-reversible PT, the optimal schedule produces significantly better round trip rates as expected. In particular when $\Lambda = 5$, it takes nearly $N = 10^5$ number of parallel chains with the uniform schedule to achieve the same round trip rate using an optimal schedule with $N = 10$. Although Corollary 4.8 implies that by increasing N, the round trip rate converges to the optimal round trip rate, this example shows that for even a large, but finite N, a poor schedule can result in very poor performance.

Example 7.3 (Ising model)**.** We now compute numerically λ for the two dimensional Ising model on a 2-dimensional lattice of size $M \times M$ with magnetic moment μ. Using the notation $x_i \sim x_j$ to indicate sites are nearest neighbours on the lattice, the target distribution is annealed by the



(a) (b)

Figure 10: (a) Monte Carlo estimate of the local communication barrier for the Ising model with $\mu = 0$ and $M = 5, 10, 20, 30$ using 5000 scans and $N = 100$. The vertical line is at the critical β_c where the phase transition occurs. (b) The global communication barrier for Ising model as a function of M.

inverse temperature β and the tempered distributions are given by

$$\pi^{(\beta)}(x) = \frac{1}{Z(\beta)} \exp\left(\beta \sum_{x_i \sim x_j} x_i x_j + \mu \sum_i x_i\right). \tag{81}$$

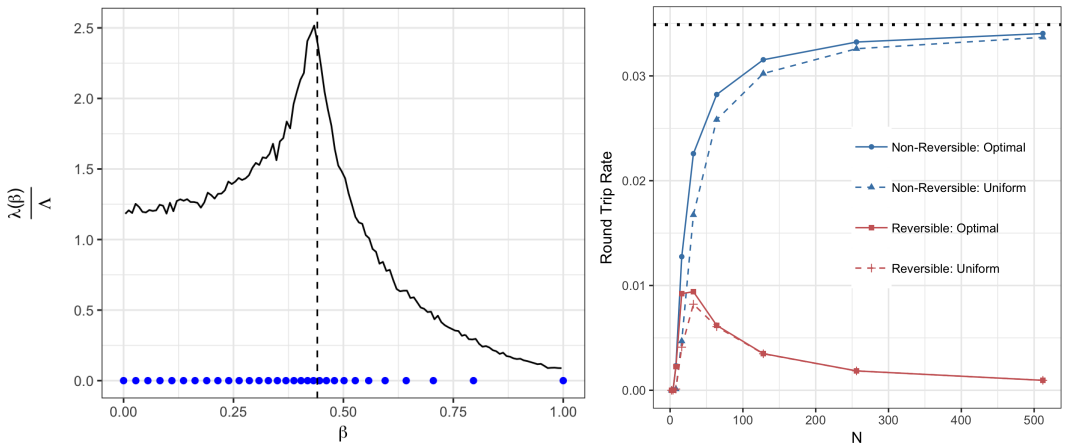
This is an M^2 dimensional model which undergoes a phase transition as $M \to \infty$ at some critical temperature β_c. When $\mu = 0$ it is known that $\beta_c = \log(1 + \sqrt{2})/2$ [Bax07]. Figure 10 shows the rejection intensity for the Ising model with $\mu = 0$ for $M = 5, 10, 20, 30$.

We observe that λ exhibits very different characteristics in this scenario compared to the discrete multimodal and Gaussian models: it is not monotonic and is maximum at the phase transition. We also note that λ increases roughly linearly with respect to M. Given Proposition 4.5, this is to be expected even if this proposition is not directly applicable here as the target distribution does not factorize.

We also approximate the optimal annealing schedule for a $M \times M$ Ising model in Figure 11. Notice how the optimal annealing schedule are denser in regions where λ is high such as the phase transition. When N is small, $\mathcal{P}_{\text{optimal}}$ results in a substantially better round trip rate than $\mathcal{P}_{\text{uniform}}$, but when N is large, the round trip rate for both schedules asymptotically increase towards $\bar{\tau}$. This is consistent with Corollary 4.8.

7.2 Effects of ELE violation

As discussed in Section 3.3, we do not expect ELE to hold exactly: the likelihoods before and after an exploration step are not independent in practice. Increasing the number n_{expl} of MCMC



(a) (b)

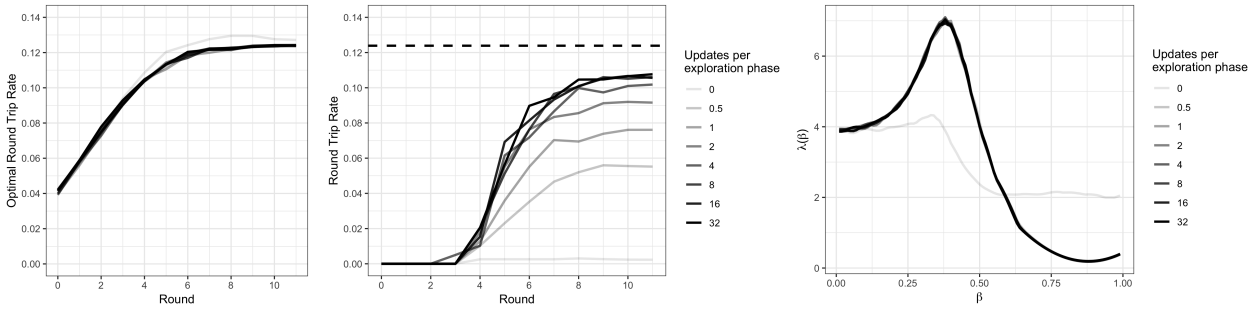
Figure 11: (a) We plot the distribution $\lambda(\beta)/\Lambda$ and optimal annealing schedule for the Ising model with $M = 20$, $\Lambda = 13.33$ $N = 30$ intervals. The vertical line indicates the phase transition. (b) The round trip rates for the Ising model with $\mu = 0$ and $M = 20$ with a uniform schedule (dashed) to the optimal schedule (solid) for both DEO (blue) and REO (red). The dotted horizontal line represents the approximation of the optimal round trip rate $\hat{\tau}$.

exploration steps taken between two communication steps (see Procedure 1) can be used to approach ELE. However a priori one may be concerned that n_{expl} would have to be very large to do so.

To investigate this question, we run the non-reversible method with different values for n_{expl}. Let d_{var} denote the number of variables in the model. We run experiments with $n_{\text{expl}} = 0$, $(1/2)d_{\text{var}}$, d_{var}, $2d_{\text{var}}$, $4d_{\text{var}}$, ..., $32d_{\text{var}}$. The fractions $0, 1/2, 1, 2, \ldots$ involved in this construction can be interpreted as the expected number of times an individual variable is updated in an exploration phase, i.e. the *expected updates per exploration phase*. The only exception is for the prior chain ($\beta = 0$), we always use $n_{\text{expl}} = 1$ since we can get exact samples from the prior distributions considered in our experiments. The case $n_{\text{expl}} = 0$ technically still yields an ergodic chain since the communication chain will ensure all chains visit the prior chain.

For each value of n_{expl} considered we ran 10 times the non-reversible PT scheme with different random seeds (a total of 80 runs). Each run uses 10 000 scans, where one scan consists in n_{expl} exploration iterations followed by one communication iteration. The 10 000 scans are organized into 12 adaptation rounds, where the last round contains 5 000 scans, the penultimate, 2 500, etc. The first round uses a uniformly-spaced annealing schedule, and the subsequent rounds approximate the optimal annealing schedule computed using the estimate of λ from the previous round. For each round and configuration, we report three quantities: (a) the estimated upper bound $\bar{\tau}$ as introduced in Section 4.2, (b) the actual restart rate, directly measured from the empirical replica trajectories, and (c) the estimated function λ. The estimation method for $\bar{\tau}$ is described in Section 5.

We show in Figure 12 the three quantities (a,b,c) described above for the Ising model. The results show that our key results are highly resilient to large violations of the ELE assumption.



(a) (b) (c)

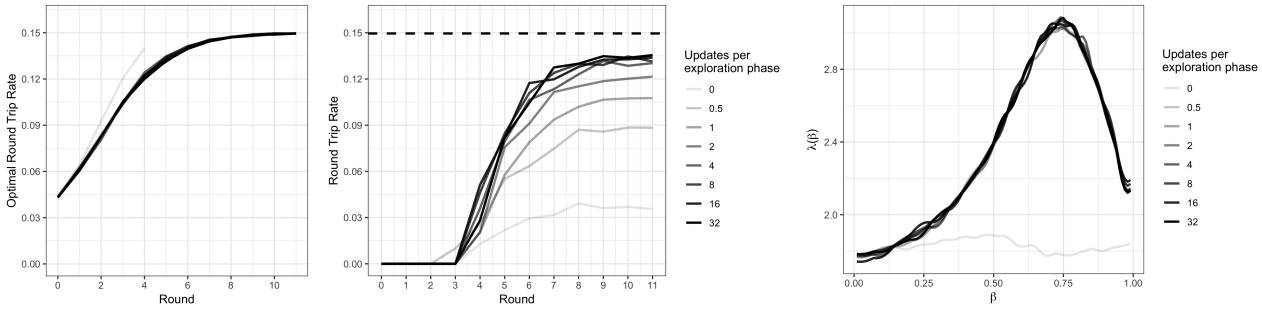
Figure 12: Results on the super-critical Ising model ($M = 5$), varying the number of expected updates per exploration phase (for $\beta = 0$, the prior chain always uses exactly one exact sample). We used $N = 16$ and estimated $\hat{\Lambda} = 3.08$. (a) The estimated upper bound $\bar{\tau}$ from Procedure 3. (b) The round trip rate directly measured from the empirical replica trajectories. The dotted line represents the estimated of $\bar{\tau}$. (c) An estimate of the local communication barrier $\lambda(\beta)$. Whenever $n_{\text{expl}} > 0$, the adaptive scheme accurately learns $\bar{\tau}, \lambda$.

First, for all $n_{\text{expl}} > 0$ considered, the estimated local communication barrier λ and therefore the global one Λ are in very close agreement and are estimated with roughly the same number of adaptation rounds. Second, for all $n_{\text{expl}} > 0$, the actual restart rate is indeed bounded by the estimated value $\bar{\tau}$. The only exception is the setting $n_{\text{expl}} = 0$, where the estimated λ is markedly off compared to the other ones.

We provided in Section 3.3 one motivating example for ELE based on symmetric multi-modal problems. To investigate if breaking these symmetries will induce more severe consequences for violating ELE, we next look at the Ising model under the effect of a magnetic field. We set the magnetic moment $\mu = 0.1$, leading to a target distribution where all marginals assign a mass of less than 0.07 to $x_i = 1$. We show the results in Figure 13. Even in this asymmetric multi-modal problem, we observe the same resilience to violations of ELE. We obtain in Figure 14 similar results for a Bayesian hierarchical model applied to a real dataset.

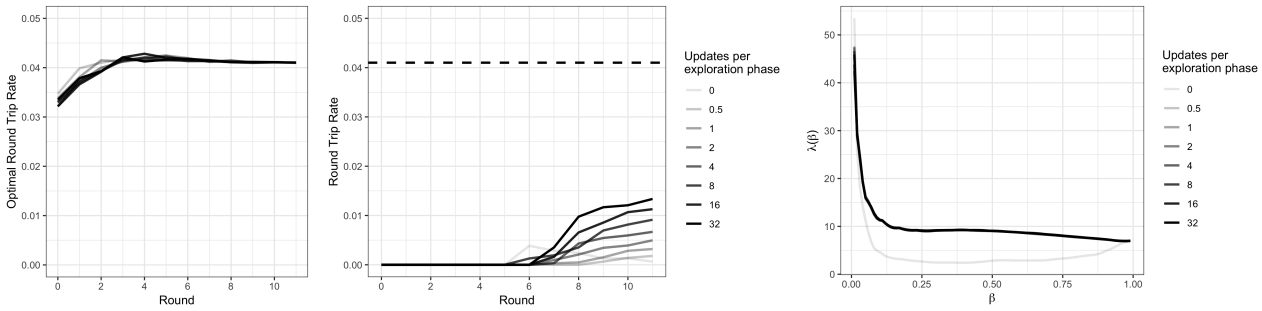
We conjecture that this resilience may come from the structure of typical neighbourhoods of non-reversible parallel tempering. Our intuition can be described using a point process defined as follows. The point process places the rejected swaps in a two-dimensional space, where one axis indexes PT communication iterations, and the other axis consists in the parallel chains. In the regime of a large number of parallel chains, for a given location in this point process, a neighbourhood will contain either zero or one rejection event. The key observation is that in both cases, no two chains interact more than once. This is true by inspection of Figure 15. As a consequence, even when a small number of exploration steps are used between swaps, with high probability they will accumulate by the time a pair of chains meet again.

Note that the same is not true for reversible PT, where the typical local neighbourhood can contain an arbitrary large number of events, and hence pairs of chain can interact more than once in the neighbourhood. As a consequence, we conjecture that for our non-reversible results, it may



(a) (b) (c)

Figure 13: Same quantities as in Figure 12, but where the Ising model ($M = 5, \mu = 0.1$) with as estimated $\hat{\Lambda} = 2.35$) is made asymmetric by adding magnetic field potentials. Note in in (a) predictions of the optimal round trip rate made were cut when $n_{\text{expl}} = 0$ not shown past round 4, because they were significantly larger than the scale of our plot.



(a) (b) (c)

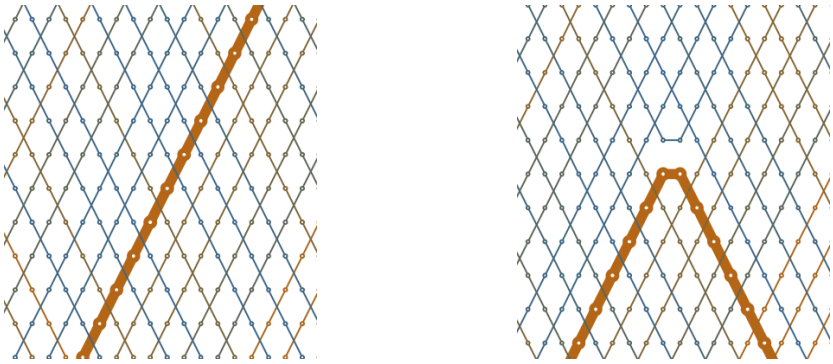
Figure 14: Same quantities as in Figure 12, but with a hierarchical Bayesian model ($\hat{\Lambda} = 12.03$) applied to the historical failure rates of 5 667 launches for 367 types of rockets.

be possible to significantly weaken the ELE assumption, but not for reversible PT. We leave the theoretical investigation of this question for future work.

To provide some empirical justification to this conjecture, we performed another experiment on the magnetic Ising model, fixing the expected updates per exploration phase to 1/2 and increasing the number of chains instead. The results are shown in Figure 16 and support that by increasing the number of parallel chains, the actual tempered restart rate still converges to the theoretical bound from below even in the face of severe ELE violation.

7.3 Comparison with other parallel tempering schemes

In this section, we present results to support that the increased round trip rates enjoyed by our method does indeed translate into increased effective sample size per compute time. The following experiment also benchmarks the empirical running time of our adaptive procedure compared to previous adaptive PT methods [ARR11, MMV13].

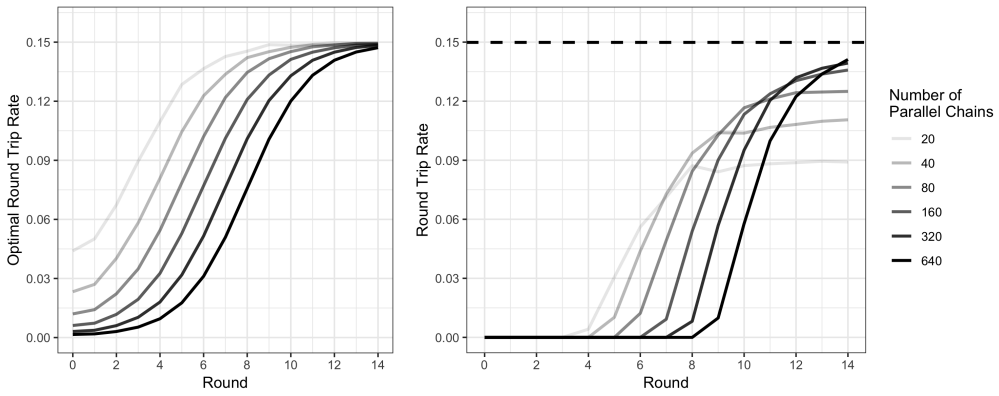


(a) (b)

Figure 15: Typical neighbourhoods of non-reversible parallel tempering in the regime of a large number of parallel chains. There are either no rejection events (a), or one rejection event (b). In both cases, no two chains interact more than once.

Benchmarked methods. The methods we considered are: (1) the stochastic optimization adaptive method for reversible schemes proposed in [ARR11]; (2), a second stochastic optimization scheme, which still selects the optimal number of chains using the 23% rule but uses an improved update scheme from [MMV13]; (3) our adaptive non reversible PT scheme; (4) our scheme, combined with a better initialization based on a preliminary execution of a sequential Monte Carlo algorithm [DMDJ06], we use this to investigate the effect on the violation of the stationarity assumption, and for fairness, we use this sophisticated initialization method for all the methods except (3); and finally, (5), as a baseline, a single-chain MCMC run. All baseline methods are implemented in Blang (`https://github.com/UBC-Stat-ML/blangSDK`), the same probabilistic programming language as used to implement our method. The code for the baseline adaption methods are available at `https://github.com/UBC-Stat-ML/blangDemos`. All methods therefore run on the Java Virtual Machine, so their wall clock running times are all comparable.

Stochastic optimization methods. Both [ARR11] and [MMV13] are based on reversible PT together with two different flavours of stochastic optimization to adaptively select the annealing schedule. In [ARR11], the chains are added one by one, each chain targeting a swap acceptance rate of 23% from the previous one. In [MMV13], the authors modify the scheme in two ways: first, they optimize all annealing parameters simultaneously, and second, they propose a different update for performing the stochastic optimization. To optimize all chains simultaneously, the authors assume that both the number of chains and the equi-acceptance probability are specified. Since this information is not provided to the other methods, in order to perform a fair comparison, for the method we label as "Miasojedow, Moulines, Vihola" we implemented a method which adds the chain one at the time while targeting the swap acceptance rate of 23% but based on the improved stochastic optimization update of [MMV13]. Specifically, both [ARR11] and [MMV13] rely on updates of the form $\rho_{n+1} = \rho_n + \gamma_n(\alpha_{n+1} - 0.23)$ where γ_n is an update schedule and ρ_n is a re-parameterization of difference in annealing parameter from the previous chain β to the one being added β'. The work



(a) (b)

Figure 16: Effect of increasing the number of parallel chains for an example where ELE is severely violated (only half the variables are updated at each exploration step). (a) Estimated upper bound $\bar{\tau}$.(b) Actual restart rate, directly measured from the empirical replica trajectories. Notice that the number of scans required for the round trip rates to in stabilize increases with N as predicted by Proposition 3.1 but eventually attains a higher round round trip.

of [ARR11] uses the update $\beta'_n = \beta(1 + \exp(\rho_n))^{-1}$, whereas the work of [MMV13] specifies the explicit parameterization used for ρ, namely $\rho = \log(\beta'^{-1} - \beta^{-1})$, from which the update becomes $\beta'_n = \beta(1 + \beta \exp(\rho_n))^{-1}$. Moreover, while [ARR11] use $\gamma_n = (n + 1)^{-1}$, [MMV13] suggest to use $\gamma_n = (n + 1)^{-0.6}$. We found that the latter set of choice was more stable. For example, in the next numerical example, the former failed to converge in 100 000 iterations while we encountered no convergence problems with the other methods.

Experimental setup. We ran all methods on a Bayesian mixture model. These experiments are performed on an EC2 instance of type `c4.8xlarge`, which uses a 2.9 GHz Intel Xeon E5-2666 v3 Processor. Since this type of CPU supports 20 threads, we set the number of chains to 16, keeping a slight buffer for garbage collection and background system tasks. With that number of chains, we obtain an average swap rejection rate of 46%, well above the reversible recommendations in the $23-40\%$ range. All methods used 10 000 scans, where a scan uses $n_{\text{expl}} = 3d_{\text{var}}$ exploration rounds. For this example, the number of latent variables to sample is equal to $d_{\text{var}} = 916$. Methods akin to [Gew04] were used to ensure correctness of the MCMC code. We computed the effective sample size using a batch estimator, see, e.g., [Fle08], which partitions the n samples into $\sqrt{n}$ subsets $B_1, B_2, \ldots, B_{\sqrt{n}}$, each of size $|B_k| = \sqrt{n} \pm 1$. To avoid the ESS estimator collapsing in cases where the estimates are too off, for example when a mode is not explored properly, we first ran a longer PT run 50 000 scans, and centre all variance computations on the estimates from that pilot run. If $f(x)$ is the test function of interest, and we have access to the true value or to a very accurate estimate of the mean $\mu = \int f(x)\pi(\mathrm{d}x)$ and variance $\sigma^2 = \int (f(x) - \mu)^2 \pi(\mathrm{d}x)$, the centred ESS estimator is given by $n / \sum_k [|B_k|^{-1} \sum_{x \in B_k} (f(x) - \mu)/\sigma]^2$. The only result qualitatively affected by this method versus standard ESS computation is the performance of the single-chain MCMC, in which standard

ESS calculations severely overestimate the quality of the samples. We ensured the ESS computation code is correct by checking we recover analytic auto-correlations values for an AR(1) process.

Results. In Figure 17, each dot summarized in the box plots represents the ESS per wall clock time in seconds for the marginal of one of the model variables. We present two versions of the plot: one where time is computed including adaptation time, and one where adaptation time is excluded. The results show that adaptation is more efficient with our proposed non-reversible scheme, as evidenced by results where the timing includes adaptation time, and also results in a more efficient sampling algorithm as measured in timing measured by sampling time. The results also show that SMC-based initialization does help PT performance, presumably by relaxing violation of the stationarity assumption. The difference in ESS per second between single chain MCMC and the PT methods underscore the actual difference in the quality of the sample: we show in Figure 18 the posterior distribution for the two mixture proportions (π_1, π_2) as inferred by our non-reversible scheme versus single chain MCMC. From symmetries induced by label switching, we know that the two posterior distributions should be symmetric around 0.5. The plot shows that the single chain MCMC is qualitatively incorrect and only explored one of the two symmetric regions of the posterior distribution whereas PT fully explores the state space. In terms of actual round trip rates, the reversible stochastic optimization-based method achieved a rate of $\tau = 0.28\%$ whereas our method achieved a rate of $\tau = 0.72\%$.

8 Discussion

PT methods are generally quite powerful when they are well tuned but they are also sensitive to design choices such as the communication scheme, the annealing schedule and the number of parallel chains to run. In particular, if PT is used in its reversible version, it is not an embarrassingly parallel algorithm in N, in the sense that adding more parallel chains eventually decreases performance.

We have shown that the situation is qualitatively different in the non-reversible case, and established for the first time that a non-reversible PT algorithm known in the physics literature as DEO can benefit from adding an arbitrary number of chains when implemented in a massively parallel computing setup. More precisely, we showed that with DEO communication, the round trip rate does not deteriorate as the number of additional cores N increases, but actually increases to an optimal round trip rate $\bar{\tau}$. This is in contrast to reversible PT where the round trip rate is $O(N^{-1})$ independent of π, π_0. We also showed that for any number of chains N, non-reversible PT dominates reversible PT. This suggests that practitioners should always use non-reversible scheme with N as large as possible.

We identified the local communication barrier $\lambda(\beta)$ as a key object to understanding the behaviour of non-reversible PT algorithms. From this rate function λ we identified an asymptotic invariant for PT called the global communication barrier Λ, which measures the deviance of π and π_0. We heuristically argued that, as the dimension of the state space d increases, we expect the

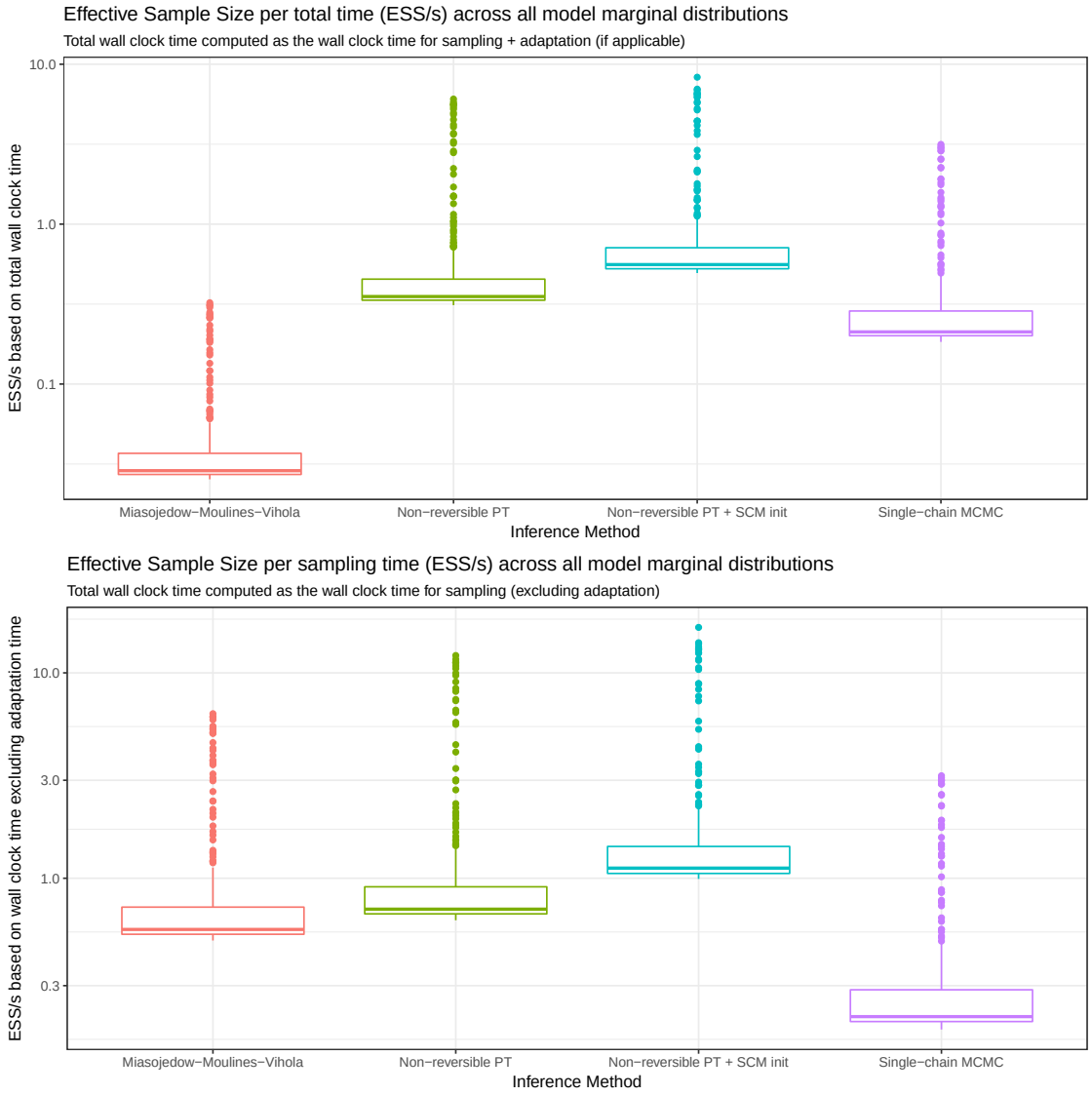


Figure 17: Benchmarking results on a Bayesian mixture model. The y-axis shows measure of efficiency in log-scale, and the x-axis, four different methods compared. The top plot shows the effective sample size per second where time is computed including the adaptation iterations needed for the first four methods. The bottom plot excludes the adaptation time from timing computation.

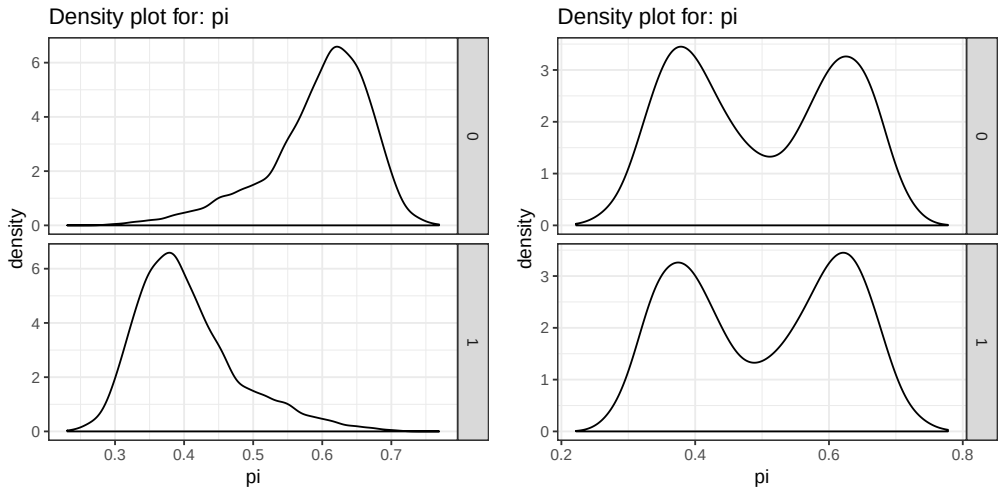


Figure 18: Estimated posterior distributions for the mixture proportion parameters π_1 and π_2 (rows). Left column is from a single-chain MCMC algorithm, right column, from our adaptive non-reversible PT algorithm.

global barrier to grow as $\Lambda = O(\sqrt{d})$ under regularity conditions. Moreover we established a connection between the round trip rate and Λ, showing that $\bar{\tau} = (2 + 2\Lambda)^{-1}$ upper bounds the round trip rate. This means that the global communication barrier Λ can be interpreted as a "sufficient statistic" for π, π_0 from the point of view of a non-reversible PT algorithm, since for N large the round trip rate only depends on π, π_0 through Λ.

Another consequence of our theory is that $\sum_{i=1}^{N} r^{(i-1,i)} \approx \Lambda$ independently of the annealing schedule. This implies that using the rejection probabilities, we can develop an estimator of Λ and $\bar{\tau} = (2 + 2\lambda)^{-1}$. These quantities are easy to approximate, so practitioners can use them to make informed decisions about how to allocate their computational resources. Importantly, note that in all our experiments, the estimate of the bound $\bar{\tau}$ converges very fast, so this allows the user to distinguish between a low round trip rate due to poor design choices versus a low rate arising from a fundamentally hard problem having high value for the global barrier Λ. This is to our knowledge the first result of this kind in the PT literature.

Using the asymptotic analysis of PT, we were able to develop a novel approach to identify the optimal annealing schedule when N is large but finite. In our experiments, our adaptive algorithm converges rapidly, is easy to implement with minimal modification to existing PT implementations, and outperforms other state-of-the-art PT adaptive schemes both in terms of round trip rate and ESS per second (see Section 7).

Finally we study the dynamics driving the qualitative differences between reversible and non-reversible PT through the scaling limits of the index process. We show that for reversible PT, as N increases, the index process for reversible PT weakly converges to a reflected Brownian motion independent of the annealing schedule, π and π_0. For non-reversible PT we show that the index process scales to a PDMP which travels in straight line trajectories and reverses direction at an inhomogeneous rate controlled by λ and the annealing schedule. When we have chosen the optimal

schedule, the rate becomes a constant equal to Λ. Unlike previous literature on PT, our analysis avoids making strong structural assumptions on either π or the state space.

Our analysis makes use of an assumption we call ELE. Empirically we have shown in Section 7.2 that our results appear robust to violation of ELE. We conjecture that this assumption can be lifted in the non-reversible setup. We view a detailed ELE-free theoretical analysis of the weak limit in N of non-reversible PT as an interesting open problem.

9 References

[ADHR04] Gautam Altekar, Sandhya Dwarkadas, John P. Huelsenbeck, and Fredrik Ronquist. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20(3):407–415, February 2004.

[AFGL05] Michael Andrec, Anthony K Felts, Emilio Gallicchio, and Ronald M Levy. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proceedings of the National Academy of Sciences*, 102(19):6801–6806, 2005.

[ARR11] Yves F. Atchadé, Gareth O. Roberts, and Jeffrey S. Rosenthal. Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing*, 21(4):555–568, October 2011.

[Bax07] Rodney J. Baxter. *Exactly Solved Models in Statistical Mechanics*. Dover books on physics. Dover Publications, 2007.

[BBCD+18] Joris Bierkens, Alexandre Bouchard-Côté, Arnaud Doucet, Andrew B Duncan, Paul Fearnhead, Thibaut Lienart, Gareth Roberts, and Sebastian J Vollmer. Piecewise deterministic Markov processes for scalable Monte Carlo on restricted domains. *Statistics & Probability Letters*, 136:148–154, 2018.

[BHH+17] Benjamin Ballnus, Sabine Hug, Kathrin Hatz, Linus Görlitz, Jan Hasenauer, and Fabian J Theis. Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Systems Biology*, 11(1):63, 2017.

[Bil13] Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013.

[BSW13] Björn Böttcher, René Schilling, and Jian Wang. Lévy matters. iii, volume 2099 of Lecture Notes in Mathematics, 2013.

[CL08] Sooyoung Cheon and Faming Liang. Phylogenetic tree construction using sequential stochastic approximation Monte Carlo. *BioSystems*, 91(1):94–107, 2008.

[CLP99] Fang Chen, László Lovász, and Igor Pak. Lifting Markov chains to speed up mixing. In *Proceedings of the 31st annual ACM symposium on Theory of computing*, pages 275–281. ACM, 1999.

[CRI10] KyungHyun Cho, Tapani Raiko, and Alexander Ilin. Parallel tempering is efficient for learning restricted Boltzmann machines. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8. IEEE, 2010.

[CS11] John D. Chodera and Michael R. Shirts. Replica exchange and expanded ensemble simulations as Gibbs sampling: Simple improvements for enhanced mixing. *The Journal of Chemical Physics*, 135(19):194110, November 2011.

[Dav93] Mark HA Davis. *Markov Models & Optimization*. Chapman and Hall, 1993.

[DHN00] Persi Diaconis, Susan Holmes, and Radford M Neal. Analysis of a nonreversible Markov chain sampler. *Annals of Applied Probability*, 10(3):726–752, 2000.

[DLCB14] Guillaume Desjardins, Heng Luo, Aaron Courville, and Yoshua Bengio. Deep tempering. *arXiv preprint arXiv:1410.0123*, 2014.

[DLPD12] Paul Dupuis, Yufei Liu, Nuria Plattner, and Jimmie D Doll. On the infinite swapping limit for parallel tempering. *SIAM Multiscale Modeling & Simulation*, 10(3):986–1022, 2012.

[DMDJ06] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[DP15] Cameron Davidson-Pilon. *Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference*. Addison-Wesley Professional, New York, 1 edition edition, 2015.

[ED05] David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

[EK09] Stewart N Ethier and Thomas G Kurtz. *Markov Processes: Characterization and Convergence*, volume 282. John Wiley & Sons, 2009.

[FC80] F. Fritsch and R. Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, April 1980.

[FFT+14] Ye Fang, Sheng Feng, Ka-Ming Tam, Zhifeng Yun, Juana Moreno, J. Ramanujam, and Mark Jarrell. Parallel tempering simulation of the three-dimensional EdwardsAnderson model with compact asynchronous multispin coding on GPU. *Computer Physics Communications*, 185(10):2467–2478, October 2014.

[Fle08] Flegal, James M. Monte Carlo Standard Errors for MCMC, 2008.

[Fri15] Jeff Friesen. *Java Threads and the Concurrency Utilities*. Apress, Berkely, CA, USA, 1st edition, 2015.

[Gew04] John Geweke. Getting it right. *Journal of the American Statistical Association*, 99(467):799–804, September 2004.

[Gey91] Charles J Geyer. Markov chain Monte Carlo maximum likelihood. *Interface Proceedings*, 1991.

[Har85] J. Michael Harrison. *Brownian Motion and Stochastic Flow Systems*. John Wiley and Sons, 1985.

[HN96] Koji Hukushima and Koji Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.

[Kal97] Olav Kallenberg. *Foundations of Modern Probability*. Springer-Verlag, 1997.

[Kal02] Olav Kallenberg. *Foundations of Modern Probability*. Springer, 2nd edition, 2002.

[KK05] Aminata Kone and David A. Kofke. Selection of temperature intervals for parallel-tempering simulations. *The Journal of Chemical Physics*, 122(20):206101, May 2005.

[Kof02] David A Kofke. On the acceptance probability of replica-exchange Monte Carlo trials. *The Journal of Chemical Physics*, 117(15):6911–6914, 2002.

[KTHT06] Helmut G Katzgraber, Simon Trebst, David A Huse, and Matthias Troyer. Feedback-optimized parallel tempering monte carlo. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(03):P03018, 2006.

[LDMT09] Martin Lingenheil, Robert Denschlag, Gerald Mathias, and Paul Tavan. Efficiency of exchange schemes in replica exchange. *Chemical Physics Letters*, 478(1-3):80–84, 2009.

[LM16] Mateusz Krzysztof Lacki and Błażej Miasojedow. State-dependent swap strategies and automatic reduction of number of temperatures in adaptive parallel tempering algorithm. *Statistics and Computing*, 26(5):951–964, 2016.

[LSR10] Tony Lelièvre, Gabriel Stoltz, and Mathias Rousset. *Free Energy Computations: A Mathematical Perspective*. World Scientific, 2010.

[LSS12] Charles E. Leiserson, Tao B. Schardl, and Jim Sukha. Deterministic parallel random-number generation for dynamic-multithreading platforms. *MIT web domain*, February 2012.

[MB12] Grigorios Mingas and Christos-Savvas Bouganis. Parallel Tempering MCMC Acceleration Using Reconfigurable Hardware. In Oliver C. S. Choy, Ray C. C. Cheung, Peter Athanas, and Kentaro Sano, editors, *Reconfigurable Computing: Architectures, Tools and Applications*, Lecture Notes in Computer Science, pages 227–238. Springer Berlin Heidelberg, 2012.

[MMV13] Błażej Miasojedow, Eric Moulines, and Matti Vihola. An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics*, 22(3):649–664, 2013.

[Nea03] Radford M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, June 2003.

[NH07] Walter Nadler and Ulrich H. E. Hansmann. Generalized ensemble and tempering simulations: A unified view. *Physical Review E*, 75(2), February 2007.

[OKOM01] Tsuneyasu Okabe, Masaaki Kawata, Yuko Okamoto, and Masuhiro Mikami. Replica-exchange Monte Carlo method for the isobaric–isothermal ensemble. *Chemical Physics Letters*, 335(5-6):435–439, 2001.

[PPC04] Cristian Predescu, Mihaela Predescu, and Cristian V Ciobanu. The incomplete beta function law for parallel tempering sampling of classical canonical systems. *The Journal of Chemical Physics*, 120(9):4119–4128, 2004.

[PS03] Jed W Pitera and William Swope. Understanding folding and design: Replica-exchange simulations of"trp-cage"miniproteins. *Proceedings of the National Academy of Sciences*, 100(13):7587–7592, 2003.

[RR14] Gareth O Roberts and Jeffrey S Rosenthal. Minimising MCMC variance via diffusion limits, with an application to simulated tempering. *The Annals of Applied Probability*, 24(1):131–149, 2014.

[SBN13] Yannick G. Spill, Guillaume Bouvier, and Michael Nilges. A convective replica-exchange method for sampling new energy basins. *Journal of Computational Chemistry*, 34(2):132–140, January 2013.

[SH16] Yuji Sakai and Koji Hukushima. Irreversible simulated tempering. *Journal of the Physical Society of Japan*, 85(10):104002, October 2016.

[SL13] Guy Steele and Doug Lea. Splittable Random application programming interface. https://docs.oracle.com/javase/8/docs/api/java/util/SplittableRandom.html, 2013. [Online; accessed 6-May-2019].

[SW86] Robert H Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607, 1986.

[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, 2011.

[Vuc16] Marija Vucelja. Lifting: a nonreversible Markov chain Monte Carlo algorithm. *American Journal of Physics*, 84(12):958–968, 2016.

[Wu17] Fan Wu. Irreversible Parallel Tempering and an Application to a Bayesian Nonparametric Latent Feature Model. Master's thesis, Oxford University, 2017.

[ZC14] Weihong Zhang and Jianhan Chen. Replica exchange with guided annealing for accelerated sampling of disordered protein conformations. *Journal of Computational Chemistry*, 35(23):1682–1689, 2014.

Appendix A Invariant distribution of $\mathbf{K}_n^{\mathrm{PT}}$

Since $\mathbf{K}_n^{\mathrm{PT}} = \mathbf{K}_n^{\mathrm{comm}}\mathbf{K}^{\mathrm{expl}}$, to show $\mathbf{K}_n^{\mathrm{PT}}$ is $\bar{\boldsymbol{\pi}}$-invariant, it is enough to verify that both $\mathbf{K}^{\mathrm{expl}}$ and $\mathbf{K}_n^{\mathrm{comm}}$ are $\bar{\boldsymbol{\pi}}$-invariant. It is clear by construction that $\mathbf{K}^{\mathrm{expl}}$ defined by (4) is $\bar{\boldsymbol{\pi}}$-stationary, so it remains to verify that this $\mathbf{K}_n^{\mathrm{comm}}$. Clearly $\mathbf{K}^{\mathrm{SEO}}, \mathbf{K}_n^{\mathrm{DEO}}$ are trivially $\bar{\boldsymbol{\pi}}$-invariant if each swap kernel $\mathbf{K}^{(i,j)}$ is. We verify this directly. Let $\bar{\mathbf{x}} = (\mathbf{x}, \sigma) \in \mathcal{X}^{N+1} \times \mathrm{Perm}([N])$, then

$$
\begin{aligned}
&\int_{\mathcal{X}^{N+1}\times\mathrm{Perm}([N])} \bar{\boldsymbol{\pi}}(\mathrm{d}\bar{\mathbf{x}})\mathbf{K}^{(i,j)}(\bar{\mathbf{x}}, A \times \{\sigma'\}) \\
&\qquad = \frac{1}{(N+1)!}\sum_{\sigma}\int_{\mathcal{X}^{N+1}} \boldsymbol{\pi}(\mathrm{d}\mathbf{x})\mathbf{K}^{(i,j)}(\bar{\mathbf{x}}, A \times \{\sigma'\}) && (82)\\
&\qquad = \frac{1}{(N+1)!}\sum_{\sigma}\int_{\mathcal{X}^{N+1}} \boldsymbol{\pi}(\mathrm{d}\mathbf{x})\left(1-\alpha^{(i,j)}(\mathbf{x})\right)\delta_{\mathbf{x}}(A)\mathbb{I}[\sigma' = \sigma] \\
&\qquad\quad + \frac{1}{(N+1)!}\sum_{\sigma}\int_{\mathcal{X}^{N+1}} \boldsymbol{\pi}(\mathrm{d}\mathbf{x})\alpha^{(i,j)}(\mathbf{x})\delta_{\mathbf{x}^{(i,j)}}(A)\mathbb{I}[\sigma' = (i,j)\circ\sigma] && (83)\\
&\qquad = \frac{1}{(N+1)!}\int_{\mathcal{X}^{N+1}} \boldsymbol{\pi}(\mathrm{d}\mathbf{x})\sum_{\sigma}\left(1-\alpha^{(i,j)}(\mathbf{x})\right)\delta_{\mathbf{x}}(A)\mathbb{I}[\sigma' = \sigma] \\
&\qquad\quad + \frac{1}{(N+1)!}\int_{\mathcal{X}^{N+1}} \boldsymbol{\pi}(\mathrm{d}\mathbf{x})\sum_{\sigma}\alpha^{(i,j)}(\mathbf{x})\delta_{\mathbf{x}^{(i,j)}}(A)\mathbb{I}[\sigma' = (i,j)\circ\sigma] && (84)\\
&\qquad = \frac{1}{(N+1)!}\int_{\mathcal{X}^{N+1}} \boldsymbol{\pi}(\mathrm{d}\mathbf{x})\left\{\left(1-\alpha^{(i,j)}(\mathbf{x})\right)\delta_{\mathbf{x}}(A) + \alpha^{(i,j)}(\mathbf{x})\delta_{\mathbf{x}^{(i,j)}}(A)\right\} && (85)\\
&\qquad = \frac{1}{(N+1)!}\int_{\mathcal{X}^{N+1}} \boldsymbol{\pi}(\mathrm{d}\mathbf{x})\mathbf{K}^{(i,j)}(\mathbf{x}, A) && (86)\\
&\qquad = \frac{1}{(N+1)!}\boldsymbol{\pi}(A) && (87)\\
&\qquad = \bar{\boldsymbol{\pi}}(A \times \{\sigma'\}). && (88)
\end{aligned}
$$

Therefore, $\mathbf{K}_n^{\mathrm{PT}}$ is $\bar{\boldsymbol{\pi}}$-invariant.

Appendix B Proof of Proposition 3.1

Proof of Proposition 3.1. To simplify notation for the rest of the proof, we define $T_\uparrow$ and $T_\downarrow$ as the hitting times to the posterior and prior defined by,

$$
T_\uparrow = \min\{n : (I_n, \varepsilon_n) = (N, 1)\}, \quad T_\downarrow = \min\{n : (I_n, \varepsilon_n) = (0, -1)\}. \tag{89}
$$

We will also denote

$$
s_i = s^{(i-1,i)} \tag{90}
$$

$$
r_i = r^{(i-1,i)}. \tag{91}
$$

(a) If we define $a_\bullet^i = \mathbb{E}_{\text{SEO}}(T_\bullet | I_0 = i)$ for $i = 0, \ldots, N$, then we have

$$\mathbb{E}_{\text{SEO}}(T) = a_\uparrow^0 + a_\downarrow^N. \tag{92}$$

By the Markov property, for $i = 1, \ldots, N-1$ we have $a_\bullet^i$ satisfies the recursion,

$$a_\bullet^i = \frac{1}{2} s_{i+1}(a_\bullet^{i+1} + 1) + \frac{1}{2} s_i (a_\bullet^{i-1} + 1) + \frac{1}{2}(r_{i+1} + r_i)(a_\bullet^i + 1) \tag{93}$$

For $i = 1, \ldots, N$ we substitute in $b_\bullet^i = a_\bullet^i - a_\bullet^{i-1}$ into (93). After simplification, we get that $b_\bullet^i$ satisfies the following recursive relation,

$$-2 = s_{i+1} b_\bullet^{i+1} - s_i b_\bullet^i. \tag{94}$$

The solutions to (94) are,

$$s_i b_\bullet^i = s_1 b_\bullet^1 - 2(i-1), \tag{95}$$

or equivalently,

$$s_i b_\bullet^i = s_N b_\bullet^N + 2(N-i) \tag{96}$$

We now deal with the case of $\uparrow$ and $\downarrow$ separately.

- To determine $a_\uparrow^0$ we note that a if $I_0 = 0$ then $I_1 = 1$ with probability $\frac{1}{2} s_1$ and $I_1 = 0$ otherwise. So $a_\uparrow^0$ satisfies,

$$a_\uparrow^0 = \frac{1}{2} s_1 (a_\uparrow^1 + 1) + \left(1 - \frac{1}{2} s_1\right)(a_\uparrow^0 + 1), \tag{97}$$

  or equivalently,

$$s_1 b_\uparrow^1 = -2. \tag{98}$$

  Substituting this into (95) implies $s_i b_\uparrow^i = -2i$. By summing $b_\uparrow^i = a_\uparrow^i - a_\uparrow^{i-1}$ from $i = 1, \ldots, N$ and noting $a_\uparrow^N = 0$ we get,

$$a_\uparrow^0 = \sum_{i=1}^N \frac{2i}{s_i}. \tag{99}$$

- Similarly to determine $a_\downarrow^N$ we note that a if $I_0 = N$ then $I_1 = N-1$ with probability

$\frac{1}{2}s_N$ and $I_1 = N$ otherwise. So $a_\downarrow^N$ satisfies,

$$a_\downarrow^N = \frac{1}{2}s_N(a_\downarrow^{N-1} + 1) + \left(1 - \frac{1}{2}s_N\right)(a_\downarrow^N + 1), \tag{100}$$

or equivalently,

$$s_N b_\downarrow^N = 2 \tag{101}$$

Substituting this into (96) implies $s_i b_\downarrow^i = 2 + 2(N-i)$. By summing $b_\downarrow^i = a_\downarrow^i - a_\downarrow^{i-1}$ from $i = 1, \ldots, N$ and noting $a_\downarrow^0 = 0$ we get,

$$a_\downarrow^N = \sum_{i=1}^{N} \frac{2(N-i)+2}{s_i}. \tag{102}$$

Substituting in (99) and (102) into (92) we get,

$$\mathbb{E}_{\mathrm{SEO}}(T) = \sum_{i=1}^{N} \frac{2i}{s_i} + \sum_{i=1}^{N} \frac{2(N-i)+2}{s_i} \tag{103}$$

$$= 2(N+1)\sum_{i=1}^{N} \frac{1}{s_i} \tag{104}$$

$$= 2N(N+1) + 2(N+1)\sum_{i=1}^{N} \frac{r_i}{s_i}. \tag{105}$$

(b) If we define, $a_\bullet^{i,\varepsilon} = \mathbb{E}_{\mathrm{DEO}}(T_\bullet | I_0 = i, \varepsilon_0 = \varepsilon)$ for $i = 0, \ldots, N$ and $\varepsilon = +, -$, then we have,

$$\mathbb{E}_{\mathrm{DEO}}(T) = a_\uparrow^{0,-} + a_\downarrow^{N,+}. \tag{106}$$

Note that for $i = 1, \ldots, N-1$ we have $a_\bullet^{i,\varepsilon}$ satisfies the recursion relations,

$$a_\bullet^{i,+} = s_{i+1}(a_\bullet^{i+1,+} + 1) + r_{i+1}(a_\bullet^{i,-} + 1) \tag{107}$$

$$a_\bullet^{i,-} = s_i(a_\bullet^{i-1,-} + 1) + r_i(a_\bullet^{i,-} + 1) \tag{108}$$

If we substitute $c_\bullet^i = a_\bullet^{i,+} + a_\bullet^{i-1,-}$, and $d_\bullet^i = a_\bullet^{i,+} - a_\bullet^{i-1,-}$ into (107) and (108) and simplify, we get,

$$a_\bullet^{i+1,+} - a_\bullet^{i,+} = r_{i+1}d_\bullet^{i+1} - 1 \tag{109}$$

$$a_\bullet^{i,-} - a_\bullet^{i-1,-} = r_i d_\bullet^i + 1 \tag{110}$$

By subtracting and adding (109) and (110) we get the joint recursion relation for $c_\bullet^i$ and $d_\bullet^i$,

$$c_\bullet^{i+1} - c_\bullet^i = r_{i+1} d_\bullet^{i+1} + r_i d_\bullet^i \tag{111}$$

$$d_\bullet^{i+1} - d_\bullet^i = r_{i+1} d_\bullet^{i+1} + r_i d_\bullet^i - 2 \tag{112}$$

Note that (112) can be rewritten as

$$s_{i+1} d_\bullet^{i+1} - s_i d_\bullet^i = -2. \tag{113}$$

If can solve $c_\bullet^i$ and $d_\bullet^i$, we can recover $a_\bullet^{i,\varepsilon}$ by using,

$$a_\bullet^{i,+} = \frac{c_\bullet^i + d_\bullet^i}{2} \tag{114}$$

$$a_\bullet^{i-1,-} = \frac{c_\bullet^i - d_\bullet^i}{2} \tag{115}$$

We now deal with the $\uparrow$ and $\downarrow$ cases separately.

- Note that $a_\uparrow^{0,-} = a_\uparrow^{0,+} + 1$. We can substitute this into (109) to get $s_1 d_\uparrow^1 = -2$, which combined with (113) implies,

$$s_i d_\uparrow^i = -2i. \tag{116}$$

  Since $a_\uparrow^{N,+} = 0$ we have $c_\uparrow^N = -d_\uparrow^N$, so by summing (111) we get,

$$2a_\uparrow^{0,-} = c_\uparrow^1 - d_\uparrow^1 \tag{117}$$

$$= c_\uparrow^N - d_\uparrow^1 - \sum_{i=1}^{N-1} (c_\uparrow^{i+1} - c_\uparrow^i) \tag{118}$$

$$= -d_\uparrow^N - d_\uparrow^1 - \sum_{i=1}^{N-1} (r_{i+1} d_\uparrow^{i+1} + r_i d_\uparrow^i) \tag{119}$$

$$= -s_N d_\uparrow^N - s_1 d_\uparrow^1 - 2\sum_{i=1}^{N} r_i d_\uparrow^i. \tag{120}$$

  After substituting in (116) into (120) and dividing by 2 we get,

$$a_\uparrow^{0,-} = N + 1 + \sum_{i=1}^{N} \frac{2ir_i}{s_i}. \tag{121}$$

- Note that $a_\downarrow^{N,+} = a_\uparrow^{N,-} + 1$. We can substitute this into (110) to get $s_N d_\downarrow^N = 2$, which combined with (113) implies,

$$s_i d_\downarrow^i = 2(N - i + 1). \tag{122}$$

Since $a_{\downarrow}^{0,-} = 0$ we have $c_{\downarrow}^{1} = d_{\downarrow}^{1}$, so by summing (111) we get,

$$2a_{\downarrow}^{N,+} = c_{\downarrow}^{N} + d_{\downarrow}^{N} \tag{123}$$

$$= c_{\downarrow}^{1} + d_{\downarrow}^{N} + \sum_{i=1}^{N-1} (c_{\downarrow}^{i+1} - c_{\downarrow}^{i}) \tag{124}$$

$$= d_{\downarrow}^{1} + d_{\downarrow}^{N} + \sum_{i=1}^{N-1} (r_{i+1} d_{\downarrow}^{i+1} + r_i d_{\downarrow}^{i}) \tag{125}$$

$$= s_1 d_{\downarrow}^{1} + s_N d_{\downarrow}^{N} + 2 \sum_{i=1}^{N} r_i d_{\downarrow}^{i}. \tag{126}$$

After substituting in (122) into (126) and dividing by 2 we get,

$$a_{\downarrow}^{N,+} = N + 1 + \sum_{i=1}^{N} \frac{2(N-i+1) r_i}{s_i}. \tag{127}$$

Finally, by substituting in (121) and (127) into (106), we get

$$\mathbb{E}_{\mathrm{DEO}}(T) = 2(N+1) + 2(N+1) \sum_{i=1}^{N} \frac{r_i}{s_i}. \tag{128}$$

□

Appendix C Proof of Proposition 4.2

Suppose V^k is integrable with respect to π_0 and π, we want to show here that $\lambda : [0,1] \to \mathbb{R}_+$ given by

$$\lambda(\beta) = \frac{1}{2} \int_{\mathcal{X}^2} |V(x) - V(y)| \pi^{(\beta)}(x) \pi^{(\beta)}(y) \mathrm{d}x \mathrm{d}y \tag{129}$$

is in $C^{k-1}([0,1])$. If we define $L(x,y) = L(x)L(y)$ and $\pi_0(x,y) = \pi_0(x)\pi_0(y)$, we can rewrite (129) as,

$$\lambda(\beta) = \frac{1}{2\mathcal{Z}(\beta)^2} \int_{\mathcal{X}^2} |V(x) - V(y)| L(x,y)^{\beta} \pi_0(x,y) \mathrm{d}x \mathrm{d}y \tag{130}$$

$$= \frac{g(\beta)}{2\mathcal{Z}(\beta)^2} \tag{131}$$

where $\mathcal{Z}, g : [0,1] \to \mathbb{R}_+$ are defined by

$$\mathcal{Z}(\beta) = \int_{\mathcal{X}} L(x)^\beta \pi_0(x) \mathrm{d}x, \tag{132}$$

$$g(\beta) = \int_{\mathcal{X}^2} |V(x) - V(y)| L(x,y)^\beta \pi_0(x,y) \mathrm{d}x \mathrm{d}y. \tag{133}$$

Since $\mathcal{Z}(\beta) > 0$ on $[0,1]$, if we can show that $\mathcal{Z}, g \in C^{k-1}([0,1])$ then it implies that $\lambda \in C^{k-1}([0,1])$.

Lemma C.1. *If V^k is integrable with respect to π_0 and π for $k \in \mathbb{N}$. Then for all $\beta \in [0,1]$, $j \leq k$, V^j is $\pi^{(\beta)}$-integrable.*

Proof. We begin by noting that for all $L > 0$, for $\beta \in [0,1]$, we have $L^\beta \leq 1 + L$. This implies,

$$\begin{aligned}
& \int_{\mathcal{X}} |V(x)|^k \pi^{(\beta)}(x) \mathrm{d}x && (134)\\
= \; & \frac{1}{\mathcal{Z}(\beta)} \int_{\mathcal{X}} |V(x)|^k L(x)^\beta \pi_0(x) \mathrm{d}x && (135)\\
\leq \; & \frac{1}{\mathcal{Z}(\beta)} \int_{\mathcal{X}} |V(x)|^k \pi_0(x) \mathrm{d}x + \frac{1}{\mathcal{Z}(\beta)} \int_{\mathcal{X}} |V(x)|^k L(x) \pi_0(x) \mathrm{d}x && (136)\\
= \; & \frac{\mathcal{Z}(0)}{\mathcal{Z}(\beta)} \int_{\mathcal{X}} |V(x)|^k \pi_0(x) \mathrm{d}x + \frac{\mathcal{Z}(1)}{\mathcal{Z}(\beta)} \int_{\mathcal{X}} |V(x)|^k \pi(x) \mathrm{d}x && (137)\\
< \; & \infty. && (138)
\end{aligned}$$

Therefore since V^k is π_0 and π-integrable, V^k is $\pi^{(\beta)}$-integrable. Finally by Jensen's inequality we have for $j \geq k$,

$$\int_{\mathcal{X}} |V(x)|^j \pi^{(\beta)}(x) \mathrm{d}x \leq \left(\int_{\mathcal{X}} |V(x)|^k \pi^{(\beta)}(x) \mathrm{d}x \right)^{\frac{j}{k}} < \infty. \tag{139}$$

□

Proposition C.2. *Suppose V^k is integrable with respect to π_0 and π for some $k \in \mathbb{N}$ then:*

(a) $\mathcal{Z} \in C^k([0,1])$ with derivatives satisfying,

$$\frac{d^j \mathcal{Z}}{d\beta^j} = \int_{\mathcal{X}} (-1)^j V(x)^j L(x)^\beta \pi_0(x) \mathrm{d}x, \tag{140}$$

for $j \leq k$.

(b) $g \in C^{k-1}([0,1])$ with derivatives satisfying,

$$\frac{d^j g}{d\beta^j} = \int_{\mathcal{X}^2} (-1)^j |V(x) - V(y)| (V(x) + V(y))^j L(x,y)^\beta \pi_0(x,y) \mathrm{d}x \mathrm{d}y, \tag{141}$$

for $j < k$.

Proof. (a) Let $h(x,\beta) = L(x)^\beta \pi_0(x) = \exp(-\beta V(x))\pi_0(x)$ which satisfies,

$$\frac{\partial^j}{\partial \beta^j} h(x,\beta) = (-1)^j V(x)^j L(x)^\beta \pi_0(x). \tag{142}$$

Note for all $\beta \in [0,1]$ and $j \le k$,

$$\sup_{\beta\in[0,1]} \left| \frac{\partial^j}{\partial \beta^j} h(x,\beta) \right| \le |V(x)|^j \pi_0(x) + |V(x)|^j L(x)\pi_0(x). \tag{143}$$

The left hand side of (143) dominates $\frac{\partial^j h}{\partial \beta^j}$ uniformly in β and is integrable by Lemma C.1. The result follows using the Leibniz integration rule.

(b) Let $\tilde{h}(x,y,\beta) = |V(x) - V(y)| L(x,y)^\beta \pi_0(x,y)$. By noting $\log L(x,y) = -V(x) - V(y)$, we get

$$\frac{\partial^j}{\partial \beta^j} \tilde{h}(x,y,\beta) = (-1)^j |V(x) - V(y)| (V(x) + V(y))^j L(x,y)^\beta \pi_0(x,y). \tag{144}$$

Similar to (a), we have for all $\beta \in [0,1]$, $j \le k-1$,

$$\begin{aligned} \sup_{\beta\in[0,1]} \left| \frac{\partial^j}{\partial \beta^j} \tilde{h}(x,y,\beta) \right| \le\ & |V(x) - V(y)||V(x) + V(y)|^j \pi_0(x,y) \\ & + |V(x) - V(y)||V(x) + V(y)|^j L(x,y)\pi_0(x,y), \end{aligned} \tag{145}$$

The left hand side of (145) dominates $\frac{\partial^j \tilde{h}}{\partial \beta^j}$ uniformly in β. It is integrable by Lemma C.1 and using the fact that V^k is integrable with respect to π_0 and π. The result follows using the Leibniz integration rule.

□

Appendix D Proof of Proposition 4.7

Proof of Proposition 4.7. Let $\mathcal{P}_N = \{\beta_0, \ldots, \beta_N\}$. There exists an i_0 such that $\mathcal{P}_{N+1} = \mathcal{P}_N \cup \{\beta\}$ for some $\beta_{i_0} < \beta < \beta_{i_0+1}$. Therefore,

$$E(\mathcal{P}_{N+1}) - E(\mathcal{P}_N) = \frac{r(\beta_{i_0},\beta)}{s(\beta_{i_0},\beta)} + \frac{r(\beta,\beta_{i_0+1})}{s(\beta,\beta_{i_0+1})} - \frac{r(\beta_{i_0},\beta_{i_0+1})}{s(\beta_{i_0},\beta_{i_0+1})} \tag{146}$$

$$= \frac{r(\beta_{i_0},\beta)s(\beta,\beta_{i_0+1}) + s(\beta_{i_0},\beta)r(\beta,\beta_{i_0+1})}{s(\beta_{i_0},\beta)s(\beta,\beta_{i_0+1})} - \frac{r(\beta_{i_0},\beta_{i_0+1})}{s(\beta_{i_0},\beta_{i_0+1})} \tag{147}$$

$$\le \frac{r(\beta_{i_0},\beta) + r(\beta,\beta_{i_0+1})}{s(\beta_{i_0},\beta)s(\beta,\beta_{i_0+1})} - \frac{r(\beta_{i_0},\beta_{i_0+1})}{s(\beta_{i_0},\beta_{i_0+1})} \tag{148}$$

$$\le \frac{r(\beta_{i_0},\beta) + r(\beta,\beta_{i_0+1})}{1 - r(\beta_{i_0},\beta) - r(\beta,\beta_{i_0+1})} - \frac{r(\beta_{i_0},\beta_{i_0+1})}{s(\beta_{i_0},\beta_{i_0+1})}. \tag{149}$$

The last inequality holds since

$$s(\beta_{i_0},\beta)s(\beta,\beta_{i_0+1}) = (1-r(\beta_{i_0},\beta))(1-r(\beta,\beta_{i_0+1})) \tag{150}$$
$$\geq 1 - r(\beta_{i_0},\beta) - r(\beta,\beta_{i_0+1}). \tag{151}$$

By Corollary 4.3 we have

$$r(\beta_{i_0},\beta) + r(\beta,\beta_{i_0+1}) = r(\beta_{i_0},\beta_{i_0+1}) + O(\|\mathcal{P}_N\|^3). \tag{152}$$

which implies,

$$s(\beta_{i_0},\beta)s(\beta,\beta_{i_0+1}) \geq s(\beta_{i_0},\beta_{i_0+1}) + O(\|\mathcal{P}_N\|^3). \tag{153}$$

and,

$$E(\mathcal{P}_{N+1}) - E(\mathcal{P}_N) \leq \frac{r(\beta_{i_0},\beta_{i_0+1}) + O(\|\mathcal{P}_N\|^3)}{s(\beta_{i_0},\beta_{i_0+1}) + O(\|\mathcal{P}_N\|^3)} - \frac{r(\beta_{i_0},\beta_{i_0+1})}{s(\beta_{i_0},\beta_{i_0+1})} \tag{154}$$

As $\|\mathcal{P}_N\| \to 0$, we have the right hand side is asymptotically equivalent to zero. Therefore $E(\mathcal{P}_{N+1}) \lesssim E(\mathcal{P}_N)$ as $\|\mathcal{P}_N\| \to 0$ and $E(\mathcal{P}_N)$ is asymptotically decreasing.

To show that $E(\mathcal{P}_N)$ asymptotically decreases to Λ, note that for all $\mathcal{P}_N$,

$$\sum_{i=1}^{N} r^{(i-1,i)} \leq E(\mathcal{P}_N) \leq \frac{1}{\min_j s^{(j-1,j)}} \sum_{i=1}^{N} r^{(i-1,i)}. \tag{155}$$

By Corollary 4.3 we have $\min_j s_j = 1 + O(\|\mathcal{P}_N\|)$ and $\sum_{i=1}^{N} r^{(i-1,i)} = \Lambda + O(\|\mathcal{P}_N\|^2)$ which combined with (155) implies

$$E(\mathcal{P}_N) = \Lambda + O(\|\mathcal{P}_N\|). \tag{156}$$

Therefore as $\|\mathcal{P}_N\| \to 0$, $E(\mathcal{P}_N)$ converges to Λ at a $O(\|\mathcal{P}_N\|)$ rate.

□

Appendix E Proof of Proposition 4.5

Proof of Proposition 4.5. For $k = 1, 2$ us define $\mathbf{V}_k^{(\beta)} \stackrel{d}{=} \mathbf{V}(X_k^{(\beta)})$ where $X_k^{(\beta)} \sim \pi_d^{(\beta)}$ and $\mathbf{V}(x) = \sum_{i=1}^{d} V(x_i)$. The independence structure from Equation (46) tells us that $\mathbf{V}_k^{(\beta)}$ can be decomposed as $\mathbf{V}_k^{(\beta)} = \sum_{i=1}^{d} V_{ki}^{(\beta)}$ where $V_{ki}^{(\beta)}$ are iid with common distribution $V^{(\beta)}$, and therefore we have,

$$\mathbf{V}_1^{(\beta)} - \mathbf{V}_2^{(\beta)} = \sum_{i=1}^{d} V_{1i}^{(\beta)} - V_{2i}^{(\beta)}. \tag{157}$$

The random variables $\{V_{1i}^{(\beta)} - V_{2i}^{(\beta)}\}_{i=1}^{d}$ are independent and identically distributed with mean zero and variance $2\sigma^2(\beta)$. By the central limit theorem,

$$\frac{\mathbf{V}_1^{(\beta)} - \mathbf{V}_2^{(\beta)}}{\sqrt{2\sigma^2(\beta)d}} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \frac{V_{1i}^{(\beta)} - V_{2i}^{(\beta)}}{\sqrt{2\sigma^2(\beta)}} \xrightarrow[d\to\infty]{} \tilde{Z} \sim N(0,1). \tag{158}$$

Thus we have

$$\lambda_d(\beta) = \frac{1}{2}\mathbb{E}\left[|\mathbf{V}_1^{(\beta)} - \mathbf{V}_2^{(\beta)}|\right] \tag{159}$$

$$= \frac{1}{2}\sqrt{2\sigma^2(\beta)d}\,\mathbb{E}\left[\left|\frac{\mathbf{V}_1^{(\beta)} - \mathbf{V}_2^{(\beta)}}{\sqrt{2\sigma^2(\beta)d}}\right|\right]. \tag{160}$$

The sequence of variables indexed by d in the expectation in (160) is also uniformly integrable. This follows by noting that the second moment of the integrand in (160) is uniformly bounded in d:

$$\sup_d \mathbb{E}\left[\left|\frac{\mathbf{V}_1^{(\beta)} - \mathbf{V}_2^{(\beta)}}{\sqrt{2\sigma^2(\beta)d}}\right|^2\right] = \sup_d \frac{1}{2\sigma^2(\beta)d} \sum_{i=1}^{d} \mathrm{Var}\left[V_{1i}^{(\beta)} - V_{2i}^{(\beta)}\right] = 1. \tag{161}$$

By $d \to \infty$ and using (158) we have,

$$\lim_{d\to\infty} \sqrt{\frac{2}{\sigma^2(\beta)d}}\lambda_d(\beta) = \mathbb{E}|\tilde{Z}| = \sqrt{\frac{2}{\pi}}, \tag{162}$$

which proves (47).

To show (48), we use Cauchy-Schwarz

$$\frac{\lambda_d(\beta)}{\sqrt{d}} = \frac{1}{2\sqrt{d}}\mathbb{E}\left[|\mathbf{V}_1^{(\beta)} - \mathbf{V}_2^{(\beta)}|\right] \tag{163}$$

$$\leq \frac{1}{2\sqrt{d}}\sqrt{\mathrm{Var}\left[|\mathbf{V}_1^{(\beta)} - \mathbf{V}_2^{(\beta)}|\right]} \tag{164}$$

$$= \frac{\sigma(\beta)}{\sqrt{2}}. \tag{165}$$

Finally, (47), (165) along with dominated convergence theorem yield

$$\lim_{d\to\infty} \frac{\Lambda_d}{\sqrt{d}} = \int_0^1 \lim_{d\to\infty} \frac{\lambda_d(\beta)}{\sqrt{d}} \mathrm{d}\beta = \int_0^1 \frac{\sigma(\beta)}{\sqrt{\pi}} \mathrm{d}\beta. \tag{166}$$

□

Appendix F Proof of scaling limit for reversible PT index process

We will prove Theorem 6.1 by using Theorem 17.25 from [Kal02].

Theorem F.1 (Trotter, Sova, Kurtz, Mackevičius). *Let $X, X^1, X^2, \dots$ be Feller processes defined on a state space S with generators $\mathcal{L}, \mathcal{L}_1, \mathcal{L}_2, \dots$ respectively. If D is a core for $\mathcal{L}$, then the following statements are equivalent:*

1. *If $f \in D$, there exist $f_N \in \mathcal{D}(\mathcal{L}_N)$ such that $\|f_N - f\|_\infty \to 0$ and $\|\mathcal{L}_N f_N - \mathcal{L} f\|_\infty \to 0$ as $N \to \infty$.*
2. *If $X^N(0)$ converges weakly to $X(0)$ in S, then X^N converges weakly to X in $D(\mathbb{R}_+, S)$.*

We will be applying Theorem F.1 with $\mathcal{L} = \mathcal{L}_W$ defined as $\mathcal{L}_W f = \frac{1}{2} f''$ for $f \in \mathcal{D}(\mathcal{L}_W)$ where

$$\mathcal{D}(\mathcal{L}_W) := \left\{ f \in C^2([0,1]) : f'(0) = f'(1) = 0 \right\}, \tag{167}$$

and $\mathcal{L}_N = \mathcal{L}_{W^N}$ defined in (68), which we recall here for the reader's sake

$$\mathcal{L}_{W^N} f(w) = \frac{N^2}{2} \sum_{\varepsilon \in \{\pm 1\}} \left(f(\Phi_\varepsilon^N(w)) - f(w) \right) s(\beta_w, \beta_{\Phi_\varepsilon^N(w)}), \quad w \in [0,1] \tag{168}$$

with $\Phi_\pm^N(w)$ defined in (62), (63) and $\beta_w = G(w)$. Also recall from the discussion just before (68) that $\mathcal{L}_{W^N}$ defines a Feller semigroup.

First notice that in [Kal02], the transition semi-group and generator of a Feller process taking values in a metric space S are defined on $C_0(S)$, the space of functions vanishing at infinity. Equivalently $f \in C_0(S)$ if and only for any $\delta > 0$ there exists a compact set $K \subset S$ such that for $x \notin K$, $|f(x)| < \delta$. In our case since $S = [0,1]$ is compact $C_0(S) = C(S)$, which justifies the definition of the generator $\mathcal{L}_W$ given above.

The Feller property of $\mathcal{L}_W$. Similarly $\mathcal{L}_W$ can be seen to define a Feller semigroup on $C([0,1])$ by the Hille-Yosida theorem (see [Kal02, Theorem 19.11]). Indeed the first condition is satisfied since any function $f \in C([0,1])$ can be uniformly approximated within $\epsilon > 0$ by a polynomial p_ϵ, that is a smooth function, by the Stone-Weirstrass theorem. We can further uniformly approximate p_ϵ within ϵ by a C^2 function $\hat{p}_\epsilon$ with vanishing derivatives at the endpoints. For example one can let, for a δ to be chosen later, $\hat{p}_\epsilon(x) = p_\epsilon(x)$ for $x \in (\delta, 1-\delta)$ and for $x \leq \delta$ set $\hat{p}_\epsilon(x) = \int_0^x \rho_\delta(y) p_\epsilon'(y) \mathrm{d}y + c$, where ρ_δ is a smooth, increasing transition function such that $\rho_\delta(x) = 0$ for $x < 0$, $\rho_\delta(x) = 1$ for $x > \delta$ [3] ; c is chosen so that $\hat{p}_\epsilon(x)$ is continuous at δ. A similar construction can be used for the right-endpoint. One can then check that indeed $\hat{p}_\epsilon \in C^2([0,1])$, $\hat{p}_\epsilon'(0) = \hat{p}_\epsilon'(1) = 0$ and that for δ small enough $\|\hat{p}_\epsilon - p_\epsilon\|_\infty < \epsilon$. The second condition of [Kal02, Theorem 19.11] also holds by [Har85, Corollary 5.2]. The third condition of [Kal02, Theorem 19.11] can also be easily seen to hold.

[3] e.g. let $\rho_\delta = \rho(x/\delta)$, $\rho(x) = g(x)/(g(x) + g(1-x))$ and $g(x) = \exp(-1/x)\mathbf{1}_{\{x>0\}}$

Now we can apply Theorem F.1 to prove Theorem 6.1. We only need to check the first condition of Theorem F.1. In this direction, first note that by definition $\Phi^N_\pm(w) = w \pm 1/N$ for $w \in [1/N, 1 - 1/N]$. Thus in this case using Taylor's theorem we have for $w^*_- \in [w-1/N, w]$ and $w^*_+ \in [w, w+1/N]$ that

$$\begin{aligned} f(\Phi^N_+(w)) - 2f(w) + f(\Phi^N_-(w)) &= f(w) + \frac{1}{N}f'(w) + \frac{1}{2N^2}f''(w^*_+) \\ &\quad + f(w) - \frac{1}{N}f'(w) + \frac{1}{2N^2}f''(w^*_-) - 2f(w) \qquad (169) \\ &= \frac{1}{2N^2}\left(f''(w^*_+) + f''(w^*_-)\right). \qquad (170) \end{aligned}$$

Since f'' is uniformly continuous it follows that as $N \to \infty$,

$$\sup_{w\in[0,1]} |f''(w^*_\pm) - f''(w)| = o(1), \qquad (171)$$

and therefore for $w \in [1/N, 1 - 1/N]$ we have

$$\sup_{w\in[0,1]} \left| f(\Phi^N_+(w)) - 2f(w) + f(\Phi^N_-(w)) - \frac{f''(w)}{N^2} \right| = o\left(\frac{1}{N^2}\right). \qquad (172)$$

When $w \in [0, 1/N)$ or $w \in (1 - 1/N, 1]$ we instead perform a Taylor expansion around 0 or 1 respectively. We only do the calculation in the first case, the other one being similar. Let $w \in [0, 1/N)$ in which case, since $f'(0) = 0$, for $w^*, w^*_-, w^*_+ \in [0, 2/N]$

$$\begin{aligned} f(\Phi^N_+(w)) - 2f(w) + f(\Phi^N_-(w)) &= f(0) + \Phi^N_+(w)f'(0) + \frac{1}{2}\left[\Phi^N_+(w)\right]^2 f''(w^*_+) \\ &\quad + f(0) + \Phi^N_-(w)f'(0) + \frac{1}{2}\left[\Phi^N_-(w)\right]^2 f''(w^*_-) \\ &\quad - 2f(0) - 2f'(0)w - 2\frac{f''(w^*)}{2}w^2 \qquad (173) \\ &= \frac{f''(0)}{2}\left\{\left[\Phi^N_+(w)\right]^2 + \left[\Phi^N_-(w)\right]^2 - 2w^2\right\} + o\left(N^{-2}\right) \qquad (174) \end{aligned}$$

where the error term is uniform in w and was obtained by combining the facts that f'' is uniformly continuous and that $|\Phi^N_\pm|, |w|/ \leq 2/N$. Finally notice that since $w \in [0, 1/N]$

$$\left[\Phi^N_+(w)\right]^2 + \left[\Phi^N_-(w)\right]^2 - 2w^2 = \left[w + \frac{1}{N}\right]^2 + \left[\frac{1}{N} - w\right]^2 - 2w^2 = \frac{2}{N^2}. \qquad (175)$$

Finally we will need the following weaker version of Theorem 4.1, whose proof we postpone until the end of the section.

Lemma F.2. *Suppose that $\pi(|V|), \pi_0(|V|) < \infty$. Then there exists a constant $C > 0$ such that*

$$\sup_{\beta} |s(\beta, \beta + \delta) - 1| \leq C\delta. \tag{176}$$

Using the above Lemma we can thus see that for some constant $C > 0$

$$\sup_{w \in [0,1]} \left| s\left(\beta_w, \beta_{\Phi^N_{\pm}(w)}\right) - 1 \right| \leq C \sup_{w} \left| G(w) - G\left(\Phi^N_{\pm}(w)\right) \right| \leq \frac{C\|G'\|_\infty}{N}, \tag{177}$$

and therefore

$$\mathcal{L}_N f(w) = \frac{N^2}{2} \sum_{\varepsilon \in \{\pm 1\}} \left(f(\Phi^N_\varepsilon(w)) - f(w) \right) s(\beta_w, \beta_{\Phi^N_\varepsilon(w)}) \tag{178}$$

$$= \frac{N^2}{2} \sum_{\varepsilon \in \{\pm 1\}} \left(f(\Phi^N_\varepsilon(w)) - f(w) \right) \left[1 + o(N^{-1}) \right] \tag{179}$$

$$= \frac{N^2}{2} \left(f(\Phi^N_+(w)) - 2f(w) + f(\Phi^N_-(w)) \right) \left[1 + o(N^{-1}) \right] = \frac{N^2}{2} \frac{f''(w)}{N^2} \left[1 + o(1) \right], \tag{180}$$

where the error term as shown above is uniform in w. Thus $\mathcal{L}_N f \to \mathcal{L} f$ uniformly.

Proof of Lemma F.2. Using the bound $0 \leq 1 - \exp(-x) \leq x$ for $x \geq 0$ we have

$$\begin{aligned}
&\left| s(\beta, \beta') - 1 \right| \\
&\leq \int\int \pi^{(\beta)}(\mathrm{d}x) \pi^{(\beta')}(\mathrm{d}y) \Big[1 - \exp\left(- \max\left\{ 0, (\beta' - \beta)[V(x) - V(y)] \right\} \right) \Big] && (181) \\
&\leq |\beta' - \beta| \int\int \pi^{(\beta)}(\mathrm{d}x) \pi^{(\beta')}(\mathrm{d}y) \max\left\{ 0, [V(x) - V(y)] \right\} \Big] && (182) \\
&\leq |\beta' - \beta| \int\int \pi^{(\beta)}(\mathrm{d}x) \pi^{(\beta')}(\mathrm{d}y) \big(|V(x)| + |V(y)| \big) \Big] \leq 2|\beta' - \beta| \sup_{\beta} \pi^{(\beta)}(|V|). \qquad \square && (183)
\end{aligned}$$

Appendix G Proof of scaling limit for non-reversible PT index process

We will prove Theorem 6.2 in a slightly round about way. We will define auxiliary processes $\{U^N(\cdot)\}$, $\{U(\cdot)\}$ living on the unit circle $\mathbb{S}^1 := \{z \in \mathbb{C} : |z| = 1\}$ along with a mapping $\phi : \mathbb{S}^1 \mapsto [0,1] \times \{\pm 1\}$ such that $Z^N = \phi(U^N)$ and $Z = \phi(U)$. We will first show that the law of U^N converges weakly to U.

Before defining the processes we point out that we will identify $\mathbb{S}^1$ with $[0, 2\pi)$ in the usual way by working in $\mod 2\pi$ arithmetic. Notice that in this way

$$C(\mathbb{S}^1) = \{ f \in C([0, 2\pi]) : f(0) = f(2\pi) \}. \tag{184}$$

The reason for working with these auxiliary processes is that we can now avoid working with PDMPs with boundaries, helping us to remove a layer of technicalities.

For any N we define $\Sigma^N : \mathbb{S}^1 \mapsto \mathbb{S}^1$ through $\Sigma^N(\theta) = \theta + 2\pi/N$. Consider then a continuous-time process U^N that jumps at the arrival times of a homogeneous Poisson process with rate N according to the kernel

$$Q^N(\theta, \mathrm{d}\theta') = s\left(\tilde{\beta}_\theta, \tilde{\beta}_{\Sigma^N(\theta)}\right)\delta_{\Sigma^N(\theta)}(\mathrm{d}\theta') + \left[1 - s\left(\tilde{\beta}_\theta, \tilde{\beta}_{\Sigma^N(\theta)}\right)\right]\delta_{2\pi-\theta}(\mathrm{d}\theta') \tag{185}$$

where

$$\tilde{\beta}_\theta = \begin{cases} G\left(\frac{\theta}{\pi}\right), & \theta \in [0, \pi), \\ G\left(\frac{2\pi-\theta}{\pi}\right), & \theta \in [\pi, 2\pi). \end{cases} \tag{186}$$

Define the map

$$\phi(\theta) = \begin{cases} \left(\frac{\theta}{\pi}, +1\right), & \theta \in [0, \pi), \\ \left(\frac{2\pi-\theta}{\pi}, -1\right), & \theta \in [\pi, 2\pi). \end{cases} \tag{187}$$

Essentially we think of the circle as comprising of two copies of $[0, 1]$ glued together at the end points. The top one is traversed in an increasing direction and the bottom one in a decreasing direction. When glued together and viewed as a circle these dynamics translate in a counter-clockwise rotation with occasional reflections w.r.t. the x-axis at the time of events. With this picture in mind it should be clear that $\phi(U^N) = Z^N$.

We also define the limiting process U as follows. First let

$$\tilde{\lambda}(\theta) = (\lambda \circ G)(\phi^1(\theta))G'(\phi^1(\theta)), \tag{188}$$

where $\phi^1(\theta)$ is the first coordinate of $\phi(\theta)$. Notice at this point that $\phi^1 : \mathbb{S}^1 \mapsto [0, 1]$ is continuous and satisfies $\phi^1(\theta) = \phi^1(-\theta)$ for any $\theta \in [0, 2\pi)$, whence we obtain that $\tilde{\lambda}(-\theta) = \tilde{\lambda}(\theta)$. Given $U(0) = \theta$, let T_1 be a random variable such that

$$\mathbb{P}[T_1 \geq t] = \exp\left\{-\int_0^t \tilde{\lambda}(\theta + s)\mathrm{d}s\right\}, \tag{189}$$

and define the process as $U(s) = \theta + s \mod 2\pi$ for all $s < T_1$ and set $U(T_1) = -U(T_1-) \mod 2\pi$. Iterating this procedure will define the $\mathbb{S}^1$-valued PDMP $\{U(\cdot)\}$. We first need the next lemma.

Lemma G.1. *Suppose V is integrable with respect to π_0 and π. The process U defined above is a Feller process, its infinitesimal generator is given by*

$$\mathcal{L}_U f(\theta) = f'(\theta) + \tilde{\lambda}(\theta)\left[f(2\pi - \theta) - f(\theta)\right], \tag{190}$$

with domain

$$\mathcal{D}(\mathcal{L}_U) = \{f \in C^1([0, \pi]) : f(0) = f(2\pi)\}, \tag{191}$$

and invariant measure $\mathrm{d}\theta/2\pi$.

Proof. First, note that since $\mathbb{S}^1$ is compact $C_0(\mathbb{S}^1) = C(\mathbb{S}^1)$ and thus to study the Feller process we consider the semi-group $\{P_U^t\}_t$ defined by the process U as acting on $C(\mathbb{S}^1)$. To prove the Feller property we can thus use [Dav93, Theorem 27.6]. Since there is no boundary in the definition of U the first assumption is automatically verified, $Qf(\theta) = f(-\theta) \in C\left(\mathbb{S}^1\right)$ for any continuous f. We also know that the rate $\tilde{\lambda}$ is bounded whereas by Proposition 4.2 and the fact that $G \in C^1[0,1]$ we know that $\tilde{\lambda}$ is also continuous. Therefore the third condition of [Dav93, Theorem 27.6] holds and thus U is Feller.

The infinitesimal generator will be defined on $\mathcal{D}(\mathcal{L}_U) \subseteq C(\mathbb{S}^1)$. The domain is defined as the class of functions $f \in C(\mathbb{S}^1)$ such that

$$g(\theta) = \lim_{h \to 0} \frac{1}{h} \left[P_U^t f(\theta) - f(\theta)\right] \in C(\mathbb{S}^1), \tag{192}$$

where the limit is uniform in θ. However by [BSW13, Theorem 1.33], we can also consider pointwise limits without enlarging the domain. Using the definition of U we then have for $\theta \in [0, 2\pi)$ that

$$\frac{1}{h}\mathbb{E}^\theta \left[f(U_h) - f(\theta)\right] = \frac{1}{h} f\left[(\theta + h) - f(\theta)\right] \mathbb{P}^\theta \left[T_1 \geq h\right] + \frac{1}{h}\mathbb{E}^\theta \left[\left(f\left(U_h\right) - f(\theta)\right) \mathbb{1}\left\{T_1 \leq h\right\}\right]. \tag{193}$$

Since for $x \geq 0$ we have $|\exp(-x) - 1 + x| \leq Cx^2$ for some constant $C > 0$, and using the continuity of $\tilde{\lambda}$ we can see that

$$\left|\exp\left\{-\int_0^h \tilde{\lambda}(\theta + s)\mathrm{d}s\right\} - 1 + \tilde{\lambda}(\theta)h\right| \leq Ch^2, \tag{194}$$

and thus

$$\frac{1}{h} f\left[(\theta + h) - f(\theta)\right] \mathbb{P}^\theta \left[T_1 \geq h\right] = \frac{1}{h} f\left[(\theta + h) - f(\theta)\right] \left(1 + o(h)\right). \tag{195}$$

In addition

$$\begin{aligned}
&\frac{1}{h}\mathbb{E}^\theta \left[\left(f\left(U_h\right) - f(\theta)\right) \mathbb{1}\left\{T_1 \leq h\right\}\right] \\
&= \frac{1}{h} \int_0^h \tilde{\lambda}(\theta + s) \exp\left\{-\int_0^s \tilde{\lambda}(\theta + r)\mathrm{d}r\right\} \mathrm{d}s \left[P_U^{h-s} Qf(\theta) - f(\theta)\right] &(196)\\
&\to \tilde{\lambda}(\theta) \left[Qf(\theta) - f(\theta)\right], &(197)
\end{aligned}$$

for any $f \in C(\mathbb{S}^1)$ by strong continuity of $\{P_U^t\}$ (Feller property) and continuity of $\tilde{\lambda}$.

Overall we thus have that $f \in \mathcal{D}(\mathcal{L}_U)$ if and only if

$$\begin{aligned}
\frac{1}{h}\mathbb{E}^\theta \left[f(U_h) - f(\theta)\right] &= \frac{f(\theta + h) - f(\theta)}{h} + \tilde{\lambda}(\theta) \left[Qf(\theta) - f(\theta)\right] + o(1) &(198)\\
&\to g(\theta) \in C(\mathbb{S}^1), &(199)
\end{aligned}$$

which is clearly equivalent to $f \in C^1(\mathbb{S}^1)$.

Finally to see that $\mathrm{d}\theta/2\pi$ is invariant, having identified the domain we can easily check that for any $f \in C(\mathbb{S}^1)$ we have

$$\int \mathrm{d}\theta P_U^t f(\theta) = \int_{s=0}^{t} \int \mathrm{d}\theta \mathcal{L}_U P_U^s f(\theta) \mathrm{d}\theta \mathrm{d}s. \tag{200}$$

Since $f \in \mathcal{D}(\mathcal{L}_U)$ we have that $P_U^s g \in \mathcal{D}(\mathcal{L}_U)$. Since for any $g \in \mathcal{D}(\mathcal{L}_U)$ we have

$$\int \mathrm{d}\theta \mathcal{L}_U f(\theta) = \int_{\theta=0}^{2\pi} f'(\theta)\mathrm{d}\theta + \int_{\theta=0}^{2\pi} \tilde{\lambda}(\theta) f(Q(\theta))\mathrm{d}\theta - \int_{\theta=0}^{2\pi} \tilde{\lambda}(\theta) f(\theta)\mathrm{d}\theta \tag{201}$$

$$= f(2\pi) - f(0) + \int_{\theta=0}^{2\pi} \tilde{\lambda}(\theta) f(Q(\theta))\mathrm{d}\theta. \tag{202}$$

□

Proposition G.2. *Suppose $U^N(0)$ converges weakly to $U(0)$, then U^N converges weakly to U in $D(\mathbb{R}_+, [0,1])$.*

Proof. We will once again use Theorem F.1. The generator of U_N is given by

$$\mathcal{L}_U^N f(\theta) = N\left[f(\theta + 1/N) - f(\theta)\right] s\left(\tilde{\beta}_\theta, \tilde{\beta}_{\Sigma^N(\theta)}\right) + N\left[f(-\theta) - f(\theta)\right] r\left(\tilde{\beta}_\theta, \tilde{\beta}_{\Sigma^N(\theta)}\right). \tag{203}$$

We will consider the two terms separately. To this end notice that by (44), the boundedness of λ and the fact that $G \in C^1[0,1]$

$$\left|1 - s\left(\tilde{\beta}_\theta, \tilde{\beta}_{\Sigma^N(\theta)}\right)\right| \leq \frac{C}{N}, \tag{204}$$

for some $C > 0$. Thus, using the mean value theorem, for each $\theta \in [0, 2\pi)$, there exists $g_N(\theta) \in [\theta, \theta + 1/N]$ such that

$$N\left[f(\theta + 1/N) - f(\theta)\right] s\left(\tilde{\beta}_\theta, \tilde{\beta}_{\Sigma^N(\theta)}\right) = f'\left(g_N(\theta)\right)\left(1 + O(1/N)\right) = f'\left(\theta\right)\left((1 + o(1)\right), \tag{205}$$

where the errors are uniformly bounded and to obtain the second equality above we have used the fact that $|g_N(\theta) - \theta| \leq 1/N$ and that f' is uniformly continuous, being continuous on a compact set.

Overall we can see that as $N \to \infty$

$$\sup_\theta \left|N\left[f(\theta + 1/N) - f(\theta)\right] s\left(\tilde{\beta}_\theta, \tilde{\beta}_{\Sigma^N(\theta)}\right) - f'(\theta)\right| \to 0. \tag{206}$$

Next, using (44) we have that

$$r\left(\tilde{\beta}_\theta, \tilde{\beta}_{\Sigma^N(\theta)}\right) = \tilde{\lambda}(\theta)\frac{1}{N} + o(N^{-1}), \tag{207}$$

where the error is uniform in θ, whence we easily conclude that

$$N\left[f(-\theta)-f(\theta)\right] r\left(\tilde{\beta}_{\theta}, \tilde{\beta}_{\Sigma^{N}(\theta)}\right) \rightarrow \tilde{\lambda}(\theta)\left[Qf(\theta)-f(\theta)\right], \tag{208}$$

uniformly in θ. □

Proof of Theorem 6.2 Now we are ready to prove the main result of this section. Notice that $Z^N(\cdot) = \phi\left(U^N(\cdot)\right)$ and $Z(\cdot) = \phi\left(U(\cdot)\right)$.

From Proposition G.2 we know that the finite dimensional distributions of U_N converge to those of U. If ϕ were continuous we could conclude using the continuous mapping theorem. Since it is not continuous at the points $\{0,1\}$, we will be using [Bil13, Theorem 2.7]. We have to check that the law of the limiting process, that is the law of $\{U(\cdot)\}$ places zero mass on finite dimensional distributions that hit $\{0,1\}$, that is for $n \in \mathbb{N}$ and $0 < t_1 < \cdots < t_n$ we want

$$\mathbb{P}\left[U(t_i) \in \{0,1\} \text{ for some } i \in \{1,\ldots,n\}\right] = 0, \tag{209}$$

when $U(0)$ is initialized according to $\mathrm{d}\theta/2\pi$. But the above follows from the fact that $\mathbb{P}[U(t_i) \in \{0,1\}] = 0$, by stationarity when $U(0)$ is initialised uniformly on $\mathbb{S}^1$.

Relative compactness of $\{Z_N(\cdot)\}_N$ can be easily seen to follow from the compact containment condition [EK09, Remark 3.7.3]. This combined with convergence of the finite dimensional distributions of Z^N to those of Z concludes the proof.